SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(90th Edition)

JUNE 2020

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 **Ninetieth** 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 **June, 2020**. 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 new 2-tuple linguistic approach for unbalanced linguistic term sets, ***6.Tanya Malhotra*** and ***3.Anjana Gupta***, Mathematics, DTU

2. A new realization scheme for dynamic PFSCL style, ***6.Ranjana Sivaram***, Kirti Gupta and ***3.Neeta Pandey***, Electronics, DTU

3. A novel fault classification-based fault-tolerant control for two degree of freedom helicopter systems, ***6.Rupam Singh*** and ***3.Bharat Bhushan***, Electrical, DTU

4. A Study on Stakeholder Management of an Engineering Project, ***8.Pranav Salhan***, Mechanical, DTU

5. A systematic literature review on empirical studies towards prediction of software maintainability, ***3.Ruchika Malhotra*** and ***3.Kusum Lata***, Computer Science, DTU

6. A systematic review of developments in the 2-tuple linguistic model and its applications in decision analysis, ***6.Tanya Malhotra*** and ***3.Anjana Gupta***, Mathematics, DTU

7. Advanced Hybrid Grid with Anti-Theft Alert and Fault Detection, ***3.J N Rai***, ***8.Naveen Verma***, ***8.Anubhav Gaba*** and ***8.Piyush Meena***, Electrical, DTU

8. Aerobic biodegradation of high explosive hexahydro-1,3,5- trinitro-1,3,5-triazine by Janibacter cremeus isolated from contaminated soil, ***6.Anchita Kalsi***, S. Mary Celin and ***3.Jai Gopal Sharma***, Biotechnoloyg, DTU

9. An experimental investigation on spray, performance and emission of hydrotreated waste cooking oil blends in an agricultural engine, ***3.Naveen Kumar***, ***6.Ankit Sonthalia***, ***6.Mukul Tomar*** and ***6.Rashi Koul***, Mechanical, DTU

10. An integrated approach for landslide susceptibility–vulnerability–risk assessment of building infrastructures in hilly regions of India, Aditi Singh, ***3.Shilpa Pal*** and D. P. Kanungo, Civil, DTU

11. An overview of the opportunities and challenges in sustaining the energy industry in Afghanistan, ***6.Shambalid Ahady***, ***3.Nirendra Dev*** and ***3.Anubha Mandal***, Civil, DTU

12. Analog Inverse Filters Using OTAs, ***7.Jyoti Srivastava***, ***3.Ram Bhagat*** and ***3.Pragati Kumar***, Electrical, DTU

13. Analysing the Effective Psychological State of Students using Facial Features, ***3.Shailender Kumar***, ***8.Deepanshu Varshney***, ***8.Gaurav Dhawan*** and ***8.Harsh Jalutharia***, Computer Science, DTU

14. Au-Nanoplasmonics-Mediated Surface Plasmon-Enhanced GaN Nanostructured UV Photodetectors, ***6.Lalit Goswami***, Neha Aggarwal, Shibin Krishna, Manjri Singh, Pargam Vashishtha, Surinder Pal Singh, Sudhir Husale, ***3.Rajeshwari Pandey*** and Govind Gupta, Electronics, DTU

15. Autonomous Model Predictive Controlled Smart Inverter with Proactive Grid Fault Ride-Through Capability, Mitchell Easley, ***6.1.Sarthak Jain***, Mohammad Shadmand and Haitham Abu-Rub, Electrical, DTU

16. Comparative Analysis of Deep Convolutional Generative Adversarial Network and Conditional Generative Adversarial Network using Hand Written Digits, ***6.Prabhat***, ***6.Nishant*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

17. Controlling Mouse Motions Using Eye Tracking Using Computer Vision, ***8.Kshitij Meena***, ***8.Manish Kumar*** and ***8.Mohit Jangra***, Computer, Science, DTU

18. Convolutional Neural Network based Automated Attendance System by using Facial Recognition Domain, ***3.Shailender Kumar***, ***8.Dhruv Kathpalia***, ***8.Dipen Singh*** and ***8.Mandeep Vats***, Computer Science, DTU

19. Design and Analysis of a Parallel Link Manipulator for Heavy Lifting in Warehouses, ***8.Rohit Kumar***, ***8.Shubham Kumar***, ***8.Siddharth*** and ***3.Vikas Rastogi***, Mechanical, DTU

20. Design and Implementation of Biquad Filter for Shunt Compensation under Normal and Distorted Grid Conditions, ***3.Ankita Arora*** and ***3.Alka Singh***, Electrical, DTU

21. Design and Implementation of Microstrip Low Pass Filter at 2.4 GHz, ***3.Cheshta Kwatra***, ***8.Kapil Gupta***, ***8.Kshitij Chhabra*** and ***3.Dr. Priyanka Jain***, Electronics, DTU

22. Design and Testing of Frequency Adaptive Zero- Crossing Detector as a Synchronizing Technique, ***6.Hemant Saxena***, ***3.Alka Singh*** and ***3.Jitendra Nath Rai***, Electrical, DTU

23. Dispersion Compensation in a 64-Channel Dense Wavelength Division Multiplexing Network, ***8.Anshul Dagar***, ***8Akhil Dixit***, ***8.Akhil Kumar*** and ***3.Dr. Gurjeet Kaur***, Electronics, DTU

24. Dynamic Traffic Management System Using Neural Network based IoT System, ***3.R.K. Yadav***, ***8.Rishabh Jain***, ***8.Sarvjit Yadav*** and ***8.Satyam Bansal***, Computer Science, DTU

25. Dynamics of Predator–Prey System with Migrating Species and Disease in Prey Population, Charu Arora and ***3.Vivek Kumar***, Mathematics, DTU

26. Effect of reinforcement architecture on the macroscopic mechanical properties of fiberous polymer composites: A review, ***6.Furkan Ahmad*** and ***3.N Yuvaraj*** and Pramendra K. Bajpai, Mechanical, DTU

27. Efficient Power Theft Detection using Smart Meter Data in Advanced Metering Infrastructure, ***8.Salil Manocha***, ***8.Vivek Bansal***, ***8.Ishan Kaushal*** and ***3.Dr. Aruna Bhat***, Computer Science, DTU

28. End-to-End Neural Network for Shadow Detection using Combined Color Spaces, ***3.Shailender Kumar***, ***8.Dhruv Ramdev***, ***8.Himanshu Aggarwal*** and ***8.Himanshu Choudhary***, Computer Science, DTU

29. Exploiting bad-smells and object-oriented characteristics to prioritize classes for refactoring, ***3.Ruchika Malhotra*** and ***6.Priya Singh***, Computer, DTU

30. Exploring Providencia rettgeri for application 1 to eco-friendly 2 paper based microbial fuel cell, ***6.Sharda Nara***, ***6.Rahul Kandpal***, ***6.Veeru Jaiswal***, ***6.Shine Augustine***, ***8.1.Srishti Wahie***, ***3.Jai Gopal Sharma***, Ryusuke Takeuchi, Shigeori Takenaka, ***3.Bansi Dhar Malhotra***, Biotechnology, DTU

31. Fabrication Techniques for Carbon Nanotubes Based ECG Electrodes: A Review, ***6.Bani Gandhi*** and ***3.N.S. Raghava***, Electronics, DTU

32. FACTS Devices Impact on Congestion Mitigation of Power System, ***8.Jitender Kumar*** and ***3.Narendra Kumar***, Electrical, DTU

33. Formation of Self-Assembled Monolayer and Characterization of AA7075-T6/B4C Nano-ceramic surface composite using Friction Stir Processing, ***3.Ravi Butola***, ***3.Qasim Murtaza*** and ***3.Ranganath M Singari***, Mechanical, DTU

34. Fractional Order Capacitors Based Filters Using Three OTAs, ***7.Gopal Singh***, ***3.Garima*** and ***3.Pragati Kumar***, Electrical, DTU

35. Fractional order Oscillator Using OTAs, ***8.Vikramjeet Singh*** and ***3.Pragati Kumar***, Electrical, DTU

36. Fuzzy-Based Model for Predicting Strength of Geogrid-Reinforced Subgrade Soil with Optimal Depth of Geogrid Reinforcement, ***8.Meenakshi Singh***, ***3.Ashutosh Trivedi*** and Sanjay Kumar Shukla, Civil, DTU

37. Genetic CoCoA++: Genetic Algorithm based Congestion Control in CoAP, ***3.Rajesh Kumar Yadav***, ***8.Navpreet Singh*** and ***8.Prince Piyush***, Computer Science, DTU

38. Global-best optimization of ANN trained by PSO using the nonextensive cross-entropy with Gaussian gain, ***3.Seba Susan***, ***8.Rohit Ranjan***, ***8.Udyant Taluja***, ***8.Shivang Rai*** and ***8.Pranav Agarwal***, IT, DTU

39. Identification of novel class I and class IIb histone deacetylase inhibitor for Alzheimer's disease therapeutics, ***6.Rohan Gupta***, ***6.Rashmi K. Ambasta*** and ***3.Pravir Kumar***, Biotechnology, DTU

40. Image Forgery Detection and Deep Learning Techniques: A Review, ***3.Ritu Agarwal***[1], ***8.Deepak Khudaniya***[1], ***8.Abhinav Gupta***[1] and ***8.Khyati Grover***[2], [1]IT and [2]Computer, DTU

41. Instant Access Memory Design based on an FPGA, ***8.Saharsh Aggarwal***, ***3.Deva Nand***, ***8.Megha Rayer*** and ***8.Saket Kr. Jha***, Electronics, DTU

42. Linking TQM Critical Success Factors to strategic goal: Impact on Organizational Performance, ***7.Rakesh Kumar*** and ***3.R. S. Mishra***, Mechanical, DTU

43. Localization of User in an Indoor Environment Using Machine Learning Classification Models, ***8.Manish Kumar***, ***8.Manish Rawat*** and ***3.Prashant Giridhar Shambharkar***, Computer Science, DTU

44. Mechanical and tribological properties of AA7075-T6 metal matrix composite reinforced with ceramic particles and aloevera ash via Friction stir processing, ***6.Lakshay Tyagi***[1], ***3.Ravi Butola***[1] and ***8.Avinash Kumar Jha***[2], [1]Mechanical and [2]Civil, DTU

45. Modeling and Development of an Autonomous Underwater Vehicle ARYA for Object Recognition, ***8.Aditya Natu***[1], ***8.1.Anuj Badhwar***[2], ***8.Vipul Garg***[1], ***8.Upasana Biswas***[3], ***8.Deepanshu Bansal***[3] and ***8.Aakash Kumar***[1], [1]Mechanical, [2]Mathematics and [3]Electrical, DTU

46. Multi-Agent based Stable Clustering and Collision Detection In VANET, ***8.Nikhil Pratap Singh Rathore***, ***8.Abhinav Yadav*** and ***3.Anamika Chauhan***, IT, DTU

47. Nanoengineered Conductive Polyaniline Enabled Sensor for Sensitive Humidity Detection***, 8.1.Samar S. Sandhu***[1], ***Saurabh Kumar***[1], ***7.Shine Augustine***[1], ***8.Udiptya Saha***[1], ***6.Kamal Arora***[2], Sayan Bayan, Samit K. Ray, ***3.Nitin K. Puri***[2] and ***3.Bansi D. Malhotra***[1], [1]Biotechnology and [2]Physics, DTU

48. Natural Language Processing based Visual Question Answering Efficient: an EfficientDet Approach, ***3.Rahul Gupta***, ***8.Parikshit Hooda***, ***8.Sanjeev*** and ***8.Nikhil Kumar Chikkara***, Computer Science, DTU

49. NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media, ***8.Ashima Suvarna*** and ***8.Grusha Bhalla***, Computer, DTU

50. Novel approach for multimodal feature fusion to generate cancelable biometric, ***8.Keshav Gupta***[1], Gurjit Singh Walia and ***3.Kapil Sharma***[2], [1]Computer and [2]IT, DTU

51. Novel approach with nature-inspired and ensemble techniques for optimal text classification, ***6.Anshu Khurana[1]*** and ***3.Om Prakash Verma[2]***, [1]Computer and [2]Electronics;, DTU

52. ON A SPECIAL TYPE OF NON-MA-MINDA FUNCTION, ***3.S. SIVAPRASAD KUMAR*** AND *6.**SHAGUN BANGA***, Mathematics, DTU

53. Optimization of Neural Network Architecture using Lions Optimization Algorithm, ***8.Kaptan Malik***, ***8.Kanishk Srivastava*** and ***8.Krishna Sharma***, Computer Science, DTU

54. Optimization of the process parameters for hydrotreating used cooking oil by the Taguchi method and Fuzzy logic, ***3.Naveen Kumar***, Ankit Sonthalia and Rashi Koul, Mechanical, DTU

55. Parametric Study of Plasma Characteristics and Carbon Nanofibers Growth in PECVD System: Numerical Modeling, ***6.Ravi Gupta*** and ***3.Suresh C. Sharma***, Physics, DTU

56. Plasma bubble evolution in laser wakefield acceleration in a petawatt regime, ***6.Monika Yadav***, Devki Nandan Gupta, ***3.Suresh C Sharma*** and Hyyong Suk, Physics, DTU

57. Real-time Pedestrian Tracking based on Deep Features, ***3.Geetanjali Bhola***, ***8.Akhil Kathuria***, ***8.Deepak Kumar*** and ***8.Chandan Das***, IT, DTU

58. REFINING YOLOV4 FOR VEHICLE DETECTION, ***8.Pooja Mahto***, ***8.Priyamm Garg***, ***8.Pranav Seth*** and ***3.J Panda***, Electronics, DTU

59. Reliability of Sub-20 nm Black Phosphorus Trench (BP-T) MOSFET in High-Temperature Harsh Environment, ***8.Ajay Kumar[1]***, Neha Gupta and ***3.Rishu Chaujar[2]***, [1]Electrical and [2]Physics, DTU

60. Silicon–air batteries: progress, applications and challenges, ***8.1.Rishabh Bansal***, ***8.1.Prajwal Menon*** and 3.R. C. Sharma, Chemistry, DTU

61. Survey Paper:Introduction to Generative Adversarial Networks in Speech Imitation, ***8.Rajashree Guha***, ***8.Vaibhav Sharma***, ***8.Yatin Yadav*** and ***3.Dr. Aruna Bhat***, Computer, DTU

62. Testing Rotating Regular Metrics as Candidates for Astrophysical Black Holes, Rahul Kumar, ***7.Amit Kumar***, and Sushant G. Ghosh, Physics, DTU

63. The prospective application of a graphene/MoS2 heterostructure in Si-HIT solar cells for higher efficiency, Chandra Kamal Borah, ***3.Pawan K. Tyagi*** and Sanjeev Kumar, Physics, DTU

64. TransSketchNet: Attention-based Sketch Recognition using Transformers, ***8.Gaurav Jain***, ***8.Shivang Chopra***, ***8.Suransh Chopra*** and ***3.Dr. Anil Singh Parihar***, Computer, DTU

65. Understanding and Analyzing Methods of Character Identification in Movies, ***3.Prashant Giridhar Shambharkar***, ***8.Anubhav Gupta***, ***8.Anirudh Gupta***, ***8.Anshuman*** and Dr. M . N. Doja, Computer Science, DTU

1. *Chancellor*

2. *Pro Vice Chancellor*

2.1. *Ex Pro Vice Chancellor*

3. *Faculty*

3.1. *Ex Faculty*

4. *Teaching-cum-Research Fellow*

4.1. *Asst. Librarian*

5. *Alumni*

6. *Research Scholar*

6.1. *Ex Research Scholar*

7. *PG Scholar*

7.1. *Ex PG Scholar*

8. *Undergraduate Student*

8.1. *Ex Undergraduate Student*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

A new 2-tuple linguistic approach for unbalanced linguistic term sets

Tanya Malhotra, Anjana Gupta

***Abstract*—Several real-world problems employ linguistic-based approaches to handle qualitative data. The set of linguistic terms that are utilized in the problems are mostly alleged to be symmetrically distributed. However, with the advent of time, as the complexity of the problem increases, the equidistant linguistic term set seems improper. Consequently, in such cases, experts often prefer to use the set of the unbalanced linguistic term to direct the appraisal for the problems. In the present work, we tend to propose a method that is newly designed to deal with a set of unbalanced linguistic terms. In this direction, we initially propose an algorithm to represent unbalanced linguistic information via a multiplicative linguistic label set that has a global inconsistent linguistic term distribution. Furthermore, in light of the Herrera and Martínez, "2-tuple linguistic model," we develop a novel 2-tuple approach for the unbalanced linguistic set, which is based on the notion of minimum distance measure. Finally, to validate the proposed model in the physical realm and to demonstrate the functioning of the method, a numerical example is being elucidated. The proposed methodology seeks to indicate a reduction in the computation time and also enhances the decision-makers' evaluations.**

***Index Terms*—Unbalanced linguistic term sets, 2-tuple linguistic variable, multiplicative linguistic term set, aggregation operators.**

I. Introduction

THE societal development engenders complex situations, consequently increases indeterminacy within the problems. In general, there may exist some intricate problems entailing quantitative aspects that are well handled through precise numerical values. However, some cases entail qualitative aspects as well that are more arduous to be managed through precise and exact values. Therefore, to handle such problems adequately, Zadeh [2]–[4] initiated a fuzzy linguistic-based approach whereby qualitative information is represented via linguistic variables. The utility of this approach aims to provide more logical and precise results in the problems encapsulating uncertainty, which is of non-probabilistic nature. Hence, on account of successful results, the fuzzy linguistic-based approach has an application of broad-spectrum, which are witnessed in distinct areas suchlike information retrieval system [5]–[9], [13], web quality [10], [11], aggregation operators [14], [16], [19], and so forth.

The presence of qualitative information in a decision problem insinuates the appropriateness for a linguistic computational models (LCMs) [12], [15], [20], [21]. According to the existing majored literature several LCMs has become widespread, for instance LCMs that are based on membership functions [22], [23], ordinal scales [16], [24], [26], type-2 fuzzy sets [32], [33], granular computing [34], [35]. Followed by the 2-tuple LCM and its extensions [15], [17], [18], [20], [27].

Since a determination of the membership function is difficult in some of the practical applications, as a consequence, symbolic models like Herrera and Martínez [17], 2-tuple model has gained considerable attention. This is because of the competence of a 2-tuple model to interpret and evaluate the results in a highly accurate and straightforward manner. Moreover, it provides a way out to handle linguistic information in a continuous range in place of a discrete one without any information loss. Most of these LCMs and their respective enhancements rely upon the term set having a uniform distribution of the linguistic terms. Such type of linguistic term set is termed as a balanced linguistic term set (BLTS). The most commonly referred BLTSs are $\mathrm{LT}_1=\{s_i \mid i=0,1,\ldots,g\}$ and $\mathrm{LT}_2=\{s_i \mid i=-g,\ldots,0,\ldots,g\}$, where each linguistic term is equidistant about their respective central cardinal. Apparently, these two BLTS follow the rule of monotonicity stating that any two linguistic terms, namely, s_α and s_β belonging to either set LT_1 or LT_2 are comparable, i.e, $s_\alpha \geq s_\beta$ iff $\alpha \geq \beta$.

However, there exist problems where information managed with linguistic assessments require linguistic variables appraised in an unevenly distributed linguistic term sets. Such type of sets is termed as unbalanced linguistic term sets (ULTSs) where the distribution of the terms are uneven. It is precisely viewed in the problems that linguistic information distributed unevenly appears due to necessity of assessing the preferences of the linguistic terms having higher granularity on either side of the central term, or maybe in light of the explicit nature of the linguistic variables which are defined in the problems for computations. In compliance with the existing literature, there are primarily 2 classes of ULTSs. In the first class of ULTSs, the distribution of the linguistic terms are uneven such that the cardinality of the terms on one side of the central term is higher than the other and corresponding distance between the consecutive terms are not equal (namely, ULTS of Ist class, see Fig. 1). Meanwhile, in the second class of ULTSs, an equal number of the linguistic terms are distributed on both the sides of the central term with unequal spacings (namely, ULTS of the second class, see Fig. 2).

Recent years mark the remarkable progress in the LCMs developed to deal with the ULTSs. Several methodologies were introduced in the literature, which ideally concerns with the presentations and applications of both the classes of

T. Malhotra and A. Gupta are with the Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India (e-mail: malhotra.tanya92@gmail.com; anjanagupta@dce.ac.in)

Manuscript received XXXX; revised XXXX. This work was supported by Delhi Technological University (DTU/IRD/619/2019/2177).

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

ULTSs. Herrera et al. [15] proposed a methodology to address the linguistic information assessed in ULTS. In that paper, authors have described a representation model to allocate semantics to the linguistic terms, which are diffused unevenly. The semantics of a linguistic term is capable of expressing each linguistic terms via parametric membership function by utilizing the concept of linguistic hierarchy. Lastly, in the paper, they have developed a computational model that uses the existing 2-tuple model [17] as its basis for ULTSs to carry out processes of computing with words (CWWs) without information loss. Wang and Hao [20] presented a new proportional 2-tuple model having "symbolic proportion" as its basis. In the proposed method, it is mentioned that linguistic information is expressed in a 2-tuple format, which composes two proportional linguistic terms. Hence, the proportional 2-tuple model deals with ULTS avoiding any information loss. Continuing further with the study of proportional 2-tuple model, Wang and Hao [36] bring forth a unifying link between the framework of Lawry [37] and the traditional 2-tuple framework [17], in parallel with the CWWs, processes for the line of reasoning with linguistic syllogisms mathematically feasible. Zou et al. [38] proposed a novel linguistic aggregation operator elicited from the model of traditional 2-tuple [17] and linguistic hierarchies [39] to express unbalanced linguistic values. Jiang et al. [40] presented the linguistic proportional 2-tuple power average operator aggregating the linguistic values of ULTSs.

Abchir and Truck [41] broadened the traditional 2-tuple model by introducing a different kind of fuzzy partition for ULTS. Bartczuk et al. [49] presented a novel approach, which is a modification of the traditional 2-tuple linguistic model for handling ULTS. In that paper, linguistic information is represented by an extended-term set which composes a pair that carries a linguistic term along with a value representing correction factor. This correction factor is used to describe the relocation of the term corresponding to its respective position in a term set distributed evenly.

Dong et al. [27] developed a new interpretation of the 2-tuple linguistic model constructed to be an extended version of it primarily based on the notion of numerical scale. Moreover, guided by the notion of the "transitive calibration matrix," along with its consistency index Dong et al. [42] introduced a novel scheme for generating individual numerical scales within the analytic hierarchy process (AHP). Further, Dong et al. [28], [43] proposed a connection link between the two different models, i.e., LCM based on the linguistic hierarchical concept [15] and the model of numerical scale [27], and lastly demonstrated the equivalence of these models to deal with ULTSs. Continuing with the noticeable advancements in the field of 2-tuples LCMs developed for addressing ULTSs, Dong, and Herrera-Viedma [44] and Dong et al. [45] introduced interval numerical scales. Wang et al. [47] proposed a methodology of normalized numerical scaling to form BLTSs or ULTSs into the unique interval $[0,1]$. M. Cai et al. [29] developed a new LCM based on symbolic models. Z. Pei and Li Zheng [31], constructed a series of a normal distribution which is based upon unbalanced linguistic scale sets and additionally, formulated a novel approach for defining unbalanced linguistic information in a format of 3-tuple. Further, M. Cai and Z. Gong [30], redefined the concept of ULTS and provided the graph for the representation model.

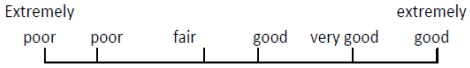


Fig. 1: Unbalanced linguistic set type I

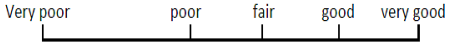


Fig. 2: Unbalanced linguistic set type II

Li et al. [48] proposed personalized individual semantics model by virtue of an interval type numerical scale and the linguistic model of 2-tuple and later on solved problems of group decision making with a process of consensus-reaching. Considering the wide-ranging decision problems existing in an uncertain decision environment providing exact values for symbolic proportions is not always easy for decision-makers. Moreover, there exist situations where linguistic distribution assessments are required to assess in ULTS. Therefore, to fix these issues, Dong et al. [46] developed the notion of unbalanced linguistic distribution assessments in the light of interval symbolic proportions and further presented the application of the proposed methodology.

However, concerning the above LCMs, it is envisioned that the representation of the ULTSs is made by using a set of linguistic terms distributed uniformly. As a consequence, the literal meaning of a ULTS is missing. Thereby in this research, we are taking the theory of ULTS a step forward. In this study, we aim to develop a methodology to represent and manage ULTS. The representation of the ULTS is given by the multiplicative unbalanced linguistic scale set initially proposed by Tang et al. [1]. Secondly, we present a novel 2-tuple model for unbalanced linguistic information based upon a notion of minimum distance measures.

Our method is advantageous because of two reasons; firstly, in our approach, a multiplicative linguistic scale set is utilized to represent unbalanced linguistic information where the terms are unevenly distributed. In the existing literature of ULTS, it is mentioned that an evenly distributed linguistic scale set is utilized to express the original unbalanced information. As a result, it leads to the absence of the existence of a formal definition of unbalanced linguistic information within the problem. Henceforth, we able to subjugate the limitation by employing a multiplicative linguistic scale set to substitute the given unbalanced information. In this way, throughout the procedure, a formal definition of ULTS is preserved. Secondly, our proposed method has a more straightforward design. As a consequence, it is computationally less expensive and less complicated in comparison to the other LCMs, suchlike, LCM

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

based on linguistic hierarchical concept [15] and numerical scale [28], and so forth.

The organization of the paper is presented as follows: Section II, provides a review of the three 2-tuple linguistic models (i.e., traditional 2-tuple model, unbalanced linguistic computational model, and numerical scale model). Then, we present a representation algorithm for ULTS and further propose a new 2-tuple unbalanced linguistic computational model (ULCM) to address ULTS in section III. Next, for the sake of completeness in section IV, we define basic aggregation operators based on a new 2-tuple ULCM. To validate the proposed methodology and illustrate the practical relevance of the method in real-world, 2 numerical examples related to the ULTS of first and second types are presented in section V. Section VI gives an analysis and comparison of the proposed model with other existing works. Lastly, Section VII covers the concluding remarks for the model presented in the paper.

II. Preliminaries

The current section briefly presents a review of the traditional 2-tuple linguistic model primarily based upon the concept of "symbolic translation." Afterward, we review the basics of the unbalanced linguistic computational model (ULCM) and the numerical scale (NS) model.

A. The 2-tuple linguistic model

The introduction of a 2-tuple model leads to the improvement in several linguistic computational techniques. The basic idea behind the design of this model is to increase the authenticity and assist the processes of "computing with words" (CWWs) by considering the continuity of the linguistic domain. Moreover, its efficacy in upgrading the linguistic solving decision-making problems and its easiness in handling linguistic information involved in a complicated framework and so forth has made it the most beneficial model among other traditional models for managing uncertainty without information distortion [17]. In this model linguistic information is represented in a 2-tuple format i.e., $(s_i, \alpha) \in LT \times [-0.5, 0.5)$ where s_i represents a linguistic term belonging to a set LT and α represents a numerical value.

Definition 1 [17]: Let $LT = \{s_i \mid i = 0 \text{ to } g\}$ be the balanced linguistic term set which is both finite and totally ordered having cardinality $g + 1$. Let β be the numerical value assessed in $[0, g]$. Let $i = \text{round}(\beta)$, $i \in \{0, \ldots, g\}$ and $\alpha = \beta - i \in [-0.5, 0.5)$ be two values. Then α is called a "*symbolic translation*".

This 2-tuple LCM is described by using a transformation function, transforming numeric values into 2-tuples as defined above to take forward the further computational processes.

Definition 2 [17]: Let $LT = \{s_i \mid i = 0 \text{ to } g\}$ be a finite balanced linguistic term set and $\beta \in [0, g]$ indicates a numerical value.Then, the 2-tuple linguistic value representing the identical numerical value to β is computed in the following aspect:

$$\Delta : [0, g] \to LT \times [-0.5, 0.5)$$

$$\Delta(\beta) = (s_{\text{round}(\beta)}, \beta - i)$$

where round(.) is the general round operation while $s_{\text{round}(\beta)}$ represents the nearest label to β, and α represents the "symbolic translation".

Remark 1: [17] Δ mapping as defined above is one-one and onto hence invertible, therefore there exist $\Delta^{-1} : \overline{LT} \equiv LT \times [-0.5, 0.5) \to [0, g]$ which returns an identical numerical value $\beta \in [0, g] \subset \mathbb{R}$ corresponding to equivalent 2-tuple such that $\Delta^{-1}(\ell_i, \alpha) = \beta = i + \alpha$.

Remark 2: A 2-tuple linguistic term is obtained from a linguistic term ℓ_i by coupling a value 0 as "symbolic translation," i.e., $\ell_i \in LT \Rightarrow (\ell_i, 0) \in \overline{LT}$.

Definition 3 [17]: The comparison of linguistic information presented by the 2-tuple is carried in accordance to an ordering based on the lexicographic approach. Let (s_i, α_i) and (s_j, α_j) be two 2-tuples with each one representing a counting of information as given below:

- If $i < j$ then $(s_i, \alpha_i) < (s_j, \alpha_j)$.
- If $i = j$ then
  1) If $\alpha_i = \alpha_j \implies (s_i, \alpha_i) = (s_j, \alpha_j)$;
  2) If $\alpha_i < \alpha_j \implies (s_i, \alpha_i) < (s_j, \alpha_j)$;
  3) If $\alpha_i > \alpha_j \implies (s_i, \alpha_i) > (s_j, \alpha_j)$.

Definition 4 [17]: Let $LT = \{s_i \mid i = 0 \text{ to } g\}$ be the finite linguistic term set with cardinality $g + 1$, then Negation operator of a 2-tuple is determined as:

$$\text{neg}((s_i, \alpha)) = \Delta(g - (\Delta^{-1}(s_i, \alpha))).$$

B. Unbalanced linguistic computational model

The proposal of a ULCM centered on the notion of linguistic hierarchy was introduced in [15]. The model was designed in a way to address ULTS, whose terms are not distributed symmetrically. In the methodology, Herrera et al. [15] take into consideration a ULTS, namely, LT as a set of minimum, maximum, and central label, and the distribution of the leftover labels on either side of the central term, i.e., $LT = LT_L \cup LT_C \cup LT_R$ such that:

- LT_L represents *left lateral set* containing all the terms except the central one.
- LT_C represents *central set* containing just one central term.
- LT_R represents *right lateral set* containing all the labels higher than the central label.

Herrera et al. [15] accustomed the conception of linguistic hierarchy to acquire linguistic representations of unbalanced linguistic values in the format of 2-tuple. A linguistic hierarchy (LH) [39] is determined as a set of levels, where each level is represented by a linguistic term set $LT^{n(t)} = \left\{s_0^{n(t)}, s_1^{n(t)}, \ldots, s_{n(t)-1}^{n(t)}\right\}$. Each level associated with a linguistic hierarchy is constructed as $l(t, n(t))$, where t is a numeric value indicating the hierarchy level and $n(t)$ indicates the linguistic term set cardinality corresponding to the level t. Here, we underline that the linguistic terms utilized in the linguistic hierarchical process are given by triangular-shaped membership functions distributed uniformly in $[0, 1]$.

Usually, we witness that the linguistic term set $LT^{n(t+1)}$ of level $t + 1$ can be easily acquired from the previous term set $LT^{n(t)}$ as

$$l(t, n(t)) \to l(t + 1, 2 \cdot n(t) - 1).$$

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

In LH [39], transformation functions are defined between labels from distinct levels to perform CWWs processes in a multi granular linguistic context avoiding any information loss.

Definition 5 [39]: For any level t and t', the transformation function is determined as $\mathrm{TF}_{t'}^{t} : l(t, n(t)) \rightarrow l(t', n(t'))$ such that

$$\mathrm{TF}_{t'}^{t}(s_i^{n(t)}, \alpha^{n(t)}) = \Delta\left(\frac{\Delta_t^{-1}(s_i^{n(t)}, \alpha^{n(t)}) \cdot (n(t') - 1)}{n(t) - 1}\right). \quad (1)$$

For the most part, any level could be chosen in the computational model characterized for the LH to bring together the multi-granular linguistic data in the model characterized for the LH. However, for the sake of convenience and generalization, we refer 't_m' as the maximum level in the LH, i.e., $l(t_m, n(t_m)) = LT^{n(t_m)} = \left\{s_i^{n(t_m)} \mid i = 0, \ldots, n(t_m) - 1\right\}$.

Next, we present below the details of the transformation process to transform any 2-tuple linguistic variable (s_i, α) into the term in $LH = \cup_t l(t, n(t))$:

(i) Representation in the LH: The representation algorithm proposed in the paper utilizes the concept of a LH to address the unbalanced terms. As a consequence, the initial step to bring about CWWs processes to convert the unbalanced terms lying in a set LT into their corresponding terms in the LH, $s_k^{n(t)} \in LH = \cup_t l(t, n(t))$ by virtue of the transformation function ψ given as:

$$\psi : \overline{LT} \rightarrow LH(\overline{LT}) \quad (2)$$

such that $\psi(s_i, \alpha) = (s_{I(i)}^{G(i)}, \alpha)$, for all $(s_i, \alpha) \in \overline{LT}$. Here, I and G represents a function assigning every unbalanced label $s_i \in S$ index of the label symbolizing it in LH and the cardinality of label set of LH where it is presented, respectively.

(ii) Computational phase: The computational model characterized for LH helps in accomplishing the process of CWWs. At first, use the Eq. 1 to transform $(s_{I(i)}^{G(i)}, \alpha)$ (for $i = 0, \ldots, g$) into linguistic 2-tuples, referred as $(s_{I'(i)}^{n(t_m)}, \lambda') \in \overline{LT^{n(t_m)}}$. Without loss of generality, if $G(i) = n(t')$, then Eq. 3 is obtained. Further, use Herrera and Martínez model over $\overline{LT^{n(t_m)}}$ to acquire desired result $(s_r^{n(t_m)}, \lambda_r') \in \overline{LT^{n(t_m)}}$.

$$(s_j^{n(t_m)}, \lambda') = TF_{t_m}^{t'}(s_{I(i)}^{G(i)}, \alpha) \quad (3)$$

(iii) Re-translation process: A process of re-translation is performed to change the obtained result $(s_r^{n(t_m)}, \lambda_r') \in \overline{LT^{n(t_m)}}$ back into the desired unbalanced term in $\overline{LT}$, by utilizing the following function, ψ^{-l}, i.e.,

$$\psi^{-1} : LH(\overline{LT}) \rightarrow \overline{LT}. \quad (4)$$

such that $\psi^{-1}(s_r^{n(t_m)}, \lambda_r') = (s_{result}, \lambda_{result}) \in \overline{LT}$.

The detailed explanation of the methodology to deal with ULTSs is given in the corresponding paper [15].

C. The Numerical scale model

Dong et al. [27] extended the current version of the Herrera and Martínez model to manage ULTSs by laying down the idea of a numerical scale.

Definition 6 [27]: Let $LT = \{s_i \mid i = 0 \text{ to } g\}$ be a finite set of linguistic terms, and $\mathbb{R}$ has the usual meaning. The function: $NS : LT \rightarrow \mathbb{R}$ is defined as a numerical scale of LT, and $NS(s_i)$ is termed as numerical index of s_i. If the function NS is defined as strictly monotone increasing, then NS is termed as an ordered numerical scale.

Definition 7 [27]: Numerical scale function $\overline{NS} : \overline{LT} \rightarrow \mathbb{R}$ is defined in the following aspect:

$$\overline{NS}((s_i, \alpha)) = \begin{cases} NS(s_i) + \alpha \times ((NS(s_{i+1})) - NS(s_i)), & \alpha \geq 0; \\ NS(s_i) + \alpha \times ((NS(s_i)) - NS(s_{i-1})), & \alpha < 0. \end{cases}$$

For simplicity of the notation, NS is used in lieu of $\overline{NS}$.

Proposition 1. [27] By defining $NS(s_i) = i$ for $i \in \{0, \ldots, g\}$ yields the Herrera and Martínez model.

Usually, linear triangular or trapezoidal membership functions are being used to express semantics of the terms belonging to a linguistic term set. However, in the proportional model introduced by Wang and Hao [20] the semantics of linguistic labels are expressed by using symmetrical trapezoidal fuzzy numbers. If the semantics of s_i is defined by $T[b_i - \sigma_i, b_i, c_i, c_i + \sigma_i]$, in the Wang and Hao model [20], the *canonical characteristic values* (CCVs) of s_i is $\frac{b_i + c_i}{2}$, i.e., $CCV(s_i) = \frac{b_i + c_i}{2}$.

Proposition 2. [27] By defining $NS(s_i) = CCV(s_i)$ for $i \in \{0, \ldots, g\}$ yields the Wang and Hao model.

Propositions 1 and 2 provide a strong relation of the NS to the model of Herrera and Martínez [17] and Wang and Hao [20].

The proof and detailed explanation of the model is provided in the paper [27].

III. Methodology for unbalanced linguistic term set

In the present section, we briefly explain the methodology for managing unbalanced linguistic information. For this, we initially design a novel approach to represent semantics to the linguistic terms that belong to a ULTS and finally propose the new 2-tuple ULCM.

A. Representation algorithm for unbalanced linguistic set

In literature, Tang et al. [1] defined a multiplicative linguistic scale set distributed throughout inconsistently as $LS = \left\{t_{a^{(-m)}}, t_{a^{-(m-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\right\}$ where $a \in \mathbb{R}$ and $m \in \mathbb{Z}^+$.

Here, before we step forward to present a representation algorithm for ULTS, we state the following definition.

Definition 8: Let $LS = \left\{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\right\}$ where, $a > 1$, $n, m \in \mathbb{Z}^+$ be a multiplicative ULTS with cardinality $n + m + 1$ such that n and m may or may not be equal. Then we call the term t_{a^i} as a plausible value for a linguistic variable. The set LS is termed as an ordered set if the following properties are satisfied:

(P1): The set LS is ordered: $t_{a^i} \geq t_{a^j}$ for $i \geq j$;
(P2): A maximization operator: $\max(t_{a^i}, t_{a^j}) = t_{a^i}$ for $i \geq j$;
(P3): A minimization operator: $\min(t_{a^i}, t_{a^j}) = t_{a^i}$ for $i \leq j$.

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

We now proceed to present the representation algorithm in order to manage unevenly distributed unbalanced linguistic information:

step 1: Partition the ULTS LS into the following three subsets, i.e, $LS = LS_{left} \cup LS_{center} \cup LS_{right}$:
- LS_{left} defines the set of linguistic terms lying left to the central term.
- LS_{center} defines the set of central term.
- LS_{right} defines the set of linguistic terms lying right to the central term.

step 2: Set $\#(LS_{left}) = n$ and $\#(LS_{right}) = m$, where $\#(LS_{left})$, $\#(LS_{right})$ are cardinalities of the set LS_{left}, LS_{right} respectively.

step 3: Representation of the central linguistic label set.

Identify the central linguistic term such that $t_{i_c} \leftarrow t_{a^0}$. For the central linguistic term, $\text{neg}(t_{a^i}) = t_{a^i}$ for $i \in \{-n, -(n-1), \ldots, 0, 1, \ldots, m\}$

step 4: Represent the left linguistic label set. The decision rule of representation of set LS_{left} is given below.

Representation of set LS_{left}

$i_L \leftarrow \#(LS_{left})$;
while $i_L \geq 1$ **do**
  $t_{i_L} = t_{a^{-i_L}}$;
  $i_L = i_L - 1$.
end

step 5: Represent the right linguistic label set. The decision rule of representation of set LS_{right} is given below

Representation of set LS_{right}

$i_R \leftarrow 1$;
while $i_R \leq \#(LS_{right})$ **do**
  $t_{i_R} = t_{a^{i_R}}$;
  $i_R = i_R + 1$.
end

Example 1. To present the working of the proposed representation method, we consider the same example of a grading system evaluation taken from [15] (see Fig. 3).

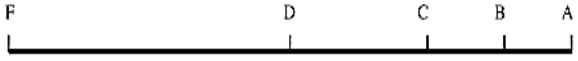


Fig. 3: Grading system evaluation [15].

We express $S = \{F, D, C, B, A\}$ with $\#(S) = 5$, in the form of set $\{t_{a^{-1}}, t_{a^0}, t_{a^1}, t_{a^2}, t_{a^3}\}$. According to the representation method, partition the set S, i.e, $S_{left} = \{F\}$, $S_{right} = \{C, B, A\}$ and $S_{center} = \{D\}$. Hence, $n = 1$ and $m = 3$. After implementing the decision rule, we obtain
$F \leftarrow t_{a^{-1}}$, $D \leftarrow t_{a^0}$, $C \leftarrow t_{a^1}$, $B \leftarrow t_{a^2}$, $A \leftarrow t_{a^3}$.

After developing a methodology to represent semantics to the terms of ULTS, we proceed next to develop a novel 2-tuple based linguistic model for ULTS to accomplish the process of CWW in a precise way without any information loss.

B. 2-tuple representation for unbalanced linguistic set

In the present subsection, we formalize the new 2-tuple based computational model for the ULTS based on a notion of minimum distance measure. In this framework, the so-called translation function defined converts a numerical value into a 2-tuple represented as (t_λ, γ) where $t_\lambda \in LS$ and $\gamma \in [-0.5, 0.5]$ is referred as a symbolic translation and vice-versa. The 2-tuple representation discussed in the present section is defined on the similar line of the representation proffered in the paper [17], however, the computation of λ and γ are different. After demarcating the translation function, we present different computational operators suchlike negation operator and comparison operator to manage the unbalanced linguistic information.

Let $LS = \{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\}$ where, $a > 1$ and $n, m \in \mathbb{Z}^+$ and n, m may or may not be equal is a multiplicative ULTS whose terms are non uniformly and inconsistently distributed. The numerical value $\beta \in [a^{-n}, a^m]$, where n and m may or may not be equal and $\beta \notin \{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\}$ is obtained as a value from a symbolic method aggregating unbalanced linguistic information. Subsequently an approximation function is utilized to represent the index of the outcome in an unbalanced set LS.

We shall now move a step forward to develop a novel approach to represent an unbalanced linguistic information in terms of 2-tuples (t_λ, γ), where $t_\lambda \in LS$ and $\gamma \in [-0.5, 0.5]$:

1) t_λ expresses the linguistic label which is the center of information where λ is of the form a^i for $i \in \{-n, -(n-1), \ldots, 0, \ldots, m\}$;
2) γ expresses a numerical value computed as a ratio of $(\beta - \lambda)$ and sum of d_L & $|d_R|$ where λ is the index of closest label and d_L & d_R represent nearest left and right distance of λ from β respectively.

Next, we define a translation function giving 2-tuple linguistic value encrypted in the space $LS \times [-0.5, 0.5]$ from a numerical value $\beta \in [a^{-n}, a^m]$.

Definition 9: Let $LS = \{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\}$ where, $a > 1$ and $n, m \in \mathbb{Z}^+$ and n, m may or may not be equal, be a multiplicative unbalanced linguistic term set having $n + m + 1$ as cardinality. Let $\beta \in [a^{-n}, a^m]$ be a numerical value that supports symbolic aggregation result. Then the translation function that translates a numeric value β into a corresponding 2-tuple unbalanced linguistic value is shown below:

$$\Omega : [a^{-n}, a^m] \longrightarrow LS \times [-0.5, 0.5]$$

$$\Omega(\beta) = (t_\lambda, \gamma)$$

$$\text{in which } \begin{cases} \lambda = a^{i_L} \text{ or } \lambda = a^{i_R} & \text{for } i_L, i_R \in \{-n, \ldots, 0, \ldots, m\}; \\ \gamma = \frac{\beta - \lambda}{d_L + |d_R|} & \text{for } \gamma \in [-0.5, 0.5]. \end{cases}$$

The decision rule for the computation of values λ and γ defined above in the definition is given in algorithm 1.

Note 1: We can infer from the Algorithm 1 that the "λ" value is evaluated by virtue of the nearest distance measure from

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

Algorithm 1 Computation of λ and γ

Input: $LS=\{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\}$ & $\beta \in [a^{-n}, a^m]$, where n and m may or may not be equal and $a>1$.
Begin:
$i_L \leftarrow \arg\min_i \{\beta - a^i \mid \beta - a^i \geq 0, i \in \{-n, \ldots, 0, \ldots, m\}\}$;
$i_R \leftarrow \arg\max_i \{\beta - a^i \mid \beta - a^i \leq 0, i \in \{-n, \ldots, 0, \ldots, m\}\}$.
Then:
$\lambda_L \leftarrow a^{i_L}$, and $\lambda_R \leftarrow a^{i_R}$, such that $\beta \in (\lambda_L, \lambda_R)$.
Next, compute:
$d_L \leftarrow \beta - \lambda_L$, and $d_R \leftarrow \beta - \lambda_R$.
Define: $d \leftarrow \min(d_L, |d_R|)$;
if $d = d_L$ **then**
  $\lambda \leftarrow \lambda_L$;
  $\gamma \leftarrow \frac{\beta-\lambda}{d_L+|d_R|}$.
else
  $\lambda \leftarrow \lambda_R$;
  $\gamma \leftarrow \frac{\beta-\lambda}{d_L+|d_R|}$.
end if

the numerical value "β" whereas "γ" is termed as a value of "symbolic translation." However, according to the algorithm 1, the range of the value "γ" belongs to $[-0.5, 0.5]$, and the proof is given in the appendix (See Proposition 4 in appendix).

Note 2: We can easily obtain our model from a numerical scale model [27] by defining $NS(t_{a^i}) = a^i$ (See proposition 5 in appendix).

We provide the following example for a better understanding of the aforementioned definition.

Consider a multiplicative unbalanced linguistic term set $LS = \left\{t_{\frac{1}{8}}, t_{\frac{1}{4}}, t_{\frac{1}{2}}, t_1, t_2, t_4, t_8\right\}$, here we have taken $a=2$ (not specific) and $n=m=3$. Suppose $\beta=5.8 \in [\frac{1}{8}, 8]$ be the value obtained as a result of symbolic aggregation. Then, the representation of β in terms of 2-tuple is given as

$$\Omega(5.8) = (t_4, 0.45).$$

The graphical representation of the 2-tuple computation calculated above is given in the Fig. 4.

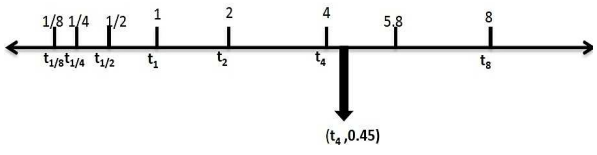


Fig. 4: Graphical representation of 2-tuple for ULTS

Proposition 3. Let $LS = \{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\}$ be an ULTS and let (t_λ, γ) be a 2-tuple corresponding to numerical value $\beta \in [a^{-n}, a^m] \subset \mathbb{R}$. Then, there always exist a function Ω^{-1} which returns an equivalent numerical information from a corresponding 2-tuple linguistic information.

Proof: It is trivial to show that a translation function Ω is a bijective. We consider the following function:

$$\Omega^{-1} : LS \times [-0.5, 0.5] \longrightarrow [a^{-n}, a^m]$$
$$\Omega^{-1}(t_\lambda, \gamma) = \gamma \times (d_L + |d_R|) + \lambda = \beta.$$

■

Remark 3: A linguistic term $t_\alpha \in LS$ can be easily converted into a 2-tuple by adding a value 0 as minimum distance measure, i.e., $t_\alpha \in LS \implies (t_\lambda, 0)$.

Remark 4: It is noted that if $t_\alpha \in LS$ then t_α is termed as actual unbalanced linguistic term and meanwhile if $t_\alpha \notin LS$ then t_α is termed as virtual unbalanced linguistic term.

Herrera and Martínez [17] has proposed a comparison of a 2-tuple linguistic information based upon classical lexicographic ordering. Analogously, we present a comparison of 2-tuple linguistic information for ULTS in compliance with the existing lexicographic approach.

Comparison of 2-tuples: Let $(t_{\lambda_i}, \gamma_i)$, $(t_{\lambda_j}, \gamma_j)$ be represented as two 2-tuples such that each of them is expressing a counting of information. Then the 2-tuples are compared as follows:

- If $\lambda_i < \lambda_j$ then $(t_{\lambda_i}, \gamma_i) < (t_{\lambda_j}, \gamma_j)$;
- If $\lambda_i = \lambda_j$ then
  1) If $\gamma_i = \gamma_j \implies (t_{\lambda_i}, \gamma_i) = (t_{\lambda_j}, \gamma_j)$;
  2) If $\gamma_i < \gamma_j \implies (t_{\lambda_i}, \gamma_i) < (t_{\lambda_j}, \gamma_j)$;
  3) If $\gamma_i > \gamma_j \implies (t_{\lambda_i}, \gamma_i) > (t_{\lambda_j}, \gamma_j)$.

Negation operator of 2-tuples: Let $LS = \{t_{a^{(-n)}}, \ldots, t_{a^0}, t_{a^1}, \ldots, t_{a^m}\}$ be an ULTS with cardinality $n+m+1$. Then we define negation operator for a 2-tuple as:

$$\text{neg}(t_{a^i}, \gamma) = (\text{neg}(t_{a^i}), -\gamma)$$

where the existence of a $\text{neg}(t_{a^i})$ for an unbalanced linguistic term in all the plausible cases is given in Algorithm 2.

After, discussing the existence of negation operator for ULTS, we wish here, to point out three observations:

- In first place, we consider the case, $n \neq m$, for instance take $n=2$ and $m=3$ such that $LS=\{t_{a^{-2}}, t_{a^{-1}}, t_{a^0}, t_{a^1}, t_{a^2}, t_{a^3}\}$. Then according to the Algorithm 2, it is stated that $\text{neg}(t_{a^1})=t_{a^{-1}}$, $\text{neg}(t_{a^2})=t_{a^{-2}}$, $\text{neg}(t_{a^3})=t_{a^0}=\text{neg}(t_{a^0})$. Henceforth, we may claim that for the case $n \neq m$, $\text{neg}(\text{neg}(t_{a^i})) \neq t_{a^i}$ for $i \in \{-2, -1, 0, 1, 2, 3\}$.
- The case $n \neq m$ can be converted it into the case $n = m$, by adding dummy linguistic variables, 'd_{a^i}' for $i \in \{-n, \ldots, 0, \ldots, m\}$ to the set LS. Hence, proceed likewise as described in Algorithm 2.
- The dummy linguistic variables which are added to the set LS are given 'no' weights. As a result, in several MCDM (Multi-criteria decision making) problems or linguistic game-theoretic models, we cannot be able to obtain an optimal solution for such a case. This validates our proposed algorithm 2.

When we have expressed all the unbalanced linguistic terms and have also defined the linguistic information in terms of 2-tuples, it, therefore, necessitates us to aggregate such information. Henceforth, the following section discusses the aggregating operators for a new 2-tuple ULCM.

IV. Aggregation operators for 2-tuple model for unbalanced linguistic term set

Several aggregation operators have been developed in literature [17] for the traditional 2-tuple model based upon "*symbolic translation*." The aggregating operation carried out by these

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

Algorithm 2 Computation algorithm for negation operator for ULTS

Input: $LS=\{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\}$ where n and m may or may not be equal and $a>1$.
$LS=LS_{left} \cup LS_{center} \cup LS_{right}$.
Begin:

```
#(LS_left) ← n, and, #(LS_right) ← m Then:
if n=m then
    for i←0 to n do
        neg(t_{a^i}) ← t_{a^{-i}};
    end for
end if
if n≠m then
    if n<m then
        for i←0 to m do
            if i≤n then
                neg(t_{a^i}) ← t_{a^{-i}};
            else
                neg(t_{a^i}) ← t_{a^{m-i}};
            end if
        end for
    else
        for i←0 to n do
            if i≤m then
                neg(t_{a^{(-i)}}) ← t_{a^i}; else
                end
                neg(t_{a^{(-i)}}) ← t_{a^{i-n}};
            end if
        end for
    end if
end if
```

so-called aggregation operators are without any information loss.

Since the primary focus of the aggregation operator is to mitigate value set into a single one that gives an overview of the inputs in a particular way. As a consequence, it plays a significant role in decision problems and further amalgamation of information. However, the information to be considered is not always balanced. Sometimes, unbalanced linguistic information is also involved in the problem. Henceforth, in the present section, we extend these operators for the developed 2-tuple ULCM.

The transformation function Ω and Ω^{-1} discussed in the previous section helps in converting a numerical value into a linguistic value given in the form of 2-tuples and vice-versa.

(A) *Arithmetic mean (AM) operator*
Definition 10: Let $X=\{(t_{\lambda_1}, \gamma_1), (t_{\lambda_2}, \gamma_2), \ldots, (t_{\lambda_k}, \gamma_k)\}$ be the set of unbalanced linguistic 2-tuples, the arithmetic mean operator $\overline{X}^{AM}$ is defined as:

$$\overline{X}^{AM} = \Omega\left(\frac{\sum_{i=1}^{k} \Omega^{-1}(t_{\lambda_i}, \gamma_i)}{k}\right).$$

The functions Ω and Ω^{-1} have their usual meaning, as discussed in the previous section. Moreover, the $\overline{X}^{AM}$ estimated for the set X is incurred in a precise way without any information loss.

(B) *Weighted average (WA) operator*
In the case of weighted average, considering the nature of the variable x, distinct values say x_i have different importance. As a consequence, each value x_i is associated with a weight w_i, which signifies the variable's nature.
In literature, the equivalent operator exists for the Herrera and Martinez' [17] 2-tuple model. Likewise, we can extend the theory concerning the WA operator for a ULCM based on 2-tuples.
Definition 11: Let $X=\{(t_{\lambda_1}, \gamma_1), (t_{\lambda_2}, \gamma_2), \ldots, (t_{\lambda_k}, \gamma_k)\}$ be the set of unbalanced linguistic 2-tuples and let $W=\{w_i \mid i=1, 2, \ldots, k\}$ be the set of weights associated with 2-tuple linguistic set, then the WA operator $\overline{X}^{WA}$ for the set X is defined as:

$$\overline{X}^{WA} = \Omega\left(\frac{\sum_{i=1}^{k} \Omega^{-1}(t_{\lambda_i}, \gamma_i) \cdot w_i}{\sum_{i=1}^{k} w_i}\right)$$

(C) *Ordered weighted average (OWA) operator*
In literature, the weighted aggregation operator is initiated by Yager [25]. In that operator, the weights which are given are not connected with a predetermined value. Instead, it is provided that the weights are connected to a fixed position.
The OWA operator X^{OWA} for dealing with unbalanced linguistic 2-tuples are defined as:
Definition 12: Let $X=\{(t_{\lambda_1}, \gamma_1), (t_{\lambda_2}, \gamma_2), \ldots, (t_{\lambda_k}, \gamma_k)\}$ be the set of unbalanced linguistic 2-tuples and let $W=\{w_i \mid i=1, 2, \ldots, k, \sum_{i=1}^{k} w_i = 1 \text{ and } w_i \in [0,1]\}$ be the set of associated weights. Then the 2-tuple OWA operator X^{OWA} is computed as:

$$X^{OWA} = \Omega\left(\sum_{j=1}^{k} w_j \cdot \beta_j^*\right).$$

where the numerical value β_j^* represents the largest of j^{th} value among all the β_i values.

V. Numerical illustration

In the present section, we consider 2 numerical examples to exhibit the efficacy of the proposed 2-tuple ULCM in the physical realm.

In the example, II, we consider the usual problem of evaluating students' expertise from variant tests to obtain a global evaluation, which is a similar example discussed in Herrera et al. [15]. The ULTS utilized here is of type I, i.e., a set where the distribution of the linguistic terms on the left and right side of the central term is unequal ($m \neq n$).

The example III is related to the stock market, which is an organized financial market and plays a significant role in raising the economic infrastructure. In the current scenario, it is envisioned that most of the car manufacturing companies in India are undergoing huge losses. Consequently, affecting Indian stock market saliently. Henceforth, the current study provides a comparative analysis of the stock prices for different car manufacturing companies in the financial year 2018. The ULTS utilized here is of type II, i.e., several linguistic

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

terms distributed on the left and right sides of the central term are equal ($m=n$).

A. Example II

Suppose a teacher wants to acquire a maximum evaluation of his pupils who have appeared in 6 different exams equally essential. The exams are completed, and the evaluations are being done using the grading system. The grading scale is considered to be a ULTS, LS, which is given in example 1. Also, the respective grades obtained by students in each exam are given in table I.

	E1	E2	E3	E4	E5	E6
J. Smith	(D,0)	(C,0)	(B,0)	(C,0)	(C,0)	(C,0)
M. Grant	(A,0)	(D,0)	(D,0)	(C,0)	(B,0)	(A,0)

TABLE I: 2-tuple linguistic assessment of unbalanced set [15].

Herrera et al [15] has solved the example to obtain global evaluations for 2 students based on the linguistic hierarchical concept for unbalanced linguistic information. The detailed calculations and required steps are provided in the paper [15].

Now, for the similar mission we solve this example by using our methodology. According to our approach, we initially apply the semantic representation algorithm for unbalanced linguistic information. The descriptive information about the working of this algorithm is given in example 1. From the analysis, we infer that $a=2$ to be the best value for the base. However, one may take any value of 'a.' Hence, we obtain $LS_{left}:\{F\leftarrow t_{2^{-1}}\}$, $LS_{center}:\{D\leftarrow t_{2^0}\}$, $LS_{right}:\{C\leftarrow t_{2^1}, B\leftarrow t_{2^2}, A\leftarrow t_{2^3}\}$.

Next, to compute the global evaluations of the students by applying arithmetic mean operator defined in section IV, we proceed as follows:

$$\begin{aligned}\bar{x}_{J.Smith}&=\Omega\left(\frac{1}{6}\left(\Omega^{-1}(D,0)+4\times\Omega^{-1}(C,0)+\Omega^{-1}(B,0)\right)\right)\\&=\Omega\left(\frac{1}{6}\left(\Omega^{-1}(t_{2^0},0)+4\times\Omega^{-1}(t_{2^1},0)+\Omega^{-1}(t_{2^2},0)\right)\right)\\&=\Omega(2.167)=(C,0.0835).\end{aligned}$$

$$\begin{aligned}\bar{x}_{M.Grant}&=\Omega\left(\frac{1}{6}\left(2\times\Omega^{-1}(A,0)+2\times\Omega^{-1}(D,0)+\Omega^{-1}(B,0)\right.\right.\\&\quad\left.\left.+\Omega^{-1}(C,0)\right)\right)\\&=\Omega\left(\frac{1}{6}\left(2\times\Omega^{-1}(t_{2^3},0)+2\times\Omega^{-1}(t_{2^0},0)+\Omega^{-1}(t_{2^2},0)\right.\right.\\&\quad\left.\left.+\Omega^{-1}(2^1,0)\right)\right)\\&=\Omega(4.00)=(B,0).\end{aligned}$$

Hence, from the above analysis, we observe that M. Grant has scored maximum grades in comparison to J. Smith.

The ranking order of the two students so obtained is similar to the order obtained in the paper [15]. Henceforth, we conclude that our proposed model is consistent with the model of Herrera et al. [15].

B. Example III

In this illustration, we have considered a fundamental real-life decision problem where the computations are made using the type II ULTS discussed in the previous section.

Suppose an investor wants to invest his/her money into a particular car manufacturing company by taking into consideration that which company has a high market value as well as high share growth in the financial year 2018 (FY2018). The alternatives of the Indian car companies that an investor is looking up in particular are:

x_1	x_2	x_3
"Maruti Suzuki India"	"Tata Motors"	"Mahindra & Mahindra"

The factors which affect the share prices of the aforementioned car companies are broadly classified into three classes: (1) Macroeconomic class, (2) Microeconomic class, and (3) Class of linguistic factors. The list of these factors is given in table II.

Macro economic factors	Micro economic factors	Class of linguistic factors
M_1: Gross domestic product (GDP)(%)	m_1: Book value per share (In INR)	ℓ_1: Investors confidence
M_2: Interest rate (%)	m_2: Depreciation (In Rs Cr)	ℓ_2: Risk of Investment
M_3: Inflation rate (%)	m_3: Dividend per share (In Rs)	
M_4: Index of industrial production (IIP)(%)	m_4: Net profit margin (%)	

TABLE II: List of factors affecting stock prices

Each factor equally affects the price of the stock market. The evaluations of the stock prices for the auto company affected by these factors are assessed in the unbalanced linguistic scale set LS={VL, L, M, H, VH} where VL: Very low, L: Low, M: Moderate, H: High, VH: Very high. The table III provides the values of different independent factors that affect the stock prices of the auto company for the FY2018. The data given in the table III is taken from https://www.moneycontrol.commoneycontrol and https://www.rbi.org.inReserve bank of India.

Apart from the listed quantitative factors, however, there exist certain qualitative factors that also affect the stock prices. So, to incorporate these factors and further obtain accurate results without any information loss, we apply a 2-tuple model for managing unbalanced linguistic information.

Based on the values recorded in table III along with the qualitative factors, we obtain the fluctuations of the stock prices for the auto companies. The variations incurred in the values of the stock prices for the companies due to these factors for the FY2018 are evaluated in the unbalanced linguistic scale set.

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

Factors	x_1	x_2	x_3	
GDP				6.8%
Interest rate				6.4%
Inflation rate				3.48%
IIP				31.4%
Book value per share	8,648.198	294.65	837.08	
Depreciation (In Rs Crs.)	727.80	770.55	431.80	
Increase/Decrease in stocks (In Rs Crs.)	−331.98	−231	98.14	
Net Profit/Loss for the period (In Rs Crs.)	1896.78	353.62	1251.58	
Dividend per share (In Rs)	80.00	0.00	7.50	
Net profit margin (%)	9.68	−1.75	8.98	

TABLE III: Values of the Independent factors affecting stock prices for the FY2018.

	Factors									
	M_1	M_2	M_3	M_4	m_1	m_2	m_3	m_4	ℓ_1	ℓ_2
x_1	VH	L	VL	VH	VH	VL	VH	VH	VH	VH
x_2	L	H	L	H	L	VH	L	VL	L	H
x_3	H	L	H	L	L	H	L	VH	H	L

TABLE IV: Evaluations of the factors for each company

Next, we apply our technique to deal with unbalanced linguistic information to obtain which company has a higher growth in share market for the FY2018. For this, we follow the similar steps as done in previous example. Henceforth, according to the semantic representation algorithm we obtain $LS = \{t_{2^{-2}} : \text{VL}, t_{2^{-1}} : \text{L}, t_{2^0} : \text{M}, t_{2^1} : \text{H}, t_{2^2} : \text{VH}\}$. Lastly, we convert the unbalanced linguistic evaluations obtained for each alternatives into 2-tuple linguistic representation for further computations.

Now, these 2-tuples are aggregated by utilizing a 2-tuple arithmetic mean operator for ULTS, thereby obtaining the collective values given in the table V.

x_1	x_2	x_3
(H, 0.45)	(M, 0.275)	(M, 0.45)

TABLE V: Aggregated results

Hence, the solution set of alternatives we obtain as $\{x_1\}$, i.e., "Maruti Suzuki India" has a high share growth for the FY2018. Therefore an investor would likely invest his money into the same company.

VI. Analysis and Comparison

The present section aims to provide a comparative study of our model with the existing works in an effort to validate the efficiency of our proposal.

So far, computational models existing in the literature [46] deal with ULTSs proposed by Herrera et al. [15]. The authors in that paper considered ULTS, $LT = LT_L \cup LT_C \cup LT_R$ where LT_L represents a minimum label set, LT_R represents a maximum label, and LT_C represents a central label set. As a consequence, the methodology depicted could be able to address only the first class of ULTS (See figure 1). In the sequent, methods portrayed in the paper [20], [27], [42] could be able to manage the second class of ULTS (See figure 2). However, in contrast to the existing methodologies, our proposal is well-suited to handle both the classes of ULTS. In our proposal, the ULTS LS is partitioned as $LS = LS_{left} \cup LS_{center} \cup LS_{right}$ where it is no longer necessary for the set LS_{left} and LS_{right} to be represented as minimum, and maximum label set respectively. The details of the representation algorithm are given in sec. III. Henceforth, we contemplate our proposed definition to handle any real-world situations involving unbalanced linguistic information.

The example V-A given in the previous section is from the paper [15]. According to the results obtained in the corresponding paper 'M. Grant' scored maximum in comparison to 'J. Smith' which is analogous to the results shown in our paper. Therefore, the comparison of our results exhibits the consistency of our proposal.

Furthermore, it is seen in the papers [15], [20], [27], [29], [30] that ULTS is expressed by a symmetrically distributed linguistic term set. So, in our proposal we work on to overcome this limitation and define ULTS as $LS = \left\{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\right\}$ where $a \in \mathbb{R}$. The terms distributed in the set LS are globally inconsistent. Our proposal is not subjected to any transformation function converting unbalanced linguistic scale set into a balanced set which might be a tedious and time taking job in several large scale decision problems.

Although any group decision problems dealing with unbalanced linguistic information can be tackled by using well-known methodologies provided in the literature, our approach is an improvement to offer acceptable results at a faster rate. Moreover, it is witnessed that the time complexity of our model is linear and further involves less number of computation steps. Consequently, it's computational cost decreases in contrast with the model-based upon linguistic hierarchy [15]. Lastly, the design of our model is much simpler, which is considered to be the most significant component of our proposal.

VII. Conclusion

In this research, we have formulated a novel approach to handle unbalanced linguistic information wherein linguistic

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

information is presupposed to be asymmetrically and non-uniformly distributed. Further, we have designed an algorithm to represent semantics to the unbalanced linguistic terms and afterward developed a 2-tuple model for ULTS based on the concept of minimum distance measure. For the sake of completeness, we provided basic aggregation operators based on Ω and Ω^{-1} functions. The utility of our methodology lies in the fact that it is less expensive and further dispels the complexity instead of other existing methods that are based on linguistic hierarchy [15]. Lastly, two numerical illustrations are provided to validate and demonstrate the usability of the suggested model in the physical realm.

The proposed model is generalized for both the types of ULTSs and also useful for the cases where linguistic information is assessed in the set suchlike $LS=$ {very low, fair, high, very high}. It is noted here that the distance between each of the consecutive terms assessed in the set is unequal. However, to carry out the computations, one may proceed by adding a dummy variable to the set LS. It is envisioned that the addition of dummy variable into the set LS does not provide us with the optimal solution in the several cases suchlike cases of multi-criteria decision problems, fuzzy linguistic-based game-theoretic models, financial problems, and so forth which may lead to a limitation.

Hence, one can work for the further extension of this model to subjugate the limitation associated with it. To complete this model, we give an exhaustive study relevant to the aggregation operators based on our model in the future and further give a comparison. Moreover, we should emphasize that this model has increased the efficacy of several linguistic solving processes and has produced results in less computation time, which, in a way, center the continuity of work.

APPENDIX

Proposition 4. The symbolic translation $\gamma \in [-0.5, 0.5]$.

Proof: We know, $\gamma = \frac{\beta-\lambda}{d_L+|d_R|}$ where $\lambda=\lambda_R$ or λ_L, and $\beta \in [\lambda_L, \lambda_R] \subseteq [a^{-n}, a^m]$. We consider the following two cases:

Case 1: $\lambda = \lambda_L$
In this case, $d_L \leq |d_R| \implies 2d_L \leq |d_R| + d_L \implies \frac{\beta-\lambda_L}{d_L+|d_R|} \leq \frac{1}{2}$ $\implies \gamma \leq \frac{1}{2}$.

Case 2: $\lambda = \lambda_R$
In this case, $|d_R| \leq d_L \implies 2|d_R| \leq d_L + |d_R| \implies \frac{-|d_R|}{d_L+|d_R|} \geq \frac{-1}{2} \implies \frac{\beta-\lambda_R}{d_L+|d_R|} \geq \frac{-1}{2} \implies \gamma \geq \frac{-1}{2}$.

■

Numerical scale based on multiplicative linguistic scale set

Before we give a proposition 5, we initially put forward the concept of numerical scale function based on ULTS.

In literature, Dong et al. [27] suggested the notion of a numerical scale based on a BLTS, $LT = \{s_i \,|\, i = 0 \text{ to } g\}$. Analogously, we provide a definition of a numerical scale based on a multiplicative linguistic scale set, LS.

Definition 13: Let $LS=\left\{t_{a^{(-n)}}, t_{a^{-(n-1)}}, \ldots, t_{a^0}, \ldots, t_{a^m}\right\}$ where, $a>1$ and $n, m \in \mathbb{Z}^+$ and n, m may or may not be equal is a multiplicative ULTS and $\mathbb{R}$ has the usual meaning in mathematical terms. Then we define a numerical scale function as $NS : LS \to \mathbb{R}$ such that $NS(t_{a^i}) = a^i$.

Definition 14: Numerical scale function $\overline{NS}$: $LS \times [-0.5, 0.5] \to \mathbb{R}$ is defined as follows:

$$\overline{NS}((t_{a^i}, \gamma)) = \begin{cases} NS(t_{a^i}) + \gamma \times ((NS(t_{a^{i+1}})) - NS(t_{a^i})), & \gamma \geq 0; \\ NS(t_{a^i}) + \gamma \times ((NS(t_{a^i})) - NS(t_{a^{i-1}})), & \gamma < 0. \end{cases}$$

Now, we proceed further to give a proposition to obtain our model from a numerical scale.

Proposition 5. When setting $NS(t_{a^i})=a^i$ where $i=-n, -(n-1), \ldots, 0, 1, \ldots, m$, we have $NS((t_{a^i}, \gamma_{a^i}))=\Omega^{-1}((t_{a^i}, \gamma_{a^i}))$, for any $(t_{a^i}, \gamma_{a^i}) \in LS \times [-0.5, 0.5]$.

Proof: Consider the following two cases:

Case 1: $\gamma \geq 0$
In this case, $a^i = a^{i_L}$ and $a^{i+1} = a^{i_R}$ for $i=-n, -(n-1), \ldots, 0, 1, \ldots, m$; $\Omega^{-1}(t_{a^i}, \gamma) = \gamma \times (d_L + |d_R|) + a^i = \gamma \times (\beta - a^i + |(\beta - a^{i+1})|) + a^i$; since $\beta < a^{i+1}$ therefore $\Omega^{-1}(t_{a^i}, \gamma) = \gamma \times (a^{i+1} - a^i) + a^i$. Now, from the definition of numerical scale function, we have $NS((t_{a^i}, \gamma)) = a^i + \gamma \times (a^{i+1} - a^i)$. Hence, $NS((t_{a^i}, \gamma)) = \Omega^{-1}((t_{a^i}, \gamma))$.

Case 2: $\gamma < 0$
In this case, $a^i = a^{i_R}$ and $a^{i-1} = a^{i_L}$ for $i=-n, -(n-1), \ldots, 0, 1, \ldots, m$; $\Omega^{-1}(t_{a^i}, \gamma) = \gamma \times (d_L + |d_R|) + a^i = \gamma \times (\beta - a^{i-1} + |(\beta - a^i)|) + a^i$; since $\beta < a^i$ therefore $\Omega^{-1}(t_{a^i}, \gamma) = \gamma \times (a^i - a^{i-1}) + a^i$. Now, from the definition of numerical scale function, we have $NS((t_{a^i}, \gamma)) = a^i + \gamma \times (a^i - a^{i-1})$. Hence, $NS((t_{a^i}, \gamma)) = \Omega^{-1}((t_{a^i}, \gamma))$.

■

REFERENCES

[1] J. Tang, Z. Xu, S. Liu, & X. Xu, "Overall unbalanced multiplicative linguistic evaluation scales," *Fuzzy Systems and Mathematics*, vol. 23, no. 5, pp. 99–102, 2009.

[2] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Part I, Information Sciences*, vol. 8, no. 3, pp. 199–249, 1975.

[3] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Part II, Information Sciences*, vol. 8, no. 4, pp. 301–357, 1975.

[4] L. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," *Part III, Information Sciences*, vol. 9, no. 1, pp. 43–80, 1975.

[5] G. Bordogna and G. Pasi, "A fuzzy linguistic approach generalizing boolean information retrieval: A model and its evaluation," *J. Amer. Soc. Inf. Sci.*, vol. 44, pp. 70–82, 1993.

[6] G. Bordogna and G. Pasi, "An ordinal information retrieval model," *Int. J. Uncertain., Fuzz. Knowl.-Based Syst.*, vol. 9, pp. 63–76, 2001.

[7] E. Herrera-Viedma, "An information retrieval model with ordinal linguistic weighted queries based on two weighting elements," *Int. J. Uncertain., Fuzz. Knowl.-Based Syst.*, vol. 9, pp. 77–88, 2001.

[8] E. Herrera-Viedma, F. Herrera, L. Martínez, J.C. Herrera, and A.G. Lopez-Herrera, "Incorporating filtering techniques in a fuzzy linguistic multi-agent model for gathering of information on the web," *Fuzzy Sets Syst.*, vol. 148, pp. 61–83, 2004.

[9] E. Herrera-Viedma, A.G. Lopez-Herrera, and C. Porcel, "Tuning the matching function for a threshold weighting semantics in a linguistic information retrieval system," *Int. J. Intell. Syst.*, vol. 20, no. 9, pp. 921–937, 2005.

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TFUZZ.2020.2994987, IEEE

[10] E. Herrera-Viedma, G. Pasi, A.G. Lopez-Herrera, and C. Porcel, "Evaluating the information quality of web sites: A methodology based on fuzzy," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 4, pp. 538–549, 2006.

[11] E. Herrera-Viedma and E. Peis, "Evaluating the informative quality of documents in SGML-format using fuzzy linguistic techniques based on computing with words," *Inf. Process. Manag.*, vol. 39, no. 2, pp. 195–213, 2003.

[12] E. Herrera-Viedma, L. Martínez, F. Mata, and F. Chiclana, "A consensus support system model for group decision-making problems with multi-granular linguistic preference relations," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 5, pp. 644–658, 2005.

[13] S. Zadrozny and J. Kacprzyk, "Computing with words for text processing: An approach to the text categorization," *Inf. Sci.*, vol. 176, no. 4, pp. 415–437, 2006.

[14] F. Herrera and E. Herrera-Viedma, "Aggregation operators for linguistic weighted information," *IEEE Trans. Syst., Man Cybern. A, Syst.*, vol. 27, pp. 646–656, 1997.

[15] F. Herrera, E. Herrera-Viedma, and L. Martínez, "A Fuzzy Linguistic Methodology to Deal with Unbalanced Linguistic Term Sets," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 2, pp. 354–370, 2008.

[16] F. Herrera, E. Herrera-Viedma, and J. L. Verdegay, "Direct approach processes in group decision making using linguistic OWA operators," *Fuzzy Sets Syst.*, vol. 79, no. 2, pp. 175–190, 1996.

[17] F. Herrera and L. Martínez, "A 2-tuple fuzzy linguistic representation model for computing with words," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 6, pp. 746–752, 2000.

[18] F. Herrera, L. Martínez, "The 2-tuple linguistic computational model. advantages of its linguistic description, accuracy and consistency," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 9 (suppl.) pp. 33–48, 2001.

[19] G.W. Wei, "Some generalized aggregating operators with linguistic information and their application to multiple attribute group decision making," *Computers and Industrial Engineering*, vol. 61, pp. 3–38, 2011.

[20] J. Wang and J. Hao, "A new version of 2-tuple fuzzy linguistic representation model for computing with words," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 3, pp. 435–445, 2006.

[21] L. Martínez, J. Liu, J.B. Yang, and F. Herrera, "A multigranular hierarchical linguistic model for design evaluation based on safety and cost analysis," *Int. J. Intell. Syst.*, vol. 20, no. 12, pp. 1161–1194, 2005.

[22] R. Degani and G. Bortolan, "The problem of linguistic approximation in clinical decision making," *Int. J. Approx. Reason.*, vol. 2, no. 2, pp. 143–162, 1988.

[23] Y.P. Jiang, Z.P. Fan, and J. Ma, "A method for group decision making with multi-granularity linguistic assessment information," *Inf. Sci. (Ny).*, vol. 178, no. 4, pp. 1098–1109, 2008.

[24] R.R. Yager, "A new methodology for ordinal multi objective decisions based on fuzzy sets," *Decision Sciences*, vol. 12, pp. 589–600, 1981.

[25] R.R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decision making," *IEEE Trans. Syst., Man, Cybern.*, no. 18, pp. 183–190, 1988.

[26] Z. Xu, "A method based on linguistic aggregation operators for group decision making with linguistic preference relations," *Inf. Sci. (Ny).*, vol. 166, no. 1âĂŞ4, pp. 19–30, 2004.

[27] Y. Dong, Y. Xu, and S. Yu, "Computing the Numerical Scale of the Linguistic Term Set for the 2-Tuple Fuzzy Linguistic," *IEEE Trans. Fuzzy Syst.*, vol. 17, no. 6, pp. 1366–1378, 2009.

[28] Y. Dong, C.C. Li, and F. Herrera, "Connecting the linguistic hierarchy and the numerical scale for the 2-tuple linguistic model and its use to deal with hesitant unbalanced linguistic information," *Inf. Sci. (Ny).*, vol. 367-368, pp. 259–278, 2016.

[29] M. Cai, Z. Gong, and X. Yu, "A Method for Unbalanced Linguistic Term Sets and its Application in Group Decision Making," *Int. J. Fuzzy Syst.*, 2016.

[30] M. Cai and Z. Gong, "Group decision making using distances between unbalanced linguistic assessments," *Appl. Soft Comput. J.*, vol. 67, pp. 613–624, 2018.

[31] Z. Pei and L. Zheng, "New unbalanced linguistic scale sets: The linguistic information representations and applications," *Comput. Ind. Eng.*, vol. 105, pp. 377–390, 2017.

[32] D. Wu, & J.M. Mendel, "A vector similarity measure for linguistic approximation: Interval type-2 and type-1 fuzzy sets," *Information Sciences*, vol. 178, pp. 381–402, 2008.

[33] I.B. Turksen, "Type 2 representation and reasoning for CWW," *Fuzzy Sets and Systems*, vol. 127, 17–36, 2002.

[34] F.J. Cabrerizo, E. Herrera-Viedma, and W. Pedrycz, "A method based on PSO and granular computing of linguistic information to solve group decision making problems defined in heterogeneous contexts," *Eur. J. Oper. Res.*, vol. 230, no. 3, pp. 624–633, 2013.

[35] W. Pedrycz and M. Song,"A granulation of linguistic information in AHP decision-making problems," *Inf. Fusion*, vol. 17, no. 1, pp. 93–101, 2014.

[36] J. Wang and J. Hao, "An Approach to Computing With Words Based on Canonical Characteristic Values of Linguistic Labels," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 4, pp. 593–604, 2007.

[37] J. Lawry, "An alternative approach to computing with words," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, pp. 3–16, 2001.

[38] L. Zou, Z. Pei, H. R. Karimi, and P. Shi, "The unbalanced linguistic aggregation operator in group decision making," *IEEE Multi-Conference Syst. Control. MSC 2012*, pp. 703–707, 2012.

[39] F. Herrera, and L. Martínez, "A Model Based on Linguistic 2-Tuples for Dealing with Multigranular Hierarchical Linguistic Contexts in Multi-Expert Decision-Making," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 31, no. 2, pp. 227–234, 2001.

[40] L. Jiang, H. Liu, and J. Cai, "The power average operator for unbalanced linguistic term sets," *Inf. Fusion*, vol. 22, pp. 85–94, 2015.

[41] M.A. Abchir and I. Truck, "Towards an extension of the 2-tuple linguistic model to deal with unbalanced linguistic term sets," *Kybernetika*, vol. 49, no. 1, pp. 164–180, 2013.

[42] Y. Dong, W. Hong, Y. Xu, and S. Yu, "Numerical scales generated individually for analytic hierarchy process," *Eur. J. Oper. Res.*, vol. 229, no. 3, pp. 654–662, 2013.

[43] Y. Dong, C.C. Li, and F. Herrera, "Connecting the numerical scale model to the unbalanced linguistic term sets," *IEEE Int. Conf. Fuzzy Syst.*, pp. 455–462, 2014.

[44] Y. Dong and E. Herrera-viedma, "Consistency-Driven Automatic Methodology to Set Interval Numerical Scales of 2-Tuple Linguistic Term Sets and Its Use in the Linguistic GDM With Preference Relation," *IEEE transactions on cybernetics*, vol. 45, no. 4, pp. 1–13, 2014.

[45] Y. Dong, G. Zhang, W. Hong, S. Member, S. Yu, and S. Member, "Linguistic Computational Model Based on 2-Tuples and Intervals," *IEEE Transactions on fuzzy systems*, vol. 21, no. 6, pp. 1006–1018, 2013.

[46] Y. Dong, Y. Wu, H. Zhang, and G. Zhang, "Multi-granular unbalanced linguistic distribution assessments with interval symbolic proportions," *Knowledge-Based Syst.*, vol. 82, pp. 139–151, 2015.

[47] B. Wang, J. Liang, Y. Qian, and C. Dang, "A normalized numerical scaling method for the unbalanced multi-granular linguistic sets," *Int. J. Uncertainty, Fuzziness Knowlege-Based Syst.*, vol. 23, no. 2, pp. 221–243, 2015.

[48] C. Li, Y. Dong, F. Herrera, E. Herrera-viedma, and L. Martínez, "Personalized individual semantics in computing with words for supporting linguistic group decision making . An application on consensus reaching," *Information Fusion*, vol. 33, pp. 29–40, 2017.

[49] Ł. Bartczuk, P. Dziwiński, and J.T. Starczewski, "A new method for dealing with unbalanced linguistic term set," *Lect. Notes Comput. Sci. (In international conference on artificial intelligence and soft computing)*, vol. 7267 LNCS, Part 1, pp. 207–212, 2012.

1063-6706 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:40 UTC from IEEE Xplore. Restrictions apply.

ARTICLE IN PRESS

Integration, the VLSI Journal xxx (xxxx) xxx

Contents lists available at ScienceDirect

Integration, the VLSI Journal

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

A new realization scheme for dynamic PFSCL style

Ranjana Sivaram [a], Kirti Gupta [b], Neeta Pandey [a,*]

[a] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India
[b] Department of Electronics and Communication Engineering, Bharati Vidyapeeth's College of Engineering, Delhi, 110063, India

ARTICLE INFO

Keywords:
low Power
Mixed-signals circuits
Source-coupled logic
PFSCL
Dynamic PFSCL

ABSTRACT

In this paper, a new scheme of logic function realization in dynamic positive feedback source-coupled logic (D-PFSCL) style is proposed. The existing scheme implements only NOR/OR based realization of a logic function. Thus, a complex function in D-PFSCL has high gate count which degrades the overall circuit performance measured in terms of power and delay. This paper therefore aims to resolve the issue by proposing a scheme which modifies the structure of a D-PFSCL gate. The modified gate exhibits AND/OR functionality and is used to realize various functions. Simulations have been carried out by implementing various functions and comparing their performance with the existing schemes at 1 GHz. The results of performance comparison with existing schemes indicates significant reuduction in gate count resulting in overall performance improvement.

1. Introduction

The research on source coupled logic (SCL) has gained momentum in the last few decades due to its inherent suitability in high speed mixed mode environment than the conventional CMOS logic counterpart. The SCL gates are used for designing wide variety of applications such as optical circuits, neural networks, front end digital signal processing units, ADCs and DACs etc [1–9]. The SCL gates are reduced swing, and use a constant current source to provide an analog friendly environment. These gates work on current steering principle and consume static power. In literature both differential and single-ended SCL gates are present [10, 11]. Positive feedback source-coupled logic (PFSCL) style is one of the promising SCL styles wherein a positive feedback is used for the speed improvement [10,11].

Dynamic source coupled logic are proposed in literature [12–16] wherein a dynamic current source is used which eliminates static power consumption. The dynamic gates offer significant power saving and reduced delays in comparison to static counterpart. The dynamic SCL gates with differential and single-ended signaling exist in literature. In Ref. [12], the differential dynamic SCL gate named as dynamic current mode logic (DyCML) was proposed. These gates use dynamic current source instead of constant current source and have similar logic realization as the static counterpart [12,13]. An alternate arrangement of DyCML gates presented in Ref. [14] modifies the dynamic current source by introducing a feedback in the circuit. The DyCML gate [12] suffer from high delay values due to the arrangement of source-coupled transistor pairs in multiple level while the DyCML gates [14] requires different path impedances for correct operation. A hybrid dynamic current mode gate is presented in Ref. [15] but its performance is limited by the transmission gates. Another arrangement of dynamic CML gate is suggested in Ref. [16] which uses static current source for improved performance. Thus, these gates considerably deviate from chronicles and consume both static and dynamic power. Analogously, dynamic gates for PFSCL are suggested in Ref. [17] by adapting the method suggested for differential dynamic gates in Ref. [12]. The dynamic positive-feedback source coupled logic (D-PFSCL) gates show improvement in speed and power consumption over its static PFSCL counterpart. The D-PFSCL supports only NOR/OR operation so the overall gate count for any complex logic function implementation is high.

In this paper, the existing D-PFSCL gate is modified with an intention to incorporate AND/OR functionality in the gate itself. This modification reduces the gate count and also intermittent buffers needed for complex logic function implementation [14]. As a consequence, the dynamic power consumption and the delay are reduced. The paper is organized in five sections including the present one. The concept of an existing D-PFSCL gate and the logic function realization scheme is elucidated in section 2. The proposed D-PFSCL gate, its design and power consumption are described in section 3. This section also puts forward a generalized approach to configure the gate for any given two input functionalities. Section 4 presents simulation results to verify the proposed scheme and its comparison with existing scheme. Lastly conclusions are drawn in Section 5.

* Corresponding author.
E-mail address: n66pandey@rediffmail.com (N. Pandey).

https://doi.org/10.1016/j.vlsi.2020.05.004
Received 18 December 2019; Received in revised form 28 April 2020; Accepted 17 May 2020
Available online xxxx
0167-9260/© 2020 Elsevier B.V. All rights reserved.

Please cite this article as: R. Sivaram et al., A new realization scheme for dynamic PFSCL style, Integration, the VLSI Journal, https://doi.org/10.1016/j.vlsi.2020.05.004

2. D-PFSCL circuits: Existing scheme

The structure of a N-input D-PFSCL gate comprise of three main parts namely a pull down network (PDN), precharge transistors and a dynamic current source as shown in Fig. 1a [17]. The PMOS transistors M_3, (M_4: optional) are the precharge transistors while the transistors M_1, M_2, the capacitor C_{dyn} forms the dynamic current source. Both of them are driven by a clock (CLK) input. The PDN consists of parallel connection of N-transistors (M_{N1}-M_{NN}) driven by inputs A_1-A_N. These transistors are further source-coupled with a transistor M_{F1} driven by a feedback connection from the output node $\overline{Q}$ voltage to provide a feedback connection. The output node, $\overline{Q}$ provides NOR output. However, in cases, where OR of the inputs is required, then it is obtained by placing an additional precharge transistor M4 as shown in Fig. 1a and then the output from its drain terminal. At any given instance, output from only one node is obtained as these gates belong to single-ended SCL gates.

A D-PFSCL gate operates in two phases known as precharge phase and the evaluation phase, depending on the CLK signal. For low value of CLK, the circuit works in precharge phase wherein the output node is charged to V_{DD} since the precharge transistors are ON. Simultaneously, the capacitor C_{dyn} gets discharged to ground potential via the conducting M_1 transistor. Further, any changes in the applied inputs A_1-A_N do not influence the output as M_2 is OFF such that no conductive path can be established between the output nodes and ground. Alternatively, for high value of CLK, the logic function is evaluated since M_2 is ON and the transistors M_1, M_3, M_4 are OFF.

The symbol of a generic D-PFSCL gate is drawn in Fig. 1b to illustrate the design of multi-stage applications using D-PFSCL gates. In such applications, D-PFSCL gates are cascaded by inserting a self-timed buffer (STB) to avoid malfunction because of the simultaneous evaluation of all stages [17]. A STB ensures that the evaluation of a stage starts only after the completion of evaluation in its preceding stage. The STB uses two clocked cascaded inverters as shown in Fig. 1c. The CLK signal (CLK_i) and the node Y voltage in dynamic current source of the preceding state (stage i) drives the first inverter. The output of STB (CLK_{i+1}) is now given as the CLK signal to the subsequent stage (stage i+1). The symbol of the STB is shown in Fig. 1d. As an example, the realization of a 2:1 multiplexer (MUX) is discussed. The NOR/OR based gate level schematic is drawn in Fig. 1e. It has two-stage structure wherein three two-input NOR/OR gates are employed. The NOR/OR gates are further substituted by their respective D-PFSCL gate and a STB is inserted in between the stages as represented by a block diagram in Fig. 1f. It is assumed that the true and the complement of all the inputs are available. The power analysis states that the total power is the sum of power of the three D-PFSCL NOR/OR gates and a STB. Similarly, it can be noted that the propagation delay is the summation of the delay in individual stages. Thus, it is evident that if any application has large number of stages then the system will become inefficient in terms of both power and delay. Therefore, to improve the performance of D-PFSCL based multi-stage applications, an alternate scheme that reduces the gate count as well the stages in cascading is presented in the paper.

3. New scheme

The main limitation in the existing scheme is the implementation of logic functions in only NOR/OR forms due to the structure of an existing D-PFSCL gate. But in practice, most of the functions are expressed as sum of products resulting in AND-OR implementation. Therefore, in such case, the existing scheme infers high gate count as well as cascading of multiple gates which degrade the performance in terms of both power and delay. A new scheme that modifies the PDN of existing D-PFSCL gate is presented for D-PFSCL circuits realizations.

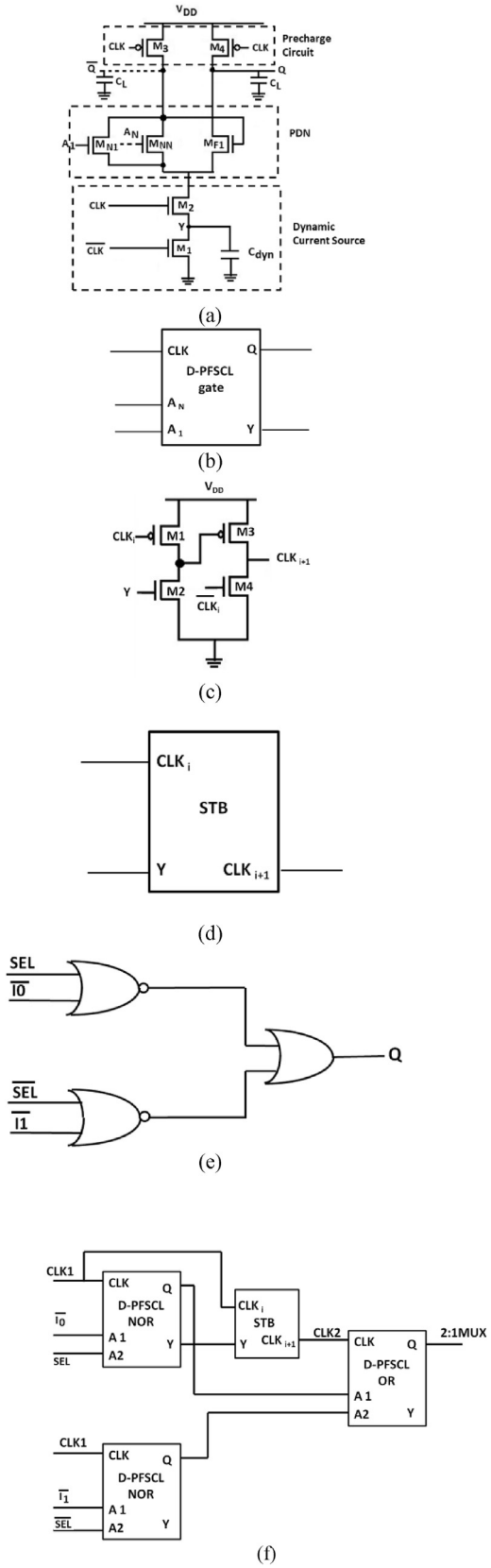


Fig. 1. (a) Schematic of N-input D-PFSCL gate [17](b) D-PFSCL gate symbol representation (c) schematic of STB [12] (d) STB symbol (e) NOR/OR based gate level schematic of 2:1 MUX (f) D-PFSCL 2:1 MUX.

3.1. Proposed D-PFSCL gate

The concept behind the proposed scheme is to generate logical AND term by modifying the existing D-PFSCL gate. An additional low threshold voltage NMOS transistor (M_{N2}) is inserted between the power supply and the source-coupled transistor pair (M_{N1}, M_{F1}) as shown in Fig. 2. For clear distinction, the low threshold voltage transistor M_{N2} is marked with bold line while the usual MOS symbol is used for remaining transistors. The threshold voltage of the middle transistor M_{N2} is maintained lower than other transistors in the PDN by a factor δ. In the Fig. 2, the output is taken from alternate branch so additional precharge

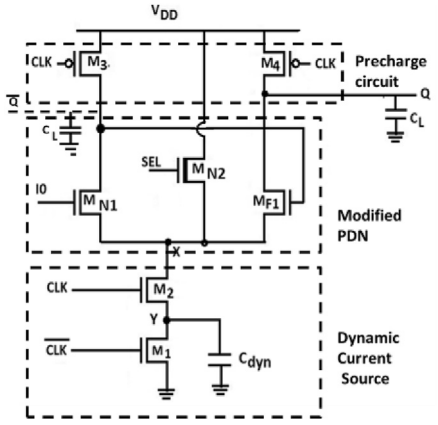


Fig. 2. Modified schematic of a D-PFSCL gate with the addition of M_{N2}.

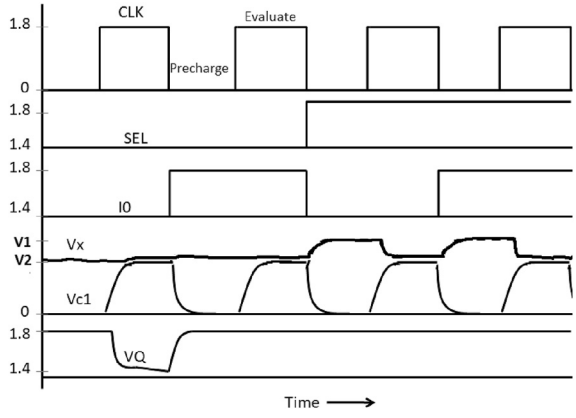


Fig. 3. Voltages at different nodes in the modified D-PFSCL gate.

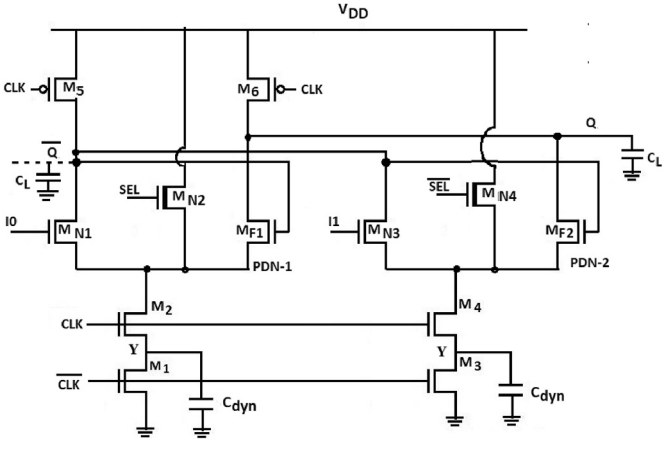


Fig. 4. Complete schematic of the proposed D-PFSCL gate.

Table 1
Summary of the proposed D-PFSCL gate operation.

INPUT(S)			State of transistors in PDN-1			State of transistors in PDN-2			OUTPUT Q
SEL	I_0	I_1	M_{N1}	M_{N2}	M_{F1}	M_{N3}	M_{N4}	M_{F2}	
L	L	L	OFF	OFF	ON	OFF	ON	ON	L
L	L	H	OFF	OFF	ON	ON	ON	OFF	L
L	H	L	ON	OFF	OFF	OFF	ON	ON	H
L	H	H	ON	OFF	OFF	ON	ON	OFF	H
H	L	L	OFF	ON	ON	OFF	OFF	ON	L
H	L	H	OFF	ON	ON	ON	OFF	OFF	H
H	H	L	ON	ON	OFF	OFF	OFF	ON	L
H	H	H	ON	ON	OFF	ON	OFF	OFF	H

Where H=High logic level (V_{DD}), L = Low logic level (V_{DD}-V_{SWING}).

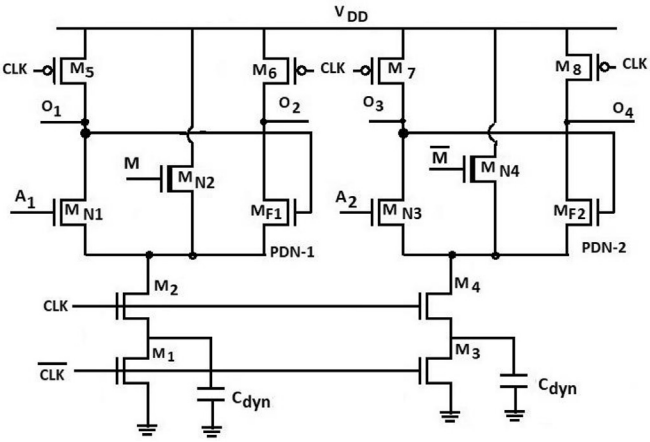


Fig. 5. Proposed generalized D-PFSCL gate based on proposed scheme.

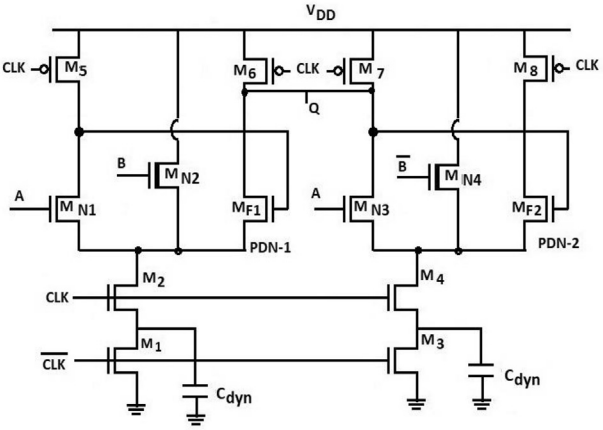


Fig. 6. XOR2 based on proposed generalized D-PFSCL gate.

Table 2
Mapping table.

Function	Inputs	Mapping in proposed D-PFSCL gate			Output Nodes
		M	A1	A2	Q
OR	A, B	B	A	B	O2, O4
NOR	A, B	B	A	B	O1, O3
NAND	A, B	B	B	A	O1, O3
AND	A,B	B	B	A	O2, O4
XOR	A, B	B	A	A	O2, O3
XNOR	A, B	B	A	A	O1, O4
MUX	SEL,I0,I1	SEL	I0	I1	O2, O4

transistor M4 is used. The effect of this modification is analysed by connecting inputs I0 and SEL to M_{N1} and M_{N2} respectively and generating the output from node Q. In the precharge phase, the output node is charged to V_{DD} and C_{dyn} is discharged to ground potential in the same way as in existing D-PFSCL gate. The potential of node X depends on the SEL input during this phase. It becomes $(V1 = V_{DD} - V_{T,n}/\ \delta)$ or $(V2 = V_{DD} - V_{T,n})$ for the high and low value of SEL input respectively. The inputs I0 and SEL do not influence the output as M_2 is OFF.

During the evaluation phase, the transistors M_1, M_3 and M_4 are OFF as the clock signal is at high logic level. The output is now evaluated since M_2 is ON. The capacitor C_{dyn} is charged till its potential becomes equal to node X potential to stop the charge transfer further. The impact of adding M_{N2} is studied by assuming low and high values of the SEL signal.

Case 1: When SEL is low

In this case, M_{N2} is OFF and the output is evaluated according to input I0. If I0 is high (low), the output node will attain high (low) logic level.

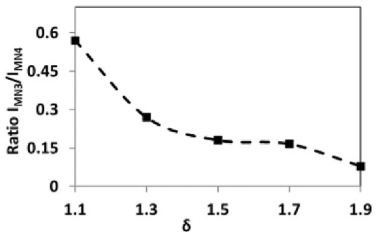


Fig. 7. I_{MN3}/I_{MN4} for different values of δ

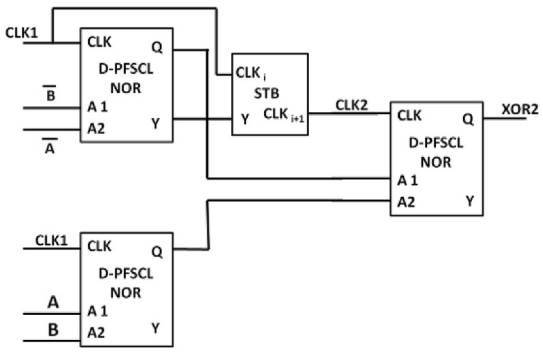


Fig. 8. D-PFSCL XOR2 gate based on the existing scheme.

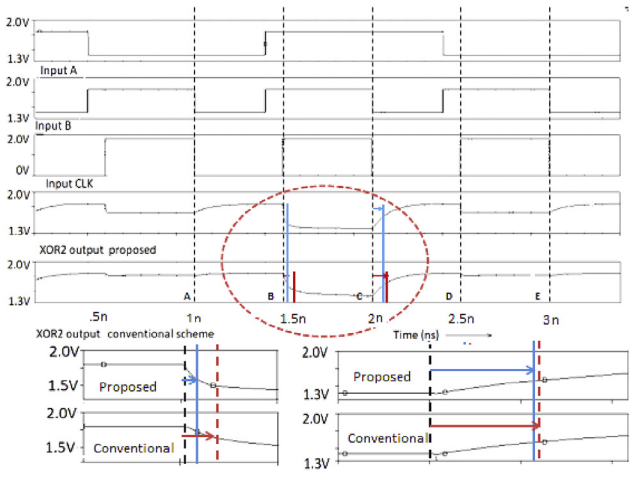


Fig. 9. Simulation waveforms of the D-PFSCL XOR2 gate.

Case 2: When SEL is high

In this condition, M_{N2} is ON and depending on the logic level at I0 either M_{N1} or M_{F1} will be ON. Let us assume that I0 is at low logic level such that M_{F1} is ON and M_{N1} is OFF. As the transistors M_{F1} and M_{N2} both conduct therefore both of them participate in the charging of C_{dyn}. But since the threshold voltage of M_{N2} is lower than M_{F1}, larger current flows through M_{N2} and charges C_{dyn} from power supply without much contribution of M_{F1}. Thus, the charge transfer from C_L to C_{dyn} does not happen and the output node remains at high logic level. Similarly, for high value of I0, the output remains high due to charging of C_{dyn} by M_{N2} instead of M_{N1}. Thus, the output remains high irrespective of I0 value. The same behaviour is elaborated in Fig. 3. It is to be noted that C_{dyn} discharges to ground in precharge phase and gets charged till its potential becomes equal to V2 of node X.

It is clear from the above discussion that for high value of SEL the output of modified gate remains independent of I0. Thus, an extra circuitry is needed such that the output node responds to transition at the input when SEL is high. This is accomplished by adding an identical set of

Table 3
Performance comparison.

Parameter	Style			
	Dynamic CMOS	Static PFSCL	Existing D-PFSCL	Proposed D-PFSCL
Circuit 1: 2:1 Multiplexer				
Gate count	1	3	3	1
STB count	0	–	1	0
Transistor count	6	18	27	14
τ_{PHL} (ps)	98	194	334	82
τ_{pre} (ps)	76	–	207	45
Power (μW)	389	600	182	88
EDP ($x10^{-27}$J.s)	933	5645	5075	147
I_{static}	–	300 μA	–	–
I_{leak} (μA)	6.37	–	0.57	0.85
Circuit 2: XOR2				
Gate count	1	3	3	1
STB count	0	–	1	0
Transistor count	6	18	27	14
τ_{PHL} (ps)	119	156	197	86
τ_{pre} (ps)	78	–	150	45
Power (μW)	364	590	290	85
EDP ($x10^{-27}$J.s)	1288	3589	2813	157
I_{static} (μA)	–	300	–	–
I_{leak} (μA)	6.62	–	0.67	0.98
Circuit 3: NAND2				
Gate count	1	1	1	1
STB count	0	–	0	0
Transistor count	4	6	7	12
τ_{PHL} (ps)	105	101	96	78
τ_{pre} (ps)	75	–	88	46
Power (μW)	243	180	54	114.3
EDP ($x10^{-27}$J.s)	669.7	459	124.4	173.8
I_{static} (μA)		100		
Il_{eak} (μA)	5.931	–	0.34	0.88
Circuit 4: XOR3				
Gate count	3	9	9	2
STB count	0	0	3	1
Transistor count	18	54	81	34
τ_{PHL} (ps)	246	369	335	134
τ_{pre} (ps)	257	398	298.7	163.5
Power (μW)	477	1620	338	191
EDP ($x10^{-27}$J.s)	7216.5	55,145	9483	857.4
I_{static} (μA)	–	900	–	–
I_{leak}(μA)	25	–	2.6	2

PDN with input I1, complement of SEL ($\overline{SEL}$) and a dynamic current source as shown in Fig. 2. The complete schematic of the proposed D-PFSCL gate is drawn in Fig. 4. It now consists of two PDNs named as PDN-1 (M_{N1}, M_{F1}, M_{N2}) and PDN-2 (M_{N3}, M_{F2}, M_{N4}). The signals (I0, I1, SEL and $\overline{SEL}$) are inputs to the two PDNs and the CLK signal drives the pre-charge and dynamic current source transistors. For high values of SEL, $\overline{SEL}$ is at low logic level, such that the output is obtained according to the I1 input. The working of the proposed D-PFSCL gate for various combinations of the inputs is summarized in Table 1. Thus, the complete functionality of the proposed D-PFSCL gate can be modelled as:

$$Q = \begin{cases} I0, & if\ SEL = 0 \\ I1, & if\ SEl = 1 \end{cases} \tag{1}$$

which can be written in Boolean equation as:

$$Q = I0\ \overline{SEL} + I1\ SEL \tag{2}$$

Hence, the proposed gate incorporates the AND/OR functionality. Further, Eq. (2) represents the function of a 2:1 MUX with data inputs I0 and I1 and select input SEL. The comparison of proposed MUX realization with the existing one (Fig. 1f) clearly indicate the reduction in gate count and STB leading to performance improvement.

3.2. Design of the proposed D-PFSCL gate

In the design of proposed D-PFSCL gate, the value of C_{dyn} plays an important role since it affects the voltage swing of the gate. It is known that in the evaluation phase the charge from the output load capacitance C_L gets transferred to C_{dyn} in the dynamic current source. Therefore, to achieve the required voltage swing V_{SWING}, the charge conservation principle is applied as:

$$V_{SWING}C_L = C_{dyn}(V_{DD} - V_{SWING}) \tag{3}$$

Thus, the value of C_{dyn} to obtain the required voltage swing at the output node is computes as:

$$C_{dyn} = \frac{V_{SWING}C_L}{(V_{DD} - V_{SWING})} \tag{4}$$

In practice, a capacitor is generally realized by a MOSFET with its source and drain terminals connected together. Therefore, by substituting, $C_{dyn} = W_{Cdyn} \times L_{Cdyn} \times C_{ox}$ and for a given value of L_{C1}, the size of transistor is calculated as:

$$W_{Cdyn} \times L_{Cdyn} = \frac{V_{SWING} \times C_L}{C_{ox} \times (V_{DD} - V_{SWING})} \tag{5}$$

where W_{Cdyn}, L_{Cdyn} are the width and length of transistor for C_{dyn}.

3.3. Power consumption of the proposed D-PFSCL gate

The static power consumption in the proposed D-PFSCL gate is negligible as the transistors $M_{1\text{-}6}$ never turn ON simultaneously, and a direct path from power supply to ground is never established. It is pertinent to mention here that leakage current may flow through the OFF transistors and will culminate into leakage power consumption ($P_{leakage}$) and will scale up with the technology scaling. Further, the capacitors will charge up during different phases of clock and therefore will consume dynamic power (P_{dyn}). During the precharge phase, the load capacitor (C_L) is precharged by the power supply (V_{DD}). In the evaluation phase, depending on the input, C_{dyn} is charged via middle transistors of PDN or by charge transfer from the C_L. Thus, the total power consumption P_{Total} of the circuit is

$$P_{Total} = P_{dyn} + P_{leakage} \tag{6}$$

where the dynamic power P_{dyn} is calculated as:

$$P_{dyn} = \alpha_{H\to L}C_L V_{DD} V_{SWING} f_{CLK} + C_{dyn} V_{DD}(V_{DD} - V_{T,n}) f_{CLK} \tag{7}$$

where f_{CLK} is the CLK signal frequency and $\alpha_{H\to L}$ corresponds to high-to-low transition probability of the output node.

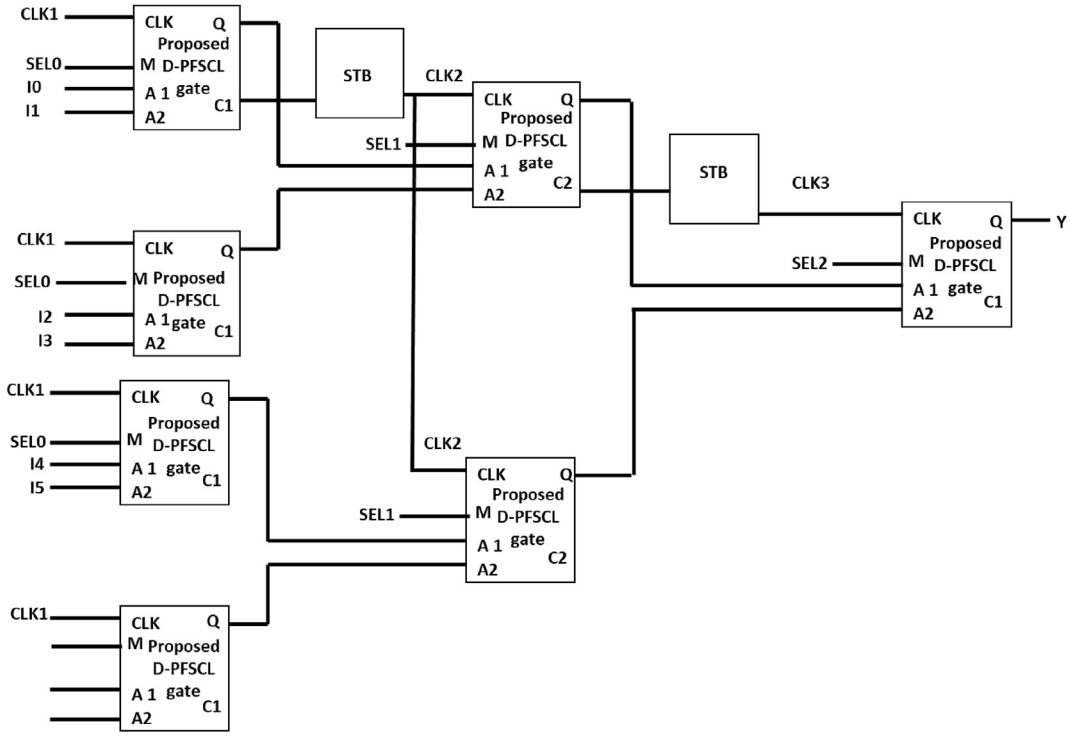


Fig. 10. Block diagram of the 8:1 multiplexer based on the proposed scheme.

3.4. A generalized approach

The proposed D-PFSCL gate (Fig. 4) can be transformed into a generalized structure by inserting separate precharge transistors in PDN-2 such that four nodes (O1–O4) are available to configure the gate for any given functionality. The complete schematic for the proposed generalized D-PFSCL gate is given in Fig. 5. At any given instant, only one output node from each PDN i.e. (either O1 or O2 from PDN-1) and (either O3 or O4 from PDN-2) is combined to define the output. The generalized structure also offers the advantage of obtaining output either in true or the compliment of the implemented function. To explain the same, the implementation of two input exclusive-OR function (XOR2) is considered. The XOR functionality with input A, B can be expressed as:

$$Q = \begin{cases} A & if\ B = 0 \\ \overline{A} & if\ B = 1 \end{cases} \quad (8)$$

So, mapping it with Eq. (1), input B and $\overline{B}$ are connected to M and $\overline{M}$ respectively. The inputs A and $\overline{A}$ are required to be connected to A_1 (PDN-1) and A_2 (PDN-2) respectively and connecting O2 (PDN-1) and O4 (PDN-2) to form output Q. An additional inverter is required to generate $\overline{A}$. This can be avoided in generalized structure by obtaining the output from other branch in PDN-2 i.e. O3 as shown in Fig. 6. Based on this, a mapping table for configuring the generalized D-PSFCL gate for various functionalities is given as Table 2.

4. Simulations

In this section, the performance of the proposed D-PFSCL gates is first compared with existing gates and is then followed by a performance examination of a D-PFSCL gate under different conditions.

4.1. Performance comparison

Different logic functions are implemented in the proposed D-PFSCL and other logic styles such as existing D-PFSCL, static PFSCL and dynamic CMOS style. The simulations are performed with a power supply, V_{SWING},

Table 4
Performance summary of 8:1 Multiplexer.

Parameter	Style			
	Dynamic CMOS	Static PFSCL	Existing D-PFSCL	Proposed D-PFSCL
Gate count	14	21	21	7
STB count	0	0	5	2
Transistor count	56	126	177	100
τ_{PHL} (ps)	754	864	831	400
τ_{pre} (ps)	562	–	620	250
Power (μW)	2268	3780	1134	700
EDP ($x10^{-24}$J.s)	322	705	195	28
I_{static} (μA)	–	2100	–	–
I_{leak} (μA)	40	–	10	6

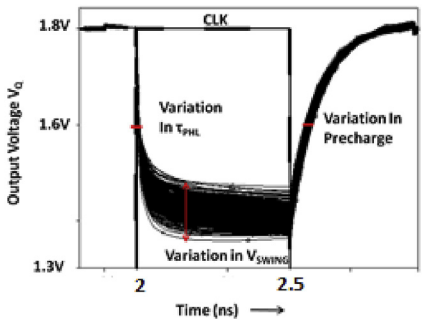


Fig. 11. Proposed D-PFSCL XOR2 output under Monte Carlo.

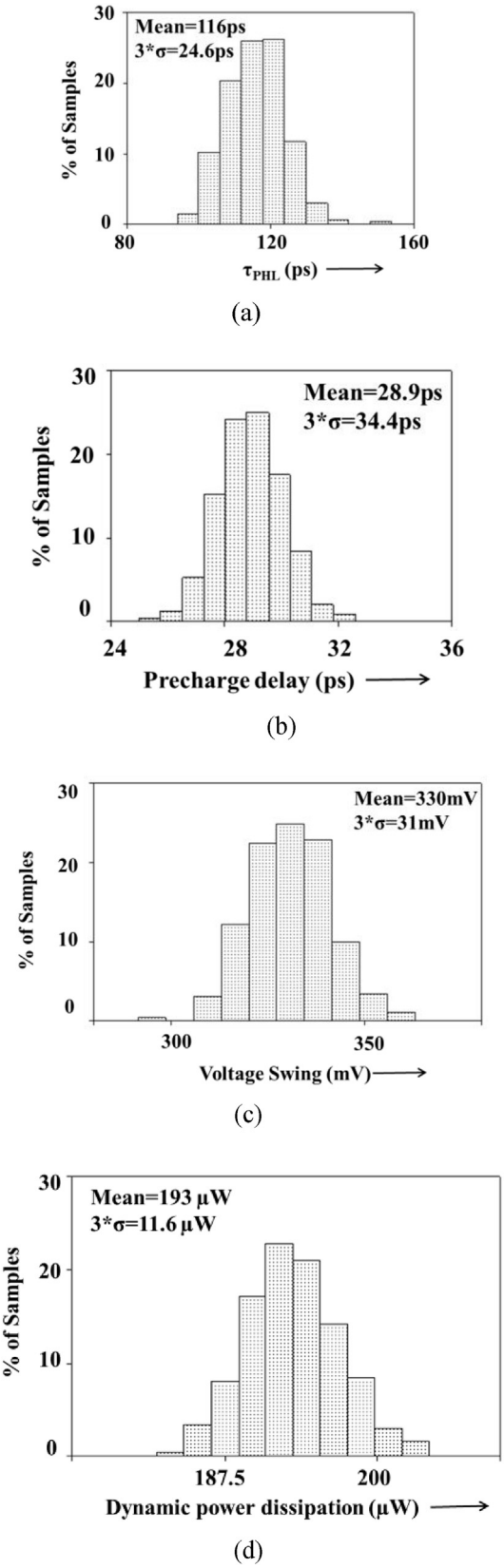


Fig. 12. Variation in (a) τ_{PHL} (b) precharge c) Voltage Swing d) Dynamic power dissipation.

Table 5
Monte Carlo simulation results for the XOR2 in different styles.

Parameter	Style							
	Dynamic CMOS		Static PFSCL		Existing D-PFSCL		Proposed D-PFSCL	
	Mean	3σ	Mean	3σ	Mean	3σ	Mean	3σ
τ_{PHL} (ps)	120	7.57	162	70.4	168.3	17.9	116	24.6
τ_{pre} (ps)	69.7	5.8	–	–	159.7	12.4	28.9	34.4
Power (μW)	321	3.45	580	23	1250	27.6	193	11.6
V_{SWING} (V)	1.8	0.01	0.388	0.057	0.36	0.1	0.33	0.031

Table 6
PVT analysis.

Parameter	Style			
	Dynamic CMOS	Static PFSCL	Existing D-PFSCL	Proposed D-PFSCL
	Simulation condition: Best Case: FF+ 1.98 V + 0°C			
τ_{PHL} (ps)	106	121	173	69
τ_{pre} (ps)	53.4	–	121	41
Power (μW)	243.78	664	325	109
Simulation condition: Worst Case: SS + 1.62V + 125°				
τ_{PHL} (ps)	163.5	207	257	137.9
τ_{pre} (ps)	121.7	–	173	67.3
Power (μW)	214.4	573	224	80

Table 7
Impact of Technology scaling on proposed XOR2.

parameter	Technology		
	180 nm	90 nm	65 nm
τ_{PHL} (ps)	86	61.7	59.3
τ_{pre} (ps)	45	17.8	23.6
Power (μW)	85	78.59	57.46
EDP ($x10^{-28}$J.s)	1.5	0.7	0.5
I_{leak} (μA)	0.98	1.3	1.5
NM (mV)	125	150	155

clock frequency and load capacitance of 1.8 V, 0.4 V, 1 GHz and 100 fF respectively using 180 nm CMOS technology parameters. During simulation, it is assumed that the compliment of all the inputs are available. The simulations of the proposed D-PFSCL XOR2 gate are carried out to get an insight on the impact of δ (the factor by which the threshold voltage of the middle transistor is reduced with respect to the outer transistors in PDN) on current flow. By assuming inputs, $A = \overline{B} = 1$ in proposed XOR2 (Fig. 6), the ratio of currents in M_{N3} and M_{N4} (I_{MN3}/I_{MN4}) is plotted in Fig. 7 for various value of δ. The correct operation requires that maximum current should flow through M_{N4}. To achieve this, it is observed that high value of δ increases the current flow through middle transistor, thereby minimizing the current ratio. A value of $\delta = 1.5$ is chosen for simulation since the further increase in its value does not cause much reduction in the current ratio.

The realization of XOR2 in existing D-PFSCL scheme is shown in Fig. 8. A STB is inserted between the stages in accordance with the existing scheme. The simulation waveform of XOR2 realized using the existing and the proposed scheme for D-PFSCL style is shown in Fig. 9 and is examined for the marked time duration (A-B-C-D-E). In the durations (A-B) and (C-D), the CLK input is at low logic level (0 V), the circuit works in precharge phase and the XOR2 gate outputs for both the schemes is precharged to high logic level (1.8 V). Also, any changes in the inputs do not influence the output. Further, in the intervals (B–C) and (D-E), the circuit works in evaluation phase. In the interval B–C, for high value of inputs A and B, the output is at low logic level (1.4 V). In the interval D-E, for low value of input A and high value of input B, the output is at high

Table 8
Pre and post layout simulation results at 65 nm technology node.

parameter	Technology		
	Pre layout results	Post layout results	Percentage Increase
τ_{PHL} (ps)	59.3	73.53	23.9%
τ_{pre} (ps)	23.6	27.56	16.8%
Power (μW)	57.46	54.464	5.5%
EDP ($x10^{-27}$J.s)	50.514	73.617	45.7%
I_{leak} (μA)	1.5	2.045	36.33%

logic level (1.8 V). So, the XOR2 gate based on both the schemes exhibits the same behaviour. However, a close observation of the waveforms reveals a reduction in delay for proposed scheme based gate in comparison to existing scheme based gate. The other logic functions namely 2:1 MUX, two input NAND (NAND2), three-input exclusive-OR (XOR3) are realized and similar behaviour is observed.

Further, to investigate further, the above logic functions are also implemented and simulated using static PFSCL and dynamic CMOS styles. Their performance is compared in terms of parameters namely gate count, transistor count, power [18], delay and other parameters. For the sake of fair comparison, the aspect ratios of transistors is maintained same in all the styles. The simulation results are listed in Table 3 and are discussed.

1. The realizations of two-input NAND2, XOR2, XOR3, 2:1 MUX using the existing scheme requires three to nine D-PFCSL gates in contrast to one to three in the new scheme. Thus, gate count in the proposed scheme is minimum among PFSCL variants, which directly affects the performance of the circuits. It can be seen that the circuit based on the proposed scheme outperforms the existing D-PFSCL circuit in all performance parameters.

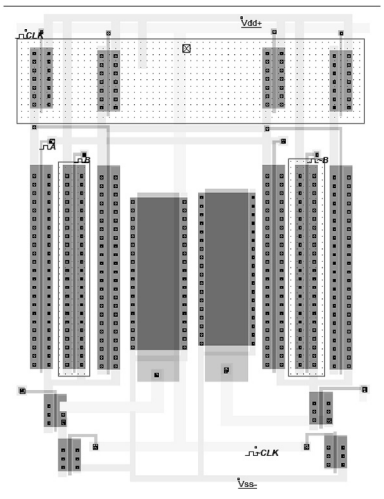


Fig. 13. Layout of the proposed D-PFSCL XOR2 gate.

2. It can also be observed that the proposed scheme circuits consume less power than the static PFSCL counterparts. The proposed circuits show both dynamic and leakage power whereas the static PFSCL circuits consume static power due to the presence of constant current source. In terms of delay, the static PFSCL circuits have large delay values as the cascading of NOR gates add to the delay.
3. The simulation results of the proposed D-PFSCL circuits indicate better performance than the dynamic CMOS circuits in terms of power, EDP and leakage current values. This is due to the fact that the proposed D-PFSCL are reduced swing in comparison to the full swing behaviour of the dynamic CMOS circuits.

At the block level, 8:1 multiplexer is considered. It is implemented and simulated in all the logic styles and their performance is compared. The block diagram of the 8:1 multiplexer in the proposed scheme is shown in Fig. 10. A performance summary is drawn in Table 4. It can be observed that the 8:1 multiplexer realized using proposed scheme outperforms in comparison to the other styles.

4.2. Examination of proposed D-PFSCL XOR2 gate

The impact of process variations is analysed by performing Monte Carlo for 500 simulation runs. The variations in output of the proposed D-PFSCL XOR2 gate when both the inputs A and B are at logic high level is shown in Fig. 11. A pictorial histogram showing the number of samples and the variation in τ_{PHL}, precharge delay, voltage swing and the dynamic power dissipation for the proposed D-PFSCL XOR2 gate is shown in Fig. 12 respectively. Similar, variations in τ_{PHL}, precharge delay, voltage swing were observed for XOR2 gates in all the styles. The mean and the variance in the performance parameters for different logic styles are listed Table 5. The results indicate that the proposed D-PFSCL gate shows more sensitivity towards variations in comparison to dynamic CMOS circuits. Further, PVT analysis was carried out to check the behaviour of the proposed circuit under best case and worst case conditions and the results in all styles are listed Table 6. It may be observed that the proposed circuit works correctly under best case and worst case PVT variation. The impact of technology scaling on the behaviour of proposed D-PFSCL XOR2 is studied by implementing it at 90 nm, 65 nm CMOS technology parameters and the simulation results are summarized in Table 7. It may be observed that the delay and the dynamic power improve with technology scaling while the leakage current increases. It is also noted that noise margin of the gate increases with lower node for constant voltage swing [19].

In the proposed XOR2 gate, the LVT transistor available in 65 nm technology is used. The value of δ was found to be 1.3. It is found that the circuit function correctly and the values of different parameters are found as τ_{PHL} (68 ps), τ_{pre} (43.8 ps), power (111 μW), I_{leak} (0.87 μA). Further, the layout of the proposed XOR2 gate is made for 65 technology node. The parasitics were extracted and the pre and post simulations results are noted and listed in Table 8. The values of τ_{PHL} and τ_{pre} have increased by 23.9% and 16.8% respectively leading to a corresponding increase in the EDP values. Further, an increase of 36.33% is measured in the post simulation results of leakage current. The layouts of the XOR2 gates in all the logic style were made. The layout of the proposed D-PFSCL gate is shown in Fig. 13. The area occupied by XOR2 gate realized using dynamic CMOS style, static PFSCL, existing scheme based and the proposed scheme based are 50 μm^2, 60 μm^2, 250 μm^2, 100 μm^2 respectively. The results clearly shows that the dynamic CMOS circuits have the least area requirement.

5. Conclusion

This paper puts forward a new scheme for logic function realization in D-PFSCL style. The proposed scheme modifies the basic structure of a D-PFSCL gate in order to implement AND/OR functionality in a single gate. This results in significant reduction of gate count which ultimately improves the performance of the complex D-PFSCL circuits in comparison to the existing scheme based circuits. Different functions based on the proposed and the existing schemes are designed and simulated by using TSMC 180 nm CMOS technology parameters. The simulation results confirms the theoretical prepositions. Hence it is expected that the proposed scheme for PFSCL circuit realizations will provide an efficient design option to designers.

Author statement

All authors have contributed equally in preparing the manuscript.

Declaration of competing interest

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

References

[1] G. Caruso, D. Di Sclafani, Analysis of compressor architectures in MOS current-mode logic, in: 2010 17th IEEE Int. Conf. Electron. Circuits Syst, IEEE, 2010, pp. 13–16, https://doi.org/10.1109/ICECS.2010.5724442.
[2] G. Caruso, A. Macchiarella, A design methodology for low-power MCML ring oscillators, in: 2007 18th Eur. Conf. Circuit Theory Des, IEEE, 2007, pp. 675–678, https://doi.org/10.1109/ECCTD.2007.4529686.
[3] S. Neshani, S.J. Azhari, A low-power low-voltage 6-bit 1.33 GS/s fully MCML all NMOS flash ADC without a front-end T/H, J. Circ. Syst. Comput. 22 (2013) 1350074, https://doi.org/10.1142/S0218126613500746.
[4] G. Shu, W.S. Choi, S. Saxena, M. Talegaonkar, T. Anand, A. Elkholy, A. Elshazly, P.K. Hanumolu, A 4-to-10.5 Gb/s continuous-rate digital clock and data recovery with automatic frequency acquisition, IEEE J. Solid State Circ. 51 (2016) 428–439, https://doi.org/10.1109/JSSC.2015.2497963.
[5] S. Han, T. Kim, J. Kim, J. Kim, A 10 Gbps SerDes for wireless chip-to-chip communication, in: 2015 Int. SoC Des. Conf, IEEE, 2015, pp. 17–18, https://doi.org/10.1109/ISOCC.2015.7401630.
[6] M. Damghanian, S.J. Azhari, A low-power 6-bit MOS CML flash ADC with a novel multi-segment encoder for UWB applications, Integration 57 (2017) 158–168. https://doi.org/10.1016/J.VLSI.2017.01.006.
[7] A.J. Mondal, A. Majudmer, B.K. Bhattacharyya, A design methodology for MOS current mode logic VCO, in: 2017 IEEE Int. Symp. Nanoelectron. Inf. Syst, IEEE, 2017, pp. 206–209, https://doi.org/10.1109/iNIS.2017.45.
[8] S. Archana, B.K. Madhavi, I.V.M. Krishna, Design of CMOS current mode and voltage mode Winner Take all circuit on 180nm technology, in: 2016 Int. Conf. Electr. Electron. Optim. Tech, IEEE, 2016, pp. 1879–1882, https://doi.org/10.1109/ICEEOT.2016.7755014.
[9] F. Saffar, M. Mirhassani, M. Ahmadi, A fault-tolerant area-efficient current-mode ADC for multiple-valued neural networks, in: 2012 IEEE 42nd Int. Symp. Mult. Log, IEEE, 2012, pp. 250–255, https://doi.org/10.1109/ISMVL.2012.31.
[10] M. Alioto, L. Pancioni, S. Rocchi, V. Vignoli, Modeling and evaluation of positive-feedback source-coupled logic, IEEE Trans. Circuits Syst. I Regul. Pap. 51 (2004) 2345–2355, https://doi.org/10.1109/TCSI.2004.838149.
[11] K. Gupta, N. Pandey, M. Gupta, Model and Design of Improved Current Mode Logic Gates, Springer Singapore, Singapore, 2020, https://doi.org/10.1007/978-981-15-0982-7.
[12] M.W. Allam, M.I. Elmasry, Dynamic current mode logic (DyCML): a new low-power high-performance logic style, IEEE J. Solid State Circ. 36 (2001) 550–558, https://doi.org/10.1109/4.910495.
[13] M.W. Allam, M.I. Elmasry, Dynamic current mode logic (DyCML), a new low-power high-performance logic family, in: Proc. IEEE 2000 Cust. Integr. Circuits Conf. (Cat. No.00CH37044), IEEE, n.d.: pp. 421–424. https://doi.org/10.1109/CICC.2000.852699.
[14] I. Hassoune, F. Mace, D. Flandre, J. Legat, Low-swing current mode logic (LSCML): a new logic style for secure and robust smart cards against power analysis attacks, Microelectron. J. 37 (2006) 997–1006.
[15] N. Pandey, D. Garg, K. Gupta, B. Choudhary, Hybrid dynamic MCML style: a high speed dynamic MCML style, J. Eng. 2016 (2016) 10. Article ID 8027150.
[16] K.P. Sai Pradeep, S. Suresh Kumar, Design and development of high performance MOS current mode logic (MCML) processor for fast and power efficient computing, Cluster Comput. 22 (2018).
[17] K. Gupta, N. Pandey, M. Gupta, Dynamic positive feedback source-coupled logic (D-PFSCL), Int. J. Electron. 103 (2016) 1626–1638, https://doi.org/10.1080/00207217.2016.1138519.
[18] S.-M. Kang, Y. Leblebici, CMOS Digital Integrated Circuits : Analysis and Design, McGraw-Hill, 2005.
[19] M. Alioto, G. Palumbo, Nanometer MCML gates: models and design considerations, in: In Proceedings of International Symposium on Circuits and Systems, Island of KOS. Greece, IEEE, 2006.



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 PhD 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 coauthored over 100 papers in international, national journals of repute and conferences.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

A Study on Stakeholder Management of an Engineering Project

Pranav Salhan
Department of Mechanical Engineering
Delhi Technological University
New Delhi, India

***Abstract* – Projects include people that may impact or be impacted by the project. If these people are not involved in the development of the project, then the project is unlikely to be successful. For the management of any Engineering project, stakeholder management is necessary for effective project relationships. Different stakeholders have different views and expectations, which calls for change and without a proper management system it can disrupt the flow of the project. The project manager needs to identify the various stakeholders and prepare a plan for their involvement in the project. This includes understanding their interest, power, and influence on the project. It is far easier to hand over deliverables to informed stakeholders when compared to someone who is unaware, which therefore leads to easier closure. The stakeholders can also help identify new risks and help understand existing ones. This study aims at formalizing the process of stakeholder management for optimum project delivery.**

Key words: Stakeholder, management, project, engagement

I. INTRODUCTION

The project stakeholder management process organizes the functions necessary to identify, classify, and manage the stakeholders in a methodical and logical way which helps in creating a plan that can be executed effectively.

What is a stakeholder?
A stakeholder is a person, group, or entity that may affect or be affected by the project. That means they have a vested interest in the project. Ultimately, this is anyone that may be impacted by the project.
There are mainly two types of stakeholders:
Internal stakeholders- They are people within the organization undertaking the project (such as manager, employees, board of directors, etc.).
External stakeholders- They are people or groups that don't work inside the business but are affected by the project. (such as customers, local government, suppliers, etc.).

One of the key problems with stakeholder management are the sheer number of people involved and the fact that their levels of power and interest differ markedly. Management of the stakeholder environment is a highly complex management task.
It is often during the early stages of a project that the stakeholder influence is felt and the need for effective communication is required for planning and smooth execution.

II. OBJECTIVE OF STUDY

-To identify project stakeholders, their roles, and interests.
-To develop a plan to manage the stakeholders and keep them informed.

III. STAKEHOLDER MANAGEMENT PLAN

There are four major components of the stakeholder management plan:
-Identifying stakeholders
-Planning stakeholder engagement
-Managing stakeholder engagement
-Monitoring stakeholder engagement

IV. IDENTIFYING STAKEHOLDERS

The first step is to identify the stakeholders. This stakeholder identification process is carried out both at the program level as well as at the project level with the program manager developing an initial identification of individual project level stakeholders. Some emerging practices in this area that have proven to be effective are to include everyone on the team in the process, to review the stakeholders continuously and to consult with them regularly. Some help can be taken from management professionals, project managers who have undertaken such projects before, or archived files for lessons learned.
Following is an example of a list of stakeholders identified for an engineering project.

Table I. Identified Stakeholders

Project	Chief Operating Officer Project Manager Project Leads Functional Manager s Project staff Project Advisors
Clients	Client-side Project Manager Client Staff Client Legal Team Client Business Operations
Contractors	Key Suppliers Operations Management
Financial Equity	Investors Debt Providers
Government and Local External	Government Agencies Special Interest Groups Labour Organizations and Trade Unions
Customers	Customers or facility Users

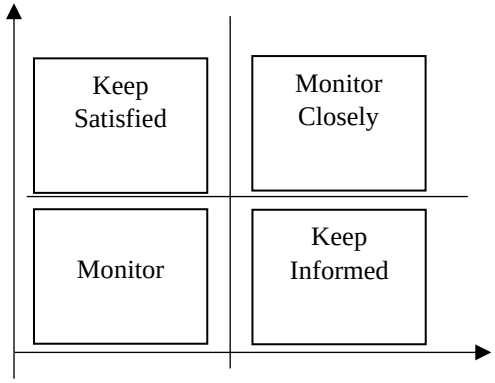
(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

V. PLANNING STAKEHOLDER ENGAGEMENT

A Stakeholder Engagement Plan is a formal strategy to communicate with project stakeholders to achieve their support for the project. It specifies the frequency and type of communications, media, contact persons, and locations of communication events. It is created at the beginning of the project and updated frequently as stakeholder communication needs change. There are many models which can be helpful in identifying the type of communication required for different type of stakeholders. Each model has its benefits, and one has to decide what is most important to them as they do their analysis. One such model often used is the power and influence grid. Using this tool, the stakeholder's power is assessed against their interest in the project. It's fairly simple but effective. If a stakeholder has little interest in the project and little power, one should monitor that person throughout the project, give them the information they need, but they don't need to be engaged in a very large way. If the stakeholder has high interest and low power, that person should be kept informed. They may not have the power to influence the project much, but people's interest and impact can change over time. Hence it is wise to keep them informed as to what is going on. If a stakeholder has little interest but lots of power, say a senior-level manager, it's wise to keep that person satisfied providing the information that shows that the project is progressing. They may not need daily updates but must be kept updated to the level appropriate. And finally, a stakeholder that has high interest and high power is someone that is needed to be monitored very closely. This person has the power to impact the project and it is necessary that this person is kept informed,
and up-to-date on the project.

Figure I. Power vs Interest Grid



In fig. I. the x-axis represents stakeholder's interest in the project and the y-axis represents power over the project. Another tool that helps organize the stakeholder is a stakeholder register. A stakeholder register helps record the stakeholders and their impact on the project through previously discussed prioritization. At a minimum, the stakeholder register should contain the person's name, role, type, and how they want to be communicated with, and how often. Given below is an example of stakeholder register entry for Shivang Bhardwaj, IT project lead working for an engineering firm.

Table II. Stakeholder Register Entry

Name of Stakeholder	Shivang Bhardwaj
Designation	Primary
Department	Information Technology
Role in project	IT Project Lead
Type of stakeholder	Internal
Type of communication	In-person
Interest	Leading
Influence on project outcome	High

VII. MANAGEMENT OF STAKEHOLDER ENGAGEMENT

Management of Stakeholder Engagement is the process of communicating and working with stakeholders to meet their needs/expectations, address issues as they occur, and foster appropriate stakeholder engagement in project activities throughout the project life cycle. The key components of this section are the project management plan, project documents, enterprise environmental factors, and any organizational requirements. The project manager needs to rely on interpersonal skills to effectively engage in meaningful communication with stakeholders. This may involve long conversations, listening to their needs, and resolving any concerns they may have. It will take management skills to facilitate negotiations and cooperation with everyone involved. It will take strong communication skills and methods to keep stakeholders informed using the appropriate modes and techniques. Management plan becomes very important at this point and should clearly articulate the who, what, how, and when people are communicating to the stakeholders.

Table III. Stakeholder Inputs, Tools and Techniques and Outputs

Inputs	-Stakeholder management plan -Communications management plan -Change log -Organizational process assets
Tools and Techniques	-Communication methods -Interpersonal skills -Management skills
Outputs	-Issue log -Change requests -Project management plan updates -Project documents updates -Organizational process updates

VIII. MONITORING STAKEHOLDER ENGAGEMENT

It is the responsibility of the project manager to monitor stakeholder relationships throughout the project with the use of best judgment and plan to assist him/her. Some additional things that can be used to help with this process are; issue logs, performance data, such as completion dates,
changes in costs and any documentation compiled up to this point as earlier. While monitoring these engagements, with the use of all information available is possible. This information can be gathered from various areas, but the bottom line is, keeping stakeholders informed using the tools implemented by creating the plan. Talking to the people on the team can help

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 06, June-2020

a lot. Gathered information should be dispersed to stakeholders as required. These engagements may result in updates to the project management plan or documentation, from any feedback received from stakeholders. The key is to control the level of information put out, since too much may become cumbersome, but not enough may be detrimental. Monitoring involves data analysis, data representation, communication skills, and decision making. This can be represented in a tabular form with the help of table III.

Table IV. Effective Monitoring Requirements

Data analysis	-Alternatives analysis -Root cause analysis -Stakeholder analysis
Decision making	-Multicriteria decision analysis -Voting
Data representation	Stakeholder engagement assessment matrix
Communication skills	-Feedback -Presentations
Impersonal and team skills	-Active listening -Cultural awareness -Leadership -Networking -Political awareness

CONCLUSIONS

Stakeholder's management has developed into an essential task of the project management process over recent years. The key benefit of stakeholder analysis is that it helps bring understanding to a complex situation and therefore helps project managers and teams to manage and communicate with stakeholders in the most effective way, enabling them to concentrate the resources where the maximum benefit will be derived and informing communications planning for the project.

REFERENCES

[1] Eyiah-Botwe, E. Aigbavboa, C.O and Thwala, W.D, "Critical Success Factors For Enhanced Stakeholder Management In Ghana" The Scientific Journal for Theory and Practice of Socio-economic Development 2016, 5(10), pp. 153-170

[2] Prof. Dr. Omar El-Naway, DR. Ibrahim Mahdi, Dr. Mohamed Badwy and Ahmed Gamal Al-Deen, "Developing Methodology for Stakeholder Management to Achieve Project Success" International Journal of Innovative Research in Science, Engineering and Technology Vol. 4, Issue 11, November 2015, pp.10652-10660

[3] Jing Yang, Geoffrey Qiping Shen, Manfong Ho, Derek S. Drew and Albert P. C. Chan, "Exploring Critical Success Factors For Stakeholder Management In Construction Projects" Journal Of Civil Engineering And Management 2009 15(4), pp.337–348

[4] Davis, J., MacDonald, A., and White, L. (2010). Problem-structuring methods and project management: an example of stakeholder involvement using Hierarchical Process Modelling methodology. The Journal of the Operational Research Society, 61(6), 893-904

[5] Eskerod, P., Huemann, M., and Savage, G. (2015). Project Stakeholder Management—Past and Present. Project Management Journal, 46(6), 6-14

[6] Hendry, J. R. (2005). Stakeholder Influence Strategies: An Empirical Exploration. Journal of Business Ethics, 61(1), 79-99

[7] Fulton, E. A., Boschetti, F., Sporcic, M., Jones, T., Little, L. R., Dambacher, J. M., . . . and Gorton, R. (2015). A multi-model approach to engaging stakeholder and modellers in complex environmental problems. Environmental Science & Policy, 48, 44-56

[8] Havard, L., Brigand, L., and Carino, M. (2015). Stakeholder participation in decision-making processes for marine and coastal protected areas: Case studies of the south-western Gulf of California, Mexico. Ocean & Coastal Management, 116, 116-131

[9] Davis, J., MacDonald, A., and White, L. (2010). Problem-structuring methods and project management: an example of stakeholder involvement using Hierarchical Process Modelling methodology. The Journal of the Operational Research Society, 61(6), 893-904

[10] Aaltonen, K. and Sivonen, R. (2009). Response strategies to stakeholder pressures in global projects. International Journal of Project Management, 27(2), 131-141

[11] Cascetta, E., Carteni, A., Pagliara, F., and Montanino, M. (2015). A new look at planning and designing transportation systems: A decision-making model based on cognitive rationality, stakeholder engagement and quantitative methods. Transport Policy, 38, 27-39

[12] Prieto, Robert. (2011). Stakeholder Management in Large Engineering & Construction Programs. PM World Today. XIII

[13] R. K., B. R. Agle, and D.J. Wood. (1997). "Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What really Counts." in: Academy of Management Review 22(4): 853 - 888

[14] Savage, G. T., T. W. Nix, Whitehead and Blair. (1991). "Strategies for assessing and managing organizational stakeholders." In: Academy of Management Executive 5(2): 61 – 75

[15] Cameron, B.G., T. Seher, E.F. Crawley (2010). "Goals for space exploration based on stakeholder network value considerations." in: Acta Astronautica doi:10.1016/j.actaastro.2010.11.003

[16] Gabriela Fernandes, Stephen Ward, Madalena Araújo, Isabel Loureiro, Ana Braga, "Perceptions of Different Stakeholders on Improving and Embendding Project Management Practice in Organisations." 2014 Procedia Technology 16 957 – 966

[17] Prof. Dr. Omar El-Naway, DR. Ibrahim Mahdi, Dr. Mohamed Badwy, Ahmed Gamal Al-Deen.,"Developing Methodology for Stakeholder Management to Achieve Project Success" November 2015 IJIRSET, ISSN (Online):2319-8753, ISSN (Print): 2347-6710, Vol. 4, Issue 11

[18] Hwang, B. G. and Ng, H. B. (2016). Project network management: risks and contributors from the viewpoint of contractors and sub-contractors. Technological and Economic Development of Economy, 22(4), 631-618

[19] Aki Aapaoja and Harri Haapasalo, "A Framework for Stakeholder Identification and Classification in Construction Projects" Open Journal of Business and Management, 2014, 2,pp. 43-55

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

Soft Computing
https://doi.org/10.1007/s00500-020-05005-4

METHODOLOGIES AND APPLICATION

A systematic literature review on empirical studies towards prediction of software maintainability

Ruchika Malhotra[1] · Kusum Lata[2]

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
Software maintainability prediction in the earlier stages of software development involves the construction of models for the accurate estimation of maintenance effort. This guides the software practitioners to manage the resources optimally. This study aims at systematically reviewing the prediction models from January 1990 to October 2019 for predicting software maintainability. We analyze the effectiveness of these models according to various aspects. To meet the goal of the research, we have identified 36 research papers. On investigating these papers, we found that various machine learning (ML), statistical (ST), and hybridized (HB) techniques have been applied to develop prediction models to predict software maintainability. The significant finding of this review is that the overall performance of ML-based models is better than that of ST models. The use of HB techniques for prediction of software maintainability is limited. The results of this review revealed that software maintainability prediction (SMP) models developed using ML techniques outperformed models developed using ST techniques. Also, the prediction performance of few models developed using HB techniques is encouraging, yet no conclusive results about the performance of HB techniques could be reported because different HB techniques are applied in a few studies.

Keywords Software maintenance · Software maintainability · Machine learning techniques · Statistical techniques · Hybridized techniques

1 Introduction

Due to innovation, advancement in technology, and an increase in customers' requirements, the complexity of software systems is continuing to increase. As a result, maintaining such complex software systems while minimizing the side effects of changes is a big challenge (Barry 1976). The majority of resources in many software organizations are consumed as a part of maintenance activity (Basili et al. 1996), and this phase accounts for more than 60% cost in software development (Somerville 2012). Thus, software maintenance being one of the pricey phases in the life cycle of software development requires the attention of researchers to accurately predict the software maintainability in the earlier stages of software development to take the decisions regarding resource allocation optimally and to develop a cost-effective, good-quality and maintainable software. In this direction, considerable effort has been put by the researchers over the years to provide various forms of measures for the accurate prediction of the maintainability of the software products. In the literature of software predictive modelling, software maintainability is seen as a software quality attribute that is defined according to numerous facets, and there is no clear correspondence to measure it. According to Coleman et al. (1994), Aggarwal et al. (2002), Schneberger (1997) and various others researchers, software maintainability is anticipated as pure subjective assessment like low, medium, and high based on combined measure of different software characteristics such as the readability of source code, software understandability, and quality of documentation. In 1993, Li and Henry (1993) proposed maintainability predictor based on object-oriented (OO) metrics that capture the characteristics of the OO software system. Their study predicted the

Communicated by V. Loia.

✉ Ruchika Malhotra
ruchikamalhotra2004@yahoo.com

Kusum Lata
kusumlata@dtu.ac.in

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

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

Published online: 28 May 2020

maintainability as a well-defined and quantifiable metric that can be determined on the basis of lines of source code added, deleted, and modified in the maintenance period. Later, various researchers (Mie et al. 2003; Aggarwal et al. 2006; Van Koten and Gray 2006; Zhou and Leung 2007; Malhotra and Chug 2012) attempted to forecast software maintainability by building the relation of software metrics with maintainability. Thus, in this way, various SMP models have been built by employing numerous ST, ML, and HB techniques to predict software maintainability at initial stages of software development to decide on resource allocation. There are many studies on maintainability prediction. To gain insight into the work done on SMP, it is essential to summarize the existing literature. Thus, this systematic review is strived to analyze and summarize the use of data sets, software metrics, techniques, performance measures, and statistical tests in the estimation of the lines of code added, deleted, and modified during the maintenance phase. To achieve this aim, we review the studies published from January 1990 to October 2019. Related reviews exists for the SMP (Riaz et al. 2009; Malhotra and Chug 2016). Riaz et al. (2009) conducted a systematic review on software maintainability prediction and metrics. Also, a systematic review of one of the co-authors of this study is published in the very recent past (Malhotra and Chug 2016). The earlier review by Malhotra and Chug (2016) provides an overview of the techniques for software maintainability. The study gave an overview of different techniques and tools to estimate the software maintainability, including maintainability index and refactoring methods. The study concluded that dynamic metrics should be proposed; agile methods should be used to improve maintainability, and analysis of effects of refactoring on maintainability should be done. As the study covered various aspects of software maintainability, it is difficult to compare the results with the results produced in the current systematic review. The present review is more concrete, as this systematic review provides a comprehensive and in-depth empirical analysis of the performance of various techniques, essentially ST, ML, and HB techniques for SMP. The study concentrates on the only prediction of maintainability. The outliers were removed before analysing the results produced by various techniques. Therefore, after rigorous selection criteria, we selected the studies measuring maintainability as a continuous variable. Unlike the earlier attempt, this provides an extensive analysis regarding (i) kind of data set used in software maintainability predictive modelling, (ii) kind of metrics used, (ii) techniques used for developing predictive models, (iii) widely used performance evaluators to judge the performance of developed models, (iv) use of statistical tests to validate the model results, (v) threats to validity reported in the empirical studies in the literature, and (vi) advantages and disadvantages of various techniques used for building models. On the ground of the outcome of the systematic review, we provide directions for future work for the researchers and software practitioners about the applicability of the prediction models for SMP. For the purpose of meeting our aim, we followed a systematic approach and 36 primary studies are selected after rigorous quality assessment.

The rest of the paper is organized as follows: Sect. 2 of the paper describes the procedure followed in conducting the review. This section includes research questions, search strategy, criteria for inclusion, and exclusion of studies, quality assessment of studies, and procedure followed for data extraction. Section 3 describes the publication sources of primary studies and their quality analysis. Section 4 presents the answers to the research questions formed in this paper. Section 5 reports the limitations of the review, and finally, Sect. 6 presents the conclusions and future work.

2 Method

This review is carried out as per the guidelines of Kitechenham (2007). Figure 1 shows the process followed for conducting this review. As per Kitechenham (2007) systematic review should consist of three stages: (i) planning the review, (ii) conducting the review, and (iii) reporting the review. The first stage incorporates identification of the need for systematic review, framing research questions to be answered by the systematic review, development protocol to conduct the review, and its evaluation. The research questions are the objectives that we keep in mind to address the demands of the systematic review. Development of review protocol encompasses designing an appropriate strategy to search relevant studies, formulating criteria for inclusion–exclusion, the formation of quality assessment criteria, preparation of forms for extraction of data, and deciding data synthesis methods. Relevant search criteria help to identify appropriate candidate studies to cater to the need for review. Based on criteria set for inclusion–exclusion, individual candidate studies are

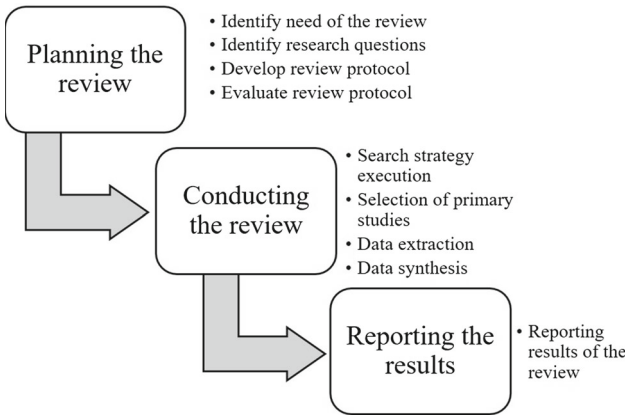


Fig. 1 Systematic review process

decided to be considered relevant or not relevant for the review. To assess the quality of each selected candidate study, assessment criteria in the form of the quality questionnaire have to be developed.

The next phase is conducting the review, which involves the execution of a search strategy to search the relevant studies, extraction of relevant facts out of the primary studies, putting the extracted facts in data extraction forms, and data synthesis. In data synthesis, tables, charts, boxplots, etc., are used to summarize the results of the systematic review. In this study, two independent researchers analysed the results separately, and inconsistencies, if any, were resolved after extensive discussions. In the following subsections, we describe the research questions, search strategy, inclusion–exclusion criteria, quality assessment criteria, and data extraction and data synthesis procedure in detail.

2.1 Research questions

As stated in Sect. 1, this review aims to systematically examine the various techniques and provide empirical evidence of their effectiveness for SMP. Thus, keeping in view the aim of the study, different research questions are formed and presented in Table 1. Firstly, we identify various techniques, data sets, and metrics used for SMP (RQ1, RQ2, and RQ3). In RQ4, we assess multiple measures to analyze the performance of prediction models. RQ5 determines the comparative performance of various techniques. In RQ6, we aim to find the generalizability of the prediction model/models. RQ7 tends to address a threat to validity reported in the studies. In RQ8, we tend to see the advantages and disadvantages of various techniques used in the literature for SMP.

2.2 The search strategy and study selection

The review aimed to select the most relevant and significant papers on the subject under study to provide concrete results of the investigation. Thus, the initial search was started by searching keywords such as "software maintainability", "software maintenance", "software maintainability prediction". After analyzing the studies retrieved after the initial search, we formed the following search string to select the most promising studies to be considered for review.

Software AND (maintainability OR maintenance effort) AND (prediction OR probability) AND (software metrics OR OO metrics OR procedural metrics) AND (classification OR regression) AND (machine learning techniques OR statistical techniques OR hybrid algorithm) AND (support vector machine OR decision tree OR neural network OR linear regression OR multiple regression OR multivariate regression OR genetic algorithm OR genetic programming OR evolutionary algorithm OR search-based techniques).

After the formation of a search string, most essential and appropriate digital libraries were selected for the extraction of relevant papers for the subject under study. The following digital libraries were used to search candidate primary studies by using the above search string.

- ACM Digital Library
- IEEE Xplore
- Science Direct
- Springer Link
- Wiley Online Library
- Web of Science
- Scopus

We searched the papers from January 1990 to October 2019. In total, 206 studies were extracted after exhaustively searching the mentioned digital libraries and by going through the reference sections of the studies extracted from the electronic libraries. During the course of our search process, we got some review studies whose reference sections were also used for extracting candidate primary studies but those review studies are not included as candidate studies in our review as these studies did not possess empirical results. After reading the abstract and introduction sections of those 206 studies, 107 studies (Basili et al. 1996; Coleman et al. 1994; Aggarwal et al. 2002; Schneberger 1997; Li and Henry 1993; Mie et al. 2003; Aggarwal et al. 2006; Van Koten and Gray 2006; Zhou and Leung 2007; Malhotra and Chug 2012; Riaz et al. 2009; Malhotra and Chug 2016; Khoshgoftaar and Szabo 1994a, b; Dagpinar et al. 2003; Elish and Elish 2009; Wang et al. 2009; Kaur et al. 2010; Jin and Liu 2010; Chandra 2012; Rana and Dash 2012; Sharawat 2012; Al-Jamimi and Ahmed 2013; Kaur and Kaur 2013; Olatunji 2013; Malhotra and Chug 2014a, b; Zhang et al. 2015; Elish and Aljamaan 2015; Bhutani and Chug 2015; Manchanda and Chug 2015; Mishra and Sharma 2015; Kumar and Rath 2015; Sharma and Chug 2015; Kumar et al. 2015; Kumar and Rath 2015; Jain et al. 2016; Kumar and Rath 2016; Tarwani and Chug 2016; Chug and Malhotra 2016; Malhotra and Lata 2017; Kumar and Rath 2017; Malhotra and Lata 2018; Kumar et al. 2019; Oman and Hagemeister 1992; Hirota et al. 1994; Daly et al. 1995; Hops and Sherif 1995; Lowther and Oman 1995; Niessink and Van Vliet 1997; Prechelt et al. 1998; Harrison and Counsell 1998; Ferneley 1999; Bengtsson and Bosch 1999; Bennett and Rajlich 2000; Harrison et al. 2000; Muthanna et al. 2000; Fioravanti and Nesi 2001; De Lucia et al. 2002; Land 2002; Arisholm 2002; Deligiannis et al. 2003; Genero et al. 2003; Bandi et al. 2003; Deligiannis and Shepperd 2004; Hayes 2004; Aggarwal et al. 2005; Hayes and Zhao 2005; Misra 2005; De Lucia et al. 2005; Broy et al. 2006; Heitlager et al. 2007; Perepletchikov 2007; Hatton 2007; Prasanth et al. 2008; Lagerström and Johnson 2008; Li et al. 2008; Zhou et al. 2008; Singh et al.

Table 1 Research questions

RQ. No	Research question	Objective
RQ1)	Which techniques have been used for SMP?	To identify techniques used for SMP so far
RQ2)	Which data sets have been used for SMP?	To identify data set for which SMP models have been developed
RQ3)	Which metrics have been used for SMP?	To investigate metrics as independent variables used for predicting software maintainability
RQ4)	Which performance evaluators have been used for SMP?	To understand how the performance of the modelling techniques was evaluated
RQ4.1)	What are the commonly used performance measures to judge performance of prediction models?	To assess performance measures on the basis of which performance of models can be judged
RQ4.2)	Which validation method has been used for developing models for SMP?	This research question identifies various cross-validation methods used for developing SMP models
RQ4.3)	Which statistical tests of significance have been used to statistically examine the predictive capability of SMP models?	To identify the studies in which results of SMP models are validated using statistical tests
RQ5)	What is the predictive ability of different techniques used for SMP?	To get insights about the predictive capabilities of various techniques for SMP
RQ5.1)	What is the performance of SMP models that are built with the help of ML techniques?	To identify the predictive capability of ML techniques
RQ5.2)	What is the overall performance of SMP of models built using ST techniques?	The aim of this question is to determine the predictive capability of ST techniques for SMP
RQ5.3)	What is the overall performance of SMP models developed using HB techniques?	This question aims to determine the predictive capability of HB techniques for SMP
RQ6)	Is there any one technique that outperforms the other in handling data sets with varying properties?	To identify the best technique for SMP
RQ7)	What are the potential threats to validity in SMP?	To identify most common threats to validity in the empirical studies for SMP
RQ8)	What are advantages and disadvantages of the techniques used for SMP?	To identify potential advantages and disadvantages of the techniques used

2009; Haiquan et al. 2009; Ping 2010; Anwar et al. 2010; Rizvi and Khan 2010; Kaur and Singh 2011; Ku et al. 2011; Bakar et al. 2012; Jeet et al. 2012; Basgalupp et al. 2012; Jeet and Dhir 2012; Singh and Balyan 2012; Bakota et al. 2012; Balogh et al. 2012; Hegedus 2013; Baqais et al. 2013; Ye et al. 2013; Al Dallal 2013; Riaz et al. 2013; Kumar et al. 2014; Kaur et al. 2014; Rai et al. 2015; Zavvar and Ramezani 2015; Jindal et al. 2015; Seref and Tanriover 2016; Sheshasaayee and Jose 2016; Kumar 2016; Jain and Sharma 2016; Almugrin et al. 2016) are considered for this review. These selected candidate studies were then subjected to the following inclusion–exclusion criteria.

Inclusion criteria:

- Empirical studies on SMP using CHANGE (line of code added, deleted, and modified) as the dependent variable.
- Empirical studies using ST techniques for developing a prediction model for SMP.
- Empirical studies using ML techniques for building a prediction model for SMP.
- Empirical studies using search-based techniques/HB techniques for developing a prediction model for SMP.
- Empirical studies comparing two or more techniques for SMP.

Exclusion Criteria:

- Studies in which the dependent variable is change-proneness.
- Studies in which no empirical evidence was provided for SMP. Studies where maintainability is a purely subjective measure like low, medium, and high.
- Studies in which software maintainability is based on expert judgment.
- Review studies.
- Studies with two similar papers of one author in conference and other in the journal, only the journal paper is considered.

Selected primary studies when subjected to the above inclusion–exclusion criteria yielded 36 studies which were

Table 2 Quality Assessment Questions

S. No	Question	Yes	Partly	No
1	Whether the objectives of the study are stated clearly?			
2	Whether the study states the maintainability facet in a clear manner?			
3	Is the scope of the study stated clearly?			
4	Does the study provide relevant literature?			
5	Are the predictor variables clearly stated and defined?			
6	Are the data collection and data pre-processing procedures described adequately?			
7	Does the study effectively justify the importance of prediction technique/techniques used?			
8	Do the research questions, purposes, or hypotheses logically flow from the introductory material?			
9	Are the research questions addressed clearly?			
10	If a study dealing with more than one prediction technique is the comparative analysis conducted?			
11	Does the study have sufficient number of citations?			
12	Does the results of study reported in clear manner?			
13	Are limitations of the study stated?			
14	Are the performance measures defined clearly and used to measure the predictive performance in the study?			

then subjected to quality assessment criteria given in the next subsection.

2.3 Quality assessment criteria

To provide concrete results of the review, it was essential that the selected primary studies should be of the utmost quality. So we designed quality assessment criteria in the form of 14 quality questions, as shown in Table 2. Each of the 36 selected primary candidate studies was then evaluated and ranked on 14 quality questions.

Each question in the quality questionnaire was assigned scores of 1(yes), 0(no), and 0.5(partially). The total score assigned to a study is defined as the aggregate of scores assigned to each quality question. Thus, the highest quality score that could be given to a study was 14, and the lowest was 0. The selected primary studies were ranked for quality by two reviewers independently. Each of the reviewers evaluated and ranked each the primary studies separately, and the final score assigned to a study was the average of the score assigned by each reviewer.

2.4 Data extraction and synthesis

For answering the RQs, it was required to hold the information extracted from all selected primary studies at one place to give conclusive and concrete results. For this purpose, we prepared data extraction forms to keep relevant information for answering all research questions. The various fields in the form were: name of the study, publication year, the name of conference or journal, data set used in the study, software metrics used, techniques used for SMP, performance measures, and statistical tests used. These data extraction forms were filled by two authors separately after going through each primary study.

After data extraction, data synthesis was necessary to answer all research questions. Two authors analysed data extraction forms separately. The disagreement, if any among the authors, was resolved by discussions, and final answers for each research question were prepared. The data were analysed keeping in mind both qualitative and quantitative perspectives. The quantitative analysis includes the performance for techniques for a particular data set, whereas qualitative analysis includes validation methods used, statistical tests used, metrics used, data set used. The results of RQs are presented with the help of tables, bar charts, line graphs, and boxplots.

3 Publication source of primary studies and quality analysis

In this section, a concise description of primary studies with their quality analysis is given. Also, year-wise distribution of studies selected for this review is presented in this section.

Table 3 Description of Primary Studies

Study#	Name	References	Study#	Name	References
SI1	Li (1993)	Li and Henry (1993)	SI19	Olatunj (2013)	Olatunji (2013)
SI2	Khoshgoftaar(1994a)	Khoshgoftaar and Szabo (1994a)	SI20	Malhotra (2014a)	Malhotra and Chug (2014a)
SI3	Khoshgoftaar (1994b)	Khoshgoftaar and Szabo (1994b)	SI21	Malhotra (2014b)	Malhotra and Chug (2014b)
SI4	Dagpinar (2003)	Dagpinar et al. (2003)	SI22	Zhang (2014)	Zhang et al. (2015)
SI5	Thwin (2005)	Mie et al. (2003)	SI23	Elish (2015)	Elish and Aljamaan (2015)
SI6	Aggarwal (2006)	Aggarwal et al. (2006)	SI24	Manchanda (2015)	Manchanda and Chug (2015)
SI7	Koten (2006)	Van Koten and Gray (2006)	SI25	Mishra (2015)	Mishra and Sharma (2015)
SI8	Zhou (2007)	Zhou and Leung (2007)	SI26	Kumar (2015)	Kumar and Rath (2015)
SI9	Elish (2009)	Elish and Elish (2009)	SI27	Sharma (2015)	Sharma and Chug (2015)
SI10	Wang (2009)	Wang et al. (2009)	SI28	Kumar (2015)	Kumar et al. (2015)
SI11	Kaur (2010)	Kaur et al. (2010)	SI29	Kumar (2015a)	Kumar and Rath (2015)
SI12	Jin (2010)	Jin and Liu (2010)	SI30	Kumar (2016)	Kumar and Rath (2016)
SI13	Chandra (2012)	Chandra (2012)	SI31	Tiwani (2016)	Tarwani and Chug (2016)
SI14	Dubey (2012)	Rana and Dash (2012)	SI32	Chug(2016)	Chug and Malhotra (2016)
SI15	Malhotra (2012)	Malhotra and Chug (2012)	SI33	Malhotra (2017)	Malhotra and Lata (2017)
SI16	Sharawat (2012)	Sharawat (2012)	SI34	Kumar(2017)	Kumar and Rath (2017)
SI17	Jamimi (2013)	Al-Jamimi and Ahmed (2013)	SI35	Malhotra (2018)	Malhotra and Lata (2018)
SI18	Kaur (2013)	Kaur and Kaur (2013)	SI36	Kumar (2019)	Kumar et al. (2019)

3.1 Description of primary studies and quality analysis

This section describes primary studies selected for this systematic review. Thirty-nine studies were considered after applying inclusion and exclusion criteria discussed in Sect. 2.2, which were subjected to quality assessment criteria (Table 2). Based on the quality assessment questionnaire presented in Table 2, the highest score that could be assigned to a study was 14, and the lowest was 0. Studies scoring greater than or equal to 8.5 were kept, while studies below a score of 8.5 were rejected. The rejected studies on the basis of quality score were (Bhutani and Chug 2015; Jain et al. 2016; Kumar et al. 2014). The studies scoring a score of greater than or equal to 8.5 are then sorted according to the ascending order of their publication year and listed in Table 3 along with their reference. Each primary study is assigned a unique identifier prefixed with SI (Study Identifier) for reference in the following subsections of paper. The maximum quality score was 13, which was gained by SI5, SI8, and a score of 12.5 was achieved by SI10, SI11, SI15, SI19, and SI23. Studies SI1, SI5, and SI8 are the high scorer in terms of citation count. The studies published in 2018 and 2019 were not having considerable citation count, so these studies were assessed for quality excluding Q11. After quality analysis, 36 studies were finally selected that were termed as primary studies. The studies (Bhutani and Chug 2015; Jain et al. 2016; Kaur et al. 2014) were rejected after quality analysis. The majority of the selected primary studies (53%) are published in the journals, and 44% of the studies are published in the proceedings of various conferences. One study is published as ACM SIGSOFT Software Engineering Notes.

3.2 Year-wise distribution of selected primary studies

Figure 2 shows the year-wise distribution of primary studies from January 1990 to October 2019. The first study for SMP was published in 1993 by Li and Henry (1993), where maintainability was taken as the dependent variable. Li & Henry proved that software maintainability could be best predicted from OO metrics. Later in 1994, two studies were published by Khoshgoftaar and Szabo (1994a, b) on maintainability prediction. Between 1994 and 2003, several studies were found, but these studies were not relevant to this review as these studies contributed towards maintainability measurement or contributed towards the measurement of maintainability index.

These studies were excluded because the objective of this systematic review is to explore the state of the art of SMP, not measurement. It has been observed that most of the work on SMP was started from the year 2003 when researchers agreed that maintainability could be predicted as the number of changes made to code during the maintenance period and software design metrics are helpful for quality prediction at early stages of software development life cycle. From 2003 onwards, researchers evaluated the effectiveness of various ST, ML, and HB techniques for maintainability prediction.

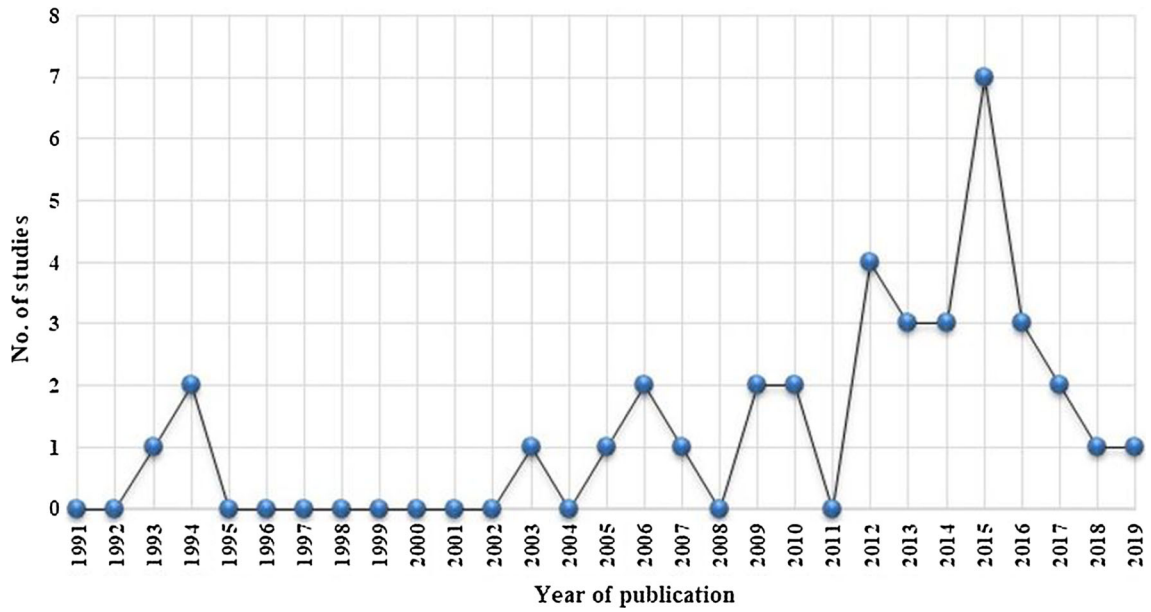


Fig. 2 Year-wise distribution of primary studies

4 Review results

4.1 RQ1. Which techniques have been used for SMP?

To accurately predict the maintainability of software, various techniques have been used by researchers and practitioners in the last two decades. These techniques tend to establish the relationship between predictor and response variables to build maintainability prediction models. We classify techniques used for SMP into the following three broad categories: (i) ST techniques, (ii) ML techniques, and (iii) HB techniques. Based on primary studies for SMP, we divide the ML techniques used for developing prediction models into the following categories:

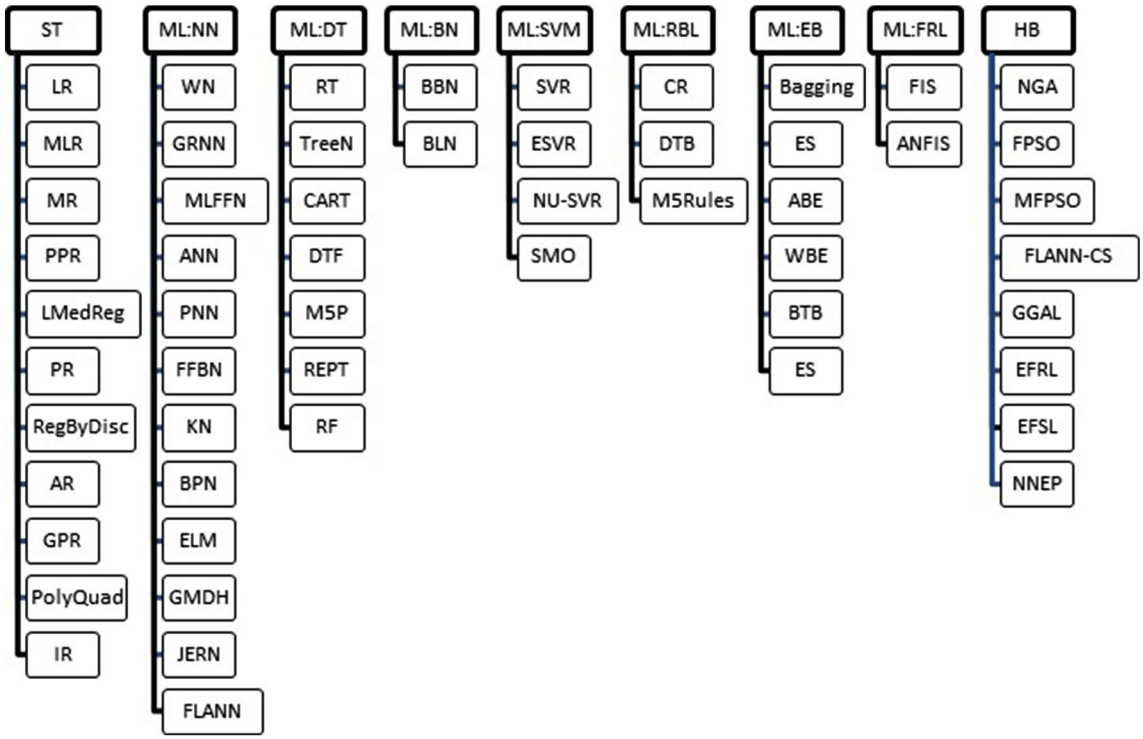
- Neural Networks (NN)
- Decision Trees (DT)
- Bayesian Networks (BN)
- Support Vector Machine (SVM)
- Instance-Based Learning (IBL)
- Rule-Based Learning (RBL)
- Ensembles (EB)
- Fuzzy Rule Learning (FRL)

Taxonomy of various techniques used (ST, ML, and HB) for SMP is presented in Fig. 3. Distribution of studies according to the above-mentioned techniques used for SMP along with study identifier is presented in Table 4.

It is evident from Table 4 that the majority of studies (44%) applied ML techniques for building maintainability prediction models; ST techniques are used in 8% of the studies, whereas HB techniques are used in 20% of the studies. The 25% of the studies have applied ML techniques and ST techniques combined, whereas 5% of the studies have applied all three types of techniques combined. There were 11 primary studies that applied more than one technique to compare the performance of a technique over the other one.

We found ML techniques to be the most prominent techniques used for SMP. Distribution of studies utilizing the above, different categories of ML techniques for developing SMP models is presented in Table 5. Among the primary studies using ML techniques, NN was the frequently used category, which was explored in 50% of the studies followed by SVM and DT, which were used in 26% and 27% of the studies, respectively. Further, studies using different categories of ML techniques for SMP given in Table 5 have applied different variants of each category. The most widely used algorithm in each category of ML technique along with study identifier is presented in Table 6.

GRNN is one of the most popular techniques in the category of NN for developing SMP models. The strengths such as ability to deal with sparse data, and to model nonlinear functions, and ability to work well in noisy environment made GRNN one of the most popular techniques in NN category. The primary studies used ST techniques for maintainability prediction, three studies applied linear regression (SI22, SI31, SI32), and another three studies applied multiple linear regression (SI3, SI7, SI19). HB techniques have been used in software engineering predictive modelling for fault prediction and predicting change proneness (Jin and Jin 2015; Kewen et al. 2014). These techniques tend to combine the advantages of (i) ST and search-based techniques



WN: Ward Network, GRNN: General Regression Neural Network, MLFFN: Multilayer Feed Forward network, ANN: Artificial Neural Network, FFBN: Feed Forward Back Propagation Network, MLP: Multilayer Perceptron, PNN: probabilistic Neural Network, KN: Kohan Network, BPN: Back Propagation Network, ELM: Extreme Learning Machines, GMDH: Group Method of Data Handling, JERN: Jordon Elman Recurrent Network, RT: Regression Tree, TreeN: TreeNet, CART: Classification and Regression Tree, DTF: Decision Tree Forest, M5P: Quinlan's M5 algorithm, REPT: Reduced Error Prone Tree, BBN: Baysian Belief Network, BLN: Baysian Learning Network, SVR: Support Vector Regression, ESVR: Epsilon Support Vector Regression, NU-SVR: NU- Support Vector Regression, SMO:Sequential Minimal Optimization, CR: Conjunctive Rules, DT: Decision Table, ES: Ensemble Selection, ABE: Average Based Ensemble, WBE: Weight Based Ensemble, BTB: Best in Training Based Ensemble, FIS: Fuzzy Inference System, ANFIS: Artificial Neuro Fuzzy Inference System, LR: Linear Regression, MLR: Multiple Linear Regression, MR: Multivariate Regression, PPR: Project Pursuit Regression, LMedReg: Least Median of Squares Regression, PR: Pace Regression, RegByDisc: Regression by Discretization, AR: Additive Regression, GPR: Gaussian Process Regression, PolyQuad: Polyquadratic, IR: Isotonic Regression, NGA: Neuro-Genetic, FLANN: Functional Link Artificial Neural Network, FPSO: Functional Link Artificial Neural Network with Particle Swarm Optimization, MFPSO: Modified Functional Link Artificial Neural Network Particle Swarm Optimization, GGAL: GRNN with Genetic Adaptive Learning, FLANN-CS: Functional Link Artificial Neural Network-Clonal Selection, EFRL: Evolutionary Fuzzy Rule Learning, EFSL: Evolutionary Fuzzy Symbolic Learning, NNEP: Neural Network with Evoluationary Programming.

Fig. 3 Taxonomy of techniques used for SMP

Table 4 Distribution of studies according to the techniques used

Technique used	Study Identifier	Count	Percentage
Only ST techniques	SI1, SI4, SI10	03	08
Only ML techniques	SI2, SI5, SI6, SI9, SI11,SI12-SI16, SI20, SI23, SI24, SI25, SI27, SI32	16	44
Only HB techniques	SI21, SI26, SI28, SI29, SI30,SI33, SI34, SI36	08	20
ST techniques & ML techniques	SI3, SI7, SI8,SI17, SI18, SI19, SI22, SI31, SI35	09	25
ST techniques, ML techniques & HB techniques	SI21, SI33	02	05

Table 5 Distribution of primary studies according to categories of ML techniques

ML Technique	No. of Studies	Percentage
NN	18	50
DT	10	27
BN	04	12
SVM	09	26
IBL	05	14
RBL	03	08
EB	03	08
FRL	03	08

and (ii) ML and search-based techniques as a single eminent technique. Eight SMP studies (SI21, SI26, SI28, SI29, SI30, SI33, SI34, SI36) explored the use of HB techniques. All HB techniques for SMP include combining ML and evolutionary or search-based techniques as a single technique. Out of eight studies using HB techniques for SMP, four studies reported the use of NGA in which NN is used for learning, and genetic algorithm (GA) is used for optimizing the weights of NN. SI21 has applied EFRL, EFSL, and NNEP techniques to develop SMP models. NGA is used is SI26, SI28, and SI29. HB techniques, namely FLANN, Adaptive FLANN, FPSO, and MFPSO, are used in SI30, whereas SI36 applied FLANN, the category of HB techniques, to develop prediction models of maintainability. We note that different studies explored different HB techniques to develop models to predict maintainability and no one HB technique can definitively be reported better than other.

4.2 RQ2. Which data sets have been used for SMP?

Over the years, researchers have experimented with a variety of data sets for predicting the maintainability of software systems. These data sets vary from proprietary software to public data sets. The data sets used for SMP are categorized as follows:

- Public data sets: In this category, User Interface Management System (UIMS) data set and Quality Evaluation System (QUES) data set were found to be the most widely used data sets for SMP. These data sets were collected by Li and Henry (1993) during the 3 years of maintenance of these two systems. Thus, UIMS and QUES data sets were together called Li & Henry data set. This data set has been used in 67% (SI1, SI5, SI6, SI7, SI8, SI9, SI10, SI11, SI13, SI14, SI15, SI16, SI17, SI18, SI19, SI23, SI25, SI26, SI28, SI29, SI30, SI33, SI34, SI36) of the primary studies selected for this review. Since these data sets are public, the results obtained using these data sets in various studies are easy to verify and replicate.
- Proprietary Software: These kinds of data sets include data sets collected from software developed by industrial professionals. Eleven per cent (SI2, SI3, SI4, SI20) of the selected primary studies have used proprietary software for SMP. Some of the proprietary software projects include File Letter Monitoring (FLM), Student Management System (SMS), Inventory Management System (IM System), Angel Bill Printing (ABP) System.
- Open-source projects: Open-source projects are used in 19% of the studies (SI21, SI22, SI24, SI27, SI31, SI32, SI35) selected for this review. Some of the open-source projects include Apache Poi, Apache Rave, Hodoku, ORDrumbox, jTDS, Art of illusion, etc. Most of these open-source projects were downloaded from software repositories: Sourceforge.net and Github.
- Student projects: This type of data set includes academic projects developed by students. Only one study (SI12) has dealt with such kind of data set.

Figure 4 shows the data set used and the corresponding percentage of studies in which these data sets are used.

4.3 RQ3. Which metrics have been used in SMP?

Various software metrics have been given in the literature to characterize and quantify various aspects of the software system such as size, complexity, modularity, complexity, etc. In the software engineering predictive modelling, the usefulness of these metrics has been very well proved. For SMP, the applicability of these lies in their ability to act as predictors or independent variables. To answer RQ3, we classified the selected primary studies based on software metrics used as predictor or independent variables.A similar kind

Table 6 The most popular algorithm from various categories of ML techniques according to the number of studies

ML Technique	Popular algorithm	Study identifier	Percentage
NN	GRNN	SI5, SI14, SI18, SI20, SI32	29
DT	RT	SI7, SI8, SI19	33
BN	BLN	SI17, SI19	50
SVM	SVR	SI12, SI13, SI17, SI23, SI31, SI32	66
IBL	KSTAR	SI18, SI22, SI24, SI32	100

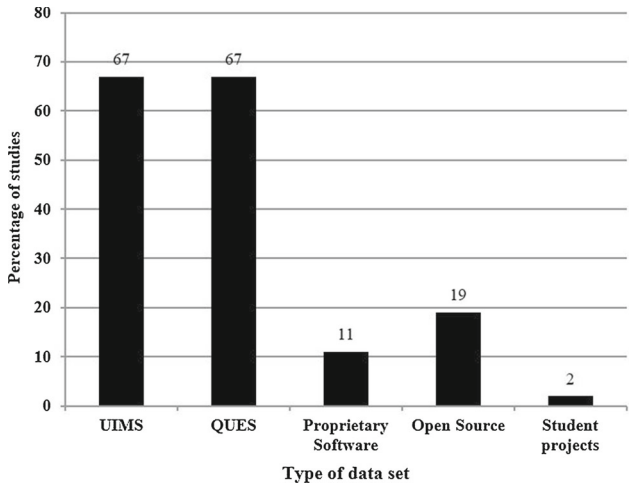


Fig. 4 Distribution according to the type of data set used for SMP

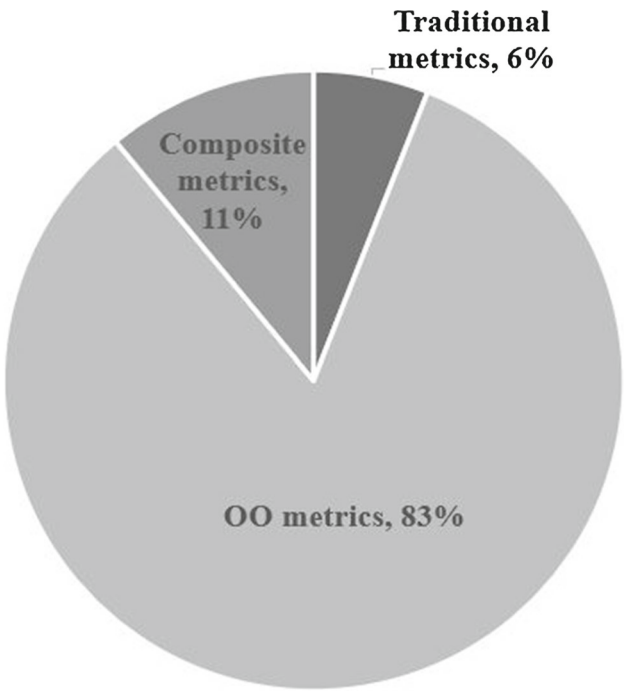


Fig. 5 Distribution of studies according to metrics used

of classification is given by Malhotra (2015) for software fault prediction. We found selected studies using both static software code metrics and dynamic metrics as independent variables for maintainability prediction. The categories of software metrics based on primary studies are given below:

- Studies using traditional software metrics as independent variables: These studies used traditional (procedural) software metrics as predictor variables given by Halstead (1977). These metrics include size metrics: lines of source code (SLOC), number of unique operands, number of unique operators, etc. Studies using traditional metrics are SI2, SI3.
- Studies using OO software metrics as independent variables: This category of studies used OO metrics as independent variables. Such metrics quantify the characteristics of OO systems such as modularity, inheritance, polymorphism, etc. Studies using OO metrics are: SI1, SI4, SI5, SI6, SI7, SI8, SI9, SI10, SI11, SI12, SI13, SI14, SI15, SI16, SI17, SI18, SI19, SI20, SI21, SI23, SI25, SI26, SI28, SI29, SI30, SI32, SI33, SI34, SI35, SI36.
- Studies using composite software metrics as independent variables: In this category, we identify two types of studies as follows: (i) studies using both traditional and OO metrics as predictors and (ii) studies using both OO metrics and dynamic metrics as predictors. The first category includes the studies that used OO metrics at the class level and traditional metrics at the method level. Studies using such a combination of metrics are SI22 and SI24. The second category includes studies that used OO metrics and dynamic metrics combined as predictors. Such studies are SI27 and SI31. The distribution of studies according to the kind of metrics used is depicted in Fig. 5.
  Among 83% studies using OO metrics, 86% of studies used the combination of Chidamber and Kemerer (CK) (1994) and Li and Henry (1993) metric suite, 7% of studies solely used CK metrics and in remaining studies CK and Li & Henry metrics are used with other metric suite like Henderson-Sellors (1996), Lorenz and Kidd (1994), Briand et al. (1999). Studies using a combination of CK, Li & Henry, and other OO metric suite are SI4 and SI32. Thus, CK and Li & Henry metric suite have been observed as one of the most popularly used OO metric suites for SMP.

4.4 RQ4. Which performance evaluators have been used for SMP?

To assess the effectiveness of the software maintainability predictor model, it is essential to examine it along with performance evaluators. This section describes the performance evaluators such as performance measures, validation methods, and statistical tests used in the literature to evaluate the effectiveness of SMP models.

4.4.1 RQ4.1 What are the commonly used performance measures to judge the performance of SMP models?

The performance of maintainability prediction models can be judged with the help of various performance measures. In the literature, multiple measures are used by the researchers to evaluate the performance of maintainability prediction models developed using different techniques discussed in

Table 7 Performance measures used in selected primary studies

Performance Measure	Description	Study identifier
Mean absolute error (MAE)	It is defined as the mean of the absolute difference between the actual and predicted values of the dependent variable	SI2, SI3, SI5, SI14, SI24, SI26, SI27, SI31, SI32, SI36
Magnitude of relative error (MRE)	It is defined as the ratio of the absolute difference between the actual and predicted values of the dependent variable to the actual value of the dependent variable.	SI6, SI7, SI8, SI11, SI12, SI13, SI15, SI17, SI18, SI19, SI20, SI21, SI23, SI26, SI28, SI29, SI30, SI33, SI35
Mean magnitude of relative error (MMRE)	It is defined as the mean of MRE's over all the data points	SI6, SI7, SI8, SI9, SI11, SI12, SI13, SI15, SI17, SI18, SI19, SI20,SI21, SI23, SI26, SI28, SI29, SI30, SI33, SI34, SI35, SI36
Maximum magnitude of relative error (maxMRE)	It measures the maximum of MRE value over all the data points	SI7,SI18,SI21,SI33
Root-mean-square error (RMSE)	It is defined as the square root of the average of the sum of squares of all of the errors	SI10, SI18, SI24, SI31, SI32, SI36
Pred(30)	It is defined as the percentage of predictions in which MRE is less than or equal to 30%.	SI7, SI8, SI9, SI15, SI17, SI18, SI19, SI20, SI21, SI23, SI33, SI35
Pred(25)	It is defined as the percentage of predictions in which MRE is less than or equal to 25%.	SI7, SI8, SI9, SI15, SI17, SI18, SI19, SI20, SI21, SI32, SI33, SI35
Coefficient of regression (R-Square)	It measures the closeness of the data point to the fitted regression line	SI1, SI4, SI5, SI8, SI14, SI15

Sect. 4.1. The definitions of the most frequently used performance measures along with the study in which they are used are given in Table 7. Figure 6 shows the percentage of studies using the performance measure given in Table 7. As shown in Fig. 6, MMRE is the most commonly used performance measure, which is used in 61% of the primary studies.

The other commonly used performance measures are Pred(25) and Pred(30), which are used in 33% and 33% of the studies, respectively. Few performance measures such as adjusted R-square, sum of squares of residuals, mean square of residuals, minimum absolute error, maximum absolute error, absolute residual error, normalized root-mean-square error, standard error of mean, true error, estimate of true errors, Spearmen rank correlation coefficient, relative absolute error (RAE) are clubbed into miscellaneous category as these were less commonly used.

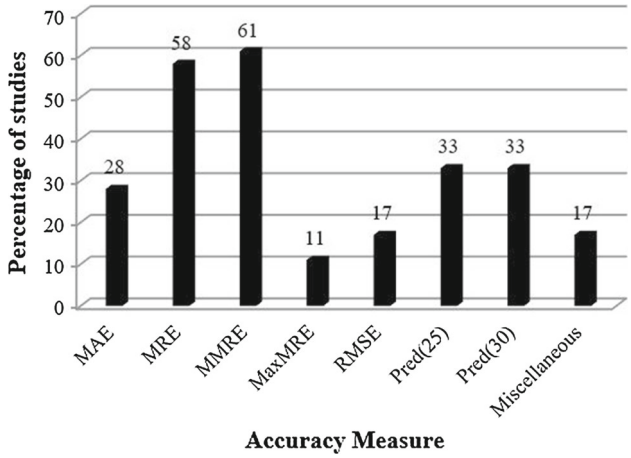


Fig. 6 Distribution of studies according to performance measures

4.4.2 RQ4.2 Which validation methods have been used for developing models for SMP?

The selection of a cross-validation method is essential to develop a prediction model effectively. Cross-validation methods tend to build a prediction model by dividing the available data set into two parts, namely: training data set and test data set. The training data set is used to train the model, while the test data set is used to evaluate the performance of the developed model. For SMP, three types of cross-validation methods have been used in the literature: K-Fold cross-validation, hold-out cross-validation, and leave one out. The brief description of these methods, along with the study identifier in which these methods have been used, is given in Table 8. Figure 7 presents the distribution of selected primary studies in accordance with the cross-validation used. Only 21 studies reported the validations methods used. K-fold cross-validation was used in 12 studies followed by hold-out and leave-one-out validation, which were used in four and five studies, respectively.

Table 8 Description of cross-validation method used in selected primary studies

Method	Description	Study Identifier
K-fold cross-validation	In this cross-validation method, the data set in hand is divided into K partitions each of equal size. In each iteration, K-1 partitions are used to train the model, while one out of K partitions is used to test the model. This process is repeated K times. The commonly used values for K were 5 and 10	SI5, SI7, SI23, SI24, SI26, SI28, SI29, SI30, SI31, SI32, SI35, SI36
Hold-out cross-validation	In hold-out cross-validation data set is split into two partitions: one partition is used for training purpose and another for validation. The commonly employed partition methods were 60–40, 70–30, and 75–25	SI15, SI19, SI22, SI27
Leave-one-out cross-validation	This validation method involves dividing data set with N observations into two partitions of sizes N-1 and 1. In each iteration, random N-1 data points are used for training and one data point is used for testing	SI8, SI9, SI10, SI18, SI34

4.4.3 RQ4.3 Which statistical tests of significance have been used to statistically examine the predictive capability of SMP models?

The use of the statistical test in predictive modelling strengthens the outcome of the developed models by statistically examining the conclusion drawn regarding the predictions given by the developed model. The use of statistical tests has been observed in 28%of the primary studies selected for this review. The distribution of the statistical test used is depicted in Fig. 8. Three studies used t test (SI23, SI30, S36); five studies used the Wilcoxon test (SI7, SI8, SI18, SI27, SI35) for comparative statistical analysis. Four studies (SI18, SI31, SI32, SI35) used Friedman ranking test, and Nemenyi test is used in one study (SI32). It is evident that the use of non-parametric tests such as the Wilcoxon test and Friedman test has been found in more number of studies as compared to parametric t test; reason is that parametric test requires data normality of the underlying data set to be true before the application of the test. In studies SI7, SI8, and SI33, the Wilcoxon test has been used to statistically prove the significant difference between the predictive performances of different techniques. In the study SI18, the Wilcoxon test is used as a post hoc analysis after the Friedman test. Nemenyi test is also used as a post hoc analysis after the application of the Friedman test in study SI32.

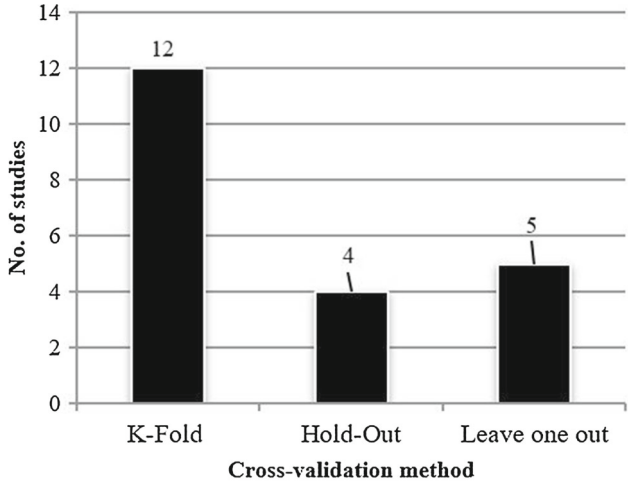


Fig. 7 Distribution of studies according to cross-validation methods

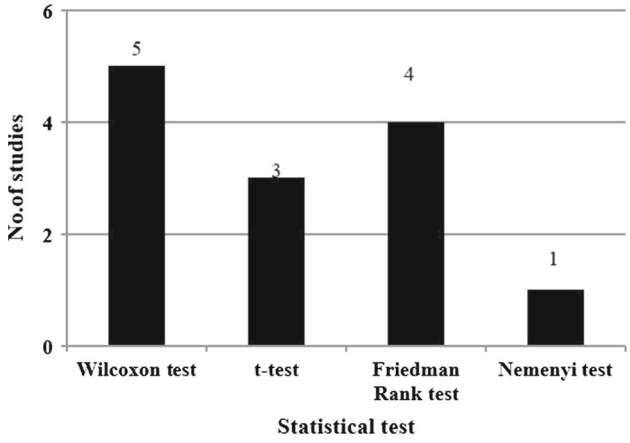


Fig. 8 Distribution of studies according to statistical tests

4.5 RQ5. What is the predictive ability of various techniques used for SMP?

This section presents the predictive capabilities of various techniques (ST, ML, and HB) used for developing SMP models. To ensure the usefulness of a technique for developing a prediction model, it is extremely important to evaluate its

Table 9 Performance statistics of SMP models using ML techniques

Technique	Count	Performance measure	Min.	Max.	Mean	SD
GRNN	9	MMRE	0.32	0.54	0.43	0.06
	10	PRED	0.26	0.78	0.59	0.15
ANN	3	MMRE	0.26	1.95	0.93	0.89
	2	PRED	0.15	0.37	0.26	0.15
FFBN	3	MMRE	0.45	0.51	0.48	0.03
MLP	4	MMRE	0.59	1.95	1.16	0.63
	2	PRED	0.13	0.37	0.25	0.16
PNN	7	MMRE	0.37	0.53	0.42	0.05
	8	PRED	0.51	0.73	0.62	0.06
KN	7	MMRE	0.35	0.56	0.46	0.07
	7	PRED	0.48	0.73	0.69	0.08
GMDH	9	MMRE	0.21	0.42	0.33	0.06
	9	PRED	0.51	0.79	0.68	0.08
RT	5	MMRE	0.49	1.53	0.91	0.54
	6	PRED	0.10	0.41	0.26	0.11
M5P	4	MMRE	0.54	1.67	1.00	0.56
	2	PRED	0.28	0.32	0.30	0.02
BN	4	MMRE	0.45	0.97	0.71	0.30
	4	PRED	0.39	0.44	0.41	0.03
SVM	9	MMRE	0.22	0.67	0.45	0.14
	8	PRED	0.31	0.65	0.46	0.10
KSTAR	9	MMRE	0.27	0.77	0.52	0.14
	9	PRED	0.36	0.72	0.53	0.10

predictive capability. The predictive performance of various techniques used for SMP is evaluated by summarizing the values of the performance measures described in Sect. 4.4.1. We report the predictive performance of techniques that have been used in at least two primary studies to get more generalized results. The following subsections describe the predictive performance of ML, ST, and HB techniques.

4.5.1 RQ5.1 What is the performance SMP models that are built with the help of ML techniques?

To evaluate the performance of ML techniques discussed in Sect. 4.1, we recorded the MMRE and Pred(25) values of different ML techniques on different data sets. We analyze the predictive performance of only those ML techniques that are used in at least two primary studies so that generalized results can be reported. MMRE and Pred (25) are selected for the analysis as these are the most commonly used performance measures, as discussed in Sect. 4.4.1. We present the descriptive statistics such as minimum, maximum, mean, and standard deviation of performance measures: MMRE and Pred (25) in Table 9.

A good maintainability prediction model should have low MMRE and high Pred(25) values. We report the MMRE and Pred values in Table 9 after removing the outliers to minimize the prediction bias as we observed that some techniques perform exceptionally different on some data sets resulting in very biased predictions. To overcome this bias, outlier analysis was essential, and we used boxplots to gain insights about the outliers. These outliers were removed before reporting the statistics of performance measures MMRE and Pred(25). Figures 9 and 10 show the results of outlier analysis done with the help of boxplots for the two prominent performance measures, namely MMRE and Pred(25), respectively. As shown in Table 9, GMDH technique reported the best mean MMRE of 0.33, followed by PNN and GRNN, which gave mean MMRE of 0.42 and 0.43, respectively. It is to be noted that the mean MMRE of GMDH, PNN, GRNN, SVM, KN, KSTAR, and FFBN ranged from 0.33 to 0.52. KN has reported the best mean Pred(25) value of 0.69 followed by GMDH with a Pred(25) value of 0.68. Again as shown in Table 9, mean Pred(25) values of GMDH, PNN, GRNN, SVM, KN, KSTAR, and FFBN ranged from 0.46 to 0.69. Therefore, it is quite evident that the performance of GMDH, PNN, GRNN, SVM, KN, KSTAR, and FFBN with respect to MMRE and Pred(25) is quite acceptable. MLP performed worst with mean MMRE of 1.16 and mean Pred(25) value of 0.25. Thus, the predictive capability of ML techniques: GMDH, PNN, GRNN, SVM, KN, KSTAR, and FFBN, is

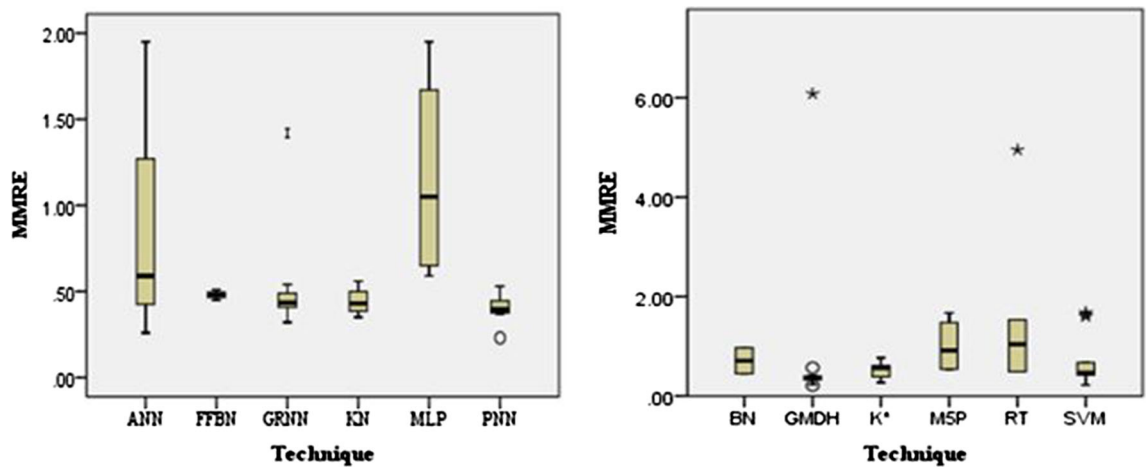


Fig. 9 Boxplots for MMRE for ML techniques according to different data sets

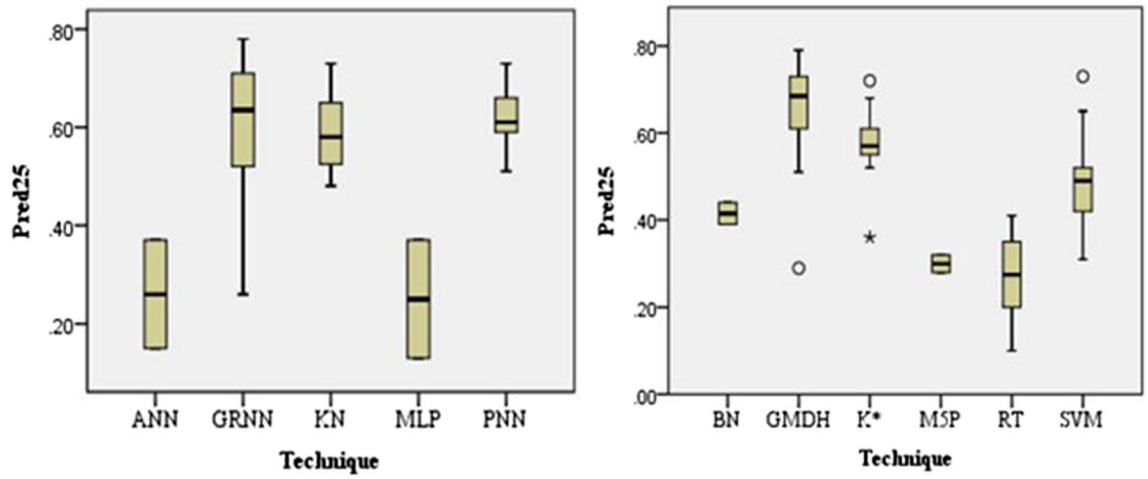


Fig. 10 Boxplots for Pred(25) for ML techniques according to different data sets

close to an acceptable level as per Conte et al. (1986) and Wen et al. (2012).

4.5.2 RQ5.2 What is the overall performance of SMP models built using ST techniques?

A variety of ST techniques described in Sect. 4.1 have been applied to develop models for SMP. Statistical techniques are used in 14 selected primary studies: SI1, SI3, SI4, SI7, SI8, SI10, SI17, SI18, SI19, SI21, SI22, SI31, SI32, and SI34. Out of studies that used ST techniques for SMP, 11 compared ST and ML models. To get generalized results about the performance of models developed using ST techniques, we evaluated the results of ST techniques, which are used in at least two primary studies on the basis of performance measures: MMRE and Pred(25). In this regard, we found LR and MLR as the prominent ST techniques used for SMP. We record the performance statistics of LR and MLR techniques in Table 10 after outlier analysis using boxplots. Figure 11 shows the result of the outlier analysis of LR and MLR techniques for performance measures MMRE and Pred(25). It is evident from Table 10 that MMRE of prominent ST techniques ranges from 0.53 to 1.50, and Pred(25) ranges from 0.28 to 0.59. On comparing the performance statistics of ST techniques (Table 10) with ML techniques (Table 9), it may be noted that ML techniques: GMDH, KN, GRNN, PNN, SVM, FFBN, and KSTAR, outperform ST techniques LR and MLR.

4.5.3 RQ5.3 What is the overall performance of SMP models developed using HB techniques for SMP?

Different kinds of HB techniques such as FPSO, NGA, MFPSO, FLANN-CS, EFRL, EFSL, and NNEP are used in various studies. For instance, SI21 has applied EFRL, EFSL, and NNEP techniques to develop SMP models. NGA is used is SI26, SI28, and SI29. HB techniques, namely FLANN, Adaptive FLANN, FPSO, and MFPSO, are used in SI30,

Table 10 Performance statistics of SMP models using ST techniques

Technique	Count	Performance Measure	Min.	Max.	Mean	Std. Dev.
LR	7	MMRE	0.4	0.66	0.53	0.09
	7	PRED	0.54	0.63	0.59	0.03
MLR	6	MMRE	0.39	2.70	1.50	1.17
	6	PRED	0.15	0.42	0.28	0.12

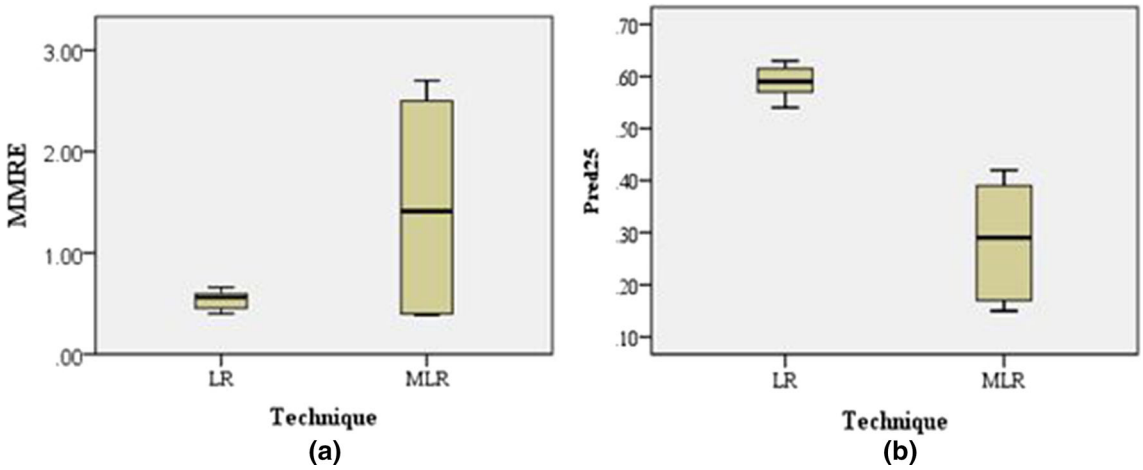


Fig. 11 Boxplots for (a) MMRE (b) Pred(25) for ST techniques according to different data sets

whereas SI36 applied FLANN, the category of HB techniques, to develop prediction models of maintainability. In total, eight selected primarily studied and used HB techniques for SMP. Unlike ML and ST techniques, cumulative performance statistics of each HB technique could not be reported as each study has applied different HB techniques to predict maintainability. It has been recorded that MMRE of HB techniques in different studies ranged from 0.22 to 2.1, whereas Pred(25) ranged from 0.19 to 0.64. Thus, by looking at these performance statistics of HB techniques, no valid conclusion about their applicability for SMP could be drawn.

4.6 RQ6 Is there any one technique that outperforms other techniques in handling multiple data sets with varying properties?

In order to get insight about which technique is best in terms of its performance on data sets with differing properties, we record the performance (in terms of MMRE) of best techniques (Sect. 4.5.1) on different data sets from different primary studies according to the same selection criteria discussed in Sect. 4.5.1 and present results in Table 11. From Table 11, it is very much clear that the GMDH technique shows lower MMRE on the majority of the diverse data sets as compared to other best techniques: GRNN, PNN, KSTAR, KN, FFBN, and SVM. On the majority of the data sets, MMRE of GMDH technique ranged from 0.21 to 0.37, which is quite close to the acceptable MMRE as suggested by Conte et al. (1986). Thus, it is quite conclusive that GMDH technique is emerged as the best technique in terms of its predictive performance irrespective of the data sets compared to other best techniques depicted in Table 11.

4.7 RQ7 What are the potential threats to validity in SMP and threat mitigation measures?

For any empirical study, the threats to the validity of the results of the study are a quite important concern as the threats limit the applicability of the results. After the analysis of the results of the empirical study, it is tremendously essential for the researchers to list down the validity threats of the results. Only 39% of selected primary studies (SI13, SI15, SI17, SI18, SI20, SI26, SI27, SI29, SI30, SI32, S33, SI34, SI35, and SI36) for this systematic review reported the threats to the validity of results. We classified the threats to validity reported by the selected primary studies into two types, namely threats to external validity and threats to internal validity as per classification given by Campbell et al. (1963). The classification of threats to validity reported by primary studies is given in Table 12. Out of 39% studies that reported threats, the threat to the applicability of results across the different programming languages and small size of data set were reported by the majority of the studies. Very few studies (SI18 and SI32) have reported the threat mitigation measures of some of the threats. Study SI18 lessened

Table 11 Performance of techniques according to different data sets

Best technique	Data set	MMRE	Best technique	Data set	MMRE
GRNN	QUES	0.41	KSTAR	Log4j	0.39
GRNN	FLM,EASY	0.54	GMDH	QUES	0.21
GRNN	Drumkit	0.38	GMDH	Jedit	0.38
GRNN	OpenCV	0.42	GMDH	Junit	0.42
GRNN	Abdera	0.45	GMDH	FLM,EASY	0.35
GRNN	Ivy	0.41	GMDH	Drumkit	0.23
GRNN	Log4j	0.32	GMDH	OpenCV	0.37
GRNN	Jedit	0.49	GMDH	Abdera	0.32
GRNN	QUES	0.41	GMDH	Ivy	0.34
FFBN	UIMS	0.51	GMDH	Log4j	0.37
FFBN	QUES	0.48	KN	Drumkit	0.56
FFBN	FLM,EASY	0.45	KN	OpenCV	0.48
PNN	Drumkit	0.38	KN	Abdera	0.37
PNN	Jedit	0.41	KN	Ivy	0.40
PNN	Ivy	0.53	KN	Log4j	0.35
PNN	OpenCV	0.42	KN	Jedit	0.43
PNN	Abdera	0.38	KN	Junit	0.52
PNN	Log4j	0.37	SVM	Junit	0.54
PNN	junit	0.47	SVM	Student Project	0.22
KSTAR	Abdera	0.55	SVM	Log4j	0.39
KSTAR	Jedit	0.56	SVM	Drumkit	0.64
KSTAR	Ivy	0.38	SVM	QUES	0.36
KSTAR	junit	0.61	SVM	Jedit	0.48
KSTAR	UIMS	0.56	SVM	Abdera	0.47
KSTAR	QUES	0.27	SVM	Ivy	0.42
KSTAR	Drumkit	0.67	SVM	OpenCV	0.67
KSTAR	OpenCV	0.77			

the threat of small size of the data set by employing leave-one-out cross-validation. Study SI32 eliminated the threat of generalizability of results by taking seven open-source software systems with different sizes, different characteristics, and maintenance requirements.

4.8 RQ8 What are the strengths and weaknesses of the techniques used for SMP?

To predict maintainability, we may concern to pick up the most appropriate estimation model developed with the application of a particular technique. For this purpose, it is crucial to present the strengths and limitations of various techniques used to create the prediction model. It has been observed in the literature that different studies have applied diverse techniques such as ML, ST, and HB techniques for building prediction models for SMP. Only a few of the ML techniques such as GMDH, FFNN, KN, GRNN, PNN, SVM, FFBN, KSTAR, and ST techniques such as LR, MLR are observed to be replicated in the literature. Hybridized techniques are hardly replicated in any study. Table 13 describes the strengths and weaknesses of different techniques used along with the study identifier in which these are reported. The strengths of different techniques are rational according to the studies in which these are reported, which makes it tough to conclude a specific situation that strengthens the application of a particular technique. Nonetheless, to cut down the study bias, we present the strengths and limitations of techniques given by at least two primary studies.

5 Limitations of review

This systematic review aims to evaluate the effectiveness of SMP models developed in the literature. To do so, digital libraries described in Sect. 2.2 are exhaustively searched, and 36 primary studies are selected for this review. Still, there may exist a threat that few relevant studies may not be included.

Table 12 Classification of threats to validity reported by selected primary studies

S. No.	Threat	Threat description	Study identifier
Threats to external validity			
1.	Applicability of results across languages.	The study considered the data set extracted for software systems written in a particular programming language	SI13, SI15, SI20, SI27, SI29, SI30, SI33, SI35, SI36
2.	Data set of inadequate size	The size of software system evaluated in the study is quite low which affects the result generalizability	SI17, SI18, SI27, SI33
3.	Application of results across different variables	Independent variables are collected from the software systems that have specific characteristics and results cannot be generalized.	SI20, SI34, SI36
4.	Result bias because techniques used	The study applied only a few techniques, and effect of other prominent techniques on the same data set needs to be judged	SI15, SI27, SI29
5.	Software system used has specific characteristics	Data set is extracted from the software system with specific characteristics, and results cannot be generalized	SI32, SI36
Threats to internal validity			
6.	Confounding effect of independent variables	The study does not account of the impact of variables on maintenance other than those selected in that particular study	SI15, SI29, SI30, SI32
7.	Influence of human factors	The study does not account that maintenance also depends on external quality attributes such as developer's competence, familiarity with code, etc.	SI21, SI36

In order to evaluate the performance of SMP models developed using techniques such as ML, ST, and HB, we considered only performance measures MMRE and Pred(25). These performance measures were selected because in the majority of the selected primary studies (Sect. 4.4.1), these metrics are used to evaluate the performance of the models. Other performance measures such as RMSE, MAE, and NRMSE are also used to assess the performance of the prediction models in different studies. So there exists a threat that the performance of models, if evaluated using other than MMRE and Pred(25), may vary the results. We have extracted the values of performance measures MMRE and Pred(25) to evaluate the performance of models developed with varying experimental settings like data sets, independent variables, and cross-validation methods employed. These elements of experimental settings may affect the prediction performance. This threat also exists in this study, but from a different frame of reference, prediction results of studies conducted with different experimental settings are considered to be more robust in comparison with the results synthesized from studies with similar experimental settings.

As a major finding of this review, we compare the overall performance of prediction models developed using the ML technique and those with ST and HB techniques. But the ST techniques are compared to those with the ML techniques in a very insufficient number of studies, which makes ML vs. ST comparison not much conclusive. Also, different HB techniques are used in different studies. Thus, no clear conclusion on the overall prediction capability of HB techniques for maintainability prediction could be drawn. Thus, the threat of result bias for the ST and HB techniques for SMP is there in this study.

6 Conclusion and future directions

This systematic review aimed to investigate the SMP models from various perspectives such as software metrics used, data set used, techniques employed, performance measures used, validation methods employed, and statistical test employed. We evaluated the performance of models developed using various techniques where maintainability is measured as the number of lines of code changed during the maintenance period and described as the continuous dependent variable in terms of the lines of code added, deleted, and modified. After an exhaustive search and going through rigorous quality assessment, 36 studies are selected for this review to answer the eight research questions. These studies have been published in the period from January 1990 to October 2019. The

Table 13 Strengths and weaknesses of techniques used for SMP

Name of the technique	Strengths	Study Identifier
GRNN	Ability to deal with sparse data. Fast to train the model and ability to perform well in the noisy environment	SI5, SI15, SI20, SI32
KN	Self-organizing network having the capability to represent multidimensional data in mush lower space	SI32
PNN	Ability to train quickly on sparse data sets, more accurate and insensitive to outliers	SI15, SI32
GMDH	It can predict outcome even with small data sets, less computation burden, and useful for constructing extremely high-order regression-type model	SI15, SI18, SI32
SVM	Ability to learn even in the case of high-dimensional feature space. Powerful regression capabilities	SI13, SI17, SI31, SI32
Name of the technique	Weaknesses	Study Identifier
PNN	Slow and requires more memory space	SI15, SI32

Table 14 Quality Assessment Questions

S. NO	Question	Yes	Partly	No
1	Whether the objectives of the study stated clearly?	89	11	0
2	Whether the study states the maintainability facet in a clear manner?	100	0	0
3	Is the scope of study stated clearly?	61	39	0
4	Does the study provide relevant literature?	86	8	6
5	Are the predictor variables clearly stated and defined?	100	0	0
6	Are the data collection and data pre-processing procedures described adequately?	92	8	0
7	Does the study effectively justify the importance of prediction technique/techniques used?	50	42	8
8	Do the research questions, purposes, or hypotheses logically flow from the introductory material?	58	39	3
9	Are the research questions addressed clearly?	50	47	3
10	If a study dealing with more than one prediction technique is the comparative analysis conducted?	50	44	6
11	Does the study have sufficient number of citations?	28	28	39
12	Are the results of the study reported in clear manner?	61	39	0
13	Are limitations of the study stated?	42	3	56
14	Are the performance measures defined clearly and used to measure the predictive performance in the study?	50	39	11

information discovered as a result of this review is concluded as follows:

- RQ1: Machine learning, ST, and few HB techniques have been applied in developing models for SMP. ML techniques are found to be the most prominent techniques applied for SMP in the studies we analysed and are divided into Neural Networks, Decision trees, Bayesian Networks, Support Vector Machine, Instance-Based Learning, Rule-Based Learning, Ensembles, and Fuzzy Rule Learning.
- RQ2: Data sets used for SMP include public data sets, proprietary software, open-source projects, and student projects used in only one study. Public data sets UIMS and QUES were found to be the most widely used data sets in the literature.
- RQ3: The majority of the studies (83%) have used OO metrics as the predictor variables; 6% of studies used procedural metrics, and the remaining 11% of the studies have used composite metrics. Among 83% of the studies using OO metrics, 86% have used CK and Li & Henry

Table 15 Research questions addressed in selected primary studies

Study No.	RQ1	RQ2	RQ3	RQ4.1	RQ4.2	RQ4.3	RQ5.1	RQ5.2	RQ5.3	RQ6	RQ7	RQ8
SI1	Yes	Yes	Yes	Yes				Yes		Yes		
SI2	Yes	Yes	Yes	Yes			Yes			Yes		
SI3	Yes	Yes	Yes	Yes			Yes	Yes		Yes		
SI4	Yes	Yes	Yes	Yes			Yes	Yes		Yes		
SI5	Yes	Yes	Yes	Yes	Yes		Yes			Yes		Yes
SI6	Yes	Yes	Yes	Yes			Yes			Yes		
SI7	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes		
SI8	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes		
SI9	Yes	Yes	Yes	Yes	Yes		Yes			Yes		
SI10	Yes	Yes	Yes	Yes	Yes			Yes		Yes		
SI11	Yes	Yes	Yes	Yes			Yes			Yes		
SI12	Yes	Yes	Yes	Yes			Yes			Yes		
SI13	Yes	Yes	Yes	Yes			Yes			Yes	Yes	Yes
SI14	Yes	Yes	Yes	Yes			Yes			Yes		
SI15	Yes	Yes	Yes	Yes			Yes			Yes	Yes	Yes
SI16	Yes	Yes	Yes	Yes			Yes			Yes		
SI17	Yes	Yes	Yes	Yes				Yes		Yes	Yes	Yes
SI18	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
SI19	Yes	Yes	Yes	Yes	Yes		Yes	Yes		Yes		
SI20	Yes	Yes	Yes	Yes			Yes			Yes	Yes	Yes
SI21	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes		
SI22	Yes	Yes	Yes	Yes	Yes		Yes	Yes		Yes		
SI23	Yes	Yes	Yes	Yes	Yes		Yes			Yes		
SI24	Yes	Yes	Yes	Yes	Yes		Yes			Yes		
SI25	Yes	Yes	Yes	Yes			Yes			Yes		
SI26	Yes	Yes	Yes	Yes	Yes				Yes	Yes		
SI27	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	
SI28	Yes	Yes	Yes	Yes	Yes				Yes	Yes		
SI29	Yes	Yes	Yes	Yes	Yes				Yes	Yes	Yes	
SI30	Yes	Yes	Yes	Yes	Yes		Yes		Yes	Yes	Yes	
SI31	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes		Yes
SI32	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes
SI33	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
SI34	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes		Yes	
SI35	Yes	Yes	Yes	Yes	Yes	Yes			Yes	Yes	Yes	
SI36	Yes	Yes	Yes	Yes	Yes	Yes			Yes	Yes	Yes	

metric suite combined. Thus, CK and Li & Henry metrics emerged as prominent metric suite used for SMP.

– RQ4: The majority of studies have used MMRE and Pred(25) performance measures. Performance measures such as mean absolute error (MAE), root-mean-square error (RMSE), and normalized root-mean-square error (NRMSE) have also been used in selected primary studies to evaluate the performance of developed models. K-fold cross-validation with K=10 is employed in most of the studies as a validation method, and leave-one-out validation and hold-out validations are used in a few studies. The use of statistical tests was observed in only 27% of the studies. Nonparametric Wilcoxon test and Friedman rank test were commonly used statistical tests.
– RQ5: Mean MMRE of ML techniques GMDH, KN, GRNN, PNN, SVM, KSTAR, and FFBN ranged from 0.33 to 0.52. Mean Pred(25) values of techniques GMDH, KN, GRNN, PNN, SVM, KSTAR, and FFBN ranged

from 0.46 to 0.69. Thus, performance measures MMRE and Pred(25) of these techniques indicate quite an acceptable level of performance of these ML techniques to develop models to predict software maintainability. The LR and MLR techniques are found to be commonly used ST techniques for developing maintainability prediction models. Overall mean MMRE values of these techniques ranged from 0.53 to 1.50, and mean Pred(25) values ranged from 0.28 to 0.59. These performance statistics are indicative of the poor performance of ST techniques. As the use of HB techniques for developing prediction models for maintainability is concerned, different HB techniques such as FPSO, MFPSO, NNEP, and EFRL are applied in different primary studies. Thus, the cumulative values of MMRE and Pred(25) for HB techniques could not be reported. However, as per values of extracted from different studies, MMRE ranged from 0.22 to 2.1, whereas Pred(25) ranged from 0.19 to 0.64. According to these statistics, no meaningful conclusions about the overall performance of HB techniques for SMP could be drawn.

- GMDH technique outperformed all techniques applied for SMP on different data sets with varying properties. On different data sets, MMRE of this technique ranged from 0.21 to 0.37.
- Only 39% studies have reported threats to the validity of their empirical results, which we classified into threats to internal validity and threats to external validity. It has been observed that the inadequate size of data sets used for experimentation and result bias may be because of the technique used; data sets extracted from software systems having specific characteristics are of major concern.
- Only a few studies discussed the advantages and disadvantages of the techniques applied for developing prediction models. In general, a specific technique may have specific characteristics that may affect the prediction performance of the models. Thus, it is imperative to analyze the technique before applying it for a specific predictive modelling task.

We propose the following future directions for researchers who want to pursue research in this field.

- It has been observed that in the majority of the selected primary studies for this review, public data sets UIMS and QUES have been used. The size of this data set is very small. Future work should focus on carrying out more empirical studies on SMP on industrial data sets, and for this, it is a critical need for the availability of these data sets to the research community. Also, researchers can use varying-sized data sets extracted from open-source software for SMP.
- It has been reflected through this review that few HB techniques were used for SMP. Hence, the generalized results about the applicability of these techniques could not be reported as these techniques were used in very few studies. The future work should focus on applying these techniques on diverse data sets to establish their applicability in this area.
- Few of the selected primary studies reported the threat of result bias as the software system from which data sets have been extracted had got specific properties. Thus, to reduce this threat, future studies should focus on cross-project validation or inter-project validation.
- This review has revealed that the overall performance of models developed using the ML technique is more as compared to models developed using ST techniques. Few ST techniques like PPR gave a good performance but were excluded from analysis due to lack of a uniform framework for experimentation. Thus, apart from exploring more techniques, future work should focus on the development of a uniform experimental scheme for developing prediction models for software maintainability.
- The review has revealed that very few studies have used dynamic metrics as predictors. More studies should be carried out to examine the effect of dynamic metrics on software maintainability.

Compliance with ethical standards

Conflict of interest Authors certify that this manuscript has not been submitted to more than one journal for simultaneous consideration and it has not been published previously (partly or in full). Authors certify that no funding has been received for the conduct of this study and preparation of this manuscript. This article does not contain any studies with human participants or animals performed by any of the authors.

Appendix

6.1 Quality Assessment Questionnaire

All candidate studies selected for this systematic review were evaluated on the basis of quality questions stated in Table 14. This table describes the percentage of studies in which the quality questions were addressed completely, partially or not addressed.

6.2 Research questions addressed in selected primary studies

Table 15 describes the research questions that are addressed in all selected primary studies individually.

References

Aggarwal KK, Singh Y, Chhabra JK (2002) An integrated measure of software maintainability. In: Proceedings of the annual reliability and maintainability symposium, 2002, pp 235–241

Aggarwal KK, Singh Y, Chandra P, Puri M, Base R (2005) Measurement of software maintainability using a fuzzy model. J Comput Sci 1(4):538–542

Aggarwal KK, Singh Y, Kaur A, Malhotra R (2006) Application of artificial neural network for predicting maintainability using object-oriented metrics. Trans Eng Comput Technol 15:285–289

Al Dallal J (2013) Object-oriented class maintainability prediction using internal quality attributes. Inf Softw Technol 55(11):2028–2048

Al-Jamimi HA, Ahmed MA (2013) Machine learning approaches for predicting software maintainability: a fuzzy-based transparent model. IET Softw 7(6):317–326

Almugrin S, Albattah W, Melton A (2016) The journal of systems and software using indirect coupling metrics to predict package maintainability and testability. J Syst Softw 121:298–310

Anwar S, Ramzan M, Rauf A, Shahid AA (2010) Software maintenance prediction using weighted scenarios: an architecture perspective. International Conf Inf Sci Appl, 1–9

Arisholm E (2002) Dynamic coupling measures for object-oriented software. Proc Int Softw Metrics Symp 20:33–42

Bakar AD, Sultan AB, Zulzalil H, Din J (2012) Applying evolution programming search based software engineering (SBSE) in selecting the best open source software maintainability metrics. In: IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE), pp 70–73

Bakota T, Hegedus P, Ladanyi G, Kortvelyesi P, Ferenc R, Gyimothy T (2012) A cost model based on software maintainability. In: 28th IEEE international conference on software maintenance, pp 316–325

Balogh G, G, Végh ÁZ, Beszédes A (2012) Prediction of software development modification effort enhanced by a genetic algorithm, SSBSE Fast Abstract track, pp 1–6

Bandi RK, Vaishnavi VK, Turk DE (2003) Predicting maintenance performance using object-oriented design complexity metrics. IEEE Trans Softw Eng 29(1):77–87

Baqais AAB, Alshayeb M, Baig ZA (2013) Hybrid intelligent model for software maintenance prediction. Lecture Notes Eng Comput Sci 1:358–362

Boehm BW (1976) Software engineering. IEEE Trans Comput 25(12):1226–1241

Basgalupp MP, Barros RC, Ruiz DD, Carlos S, Brazil SP (2012) Predicting software maintenance effort through evolutionary-based decision trees. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing, pp 1209–1214

Basili V, Briand L, Condon S (1996) Understanding and predicting the process of software maintenance releases. In: Proceedings—international conference on software engineering, pp 464–474

Bengtsson PO, Bosch J (1999) Architecture level prediction of software maintenance. In: Proceedings of the 3rd European conference on software maintenance and reengineering, pp 139–147

Bennett KH, Rajlich VT (2000) Software maintenance and evolution: a roadmap. Proc Conf Future Softw Eng 225:73–87

Bhutani S, Chug A (2015) Prediction of software maintainability using neural networks. Int J Comput Sci Commun Netw 5(2):92–95

Briand L, Daly W, Wust J (1999) A unified framework for coupling measurement in object-oriented systems. IEEE Trans Softw Eng 25(1):91–121

Broy M, Deissenboeck F, Pizka M (2006) Demystifying maintainability. In: Proceedings of the 2006 international workshop on Software quality—WoSQ '06, pp 21–26

Campbell DT, Stanley JC, Gage NL (1963) Experimental and quasi-experimental designs for research. Houghton Mifflin, Boston

Chandra D (2012) Support vector approach by using radial kernel function for prediction of software maintenance effort on the basis of multivariate approach. Int J Comput Appl 51(4):21–25

Chidamber SR, Kemerer CF (1994) A metrics suite for object oriented design. IEEE Trans Softw Eng 20:476–493

Chug A, Malhotra R (2016) Benchmarking framework for maintainability. Int J Innovat Comput Inf Control 12(2):615–634

Coleman D, Ash D, Lowther B, Oman P (1994) Using metrics to evaluate software system maintainability. Computer 27(8):44–49

Conte SD, Dunsmore HE, Shen VY (1986) Software engineering metrics and models. Benjamin-Cummings Publishing Co., Redwood City

Dagpinar M, Jahnke JH, Canada BC (2003) Predicting Maintainability with object-oriented metrics—an empirical comparison. In: Proceedings of the 10th working conference on reverse engineering, WCRE 2003, pp 155–164

Daly J, Brooks A, Miller J, Roper M, Wood M (1995) The effect of inheritance on the maintainability of object-oriented software: an empirical study. Proc Int Conf Softw Maint 44:20–29

De Lucia A, Pompella E, Stefanucci S (2005) Assessing effort estimation models for corrective maintenance through empirical studies. Inf Softw Technol 47(1):3–15

De Lucia A, Pompella E, Stefanucci S (2002) Effort estimation for corrective software maintenance. In: Proceedings of the 14th international conference on Software engineering and Knowledge Engineering, pp 409–416

Deligiannis I, Shepperd M (2004) A controlled experiment investigation of an object-oriented design heuristic for maintainability. J Syst Softw 72(2):129–143

Deligiannis I, Shepperd M, Roumeliotis M, Stamelos I (2003) An empirical investigation of an object-oriented design heuristic for maintainability. J Syst Softw 65(2):127–139

Elish MO, Elish KO (2009) Application of TreeNet in predicting object-oriented software maintainability: a comparative study. In: Proceedings of the european conference on software maintenance and reengineering, CSMR 2009, pp 69–78

Elish MO, Aljamaan H (2015) Three empirical studies on predicting software maintainability using ensemble methods. Soft Comput 19(19):2511–2524

Ferneley EH (1999) Design metrics as an aid to software maintenance: an empirical study. J Softw Evolut Process 11(1):55–72

Fioravanti F, Nesi P (2001) Estimation and prediction metrics for adaptive maintenance effort of object-oriented systems. IEEE Trans Softw Eng 27(12):1062–1083

Genero M, Genero M, Cantone G, Informatica D (2003) Building UML class diagram maintainability prediction models based on early metrics. In: Proceedings of 9th international software metrics symposium, pp 263–275

Haiquan Y, Gaoliang P, Wenjian L (2009) An application of case based reasoning to predict structure maintainability. Int Conf Comput Intell Softw Eng, pp 1–5

Halstead M (1977) Elements of software science. Elsevier, New York

Harrison R, Counsell S, Nithi R (2000) Experimental assessment of the effect of inheritance on the maintainability of object-oriented systems. J Syst Softw 52(2):173–179

Harrison R, Counsell S (1998) The role of inheritance in the maintainability of object-oriented systems. In: Proceedings of ESCOM, pp 449–458

Hatton L (2007) How accurately do engineers predict software maintenance tasks? Computer 40(2):64–69

Hayes JH (2004) A metrics-based software maintenance effort model. In: Proceedings of 8th Europian conference on software maintenance and re-engineering, pp 254–258

Hayes JH, Zhao L (2005) Maintainability prediction: a regression analysis of measures of evolving systems. In: 21st IEEE international conference on software maintenance (ICSM'05), pp 601–604
Hegedus P (2013) Revealing the effect of coding practices on software maintainability. In: IEEE International Conference on Software Maintenance, pp 578–581
Heitlager I, Kuipers T, Visser J (2007) A practical model for measuring maintainability. In: 6th International conference on the on quality of information and communications technology, pp 30–39
Henderson-Sellors B (1996) Software metrics. Prentice Hall, Upper Saddle River
Hirota T, Tohki M, Overstreet CM, Hashimoto M, Cherinka R (1994) An approach to predict software maintenance cost based on ripple complexity. In: Proceedings of 1st asia-pacific software engineering conference, pp 439–444
Hops JM, Sherif JS (1995) Development and application of composite complexity models and a relative complexity metric in a software maintenance environment. J Syst Softw 31(2):157–169
IEEE: IEEE Standard Glossary of Software Engineering Terminology. IEEE Press, Piscataway (1990)
Jain R, Sharma A (2016) Critical insight into maintainability and reliability assessment aspects of software systems using soft computing approaches. Int J Control Theory Appl 9(20):9397–9404
Jain A, Tarwani S, Chug A (2016) An empirical investigation of evolutionary algorithm for software maintainability prediction. In: 2016 IEEE Students' conference on electrical, electronics and computer science, SCEECS-2016, pp 1–6
Jeet K, Dhir R, Verma H (2012) A comparative study of bayesian and fuzzy approach to assess and predict maintainability of the software using activity-based quality model. SIGSOFT Softw Eng Notes 37(3):1–9
Jeet K, Dhir R (2012) Bayesian and fuzzy approach to assess and predict the maintainability of software: a comparative study. ISRN Software Engineering, pp 1–10
Jin C, Jin SW (2015) Prediction approach of software fault-proneness based on hybrid artificial neural network on quantum particle swarm optimization. Appl Soft Comput 35:717–725
Jindal R, Malhotra R, Jain A (2015) Predicting software maintenance effort using neural networks. In: 2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, pp 1–6
Jin C, Liu JA (2010) Applications of support vector mathine and unsupervised learning for predicting maintainability using object-oriented metrics. In: International conference on multimedia and information technology, MMIT 2010 vol 1, pp 24–27
Kaur A, Kaur K (2013) Statistical comparison of modelling methods for software maintainability prediction. Int J Softw Eng Knowl Eng 23(6):743–747
Kaur K, Singh H (2011) Determination of maintainability index for object oriented systems. ACM SIGSOFT Softw Eng Notes 36(2):1–6
Kaur A, Kaur K, Malhotra R (2010) Soft computing approaches for prediction of software maintenance effort. Int J Comput Appl 1(16):69–75
Kaur A, Kaur K, Pathak K (2014) Software maintainability prediction by data mining of software code metrics. In: 2014 International conference on data mining and intelligent computing, ICDMIC 2014, pp 1–6
Kewen L, Chenxi C, Wenying L, Xianghua F, Qinghua L (2014) Software defect prediction using fuzzy integral fusion based on GA-FM. Wuhan Univ J Natural Sci 19(5):405–408
Khoshgoftaar TM, Szabo RM (1994a) Improving code churn predictions during the system test and maintenance phases. In: Proceedings international conference on software maintenance, ICSM 1994 vol 7, pp 58–67
Khoshgoftaar TM, Szabo RM (1994b) Improving neural network predictions of software quality using principal components analysis. In: Proceedings of IEEE international conference on neural networks, ICNN 1994 vol 5, pp 3295–3300
Kitechenham BA (2007) Guidelines for performing systematic literature review in software engineering, Technical report EBSE-2007-001, UK
Ku Y, Du J, Yang Y, Wang Q (2011) Estimating software maintenance effort from use cases: an industrial case study. IEEE international conference on software maintenance, ICSM, pp 482–491
Kumar L (2016) Maintainability prediction of web service using support vector machine with various kernel methods. Int J Syst Assur Eng Manag 8(2):205–222
Kumar L, Rath SK (2016) Hybrid functional link artificial neural network approach for predicting maintainability of object-oriented software R. J Syst Softw 121:170–190
Kumar L, Rath SK (2017) Software maintainability prediction using hybrid neural network and fuzzy logic approach with parallel computing concept. Int J Syst Assur Eng Manag 8(2):1487–1502
Kumar V, Kumar R, Sharma A (2014) Maintainability prediction from project metrics data analysis using artificial neural network. Middle-East J Sci Res 19(10):1412–1420
Kumar L, Kumar D, Kumar S (2015) Validating the effectiveness of object-oriented metrics for predicting maintainability. Procedia Comput Sci 57:798–806
Kumar L, Lal S, Murthy LB (2019) Estimation of maintainability parameters for object-oriented software using hybrid neural network and class level metrics. Int J Syst Assur Eng Manag, 1–31
Kumar L, Rath S (2015) Neuro-Genetic Approach for predicting maintainability using Chidamber and Kemerer software metrics suite. Recent Advances in Information and Communication Technology, Springer International Publishing, pp 31–40
Kumar L, Rath SK (2015) Predicting object-oriented software maintainability using hybrid neural network with parallel computing concept. In: Proceedings of the 8th India software engineering conference, ISEC-2015, pp 100–109
Lagerström R, Johnson P (2008) Using architectural models to predict the maintainability of enterprise systems. In: 12th european conference on software maintenance and reengineering, IEEE, pp 248–252
Land R (2002) Measurements of Software Maintainability. In: proceedings of ARTES Graduate Student Conference
Li YF, Xie M, Goh TN (2008) A bayesian inference approach for probabilistic analogy based software maintenance effort estimation. In: Proceedings of the 14th IEEE pacific rim international symposium on dependable computing, PRDC 2008, pp 176–183
Li W, Henry S (1993) Object-oriented metrics that predict maintainability. J Syst Softw 23(2):111–122
Lorenz M, Kidd J (1994) Object-oriented software metrics. Prentice-Hall, Upper Saddle River
Lowther B, Oman P (1995) The application of software maintainability in industrial software systems. J Syst Softw 29(1):3–16
Malhotra R (2015) A systematic review of machine learning techniques for software fault prediction. Appl Soft Comput 27:504–518
Malhotra R, Chug A (2012) Software maintainability prediction using machine learning algorithms. Softw Eng nt J (SEIJ) 2(2):19–36
Malhotra R, Chug A (2014a) Application of Group Method of Data Handling model for software maintainability prediction using object oriented systems. IntJ Syst Assur Eng Manag 5(2):165–173
Malhotra R, Chug A (2014b) Application of evolutionary algorithms for software maintainability prediction using object-oriented metrics. In: Proceedings of the 8th international conference on bioinspired information and communications technologies, pp 348–351
Malhotra R, Chug A (2016) Software maintainability: systematic literature review and current trends. Int J Softw Eng Knowl Eng 26(8):1221–1253

Malhotra R, Lata K (2017) An exploratory study for predicting maintenance effort using hybridized techniques. In: Proceedings of the 10th innovations in software engineering conference. ACM, pp 26–33

Malhotra R, Lata K (2018) On the application of cross-project validation for predicting maintainability of open source software using machine learning techniques. In: 7th IEEE International conference on reliability, infocom technologies and optimization (ICRITO), pp 175–181

Manchanda S, Chug A (2015) CFS based feature subset selection for software maintainance prediction. Int J Adv Found Res Comput 2(5):13–23

Mie M, Thwin T, Quah T (2003) Application of neural networks for software quality prediction using object-oriented metrics. In: Proceedings of the international conference on software maintenance, ICSM 2003, vol 76, pp 116–125

Mishra S, Sharma A (2015) Maintainability prediction of object oriented software by using adaptive network based fuzzy system technique. Int J Comput Appl 119(9):24–27

Misra SC (2005) Modelling design/coding factors that drive maintainability of software systems. Softw Q J 13(3):297–320

Muthanna S, Ponnambalaml KKK, Stacey B (2000) A maintainability model for industrial software systems using design level metrics. In: Proceeding 7th working conference on reverse engineering, pp 248–256

Niessink F, Van Vliet H (1997)Predicting maintenance effort with function points. Int Conf Softw Maint 32–39

Olatunji SO (2013) Sensitivity based linear learning method and extreme learning machines compared for software maintainability prediction of object orinented software system. ICTACT J Soft Comput 3:514–523

Oman P, Hagemeister J (1992) Metrics for assessing a software system's maintainability. In: Software Maintenance procedings, pp 337–334

Perepletchikov M (2007) Coupling metrics for predicting maintainability in service-oriented designs. In: 18th Australian conference on software engineering, pp 329–340

Ping L (2010) A quantitative approach to software maintainability prediction. In: Proceedings of 2010 international forum on information technology and applications, IFITA 2010, vol 1, pp 105–108

Prasanth NN, Ganesh S, Dalton GA (2008) "Prediction of maintainability using software complexity analysis: an extended FRT. In: Proceedings of the 2008 international conference on computing, communication and networking, ICCCN 2008, pp 1–9

Prechelt L, Unger B, Philippsen M, Tichy W (1998) Re-Evaluating inheritance depth on the maintainability of object-oriented software. Int J Empir Softw Eng pp 1–16

Rai V, Srivastava AM, Pandey H, Singh VK (2015) Estimation of maintainability in object oriented design phase: state of the art. Int J Sci Eng Res 6(9):25–35

Rana SKA, Dash Y (2012) Maintainability prediction of object-oriented software system by multilayer perceptron model. ACM SIGSOFT Softw Eng Notes 37(5):1–4

Riaz M, Mendes E, Tempero E (2009) A systematic review of software maintainability prediction and metrics. In: Proceedings of the 3rd international symposium on empirical software engineering and measurement. IEEE Computer Society, pp 367–377

Riaz M, Mendes E, Tempero E, Sulayman M (2013) Using CBR and CART to predict maintainability of relational database-driven software applications. In: Proceedings of the 17th international conference on evaluation and assessment in software engineering—EASE '13, p 132

Rizvi SWA, Khan RA (2010) Maintainability estimation model for object-oriented software in design phase. J Comput 2(4):26–32

Schneberger SL (1997) Distributed computing environments: effects on software maintenance difficulty. J Syst Softw 37(2):101–116

Seref B, Tanriover O (2016) Softwarecode maintainability: a litreature review. Int J Softw Eng Appl (IJSEA) 7(3):69–87

Sharawat S (2012) Software maintainability prediction using neural networks. Int J Eng Res Appl (IJERA) 2(2):750–755

Sharma H, Chug A (2015) Dynamic metrics are superior than static metrics in maintainability prediction: an empirical case study. In: 4th International conference on reliability, infocom technologies and optimization, ICRITO-2015, pp 2–7

Sheshasaayee A, Jose MR (2016) Deployment of a fuzzy XNOR based evaluative system for the maintainability features of aspect oriented systems. Int J Appl Eng Res 11(9):6637–6642

Singh RK, Balyan A (2012) Approach to software maintainability prediction versus performance. Int J Soft Comput Eng 2(2):51–54

Singh Y, Bhatia PK, Sangwan O (2009) Predicting software maintenance using fuzzy model. ACM SIGSOFT Softw Eng Notes 34(4):1–6

Somerville I (2012) Software engineering. Addison-Wesley, Boston

Tarwani S, Chug A (2016) Predicting maintainability of open source software using gene expression programming and bad smells. In: 5th International conference on reliability, in infocom technologies and optimization (Trends and Future Directions), ICRITO-2016, pp 452–459

Van Koten C, Gray AR (2006) An application of Bayesian network for predicting object-oriented software maintainability. Inf Softw Technol 48(1):59–67

Wang LJ, Hu XX, Ning ZY, Ke WH (2009) Predicting object-oriented software maintainability using projection pursuit regression. In: 1st International conference on information science and engineering, ICISE 2009, pp 3827–3830

Wen J, Li S, Lin Z, Hu Y, Huang C (2012) Systematic literature review of machine learning based software development effort estimation models. Inf Softw Technol 54(1):41–59

Ye F, Zhu X, Wang Y (2013) A new software maintainability evaluation model based on multiple classifiers combination. IEEE International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, pp 1588–1591

Zavvar M, Ramezani F (2015) Measuring of software maintainability using adaptive fuzzy neural network. Int J Modern Edu Comput Sci 7(10):27–32

Zhang W, Huang LG, Ng V, Ge J (2015) SMPLearner: learning to predict software maintainability. Autom Soft Eng 22(1):111–141

Zhou Y, Leung H (2007) Predicting object-oriented software maintainability using multivariate adaptive regression splines. J Syst Softw 80(8):1349–1361

Zhou Y, Xu B, Yuming Z, Baowen XU (2008) Predicting the maintainability of open source software using design metrics. Wuhan Univ J Natural Sci 13(1):14–20

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

Soft Computing
https://doi.org/10.1007/s00500-020-05031-2

FOCUS

A systematic review of developments in the 2-tuple linguistic model and its applications in decision analysis

Tanya Malhotra[1] · **Anjana Gupta**[1]

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
Several practical problems incorporate uncertainty and erroneous information within its framework of definition. Many distinguished models have already been developed to manage such unreliability, like probabilistic models, and so forth. However, to revitalize the problems where uncertainty is not probabilistic, or the information is vaguely accessible, other tools like fuzzy logic and fuzzy linguistic-based approach have emerged. The efficacy of a linguistic-based approach pertains to a problem of decision making involving qualitative information where precise numerical modeling appears to be inappropriate. The 2-tuple linguistic model formalized by Herrera and Martínez in year 2000, has upgraded several linguistic processes for solving complex decision-making issues. Its successful application in different fields has impelled several researchers to work in its extension. This study aims to offer a systematic literature review on a 2-tuple linguistic model as well as discuss its extensions. The manuscript lays out a brief review of 188 articles published between 2000 and 2019. Based on the 188 selected articles, the following questions can be justified: (1) What is the fuzzy linguistic approach? (2) What is the linguistic model based on the 2-tuples and its foundation? (3) Why the 2-tuple model preferred over other existing classical linguistic models? (4) What are the extensions of the 2-tuple model? Moreover, (5) What are the major applications of a 2-tuple model witnessed in varied disciplines, specifically decision analysis? Therefore, seeking answers to these questions indicate current trends in research. Further, it provides the direction for scholars to work in this field and overcome the challenges associated with the prevailing 2-tuple model.

Keywords Linguistic variables · Computing with words (CWW) · 2-Tuple linguistic model · Multi-attribute decision making (MADM) · Multiple criteria group decision making (MCGDM) · Unbalanced linguistic term sets (ULTSs) · Multi-granularity linguistic term set (MGLTS) · Linguistic distribution assessment · Heterogeneous information

1 Introduction

Several real-world issues deal with information that is uncertain and vague. The more possibility in a question, the less accurate we can be in our understanding of that problem, and as a result, it should be obstructed apparently. Problems at the molecular level are not addressed in the late nineteenth century. As a result, to achieve a more credible solution and simultaneously quantifying the amount of risk leads to the development of probability theory and more advanced statistical methods in the late twentieth century. The question of possibility was challenged by the studies of Max-Black in vagueness persisting in the problem, first in 1937. Later on, it was challenged by Zadeh (1965), who introduces the concept of fuzzy sets. The introduction of the fuzzy set theory has shown great alacrity and competence in modeling the uncertainty abiding within the problems in an attempt to challenge the existing probability theory along with the very basis on which it depends, i.e., classical binary logic.

Traditional set theory is based primarily on the idea that either an element belongs to a collection or does not belong to it. There is no significant presence of partial membership of elements of a set. However, fuzzy set theory permits us to extend the fundamental concepts of classical set theory and logic in an infinite way that seems hugely sensible. The conviction of binary membership was extended by Zadeh to accommodate varied " degrees of membership" on the unit interval [0, 1] where 0 and 1 represent "non-membership"

Communicated by Hemen Dutta.

✉ Anjana Gupta
anjanagupta@dce.ac.in

[1] Department of Applied Mathematics, Delhi Technological University, Bawana Road, Rohini, Delhi, India

Published online: 08 June 2020

and "full membership," respectively, and the infinite number of values lying in between 0 and 1 represent partial membership. For more details about the contributions made to the theory of fuzzy sets and applications, one can see Ref. Zadeh (1965, 1968, 1971, 1972a, b, 1973, 1975a), Zadeh et al. (1975).

Decision-making and decision analysis is considered one of the most prominent application areas and has been studied extensively in the literature. It is observed that most of the decision-making problems are ill-structured in dealing with imprecise data. Probabilistic models able to handle decision problems under uncertainty, but in many situations, non-probabilistic models are required to offer better results. A plethora of fuzzy MCDM approaches has widely studied in the literature. The motivation towards the development of diverse fuzzy MCDM approaches is because of the desire of researchers and practitioners for reinforcing decision-making methods via a recent enhancement in computer technology, scientific computing, and mathematical optimization. In paper Mardani et al. (2015), the systematic review of fuzzy MCDM methods and their application in different fields from 1994 to 2014 has been given.

Several aspects of practical, real-world problems require human intervention, where they need to select among different alternatives through rationale and cognitive process. Such decision problems are ill-structured and are not submissive to quantitative characterization. Meanwhile, there may exist some situations where the present information is not quantifiable because of its nature (e.g., when categorizing the quality of research paper, the terms like "Excellent," "Good," "Average" can be used), or the cost of computation for quantitative information present in the problems is very high that an "approximate value" is bearable. Therefore, experts often prefer to express the results in a natural language rather than precise numerical value. This leads to the notion of a linguistic variable. Linguistic variables are defined as variables whose values are accepted as either words or an ordered blend of words used in a natural or artificial language, whose meaning is characterized by a semantic principle (Zadeh 1975b). For instance, age can be a linguistic variable whose values are considered as linguistics, such as very young, young, not young, not and very young instead of numbers like 1, 2, 3,…, 100.

The linguistic approximation was initiated ideally by Zadeh (1975b) and was considered as the most prominent application area of the linguistic variable concept. He presented in the paper that the linguistic-based approach was not associated with the conventional non-mathematical ways of dealing with humanistic systems. Instead, it was represented as an amalgam between the quantitative and qualitative aspects, based on the use of words when numerical characterizations were inappropriate. The progress in the field of fuzzy linguistic-based approach is remarkable with the growing number of years and considered as a core area in different disciplines like decision making, supply chain, pattern recognition, energy optimization, approximate reasoning, and so forth.

The usefulness of linguistic-based information implies a broad perspective in the current specialized literature on linguistic decision analysis and CWWs-related methodologies. Furthermore, distinct linguistic computational models (LCMs), namely computational models based on membership function (Bonissone and Decker 1986; Degani and Bortolan 1988; Dubois and Prade 1980; Klir and Yuan 1995) and based on ordinal scales (Delgado et al. 1993; Xu 2004a; Yager 1981), were also developed. However, both models simultaneously encounter a lack of accuracy due to the information loss that occurs primarily due to the prerequisite for enumerating the ultimate results in the primary expression domain. As a result, to overcome the shortcomings of the existing models and to facilitate the process of CWW, Herrera and Martínez (2000a) introduced a linguistic model that presents the linguistic information in terms of a couple of values, known as a 2-tuple. The 2-tuple linguistic model is created with the help of a linguistic term belonging to the set of linguistic terms defined previously and a symbolic translation expressed by a numerical value evaluated in $[-0.5, 0.5)$. The ascendancy of this model lies within the fact that the representation of this model remain continuous in its domain, while it is considered as discrete in the other existing conventional models. Therefore, the 2-tuple model specifies any number of information in the universe of discourse. The utilization of the model, as mentioned above, is able to manage the problem of multi-granularity, which is presented in Herrera et al. (2000), Herrera and Martínez (2001b). Further, the combination of linguistic-numerical information is introduced in Delgado et al. (1998), Herrera and Martínez (2000a, b). Hence, in problems, especially where the information available is uncertain and non-probabilistic, linguistic modeling seems considerably logical and, in this manner, has produced significant results in diverse fields, as mentioned in Fig. 1.

A profound review of CWW's distinguished literature shows the evolution and usefulness of the linguistic model based on a 2-tuple applicable to many real-world problems. Therefore, it is evident that this model is more prominent than other existing computational models for CWW. It would be more interesting to make a brief review of the model concentrating precisely on the prerequisite and foundations and advantages of the model (Herrera and Martínez 2000a, 2001a), methodologies for CWW in distinct problems, and complicated frameworks (Espinilla et al. 2013; Herrera et al. 2000; Herrera and Martínez 2001b; Herrera et al. 2005, 2008). Furthermore, we tried to summarize various extensions and the newer version of linguistic compu-

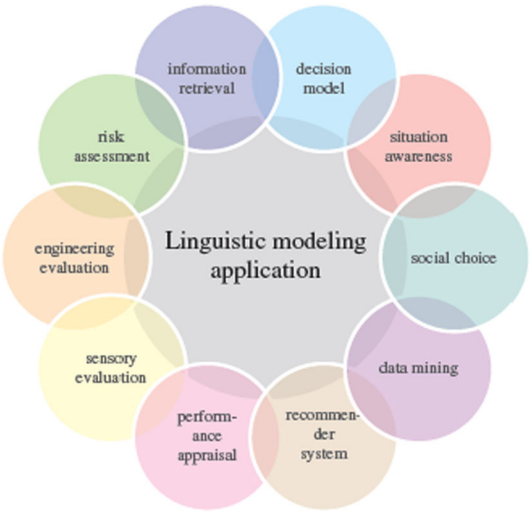


Fig. 1 Linguistic modeling application

tational models (LCMs) based on 2-tuple (Dong et al. 2009a; Martínez and Herrera 2012; Wang and Hao 2006).

The primary objective of this review is to classify and summarize categorically the literature relevant to CWW's 2-tuple-based linguistic model as well as its primary application envisaged in different decision-making models. This review also looks at the weaknesses of traditional linguistic computational models and presents pertinence of the traditional 2-tuple model in various fields. Besides, this contribution is also made up of the literature relevant to numerous aggregation operators centered upon the linguistic model of 2-tuple.

The remaining structure of the paper set out as follows: Sect. 2 summarizes literature relevant to the fuzzy linguistic-based approach as well as its application in various disciplines. Section 3 presents a brief research methodology adopted for reviewing the literature. Section 4 provides a brief introduction about the 2-tuple linguistic model for CWW, along with the extensions and utilization of this model in MCDM and MCGDM problems. Section 5 presents an in-depth review of the application concerning the 2-tuple linguistic model. Finally, the conclusion and summarization of the contribution are presented in Sect. 6.

The work presented in this contribution is very much useful for researchers around the world who are working in this field and is also looking for exploration. This study is related to the usage of the traditional 2-tuple linguistic model in various decision-making problems. Our study has been driven out of curiosity to examine several decision models primarily based upon the 2-tuple linguistic model.

2 Fuzzy linguistic approach

Before summarizing the main works of the 2-tuple model, it is initially essential to provide insights about the evolution of fuzzy linguistic-based approaches. So this section presents necessary information about the fuzzy linguistic approach followed by a systematic review of its application in different research fields.

We generally work in a quantitative environment where the data available are represented numerically. Even the models that employ Boolean variables with either "true" or "false" values are replaced by 0 and 1. However, it is envisioned that the results acquired in binary form have given inadmissible results in many human decision problems and are therefore incompatible in a situation where the accessible information is incomplete or vague. For handling such type of complex decision problems where data are unquantifiable, the methodology based on linguistic approximation and truth qualification offers excellent results in the physical realm.

Zadeh (1976) pioneered the linguistic-based approach, which served as a way to provide a language for inaccurate estimation of the components that are involved in the process of decision making which are either intrinsically fuzzy or are inadequate of precise measurement. Henceforth, it provides many reliable results and, consequently, successfully applied in different research areas. For more details on the fuzzy linguistic approach, one can refer to the papers (Bordogna et al. 1997; Delgado et al. 1998, 1992; Eshragh and Mamdani 1979; Herrera et al. 1995, 1996; Martin and Klir 2006; Tong and Bonissone 1980; Wang 2001; Wenstop 1980; Zadeh 1975b, 1976, 1996; Zadeh and Kacprzyk 1999; Zeeny 1982). Here, Table 1 shows a summary of the existing literature based on the application area of the fuzzy linguistic approach. In this table, the chronological order of the publication and summary of the significant contributions made by the fuzzy linguistic-based approach is given.

After discussing about the literature relevant to the fuzzy linguistic approach, we establish a brief research methodology carried out for reviewing the journal articles published between 2000 and 2019 related to the 2-tuple linguistic model in the subsequent section.

3 Research methodology for literature review

The literature review on the 2-tuple linguistic model and its application in various fields is conducted to present articles that are published in well-established journals providing essential insights to researchers and practitioners working in this area. In our manuscript, we tried to summarize the results published in the journal articles between 2000 and 2019. For the sake of understanding, the flowchart of the research methodology is demonstrated in Fig. 2.

We searched web of science to obtain an extensive bibliography of the vast literature related to the 2-tuple linguistic model as well as its contributions to different fields. The web

Table 1 Summary of applications based on a fuzzy linguistic approach

Title	Year	Paper	Description	Why fuzzy linguistic approach	Strong/weak points
The Concept of a Linguistic Variable and Its Applications to Approximate Reasoning Part I, Part II, Part III	1975	Zadeh (1975b, c, d)	This paper demonstrates an idea of a linguistic variable and its application in approximate reasoning, which is considered to be a mode of reasoning that is not exact nor very inexact	Linguistic variables provides a method to estimate characterization of phenomena which are really complex or obscure to be amenable to description in standard quantitative terms	This approach offers a more realistic framework for human reasoning than classical binary logic. Also, computation with linguistic variables may require the solution of nonlinear programs and thereby leads to results that are imprecise to the same degree as the underlying probabilities
A linguistic approach to decision making with fuzzy sets	1980	Tong and Bonissone (1980)	This paper proposes a method for fuzzy decision making based on linguistic approximation and truth qualification	Fuzzy linguistic approach is considered to be the most expropriate in a situation where information is vague or imprecise. Hence, decision expressed in such cases should be linguistic rather numerical	The proposed approach creates a linguistic evaluation of the choices, and therefore results acquired from this approach can be used in fuzzy decision problems. Decision-makers associated with the choice handle at every level, and as a result, the proposed approach is of great practical importance
Fuzzy set based linguistic approach: theory and application	1980	Bonissone (1980)	Fuzzy sets and logic are considered to be premise for the linguistic approach, and this paper very well explains this approach and further illustrates it with example	Proposed approach helps in developing models that can able to mimic approximate reasoning	The linguistic approach can able to handle two major issues: Firstly, the proposed methodology relates a name to an unbalanced fuzzy set on the premise of semantic similitude and, secondly, performs mathematical operations with fuzzy numbers
The problem of linguistic approximation in clinical decision making	1988	Degani and Bortolan (1988)	This paper explains linguistic approximation method in computerized framework which is the background of medical decision making	The proposed method is useful in expressing linguistic outputs that are simple and reliable. Moreover, the choices can be completely represented by graphical representation of the suitability sets over a decision space [0, 1] to avoid uncertainty	The linguistic approximation approach is useful for clinically recognized linguistic terms whose meaning is well defined in the medical community. However, some problems still persist with the existing approach. The first concern is regarding a choice of the threshold, which is used to differentiate good and bad approximations. The second concern is about a limited number of terms at the user's disposal

Table 1 continued

Title	Year	Paper	Description	Why fuzzy linguistic approach	Strong/weak points
Linguistic decision-making models	1992	Delgado et al. (1992)	This paper proposes the model for decision-making problems in which knowledge about gains and true state of nature is given as linguistic rather than numeric.	Linguistic decision models proved to be an edge over other models based on classical decision theory. Linguistic decision models could very well handle representation of the true state of nature, and gains are represented by means of linguistic terms, and the objective of decision-makers are vaguely established.	This article is restricted to single objective decision-making problems
A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation	1993	Bordogna and Pasi (1993)	In this article, an extension of the weighted Boolean retrieval model is being formalized within the fuzzy linguistic framework. In this model, linguistic descriptors replace numeric query weight. This fuzzy linguistic model is explained, and further evaluation is being carried out of its implementation on a Boolean information retrieval system	Proposed approach has an ability to deal with the imprecision and subjectivity characterizing retrieval activity	The retrieval model defined within the fuzzy framework provides an understandable and convenient means of dealing with qualitative and imprecise criteria. The introduction of linguistic weights in queries has improved the expressive power of the weighted Boolean query language
Non-numeric multi-criteria multi-person decision making	1993	Yager (1993b)	The objective of this paper is to describe a multi-expert MCDM strategy for estimating and selecting the alternatives based on a non-numeric scale	The non-numeric linguistic scale has an ability to deal with the imprecision and thereby provides a more flexible framework for an evaluator	The central idea of the proposed methodology in using linguistic terms linked with the scores is that it makes it simple for an evaluator to manipulate. The process explained in this paper allows experts for the multi-criteria evaluation of every object, followed by an aggregation of these individual experts in order to obtain an overall object evaluation

Table 1 continued

Title	Year	Paper	Description	Why fuzzy linguistic approach	Strong/weak points
A fuzzy multi-criteria decision-making method for technology transfer strategy selection in biotechnology	1994	Chang and Chen (1994)	In this article, the potential application of fuzzy set theory to technology transfer strategy selection in the area of biotechnology management is described. This article also suggests a new look at the concept of the index of optimism as implicated by Kim and Park (1990)	Proposed approach is beneficial in describing subjective assessments of the appropriateness of alternatives verses criteria and the importance weightings of criteria	The decision algorithm devised in this paper is enforced to the varied management decision process and relevant areas. But still, their persist many lookup issues halting in the enrichment of this approach, few among them are stated below: (1) The definition of the suitable fuzzy linguistic variables, their numbers, their values, and their universe of discourse for conventional use in the algorithm. (2) A method for accumulating raw records and estimating the suited membership functions for the base linguistic variable values. (3) The relative importance of every decision-maker is not viewed in this study. Therefore, in many multi-judge decision problems, weights are not only decided by the decision-maker and criterion; however, it can be different with respect to the decision environment and structure
FUZMAR: An approach to aggregating market research data based on fuzzy reasoning	1994	Yager et al. (1994)	This article introduces an approach for validating models that contain variables having linguistic values that are drawn from a linear ordering scale. The methodology proposed in the paper fundamentally depends on operations from fuzzy logic to construct the models and the principal of minimal entropy to choose between competing models.	In the proposed methodology, linguistic values assumed to give accurate and reliable results. Also, computations made on linguistic values are simpler than numerical ones	The key idea of the work is to indicate a methodology to use the information provided by market surveys to predict the values of other related variables of interest to market research analysts. The proposed methodology based on the use of Shannon's entropy. Further, an approach to a simple aggregate idea to form a complicated one is also introduced
A multiple criteria linguistic decision model for human decision making	1994	Liu et al. (1994)	This article proposes a multiple criteria linguistic decision model (MCLDM) for human decision-making problems. MCLDM is beneficial to appraise a set of alternatives over varied criteria by using linguistic variables and then presents linguistic decision output for selecting the best alternatives	The proposed approach is beneficial in describing a systematic way of making linguistic decisions within a fuzzy environment, so as to tackle imprecise information well, that cannot be handled by probability theory.	MCLDM method is one of the most powerful methods for managing decision-makers engaged in making decisions in a situation where the information available is vague or uncertain. Moreover, MCLDM is useful in performing sensitivity analysis

Table 1 continued

Title	Year	Paper	Description	Why fuzzy linguistic approach	Strong/weak points
A fuzzy linguistic approach to a multi-criteria sequencing problem	1996	Adamopoulos and Pappis (1996)	This paper considers a fuzzy approach to a single machine scheduling problem, where the system's variables are defined with the help of linguistic terms. The fundamental point of the study is to figure out the period of the transforming times, on grouping the jobs in the machine. And lastly, to determine the common expected rate in a near-optimal way	Fuzzy set theoretic linguistic framework overcomes the difficulty in approximating the parameters which best describe the system under consideration. The proposed approach proves to be beneficial in a situation where parameters such as processing times of tasks are characterized by vagueness. Such vagueness is well handled qualitatively rather numerically	The advantage of using the proposed approach in many real production situations is to give promising results when the information available is vague. It could able to deal with several scheduling models in which processing times of jobs are not precisely known and henceforth give considerable results
Using fuzzy numbers in educational grading system	1996	Law (1996)	This paper frames out a structural model of a fuzzy academic grading system and further proposes an algorithmic program for it. This study also comprised of a strategy to frame out membership functions of various linguistic values with dissimilar weights.	Observed scores of students, as well as scores of distinct questions in an examination, are vague; thereby, fuzzy set theory is beneficial in aggregating scores. Moreover, the linguistic variable is being used to judge student's performance via grade and hence is beneficial in aggregating scores well	The proposed approach offers a result that is quite reliable and uses the fuzzy set theory as its basis to provide us with the information needed to develop a new grading method. In the proposed methodology, the grades to be assigned correspond to the degree of membership so that aggregation of different test scores into a single score is plausible. Further, it provides us means to strengthen the quality of the grading system, thereby giving a clear indication about the student's academic performance
Direct approach processes in group decision making (GDM) using linguistic ordered weighted average operators	1996	Herrera et al. (1996)	In this paper, the properties and the axiomatic of the linguistic ordered weighted averaging operator (LOWA) by presenting the rationality of its aggregation way is being presented. Further, the use of the LOWA operator is also being introduced in order to solve GDM problems from individual linguistic preference relations.	Proposed approach helps in incorporating high human consistency in decision models	The focal thought created in the paper is to find out evidence of rationality of the LOWA operator. Also, further indicating its requirement in processes of GDM in a linguistic environment
Application of a fuzzy linguistic approach to analyse asian airports competitiveness	1997	Park (1997)	This paper proposes an analysis of the potential competitiveness of 9 major airports, which is based on the fuzzy linguistic approach and also on the airport experts' point of view	The proposed approach provides a practical and applicable assessment of airport competitiveness in the East Asia region. Also, it is even more flexible and adaptable to deal with competitiveness associated with designated influencing factors	The central idea of our proposed approach is that it helps in converting influencing factors to the competitiveness of finite scales of using linguistic variables

Table 1 continued

Title	Year	Paper	Description	Why fuzzy linguistic approach	Strong/weak points
A new method for tool steel materials selection under fuzzy environment	1997	Chen (1997)	This paper provides a new approach in solving the tool steel materials selection problem under the framework of fuzzy, where the significance weights of distinctive criteria and the ranking of different preferences underneath distinctive criteria are assessed in linguistic phrases expressed by way of trapezoidal fuzzy numbers. Further, an example is given to illustrate the newer methodology properly	The proposed technique is much more proficient than rather a traditional method presented by Wang and Chang in 1995. Henceforth, it helps the designer to choose the most relevant tool steel materials systematically. Also, it makes the execution much faster	The proposed technique makes use of easy arithmetic operations instead of complex arithmetic operations for aggregation and ranking of fuzzy numbers. As a result, its execution is much quicker
Linguistic labels for expressing fuzzy preference relations in fuzzy group decision making	1998	Umano et al. (1998)	This paper proposes improvements to pairwise group decision making based on fuzzy preference relations and then proposes the use of linguistic labels for expressing the fuzzy preference relation	The proposed approach helps represent preferences in linguistic labels rather than numerical, thereby producing the results which are more realistic in presenting imprecise preference relations	The linguistic labels representation method proposed has helped in improving simplicity and flexibility in the applications of the group decision-making method
Multi-criteria Multi-stages linguistic evaluation and ranking of machine tools	1999	Devedzić and Pap (1999)	This paper shows a model for the linguistic evaluation of machine tools parameters and a procedure for alternatives ranking	The proposed approach is helpful in modeling and ranking linguistically evaluated characteristics of machine tools. It is based on the application of fuzzy sets theory and represents an up-gradation to the former approaches by overcoming the inconsistency that appears in the final result	Suggested methodology for machine tools evaluation and ranking may be used for both machine tools selection and linguistic quantification of an already selected machine tool. By linguistic quantification of machine tools elements rigidity and by setting a significance of these elements for given machining conditions, it is possible to generate a linguistic value of rigidity of machine tool as a whole
Linguistic decision analysis: steps for solving decision problems under linguistic information	2000	Herrera and Herrera-Viedma (2000)	This paper displays a study on the steps to be embraced in linguistic decision analysis under linguistic data	The linguistic-based approach provides a smoother framework for solving decision-making problems such that it helps in representing information in a more straightforward and adequate manner	The use of linguistic models in decision problems is highly beneficial in the decision-making problems where results cannot be expressed in numerical form

Table 1 continued

Title	Year	Paper	Description	Why fuzzy linguistic approach	Strong/weak points
Extensions of the TOPSIS for group decision making (GDM) under fuzzy environment	2000	Chen (2000)	This paper aims to present the development of the TOPSIS method for solving the GDM problem under the framework of fuzzy. In this paper, linguistic terms are in a way used to represent the ranking of every alternative and the weight of each criterion, which can be described accordingly with the triangular fuzzy number. Vertex method is also designed to estimate the distance between the two triangular fuzzy ratings. Finally, concerned with the notion of TOPSIS, a closeness coefficient is depicted to select the ranking order of all alternatives	Crisp values are insufficient to handle practical, real-life situations. Some amount of fuzziness is linked with decision information and GDM problems. Thereby, the use of a linguistic approach to appraising the weights of all criteria and ranking of every alternative concerning every criterion has provided a more flexible and realistic framework than numerical value	The proposed approach is useful in solving MCDM problems within a fuzzy framework. The need of linguistic variables in various decision problems is highly beneficial in assessing alternatives concerning criteria and significance weights. Further, with the help of a vertex method distance between the two fuzzy triangular numbers can be estimated. Henceforth, it extends the TOPSIS technique to the fuzzy environment
A linguistic decision model for personnel management solved with a linguistic bi-objective genetic algorithm	2001	Herrera et al. (2001)	This paper exhibits a model for staff selection in the state of uncertainty with the goal that it will both limit the risks emerging from the performance of undertakings by unsuitable personnel and boost the limit of the firm via optimal assignment of workers	The linguistic formulation is very much useful in a situation where the information available is vague or imprecise. Thereby working in qualitative areas suchlike a personnel management linguistic approach has provided a more flexible and realistic framework than numerical value	The proposed method has given promising results henceforth; it can be adopted and further implemented to different practical, real-life problems under the same consideration

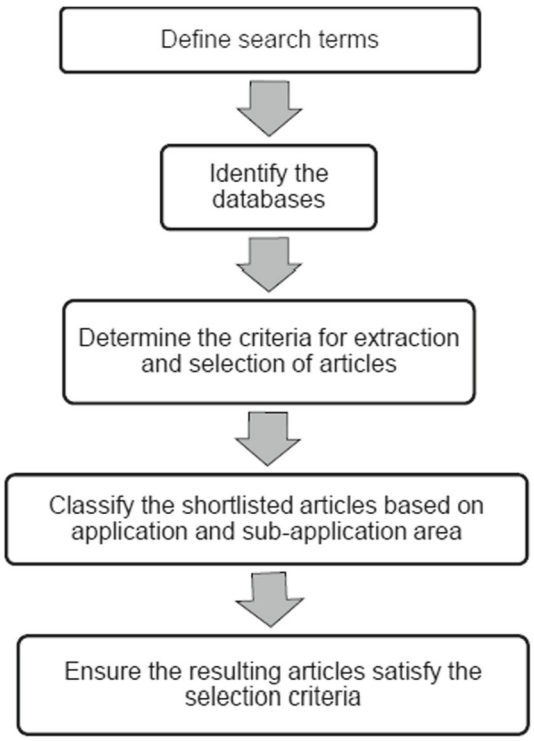


Fig. 2 Flowchart of research methodology

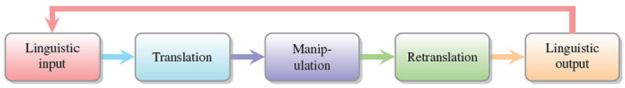


Fig. 3 Computing with words scheme (Martínez and Herrera 2012; Yager 1999)

of science core collection information provides us with fast and impactful access to the world's leading citation database such as ScienceDirect, IEEEXplore, Springer link online libraries, ACM digital library, Willey, Research Gate, and Taylor & Francis and so forth.

The following subsections clearly outline the procedure we followed in extracting articles along with selection criteria, classification method, and further explains the filtering processes.

3.1 Procedure for extraction of articles and selection criteria

We conducted a comprehensive search for articles based upon the 2-tuple linguistic model from the year 2000–2019. For this purpose, we used basic search to look for the relevant articles by using terms and phrases such as "2-tuple linguistic model," "fuzzy linguistic approach," "multi-criteria group decision making," "computing with words," "multi-granularity linguistic term set," and "fuzzy linguistic modeling." The number of potential scholarly articles obtained at the initial step by using keywords described above in the ScienceDirect database was approximately 1192, published in the year between 2000 and 2019.

Three criteria were determined and used to choose and acknowledge the journal articles for additional consideration. Articles were excluded if they did not meet the accompanying selection criteria.

- Decision making is considered to be the most important and extensively carried out activities in the literature. In this context, we tried to choose articles related to the decision models based on 2-tuple and its extension published from 2000 to 2019.
- The study excluded the articles that were based on decision models defined in respect of fuzzy sets, linguistic expressions based on hesitant fuzzy environment, or intuitionistic fuzzy linguistic environment. The papers related to hierarchical algorithms for incomplete linguistic preference relation, aggregation operators centered upon Choquet integral, 2-tuple linguistic hybrid arithmetic aggregation operators, and harmonic operators and their applications in MCGDM problems are also excluded. Finally, the Journal articles concerning the consensus model primarily based upon the 2-tuple linguistic model and its extension, as well as decision models based upon incomplete weight information, are also not considered in this study.
- Only scholarly articles written in English and published in international journals had been chosen to accomplish the top level of aptness. This review excluded master's and doctoral thesis, textbooks, book chapters, and notes.

After examining the title and abstract of the 1192 articles so obtained at the very initial step, we retrieved near about 739 articles, which satisfies the selection criteria. These articles further undergo the next round of scrutiny for further filtration and deep analysis.

3.2 Classification criteria

In the second step, the research articles that we have extracted in the previous step further undergoes into the classification of articles based on application and sub-application area. It is well known that the 2-tuple model has been an improvement over the existing LCMs because of its accuracy and interpretability. As a result, over the subsequent years, the growth of the 2-tuple model is tremendous and hence shows the applicability in distinct problems, specifically decision-making and decision analysis.

The significant contribution of the use of this linguistic model is visualized within decision-making problems, which is broadly categorized into MCDM and MCGDM problems. Further in this study, we have outlined about the key findings of the 2-tuple model, its extension (i.e., numerical scale

model (Dong et al. 2009a), personalized individual semantics model (Li et al. 2018b), linguistic distribution (Zhang et al. 2014), hierarchical linguistic model (Herrera et al. 2008), and so forth), numerous aggregation operators, and their applicability into the field of decision making. The complete information has been classified systematically in the following section arranged in the tabular form, namely Tables 2 and 3.

3.3 Reviewing process

With the growing years, there is a tremendous and extensive surge of papers primarily based upon the development of the 2-tuple model along with its application in the various fields. In this step, each article is manually scrutinized in two rounds; the first round of review involves manual articles scrutiny by considering the title, abstract, introduction, and conclusion so that we can categorize the article according to the desired application area. In the final round, we read the full text and analyze each article, its methodology, its weak and strong points, and then classify it into relevant areas accordingly. This analysis finally resulted in 188 articles.

4 2-Tuple linguistic model for computing with words (CWWs)

The objective of the present section is to initially clarify the motivation and basic definitions concerning the 2-tuple model followed by its extensions. In the specialized literature of linguistic decision making, it could be considered that the contribution of the 2-tuple model is vast and extensive. As a consequence, in our study, we shall focus on the use of this model in the area of decision making and other distinguished research branches.

4.1 Motivation

It is a well-known fact that some of the human decision-making problems are too complex to be handled via traditional quantitative models (Huynh and Nakamori 2005; Lawry 2001a, b). Hence, the use of linguistic descriptors camouflages the uncertainty embedded in those problems. For improving the accuracy along with facilitating the methodologies of CWW by handling the problems involving linguistic information, distinct LCMs have been designed within the majored literature suchlike:

– *The extension principle-based computational model* (Bonissone and Decker 1986; Degani and Bortolan 1988; Dubois and Prade 1980; Klir and Yuan 1995). This computational model supports the fuzzy linguistic approach and made operations directly on the membership function defined for the linguistic terms by utilizing the extension principle (Dubois and Prade 1980; Klir and Yuan 1995). Consideration of the membership function in this computational model is essential, which is usually a tedious step in a real application.
– *The ordinal scale-based computational models* (Yager 1981). In this symbolic computational model, the information is represented in line with the fuzzy linguistic approach. It further enhances the use of the ordered structure of the set of linguistic terms to facilitate the required symbolic computations. Different ordinal scale based computational models having similar computations are presented in Delgado et al. (1993), Xu (2004a). Due to simple adaptation and ease for decision-makers (Yager 1981, 1993a, 1995, 1999), symbolic models have been extensively enforced to various problems of decision making.

However, it can be found that in both the former models, the final results so obtained do not lie within the primary expression domain. As a result, it leads to the development of an approximation process that helps in expressing a fair result, which finally matches an initial linguistic term. Nevertheless, it has witnessed quite often that there is a subsequent information loss while performing this approximation process shown as a re-translation step. Hence, it leads to an inability to provide precise results, thus provoking inaccuracy in real-life situations. So in order to resolve the shortcomings of attaining inaccuracy in the re-translation step of the CWW scheme given in Fig. 3 [taken from Martínez and Herrera (2012)], a newer fuzzy linguistic representation model has been developed. This new linguistic computational model is referred as 2-tuple which is created by a linguistic term and a number that is $(s_i,\alpha) \in L \times [-0.5, 0.5)$, where $s_i \in L$, a linguistic term and $\alpha \in [-0.5, 0.5)$, a crisp value portraying the "symbolic translation."

The central idea behind the foundation of the linguistic model based on 2-tuple is to boost the authenticity and assist the processes of CWW by considering the continuity of the linguistic domain (Martínez and Herrera 2012). This model also referred to as a continuous 2-tuple linguistic model. Further, Herrera and Martínez (2001a), in their paper, has solved a multi-expert MCDM problem that is elucidated in a linguistic context of multi-granularity using the aforementioned linguistic computational techniques. In that paper, authors have presented a comparison of the linguistic description, its authenticity, and the consistency of the results obtained through the use of these computational models. It is thereby showing that the 2-tuple model is highly efficient in managing the accuracy of linguistic information.

Table 2 Summary of MCDM model based on 2-tuple linguistic model

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
A model based on linguistic 2-tuples for dealing with multi-granular hierarchical linguistic contexts in multi-expert decision making (MEDM)	2001	Herrera and Martínez (2001b)	This paper provides a technique for dealing with the problem of MEDM, defined within a multi-granular linguistic context so that we can unify the knowledge evaluated in it without losing any information	Multi-granular linguistic framework	The proposed model is based on the 2-tuple linguistic model. The development of different functions leads to the transformation of LTSs of a linguistic hierarchy without losing any information. The proposed functions are further applied to decision models based on multi-granular linguistic information
Multi-granular hierarchical linguistic model for design evaluation based on safety and cost analysis	2005	Martínez et al. (2005)	This paper proposes a method dependent on multi-expert MCDM that handle multi-granular linguistic data without any information loss, to evaluate different design options for large engineering system is related to safety and cost criteria	Multi-granular hierarchical linguistic model	This article presents an evaluation approach for the design assessment of complicated engineering systems based on safety and cost analysis. The usage of the linguistic model has proved beneficial than traditional probability models and tools as it could able to handle vagueness and uncertainty well. Thereby linguistic approach provides a more realistic and favorable means in supporting solutions related to such complex decision problems
A fuzzy model for design evaluation based on multiple criteria analysis in engineering systems	2006	Martínez et al. (2006)	The central idea of the paper is to develop a fuzzy evaluation model depending on multi-criteria decision analysis. The proposed model deals with the information, i.e., both numerical and linguistic, in order to evaluate different design options available for an engineering system following the safety, cost, and technical performance criteria assessed within the distinct utility spaces	Fuzzy rule-based evidential reasoning	In the proposed model, to deal with the evaluation framework where information assessed in distinct expression domains and scales, that is by using the potential of heterogeneous information. Distinct fuzzy transformation functions are presented so that it permits to conduct this heterogeneous information into common utility space by way of 2-tuple linguistic information. Further, an evaluation procedure primarily based on a multi-expert MCDM model to estimate the suitability of distinctive pattern options have been developed

Table 2 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
A method for multi-attribute decision-making (MADM) with incomplete weight information in linguistic setting	2007	Xu (2007)	The central idea of the study is to propose a method to solve MADM problems well equipped with linguistic information, wherein the attribute weight information is inadequately known. In order to determine the attribute weights, an optimization model primarily based on the ideal point of attribute values is also established in the paper	ideal point of attribute values	The proposed method is well sufficient for solving MADM problems with incomplete information. Moreover, simple optimization models established have provided a simple and exact formula for obtaining the attribute weights, henceforth making the proposed method more favorable for practical application
Linguistic multi-person decision making based on the use of multiple preference relations	2009	Dong et al. (2009b)	This paper proposes a linguistic multi-person decision-making model that depends on linguistic preference relations, integrating fuzzy preference relations, distinctive types of multiplicative preference relations, and multi-granular linguistic preference relations. Furthermore, conditions underneath which the proposed decision model satisfy social choice axiom have also been discussed	Fuzzy majority and 2-tuple ordered weighted averaging (TOWA) operator, linguistic preference relation	Preference relations are extensively used in decision models. This paper has given considerable contributions to present ongoing research based on linguistic preference relations and TOWA operators. In the proposed model, a set of transformation functions has been developed to relate fuzzy preference relations and distinctive types of multiplicative preference relations with multi-granular linguistic preference relations. Further, the linguistic version of the selection procedure based on the OWA like operator and the fuzzy majority have been discussed. Finally, the internal consistency of the proposed transformation function has been analyzed in detail
Fuzzy multi-criteria decision-making (MCDM) approach for personnel selection	2010	Dursun and Karsak (2010)	This paper proposes a fuzzy MCDM algorithm depending on the principles of fusion of fuzzy information and 2-tuple linguistic model. Further, "technique for order preference by similarity to ideal solution (TOPSIS)" is also being established	Fuzzy multi-criteria decision-making approach	The proposed methodology is pertinent in managing information appraised using both linguistic as well as numerical scales, making problems with multiple information. Also, the proposed method could able to tackle heterogeneous information and thereby permits the use of distinct semantic types of the decision-maker

Table 2 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
A multi-granular linguistic model for management decision making in performance appraisal	2010	de Andrés et al. (2010)	The paper aims to propose a performance appraisal model, wherein assessments are modeled by utilizing linguistic information described by sets of distinctive reviewers to manage the uncertainty and subjectivity of such assessments	Multi-granular linguistic scheme	The proposed model provides a pretty flexible framework and thereby permits the management group to customizes how to aggregate the individual speculations and how to classify employees. Subsequently, this model suggests an increment of flexibility and an improvement in the cure of information with uncertainty and vagueness in the performance appraisal model
An innovative multi-criteria supplier selection based on 2-tuple MULTIMOORA and hybrid data	2011	Balezentis and Balezentis (2011)	This paper is an extension of multi-criteria decision-making (MCDM) method MULTIMOORA (multi-objective analysis by ratio analysis plus the full multiplicative form) to handle fuzzy supplier selection problem. This extension of MULTIMOORA primarily based on the 2-tuple linguistic representation method. Thus varied crisp and fuzzy numbers are represented, converted, and mapped into a basic linguistic term set with the help of a 2-tuple linguistic model	Multi-objective analysis by ratio analysis plus the full multiplicative form (MULTIMOORA)	MULTIMOORA is a pretty effective tool for assessing the sustainability of varied phenomena resulting in an unbiased rating of alternatives. New MCDM method MULTIMOORA-2T proposed in the paper aimed at a combination of hybrid data, specifically real numbers, interval numbers, and linguistic variables. Thus, the proposed strategy is able to handle both objective and subjective criteria
Uncertain linguistic Bonferroni mean operators and their application to multiple attribute decision making	2013	Wei et al. (2013)	This paper aims to develop two aggregation techniques called the "uncertain linguistic Bonferroni mean (ULBM)" operator and the "uncertain linguistic geometric Bonferroni mean" (ULGBM) operator for aggregating the uncertain linguistic information. The "uncertain linguistic weighted Bonferroni mean" (ULWBM) operator and the "uncertain linguistic weighted geometric Bonferroni mean" (ULWGBM) operator are defined based on which two procedures for MADM under the uncertain linguistic environments are developed	Bonferroni mean and geometric Bonferroni mean operator	The prominent advantage of the proposed operators and approaches over the traditional uncertain linguistic operators and approaches is because these operators accommodate the uncertain linguistic environment. Also, due to the consideration of the interrelationship among the input arguments, therefore making it more feasible and practical

Table 2 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
A method based on induced aggregation operators and distance measures to multiple attribute decision making under 2-tuple linguistic environment	2014	Li et al. (2014)	The objective of this article is to develop the 2-tuple linguistic induced generalized ordered weighted averaging distance (2LIGOWAD) operator, which is considered as newer multi-attribute decision-making (MADM) approach for dealing with 2-tuple linguistic information. The paper consists of some of its essential properties and further by using quasi-arithmetic mean this operator is generalized. Finally, the paper concludes with an application of a proposed operator to a group decision-making problem about the selection of strategies	Induced aggregation operators and distance measures	The 2LIGOWAD operator is capable of dealing with complicated reordering processes that represent a broad spectrum of factors in an uncertain linguistic environment. Consequently, deal in situations where the available information has a high degree of uncertainty. Further, the proposed operator is capable of dealing with complicated attitudinal characters in the decision process by utilizing order inducing variables. The 2LIGOWAD operator is generalized by using the quasi-arithmetic means to obtain the quasi-arithmetic 2LIOWAD (Q2LIOWAD). This approach includes the 2LIGOWAD operator as a special case and a lot of other cases. Thereby, we obtain a more robust formulation of this model
New linguistic aggregation operator and its application to multiple attribute decision-making	2015	Lan et al. (2015)	This paper proposes a new linguistic aggregation operator in order to deal with qualitative linguistic information, and further desirable properties like monotonicity, idempotent, commutative, and boundedness have also been studied	Uncertain linguistic aggregation operator	The proposed aggregation operator depends on the linguistic scale function that helps in converting linguistic variables to the [0,1] and, therefore, proficiently avoids the problem caused by the product of numerical values and linguistic variables, aggregated results belong to initial linguistic expression domain. Uncertain linguistic aggregation operator could able to aggregate linguistic variable with non-equidistant labels

Table 2 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
Linguistic discriminative aggregation in multi-criteria decision-making (MCDM)	2016	Aggarwal (2016)	This paper aims to develop and discuss in detail new linguistic aggregation operators termed as linguistic discriminative averaging (LDA) and linguistic discriminative ordered geometric (LDOG) operator along with their properties. Further, ordered variants and generalized variants of the proposed operators have also been investigated	Aggregation operators	The proposed linguistic aggregation operators are modeled to give weight to a criterion according to the degree of variation in the various evaluations of alternatives against it. Further, these operators are very much useful in discriminating among alternatives in MCDM problems
Adaptive linguistic weighted aggregation operators for multi-criteria decision making	2017	Aggarwal (2017)	This paper proposes new aggregation operators for MCDM under linguistic setting, and further its properties are investigated	Aggregation operators	The proposed operators are mainly designed for a MCDM environment, where the experts provide the evaluations of the alternatives against criteria in the linguistic terms, and the management decides the primary criteria weights. It is useful in a situation where primary weight information is not available

Table 3 Summary of GDM model based on a 2-tuple linguistic model and its extension

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
A group decision support approach to evaluate experts for R and D project selection	2008	Sun et al. (2008)	This paper aims to propose a group decision support approach in order to assess experts for R and D project selection wherein the criteria and their attributes are summarized majorly grounded on the experience with the "National Natural Science Foundation of China (NSFC)"	Group decision support approach	The aim of the proposed method is to present a feasible expert evaluation approach that can be used for R and D project selection and thereby both GDM approach and group decision support system are required to help aggregating and eliciting group judgments
A method for group decision making with multi-granularity linguistic assessment information	2008	Jiang et al. (2008)	This paper aims to develop a methodology to solve the GDM problems subjected to multi-granularity linguistic assessment data	Linguistic assessment information, goal programming	The methodology proposed in the paper able to express the multi-granularity linguistic information in terms of fuzzy numbers. A linear goal programming model is designed to integrate the information of fuzzy assessment and, lastly, to compute the collective ranking values of alternatives without the requirement of information transformation
Group decision making based on computing with linguistic variables and an example in information system selection	2008	Chen and Cheng (2008)	This paper proposes a transformation technique to transform the non-homogeneous linguistic information into a standard linguistic term set. Then, grounded on computing with the 2-tuple linguistic variables, a decision-making model is developed to manage the GDM problems	2-tuple linguistic variables	The linguistic variables defined in terms of 2-tuple are utilized in the proposed model representing the subjective judgment of every decision-maker. After that, the 2-tuple model and operation method are applied to deal with the aggregation of ranking and weighting among items and criteria effectively. An extended version of TOPSIS within a fuzzy environment is also developed primarily based on CWWs

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
PROMETHEE-MD-2T method for project selection	2009	Halouani et al. (2009)	The objective of this paper is to develop two newer multi-criteria 2-tuple group decision methods called "Preference Ranking Organisation Method for Enrichment Evaluation Multi Decision maker 2-Tuple-I and II (PROMETHEE-MD-2T-I and II)." These methods are capable enough to integrate inside their procedure, both quantitative and qualitative information in an uncertain context	Promethee method	The proposed methods apply to all kinds of decision-making problems with heterogeneous and multi-granular information. Therefore, the applicability of these techniques to real problems provides better outcomes in MCGDM. It offers to the decision-makers an extensive and simpler application of the aggregation operators of the PROMETHEE multi-criteria method
Group decision making with incomplete fuzzy linguistic preference relations	2009	Alonso et al. (2009)	This paper aims to develop a method to estimate missing preference values while dealing with incomplete fuzzy linguistic preference relations appraised employing a two-tuple linguistic approach	Incomplete fuzzy linguistic preference relations	The proposed methodology is useful in estimating missing values in the incomplete 2-tuple fuzzy linguistic preference relations based on the additive consistency property
The multi-attribute group decision-making method based on aggregation operators with interval-valued 2-tuple linguistic information	2012	Zhang (2012)	The paper aims to put forward an extension of the 2-tuple linguistic model known as "interval-valued 2-tuple linguistic variable." For comparing two interval-valued 2-tuples, authors have described the score and accuracy functions. Further, some aggregation operators of interval-valued 2-tuples, together with their properties and mathematical formulations, have been explained in the paper	interval valued 2-tuple linguistic information, aggregation operators	This paper develops an extended version of 2-tuple linguistic model, i.e., interval-valued 2-tuple linguistic information wherein decision-maker can express his decision in the form of an interval (i.e., $[(s_i, \alpha_i), (s_j, \alpha_j)]$, where s_i and s_j denote linguistic labels which belong to the LTS and α_i, α_j denote the symbolic translation). The proposed model obtains successful results on application to complex MCGDM problems where the decision-maker wants to express his information in a range

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
The unbalanced linguistic aggregation operator in group decision making	2012	Zou et al. (2012)	This paper aims to propose an updated linguistic aggregation operator to handle unbalanced linguistic values in GDM problems and linguistic hierarchies to express unbalanced linguistic values. Finally, the unbalanced linguistic ordered weighted geometric operator to aggregate unbalanced linguistic evaluation values is also presented	Ordered weighted geometric (OWG) operator	The unbalanced linguistic ordered weighted geometric (ULOWG) operator proposed in the paper to solve linguistic group decision-making problems where experts use unbalanced linguistic values to express their evaluation for problems. The proposed operator has several enthralling properties
Linguistic power aggregation operators and their application to multiple attribute group decision making	2012	Xu et al. (2012)	This paper aims to propose an updated linguistic aggregation operator based on a power-average (PA) operator. For the unknown weighting vector of the decision-maker, a different approach based on a linguistic power ordered weighted average operator is also proposed. Further, these approaches are extended to solve GDM problems	Power-average (PA) operator, linguistic weighted average operator	The operators that are developed in the paper can easily relieve the influence of petty arguments on the aggregated outcomes, and thereby the obtained aggregated outcomes are more reasonable. Two approaches have been developed to deal with the GDM problems under linguistic preference values when the weighting vector of the decision-maker is known and when it is unknown. The prominent characteristic of the approaches that have been developed is that they take all the decision arguments and their relationships into account. Further, an extension of the operators to the uncertain linguistic environment has also been developed
Some interval-valued 2-tuple linguistic aggregation operators and application in multi-attribute group decision making (MAGDM)	2013	Zhang (2013)	This paper proposes a method that deals with MAGDM problems based on interval-valued 2-tuple linguistic information. Some new aggregation operators and their desirable properties have been discussed in detail. Furthermore, an updated methodology to decide the weight vector of the interval-valued 2-tuple aggregation operator primarily based on the idea of the degree of precision has also been presented	Interval-valued 2-tuple linguistic information, degree of precision	Interval-valued 2-tuple linguistic model is suitable in handling the MAGDM problem described under a multi-granular linguistic notion as it is capable of fully expressing the information and also able to unify it easily without involving any tedious aggregation steps

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
Multi-attribute group decision making (MAGDM) with multi-granularity linguistic assessment information: an improved approach based on deviation and TOPSIS	2013	Liu et al. (2013)	This paper aims to propose an updated method concerning GDM problems associated with multi-granularity linguistic assessment information. The computational formula is developed to transform and unify the multi-granularity linguistic comparison matrices. Furthermore, the technique of standard and mean deviation to decide the attribute weights that are not known is applied. Finally, the weights of the decision-makers will be decided by using the extended TOPSIS method	"TOPSIS (technique for order preference by similarity to an ideal solution) method", standard and mean deviation	The proposed method could able to overcome the limitation of the classical approaches. It can combine a standard and mean deviations as well as the TOPSIS method. Further, the MAGDM approach transforms multi-granularity linguistic information simpler. Along with this, it takes into account the weights of the attributes and decision-makers by the process of aggregating the assessment information and objectively weighing the attributes and decision-makers
Multi-attribute group decision making with aspirations: A case study	2014	Feng and Lai (2014)	This paper aims to propose an integrated multi-attribute decision-making approach for problems with consideration of the decision maker's aspirations. By solving the case problem of China Southern Airlines (CAS)	Aspiration utility function	This study presents an integrated decision-making approach for solving multi-attribute group decision-making problems with aspirations. The proposed approach integrates aspirations into the utility theory for MADM. Further, it uses optimization techniques for reducing the gap between group opinion and individual opinions, thereby allowing managers to make more reliable decisions. It also provides a consistency coefficient to managers for checking the group decision-making quality. Lastly, it can accommodate complex decision data by incorporating numerical values, interval numbers, linguistic terms, and uncertain linguistic terms

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
Method of multi-criteria group decision making based on cloud aggregation operators with linguistic information	2014	Qiang et al. (2014)	The central idea of the paper is to define cloud operators (such as "cloud weighted arithmetic averaging (CWAA)" operator, "cloud-ordered weighted arithmetic averaging (COWA)" operator, and "cloud hybrid arithmetic (CHA)" operator), followed by its application in the MCGDM problem. Further, in the paper, a section is devoted, which explains the conversion between linguistic variables and clouds	cloud model, aggregation operator	Traditional methods existing in the literature are not robust enough to convert qualitative concepts to quantitative information in linguistic MCDM problems; neither can they completely reflect the fuzziness and randomness inherent in qualitative concepts. However, the cloud model proposed in the paper overcome these difficulties. The three numerical characteristics of the cloud model render the transformation between this qualitative concept and quantitative information smoothly and effectively
Consensus-based group decision making under multi-granular unbalanced 2-tuple linguistic preference relations	2015	Dong et al. (2015a)	In this paper, the GDM model with multi-granular unbalanced 2-tuple linguistic preference relations is proposed. Both balanced and unbalanced multi-granular linguistic preference relations are related with the help of transformation function. Further, a consensus model is presented to help the decision-makers reach a consensus	Proportional 2-tuple linguistic model	The proposed multi-granular linguistic GDM model is based on the linguistic model of 2-tuple. The transformation function presented in the paper can be reduced into a traditional transformation function given in Herrera and Martínez (2001b) beneath the context of a multi-granular balanced linguistic environment. The consensus model so proposed sets out a new means to manage both the consistency of an individual and group via a linear programming model. The model also stipulates a novel means by minimizing the loss of information when the consensus has been established

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
A model based on subjective linguistic preference relations for group decision-making problems	2016	Massanet et al. (2016)	In this paper, a novel definition of preference relation, the so-called subjective linguistic preference relation, is proposed. These preference relations are based on the concept of subjective evaluations, introduced in the LCM based on discrete fuzzy numbers. Further, an example of a multi-expert decision-making problem with a hierarchical multi-granular linguistic context is analyzed to illustrate the potential of the proposed method and its advantages concerning other methods	Subjective linguistic preference relation	The advantage of the proposed model is that it offers experts more flexibility to express their opinions, and also, this model guarantees no loss of information
A method for multi-criteria group decision making with 2-tuple linguistic information based on cloud model	2017	Zhang et al. (2017a)	This paper aims to propose a 2-tuple hybrid ordered weighted geometric (THOWG) operator, which synthetically considers the importance of both individual and the ordered position to overcome the defect of the existing operators. Further, this paper develops a new cloud generation method to transform 2-tuple linguistic variables into clouds	Cloud model	The advantage of the proposed method is the development of new THOWG operators to overcome the limitations of existing 2-linguistic power aggregation operators. Traditional 2-tuple linguistic operators either ignore the importance of the individual or neglect the importance of the ordered position. Further new cloud generating methods to transform 2-tuple linguistic variables into clouds have been presented. This method integrates the significant advantages of the cloud model so that it can handle the randomness of natural languages, which will significantly improve the decision quality. Finally, some new cloud algorithms: cloud distance, cloud possibility degree, and cloud support degree have also been developed

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
Personalized individual semantics in computing with words for supporting linguistic group decision making. An application on consensus reaching	2017	Li et al. (2017)	This paper introduces a personalized individual semantics (PIS) model to tailor individual semantics by a numerical scale model primarily based on intervals and the traditional linguistic model of 2-tuple. Further, to obtain and express the PIS model, a consistency-driven optimization-based model is also proposed. Finally, to justify the feasibility and validity of the PIS model, it is applied to a linguistic GDM problem with a consensus reaching process, by defining the notion of the individual linguistic understanding	Interval numerical scale, 2-tuple linguistic model	The introduction of PIS methodology to model and solve linguistic GDM problems with preference relations could able to improvise the management of different meanings of words for different people. The proposed model is based on the numerical interval scale. This is because it could able to handle distinct linguistic representations in a more precise way. Moreover, a novel frame of work developed for handling PIS could able to redesign the phases of CWW, which helps in obtaining tailored linguistic results easier to interpret and understandable by humans
The 2-rank consensus reaching model in the multi-granular linguistic multiple-attribute group decision making (MAGDM)	2017	Zhang et al. (2017b)	In this paper, a 2-rank consensus reaching model with minimum adjustments for the 2-rank MAGDM problem is developed. An optimization model to support a 2-rank consensus rule is also proposed, which is further transformed into a mixed 0–1 linear programming model. Finally, a comparative analysis of the proposed model with the existing consensus models is also presented	2-Rank consensus reaching model	Preferences of the decision-makers are expressed by using the context of multi-granular linguistic term sets. Optimization-based model is put forwarded to support the consensus rule. The optimal adjusted uniform linguistic decision matrices are transformed into multi-granular linguistic decision matrices, which are used as the references for decision-makers to modify their preferences. Further, an iterative 2-rank consensus reaching process with the minimum adjustments is also developed. Moreover, many useful properties of the proposed model are also discussed

Table 3 continued

Title	Year	Paper	Description	Approach/strategy	Strong/weak points
Group decision making based on linguistic distributions and hesitant assessments: Maximizing the support degree with an accuracy constraint	2018	Wu et al. (2018a)	In this paper, a novel approach is designed to propose a new linguistic group decision model called the maximum support degree model (MSDM). This model aims to ensure maximum support degree of the group opinion and further guarantying the accuracy of the group opinion. Further, a mixed 0–1 linear programming approach is also presented to solve the MSDM. Finally, the utility and validity of the MSDM in multiple attribute group decision making are proven	Linguistic distributions and hesitant fuzzy linguistic term sets (HFLTSs)	The aggregation methodology proposed in this study ensures maximum degree support of the group opinion in linguistic distributions based on GDM problems and further guarantees that the opinion of the group is an HFLTS which has a certain degree of accuracy. In the MSDM, linguistic distributions are used to express decision-makers' individual preferences to provide more probabilistic preference information over linguistic terms, and the HFLTS with certain accurate constraint is employed to guarantee the accuracy of the group opinion

4.2 Definitions

Herrera and Martínez (2000a) has initially introduced the notion of a 2-tuple model where the linguistic information is represented in a continuous domain. Henceforth, any counting of information is expressed within a process of aggregation. This subsection mainly describes the definitions and mathematical formulations of the 2-tuple model.

Definition 1 (Herrera and Martínez 2000a) Let $L = \{\ell_i \,|i{=}0, \ldots, g\}$ be defined as the finite and totally ordered linguistic term set (LTS) having cardinality $g+1$ and let $\beta \in [0, g]$ be the numeric value. Let $i \in \{0, \ldots, g\}$ and $\alpha \in [-0.5, 0.5)$ be two values then α is called a *symbolic translation* which means a numerical value representing the value of translation from initial result β to the nearest index label i. Here, i is denoted as round(β) while $\alpha = \beta - i$.

This 2-tuple linguistic model is described via transformation function established between numeric values and 2-tuples, as defined above, to carry forward the linguistic computational processes.

Definition 2 (Herrera and Martínez 2000a) Let $L = \{\ell_i \,|i{=}0 \text{ to } g\}$ be defined as a finite LTS and let $\beta \in [0, g]$ be a numeric value that express the outcome of a symbolic aggregation operation, then the 2-tuple which represents the identical information to β is given below:

$$\Delta : [0, g] \to \overline{L}$$
$$\Delta(\beta) = (\ell_{\text{round}(\beta)}, \beta - i)$$

where $\overline{L} \equiv L \times [-0.5, 0.5)$, round(.) is the usual round operation which assigns to crisp value β an integer value $i \in \{0, 1, \ldots, g\}$ nearest to β, and $\alpha = \beta - i$, is termed as the "symbolic translation."

Remark 1 (Herrera and Martínez 2000a) Δ mapping as defined above is one-one and onto hence invertible; therefore, there exist $\Delta^{-1} : \overline{L} \to [0, g]$ which returns an identical crisp value $\beta \in [0, g] \subset \mathbb{R}$ corresponding to equivalent 2-tuple such that $\Delta^{-1}(\ell_i, \alpha) = \beta = i + \alpha$.

Remark 2 2-Tuple linguistic term is obtained from a linguistic term ℓ_i by merely including a value 0 as "symbolic translation": $\ell_i \in L \Rightarrow (\ell_i, 0) \in L \times [-0.5, 0.5)$.

Remark 3 Linguistic model based on 2-tuple can be utilized from any membership function maintaining the semantics of the given linguistic terms and also upgrade the authenticity of the traditional symbolic ways. Nevertheless, in the paper (Herrera and Martínez 2000a, 2001a) it was demonstrated that the utilization of the triangular shaped membership functions brings about the most likely outcome with high precision.

Further, Herrera and Martínez (2000a) presented the comparison of 2-tuple linguistic information by using conventional lexicographic order.

Definition 3 Let (ℓ_i,α_i) and (ℓ_j,α_j) be two 2-tuple linguistic variables.

- if $i < j \implies (\ell_i,\alpha_i) < (\ell_j,\alpha_j)$.
- if $i = j$ then,
  1. if $\alpha_i = \alpha_j \implies (\ell_i, \alpha_i) = (\ell_j, \alpha_j)$;
  2. if $\alpha_i < \alpha_j \implies (\ell_i, \alpha_i) < (\ell_j, \alpha_j)$;
  3. if $\alpha_i > \alpha_j \implies (\ell_i, \alpha_1) > (\ell_j, \alpha_2)$.

Definition 4 (Herrera and Martínez 2000a) Let $L = \{\ell_i \,|i{=}0, \ldots, g\}$ be the finite LTS with cardinality $g+1$, then negation operator of a 2-tuple is given as

$$Neg((\ell_i, \alpha)) = \Delta(g - (\Delta^{-1}(\ell_i, \alpha))).$$

Furthermore, the generalized version of the 2-tuple linguistic variable and the translation function is proposed by Chen and Tai (2005).

Definition 5 (Chen and Tai 2005) Let $L = \{\ell_i \,|\, i{=}0, \ldots, g\}$ be the finite LTS with cardinality $g+1$ and let β be the crisp value such that $\beta \in [0, 1]$, then the transformation of β to 2-tuple linguistic variable is procured in the following aspect:

$$\Delta : [0, 1] \to \overline{L}$$
$$\Delta(\beta) = (\ell_i, \alpha) \quad \text{with} \quad \begin{cases} \ell_i, & i = \text{round}(\beta\, g), \\ \alpha = \beta - i/g, & \alpha \in [-0.5/g, 0.5/g)\,. \end{cases}$$

where $\overline{L} \equiv L \times [-0.5, 0.5)$, round(.) is the general round operation, ℓ_i being the nearest index label to β, and α represents the "symbolic translation."

On the contrary, $\Delta^{-1} : \overline{L} \to [0, 1]$ can also be defined to convert the 2-tuple linguistic variable into an equivalent crisp value $\beta(\beta \in [0, 1])$ in the following aspect:

$$\Delta^{-1}(\ell_i, \alpha) = \beta = i/g + \alpha.$$

4.3 Extensions of the 2-tuple linguistic model

Being the most beneficial model for CWW due to its more straightforward interpretation and competence of expressing any linguistic data in a continuous domain, the 2-tuple model has a wide range of applications. Over the subsequent years, many researchers have proposed variant versions of this model, stating its mathematical formulations, operations, and aggregation operators. This subsection has brought in light about the extensions of the 2-tuple linguistic model in

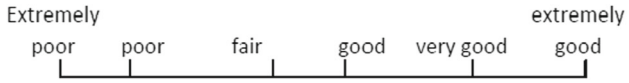


Fig. 4 Unbalanced linguistic set type I



Fig. 5 Unbalanced linguistic set type II

addition to its applications in decision making. Below we summarize some of the extensions of 2-tuple model individually by highlighting the main results.

Unbalanced linguistic term sets (ULTSs)

It is known that Herrera and Martínez' 2-tuple model is alleged to address only the LTS having the symmetric distribution of linguistic labels. However, some problems are subjected to term sets having uneven distribution of linguistic labels, such type of set is termed as ULTS Herrera et al. (2008). There are mainly two categories of ULTS. In the first category, the distribution of the linguistic terms is uneven such that the cardinality of the terms on one side of the intermediate term is higher than the other and corresponding distance between the consecutive terms is not equal (namely ULTS of Ist type, see Fig. 4). Meanwhile, in the second category, an equal number of the linguistic terms are distributed on the left and right sides of the intermediate term with unequal spacings (namely ULTS of the 2nd type, see Fig. 5).

Herrera et al. (2008) presented a fuzzy linguistic approach to handle a ULTS. In that paper, authors have introduced the hierarchical linguistic concept used as a basis of representation for portraying unbalanced linguistic information. Also, in the paper, an unbalanced linguistic computational model (ULCM) has been proposed primarily based upon the framework of a 2-tuple model to execute processes of CWW in a more precise way without losing a single bit of information.

In sequence with the computational models to address unbalanced linguistic information, Wang and Hao (2006, 2007) presented an updated formulation of the current 2-tuple linguistic model that is known as a proportional 2-tuple model. This proportional 2-tuples are subjected to a notion of symbolic proportion and canonical characteristics values (CCV) of linguistic terms. Dong et al. (2009a) proposed the idea of a numerical scale (NS) and further extended the 2-tuple linguistic models beneath the NS. In that paper, the notion of the transitive calibration matrix and its consistent index was presented and henceforth developed an optimization model to evaluate the NS of the LTS. The desired properties of the optimization model have also been presented.

Guo and Huynh (2013) suggested a proportional 2-tuple fuzzy linguistic screening evaluation model and the so-called preference-preserving transformation based on CCV. In that paper, the authors presented a newer screening assessment model that not only refines the standard requirement of equal distance between labels. It also can reflect evaluators' level of confidence in decisions, enriching the knowledge of results, and thus providing decision-makers with more detailed guidance.

Interval judgments thoroughly endorse several decision-making models. As a consequence, Dong et al. (2013) developed a new LCM primarily based upon the 2-tuple and intervals, which was termed as an interval version of the 2-tuple linguistic model. Authors have defined the notion of interval NS that converts linguistic terms into interval numbers and has further introduced a generalized inverse operation of interval NS. By continuing the study of the interval 2-tuple linguistic model, Dong and Herrera-Viedma (2014) introduces a novel consistency-driven methodology in the problem of decision making along with the linguistic preference relations. The proposed methodology stipulates a peculiar way out to set the interval NS without requiring the semantics given by interval type-2 fuzzy sets (IT2FSs). Furthermore, Dong et al. (2016) proposed a methodology to connect two distinct models designed to address ULTS, i.e., LCM based on hierarchical linguistic concept (Herrera et al. 2008) and the NS model (Dong et al. 2009a). Moreover, authors have designed a novel CWW methodology relied on the NS model and ULTS to handle hesitant unbalanced linguistic information. Henceforth, the NS model establishes an integrated framework to link the classical linguistic model of 2-tuples, the proportional 2-tuples, and the hierarchical linguistic model.

Cai et al. (2016) developed a new LCM based on symbolic models. Pei and Zheng (2017) constructed a series of the normal distribution, which is based on unbalanced linguistic scale sets and has furthermore developed a novel approach for defining unbalanced linguistic information in a format of 3-tuple. Lastly, Cai and Gong (2018) redefined the concept of ULTS and provided the graph for the representation model.

Linguistic distribution (LD)

According to Wang and Hao model Wang and Hao (2006), the linguistic information is provided in terms of proportional 2-tuples where symbolic proportions are given to two consecutive ordinal linguistic terms. However, it is envisaged that the proportional 2-tuple model seems improper to handle decision situations where performances of the alternatives are evaluated via uncertain linguistic judgments as examined in the papers (Huynh et al. 2006; Xu 2012; Yan et al. 2011; You et al. 2015). As a consequence, to overcome the limitations offered by the mentioned model, Zhang et al. (2014) has introduced a notion of the distribution assessment in an LTS where symbolic proportions are assigned to all the linguistic terms. In that paper, authors have developed a consensus

model to facilitate decision-makers in improving the level of consensus. Followed by discussing the desirable properties of the model as well as measures of consistency and consensus in favor of linguistic preference relations (LPRs) centered upon distribution assessments. Lastly, some operational laws and aggregation operators on account of linguistic distribution assessments are also introduced.

Meanwhile, considering both Wang and Hao's model (Wang and Hao 2006) and Zhang et al.'s model (Zhang et al. 2014) it is envisioned that the symbolic proportions involved in them are summing up to 1. Henceforth, both these models are inadequate to handle incomplete information. So, to overcome this limitation, Guo et al. (2017) proposed a new proportional LD model for modeling incomplete linguistic assessments addressed in MADM problems. However, given varying subjective conditions, it found challenging for decision-makers in providing exact symbolic proportions in LD assessments. As a consequence, Dong et al. (2015b) developed a notion of LD assessments in terms of interval symbolic proportions under a multi-granular unbalanced linguistic context. Furthermore, Zhang et al. (2016) introduced a new LCM for solving large-scale group decision-making (LGDM) problem to model linguistic information in multi-granular context by allowing the maximum information at the first stage and further removal of the initial processes of aggregation as well as managing the required information given by experts along with the utilization of LD assessments to acquire accurate results. Wu et al. (2018a) presented a novel linguistic GDM model referred to as the maximum support degree model (MSDM) primarily based upon the utilization of LD Zhang et al. (2014) and hesitant fuzzy linguistic term sets (HFLTSs) Rodríguez et al. (2011). The proposed model aims at providing maximum degree support of the group opinion and further guarantees the accuracy.

Although the HFLTSs and LDs are appropriate enough to represent linguistic preferences in a more realistic form in an uncertain decision environment, they are not flexible enough due to some limitations. In the case of the HFLTSs, several consecutive linguistic labels are used to extract the preferences of the decision-makers' while LDs present certain probabilistic preference information over linguistic terms. However, the elicitation of preferences is not closer to humans' cognitive processes. As a consequence, Wu et al. (2019) recently developed a more comprehensive format of linguistic expressions ideally termed as flexible linguistic expression (FLE) to substantially improve the flexibility in constructing complex linguistic expressions as well as eliciting linguistic preferences. Further, in the paper, authors developed a novel linguistic GDM model along with FLEs, termed as FLE-based GDM (FLEGDM). For the detailed explanation of the proposed methodology, one can refer to the corresponding paper (Wu et al. 2019). In the sequent, Wu et al. (2020) proposed a notion of flexible linguistic preference relations (FLPRs) where linguistic expressions are expressed in a more flexible way to generalize all kinds of linguistic preference relations (LPRs). Further, in that paper, the authors presented a novel methodology to rank alternatives centered upon preference information in FLPRs by envisaging the LD as well as priority approximation (PA) of FLPRs. Lastly, a comparative study for the obtained results of priority vectors is put forward to validate the proposal.

Multi-granularity linguistic term sets (MGLTSs)

The problems of GDM addressing linguistic information incorporates the complexity and further necessitates the involvement of several experts in the decision-making process. In the specialized literature, the majority of the proposals designed for solving GDM problems are subjected to situations where experts express linguistic information within the same LTS (Herrera and Martínez 2000a; Lawry 2001a). However, in many practical, real-life problems, it is envisioned that most of the experts have their distinct cultures, qualifications, knowledge background, and skills. Also, their decisions are taken according to different circumstances. As a consequence, the representation of the information via unique LTS seems inadequate. Thus, in such cases, experts often use LTSs with different cardinalities, commonly referred to as MGLTSs, in order to represent their individual assessment information without any information loss.

The research on the problems of decision making assessed with information in a multi-granular context holds considerable importance in practical applications, and several methods have been developed in the literature to solve problems involving MGLTSs. Morente-Molinera et al. (2015) presented a systematic review of the recent enhancements in multi-granular linguistic methodologies from 2000 to 2014. In that review, the authors have presented an in-depth analysis of the respective methodologies designed for solving GDM problems as well as more relevant applications. Further, Morente-molinera et al. (2017) proposed a novel methodology for solving MCGDM problems involving a high amount of alternatives. In that paper, fuzzy ontologies reasoning procedures are utilized in order to obtain the alternatives ranking classification, and multi-granular linguistic methodology automatically is being used to improve human–computer communication. The proposed method is highly advantageous and is useful for cases involving a high amount of alternatives. For a detailed understanding of the method, one can refer to paper (Morente-molinera et al. 2017).

Tian et al. (2018) presented a novel approach to managing hesitant fuzzy linguistic MCGDM problems with multi-granular unbalanced LTS. In that paper, authors have initially developed a signed distance measure, which is considered as the support tool for hesitant fuzzy linguistic term set (HFLTS). This measure is primarily based on the ordinal

semantics of linguistic terms and the possibility distribution method. For the unification of the multi-granular hesitant unbalanced linguistic information, a signed-distance-based transformation function is proposed. Furthermore, a comprehensive consensus measure is presented so that it can robustly measure the consensus degree among decision-makers and also give guidance to members in order to modify their judgments before the selection process. Lastly, a numerical example validates the applicability of the proposed methodology in the practical scenario. The results and an in-depth comparative analysis of the proposed approach with the existing methods (Wu and Xu 2015) demonstrate the possibility and effectiveness of the proposed methodology in tackling MCGDM problems with HFLTSs based on balanced LTSs to some degree. The proposed approach is an upgraded and generalized version of Wu and Xu (2015) method.

Heterogeneous Information

Nowadays, the decision-making process has become one of the essential human activities. In many complex real-life decision problems, the attributes can be either quantitative or qualitative. The quantitative attribute values can be given using distinct numerical types suchlike real numbers, fuzzy numbers (triangular or trapezoidal), interval numbers, and so forth. Complex GDM problems often contain heterogeneous information, i.e., the information in a mixed form. However, heterogeneous GDM problems are specified in 3 frameworks. The very first framework is related to different preference formats where opinions of the decision-makers are expressed by different preference relations such as preference orderings, utility functions, multiplicative preference relations, and fuzzy preference relations. The second framework appears when each expert has different levels of knowledge and background related to the problem, or when the experts have different linguistic label sets to assess preferences, i.e., the multi-granular and unbalanced linguistic contexts. The third framework focuses majorly on the heterogeneous expressions of the experts, which are used to express or provide their particular preferences for the attributes of each alternative. It provides information about attributes, which consists of not only crisp or uncertain information, but also interval numbers, fuzzy numbers, and linguistic data (Li et al. 2016). Several authors have proposed different methodologies to solve GDM problems assessing information in a heterogeneous form (Li et al. 2016; Liu et al. 2017; Martínez et al. 2007a; Zhang and Guo 2014). For a more detailed study of the methodology, one can refer to Li et al. (2016), Liu et al. (2017), Martínez et al. (2007a), Zhang and Guo (2014).

4.4 Contribution of 2-tuple linguistic model

Human beings are always engaged in making decisions beneath the linguistic environment to tackle the ill-structured problems successfully. Several methods have developed in the literature for managing linguistic information like method grounded on extension principles and symbols. Both these techniques have certain limitations, which is resolved by the introduction of a 2-tuple model. The 2-tuple model has improved the accuracy and facilitated the processes of CWW by managing the linguistic information in the continuous linguistic domain. In the present section, we emphasize the decision models centered upon the 2-tuple model as well as its extension.

On the broader prospect, MADM and MCGDM problems are two majorly considered decision-making problems. MADM process ideally composes of two steps: (1) collection of decision information about attribute weights as well as attribute values. In the MADM problem, there exists a finite set of alternatives. The attributes are pointers that measure the given alternatives and, thereby making each attribute as necessary, determine within the procedure of decision making. (2) Aggregating the decision information. It is one of the most prominent concepts in the decision-making problem as different experts come up with different knowledge areas and preferences. As a consequence, it is the process of merging an entire collection of information into a particular representative value. Hence, aggregation operators have become one of the most prominent tools while merging of information in a decision-making problem. Henceforth, it has become the most populous studied field among researchers and has drawn a great deal of consideration from practitioners among various disciplines (Chen and Hwang 1992; Fodor and Roubens 1994; Hwang and Yoon 1981; Saaty 1980; Yager and Kacprzyk 1997). Table 2 shows the summary of the existing literature related to the MADM model primarily based upon the 2-tuple model.

Over the years, as the complexity of the socioeconomic environment increases (Chen and Hwang 1992). Decision-making processes have become one of the most wide-ranging and prominent application areas and have attracted various practitioners and researchers. Many institutions have turned from single decision-makers to group decision-makers to successfully and proficiently complete the task. A simple GDM problem is defined as a problem where the decision making is considered with several alternatives and experts who are trying to achieve a peculiar solution by considering their preferences or point of view. Ideally, the GDM problem is composed of two steps:

1. *Aggregation Phase* A phase that aggregates experts' preferences.
2. *Exploitation Phase* A phase that helps in obtaining a solution set of alternatives in the decision problem.

Table 3 is arranged in a chronological order of publication and outline briefly about the summary of the existing literature

Table 4 Application based on 2-tuple linguistic model

Title	Paper
Supply chain	Li and Xie (2009), Yeh et al. (2007)
Nuclear safeguards	Liu et al. (2002), Rodríguez et al. (2010a)
Selection process	Balezentis and Balezentis (2011), Halouani et al. (2009), Sun et al. (2008)
Engineering systems	Alcalá et al. (2009), Martínez et al. (2006)
Recommender system	Martínez et al. (2008a), Martínez et al. (2007b), Porcel and Herrera-Viedma (2010), Porcel et al. (2012), Rodríguez et al. (2010b)
Risk evaluation	Chang and Wang (2010), Liu et al. (2011), Pei and Shi (2011), Yu (2009), Zhou et al. (2011)

related to a GDM model based on the linguistic model of 2-tuple.

Due to the societal developments and an increase in the complexity of the problems, we come across several decision problems where a group of individuals is having differences in cultural and knowledge background. As a result, decision-makers may represent their respective evaluations using MGLTSs. In the literature, several methodologies have been developed to tackle the problems of GDM involving multi-granularity linguistic information. Zhang and Guo (2012) proposed a GDM method where decision-makers are required to express their assessment using multi-granularity in an uncertain linguistic environment along with insufficient information of the attributes' weight. The information of the evaluation is transformed into a trapezoidal fuzzy number (TFN). Optimization models are established to minimize the deviation between the evaluation of each decision-maker and the group's collective evaluation for each alternative. These models of optimization are solved to obtain the group's collective evaluation value of the alternatives. Finally, the closeness coefficient and ranking for each alternative are evaluated. The optimal alternative based on the ranking can be further selected. The method proposed is simple and easy to understand. Henceforth, it is well applicable in the GDM problems involving multi-granularity uncertain linguistic and insufficient weight information.

Li (2009) introduced the concept of extended linguistic variables to propose a new methodology for solving multi-attribute GDM problems under linguistic assessments. Further, Ju (2014) proposed an upgraded method to solve MCGDM problems, where both the criteria values and criteria weights take the form of linguistic information, and the information about linguistic criteria weights is partly known or completely unknown. Morente-Molinera et al. (2015) proposed a systematic literature review on multi-granular fuzzy linguistic modeling in GDM from the year 2000 and 2014 and further discussed their drawbacks and advantages.

With the developing headway of the 2-tuple linguistic model in decision-making processes and other disciplines, it had been scrutinized that multi-objective linguistic optimization problem concomitant with the 2-tuple linguistic model was a significant advantage. Multi-objective optimization problems helped manage problems that were engrossed with conflicting objectives. There exist class of issues, where every objective function was self-governed and thereby cannot be solved by transforming into multiple objective optimization problems. Numerous methods had been developed in the literature to solve such problems (Carlsson and Fullér 2000, 2001). These methods were primarily based on Tsukamoto's fuzzy reasoning method, where the objective function was converted into a crisp form, and further, the objective function was solved by using any conventional optimization technique. Nevertheless, this methodology endured through a limitation as the resulting output of the problem was attained in crisp form rather than linguistic. To overcome this limitation, Gupta and Muhuri (2017) developed a newer method grounded upon the 2-tuple linguistic model in solving multiple linguistic optimization problems.

In certain situations, for instance, the problem of the social network, e-democracy, and so forth decisions are required to be taken by relatively larger group size, categorized as the large-scale group decision making (LSGDM). By defining a consensus model based on PIS for LSGDM, Li et al. (2018a) gave the first proposal in this field. For the detailed knowledge of the model, one can refer to the corresponding paper (Li et al. 2018a).

5 Application of the 2-tuple linguistic model and its extension

The linguistic model of 2-tuple is prudent over other existing approaches by overcoming the restrictions experienced by conventional methods. After having reviewed the dominant position that the 2-tuple linguistic model holds among variant LCMs as well as other decision models, we move forward to show extensive applications of the model in various disciplines. Table 4 shows the application depending on the 2-tuple linguistic model as well as its extension.

The 2-tuple linguistic model is used for segmentation of color images (Orduna et al. 2014). For this, a set of experts are being provided such that every object of the con-

sidering image is assigned degree of every pixel by each expert. The objects are represented by means of fuzzy linguistic labels and then by using the decision-making model based on 2-tuples an aggregation phase to classify each pixel is applied. Other application areas of 2-tuple linguistic model are social judgment analysis (Wu and Mendel 2010a), weapon system evaluation (Wu and Mendel 2010b), cooking recipe recommendation (Bilgin et al. 2013, 2016), decision making (Chen and Ben-Arieh 2006; Dong et al. 2015a; Herrera et al. 2009; Li et al. 2017; Lin et al. 2009; Liu and Chen 2018; Liu and Wang 2017; Liang et al. 2019; Nie et al. 2017; Pedrycz et al. 2010; Sohaib et al. 2019; Wu et al. 2018b; Xu 2004b), personalized individual semantics (Li et al. 2018b), renewable energy (Doukas et al. 2010), energy optimization(Gupta and Muhuri 2014; Muhuri et al. 2017), natural language generation (Kacprzyk and Zadrozny 2010), supplier selection problems (Wan et al. 2017; Wang 2008; Wen et al. 2020), sensory evaluation model (Martínez et al. 2008b, 2009; Martínez 2007), a cloud-based decision model for low-speed wind farm site selection (Wu et al. 2017), emergency management (Qi et al. 2018), emergency decision making (Zhang et al. 2018), financial technologies selection (Mao et al. 2019), analysis of information and communication technology (Cid-López et al. 2016), information retrieval system (Herrera-Viedma and López-Herrera 2007), hotel selection (Yu et al. 2018), non-cooperative matrix games (Singh et al. 2018; Singh and Gupta 2018).

6 Conclusion and summarization

Many practical situations certainly deal with uncertainty and vagueness and further increase the complexity of the problem. The use of probabilistic models to deal with such uncertainty has resulted in unsatisfactory results. There are different computational models based on the linguistic term set to accomplish CWW processes, among them the 2-tuple linguistic model CWW has produced successful results without any information loss and additionally applicable in the complicated situations. It has been extensively applied in the field of decision making and many other related areas due to its accuracy and simplicity. This review presents a comprehensive study of this 2-tuple model, its foundation, and use in various problems of decision making. This study also provides a brief outline of the various extensions of the 2-tuple model.

The implication of this contribution is to provide a better understanding of this model and further bring forth its broad application witnessed in the field of decision analysis. Although the 2-tuple model has gained considerable attention among other LCMs, there are still some challenges persisting with the model. Many times over, it is examined that in solving complicated decision problems, the 2-tuple model may not deliver results with greater precision and flexibility. Nevertheless, the inflexible results exist due to the explicit nature of the linguistic variables. Henceforth, the most significant challenge is to increase the efficiency of the model. Moreover, it necessitates improving the model in a way to enrich the vocabulary to express the results further. It is noteworthy that the process of aggregation in LSGDM problems is usually a tedious and time-consuming job. As a consequence, decreasing computation time during the aggregation framework in such problems and together with increasing authenticity in results without loss of information remains a significant area of research.

It is prudent to say that the literature presented in this contribution provides a systematic categorization of various articles based upon the 2-tuple linguistic model and its extension in the field of decision making. The review is designed in a way to help researchers and practitioners working in this area to improve this model further substantially, henceforth developing ahead to enhance its feasibility and applicability in other areas like game theory, goal programming, bimatrix, data envelopment analysis (DEA), etc. The significance of this contribution can be assessed from the industrial, academic, and end-user levels; as decision analysis and uncertainty theory and decision models centered upon the 2-tuple linguistic approach have been widely studied, academics and young students will benefit. In young minds, the ideas that have been presented, and the literature reviewed may inculcate novel ideas that may originate at an early stage. As a result, researchers and practitioners around the globe will be attracted to devise and extend the 2-tuple linguistic model further.

Acknowledgements The work was supported by Delhi Technological University (Grant number DTU/IRD/619/2019/2177).

Compliance with ethical standards

Conflict of interest The authors declare that they have 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

Adamopoulos GI, Pappis GP (1996) A fuzzy linguistic approach to a multi criteria sequencing problem. Eur J Oper Res 92:628–636

Aggarwal M (2016) Linguistic discriminative aggregation in multi-criteria decision-making. Int J Intell Sys 31(6):529–555

Aggarwal M (2017) Adaptive linguistic weighted aggregation operators for multi-criteria decision-making. Appl Soft Comput 58:690–699

Alcalá R, Alcalá-Fdez J, Gacto MJ, Herrera F (2009) Improving fuzzy logic controllers obtained by experts: a case study in HVAC systems. Appl Intell 31(1):15–30

Alonso S, Cabrerizo FJ, Chiclana F, Herrera F, Herrera-Viedma E (2009) Group decision making with incomplete fuzzy linguistic preference relations. Int J Intell Syst 24(2):201–222

Balezentis A, Balezentis T (2011) An innovative multi-criteria supplier selection based on two-tuple multimoora and hybrid data. Econ Comput Econ Cybern Stud Resarch 2:1–20

Bilgin A, Hagras H, Malibari A, Alhaddad MJ, Alghazzawi D (2013) Towards a linear general type-2 fuzzy logic based approach for computing with words. Soft Comput 17(12):2203–2222

Bilgin A, Hagras H, Helvert JV, Alghazzawi D (2016) A linear general type-2 fuzzy logic based computing with words approach for realizing an ambient intelligent platform for cooking recipe recommendation. IEEE Trans Fuzzy Syst 24(2):306–329

Bonissone PP (1980) A fuzzy sets based linguistic approach: theory and application. In: Oren TI, Shub CM, Roth PF (eds) Proceedings of the 1980 winter simulation conference, 80CH1617-0/80/0000-0099. IEEE

Bonissone PP, Decker KS (1986) Selecting uncertainty calculi and granularity: an experiment in trading-off precision and complexity. In: Kanal LH, Lemmer JF (eds) Uncertain Artif Intell. North-Holland, Amsterdam, pp 217–247

Bordogna G, Pasi G (1993) A fuzzy linguistic approach generalizing Boolean information retrieval: a model and its evaluation. J Am Soc Inf Sci 44:70–82

Bordogna G, Fedrizzi M, Pasi G (1997) A linguistic modeling of consensus in group decision making based on OWA operators. IEEE Trans Syst Man Cybern 27:126–132

Cai M, Gong Z (2018) Group decision making using distances between unbalanced linguistic assessments. Appl Soft Comput J 67:613–624

Cai M, Gong Z, Yu X (2016) A method for unbalanced linguistic term sets and its application in group decision making. Int J, Fuzzy Syst 19:671–682

Carlsson C, Fullér R (2000) Multiobjective linguistic optimization? Fuzzy Sets Syst 115:5–10

Carlsson C, Fullér R (2001) Optimization under fuzzy if then rules? Fuzzy Sets Syst 119:111–120

Chang P, Chen Y (1994) A fuzzy multi-criteria decision making method for technology transfer strategy selection in biotechnology. Fuzzy Sets Syst 63:131–139

Chang TH, Wang TC (2010) A novel efficient approach for DFMEA combining 2-tuple and the OWA operator. Exp Syst Appl 37(3):2362–2370

Chen SM (1997) A new method for tool steel materials selection under fuzzy environment. Fuzzy Sets Syst 92:265–274

Chen CT (2000) Extensions of the TOPSIS for group decision making under fuzzy environment. Fuzzy Sets Syst 114:1–9

Chen Z, Ben-Arieh D (2006) On the fusion of multi-granularity linguistic label sets in group decision making. Comput Ind Eng 51(3):526–541

Chen CT, Cheng HL (2008) Group decision making based on computing with linguistic variables and an example in information system selection. IEEE. https://doi.org/10.1109/CIT.2008.4594716

Chen SJ, Hwang CL (1992) Fuzzy multiple attribute decision-making methods and applications. Springer, Berlin

Chen CT, Tai WS (2005) Measuring the intellectual capital performance based on 2-tuple fuzzy linguistic information. In: Proceedings of the 10th annual meeting of Asia pacific region of decision sciences institute, Taiwan

Cid-López A, Hornos MJ, Carrasco RA, Herrera-Viedma E (2016) Applying a linguistic multi-criteria decision-making model to the analysis of ICT suppliers' offers. Exp Syst Appl 57:127–138

de Andrés R, García-Lapresta JL, Martínez L (2010) A multi-granular linguistic model for management decision-making in performance appraisal. Soft Comput 14(1):21–34

Degani R, Bortolan G (1988) The problem of linguistic approximation in clinical decision making. Int J Approx Reason 2:143–162

Delgado M, Verdegay JL, Vila MA (1992) Linguistic decision making models. Int J Intell Syst 7:479–492

Delgado M, Verdegay JL, Vila MA (1993) On aggregation operations of linguistic labels. Int J Intell Syst 8(3):351–370

Delgado M, Herrera F, Herrera-Viedma E, Martínez L (1998) Combining numerical and linguistic information in group decision making. Inf Sci 107:177–194

Devedzić GB, Pap E (1999) Multicriteria-multistages linguistic evaluation and ranking of machine tools. Fuzzy Sets Syst 102:451–461

Dong Y, Herrera-Viedma E (2014) Consistency-driven automatic methodology to set interval numerical scales of 2-tuple linguistic term sets and its use in the linguistic GDM with preference relation. IEEE Trans Cybern 45(4):780–792

Dong Y, Xu Y, Yu S (2009a) Computing the numerical scale of the linguistic term set for the 2-tuple fuzzy linguistic representation model. IEEE Trans Fuzzy Syst 17(6):1366–1378

Dong Y, Xu Y, Yu S (2009b) Linguistic multi person decision making based on the use of multiple preference relations. Fuzzy Sets Syst 160(5):603–623

Dong Y, Zhang G, Hong WC, Yu S (2013) Linguistic computational model based on 2-tuples and intervals. IEEE Trans Fuzzy Syst 21(6):1006–1018

Dong Y, Li CC, Xu Y, Gu X (2015a) Consensus-based group decision making under multi-granular unbalanced 2-tuple linguistic preference relations. Group Decis Negot 24(2):217–242

Dong Y, Wu Y, Zhang H, Zhang G (2015b) Multi-granular unbalanced linguistic distribution assessments with interval symbolic proportions. Knowl Based Syst 82:139–151

Dong Y, Li CC, Herrera F (2016) Connecting the linguistic hierarchy and the numerical scale for the 2-tuple linguistic model and its use to deal with hesitant unbalanced linguistic information. Inf Sci 367:259–278

Doukas H, Karakosta C, Psarras J (2010) Computing with words to assess the sustainability of renewable energy options. Exp Syst Appl 37(7):5491–5497

Dubois D, Prade H (1980) Fuzzy sets and systems: theory and applications. Kluwer Academic, New York

Dursun M, Karsak EE (2010) A fuzzy MCDM approach for personnel selection. Exp Syst Appl 37(6):4324–4330

Eshragh F, Mamdani EH (1979) A general approach to linguistic approximation. Int J Man Mach Stud 11(4):501–519

Espinilla M, Liu J, Martínez L (2013) An extended hierarchical linguistic model for decision-making problems. Comput Intell 27(3):489–512

Feng B, Lai FJ (2014) Multi-attribute group decision making with aspirations: a case study. Omega 44:136–137

Fodor J, Roubens M (1994) Fuzzy preference modelling and multicriteria decision support. Kluwer Academic Publishers, Dordrecht

Guo W, Huynh V (2013) A new version of 2-tuple fuzzy linguistic screening evaluation model in new product development. In: Proceedings of the 2013. IEEE. https://doi.org/10.1109/IEEM.2013.6962396

Guo WT, Huynh VN, Sriboonchitta S (2017) A proportional linguistic distribution based model for multiple attribute decision making under linguistic uncertainty. Ann Oper Res 256(2):305–328

Gupta PK, Muhuri PK (2014) Perceptual computing based performance control mechanism for power efficiency in mobile embedded systems. FUZZIEEE. https://doi.org/10.1109/FUZZ-IEEE.2014.6891865

Gupta PK, Muhuri PK (2017) Multi-objective linguistic optimization: extensions and new directions using 2-tuple fuzzy linguistic representation model. https://doi.org/10.1109/FUZZ-IEEE.2017.8015772

Halouani N, Chabchoub H, Martel JM (2009) PROMETHEE-MD-2T method for project selection. Eur J Oper Res 195(3):841–849
Herrera F, Herrera-Viedma E (2000) Linguistic decision analysis: steps for solving decision problems under linguistic information. Fuzzy Sets Syst 115:67–82
Herrera F, Martínez L (2000a) 2-Tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8:746–752
Herrera F, Martínez L (2000b) An approach for combining linguistic and numerical information based on 2-tuple fuzzy linguistic representation model in decision-making. Int J Uncertain Fuzziness Knowl Based Syst 8(5):539–562
Herrera F, Martínez L (2001a) The 2-tuple linguistic computational model. Advantages of its linguistic description, accuracy and consistency. Int J Uncertain Fuzziness Knowl Based Syst 9:33–48
Herrera F, Martínez L (2001b) A model based on linguistic 2-tuples for dealing with multi-granularity hierarchical linguistic contexts in multi expert decision-making. IEEE Trans Syst Man Cybern B Cybern 31:227–234
Herrera F, Herrera-Viedma E, Verdegay JL (1995) A sequential selection process in group decision making with linguistic assessment. Inf Sci 85:223–239
Herrera F, Herrera-Viedma E, Verdegay JL (1996) Direct approach processes in group decision making using linguistic OWA operators. Fuzzy Sets Syst 79:175–190
Herrera F, Herrera-Viedma E, Martínez L (2000) A fusion approach for managing multi-granularity linguistic terms sets in decision making. Fuzzy Sets Syst 114(1):43–58
Herrera F, López E, Mendaña C, Rodríguez MA (2001) A linguistic decision model for personnel management solved with a linguistic bi-objective genetic algorithm. Fuzzy Sets Syst 118:47–64
Herrera F, Martínez L, Sánchez PJ (2005) Managing non-homogeneous information in group decision making. Eur J Oper Res 166:115–132
Herrera F, Herrera-Viedma E, Martínez L (2008) A fuzzy linguistic methodology to deal with unbalanced linguistic term sets. IEEE Trans Fuzzy Sys 16(2):354–370
Herrera F, Alonso S, Chiclana F, Herrera-Viedma E (2009) Computing with words in decision making: foundations, trends and prospects. Fuzzy Optim Decis Mak 8(4):337–364
Herrera-Viedma E, López-Herrera AG (2007) A model of an information retrieval system with unbalanced fuzzy linguistic information. Int J Intell Syst 22(11):1197–1214
Huynh VN, Nakamori Y (2005) A satisfactory-oriented approach to multi-expert decision-making under linguistic assessments. IEEE Trans Syst Man Cybern SMC 35(2):184–196
Huynh VN, Nakamori Y, Ho TB, Murai T (2006) Multiple-attribute decision making under uncertainty: the evidential reasoning approach revisited. IEEE Trans Syst Man Cybern Part A Syst Hum 36(4):804–822
Hwang CL, Yoon K (1981) Multiple attribute decision making. Springer, Berlin
Jiang YP, Fan ZP, Ma J (2008) A method for group decision making with multi-granularity linguistic assessment information. Inf Sci 178(4):1098–1109
Ju Y (2014) A new method for multiple criteria group decision making with incomplete weight information under linguistic environment. Appl Math Model 38:5256–5268
Kacprzyk J, Zadrozny S (2010) Computing with words is an implementable paradigm: fuzzy queries, linguistic data summaries, and natural language generation. IEEE Trans Fuzzy Syst 18(3):461–472
Kim K, Park KS (1990) Ranking fuzzy numbers with index of optimism. Fuzzy Sets Syst 35:143–150
Klir GJ, Yuan B (1995) Fuzzy sets an fuzzy logic: theory and applications. Prentice-Hall PTR, Upper Saddle River
Lan J, Chen Y, Ning M, Wang Z (2015) A new linguistic aggregation operator and its application to multiple attribute decision making. Oper Res Perspect 2:156–164
Law CK (1996) Using fuzzy numbers in educational grading system. Fuzzy Sets Syst 83:311–323
Lawry J (2001a) An alternative approach to computing with words. Int J Uncertain Fuzziness Knowl Based Syst 9:3–16
Lawry J (2001b) A methodology for computing with words. Int J Approx Reason 28:51–89
Lawry J (2004) A framework for linguistic modelling. Artif Intell 155(1–2):1–39
Li DF (2009) Multiattribute group decision making method using extended linguistic variables. Int J Uncertain Fuzziness Knowl Based Syst 17(6):793–806
Li YF, Xie QH (2009) A method of identifying supply chain risk factors. In: Tran D, Zhou SM (eds) World congress on software engineering, vol 4, pp 369–373
Li C, Zeng S, Pan T, Zheng L (2014) A method based on induced aggregation operators and distance measures to multiple attribute decision making under 2-tuple linguistic environment. J Comput Syst Sci 80:1339–1349
Li G, Kou G, Peng Y (2016) A group decision making model for integrating heterogeneous information. IEEE Trans Syst Man Cybern Syst 48(6):982–992
Li CC, Dong Y, Herrera F, Herrera-Viedma E, Martínez L (2017) Personalized individual semantics in computing with words for supporting linguistic group decision making. An application on consensus reaching. Inf Fusion 33:29–40
Li CC, Dong Y, Herrera F (2018a) A consensus model for large-scale linguistic group decision making with a feedback recommendation based on clustered personalized individual semantics and opposing consensus groups. IEEE Trans Fuzzy Syst 27(2):221–233
Li CC, Rodríguez RM, Martínez L, Dong Y, Herrera F (2018b) Personalized individual semantics based on consistency in hesitant linguistic group decision making with comparative linguistic expressions. Knowl Based Syst 145:156–165
Liang Y, Tu Y, Ju Y, Shen W (2019) A multi-granularity proportional hesitant fuzzy linguistic TODIM method and its application to emergency decision making. Int J Disaster Risk Reduct 36:101081. https://doi.org/10.1016/j.ijdrr.2019.101081
Lin J, Lan JB, Lin YH (2009) Multi-attribute group decision-making method based on the aggregation operators of interval 2-tuple linguistic information. J Jilin Normal Univ 1:5–9
Liu P, Chen SM (2018) Multi attribute group decision making based on intuitionistic 2-tuple linguistic information. Inf Sci 430:599–619
Liu P, Wang P (2017) Some improved linguistic intuitionistic fuzzy aggregation operators and their applications to multiple-attribute decision making. Int J Inf Technol Decis Mak 16(3):817–850
Liu CM, Wang MJ, Pang YS (1994) A multiple criteria linguistic decision model (MCLDM) for human decision making. Eur J Oper Res 76:466–485
Liu J, Ruan D, Carchon R (2002) Synthesis and evaluation analysis of the indicator. Int J Appl Math Comput Sci 12(3):449–462
Liu Y, Xu J, Nie W (2011) Assessment of capacity of flood disaster prevention and reduction with 2-tuple linguistic information. J Converg Inf Technol 6(7):268–273
Liu S, Chan FT, Ran W (2013) Multi-attribute group decision-making with multi-granularity linguistic assessment information: an improved approach based on deviation and TOPSIS. Appl Math Model 37(24):10129–10140
Liu W, Dong Y, Chiclana F, Javier F, Enrique C (2017) Group decision-making based on heterogeneous preference relations with self-confidence. Fuzzy Optim Decis Mak 16(4):429–447
Mao XB, Wu M, Dong JY, Wan SP, Jin Z (2019) A new method for probabilistic linguistic multi-attribute group decision making:

application to the selection of financial technologies. Appl Soft Comput 77:155–175
Mardani A, Jusoh A, Zavadskas EK (2015) Fuzzy multiple criteria decision making techniques and applications- Two decades review from 1994 to 2014. Exp Syst Appl 42(8):4126–4148
Martin O, Klir GJ (2006) On the problem of re-translation in computing with perceptions. Int J Gen Syst 35(6):655–674
Martínez L (2007) Sensory evaluation based on linguistic decision analysis. Int J Approx Reason 44(2):148–164
Martínez L, Herrera F (2012) An overview on the 2-tuple linguistic model for computing with words in decision making: extensions, applications and challenges. Inf Sci 207:1–18
Martínez L, Liu J, Yang JB, Herrera F (2005) A multi-granular hierarchical linguistic model for design evaluation based on safety and cost analysis. Int J Intell Syst 20:1161–1194
Martínez L, Liu J, Yang JB (2006) A fuzzy model for design evaluation based on multiple-criteria analysis in engineering systems. Int J Uncertain Fuzziness Knowl Based Syst 14:317–336
Martínez L, Liu J, Ruan D, Yang JB (2007a) Dealing with heterogeneous information in engineering evaluation processes. Inf Sci 177:1533–1542
Martínez L, Pérez LG, Barranco MJ (2007b) A multi-granular linguistic based-content recommendation model. Int J Intell Syst 22(5):419–434
Martínez L, Barranco MJ, Pérez LG, Espinilla M (2008a) A knowledge based recommender system with multigranular linguistic information. Int J Comput Intell Syst 1(3):225–236
Martínez L, Espinilla M, Perez LG (2008b) A linguistic multigranular sensory evaluation model for olive oil. Int J Comput Intell Syst 1(2):148–158
Martínez L, Espinilla M, Liu J, Pérez LG, Sánchez PJ (2009) An evaluation model with unbalanced linguistic information:applied to olive oil sensory evaluation. J Multi Valued Logic Soft Comput 15(2–3):229–251
Massanet S, Riera J, Torrens J, Herrera-Viedma E (2016) A model based on subjective linguistic preference relations for group decision making problems. Inf Sci 355–356:249–264
Morente-Molinera J, Pérez I, Ureña M, Herrera-Viedma E (2015) On multi-granular fuzzy linguistic modeling in group decision making problems: a systematic review and future trends. Knowl Based Syst 74:49–60
Morente-molinera JA, Kou G, González-crespo R, Corchado JM (2017) Solving multi-criteria group decision making problems under environments with a high number of alternatives using fuzzy ontologies and multi-granular linguistic modelling methods. Knowl Based Syst 137:54–64
Muhuri PK, Gupta PK, Mendel J (2017) User satisfaction-aware power management in mobile devices based on perceptual computing. IEEE Trans Fuzzy Syst 26(4):2311–2323. https://doi.org/10.1109/TFUZZ.2017.2773020
Nie RX, Wang JQ, Li L (2017) A shareholder voting method for proxy advisory firm selection based on 2-tuple linguistic picture preference relation. Appl Soft Comput 60:520–539
Orduna R, Jurio A, Paternain D, Bustince H, Pinto PM, Barrenechea E (2014) Segmentation of color images using a linguistic 2-tuples model. Inf Sci 258:339–352
Park Y (1997) Application of a fuzzy linguistic approach to analyse asian airports competitiveness. Transp Plan Technol 20:291–309
Pedrycz W, Ekel P, Parreiras R (2010) Fuzzy multicriteria decision-making: models. methods and applications. Wiley, Chichester
Pei Z, Shi P (2011) Fuzzy risk analysis based on linguistic aggregation operators. Int J Innov Comput Inf Control 7(12):7105–7117
Pei Z, Zheng L (2017) New unbalanced linguistic scale sets: the linguistic information representations and applications. Comput Ind Eng 105:377–390
Porcel C, Herrera-Viedma E (2010) Dealing with incomplete information in a fuzzy linguistic recommender system to disseminate information in university digital libraries. Knowl Based Syst 23(1):32–39
Porcel C, Tejeda-Lorente A, Martínez MA, Herrera-Viedma E (2012) A hybrid recommender system for the selective dissemination of research resources in a technology transfer office. Inf Sci 184(1):1–19
Qi K, Wang Q, Duan Q, Gong L, Sun J, Liew KM, Jiang L (2018) A multi criteria comprehensive evaluation approach for emergency response capacity with interval 2-tuple linguistic information. Appl Soft Comput 72:419–441
Qiang WJ, Peng L, Yu ZH, hong CX (2014) Method of multi-criteria group decision-making based on cloud aggregation operators with linguistic information. Inf Sci 274:177–191
Rodríguez RM, Martínez L, Ruan D, Liu J (2010a) Using collaborative filtering for dealing with missing values in nuclear safeguards evaluation. Int J Uncertain Fuzziness Knowl Based Syst 18(4):431–449
Rodríguez RM, Espinilla M, Sánchez PJ, Martínez L (2010b) Using linguistic incomplete preference relations to cold start recommendations. Internet Res 20(3):296–315
Rodríguez RM, Martínez L, Herrera F (2011) Hesitant fuzzy linguistic term sets for decision making. IEEE Trans Fuzzy Syst 20(1):109–119
Saaty TL (1980) The analytic hierarchy process. McGraw-Hill, New York
Singh A, Gupta A (2018) Matrix games with interval-valued 2-tuple linguistic information. Games 9(3):62
Singh A, Gupta A, Mehra A (2018) Matrix games with 2-tuple linguistic information. Ann Oper Res 287:1–16
Sohaib O, Naderpour M, Hussain W, Martínez L (2019) Cloud computing model selection for e-commerce enterprises using a new 2-tuple fuzzy linguistic decision-making method. Comput Ind Eng 132:47–58
Sun YH, Ma J, Fan ZP, Wang J (2008) A group decision support approach to evaluate experts for R and D project selection. IEEE Trans Eng Manag 55(1):158–170
Tian Z, Wang J, Zhang H, Wang T (2018) Signed distance-based consensus in multi-criteria group decision-making with multi-granular hesitant unbalanced linguistic information. Comput Ind Eng 124:125–138
Tong RM, Bonissone PP (1980) A linguistic approach to decision making with fuzzy sets. IEEE Trans Syst Man Cybern 10(11):716–723
Umano M, Hatono I, Tamura H (1998) Linguistic labels for expressing fuzzy preference relations in fuzzy group decision making. IEEE Trans Syst Man Cybern Part B Cybern 28:205–218
Wan SP, Xu GL, Dong JY (2017) Supplier selection using ANP and ELECTRE II in interval 2-tuple linguistic environment. Inf Sci 385:19–38
Wang PP (2001) Computing with words. Wiley series on intelligent systems. Wiley, New York
Wang SY (2008) Applying 2-tuple multigranularity linguistic variables to determine the supply performance in dynamic environment based on product-oriented strategy. IEEE Trans Fuzzy Syst 16(1):29–39
Wang JH, Hao J (2006) A new version of 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 14(3):435–445
Wang JH, Hao J (2007) An approach to computing with words based on canonical characteristic values of linguistic labels. IEEE Trans Fuzzy Syst 15(4):593–604
Wei GW, Zhao XF, Lin R, Wang HJ (2013) Uncertain linguistic Bonferroni mean operators and their application to multiple attribute decision making. Appl Math Model 37(7):5277–5285

Wen TC, Chang KH, Lai HH (2020) Integrating the 2-tuple linguistic representation and soft set to solve supplier selection problems with incomplete information. Eng Appl Artif Intell 87:103–248
Wenstop F (1980) Quantitative analysis with linguistic values. Fuzzy Sets Syst 4:99–115
Wu D, Mendel JM (2010a) Social judgment advisor: an application of the perceptual computer. FUZZ-IEEE, pp 1–8
Wu D, Mendel JM (2010b) Computing with words for hierarchical decision making applied to evaluating a weapon system. IEEE Trans Fuzzy Syst 18:441–460
Wu Z, Xu J (2015) Possibility distribution-based approach for MAGDM with hesitant fuzzy linguistic information. IEEE Trans Cybern 46(3):694–705
Wu Y, Chen K, Zeng B, Yang M, Li L, Zhang H (2017) A cloud decision framework in pure 2-tuple linguistic setting and its application for low-speed wind farm site selection. J Clean Prod 142:2154–2165
Wu Y, Li CC, Chen X, Dong Y (2018a) Group decision making based on linguistic distributions and hesitant assessments: maximizing the support degree with an accuracy constraint. Inf Fusion 41:151–160
Wu H, Xu Z, Ren P, Liao H (2018b) Hesitant fuzzy linguistic projection model to multi-criteria decision making for hospital decision support systems. Comput Ind Eng 115:449–458
Wu Y, Dong Y, Qin J, Pedrycz W (2019) Flexible linguistic expressions and consensus reaching with accurate constraints in group decision-making. IEEE Trans Cybern. https://doi.org/10.1109/TCYB.2019.2906318 (**in press**)
Wu Y, Dong Y, Qin J, Pedrycz W (2020) Linguistic distribution and priority-based approximation to linguistic preference relations with flexible linguistic expressions in decision making. In: IEEE transactions on cybernetics
Xu ZS (2004a) A method based on linguistic aggregation operators for group decision making with linguistic preference relations. Inf Sci 166(1–4):19–30
Xu ZS (2004b) Uncertain multiple attribute decision making: methods and applications. Tsinghua University Press, Beijing
Xu Z (2007) A method for multiple attribute decision making with incomplete weight information in linguistic setting. Knowl Based Syst 20(8):719–725
Xu DL (2012) An introduction and survey of the evidential reasoning approach for multiple criteria decision analysis. Ann Oper Res 195(1):163–187
Xu YJ, Merigo JM, Wang HM (2012) Linguistic power aggregation operators and their application to multiple attribute group decision making. Appl Math Model 36(11):5427–5444
Yager RR (1981) A new methodology for ordinal multiobjective decisions based on fuzzy sets. Decis Sci 12:589–600
Yager RR (1993a) Fuzzy screening systems. In: Lowen R (ed) Fuzzy logic: state of the art. Kluwer Academic Publishers, Dordrecht, pp 251–261
Yager RR (1993b) Non-numeric multi-criteria multi-person decision making. Group Decis Negot 2(1):81–93
Yager RR (1995) An approach to ordinal decision making. Int J Approx Reason 12:237–261
Yager RR (1999) Computing with words and information/intelligent systems 2: applications, chapter approximate reasoning as a basis for computing with words. Physica Verlag, Heidelberg, pp 50–77
Yager RR, Kacprzyk J (1997) The ordered weighted averaging operators: theory and applications. Kluwer, Boston
Yager RR, Goldstein LS, Mendels E (1994) FUZMAR: an approach to aggregating market research data based on fuzzy reasoning. Fuzzy Sets Syst 68:1–11
Yan HB, Huynh VN, Nakamori Y (2011) A probabilistic model for linguistic multi-expert decision making involving semantic overlapping. Exp Syst Appl 38(7):8901–8912
Yeh DY, Cheng CH, Chi ML (2007) A modified two-tuple FLC model for evaluating the performance of SCM: by the six sigma DMAIC process. Appl Soft Comput 7(3):1027–1034
You X, You J, Liu H, Zhen L (2015) Group multi-criteria supplier selection using an extended vikor method with interval 2-tuple linguistic information. Exp Syst Appl 42(4):1906–1916
Yu L (2009) Method for risk evaluation of high-technology with 2-tuple linguistic information. In: Third international symposium on intelligent information technology application, pp 261–264
Yu SM, Wang J, Wang JQ, Li L (2018) A multi-criteria decision-making model for hotel selection with linguistic distribution assessments. Appl Soft Comput 67:741–755
Zadeh LA (1965) Fuzzy sets. Inf. Control 8:338–353
Zadeh LA (1968) Fuzzy algorithms. Inf Control 12:94–102
Zadeh LA (1971) Quantitative fuzzy semantics. Inf Sci 3:159–176
Zadeh LA (1972a) A fuzzy-set-theoretic interpretation of linguistic hedges. J Cybern 2:4–34
Zadeh LA (1972b) A new approach to system analysis. In: Marois M (ed) Man and computer. North-Holland, Amsterdam, pp 55–94
Zadeh LA (1973) Outline of a new approach to the analysis of complex systems and decision processes. IEEE Trans Syst Man Cybern SMC 3:28–44
Zadeh LA (1975a) Fuzzy logic and approximate reasoning (in Memory of Grigore Moisil). Synthese 30(3):407–428
Zadeh LA (1975b) The concept of a linguistic variable and its application to approximate reasoning. Part III. Inf Sci 8(4):301–357
Zadeh LA (1975c) The concept of a linguistic variable and its application to approximate reasoning. Part I. Inf Sci 8(3):199–249
Zadeh LA (1975d) The concept of a linguistic variable and its application to approximate reasoning. Part II. Inf Sci 8(4):301–357
Zadeh LA (1976) The linguistic approach and its application to decision analysis. Springer, Berlin. https://doi.org/10.1007/978-1-4684-2259-7-20
Zadeh LA (1996) Fuzzy logic = computing with words. IEEE Trans Fuzzy Syst 94(2):103–111
Zadeh LA, Kacprzyk J (eds) (1999) Computing with words in information/intelligent systems 1 (foundations). Studies in fuzziness and soft computing, vol 33. Springer, Berlin
Zadeh LA, Fu KS, Tanaka K, Shimura M (1975) Fuzzy sets and their application to cognitive and decision processes. Academic Press, New York
Zeeny M (1982) Multiple criteria decision making. McGraw-Hill, New York
Zhang H (2012) The multiattribute group decision making method based on aggregation operators with interval-valued 2-tuple linguistic information. Math Comput Model 56(1–2):27–35
Zhang HM (2013) Some interval-valued 2-tuple linguistic aggregation operators and application in multiattribute group decision making. Appl Math Model 37(6):4269–4282
Zhang Z, Guo C (2012) A method for multi-granularity uncertain linguistic group decision making with incomplete weight information. Knowl Based Syst 26:111–119
Zhang Z, Guo C (2014) An approach to group decision making with heterogeneous incomplete uncertain preference relations. Comput Ind Eng 71:27–36
Zhang G, Dong Y, Xu Y (2014) Consistency and consensus measures for linguistic preference relations based on distribution assessments. Inf Fusion 17:46–55
Zhang Z, Guo C, Martínez L (2016) Managing multigranular linguistic distribution assessments in large-scale multi attribute group decision making. IEEE Trans Syst Man Cybern Syst 47(11):3063–3076
Zhang HM, Wu Y, Gao J, Xu C (2017a) A method for multi-criteria group decision making with 2-tuple linguistic information based on cloud model. Information 8(2):54. https://doi.org/10.3390/info8020054

Zhang H, Dong Y, Chen X (2017b) The 2-rank consensus reaching model in the multigranular linguistic multiple-attribute group decision-making. IEEE Trans Syst Man Cybern Syst 48(12):2080–2094

Zhang L, Wang Y, Zhao X (2018) A new emergency decision support methodology based on multi-source knowledge in 2-tuple linguistic model. Knowl Based Syst 144:77–87

Zhou S, Chang W, Xiong Z (2011) Risk assessment model with 2-tuple temporal linguistic variable. Appl Mech Mater 58–60:2540–2545

Zou L, Pei Z, Karimi HR, Shi P (2012) The unbalanced linguistic aggregation operator in group decision making. In: IEEE international symposium on intelligent control (ISIC). IEEE. https://doi.org/10.1109/ISIC.2012.6398255

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

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Advanced Hybrid Grid with Anti-Theft Alert and Fault Detection

J N Rai
Electrical Engineering Department
Delhi Technological University
Delhi, India
jnrai@dce.ac.in

Naveen Verma
Electrical Engineering Department
Delhi Technological University
Delhi, India
21naveen.verma@gmail.com

Anubhav Gaba
Electrical Engineering Department
Delhi Technological University
Delhi, India
anubhavgaba@gmail.com

Piyush Meena
Electrical Engineering Department
Delhi Technological University
Delhi, India
piyushmeena773@gmail.com

***Abstract*—The utility providers are estimated to lose billions of dollars annually due to non-technical losses, typically energy theft. To alleviate losses due to electricity thefts and smart meter tampering, utility providers are leveraging various techniques implemented in advanced grids to identify the possible defect in the line. This paper intends to integrate the concepts of efficient hybrid power generation, advanced grid, fault detection, and protection systems and electrical pilferage in a power system and utilize them to create an operational miniature of an "Advanced Hybrid Grid" which combines the previously mentioned concepts into a single model and also allows real-time data and status access using advanced technology like wireless radio transmission or Radio Frequency(RF).**

Keywords—advanced hybrid grid, wireless radio transmission, fault detection, protection system, anti-theft alert system, Radio Frequency (RF).

I. INTRODUCTION

The energy requirement is growing at a very high rate all over the globe. The energy generation of electrical energy to fulfill the power needs is mainly met via conventional resources of energy. However, the world is facing an energy crunch due to diminishing fossil fuels along with the problem of climate change due to Greenhouse Gas emission. Thus researchers are contemplating various non-conventional resources like wind, hydro, biomass, and solar for energy production.

Photovoltaic cells convert solar radiation into DC electricity and this energy could be utilized in fulfilling power requirements [1]. Generally, the wind is characterized by both its speed and direction which is however very random, but all these accounts to its distinctive features in fulfilling all these energy requirements. Wind turbines vary from a few kWs to MWs capacity. Here the length of the blades and diameter of the turbine plays a major role in determining capacity [2].

However, while transmitting this energy from generating station to the distribution network, various types of energy losses and faults occur. Above all that all this rising demand for energy has also led to an increase in energy theft. Even as per the Central Authority of India, there is as much loss in power produced which is enough to light up a city like New York in the US for like 2 years due to energy theft and other losses.

Hence in this paper, the hybrid grid is used which is essentially a grid employing two or more modes of electrical power generation together, we have used solar and wind power for the portion generated from renewable power combined with single-phase $230V, 50Hz$ ac supply for representing the conventional source of power.

Also, this advanced grid is a special type of a grid that is employed digitally, intended to act on the information and data furnished by the behavior of all supplying participants with integrated real-time detection and clearance of over-voltage fault and over-temperature (hotspots) faults in the power system. Advanced hybrid grid with anti-theft alert and fault detection intends to introduce and study the concept of a hybrid grid, fault detection, protection systems, and prevention of illegal usage of electrical power, all combined in one particular power system. The advanced hybrid grid can be proven to be an essential constituent in contemporary power systems to supply the required power in abnormal conditions and make sure the continuity of power throughout the day. It is the straight-forward installation integrated with renewable energy sources like solar radiation, wind power modulators, and fashionable communication systems. This following work is concentrated with a selection of the most effective and efficient power source, detection of faults occurred, and pilferage of electrical energy in the power system with real-time data access.

The hybrid grid uses solar and wind energy systems combined with single-phase AC Supply and supplying energy via transmission and distribution networks. Also, the communication systems are employed for real-time data access and transfer of data, and further, the technology of micro-controllers is utilized for detection of faults and the user trying for pilferage of power and hence, also maintain the economic stability of the system. For the complete operation, the system or the model can be subdivided into two major sections; one is the generating section or the sending end were the sources viz. solar, wind, and AC sources are present and customer end that is the receiving section.

In the generating section, we have two types of non-conventional power generating sources, solar and wind and a conventionally generated single-phase AC supply.

In the receiving section, the generated power supply is received via a transmission line integrated with pilferage detection system and fault detection system.

II. SYSTEM CONFIGURATION

A. Solar and Wind Hybrid System

In our miniature, there is a solar panel of fixed on a solar tracker system (uses LDR to track the sunlight and operate at, nearly peak power all the time) which produces DC output and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:46 UTC from IEEE Xplore. Restrictions apply.

is connected to a DC/DC converter to give regulated power supply to the battery [3].

As depicted in Fig. 1, a dynamo is connected to the wind-mill which produces variable DC output. The output of the dynamo is connected to a DC/DC converter to provide a regulated power supply to the charging circuit [4]-[7].

The charging circuit consists of a p-n junction diode to block the reverse flow of current and capacitive filters to produce ripple-free DC output, thus acting as a DC filter [6]. A 16V battery bank is getting charged form the charging circuit.

The output of $16V$ battery is passed through an inverter to convert $16V$ DC to $16V$ AC. $16V$ AC is stepped up by passing through an autotransformer $(16V - 220V)$ to $220V$ AC to lighten up a bulb (load) [6]-[8].

Fig. 1. Block diagram for the wind-PV hybrid system

B. Fault Detection and Protection System

a) For Over-Temperature Fault: There is a step-down $(12 - 0 - 12V)$ which is taking $220V$ AC supply and the voltage is stepping to $12V$ which is further passed through a full-wave rectifier to produce $12V$ DC output. The output is then passed through the voltage regulator (7805) which regulates the voltage, converts it into $5V$ and removes any fluctuations from the voltage so that no device is damaged. The output $5V$ is now supplied to many devices like micro-controller, light-emitting diode, temperature sensor, etc. There is a temperature sensor (LM35) to sense the over-temperature fault. If the temperature is exceeding the predefined value, then it gives analog output voltage which is in direct proportion to temperature. The sensitivity of LM35 is $10\ mV/°C$. If the temperature of any electrical machine exceeds the pre-defined value then the output of the temperature sensor is passed through the cascade amplifier to amplify current and then is passed to micro-controller as a logic 1. The output of micro-controller is passed to relay through Darlington pair, then the relay operates the fan (connected to a DC motor of rating $5.3 - 9.0V, 300rpm$) for cooling down the heated machine and it also gives the logic 1 to RF transmitter which gives the data to the RF receiver that a certain machine is overheated. And when the temperature of the machine becomes lower than the predefined value, then the same process as mentioned before takes place and a signal is sent to the relay to turn the fan off [9].

b) Anti-Theft Alert Systems: When any external tapping or hook-line activity takes place, a signal will be sent to the micro-controller. Micro-controller will immediately cut-off the supply to the sub-station and will send the information to the transmitter module cutting off supply due to power theft. Transmitter module transmits this information to the receiver end through RF communication. The receiver end receivesinformation from transmitter and this information is sent to the control room of the sub-station ,indicated by LED on the module. This LED will inform the control room that an act of power theft has taken place and thus the supply has been cut out [10]-[12].

Fig. 2. Flowchart for over-temperature fault detection

Fig. 3. Flowchart for Anti-Theft alert system

c) For Underground Fault Detection: The circuit for underground fault detection system consists of a $12V$ power supply, $16X2$ Liquid Crystal Display (LCD), Arduino UNO, 3 relays, 3 Light Emitting Diodes (LEDs), DPDT switches and voltage and resistance measurement circuits. Faults are prompted manually in the system by using DPDT fault switches. A total of 12 DPDT fault switches are used with 4 switches in each of the 3 phases -R, Y, and B respectively. DPDT switches have two positions, i.e. fault position (representing fault at the location) and no-fault position (NF). One of the main components of the underground fault detection circuit is ULN2003A, each channel or Darlington pair in ULN2003 is rate $500\ mA$ and can withstand peak current of $600\ mA$. Due to low cable resistance per meter, nearly $0.01\ \Omega/m$, it can measure very low resistance value. The circuit can check the fault for up to a length of $8\ km$, as it can measure $100\ \Omega$ of cable resistance. Starting from the reference point, the circuit is assembled with a set of resistors in series representing the resistance of the underground cable for a particular distance in km, in our case the underground cable is $8\ km$ long with one resistor at every $2\ km$ and four such sets of resistance array are placed which represents the three phases and the neutral. The value of each resistance is $10\ K\Omega$. Each resistor is connected in series with the resistors in the same phase and is connected in parallel with the corresponding resistances of the other phases. This provides an alternate path for the current to flow even when there's a fault in a phase cable and hence maintains the continuity of the supply [13]-[14]. Faults such as short-circuit, symmetrical, unsymmetrical faults can be easily determined by this method, as inductance and capacitance are taken as distributed parameters, and resistance is considered as lumped [15].

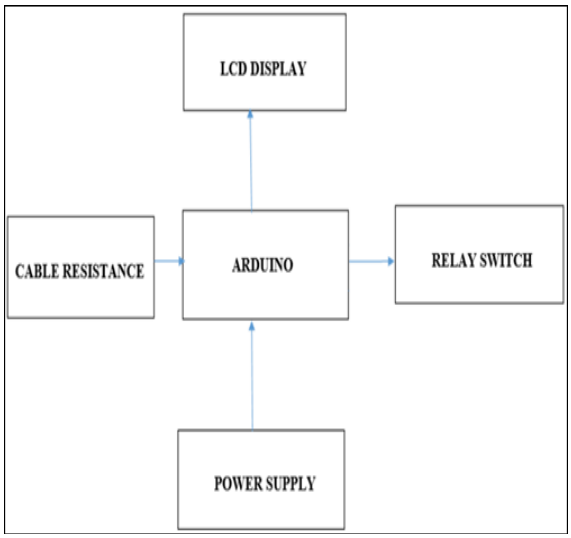


Fig. 4. Block diagram for Underground Fault detection

d) Radio Frequency (RF) Module: This consists of an RF-transmitter and an RF-receiver which operates at a frequency of $434MHz$. The microcontroller transmits the serial data to the RF Transmitter and it transfers the data to the RF receiver wirelessly. The baud rate for transmitting the data is $1 - 10\ Kbps$. The RF transmitter and the RF receiver operate at the same frequency. When logic 0 is sent to the RF Transmitter then it does not transfer any data to the receiver

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:46 UTC from IEEE Xplore. Restrictions apply.

and doesn't consume any power but when logic 1 is sent to the RF Transmitter then the serial data is sent to the RF Receiver, The RF module works along with the encoder and decoder because the data that has to be transmitted has to be converted into analog form and then it is to be converted back in digital form at the receiver end. Here in the circuitry HT12E encoder and HT12D decoder is used [11].



Fig. 5. Transmitter Module

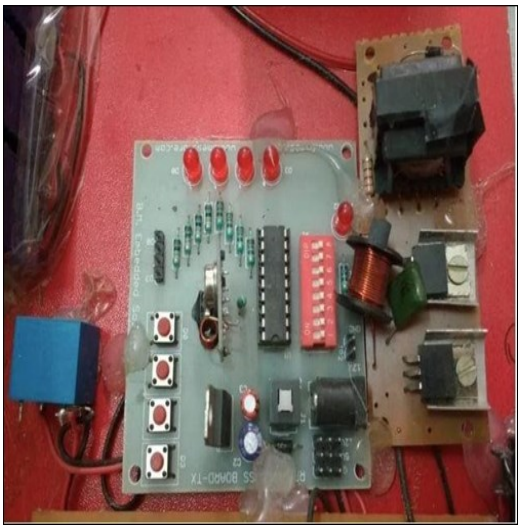


Fig. 6. Receiver Module

III. COMPUTATIONAL DIAGRAMS

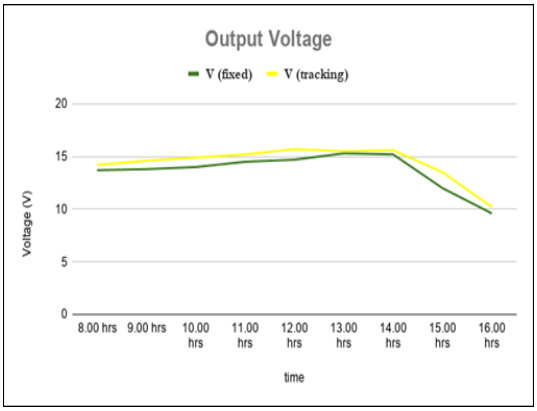


Fig. 7. Comparison of Ouput Voltage between fixed solar panel and tracking solar panel

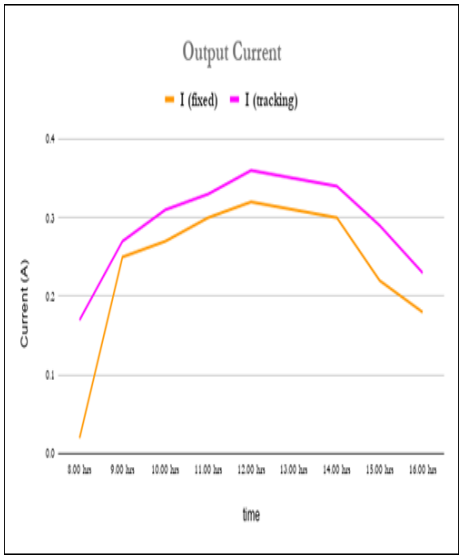


Fig. 8. Comparison of Output Current between fixed solar panel and tracking solar panel

The comparison of output voltage and output current is graphically shown in Fig. 7 and Fig. 8 respectively, in case of fixed and tracking solar panel at different times of the day. The wind turbine generator power vs wind speed curve is shown in Fig. 9 which depicts the cubical relation between wind speed and produced power [16].

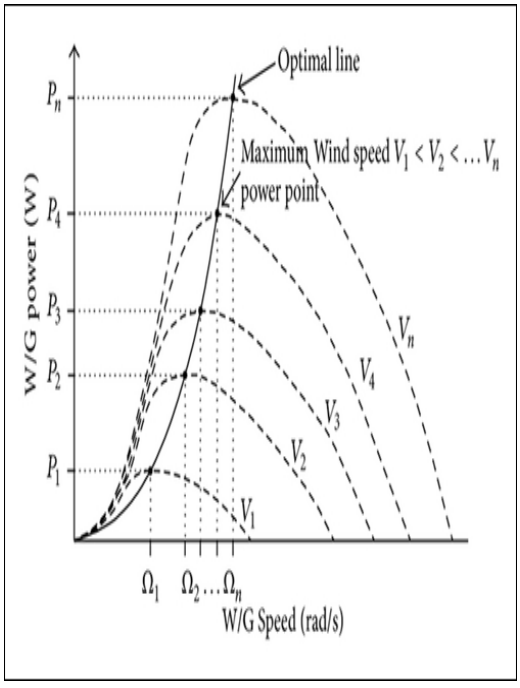


Fig. 9. Wind turbine generator power-wind speed curve

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:46 UTC from IEEE Xplore. Restrictions apply.

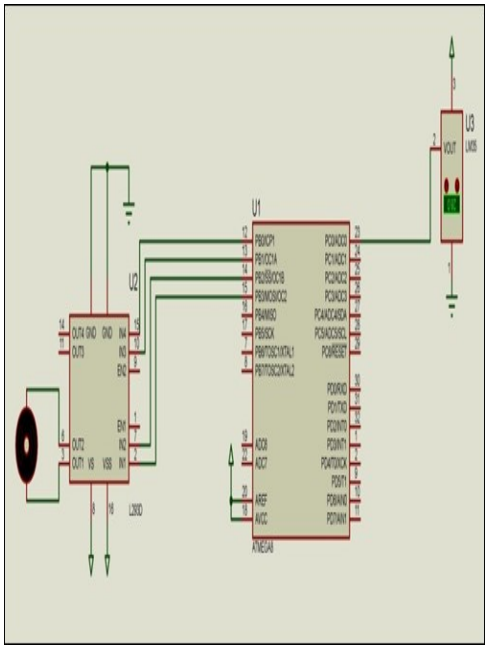


Fig.10. Proteus model of the Over-Temperature Fault Detection System.

The Proteus model of over-temperature fault detection the system is shown in Fig. 10 which shows the connection between LN35 temperature sensor, L239D motor driver and controller to sense the temperature and perform task accordingly, while in Fig. 11 the Proteus model of underground fault detection is shown which depicts the connection and circuit diagram of the same as discussed earlier.

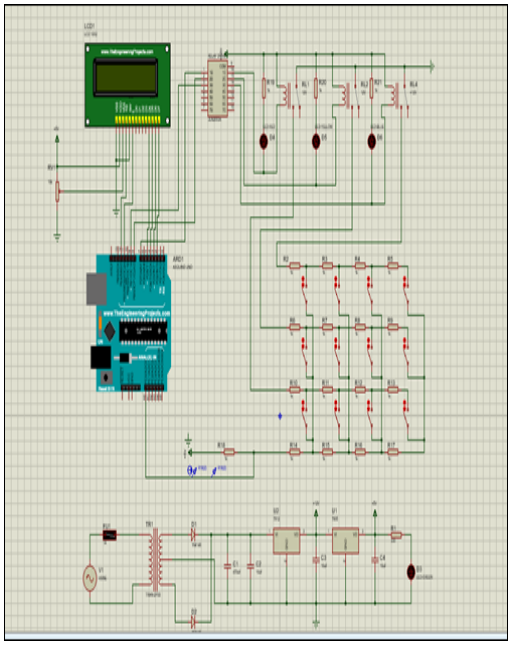


Fig.11. Proteus model of Underground Fault Detection System.

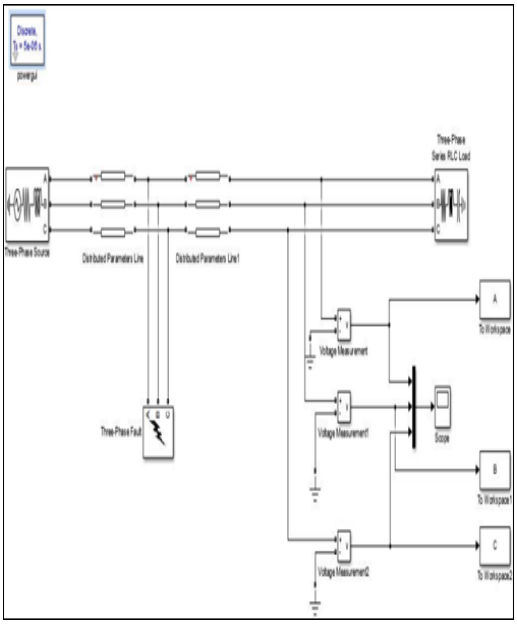


Fig.12. Simulink model for V-t curve of Underground Fault Detection System.

The Simulink model depicted in Fig. 12 shows the basic idea behind an underground fault detection system that is when the fault occurs in any of the phases. the voltage of all phase lines gets disturbed and thus exhibiting the occurrence of fault.

IV. Results and Observations

This section demonstrates the observations made in the project and the results associated with them. In the miniature model of the project, it has been observed that the system was able to execute the following operations:

A. Hybrid Grid Operation

The system was supplied with conventional source supply ($12V$ AC) whilst the non-conventional energy (solar and wind) was being stored in the $16V$ battery, indicated by two LEDs, one for each non.-conventional source. The Azimuth single-axis solar-tracker used with PV panel rotates along the vertical axis ensures that the maximum intensity of sun rays falls on the PV panel and hence maximum solar power is produced during the day. At night time or when there's darkness around the system, the conventional source supply automatically turns off and the stored non-conventional energy is supplied to the system after passing through the inverter, and the bulb (which was used as the load) started glowing.

B. Over-Temperature Fault and Protection

When the temperature of the electrical equipment rises above a predetermined value (38˚C in this case), the cooling fan is turned ON by the micro-controller and RF signal is sent from the transmitting station to the receiving station of the RF Module that an over-temperature fault has occurred in the system, it is indicated at the receiving station by a glowing LED. Once the temperature of the equipment becomes lower than the predetermined value, another signal is sent to the receiving station that the fault has been removed and the LED stops glowing.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:46 UTC from IEEE Xplore. Restrictions apply.

C. *Anti-Theft Operation*

When pilferage of electricity takes place i.e., when there's any external tapping in the line, the power of the whole system is disconnected by the micro-controller and RF signal is sent from the transmitting station to the receiving station of the RF Module that pilferage of electricity is taking place. This turns ON the LED and the alarm at the receiving station. Once the external tapping is removed from the line, another signal is sent to the receiving station thus deactivating the LED and alarm.

D. *Underground Fault Detection*

The supply from a $12V$ battery is stepped down to $5V$ by the Arduino UNO. Then the Arduino sends the supply to the three phases of the underground cable, after adding a phase difference of 120° between the phase voltage and current. Under No-fault condition, the LCD shows NF in all the three phases R, Y and B. When there's a fault(produced manually in the model by using DPDT switches in any of the phases of the underground cable, the distance at which the fault has occurred is displayed along with the phase, in the LCD. The data in the LCD resets to the No-fault values once the fault is removed from the Under-ground cable.

TABLE I. SOLAR PANEL TRACKING

Day-Time	Normal Solar Panel (fixed)		Normal Solar Panel (tracking)	
	Output			
	Voltage(V)	*Current(A)*	*Voltage(V)*	*Current(A)*
8 AM (08:00)	13.70	0.02	14.20	0.17
9 AM (09:00)	13.80	0.25	14.60	0.27
10 AM (10:00)	14.00	0.27	14.90	0.31
11 AM (11:00)	14.50	0.30	15.20	0.33
12 PM (12:00)	14.70	0.32	15.70	0.36
1 PM (13:00)	15.30	0.31	15.50	0.35
2 PM (14:00)	15.20	0.30	15.60	0.34
3 PM (15:00)	12.00	0.22	13.50	0.29
4 PM (16:00)	9.60	0.18	10.20	0.23

The readings of solar panel tracking system are shown in Table.1. which is graphically depicted in Fig. 7 and Fig. 8. It is evident from both these curves that the solar panel is delivering more power when LDR is operational at 1000 W/m2 (ideally) solar irradiance with maximum output voltage observed to be $15.70V$ and maximum output current observed to be $0.36A$, thus showing that solar tracker optimises efficiency of the PV panel throughout day time.

Fig. 9 shows wind turbine generator power vs wind speed curve. It is evident from the curve that with increasing speed of the wind, the produced power also increases and the desirable line shows highest points of power with varied speeds of wind.

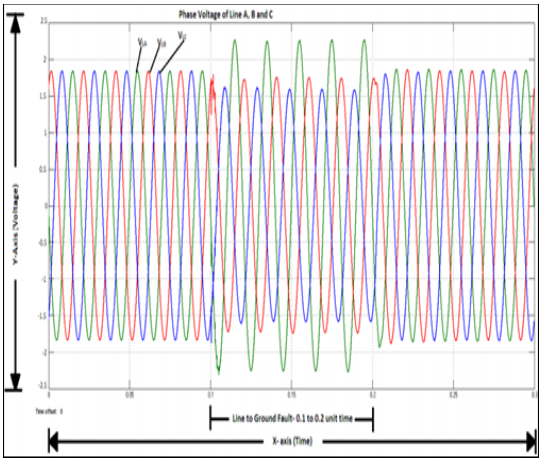


Fig.13. V-t curve of Underground Fault Detection System.



Fig.14. Compact Model Layout.

As shown in Fig. 13 that due to fault in any one phase, the voltage of all three phases is disturbed from the steady-state voltage of the system. In this case, the occurrence of fault was observed at $t = 0.1s$ making voltages of the three phases unstable until the fault is removed later.

After successfully testing all units, they were integrated to form an overall advanced hybrid grid model as shown in Fig. 14, thus giving a compact presentation.

CONCLUSION

This paper designed, implemented a novel integrated architecture of advanced hybrid grid with anti-theft alert and fault detection. We have successfully emulated and tested an advanced hybrid power system. The proposed system is feasible to those areas where solar and wind energies are available at moderate nature. It also promulgates the importance of the growing need of shifting towards renewable sources of power from the conventional fossil fuel-based power generation methods and in their indispensability for a greener and cleaner future and also it was also able to achieve

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:46 UTC from IEEE Xplore. Restrictions apply.

the real-time transfer of data in the form of RF signals using the RF module allowing seamless transfer of data. Also for the future, even better-pioneered techniques can be proposed and concentrated to develop fault detection systems with improved complexity and accuracy.

REFERENCES

[1] Shaikh, Mohd Rizwan & Shaikh, Sirajuddin & Waghmare, Santosh & Labade, Suvarna & Tekale, Anill: "A Review Paper on Electricity Generation from Solar Energy", International Journal for Research in Applied Science and Engineering Technology, 887.10.22214/ijraset.2017.9272, 2017.

[2] Chakraborty, S. & Sinha, Dr. Abhijit & Dutta, Saptarshi & Biswas, Nabarun: "A Review Paper of Wind Energy", Advances & Research in Electrical System Technologies (AREST-11), 2011.

[3] Shivanshu Tiwari, Vivek Kumar Soni: "Single Axis Solar Tracking with LDR", www.ijraset.com, International Journal for Research in Applied Science & Engineering Technology, ISSN: 2321-9653; IC Value: 45.98; SJ Impact.

[4] Ali Mekkaoui, Mohammed Laouer, Younes Mimoun: "Modeling and Simulation for Smart Grid Integration of Solar/Wind Energy " Leonardo Journal of Sciences, ISSN 1583-0233 Issue 30, January-June 2017, p. 31-46.

[5] Ahmed Saidi, Benachaiba Chellali: "Simulation and Control of Solar/Wind Hybrid Renewable Power System", Proceedings of the 6th International Conference on Systems and Control, IEEE, 978-1-5090-3960-9/17, 2017.

[6] Rajesh K, A.D. Kulkarni, T. Ananthapadmanabha: "Modelling and Simulation of Solar PV and DFIG Based Wind Hybrid System", Procedia Technology 21 (2015), 667-675,SMART GRID Technologies, August 6-8,2015.

[7] Giuseppe Marco Tina, Salvina Gagliano "Probabilistic modelling of hybrid solar/wind power system with solar tracking system", www.elsevier.com/locate/renene, RenewableEnergy 36 (2011),1719-1727.

[8] Her-Terng Yau, Chih-Jer Lin, Qin-Cheng Liang: "PSO Based PI Controller Design for a Solar Charger System", The Scientific World Journal 2013(3):815280, May 2013.

[9] Blessed Olalekan Oyebola, Odueso Victor Toluwani: "LM35 Based Digital Room Temperature Meter: A Simple Demonstration", Equatorial Journal of Computational and Theoretical Science, 2 (1): 6-15, www.erjournals.com, ISSN: 0184-7937, 2017

[10] Priyamvada Chandel, Tripta Thakur, B.A. Sawle, Rakhi Sharma: "Power theft: Major cause of non-technical losses in Indian distribution sector", IEEE 7th Power India International Conference (PIICON), 2016.

[11] Mohammed Samad A., Harish Shettikeri, Prof. Rakesh L., Nandish B M:"Wireless Power theft detection using ZIGBEE Technology", www.irjti.org, IJRTI1705043, International Journal for Research Trends and Innovation.

[12] Dr. Thida Aung: "Wireless Remote using RF Module", International Journal of Trend in Scientific Research and Development, Volume 3, Issue 5, August 2019, www.ijtsrd.com, e-ISSN: 2456-6470.

[13] Emmanuel Gbenga Dada, Abdulkadir Hamidu Alkali, Stephen Bassi Joseph, Umar Abba Sanda: " Design and Implementation of Underground Cable Fault Detector", International Journal of Science and Engineering Investigations, Volume 8, Issue87, April 2019, www.IJSEI.com, ISSN: 2251 8843, Paper ID: 88719-07.

[14] T. Nandhini, J. Shalini, T. Sai Sangeeetha D. Gnanaprakasam: "Underground Cable Fault Detection using Arduino", International Journal of Engineering Science and Computing, April 2017.

[15] Sahilbhai Vhora, Jaimin Atodaria, Ashutosh Singh, Hardik Kotak, Ankur K. Gheewala: "Underground Cable Fault Detection Using Matlab", International Journal of Innovative Research in Science, Engineering and Technology, www.ijirset.com, ISSN(Online): 2319-8753, ISSN (Print) : 2347-6710, Vol. 7, Issue 12, December 2018.

[16] Wang, Chia-Nan & Lin, Wen-Chang & Le, Khoa:"Modelling of a PMSG Wind Turbine with Autonomous Control", Mathematical Problems in Engineering. 2014. 1-9. 10.1155/2014/856173.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:46 UTC from IEEE Xplore. Restrictions apply.

Biotechnol Lett
https://doi.org/10.1007/s10529-020-02946-6

ORIGINAL RESEARCH PAPER

Aerobic biodegradation of high explosive hexahydro-1,3,5-trinitro-1,3,5-triazine by *Janibacter cremeus* isolated from contaminated soil

Anchita Kalsi · S. Mary Celin · Jai Gopal Sharma

Received: 28 January 2020 / Accepted: 16 June 2020
© Springer Nature B.V. 2020

Abstract
Objective To evaluate the ability of *Janibacter cremeus* a soil bacterium isolated from explosive contaminated site in degradation of hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) and to study enzyme responsible for degradation.
Results The isolate exhibited 88% degradation of RDX in 30 days of incubation. The biodegradation process followed the first order kinetics. The half- life of RDX was calculated to be 11.088 days. The RDX degradation process was complemented by concomitant release of nitrite ions with 0.78 mol of nitrite released per mole of RDX. The metabolites; Trinitroso- RDX, diamino-RDX, trimino-RDX, bis- (hydroxymethyl) nitramine and methylenedintramine derivative, viz, methylene- *N*- (hydroxy- methyl)-hydroxylamine- *N*-(hydroxymethyl) nitroamine corresponding to the molecular weights 174, 162, 132, 122 and 167 Da respectively were also detected. Nitroreductase enzyme was found to be responsible for RDX degradation.
Conclusion *J. cremeus* could degrade RDX as sole source of nitrogen, via three different pathways wherein, Nitroreductase enzyme was found to play a major role. The efficient degradation of RDX makes *J. cremeus* suitable in treatment of contaminated water and soil at field scale levels.

Keywords Bioremediation · Degradation pathway · Explosives · Nitroreductase enzyme

List of symbols
ln Natural logarithm
k Rate constant

Introduction

Explosives are toxic compounds predominantly found in the military areas. High explosive, Hexahydro-1,3,5- trinitro-1,3,5-triazine (RDX) is one of the most commonly used military explosive. It's a heterocyclic nitramine compound which finds its way into the soil and water generally through improper disposal of manufacturing and process waste, loading, assembly, packaging (LAP) and testing (Oh et al. 2003). RDX has been classified as a potential human carcinogen by United States Environment Protection Agency (USEPA 2011) based on animal studies. Because of

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10529-020-02946-6) contains supplementary material, which is available to authorized users.

A. Kalsi · S. M. Celin (✉)
Centre for Fire Explosives and Environment Safety (CFEES), DRDO, Delhi, India
e-mail: marycelin@cfees.drdo.in

A. Kalsi · J. G. Sharma
Delhi Technological University, Delhi, India

Published online: 22 June 2020

its detrimental effects, the treatment of RDX contaminated waters is of utmost importance. Many physical methods such as adsorption, ozonation, exposure to ultraviolet radiation (UV) and chemical methods such as alkaline/ acid hydrolysis, catalyst mediated and advanced oxidation processes have been developed for the treatment of contaminated waters. But, these approaches are expensive, non-specific and can lead to secondary pollution (Singh et al. 2011a, b). Hence, a need for eco-friendly approach to mitigate the toxic effects of RDX has been paved leading to a shift to bioremediation as a possible treatment technology.

Microbial communities present in the contaminated sites (primarily soil, groundwater, aquifer) exhibit the inherent ability to thrive in presence of the contaminant (Cupples 2013). The use of such bacterial communities for remediation purpose provides an advantage to the approach. Many researchers have isolated explosive degraders from contaminated sites and utilized them for the bioremediation experiments. Arnett and Adrian (2009) isolated *Geobacter sp*., *Acetobacterium sp*., *Desulfovibrio sp.*, from an RDX waste water treatment plant. Roh et al. (2009) also isolated bacteria belonging to *Pseudomonas, Enterobacter* and *Azospirillum* species from contaminated groundwater and aquifer. HAW-EB21 was isolated from the marine sediments of the dumpsite of a former ammunition plant (Zhao et al. 2004). Singh et al. (2011a, b) isolated *Bacillus cereus* strain PU isolated from firing range for biodegradation of 2,4,6- trinitrotoluene. Use of such resistant microbes in the bioremediation of explosive is being explored worldwide.

The earliest work in RDX biodegradation cites the use of anaerobic process (McCormick et al. 1981). But, plethora of work since then have indicated the use of aerobic biodegradation process for the mineralization of RDX (Binks et al. 1995; Coleman et al. 1998; Seth-Smith et al. 2008; Fournier et al. 2002). Seth-Smith et al. (2008) elucidated the pathway responsible for aerobic degradation of RDX by *Rhodococcus rhodochrous* strain 11Y and also identified the enzyme responsible for the same. Kitts et al. (2000) studied the involvement of type I (oxygen insensitive) nitroreductases in degradation of RDX, thereby proving that this enzyme could also be responsible for RDX degradation under aerobic conditions.

Janibacter cremeus was first isolated by Hamada et al. (2013) from the foreshore of an island in Japan. It is one of the ten species under the genus *Janibacter. The other species includes, Janibacter limosus, Janibacter terrae, Janibacter brevis, Janibacter anopheles*, *Janibacter melonis, Janibacter corallicola*, *Janibacter hoylei, Janibacter alkaliphilus*, *Janibacter indicus.* Not much work has been reported so far in respect to degradation of hazardous compounds. Khessairi et al. (2014), used *Janibacter sp*. strain FAS23, in remediation of Pentachlorophenol. *Janibacter cremeus,* on the other hand hasn't been explored for its remediation capacity for any xenobiotics. This research was undertaken to evaluate the native microbe *Janibacter cremeus* in remediation of RDX under aerobic conditions. We, for the first time explore the capability of *Janibacter* sp. in remediation of RDX, thereby paving a path for its future applications in field scale remediation.

Material and methods

Chemicals and media components

RDX (> 99% purity) was provided by an RDX manufacturing facility in India. All other chemicals used throughout this study were of analytical and gradient grade and were obtained from standard manufacturers.

The primary cultures of *Janibacter cremeus* were maintained on Tryptic Soy Agar (TSA) plates. During the aqueous phase study the culture of *J. cremeus* was grown and maintained in Minimal Salt Medium (MSM) (media compositions have been detailed in Online Resource 1).

Microbial culture characterization and growth conditions

The organism *Janibacter cremeus*, was isolated from explosive contaminated soil and the lyophilized culture of the same (MTCC NO. 12869) was obtained from Institute of Microbial Technology, (IMTECH), Chandigarh, India. The culture was revived on TSA plates and were then sub cultured for three generations in MSM. The growth of the culture was monitored by observing its absorbance at 600 nm using Perkin Elmer UV- Visible spectrophotometer (Model Lambda 650S). The third generation culture acclimatised to growth in minimal medium was grown in MSM without any nitrogen source. The cultures in log

phase with an absorbance of 0.8 at 600 nm was used in the degradation experiments. The bacterium was assessed for its growth in different concentrations of RDX (20, 40 and 60 mg/L) in MSM, where RDX served as sole source of nitrogen. The degradation studies were undertaken at the water solubility limit of RDX (60 mg/L).

Aqueous phase RDX degradation experiments

The bacteria, *J. cremeus,* in its log phase was inoculated in the minimal medium (MSM) spiked with 60 mg/L of RDX (sole nitrogen source) under sterile conditions. The bacterium was inoculated at 5% (v/v) in the medium containing RDX. Two control experiments were also set up, viz, MSM (without any nitrogen source) with the bacterial inoculation and MSM with RDX as nitrogen source but without inoculation. The study was conducted for a period of 30 days under sterile conditions on an incubator shaker at 35 °C and 120 rpm. The samples were withdrawn on regular intervals to monitor the degradation capacity of the selected bacteria. The growth of the bacteria was monitored at 600 nm using UV–Visible spectrophotometer.

Monitoring the RDX degradation efficiency

The RDX degradation efficiency of the microbe was monitored by measuring the residual RDX concentrations in the treated samples. The residual RDX concentration was measured by USEPA 8330A (EPA 2006) method using High Performance Liquid Chromatography (HPLC). Briefly, 2 mL aliquot of samples withdrawn were centrifuged to discard the cellular content. The supernatant was then extracted for RDX by addition of equal volume of acetonitrile. The prepared sample was then filtered through 0.45 µm Teflon filters (Millex filters). Analysis of RDX was carried out using Dionex model HPLC (Thermo Fisher make). C18 column, (3 µm, 150 × 4.6 mm) functioned as the stationary phase. The mobile phase was a mixture of acetonitrile and water (50:50 v/v). Separation was performed at a flow rate of 1 mL/min. RDX was detected at 254 nm.

Analysis of metabolites

Nitrite, an important metabolite in the degradation of RDX was analysed spectrophotometrically at 540 nm (Mercimek et al. 2013) briefly described in online resource 2. 1 mL of aliquot was centrifuged at 9500×*g* for 10 min to obtain cell free supernatant. The supernatant was used for the analysis of nitrite.

Mass spectrometric (MS) analysis of the acetonitrile extracted sample was performed using Synapt G2 (Waters ACQUITY QSM) MS system in positive ion mode (ES +). RDX and its degradation products were separated on an Acquity UPLC® BEH column (C18, 1.7 µm) at 0.75 mL/min flow rate over a period of 20 min.

Results and discussion

Janibacter cremeus was tested for its efficiency to degrade RDX in aqueous medium so as to predict its application in further studies for treatment of actual contaminated sites. The bacteria was inoculated in MSM containing RDX as lone nitrogen source and glycerol, glucose and succinate as carbon sources. The studies were carried out at 35 °C at 120 rpm in an orbital incubator shaker.

Growth of *J. cremeus* on RDX

Janibacter cremeus was assessed for its growth in different concentrations of RDX after it was acclimatised to MSM for three generations. Figure 1 compares the growth of the isolate in different concentrations of RDX (20, 40 and 60 mg/L) in MSM. The increasing concentration of RDX lead to a comparative lower growth of the bacterium. Maximum absorbance at 600 nm was recorded at day 5. The cultures at 20 mg/L showed maximum growth of 1.426, followed by 1.327 at 40 mg/L and lastly 1.241 at 60 mg/L concentration. The degradation efficacy of the bacteria was monitored at 60 mg/L of RDX. *J. cremeus* was found to grow well even in the control bottles without any nitrogen source. Figure 2 compares the growth of the bacteria in presence and absence of RDX (60 mg/L). As can be seen, the cells of the bacterium follow a sigmoid growth pattern in both conditions. The bacteria is characterized by a short lag phase of 2 days. The cells reach maximum

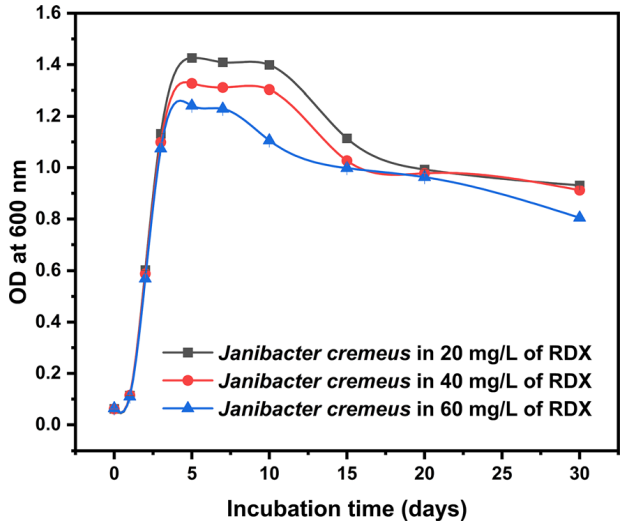


Fig. 1 Comparison of growth of *J. cremeus* in different concentrations of RDX. Error bars denote standard deviation of mean (n = 3) at a given time

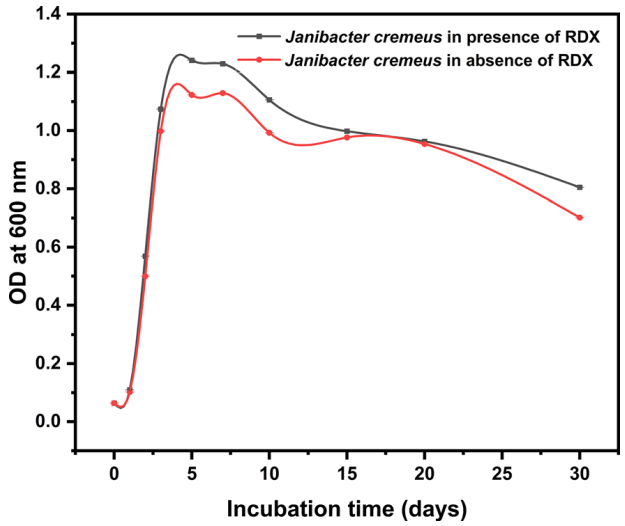


Fig. 2 Comparison of growth of *J. cremeus* in presence and absence of RDX. Error bars denote standard deviation of mean (n = 3) at a given time

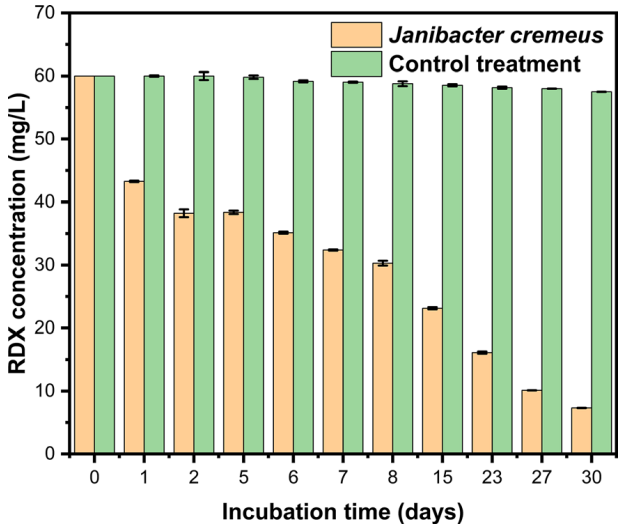


Fig. 3 RDX degradation in treatments with and without *J. cremeus*. Error bars denote standard deviation of mean (n = 3) at a given time

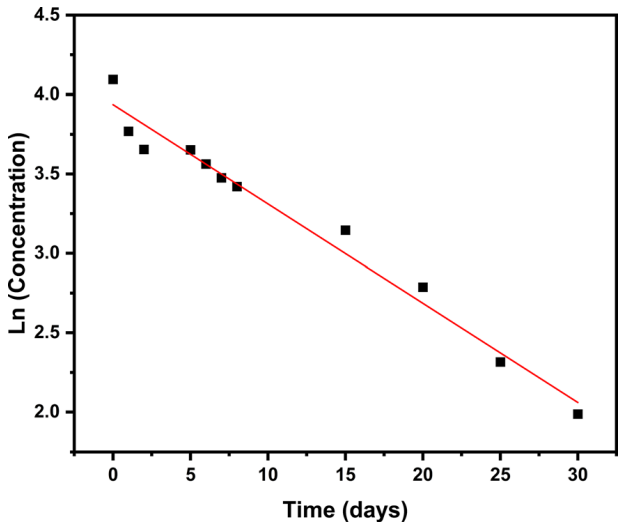


Fig. 4 First- order rate kinetics of RDX degradation by *J. cremeus*

growth of 1.241 and 1.123 for with and without RDX (respectively) on fifth day. The proliferation of *J. cremeus* in presence of RDX proves that, the explosive compound doesn't negatively affect its growth. Growth of the isolate in presence and absence of RDX as well as degradation of the explosive compound later shows that *J. cremeus* utilizes RDX as a non-growth substrate. This proves that, the bacterium degrades RDX via cometabolism (Van Aken et al. 2004; Solyanikova et al. 2011; Nagar et al. 2018).

Degradation of RDX in aqueous phase

RDX degradation efficacy of the bacteria was monitored using HPLC. It was observed that *J. cremeus* could degrade 88% of 60 mg/L of RDX in MSM in 30 days. The controls without any bacterial treatment reported negligible degradation (Fig. 3).

The reaction kinetics for the biodegradation was also calculated to determine the rate of reaction. First order kinetic model as given by Eq. (1), was applied to the residual RDX concentrations (in treatment sets).

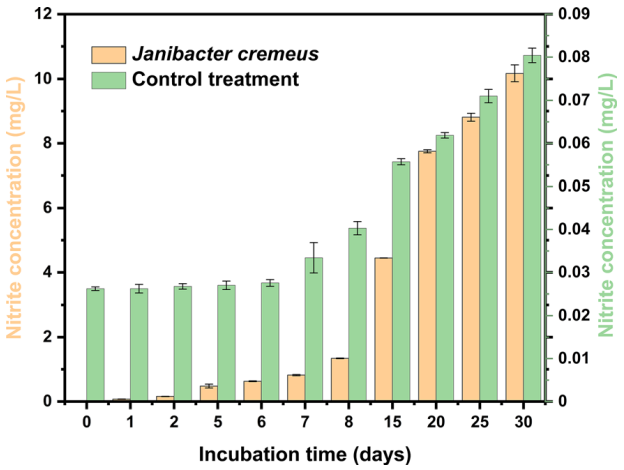


Fig. 5 Nitrite release in treatments with and without *J. cremeus* during RDX degradation. Error bars denote standard deviation of mean (n = 3) at a given time

$$\ln A = -kt + \ln A_0 \quad (1)$$

Equation (1) represents the linear form of the first-order kinetic model, wherein,

A_0 represents initial RDX concentration (mg/L), A is the RDX concentration (mg/L) at time t (days), k is the RDX degradation rate constant. The rate constant for degradation of RDX by *J. cremeus* was obtained from the graph shown in Fig. 4 as 0.0625 day^{-1}. The half-life for RDX degradation was then calculated using the following Eq. (2)

$$t_{\frac{1}{2}} = \frac{0.693}{k} \quad (2)$$

where, $t_{1/2}$ is the half-life of RDX, that is the time required for degradation of RDX by half.

The half-life of RDX on treatment with *J. cremeus* was calculated to be 11.088 days.

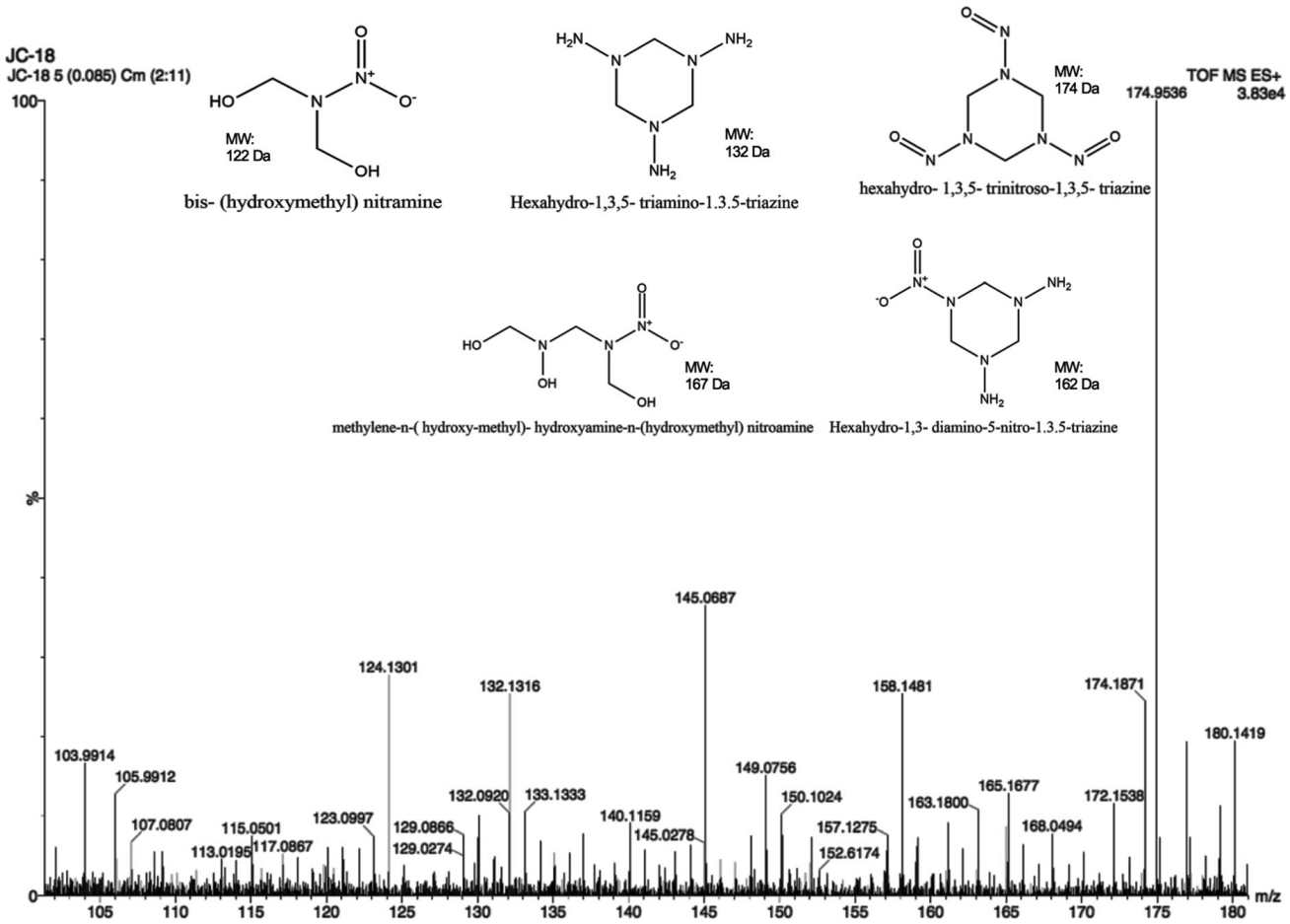


Fig. 6 Mass spectrum showing the intermediates obtained during RDX degradation by *J. cremeus*

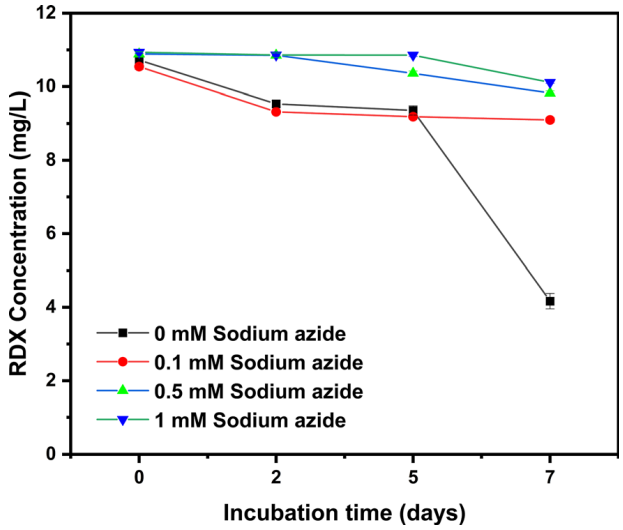


Fig. 7 Effect of increasing concentrations of sodium azide, a specific inhibitor of nitroreductase enzyme on RDX degradation. Error bars denote standard deviation of mean (n = 3) at a given time

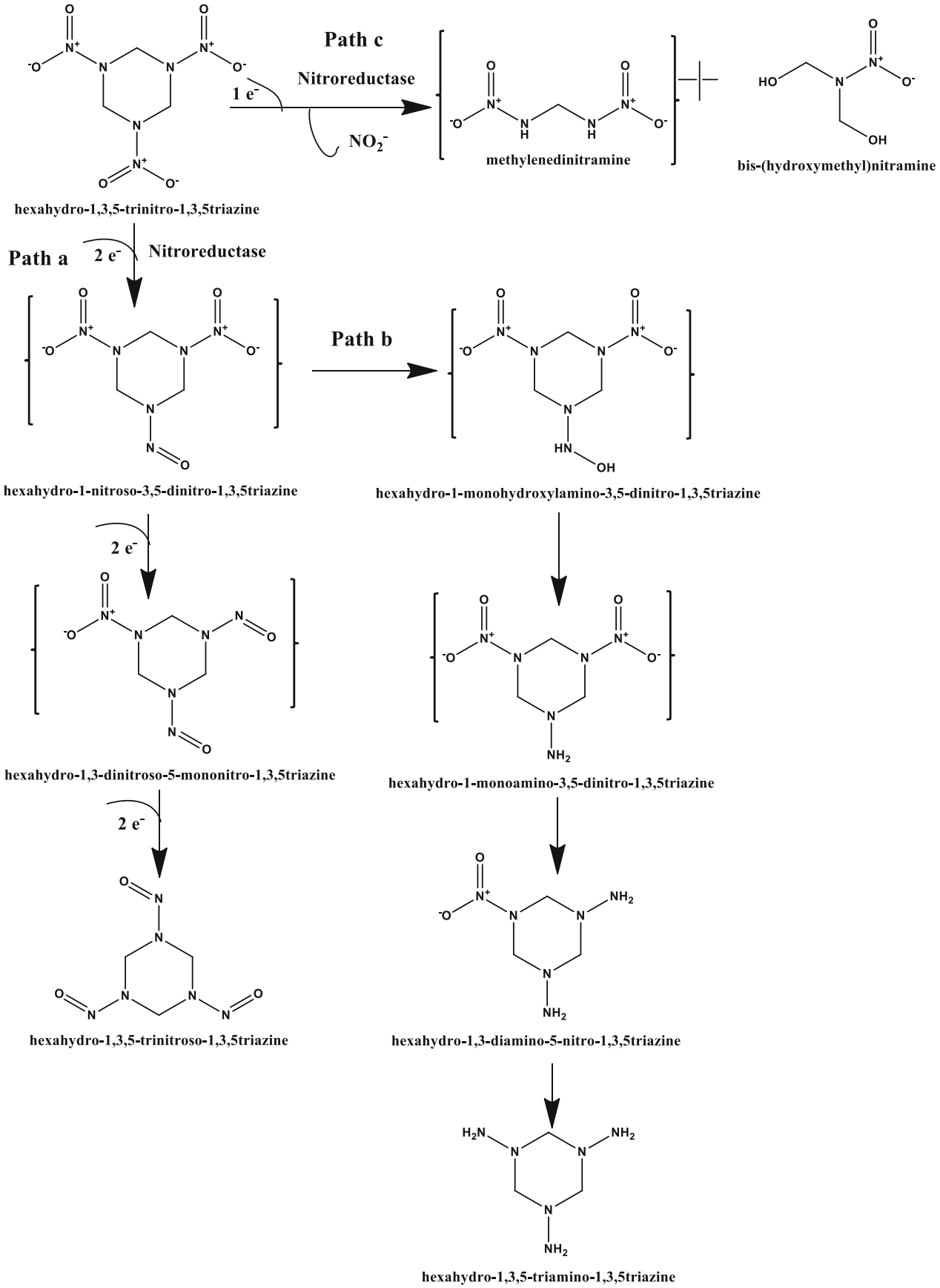
Analysis of metabolites

Nitrite is an important metabolite released during the biodegradation of RDX. It is an important initial step, which involves the removal of nitro groups present in the RDX molecule (Coleman et al. 1998). Figure 5 shows the nitrite released on incubation of RDX with *Janibacter cremeus* and control experiments. As is evident, the control experiments show negligible release of nitrite. The negligible nitrite release could be due to the photodegradation of RDX. Whereas, incubations with *J. cremeus* led to release of 0.78 mol of nitrite for a mole of RDX at the end of 30 days of incubation. The release of nitrite indicates that, RDX degradation in *J. cremeus* follows the single electron transfer/ denitration pathway (Crocker et al. 2006).

To further elucidate the pathway and identify the ring cleavage products of RDX, the samples were subjected to MS, using positive electron spray ionization (ESI). The chromatogram reveals the presence of intermediate with peak at *m/z* of 175 Da (Fig. 6). This fragment indicates the presence of Trinitroso RDX (TNX), which is formed during the double electron reductive pathway as explained by McCormick et al. Though no peaks for Mononitroso and Dinitroso-RDX (MNX and DNX) were observed which could be due to the instability of these metabolites in water (Crocker et al. 2006). But presence of TNX showed that the RDX was transformed to TNX via the mono and dinitroso- intermediates. Bhatt et al. (2006) reported the formation of nitroso derivatives by *Acremonium sp.* HAW-OCF3 under aerobic conditions. Recently, Yadav et al. (2020) also reported the formation of nitroso derivatives during aerobic degradation of RDX using, *Microbacterium esteraromaticum*. Another intermediate metabolite ion peak was obtained at *m/z* of 123 Da, which refers to the presence of bis-(hydroxymethyl) nitramine (BHNA) with a molecular weight of 122 Da. MDNA (methylenedinitramine) derivative, methylene- *N*- (hydroxy- methyl)- hydroxylamine- *N*-(hydroxymethyl) nitroamine, with an empirical formula $C_3H_9N_3O_5$ (molecular weight of 167 Da) was also observed, showing that, MDNA could have been formed by single electron reduction or denitration of RDX, as observed by Binks et al. (1995) in *Stenotrophomonas maltophilia* PB1.The presence of nitrite also proves the denitration step occurring in RDX degradation. Presence of MDNA and BHNA shows that the bacteria also follows the direct enzymatic cleavage of RDX (Hawari et al. 2000). Also, observed were peaks at *m/z* of 163 and 133 Da. These intermediate peaks correspond to presence of Diamino-RDX (162 Da) and Triamino-RDX (132 Da) respectively. A similar RDX degradation pathway was observed by Zhang and Hughes (2003) in *Clostridium acetobutylicum*. The formation of Triamino- RDX leads to no ring cleavage products. Thus the study shows that RDX is degraded by *Janibacter cremeus* by multiple degradation routes.

To observe the role of nitroreductase in RDX degradation, its activity was inhibited using specific inhibitor, sodium azide (Villanueva 1959; Roldán et al. 2008). Sodium azide at different concentrations, viz, 0.1 mM, 0.5 mM, 1 mM, was used to study the effect on RDX degradation in resting cell incubations. *J. cremeus* grown in presence of RDX as sole nitrogen source was harvested in its late log phase and resuspended in RDX at 10 mg/L concentration. The RDX degradation was monitored using HPLC. Figure 7 shows the comparative RDX degradation at different inhibitor concentrations. With no inhibitor, RDX degradation was found to be 61.13% in 7 days. It

Fig. 8 The proposed mechanism for RDX degradation by *J. cremeus*. The intermediates shown in brackets {} were not detected in this study. Path a: McCormick et al. (1981); Path b: Zhang and Hughes (2003); Path c: Hawari et al. (2000)


Path c
Nitroreductase
1 e⁻
NO2⁻
methylenedinitramine
bis-(hydroxymethyl)nitramine
hexahydro-1,3,5-trinitro-1,3,5triazine
Path a
2 e⁻
Nitroreductase
Path b
hexahydro-1-nitroso-3,5-dinitro-1,3,5triazine
hexahydro-1-monohydroxylamino-3,5-dinitro-1,3,5triazine
2 e⁻
hexahydro-1,3-dinitroso-5-mononitro-1,3,5triazine
hexahydro-1-monoamino-3,5-dinitro-1,3,5triazine
2 e⁻
hexahydro-1,3,5-trinitroso-1,3,5triazine
hexahydro-1,3-diamino-5-nitro-1,3,5triazine
hexahydro-1,3,5-triamino-1,3,5triazine

was observed that, with increasing concentration of sodium azide, RDX degradation decreased, showing that, nitroreductase was a key player in RDX degradation pathway in *J. cremeus*. Type I nitroreductases (oxygen insensitive) may be responsible for the two electron transfer process leading to the formation of MNX, DNX and TNX, denitration and ring cleavage leading to the formation of BHNA and MDNA (Kitts et al. 1994, 2000; Young et al. 1997a, 1997b).

On basis of the above results, the proposed degradation pathway has been given in Fig. 8. Due to the substantial release of nitrite, we propose the denitration pathway (leading to the formation of BHNA and MDNA) could be a major degradation pathway for RDX transformation.

Conclusion

RDX is a very commonly used military explosive. It is found to be toxic to human and environmental health. Due to its recalcitrant nature, RDX biodegradation is a major hurdle in eco-friendly treatment of the munition compound. *Janibacter cremeus*, a bacterial isolate from an explosive contaminated soil was evaluated for its ability to degrade RDX aerobically in aqueous medium. This study showed that, *J. cremeus* was a potential degrader, which exhibited 88% degradation in 30 days. The degradation followed the first order with a half-life of 11.088 days. The pathway for degradation was also elucidated. The bacterium followed three different pathways for RDX degradation. It was also observed that, inhibition of Nitroreductase led to the inhibition of degradation, thereby proving that nitroreductase was the major enzyme involved in RDX degradation. The results of the study prove that *J. cremeus* can find further application in site remediation as it could effectively reduce RDX aerobically.

Acknowledgements Authors wish to thank Sh. Rajiv Narang, OS and Director and R.K Tanwar of Centre for Fire Explosive and Environment Safety (CFEES) DRDO for their constant support and encouragement. The authors would also like to thank Dr. S. Mayil Raj and Dr. S. Krishnamurthi, IMTECH, Chandigarh for isolation and identification of culture.

Supporting information Online Resource 1—Media composition.

Online Resource 2—Analysis of nitrite released.

Funding This work was supported by Centre for Fire, Explosive and Environment Safety (CFEES), Defense Research and Development Organization, India.

Compliance with ethical standards

Conflict of interest All the authors declare that they have 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

Arnett CM, Adrian NR (2009) Co-substrate independent mineralization of hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) by a *Desulfovibrio* species under anaerobic conditions. Biodegradation 20:15–26. https://doi.org/10.1007/s10532-008-9195-1

Bhatt M, Zhao JS, Halasz A, Hawari J (2006) Biodegradation of hexahydro-1,3,5-trinitro-1,3,5-triazine by novel fungi isolated from unexploded ordnance contaminated marine sediment. J Ind Microbiol Biot 33(10):850–858. https://doi.org/10.1007/s10295-006-0136-x

Binks PR, Nicklin S, Bruce NC (1995) Degradation of hexahydro- 1,3,5-trinitro-1,3,5-triazine (RDX) by *Stenotrophomonas maltophilia* PB1. Appl Environ Microbiol 61:1318–1322

Coleman NV, Nelson DR, Duxbury T (1998) Aerobic biodegradation of hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) as a nitrogen source by a *Rhodococcus* sp., strain DN22. Soil Biol Biochem 30:1159–1167. https://doi.org/10.1016/S0038-0717(97)00172-7

Crocker FH, Indest KJ, Fredrickson HL (2006) Biodegradation of the cyclic nitramine explosives RDX, HMX, and CL-20. Appl Microbiol Biotechnol 73:274–290. https://doi.org/10.1007/s00253-006-0588-y

Cupples AM (2013) RDX degrading microbial communities and the prediction of microorganisms responsible for RDX bioremediation. Int Biodeter Biodegr 85:260–270. https://doi.org/10.1016/j.ibiod.2013.08.002

Environmental Protection Agency (2006) Method-8330B (SW-846): nitroaromatics, nitramines, and nitrate esters by high performance liquid chromatography (HPLC). U.S, Environmental Protection Agency Washington, DC

Environmental Protection Agency (2011) Edition of the drinking water standards and health advisories. EPA 820-R-11–002. Office of Water U.S. Environmental Protection Agency Washington, DC

Fournier D, Halasz A, Spain JC, Fiurasek P, Hawari J (2002) Determination of key metabolites during biodegradation of hexahydro-1,3,5- trinitro—1,3,5-triazine with *Rhodococcus* sp. strain DN22. Appl Environ Microbiol 68:166–172. https://doi.org/10.1128/AEM.68.1.166-172.2002

Hamada M, Shibata C, Tamura T, Yamamura H, Hayakawa M, Suzuki K (2013) *Janibacter cremeus sp. nov.,* an

actinobacterium isolated from sea sediment. Int J Syst Evol Microbiol 63:3687–3690. https://doi.org/10.1099/ijs.0.051532-0

Hawari J, Halasz A, Sheremata T, Beaudet S, Groom C, Paquet L, Rhofir C, Ampleman G, Thiboutot S (2000) Characterization of metabolites during biodegradation of hexahydro-1,3,5-trinitro- 1,3,5-triazine (RDX) with municipal anaerobic sludge. Appl Environ Microbiol 66:2652–2657. https://doi.org/10.1128/AEM.66.6.2652-2657.2000

Khessairi A, Fhoula I, Jaouani A, Turki Y, Cherif A, Boudabous A, Hassen A, Ouzari H (2014) Pentachlorophenol degradation by *Janibacter* sp., a new actinobacterium isolated from saline sediment of arid land. BioMed Res Int. https://doi.org/10.1155/2014/296472

Kitts CL, Cunningham DP, Unkefer PJ (1994) Isolation of three hexahydro-1,3,5-trinitro-1,3,5-triazine-degrading species of the family *Enterobacteriaceae* from nitramine explosive-contaminated soil. Appl Environ Microbiol 60:4608–4611

Kitts CL, Green CE, Otley RA, Alvarez MA, Unkefer PJ (2000) Type I nitroreductases in soil enterobacteria reduce TNT (2,4,6-trinitrotoluene) and RDX (hexahydro-1,3,5-trinitro-1,3,5-triazine). Can J Microbiol 46:278–282. https://doi.org/10.1139/w99-134

McCormick NG, Cornell JH, Kaplan AM (1981) Biodegradation of hexahydro-1,3,5-trinitro-1,3,5-triazine. Appl Environ Microbiol 42:817–823

Mercimek HA, Dincer S, Guzeldag G, Ozsavli A, Matyar F (2013) Aerobic biodegradation of 2,4,6-Trinitrotoluene (TNT) by *Bacillus cereus* isolated from contaminated soil. Microb Ecol 66:521–521. https://doi.org/10.1007/s00248-013-0248-6

Nagar S, Shaw AK, Anand S, Celin SM, Rai PK (2018) Aerobic biodegradation of HMX by *Planomicrobium flavidum*. 3 Biotech. https://doi.org/10.1007/s13205-018-1479-5

Oh SY, Chiu PC, Kim BJ, Cha DK (2003) Enhancing Fenton oxidation of TNT and RDX through pre-treatment with zero-valent iron. Water Res 37:4275–4283. https://doi.org/10.1016/S0043-1354(03)00343-9

Roh H, Yu CP, Fuller ME, Chu KH (2009) Identification of hexahydro-1,3,5- trinitro-1,3,5-triazine-degrading microorganisms via ^{15}N-stable isotope probing. Environ Sci Technol 43:2505–2511. https://doi.org/10.1021/es802336c

Roldán MD, Pérez-Reinado E, Castillo F, Moreno-Vivián C (2008) Reduction of polynitroaromatic compounds: the bacterial nitroreductases. FEMS Microbiol Rev 32:474–500. https://doi.org/10.1111/j.1574-6976.2008.00107.x

Seth-Smith HMB, Edwards J, Rosser SJ, Rathbone DA, Bruce NC (2008) The explosive-degrading cytochrome P450 system is highly conserved among strains of *Rhodococcus spp*. Appl Environ Microb 74:4550–4552. https://doi.org/10.1128/aem.00391-08

Singh B, Kaur J, Singh K (2011a) 2,4,6-Trinitrophenol degradation by *Bacillus cereus* isolated from a firing range. Biotechnol Lett 33:2411–2415. https://doi.org/10.1007/s10529-011-0726-1

Singh S, Kang SY, Mulchandani A, Chen W (2011b) Bioremediation: environmental clean-up through pathway engineering. Curr Opin Biotech 19:437–444. https://doi.org/10.1016/j.copbio.2008.07.012

Solyanikova IP, Baskunov BP, Baboshin MA, Saralov AI, Golovleva LA (2011) Detoxification of high concentrations of trinitrotoluene by bacteria. Appl Biochem Microbiol 48:21–27. https://doi.org/10.1134/S0003683812010152

Van Aken B, Yoon JM, Schnoor JL (2004) Biodegradation of nitro-substituted explosives 2,4,6-trinitrotoluene, hexahydro-1,3,5-trinitro-1,3,5-triazine, and octahydro-1,3,5,7-tetranitro-1,3,5-tetrazocine by a phytosymbiotic *Methylobacterium sp*. associated with poplar tissues (populus deltoides x nigra DN34). Appl Environ Microb 70(1):508–517. https://doi.org/10.1128/aem.70.1.508-517.2004

Villanueva JR (1959) Interference by azide with diazotization procedures used in biological assay systems. Nature 184:549–549. https://doi.org/10.1038/184549a0

Yadav S, Sharma A, Khan MA, Sharma R, Celin M, Malik A, Sharma S (2020) Enhancing hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine (RDX) remediation through water-dispersible *Microbacterium esteraromaticum* granules. J Environ Manag 264:110446. https://doi.org/10.1016/j.jenvman.2020.110446

Young DM, Kitts CL, Unkefer PJ, Ogden KL (1997a) Biological breakdown of RDX in slurry reactors proceeds with multiple kinetically distinguishable paths. Biotechnol Bioeng 56:258–267

Young DM, Unkefer PJ, Ogden KL (1997b) Biotransformation of hexahydro-1,3,5-trinitro-1,3,5-triazine (RDX) by a prospective consortium and its most effective isolate *Serratia marcescens*. Biotechnol Bioeng 53:515–522

Zhang C, Hughes JB (2003) Biodegradation pathways of hexahydro- 1,3,5-trinitro-1,3,5-triazine (RDX) by *Clostridium acetobutylicum* cell-free extract. Chemosphere 50:665–671. https://doi.org/10.1016/S0045-6535(02)00639-2

Zhao JS, Greer CW, Thiboutot S, Ampleman G, Hawari J (2004) Biodegradation of the nitramine explosives hexahydro-1,3,5-trinitro-1,3,5-triazine and octahydro- 1,3,5,7-tetranitro-1,3,5,7-tetrazocine in cold marine sediment under anaerobic and oligotrophic conditions. Can J Microbiol 50:91–96. https://doi.org/10.1139/w03-112

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

Check for updates

International Journal of ENGINE RESEARCH

Standard Article

International J of Engine Research
1–13
© IMechE 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1468087420928734
journals.sagepub.com/home/jer
SAGE

An experimental investigation on spray, performance and emission of hydrotreated waste cooking oil blends in an agricultural engine

Naveen Kumar[1], Ankit Sonthalia[1,2], Mukul Tomar[1] and Rashi Koul[1,3]

Abstract

Biomass-based fuels are gaining importance for operating a compression ignition engine as they can curb greenhouse gases and are a key for addressing the energy security. Hydrotreated oil is considered to be a potential drop-in fuel for the compression ignition engine as its cetane number is higher than fossil diesel. In this study, hydrotreated waste cooking oil and its blends (10%, 20%, 30%, 40% and 50% by volume) with diesel were prepared. The ignition probability of the test fuel samples was found using a hot-plate test setup. The neat hydrotreated fuel has higher ignition probability at a particular temperature than the other test fuels. The Sauter mean diameter of the test fuels was also observed using a Malvern Spraytec test setup. The results reveal that the neat hydrotreated fuel has higher Sauter mean diameter due to its high viscosity. As the percentage of the hydrotreated fuel in the blend decreases, the Sauter mean diameter decreases and diesel has the lowest Sauter mean diameter. The test fuels were also used to run a compression ignition engine. The results reveal a decrease in brake thermal efficiency with the increase in the hydrotreated fuel share in the blend. The heat release for the blends starts earlier than diesel and the peak heat release is also lower than diesel. The HC, CO and smoke emissions for the test blends decreases up to 30% blend. When the percentage of the hydrotreated oil is further increased, the emissions starts increasing. The NO emissions were lower than diesel for all the test samples. As compared to diesel, the maximum reduction in NO (neat), HC (30% blend), CO (30% blend) and smoke emissions (30% blend) is 23.2%, 14.4%, 13.83% and 13.3%, respectively.

Keywords

Hydrotreated waste cooking oil, diesel engine, Sauter mean diameter, hot-plate test, ignition probability

Date received: 3 March 2020; accepted: 1 May 2020

Introduction

The growth of any economy is fueled by energy. According to International Energy Agency (IEA), the energy requirement of the world will increase by 25% in the year 2040 and the energy requirement of India will be just double of its current requirement.[1] Currently, India is importing crude oil from Middle East countries especially from Iran, Saudi Arabia and Iraq to satisfy its demand. India is a major importer of Iranian oil; however, the recent sanctions on Iran by the US government and slowdown in the production of oil by other countries are threatening the supply of crude oil. Therefore, to keep its economy going at the current rate, India has to look for other sources of energy. Biofuels provide some hope of self-reliance as they are renewable, biodegradable and environment friendly. In the last two decades, a lot of emphasis has been laid on operating the compression ignition (CI) engine with biodiesel, since they are the preferred prime movers for the transport sector and off-road vehicles.[2–4] Diesel-like fuel produced by hydroprocessing the vegetable oil is also a prospective alternate for CI engines. In this process, hydrogen in the presence of a conventional hydrotreating catalyst removes the oxygen atoms and breaks

[1]Center for Advanced Studies and Research in Automotive Engineering, Delhi Technological University, Delhi, India
[2]Department of Automobile Engineering, SRM Institute of Science and Technology, Ghaziabad, India
[3]Department of Mechanical Engineering, Amity University, Gurgaon, India

Corresponding author:
Ankit Sonthalia, Department of Automobile Engineering, SRM Institute of Science and Technology, Delhi-NCR Campus, Ghaziabad 201204, India.
Email: ankit_sont@yahoo.co.in

the double bond from the triglycerides. The physical and chemical properties of the fuel are similar to diesel and it can be stored and transported using the existing infrastructure of diesel.[5]

The rate of rise of pressure and peak pressure in the cylinder is low for the engines fueled with hydrotreated oil in comparison to diesel.[6] This is due to the reduction in the delay of ignition causing smaller premixed fraction to burn, thus lowering the initial rate of combustion. The engine performance of hydrotreated vegetable oil (HVO), iso-hydrotreated oil and biodiesel blends with fossil diesel was compared by Kim et al.[7] At maximum load, power loss occurs with the three fuels, which increases as the blending ratio increases. Among the three fuels, engine operation with iso-hydrotreated oil resulted in lowest power loss, whereas operation with biodiesel resulted in the highest power loss. The consumption of fuel was also lower with hydrotreated oil in comparison to biodiesel blends. Caprotti et al.[8] blended biodiesel (5%) in varying blends of iso-hydrotreated oil and diesel to run a pickup truck. The fuel consumption was found to decrease with the increase in hydrotreated oil in the blend. The field test also shows that the cold start ability and drivability are equivalent to fossil diesel. Moreover, with correct cold flow additives, 20% hydrotreated oil in diesel can be used harm free for severe driving conditions.

The emission characteristics of diesel engines or vehicles fuelled with hydrotreated oil or its blends has been studied by many researchers. Hydrotreated oil, rapeseed methyl ester (RME), jatropha methyl ester (JME) and fossil diesel were compared for their bacterial mutagenicity and emission.[9] JME and RME had high mutagenicity, whereas hydrotreated oil has low mutagenicity. The particulate matter (PM) emission was lowest with JME and RME, while HC and CO emissions were lowest with RME engine operation. The NOx emissions were lower with hydrotreated oil operation in comparison to RME and JME. Singh et al.[10] used European Stationary cycle for studying the emission characteristics of a heavy-duty diesel engine. The NOx emissions were 29% lower with hydrotreated oil than jatropha biodiesel but substantially higher than diesel. The smoke emissions were higher than biodiesel but far lower than the European stationary cycle limit. The HC and CO emissions were also lower. The same group conducted another study using a small diesel engine which finds applications in three wheelers for carrying goods and passenger.[11] The NOx and PM emissions with hydrotreated oil are substantially lower than diesel. The authors also studied the morphology of soot particles using a tunneling electron microscope (TEM). They found that the soot particle size with hydrotreated oil was smaller than that of diesel. The particle size is dependent upon the formation of the particle during the nucleation phase which is affected by the fuels' aromatic content. Since the aromatic content of hydrotreated oil is low, the soot particle size is also smaller. Happonen et al.[12] also studied the soot particles structure and found that the characteristics of soot particles are same with both hydrotreated oil and diesel but the chemical composition is different.

Pflaum et al.[13] operated a common rail diesel engine with neat and blends of hydrotreated oil in diesel. The smoke emissions reduced by up to 50% with HVO at low loads, the particle size also reduced but the particle number was same. The authors also found HC and CO emission reduction with hydrotreated oil. The NOx emission reduced slightly with engine dynamometer test due to reduction in adiabatic flame temperature, whereas with chassis dynamometer, there was no significant variation. Since the soot emission was lower with hydrotreated oil, the vehicle was able to operate for a long duration before the regeneration of diesel particulate filter (DPF) could start. The exhaust temperature should be higher for regenerating the DPF, which is achieved by post injection of the fuel. With the injection of fuel late in the power stroke, there is a possibility that the fuel may hit the cylinder walls and dilute the engine oil. The effect of the hydrotreated oil on DPF was also studied by Kopperoinen et al.[14] The authors found low rate of soot accumulation with neat HVO and the DPF regeneration did not began until the vehicle was operated for at least 400 km. Erkkilä et al.15 while operating a diesel engine with HVO found that the particulate filter is more effective due to favorable (higher) NOx-to-PM ratio. The authors also compared unregulated emissions for three city buses and found 55% reduction in two- and three-ring hydrocarbons with hydrotreated oil, whereas 85% reduction in heavier compounds containing four or more rings was observed.

The literature study shows that the properties, such as cetane number and calorific value of hydrotreated oil, have an influence on the combustion characteristics which can in turn affect the engine exhaust emissions except NOx emissions. Since NOx emissions are also influenced by the engine type (light duty or heavy duty), injection system and timing, combustion chamber design and test cycle.[16] In this work, the hydrotreated oil and its blends with diesel were evaluated by carrying out engine experiments. An agricultural engine which also doubles up as a power backup was utilized in the study. The engine was evaluated for its performance and emission characteristics at different loads. Since it is proposed that hydrotreated oil may be used as a drop-in fuel for supplementing diesel; therefore, the characteristics of such small diesel engines which have a huge population in India is worth studying. The study also characterizes the microscopic spray characteristics such as Sauter mean diameter (SMD) of the fuel blends. A hot-plate test was also carried out to study the probabilistic ignition of the blends.

Table 1. Comparison of the physiochemical properties of neat diesel and the test fuels.

Properties	Standard	Diesel	Hydrotreated waste cooking oil
Density at 15 °C (kg/m^3)	ASTM D1298	835.07	817.7
Kinematic viscosity at 40 °C (cST)	ASTM D445	2.51	2.9
Flash point (°C)	ASTM D93	58	68
Calorific value (MJ/kg)	ASTM D240	42.57	41.79
Cetane index	ASTM D4737	48.78	70.2

Production of hydrotreated oil

In this work, used cooking oil was procured from the cafeteria of Delhi Technological University. The hydrotreating of the oil was carried out at 330 °C reaction temperature and 4 MPa hydrogen pressure. The catalyst used for the reaction was alumina supported 2% ruthenium, cobalt and manganese. The ratio of the catalyst and the used cooking oil was maintained at 2.5%. From the reaction, nearly 89% bio-crude was produced. The obtained liquid was then distilled according to ASTM D2892 standard. The oil obtained in the temperature range of 180 °C–370 °C was considered as the hydrotreated oil. The gas chromatography–mass spectrometry (GC-MS) analysis of the fuel was carried out and it contained carbon atoms in the range of C_{11}–C_{18} with heptadecane as the predominant hydrocarbon.[17] Table 1 compares the property of the hydrotreated oil with fossil diesel.

Experimental setup for hot-plate ignition

For this work, a stainless-steel plate with thickness of 8 mm and diameter of 152 mm was selected. A nichrome wire electrical heater was used to heat the plate. The plate's temperature was measured using K-type thermocouple and its temperature was varied by varying the voltage supply (variac) to the heater. A pipette was mounted firmly on a stand so that the fuel droplet fell on the center of the hot plate. The steel plate was placed directly above the heater and using glass wool, the space between heater and steel plate was covered so that the heat loss to the surroundings was minimum. The steel plate surface was finely polished and the plate had a small concave depression in the center so that the droplet stays in the middle of the plate. The pipette was placed perpendicularly to the plate, and the distance of the pipette nozzle was fixed at 20 mm from the steel plate. The distance was so chosen such that the chance of the flame igniting the fuel in the pipette was minimized and the chance of the drop breakup while falling was also less. Figure 1 shows the schematic of the setup.

For the experiments, a single droplet of fuel was allowed to fall on the plate. Two possible events, namely, ignition of droplet or no ignition of droplet, were recorded. At a particular surface temperature, the fuel was dropped ten times to find that the ignition occurred or not. The temperature range of 280 °C–640 °C in increments of 20 °C was selected for carrying out the experiments. After every experiment, the plate was glazed with 90 grit size sandpaper. The test was performed in a closed room at constant room and fuel temperature.

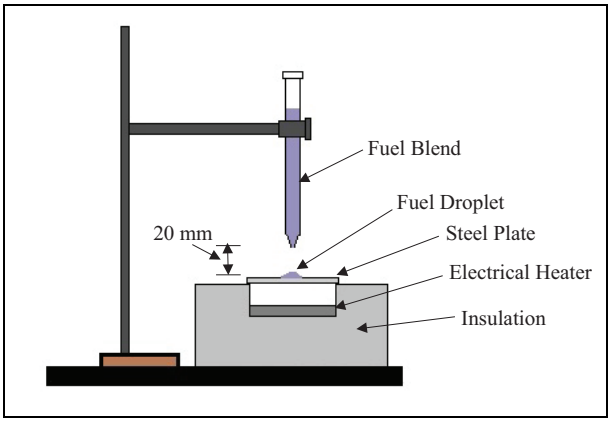


Figure 1. Schematic of hot-plate test setup.

Experimental setup for spray characteristics

The fuel droplet size distribution was measured by a laser diffraction system (Malvern Spraytec). The technical specification of the system is given in Table 2. The system uses Mie theory and Fraunhofer approximation for determining the particle size. The Fraunhofer approximation is used for measurement of particle size of different materials and it does not require any prior knowledge of the optical properties of the material. However, the Fraunhofer approximation gives good results when the particle size is greater than 50 μm. The Mie theory predicts the scattering of the light by spherical particles and deals with the way light passes or is absorbed by the particles. The theory is more accurate but the user should have knowledge of refractive index of the particle.[18]

In this work, a solenoid controlled custom built injector with single orifice and having dimensions similar to that of the injector present in the engine was used for the spray characterization. A high pressure pump supplied the fuel to the injector through a common rail. The injection pressure was fixed at 200 bar for all the experiments. The injection system of the engine used in this study works at this pressure. Figure 2 shows the experimental setup schematic. Five data sets were

Table 2. Specifications of Malvern Spraytec.

Measurement principle	Laser diffraction
Size range	0.1–2000 μm
Optical models	Mie theory and Fraunhofer approximation including patented multiple scattering correction
Lens range	750 mm lens: 2.0–2000 μm (Dv50: 5–1600 μm)
Working range	100 mm at 0.5 μm and > 1 m at 10 μm
Concentration range	Minimum acceptable transmission: 5%
Detection system	35 element log-spaced silicon diode detector array
Light source	632.8 nm, 5 mW helium–neon laser
Maximum acquisition rate	Continuous mode: 1 HzRapid mode: 2.5 kHz as standard, 10 kHz with additional software feature key
Accuracy	Better than ±1% on the Dv50
Precisions/repeatability	Better than ±1% COV on the Dv50

COV: coefficient of variation.

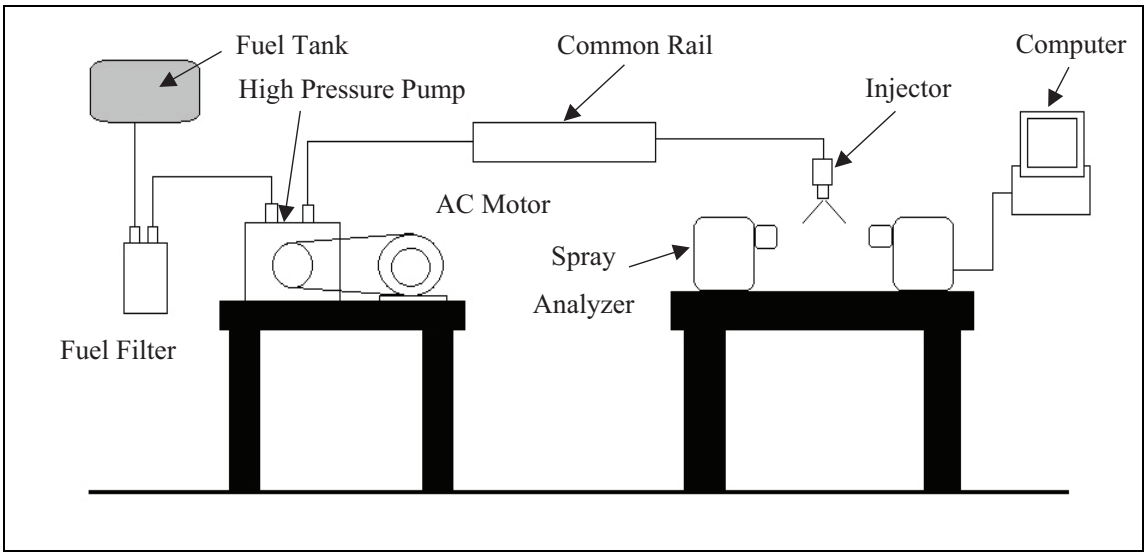


Figure 2. Schematic of the spray analysis setup.

recorded for each test fuel and the values were averaged. The position of the nozzle with the respect to the laser beam plays a major role on the particle size distribution. If the nozzle is far away from the laser beam, the spray plume will expand as a result of which the data of only a small portion of the plume will be captured. The particle size also varies when one moves away from the nozzle. The authors in this study are interested in the particle size distribution inside the engine cylinder. Therefore, the nozzle was placed 25 mm away from the laser beam. The data were collected for 200 ms, although the injection lasted for about 12 ms. The data shown in the study are at 6 ms such that a clear comparison can be made of the droplet distribution in the middle of injection. A low-light room with constant atmospheric conditions was used to carry out the tests and the fuel was injected directly in air. These conditions were chosen, since the primary atomization of the fuel needs to be studied. The primary atomization is dependent upon the geometry of the injector, the fuel properties, and the density of air in which the fuel is sprayed. Since the injector geometry and air density are nearly constant, the fuel properties will play a major role in the fuel spray characteristics. The procedure for data collection was similar to Daho et al.[19]

Experimental setup for engine trial

In this work, a single cylinder diesel engine was used to study the performance and emission of the test fuels. For loading the engine, it was coupled to an eddy current based water-cooled dynamometer. In addition to this, test setup is equipped with electronic data acquisition system to measure the engine parameters such as applied load, fuel flow and temperature value at different points. Figure 3 shows the test-rig schematic and Table 3 tabulates the engine specifications. The fuel flow rate was measured using a differential pressure transmitter (Yokogawa, Japan make) which works on the hydrostatic head principle. The transmitter was placed at the end of the burette and it produced an output voltage which is proportional to the weight of the fuel in the column. The fuel flow rate was calculated by measuring the difference in transmitter output at

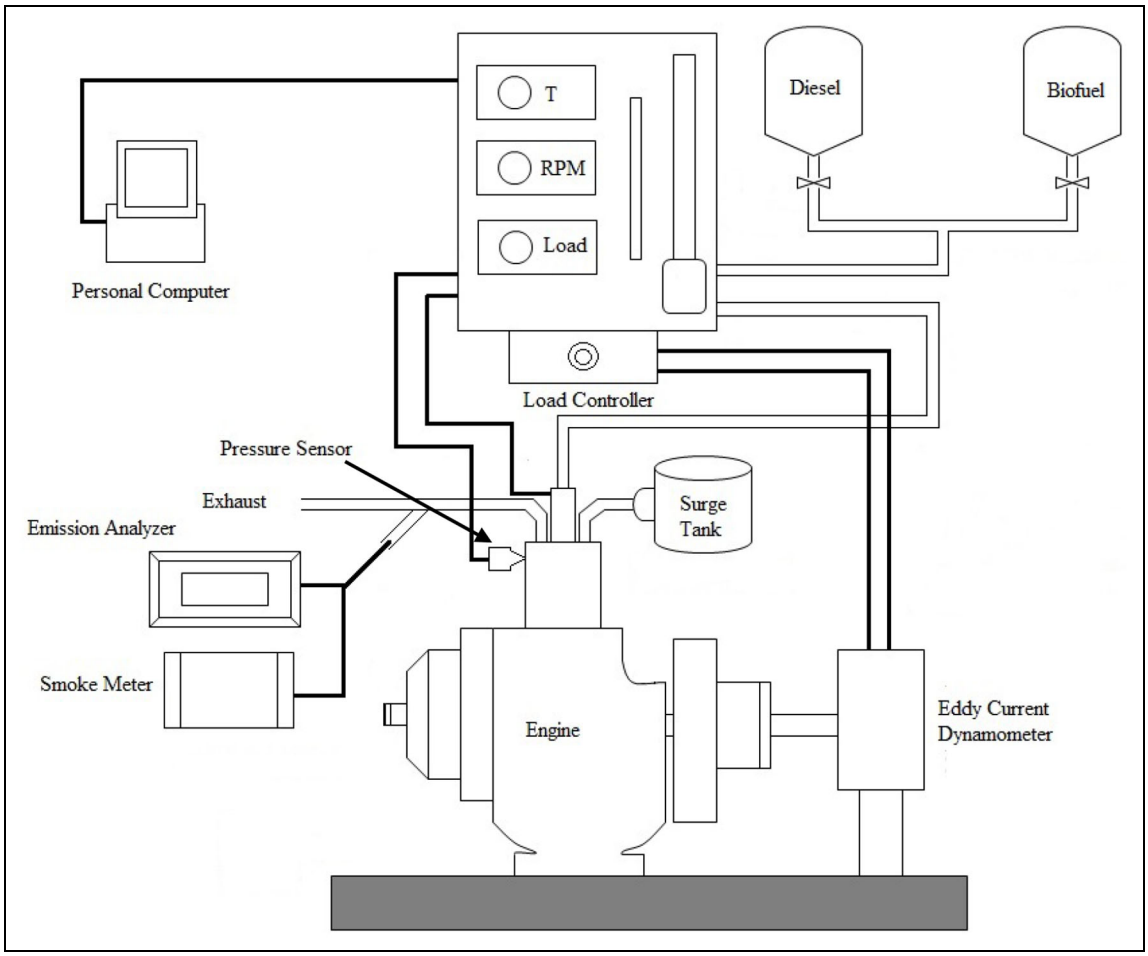


Figure 3. Schematic of the engine experiment setup.

Table 3. Specification of the test engine.

Specification	Value
Rated power	3.5 kW
Rated speed	1500 rpm
No. of cylinders	1
Stroke	4
Displacement	661 cc
Stroke length	110 mm
Bore diameter	87.5 mm
Compression ratio	17.5
Cooling system	Water cooled
Injection pressure	23°bTDC

specific intervals and multiplying it with fuel factor. The fuel factor is the amount of fuel (in kg) that comes out from the burette per unit change in the output voltage of the transmitter (kg/V). The fuel factor is given by equation (1)

$$\text{Fuel factor} = \rho \times g \times h \times A \quad (1)$$

where ρ is the density of fuel in (kg/m^3), g is the gravitational acceleration, h is the height of the column above the transmitter (m) and A is the cross-sectional area of the fuel pipe (m^2). It is seen that the density of the fuel is taken into account in the fuel factor; therefore, different fuels can be used.

To measure the engine exhaust emission AVL di-gas analyzer and AVL smoke meter was placed in the tailpipe of the engine which records the concentrations of CO, HC, NO emissions and smoke opacity released during burning of different fuel blends. The in-cylinder pressure was measured by PCB USA make piezoelectric pressure sensor (350 bar capacity) which was air-cooled and was placed on the cylinder head. The crank angle was measured by a Kubler crank angle sensor. The data acquisition system receive signals from all the sensors in the analog form and send further for processing to the engine software known as "Engine Soft." The software processes and converts the signals into digital form so that it is easily readable. The apparent heat release rate was also calculated from the in-cylinder pressure with the help of first law of thermodynamics using equation (2). For determining the heat release, it was assumed that the mixture inside the combustion chamber behaves as an ideal gas, the contents of the combustion chamber are homogeneous in nature and the system is a closed system, that is, mass is constant throughout the cycle and the flow of the mass to the crevices, blow-by and fuel injection is neglected[20,21]

$$\frac{dQ}{d\theta} = \frac{1}{\gamma - 1} v \frac{dP}{d\theta} + \frac{\gamma}{\gamma - 1} P \frac{dv}{d\theta} \quad (2)$$

where $dQ/d\theta$ is the apparent heat release, γ is the ratio of specific heats, p is the cylinder pressure and v is the cylinder volume.

The engine parameters were recorded at the rated speed of 1500 rpm under different loading conditions. The readings were noted at no load, 20%, 40%, 60%, 80% and full load for different blending ratios of the hydrotreated oil and diesel. The in-cylinder pressure was recorded for 50 continuous cycles and averaged. The performance and emission data were recorded in triplicate and averaged for each experiment. The blends contain 10%, 20%, 30%, 40% and 50% of hydrotreated oil on volume basis and the rest of the blend contained diesel. The 10% blend of hydrotreated oil and diesel was designated as H10D90, similarly other blends were designated as H20D80, H30D70, H40D60 and H50D50 for 20%, 30%, 40% and 50%, respectively, for blends of HVO and diesel. The engine was operated with the test fuels and their characteristics were recorded and compared with the diesel (baseline).

The detailed examination of the uncertainties in the physical measurement is error analysis. The error occurs due to the deviation in the measurement of the actual value. The precision and accuracy of the instrument and the operating conditions are some of the reasons for experimental uncertainty. The accuracy and the uncertainty of the instruments used in the study are shown in Table 4.

Table 4. Uncertainty of instruments and parameters.

Measurement	Accuracy	% Uncertainty
Load	±0.1 kg	±0.2
Speed	±10 rpm	±0.3
Pressure pickup	±0.1 bar	±0.08
Crank angle encoder	±1	±0.08
Fuel measurement	±0.1 cc	±1
CO	±0.03%	±0.1
HC	±10 ppm	±0.2
NO	±50 ppm	±2
Smoke	±0.1%	±1
EGT	±1 °C	±0.2

EGT: exhaust gas temperature.

Results and discussion

The results of carrying out a hot-plate test to find the ignition probability of the fuel samples are discussed. The SMD of the test fuels was found using the Malvern Spraytec and a detailed discussion on the results is given in this section. Finally, the fuels were used to operate a diesel engine. The effect of the fuels on the engine combustion, performance and emission is discussed in detail in this section.

Ignition probability

Figure 4 shows the ignition probability of the fuel samples with respect to surface temperature. The tests were conducted in the surface temperature range of 280 °C–640 °C, since our preliminary study suggests that film boiling occurs in this region for the test fuels.[17] Previous studies[22–24] also show that the ignition probability is the highest in the film boiling region. In this region, the droplet is covered by a vapor film and the heat transfer takes place through both radiation and conduction. With the rise in surface temperature, the radiation of heat through the vapor increases, thereby increasing the ignition probability. It is seen in Figure 4 that H100 has a higher ignition probability at all the temperatures than diesel and the blends of the hydrotreated oil and diesel. Its ignition probability reached 100% at 560 °C, whereas for diesel, the ignition probability reached 100% at 600 °C. The blends had ignition probability in between the two test fuels. The fuels' chemistry plays an important role in improving the ignition chances. The HVO contains large amount of paraffins and very low quantity of aromatics, which improves its cetane number, hence improving the ignition probability.[5,25] Also, the carbon atoms in H100 ranges from C_{11} to C_{18} whereas a typical diesel fuel contains carbon atoms in the range of C_8–C_{22}.[6] Fuels having low amount of long chain carbon atoms have relatively less boiling point and higher volatility as compared to fuels with high amount of long chain carbon atoms.[26] Therefore, H100 has a lower boiling point and higher volatility than diesel, thereby improving the ignition and hence increasing the ignition probability.

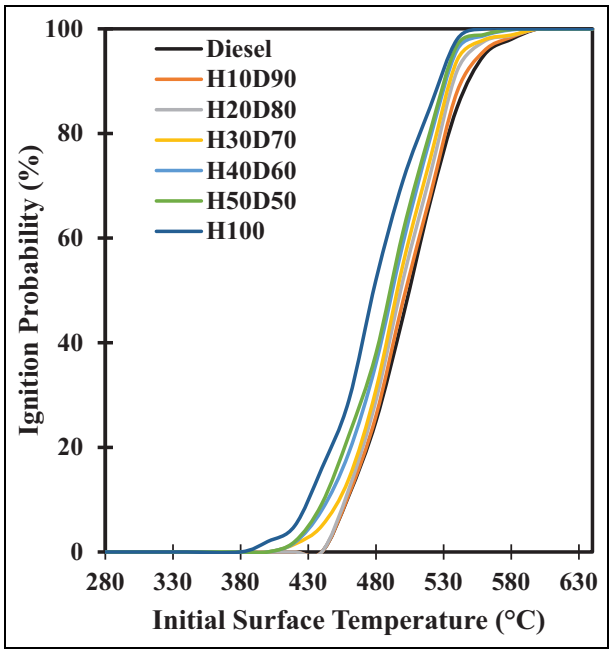


Figure 4. Variation of ignition probability with surface temperature of plate for different test fuels.

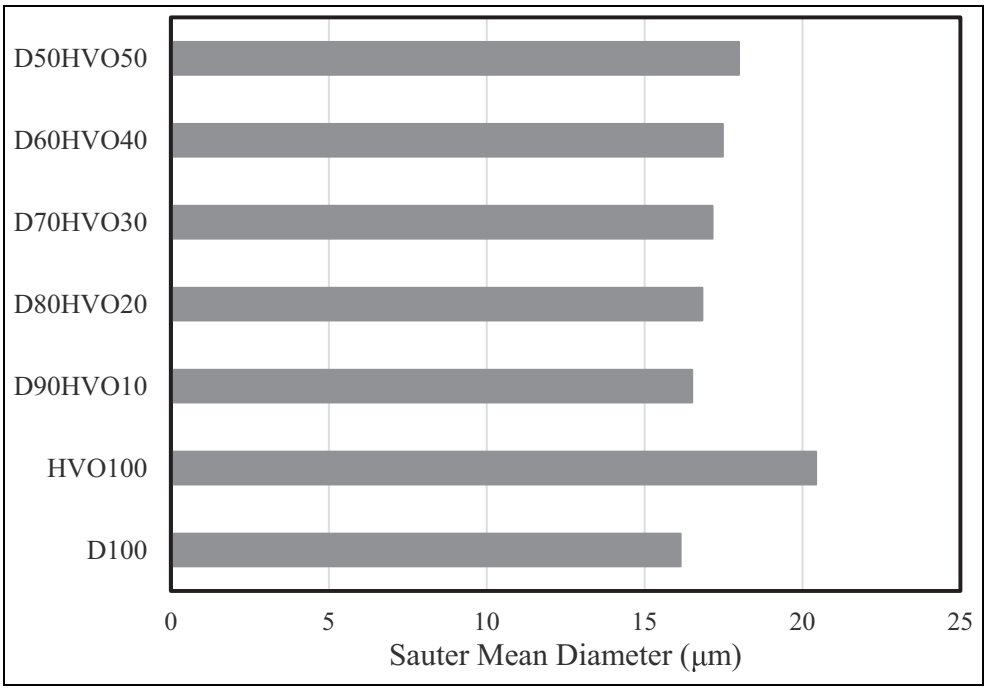


Figure 5. Sauter mean diameter of the test fuels.

Particle size

The fuel particles sprayed through an injector are of varying sizes, which makes it difficult to find the exact size of the fuel particles. Therefore, the SMD of the particles is statistically calculated. The combustion quality is invariably dependent upon the particle size. If the particle size is small, then its vaporization will be good, and the fuel will burn earlier. However, the fuel penetration will be less as the small particles have less momentum. If the particle size is large, then there will be delay in ignition and the vaporization will be slow. Therefore, the particle size should be optimum for efficient running of the engine.[21] The fuel viscosity, surface tension and injection pressure affect the particle size. The SMD of diesel, H100 and blends of diesel and H100 is shown in Figure 5. It is seen that the hydrotreated oil has a higher SMD than diesel due to its higher viscosity. The sprayed fuel breaks up into fine particles due to instabilities in the fuel jet.[27] However, the molecules of fuel with high viscosity are closely packed. When they are sprayed, they are unable to break into small particles as the force of collision among the molecules is less. Therefore, the particle movement is inhibited; hence, the instabilities needed for breaking the fuel jet is less, thereby increasing the particle size. Gao et al.[28] also found that biodiesel of various inedible oils had higher SMD due to their higher viscosity than diesel.

Combustion characteristics

In Figure 6, the heat release rate at full load is compared for diesel, H10D90, H20D80, H30D70, H40D60, H50D50 and H100. Engine operation with diesel resulted in the peak heat release of 56.13 J/°CA, whereas for H100 it is 45.96 J/°CA. The blends had peak heat release in between the two fuels and it decreases as the hydrotreated oil percentage in the blend increases. The curves also show that the heat release starts earlier for the neat hydrotreated oil and its blends as the delay in ignition is less with these fuels due to high cetane number. Moreover, it is seen that high heat release occurs in the diffusion combustion for H100 and its blends. Due to short ignition delay, less fuel burns in the initial phase of combustion. Similar results were obtained by Aatola et al.,[29] McDaniel et al.,[30] Sugiyama et al.,[31] Mattson and Depcik,[32] and Singh et al.[11] All the researchers concluded that the

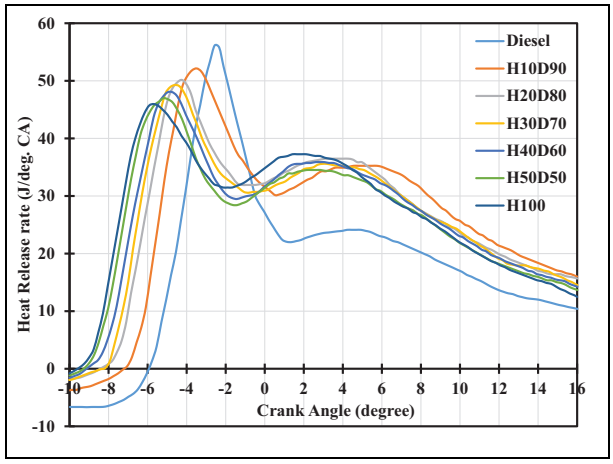


Figure 6. Variation in heat release rate at full load.

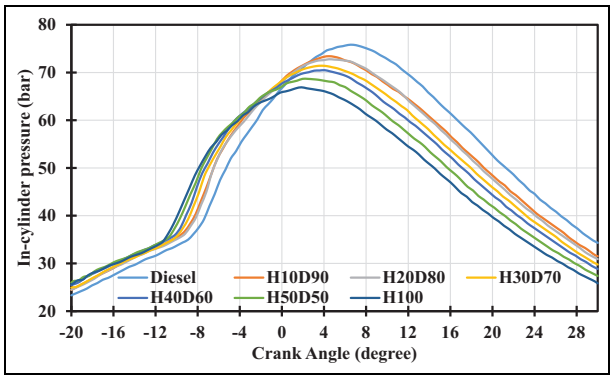


Figure 7. Variation in in-cylinder pressure at full load.

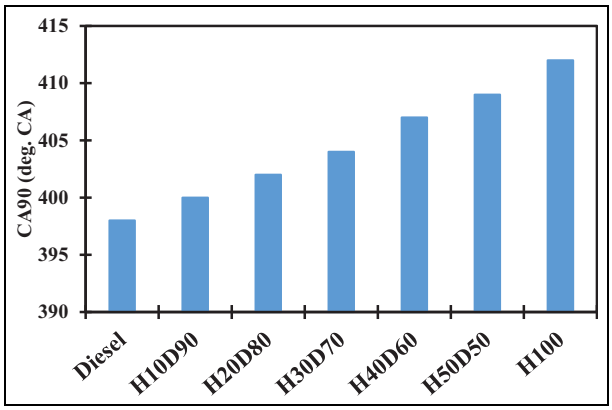


Figure 9. Variation in CA90 at full load.

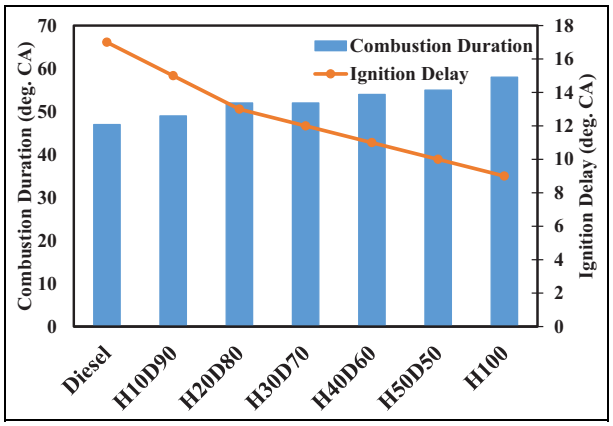


Figure 8. Variation in combustion duration and ignition delay at full load.

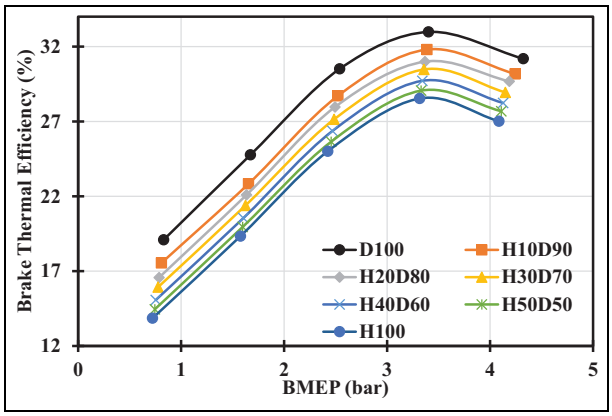


Figure 10. Variation in brake thermal efficiency at various loads.

lower peak heat release with the hydrotreated oil and its blends is caused by short ignition delay.

Figure 7 presents the in-cylinder pressure for the fuel samples. The peak in-cylinder pressure for diesel, H10D90, H20D80, H30D70, H40D60, H50D50 and H100 is 75.8 bar, 73.41 bar, 72.78 bar, 71.4 bar, 70.48 bar, 68.6 bar and 66.8 bar, respectively. The combustion with H100 and its blends with diesel resulted in lower peak pressures, since the heat released in the initial stage of combustion is lower in comparison to diesel. This is due to higher cetane index of the hydrotreated oil. As the hydrotreated oil percentage increases in the blend, its cetane number increases, thereby reducing the peak pressure.

The variation in combustion duration and ignition delay at full load is indicated in Figure 8. It is seen that as the percentage of the hydrotreated oil increases in the blend, the combustion duration increases and the ignition delay reduces. The reduction in ignition delay is due to higher cetane number of the blends. The increase in the combustion duration with the blends is due to low heat release in the first phase of combustion followed by a longer diffusion combustion phase (Figure 6). As seen in Figure 10, the thermal efficiency of the engine with the blends is lower than diesel, that is, more amount of fuel will be injected to produce the same power. This extra fuel requires more time to burn, thereby extending the combustion duration.

The crank angle at which 90% of the total heat (CA90) is released is shown in Figure. 9. The CA90 was found to increase as the percentage of the hydrotreated oil in the blend increases and it was highest for neat hydrotreated oil engine operation. The initial temperature of the combustion chamber is low with the blends, thereby decreasing the molecular activity. The reduction in initial temperature is caused by higher cetane number of the fuel which lowers the heat release rate in premixed-combustion phase. Moreover, due to large amount of fuel injection (low BTE), the heat release of the excess fuel takes place in the diffusion combustion phase, thereby increasing the CA90.

Performance characteristics

The influence of the test fuels on the engine brake thermal efficiency (BTE) is displayed in Figure 10. The engine's ability in converting the fuel's chemical energy into useful shaft work is known as BTE. With the increase in engine load, the BTE increases as the amount of fuel injected increases which in turn releases

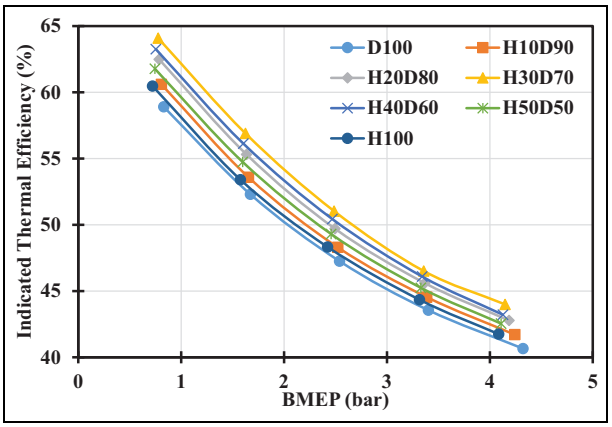


Figure 11. Variation in indicated thermal efficiency at various loads.

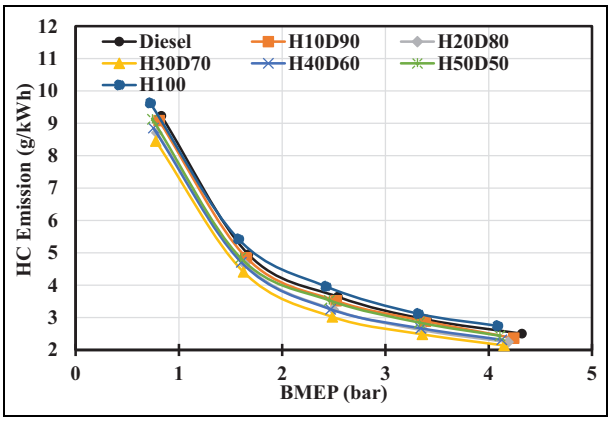


Figure 13. Variation in brake-specific unburned hydrocarbon emission at various loads.

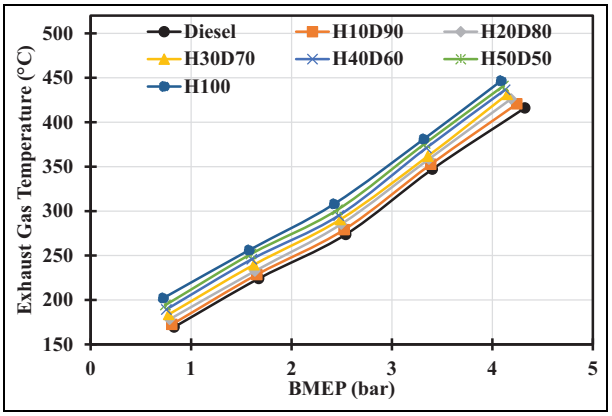


Figure 12. Variation in exhaust gas temperature at various loads.

more energy thereby increasing the produced power. The trend of all the test fuels is similar to diesel. Moreover, with the rise in the percentage of the hydrotreated oil in the blend, the BTE decreases. The BTE at full load for H100, H50D50, D40D60, H30D70, H20D80, H10D90 and diesel is 27%, 27.6%, 28.2%, 28.9%, 29.6%, 30.17% and 31.18%, respectively. The calorific value of the hydrotreated oil is low and its viscosity is slightly higher, which causes a decrease in the thermal efficiency. Due to the high viscosity of the oil, the size of the sprayed fuel particle is larger than diesel (Figure 5) which may affect the fuel–air mixing tendency and hence delay the ignition.[33,34] However, it is seen in Figure 6 that the heat release for the blends is earlier than diesel, which suggests that the particle size does not have much effect on the ignition delay.

The variation in indicated thermal efficiency with the test fuels at different loads is shown in Figure 11. The indicated thermal efficiency of the hydrotreated oil and its blends is higher than diesel at all loads. The improvement in efficiency is caused by better combustion quality of the hydrotreated oil due to its high cetane number, and the energy requirement is less due to their high paraffinic content and short chain hydrocarbons. Moreover, the thermal efficiency increases till 30% hydrotreated oil mixed in the blend, further addition of the oil resulted in the decrease in efficiency. As the percentage of the hydrotreated oil in the blend increases, the cetane number increases resulting in early start of heat release which reduces the combustion pressure and temperature. This decrease in temperature and pressure reduces the combustion efficiency.

Figure 12 shows the exhaust gas temperature (EGT) with load for different test fuels. It is seen that the EGT increases for all the test fuels as the engine load increases. The lowest EGT was observed with diesel fuel engine operation at all the loads. Furthermore, the EGT increases as the hydrotreated oil percentage increases in the blend. At full load, diesel, H10D90, H20D80, H30D70, H40D60, H50D50 and H100 had 416 °C, 420 °C, 426 °C, 431 °C, 437 °C, 442 °C and 446 °C, respectively, as the EGT. The EGT increases when the heat released in the diffusion combustion phase is higher in comparison to premixed-combustion phase. Since this heat cannot produce useful power, it leaves the engine as a waste.[35,36] Figure 6 shows that the heat released for the neat oil and its blends is higher in diffusion combustion phase, as compared to diesel. Moreover, the trend of CA90 in Figure 9 confirms that the 90% of the heat released with the blends is late in comparison to diesel which is unutilized and goes into the engine exhaust as waste heat.

Emission characteristics

Figure 13 portrays the unburned hydrocarbon emission for the fuel samples. The emission is high at low loads; with the increase in engine load, the emission level decreases. Same trend was followed by all the tested fuel samples. Due to incomplete combustion, the hydrocarbon emission takes place as the charge present near the cylinder walls is unable to burn because of the quenching effect. This effect is more dominant at low loads,

since the temperature in the cylinder is low and the combustion flame is unable to reach the far ends of the combustion chamber.[37] At full load, the HC emissions as compared to diesel reduced by 5.5%, 10.3%, 14.4%, 7.4% and 2.5% with H10D90, H20D80, H30D70, H40D60 and H50D50, respectively, whereas with H100, the HC emission increases by 9.7%. Although the heat released during the combustion of the hydrotreated oil blends is lower than diesel, the HC emission decreases. This is due to the improvement in the combustion because of its high paraffinic and very low aromatic content. Since it is easier to breakdown straight chain carbon atoms into smaller chain hydrocarbons and gaseous components as the energy requirement is less as compared to fuel containing high amount of aromatic rings.[6] Also the hydrotreated oil has shorter chain carbon atoms as compared to diesel resulting in lower energy requirement. The figure shows that the HC emissions decreases till 30% of the hydrotreated oil is present in the blend. Further increase in the percentage of the hydrotreated oil resulted in an increase in HC emission at all the loads. Till 30% blending, the heat release is high enough for proper combustion, thereby reducing the HC emissions. Further increase in blending percentage reduces the heat release to an extent such that the combustion is unable to sustain, hence the HC emission starts increasing and engine operation with H100 resulted in the highest HC emission.

The effect on CO emission under the influence of test fuels and various loads is shown in Figure 14. The emission decreases with the increase in load up to 50% engine load and further rise in load tends to increase the CO emission and highest CO emission is observed at full load. At low loads, the injected fuel may hit the walls of the relatively cool combustion chamber which reduces the fuel–air mixing rate, resulting in incomplete combustion, whereas, at high loads, the mixture is relatively rich; therefore, the carbon molecules may not find enough oxygen to be completely converted into carbon dioxide.[38] The CO emissions at full load for diesel, H10D90, H20D80, H30D70, H40D60, H50D50 and H100 is 31.9, 29.7, 27.9, 26.55, 27.41, 30.7 and 34 g/kW h, respectively. The trend of CO emissions is similar to HC emissions, that is, it decreases for the blends and the lowest emission is observed with H30D70. As the hydrotreated oil percentage further increases in the blend, the CO emissions start increasing. The highest CO emissions are observed with H100 fuel operated engine. Carbon monoxide is formed when the carbon containing compounds combustion in air is incomplete. Since the combustion of the blends of the hydrotreated oil is good due to their higher paraffinic content and cetane number, a decrease in carbon monoxide emission is observed. Also, small amount of oxygen present in the hydrotreated oil aids the conversion of CO to CO_2. The increase in CO emission with H100 engine operation is because of lower combustion temperature which may cause incomplete combustion.

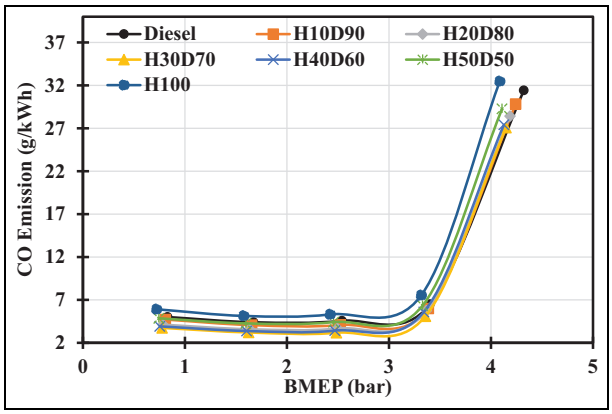


Figure 14. Variation in brake-specific carbon monoxide emission at various loads.

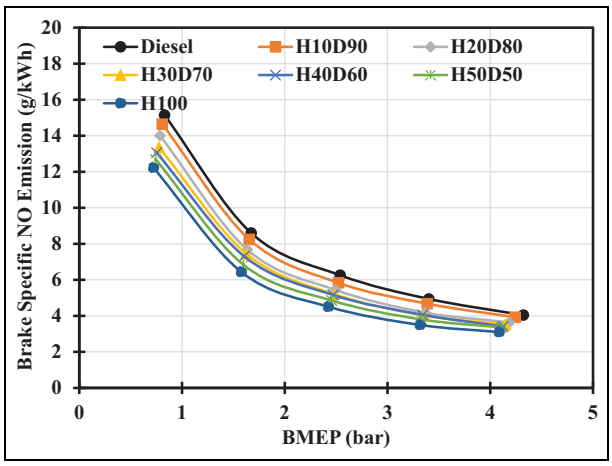


Figure 15. Variation in brake-specific oxides of nitrogen emission at various loads.

The NO emission depends upon the temperature of combustion (thermal NO), the time available for the chemical reaction at high temperature and the availability of oxygen rich zones. The molecular nitrogen can also react with the hydrocarbon radicals in the fuel rich flame zone resulting in the formation of prompt NO. Nitrogen contained in the fuel is also responsible for NO formation. However, high combustion temperature is the main reason for the NO formation in a CI engine. Extended Zeldovich mechanism can be used to explain the NO formation.[10,21] Figure 15 depicts the brake-specific NO emission with load. The NO emission at full load is 4, 3.9, 3.6, 3.5, 3.4, 3.3 and 3.1 g/kW h for diesel, H10D90, H20D80, H30D70, H40D60, H50D50 and H100, respectively. The decrease in NO emissions with the test fuels is due to low premixed combustion as compared to diesel resulting in lower combustion temperatures. Due to high cetane number of the test fuels, the delay in ignition is reduced thereby decreasing the premixed-combustion phase. Singh et al.[11] used a single cylinder CI engine with 19:1 compression ratio and observed NO to be lower than diesel for the

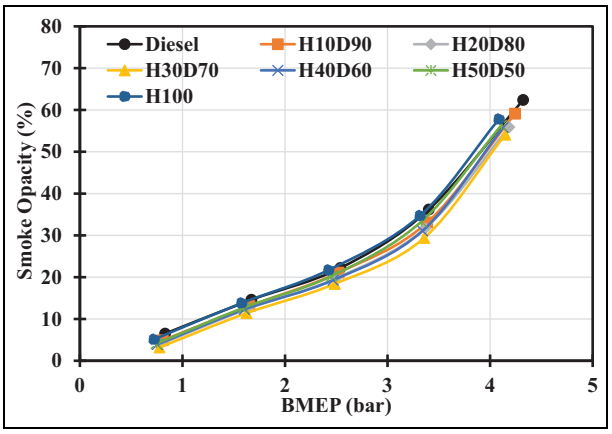


Figure 16. Variation in smoke opacity at various loads.

hydrotreated oil. In another study, Singh et al.[10] used a turbocharged diesel engine with 17:1 compression ratio and observed the NO emission to be higher than diesel. The turbocharged engine had a lower compression ratio; therefore, the ignition delay would be higher than the single cylinder engine, thereby increasing the time for the hydrotreated oil to mix properly and release more heat during the premixed-combustion phase causing higher NO emission. In this study, the engine used has a compression ratio of 17.5:1; therefore, it is expected that due to shorter delay in ignition and small heat release with the hydrotreated oil and its blends the NO emissions are lower in comparison to diesel.

Figure 16 demonstrates the effect of the test fuels at various loads on the engine exhaust smoke. The soot formation phenomenon during combustion is a complex one and it depends upon factors like the fuel composition, the ignition delay and the air–fuel ratio. The exhaust gas opacity was measured in this study, since it is an indirect measure of the amount of soot present in the exhaust gas. It is seen that the soot emission increases with the increase in load on the engine, for all the test fuels. H100 and its blends with diesel had lower smoke opacity in comparison to diesel. At full load, the smoke opacity is 62.4%, 59.1%, 55.9%, 54.1%, 55.8%, 56.3% and 57.8% with diesel, H10D90, H20D80, H30D70, H40D60, H50D50 and H100, respectively. The aromatic content of the fuel is considered as the main reason behind the soot formation. Since the hydrotreated oil has low aromatic content, the soot formation tendency is less. Moreover, due to lower premixed-combustion phase, the fuel and air have enough time to mix, thereby improving the combustion quality leading to complete combustion. However, it is also seen that the particle size of the test fuels was slightly larger than diesel; therefore, the initial mixing of the fuel and air may not be proper. This has two effects, such as the initial vaporization may be slow and the bigger fuel particles will travel a large distance, thereby reaching oxygen rich places in the combustion chamber.[21] Since the soot emission is low which shows that the latter effect is more dominant, also due to better ignition quality and high cetane number the fuel particles in the far places of the combustion chamber will also burn. Singh et al.,[39,11] Prokopowicz et al.,[40] and Ogunkoya et al.[6] also observed reduction in smoke emissions. By blending the hydrotreated oil with oxygenates, even further reduction in PM emissions is possible.[41] The trends also show an increase in smoke emissions when the hydrotreated oil percentage in the blend is higher than 30%. The probable reason for the trend is the decrease in the combustion temperature which is necessary for sustaining the chemical reactions.

Conclusion

Experiments were conducted in this work to study the effect of hydrotreated oil and its blends with diesel on a single cylinder diesel engine. The performance and emission results of the test fuels were compared with the fossil–diesel. The effect of fuel properties on ignition probability and SMD was also studied. The conclusions drawn from the study are as follows:

- As the percentage of the hydrotreated oil in the blend with diesel increases, its ignition probability increases with neat hydrotreated oil having the highest ignition probability.
- The SMD of diesel is lowest and as the hydrotreated oil percentage increases the SMD increases. The neat hydrotreated oil has an SMD of 20.4 μm, whereas diesel has 16.1 μm SMD.
- The peak heat release with the hydrotreated oil and its blends were 4% to 15% lower in comparison to diesel.
- As the heat release was low, the peak in-cylinder pressure was also lower for the blends and neat fuel.
- The BTE and the brake-specific fuel consumption of the test fuels were lower than diesel on account of lower heat release.
- The unburned hydrocarbon and carbon monoxide emissions were lower than diesel till 30% blend of the hydrotreated oil; further increase in the hydrotreated oil percentage resulted in increase in HC emissions. Neat hydrotreated oil had higher HC and CO emissions than diesel.
- The NOx emissions were lower than diesel for all the test fuel samples due to lower combustion temperature.
- The soot emission was also lower for the test fuels, since the fuel contains low amount of aromatics and small amount of oxygen which helps in complete combustion. Moreover, due to lower combustion temperature, the smoke emissions start increasing as the percentage of the hydrotreated oil in the blend exceeds 30%.

The study shows that the use of hydrotreated oil and its blends in a CI engine reduces the regulated emissions

because of its paraffinic nature and low aromatic content. However, the HC and CO emission reduction is observed till 30% blend of the hydrotreated oil. The flipside of using the hydrotreated oil is the decrease in the BTE due to lower heat release and increase in fuel consumption. Therefore, it is recommended that till 30%, the hydrotreated oil be allowed to blend with diesel. This will help in reducing the emissions as well as keep the fuel consumption closer to diesel.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ankit Sonthalia https://orcid.org/0000-0002-4766-4432

References

1. International Energy Agency. World energy outlook, https://www.iea.org/reports/world-energy-outlook-2019 (accessed 13 November 2019).
2. Tilli A, Hulkkonen T, Kaario O, et al. Biofuel blend late post-injection effects on oil dilution and diesel oxidation catalyst performance. *Int J Engine Res* 2017; 19(9): 941–951.
3. Singh AP and Agarwal AK. Performance and emission characteristics of conventional diesel combustion/partially premixed charge compression ignition combustion mode switching of biodiesel-fueled engine. *Int J Engine Res*. Epub ahead of print 16 July 2019. 10.1177/1468087419860311.
4. Lapuerta M, RamosÁ, Rubio S, et al. Optimization of a diesel engine calibration for operating with a residual glycerol-derived biofuel. *Int J Engine Res*. Epub ahead of print 20 December 2019. DOI: 10.1177/1468087419891535.
5. Sonthalia A and Kumar N. Hydroprocessed vegetable oil as a fuel for transportation sector: a review. *J Energy Inst* 2019; 92(1): 1–17.
6. Ogunkoya D, Roberts WL, Fang T, et al. Investigation of the effects of renewable diesel fuels on engine performance, combustion, and emissions. *Fuel* 2015; 140: 541–554.
7. Kim D, Kim S, Oh S, et al. Engine performance and emission characteristics of hydrotreated vegetable oil in light duty diesel engines. *Fuel* 2014; 125: 36–43.
8. Caprotti R, Tang T, Ishibe N, et al. Performance of diesel containing bio-hydrogenated component. *SAE technical Paper series no*. 2011-01-1953, 2011.
9. Westphal GA, Krahl J, Munack A, et al. Combustion of hydrotreated vegetable oil and jatropha methyl ester in a heavy duty engine: emissions and bacterial mutagenicity. *Environ Sci Technol* 2013; 47(11): 6038–6046.
10. Singh D, Subramanian KA and Singal SK. Emissions and fuel consumption characteristics of a heavy duty diesel engine fueled with hydroprocessed renewable diesel and biodiesel. *Appl Energy* 2015; 155: 440–446.
11. Singh D, Subramanian KA, Bal R, et al. Combustion and emission characteristics of a light duty diesel engine fueled with hydro-processed renewable diesel. *Energy* 2018; 154: 498–507.
12. Happonen M, Lähde T, Messing ME, et al. The comparison of particle oxidation and surface structure of diesel soot particles between fossil fuel and novel renewable diesel fuel. *Fuel* 2010; 89(12): 4008–4013.
13. Pflaum H, Hofmann P, Geringer B, et al. Potential of hydrogenated vegetable oil (HVO) in a modern diesel engine. *SAE technical Paper series no*. 2010-32-0081, 2010.
14. Kopperoinen A, Kytö M and Mikkonen S. Effect of hydrotreated vegetable oil (HVO) on particulate filters of diesel cars. *SAE technical Paper series no*. 2011-01-2096, 2011.
15. Erkkilä K, Nylund N, Hulkkonen T, et al. Emission performance of paraffinic HVO diesel fuel in heavy duty vehicles. *SAE technical Paper series no*. 2011-01-1966, 2011.
16. Singh D, Subramanian KA and Garg M. Comprehensive review of combustion, performance and emissions characteristics of a compression ignition engine fueled with hydroprocessed renewable diesel. *Renew Sustain Energy Rev* 2018; 81: 2947–2954.
17. Sonthalia A and Kumar N. Comparison of fuel characteristics of hydrotreated waste cooking oil with its biodiesel and fossil diesel. *Environ Sci Pollut Res*. Epub ahead of print 17 December 2019. DOI: 10.1007/s11356-019-07110-w.
18. Haynes A, Shaik MS, Krarup H, et al. Evaluation of the Malvern Spraytec with inhalation cell for the measurement of particle size distribution from metered dose inhalers. *J Pharm Sci* 2004; 93(2): 349–363.
19. Daho T, Vaitilingom G, Ouiminga SK, et al. Influence of engine load and fuel droplet size on performance of a CI engine fueled with cottonseed oil and its blends with diesel fuel. *Appl Energy* 2013; 111: 1046–1053.
20. Soloiu V, Duggan M, Harp S, et al. PFI (port fuel injection) of n-butanol and direct injection of biodiesel to attain LTC (low-temperature combustion) for low-emissions idling in a compression engine. *Energy* 2013; 52: 143–154.
21. Heywood JB. *Internal combustion engine fundamentals*. New York: *Mc-graw Hill*, 1988.
22. Xiong TY and Yuen MC. Evaporation of a liquid droplet on a hot plate. *Int J Heat Mass Transf* 1991; 34(7): 1881–1894.
23. Arifin YM and Arai M. The effect of hot surface temperature on diesel fuel deposit formation. *Fuel* 2010; 89(5): 934–942.
24. Arifin YM, Furuhata T, Saito M, et al. Diesel and biodiesel fuel deposits on a hot surface. *Fuel* 2008; 87(8): 1601–1609.
25. No SY. Application of hydrotreated vegetable oil from triglyceride based biomass to CI engines—a review. *Fuel* 2014; 115: 88–96.
26. Bacha J, Freel J, Gibbs A, et al. *Diesel fuels technical review*. San Ramon, CA: Chevron, 2007: pp1–116.

27. Ejim CE, Fleck BA and Amirfazli A. Analytical study for atomization of biodiesels and their blends in a typical injector: surface tension and viscosity effects. *Fuel* 2007; 86(10–11): 1534–1544.
28. Gao Y, Deng J, Li C, et al. Experimental study of the spray characteristics of biodiesel based on inedible oil. *Biotechnol Adv* 2009; 27(5): 616–624.
29. Aatola H, Larmi M, Sarjovaara T, et al. Hydrotreated vegetable oil (HVO) as a renewable diesel fuel: trade-off between NOx, particulate mission, and fuel consumption of a heavy duty engine. *SAE technical Paper series no*. 2008-01-2500, 2008.
30. McDaniel A, Dickerson T, Luning-Prak D, et al. A technical evaluation of new renewable jet and diesel fuels operated in neat form in multiple diesel engines. *SAE technical Paper series no*. 2016-01-0829, 2016.
31. Sugiyama K, Goto I, Kitano K, et al. Effects of hydrotreated vegetable oil (HVO) as renewable diesel fuel on combustion and exhaust emissions in diesel engine. *SAE Int J Fuels Lubr* 2011; 5(1): 205–217.
32. Mattson J and Depcik C. First and second law heat release analysis in a single cylinder engine. *SAE Int J Engines* 2016; 9: 536–545.
33. Geo VE, Sonthalia A, Nagarajan G, et al. Effect of port premixed liquefied petroleum gas on the engine characteristics. *J Energy Resour Technol Trans ASME* 2019; 141(11): 1–8.
34. Varuvel EG, Sonthalia A, Subramanian T, et al. NOx-smoke trade-off characteristics of minor vegetable oil blends synergy with oxygenate in a commercial CI engine. *Environ Sci Pollut Res* 2018; 25(35): 35715–35724.
35. Boopathi D, Thiyagarajan S, Edwin Geo V, et al. Effect of geraniol on performance, emission and combustion characteristics of CI engine fuelled with gutter oil obtained from different sources. *Energy* 2018; 157: 391–401.
36. Boopathi D, Thiyagarajan S, Geo EV, et al. Effect of the second generation and third generation biofuel blend on performance, emission and combustion characteristics of CI engine. *Int J Ambient Energy*. Epub ahead of print 11 July 2018. DOI: 10.1080/01430750.2018.1492439.
37. Subramanian T, Sonthalia A and Varuvel EG. Effect of calcite/activated carbon-based post-combustion CO_2 capture system in a biodiesel-fueled CI engine—an experimental study. *Energy Sour Part A Recover Util Environ Eff* 2019; 41(16): 1972–1982.
38. Ashok B, Nanthagopal K, Perumal AD, et al. An investigation on CRDi engine characteristic using renewable orange-peel oil. *Energy Convers Manag* 2019; 180: 1026–1038.
39. Singh D, Sarma AK and Sandhu SS. A comprehensive experimental investigation of green diesel as a fuel for CI engines. *Int J Green Energy* 2019; 16(14): 1152–1164.
40. Prokopowicz A, Zaciera M, Sobczak A, et al. The effects of neat biodiesel and biodiesel and HVO blends in diesel fuel on exhaust emissions from a light duty vehicle with a diesel engine. *Environ Sci Technol* 2015; 49(12): 7473–7482.
41. Happonen M, Heikkilä J, Aakko-Saksa P, et al. Diesel exhaust emissions and particle hygroscopicity with HVO fuel-oxygenate blend. *Fuel* 2013; 103: 380–386.

Environment, Development and Sustainability
https://doi.org/10.1007/s10668-020-00804-z

An integrated approach for landslide susceptibility–vulnerability–risk assessment of building infrastructures in hilly regions of India

Aditi Singh[1] · Shilpa Pal[2] · D. P. Kanungo[3]

Received: 27 December 2019 / Accepted: 2 June 2020
© Springer Nature B.V. 2020

Abstract

Considering the ever-increasing landslide incidences in Indian Himalayas, a methodology has been presented to assess the risk to buildings constructed in the landslide-prone areas. Since landslide is a dynamic phenomenon, an inter-disciplinary approach is required for the assessment of elements at risk (buildings in this case). Therefore, a novel remote sensing and GIS-based semi-quantitative technique has been developed by integrating the concepts of landslide susceptibility zonation (LSZ), physical vulnerability (PV) and the proximity (P_{rox}) of buildings from the influence zone (i.e. LSZ and drainage channels). In order to understand the acceptability of risk, the landslide risk (LR) has been categorized into three risk classes as class I (low risk), class II (moderate risk) and class III (high risk). This study aims to develop a systematic and easy to adopt methodology for hilly terrains of India in a scenario of historical data scarcity and also in line with the codal provisions of the country as well as the geographical conditions. The developed methodology is implemented in a test site of Gopeshwar Township, Chamoli District Headquarter, Uttarakhand State of India, covering an area of 8.39 km^2 situated in the upper Alaknanda valley. This study will be useful in increasing the safety aspects of the infrastructures and lives and also for strategic governance of developmental activities in the times ahead, especially in developing countries.

Keywords Building resistance · Landslide · Physical vulnerability · Proximity · Risk · Susceptibility

✉ D. P. Kanungo
debi.kanungo@gmail.com

[1] Department of Civil Engineering, Sarala Birla University, Ranchi, Jharkhand, India

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

[3] Geo-Hazard Risk Reduction (GHRR) Group, CSIR—Central Building Research Institute (CBRI), Roorkee, Uttarakhand, India

Published online: 09 June 2020

1 Introduction

Landslides in the Indian Himalayas have gained global attention primarily because of escalating pressure of urbanization on the mountainous environment, thereby leading to overexploitation of natural resources. Apart from the rise in ecological concerns, unplanned urbanization also gave rise to dysfunctional and aesthetically unattractive cities (Cetin et al. 2019; Adiguzel et al. 2019). Urban spaces are intended to be managed in light of ecological concerns in order to provide sustainable urbanization. According to Cetin et al. (2019), green concept-based solutions for urban space utilization for community parks, playgrounds, sports fields, etc., may be introduced within the city. This study also highlights the significance of ecological aspects in urban landscape design for development of sustainable cities. A holistic approach in terms of aesthetics and sustainability, quality and practicality from ecological and socio-economic points of view are always welcomed in urban design (Kaya et al. 2019; Adiguzel et al. 2019; Cetin et al. 2019). Keeping these vital aspects of safe and sustainable urbanization in mind specifically for hilly terrains, landslide susceptibility–vulnerability–risk assessment of the existing urban built spaces should be given paramount importance and primary focus to adopt corrective measures for ensuring ecological sustainability towards landslide disaster risk reduction (LDRR).

In India, nearly 0.42 km^2 (i.e. 12.6% of its territory) is subjected to various degrees of landslide (www.gsi.gov.in, retrieved on 07th August, 2018). The North-East Himalaya, North-West Himalaya, Western Ghats and Konkan hills and Eastern Ghats of Andhra Pradesh are the regions prone to landslide hazard (BMTPC 2019). Froude and Petley (2018) have reported based on non-seismic landslide events during 2004–2016 across the globe that India contributes to about 28% towards construction-triggered landslide events. Further, due to rapid urbanization in hilly terrains, the roadways network of India has almost expanded threefold in length in the last 24 years (Govt. of India 2016) leading to 30% of the landslide events due to construction activities along the roads. Not only this, the catastrophic landslides initiated by illegal mining practices have emerged in nearly 32 countries with India ranking at 15th. Therefore, in order to reduce the probable losses of infrastructures and people living in it, assessment of landslide risk becomes crucial.

In spite of all technological progress by various scientists and researchers in the field of landslides, majority of the techniques applied for executing landslide vulnerability and risk assessment possesses certain shortcomings. Nevertheless, assessment of landslide vulnerability and risk is a bit complicated process, as it involves the integration of different techniques, methodology and tools and the amalgamation of various expertise. In most of the countries, there is a deficiency in standardized parameters to assess landslide vulnerability and risk. As of now, very limited research has been carried out to assess vulnerability and risk to buildings and infrastructures exposed to landslide disasters. However, only four countries (i.e. Switzerland, France, Italy and Sweden) have developed risk assessment framework and are practically using it (SafeLand 2010). Hence, suitable and applicable methodology for landslide vulnerability and risk assessment is still in its developmental stage for many countries.

The present study intends to develop a landslide risk assessment (LRA) methodology applicable to hilly terrains in Indian scenarios by considering the data availability, codal provisions and the geographical conditions of Indian sub-continent. A novel remote sensing and GIS-based semi-quantitative technique for LRA has been developed by integrating the concepts of landslide susceptibility zonation (LSZ), physical vulnerability (PV) and the proximity (P_{rox}) of buildings from the influence zone (i.e. LSZ and drainage channels). The

developed methodology is tested on 542 buildings located under 11 wards in Gopeshwar Township of Chamoli District in Uttarakhand State (India). This particular work will be useful to identify the areas prone to landslide disasters along with the vulnerability and risk to different types of building infrastructures against landslides in the area posing different levels of threat to the people residing over there. Planned urbanization allows the progress in socio-economic and physical context (Cetin et al. 2019). Therefore, this information will not only provide an ideal framework for outlining developmental actions in the area but will also help in preparing the building codes/guidelines/ordinances for engineering practices.

2 Review of existing methodologies to assess landslide risk

In general, risk is a degree of the likelihood and severeness of devastating impact on the health, property or environment (AGS 2000). A number of definitions of risk in context to landslides have been proposed by various researchers. Varnes (1984) defined landslide risk as the probable lives lost, persons injured, destruction to property and hampering of economic activity due to a specific landslide event in a particular area and reference period. However, Royal Society (1992) defined risk as a blend of the likelihood or frequency of happening of a particular hazard and the magnitude of the aftermath of occurrence. This definition is quite essential as it indicates the importance of a phenomenon (i.e. landslides in the present case) in generating risk and the significance of consequences (i.e. building infrastructures in the present case) in the assessment of risk. In fact, it has been recognized that landslide risk should be expressed as the function of landslide hazard and the vulnerability of elements at risk during last more than three decades (Varnes and IAEG 1984; Michael-Leiba et al. 1999; Cardinali et al. 2002; Remondo et al. 2005; Uzielli et al. 2008; Van Westen et al. 2008; Zêzere et al. 2008; Garcia et al. 2016).

Formulated on the method of processing and input details, landslide risk assessment is segmented into qualitative risk assessment (Cardinali et al. 2002; Van Westen et al. 2006; Fell et al. 2008; Zêzere et al. 2008) and quantitative risk assessment (Dai et al. 2002; Remondo et al. 2005; Jaiswal and Van Westen 2009). However, the suitability of either qualitative or quantitative assessments leans on the required accuracy of the results and the type of problem. Generally, qualitative risk assessment is based on the experience of expert judgements and generally classified into three (low, moderate and high) or five (very low, low, moderate, high and very high) risk classes. One of the earliest examples of qualitative risk assessment developed in the late 1970s by the Hong Kong GCO (Geotechnical Control office) is the Cut Slope Ranking System (Koirala and Watkins 1988). Therefore, since 1970s a variety of qualitative landslide risk assessment methods have been developed, which includes (a) risk registers (Lee 1999; Lee and Clark 2000); (b) relative risk scoring or rating (Bogget et al. 2000; Rautela and Lakhera 2000; Chau et al. 2004; Kelly et al. 2005; Saldivar-Sali and Einstein 2007); (c) risk ranking matrices (Anbalagan and Singh 1996; AGS 2000; Van Dine et al. 2002; Cardinali et al. 2002); (d) relative risk rating (Palmer et al. 2002; Wong 2005); and (e) the FMECA (failure modes, effects and criticality analysis) approach (Sandilands et al. 1998; Hughes et al. 2000). On the other hand, quantitative landslide risk assessment depends on the measurable values of the likelihood, vulnerability and repercussions. The results thus acquired are presented as a binary value of the risk (Fell and Hartford 1997; SafeLand 2010; Martha et al. 2013; Pereira et al. 2016; Zou et al. 2016) that can be correlated with other study areas. However, more recently

risk assessment techniques have been further classified as: qualitative, semi-quantitative and quantitative (AGS 2000; Chowdhury and Flentje 2003). The fundamental distinctness between qualitative and semi-quantitative LRA method is that in the semi-quantitative approach (Kanungo et al. 2008; Abdulwahid and Pradhan 2016), risk classes are arranged by a hierarchy system and weighted within a specified condition.

Several different parameters are required to assess the landslide risk, because of which its actual quantification is difficult (Van Westen et al. 2006). A number of studies have been conducted on LRA worldwide; however, the methodologies to assess the landslide risk vary according to the applied scale, the type of mechanism and also on the accessibility of the input data. Also, due to wide diversity in topography, geology, climatic conditions, etc., around the globe, no universal framework is yet available to assess the landslide risk (Singh et al. 2017).

Several guidelines, overviews and discussions on the landslide risk assessment have also been published (Dai et al. 2002; Van Westen 2004; Catani et al. 2005; AGS 2007). Dai et al. (2002) presented a framework and identified different major steps to assess the landslide risk. The study also stated that a variety of risk (i.e. distributed landslide risk, site-specific landslide risk and global landslide risk) needs to be incorporated in the assessment of landslide risk. Van Westen (2004) presented an overview of recent developments for assessing the landslide risk by using geo-information tools. Catani et al. (2005) presented a complete set of methodology that includes the assessment of susceptibility, hazard, vulnerability and exposure and also a case study for assessing the landslide risk in Central Italy. The national committee of the Australian Geomechanics Society (AGS 2007) has also set up guidelines for landslide risk management.

In previous studies, scientists and researchers have expressed LRA in a number of ways. Chung and Fabbri (2003) stated that risk assessment approaches can be addressed qualitatively by expert knowledge or quantitatively by using statistical or complex mathematical analysis. Kanungo et al. (2008) carried out a semi-quantitative landslide risk assessment study in Darjeeling Himalayas (India) using the fundamentals of danger pixels and fuzzy set theory. Vranken et al. (2015) developed an approach to assess the economic risk of landslides (quantitatively) when the availability of information on landslide hazard is scarce. Zou et al. (2016) suggested a quantitative technique to assess risk due to debris flow on a local scale in Qipan gully (Sichuan Province, China).

Table 1 presents a list of expressions used for assessing landslide risk in previous studies. It is interesting to note that these expressions may appear similar with respect to the combination of the possibility of landslide hazard in a particular area and vulnerability of elements at risk, but the methodologies are case-dependent and differ significantly in context to certain abbreviations and the process of estimation of each component within.

3 Methodology for landslide risk assessment (LRA)

In this study, landslide risk (LR) has been assessed spatially as a function of physical vulnerability (PV) and proximity (P_{rox}) of the buildings from landslide susceptibility zonation (LSZ) class and drainage channels as given in Eq. (1) by using a semi-quantitative approach in a GIS platform.

$$\text{LR} = f\left(\text{PV}, P_{\text{rox}}\right) \tag{1}$$

Table 1 Literature review on landslide risk assessment (LRA)

S. no.	Author and year	Country	Purpose of the study	Risk formulation used in study
1	Fell (1994)	–	Overview	$R_t = \sum \left(E \times R_s\right) = \sum (E \times P \times V)$
2	Anbalagan and Singh (1996)	Kumaun Himalaya, India	Proposed a Methodology	$R = f(HP, DP)$
3	Australian Geomechanics Society (AGS) (2000)	Australia	Recommended guidelines	$R_{Prop} = P_H \times P_{S:H} \times V_{Prop:S} \times E$ $R_{DI} = P_H \times P_{S:H} \times V_{T:S} \times V_{D:T}$
4	Dai et al. (2002)	–	Framework	$R_{Prop} = P_H \times P_{S:H} \times V_{Prop:S} \times E$ $R_{DI} = P_H \times P_{S:H} \times V_{T:S} \times V_{D:T}$
5	Fabera and Stewart (2003)	–	Overview	$R = P \times C$
6	Bell and Glade (2004)	B´ıldudalur, NW-Iceland	Proposed a methodology on regional scale	$R = H \times C \times E$ $R_{ipe} = \left(H \times P_s \times P_t \times V_p \times V_{pe} \times P_{so}\right) \times E_{ipe}$ $R_{pe} = \left(H \times P_s \times P_t \times V_p \times V_{pe} \times P_{so}\right) \times E_{pe}$
7	Catani et al. (2005)	Arno River (Central Italy)	Framework—case study	$R(T, I) = D(I) \times H(T, I)$
8	Van Westen et al. (2006)	–	Literature Review	$Risk = \sum (H \sum (VA))$
9	Saldivar-Sali and Einstein (2007)	Baguio, Philippines	–	RR = HR*LUM*PM
10	Kanungo et al. (2008)	Darjeeling Himalayas, India	Proposed a semi-quantitative methodology	$LR = LP \times RDP$
11	Bednárik et al. (2012)	South-western Slovakia	Case-study	$r = h * v$
12	Martha et al. (2013)	India	Quantitative approach—Case study	Landslide risk = landslide hazard vulnerability ×value of exposed elements at risk
13	Vranken et al. (2015)	Belgium	Proposed a methodology	$R(PD) = \sum_{i=1}^{I} P_i\left(H_s\right) * \sum_{j=1}^{J} P_j(D) * V_j(P\|D) * E_j$
14	Corominas et al. (2014)	–	Recommended methodology	$R = P\left(M_i\right)P\left(X_j\|M_i\right)P\left(T\|X_j\right)V_{ij}C$
15	Uzielli et al. (2015)	Ancona Municipality	Quantitative approach—case study	$R = R_S \cdot E$, where $R_s = H \cdot V$
16	Pereira et al. (2016)	Portugal	Quantitative approach—case study	$Rb^{mi} = MaxPjyz^{mi}b \times PV \times EV$
17	Zou et al. (2016)	China	Quantitative approach—case study	$R = H \times V$
18	Abdulwahid and Pradhan (2016)	Malaysia	Semi-quantitative approach—case study	$R = H \times V$

Refer respective papers for abbreviations used in the section 'risk formulation used in the study'

Further, the landslide risk in this study is referred as overall risk (R_o), which deals with assessing the worst or higher risk class for individual building due to different landslide intensity scenarios in the whole study area. The overall risk (R_o) map represents the worst or higher risk class for each building out of all the four specific risk classes obtained for four different intensity scenarios and is estimated in two steps.

Firstly, specific risk (R_s) of the individual building exposed to different landslide intensities (low, moderate, high and very high) relating to different landslide susceptibility zonation (LSZ) classes (i.e. very low and low susceptibility class relate to low intensity (LI); moderate, high and very high susceptibility classes relate to moderate intensity (MI), high intensity (HI) and very high intensity (VHI), respectively) is assessed using Eq. (2). Hence, different specific risks for each building against different landslide intensities are obtained.

$$R_s = \mathrm{PV_i} - P_{rox} \tag{2}$$

where R_s: specific risk at different intensities (i.e. LI, MI, HI and VHI); PV_i: physical vulnerability due to a particular intensity such as LI, MI, HI and VHI; P_{rox}: proximity of the buildings from LSZ and drainage channels.

Secondly, the overall risk of a particular building is assigned as the worst or higher risk class as obtained amongst all risk classes against different intensities.

Hence, in the present study, overall risk (R_o) is determined for buildings in the study area exposed to landslides by first assessing risk at different intensities (LI, MI, HI and VHI) referred here as $(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$, respectively. The scenario-based risk at different intensity scenarios is obtained on the scale of 0 and 1. Specific risk at different intensities $(R_s)_i$ is estimated using Eq. (2). The $(R_s)_i$ thus obtained is also categorized into three risk classes as risk class I ($(R_s)_i \leq 0.5$), risk class II ($0.5 < (R_s)_i \leq 0.9$) and risk class III ($0.9 < (R_s)_i \leq 1$). In the process, four different specific risk maps ($(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$) for different intensity scenarios are prepared for all the surveyed buildings in the study area on GIS platform. Initially, the risk classes I, II and III (corresponding to low, moderate and high risk classes, respectively) of all the four risk maps were assigned numerical risk values (R_r) 1, 2 and 3, respectively. In the second step, weightages (W_i) of 1, 10, 100 and 1000 were assigned to specific risk maps corresponding to $(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$, respectively. Further, integration of these four specific risk maps has been performed in GIS using arithmetic weighted overlay technique as per Eq. (3), to generate an overall risk map of the study area.

$$(R_o)\mathrm{score} = \sum W_i \times R_r \tag{3}$$

where $(R_o)_{score}$: specific risk score; W_i: weightage assigned to each specific risk map (i.e. $(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$) corresponding to a particular landslide intensity (LI, MI, HI and VHI) and carries a value of 1, 10, 100 and 1000, respectively; R_r: rating assigned to each risk class (I, II and III) and carries a value of 1, 2 and 3, respectively.

Fell (1994) considered the acceptability of risk as an essential parameter in assessing risk. Therefore, in the present study landslide risk (i.e. overall risk, R_o) has also been classified into three risk classes as risk class I ($R_o \leq 0.5$), risk class II ($0.5 < R_o \leq 0.9$) and risk class III ($R_o \leq 1$) depending upon its severity in implying low risk, moderate risk and high risk, respectively.

The detailed methodology for the assessment of overall risk (R_o), in the study area, is presented as a flow chart in Fig. 1.

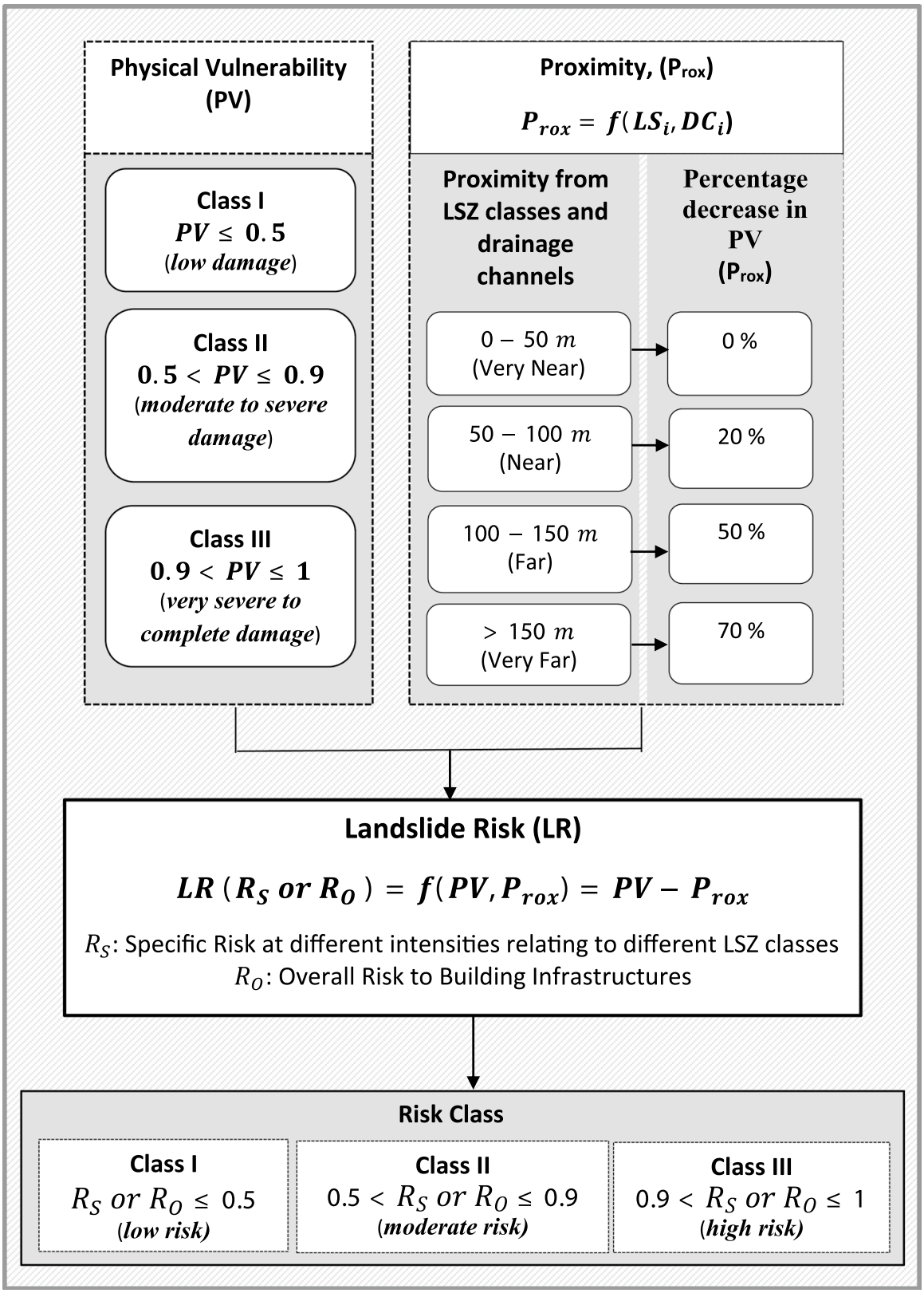


Fig. 1 Methodological flow chart to assess landslide risk (LR) of buildings exposed to landslides (where LS_i: landslide susceptibility zonation (LSZ) classes and DC_i: drainage channels feeding from a particular LSZ class)

3.1 Physical vulnerability (PV)

In the present case, physical vulnerability (PV) of the buildings is assessed by following a semi-quantitative indicator-based methodology developed for Indian scenarios by Singh et al. (2018, 2019) using Eq. (4):

$$\mathrm{PV} = f(I,R)\begin{cases} 2\frac{I_m^2}{R^2} & \frac{I_m}{R} \le 0.5 \\ 1 - \frac{2(R-I_m)^2}{R^2} & 0.5 < \frac{I_m}{R} \le 1.0 \\ 1.0 & \frac{I_m}{R} > 1 \end{cases} \tag{4}$$

where I refers to the intensity of landslide; R is the resistance of the buildings and I_m is the intensity of landslide at different degree m (m ranges from low intensity (LI) to very high intensity (VHI)).

3.2 Proximity (P_{rox})

Proximity (P_{rox}) is a relatively new factor for assessing landslide risk, which has not been taken into consideration in any of the landslide studies until now. This factor is overlooked in many published studies. Besides other researchers, Kanungo et al. (2008) suggested that vulnerability of landslide relies on various factors such as the resources at risk, their type and the proximity to the landslide. For the effective assessment of landslide risk, the details on susceptible zones and the drainage channels feeding from each susceptibility zone should also be known. The purpose of including the drainage channel feeding from landslide susceptible zone is to identify the particular regions that is also most likely be impacted by the debris flow events initiated from the landslide susceptible zones through associated drainage channels in the present case. Since the drainage channels put significant levels of stress on the environment and, in many situations, decrease the stability of surrounding slopes, a factor for proximity (P_{rox}) parameter has been introduced for assessing the landslide risk.

Therefore, proximity of buildings in this study is taken as the function of LSZ class and the drainage channel leading from that particular LSZ class as shown in Eq. (5):

$$P_{\mathrm{rox}} = f(\mathrm{LS}_i, \mathrm{DC}_i) \tag{5}$$

where P_{rox}: proximity of the buildings from LSZ class and drainage channels; LS_i: LSZ class; and DC_i: drainage channel feeding from a particular LSZ class.

Further, the drainage channel feeding from different LSZ classes may affect a building differently depending on its proximity (or position). Hence, the buildings having same physical vulnerability (PV) may pose different landslide risks due to their varying proximity to the drainage channel and different LSZ classes. Keeping in view the real field scenario and the damage patterns of infrastructures due to landslide disasters, the range of proximity of buildings from landslides presented by Singh et al. (2019) has been modified in this study. Therefore, a range of distances (very near, i.e. ≤ 50 m, to very far, i.e. ≥ 151 m) of buildings from LSZ classes and drainage channels that will affect the PV of the buildings in terms of a proximity factor (P_{rox}) has been introduced as a percentage decrease in PV as listed in Table 2.

Table 2 Percentage decrease in physical vulnerability (PV) of buildings depending upon its proximity from LSZ classes and drainage channels

S. no.	Proximity of the buildings from LSZ classes and drainage channels	% decrease in PV(P_{rox})
1	0–50 m (very near)	0
2	50–100 m (near)	20
3	100–150 m (far)	50
4	≥ 150 m (very far)	70

3.3 Landslide susceptibility zonation (LSZ)

The preparation of LSZ mapping is carried out in advance to rank the region into distinct zones of varying degrees of landslide susceptibility, derived from the influence of causative factors (such as aspect, slope, drainage, elevation, etc.) and their categories in landslide occurrences. Literature shows that LSZ mapping has been extensively studied in the past two decades and several significant principles, concepts, techniques, models and methodologies have been proposed by various engineers, scientists and researchers globally. These methods and techniques have been reviewed in detail from time to time (Soeters and Van Westen 1996; Guzzetti et al. 1999; Kanungo et al. 2009).

Therefore, in the present case, seven landslide causative factors (slope, aspect, curvature, drainage buffer, land use land cover, topographic wetness index and elevation) along with the existent landslide distribution layer were carefully prepared in GIS as thematic data layers using remote sensing image and field data interpretation to generate the LSZ map.

All the selected thematic layers were projected using UTM (Universal Transverse Mercator) system with zone 44 N. The numerical weights of different categories of thematic layers were determined using information value (InfoVal) technique (Van Westen et al. 1997; Saha et al. 2005; Sarkar et al. 2008; Oliveira et al. 2015) as shown in Eq. (6):

$$W_i = \ln\left(\frac{\text{Densclas}}{\text{Densmap}}\right) = \ln\frac{\text{Npix}(S_i)/\text{Npix}(N_i)}{\sum_{i=1}^{n}\text{Npix}(S_i)/\sum_{i=1}^{n}\text{Npix}(N_i)} \tag{6}$$

where W_i: weight of the ith class of a particular thematic data layer (e.g. convex or flat in the thematic layer 'curvature'); Densclas: landslide density within the thematic class; Densmap: landslide density within the entire thematic layer; Npix(S_i): number of landslide pixels in a certain thematic class; Npix(N_i): total number of pixels in a certain thematic class; n: number of classes in a thematic map.

The weights are then designated to the classes of individual thematic data layer, in order to generate weighted thematic maps, which are further overlaid and added (numerically) as per Eq. (7) to develop a Landslide Susceptibility Index (LSI) map. The LSI values represent the relative susceptibility to landslide occurrences (i.e. the higher the LSI value, more will be the susceptibility of landslide events) in the study area (Balamurugan et al. 2016):

$$\text{LSI} = \text{Sl} + \text{As} + \text{TWI} + \text{Db} + \text{El} + \text{LULC} + \text{Cu} \tag{7}$$

where Sl, As, TWI, Db, El, LULC, Cu refer to slope, aspect, topographic wetness index, lithology, drainage buffer, elevation, land use and land cover and curvature, respectively.

The LSI values thus obtained are later divided into five different landslide susceptibility zones (i.e. very low susceptibility (VLS), low susceptibility (LS), moderate susceptibility (MS), high susceptibility (HS) and very high susceptibility (VHS)) to prepare LSZ map of

the study area by employing success rate curve method. Success rate is expressed as the percentage of landslide occurrence in any susceptibility zone. It is one of the novel ways of segmenting the range of LSI values (Kanungo et al. 2006; Ghosh et al. 2009; Kannan et al. 2015). This method has been selected in order to ascertain the LSZ boundaries accurately and also to dodge the uncertainties (if any, in the segmentation process).

4 A case study

4.1 Study area

Gopeshwar Township of Chamoli District in the Garhwal Himalayas of Uttarakhand state has been selected as the test site for implementation of the proposed methodology, as shown in Fig. 2. The Chamoli District is in the north-eastern part of Uttarakhand state and covers an area of 8030 km^2 (https://chamoli.gov.in). It is the second largest district of Uttarakhand state and lies within the latitudes 29°55′00″ N–31°03′45″ N and longitude 79°02′39″ E–80°03′29″ E. National Highway (NH) 58 is the only lifeline that connects the study area to the township, and the route is very important since it connects to famous pilgrim destinations like Joshimath, Badrinath and Kedarnath. Chamoli District lies in the Lesser Himalayas and falls under seismic zone IV (as per IS 1893 2016). Alaknanda River flowing towards south-east is the main source of drainage in the Gopeshwar Township of Chamoli District. The area under study, situated in the upper Alaknanda valley, is highly susceptible specifically to debris slides and debris flow type of landslides (Sajwan and Sushil 2016).

Gopeshwar Township is densely populated as it is the district headquarter of Chamoli. This township is well known for its fragile landscape. Hence, it comes under the threat of heavy rainfall, consequently leading to the occurrence of landslides. According to the

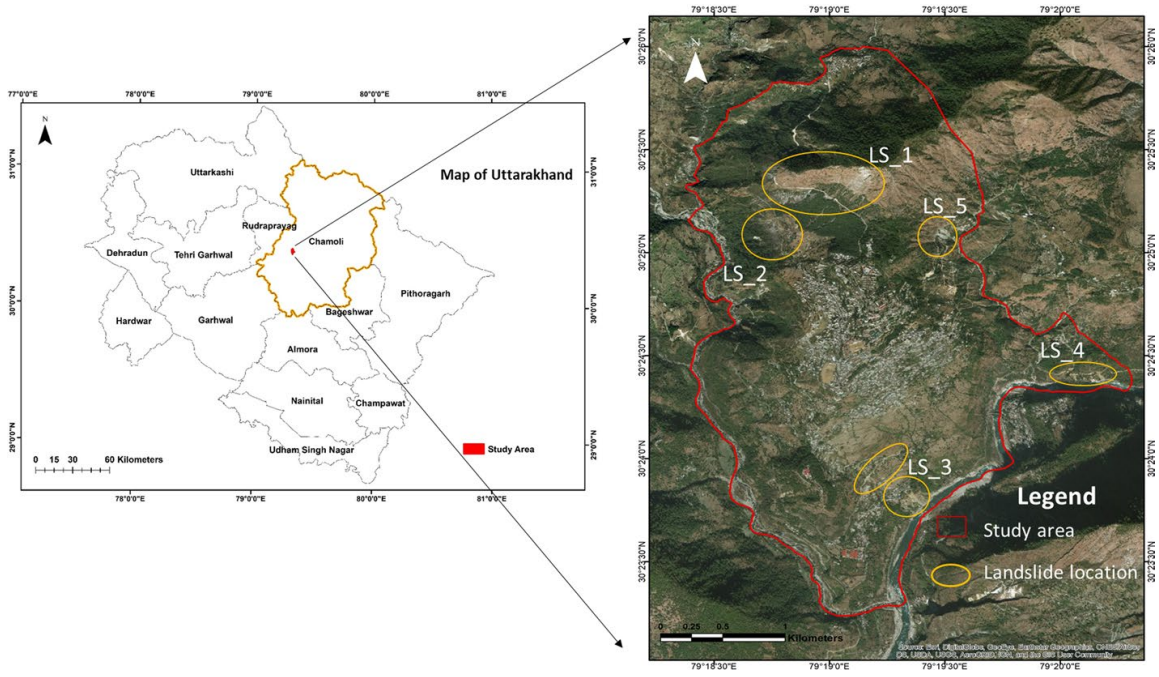


Fig. 2 Study area of Gopeshwar township in Chamoli District of Uttarakhand State (India) showing locations of landslides (referred as LS_1, LS_2, LS_3, LS_4 and LS_5)

municipality, Gopeshwar Township has been divided into eleven wards, of which some wards are affected by landslides (specifically by debris slides and flows).

4.2 Data sources

In this present study, remote sensing data such as P5 PAN CARTODEM with 2.5 m resolution (acquired from NRSA-National Remote Sensing Agency, Hyderabad, India), Google Earth (GE) images and tools, online maps provided by ESRI on GIS platform (i.e. the basemap Imagery on ArcMap), published literature (Saha et al. 2005; Chauhan et al. 2010; chamoli.gov.in; http://uttarakhandtourism.gov.in/gopeshwar) on the study area and field-based information (collected during the years 2016–2018) including interviews of local people have been utilized to extract pertinent data and evidences for the generation of landslide inventories, thematic data layers, building inventories database, etc. These data have been further utilized to achieve different objectives of the present research.

4.3 Landslides and building infrastructures in study area

In the present study, landslides are mapped using Google Earth (GE) images and base map in ArcGIS (a collection of ortho-rectified satellite and aerial imageries in natural colour, at 1 m or less resolution). Moreover, the 'historical imagery' feature allows to view old satellite images to gain knowledge on former landslide and detect changes in different time frames. Since scarcity of specific data regarding the occurrence of landslides has always been a setback in the field of landslide studies in India, the 'historical imagery' feature has been very crucial in this study. The five major identified landslide locations in the study area are referred here as LS_1, LS_2, LS_3, LS_4 and LS_5 as shown in Fig. 2. LS_1 is a rock fall type of potential failure, while other locations LS_2, LS_3, LS_4 and LS_5 are debris flow types of potential failures. The landslides identified in Fig. 2, were also analysed via 'history imagery tool'. From the GE 'historical imagery tool,' it was observed that most of the landslides such as LS_1, LS_2, LS_4 and LS_5 are old landslides and their existence can be traced even before 2006, as shown in Fig. 3. However, LS_3 (Fig. 3c) was initiated soon after the 2013 flash flood that took place between June 15 and 17 in Kedarnath (Uttarakhand). The landslides identified on GE are further aided by extensive field survey in the study area. The smallest area of the landslide is about 119 m^2 and that of the largest is about 101,272 m^2. Therefore, the landslide and other related features observed on GE could be easily digitized and exported to excel and ArcGIS.

Landslide distribution layer, commonly known as the landslide inventory, is the compilation of characteristics and location of landslide events in the study area in the past. In the present study, the landslides identified on GE image are first digitized as polygons and saved in KML format, which is then exported to ArcGIS using 'KML To Layer' (conversion) tool. The exported landslide polygon layer thus obtained is represented in a single vector layer. The landslide polygons are reclassified and assigned the values 1. Therefore, 1 value will represent the presence of landslides and 0 value will represent the absence of landslide. This layer is then converted to a rasterized landslide distribution layer with a pixel size of 10 m × 10 m as shown in Fig. 4.

The eleven wards of Gopeshwar Township are spread in an area of about 14.08 km^2 and hold a total population of 21,444 as per report released by Census of India 2011 and Nagar Palika Parishad, Gopeshwar. Further, the township has noticed an exceptional growth in the last decade in its urban population. As per the Census of India (2011), the township

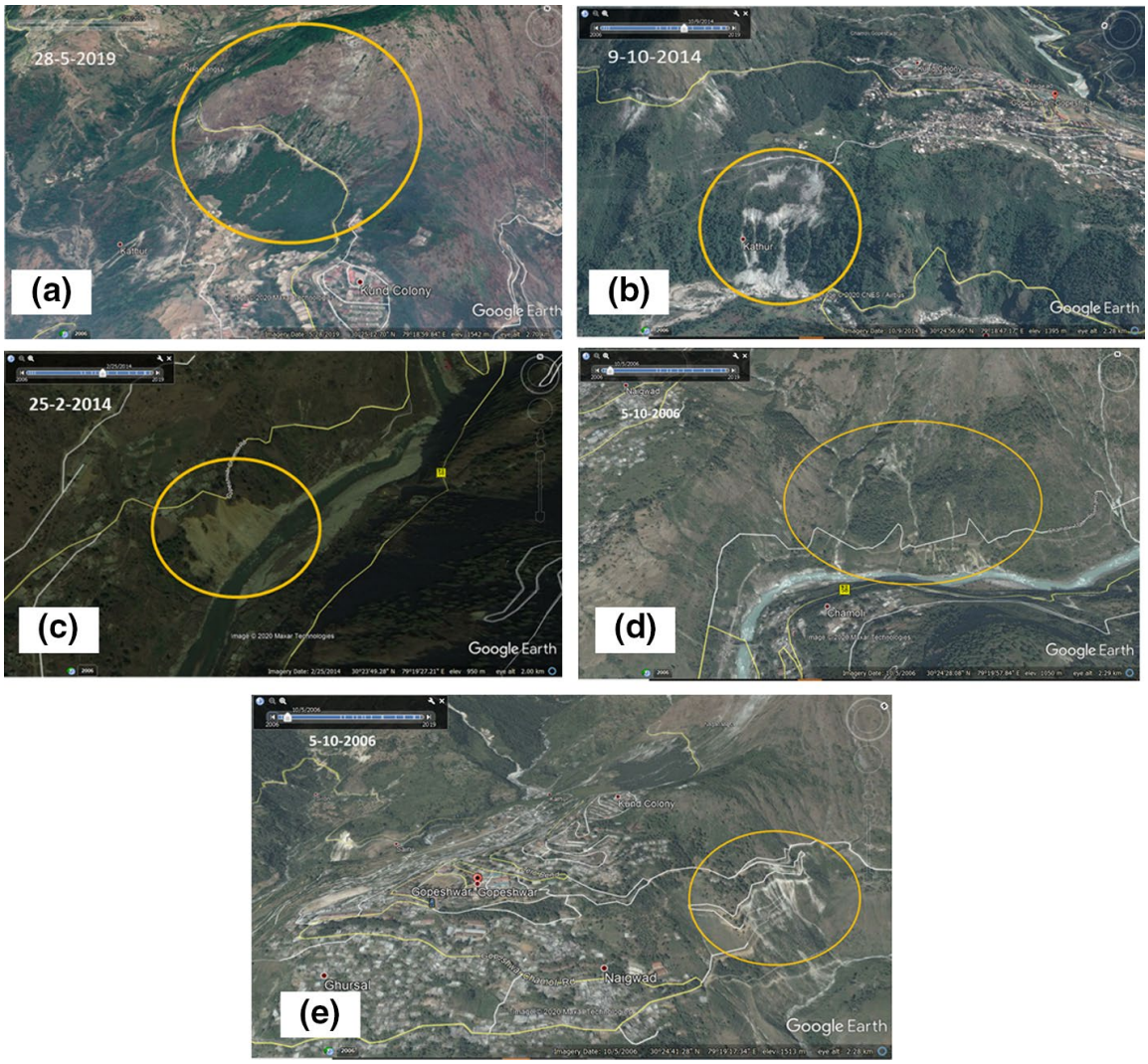


Fig. 3 Google Earth imageries showing the dates of occurrences of **a** LS_1, **b** LS_2, **c** LS_3, **d** LS_4, and **e** LS_5

population escalated in the period 1991–2001 by 28.96% at a rate of 2.89% per annum. However, during 2001–2011 the population soared by 81.3% at the rate of 8.13% per annum.

Table 3 presents a ward-wise distribution of the population, geographical area and number of existing buildings in Gopeshwar Township according to Nagar Palika Parishad (Gopeshwar). As observed (Table 3), the overall average density of buildings in the township is about 379 buildings per km^2. It is also noticed that the ward no. 5 (i.e. Lower Bazar) has the highest density of 2028 buildings per km^2 and a total of 6 wards (i.e. ward nos. 3–8) out of 11 Wards hold more than the average density of buildings in the township. The buildings in the township have been built in clusters haphazardly and without any town planning scheme (Fig. 5). Therefore, it can be perceived from Fig. 5 that majority of buildings in the hilly area are non-engineered, and low in the construction quality. Different typologies of the buildings as observed in the study area are presented in Fig. 6.

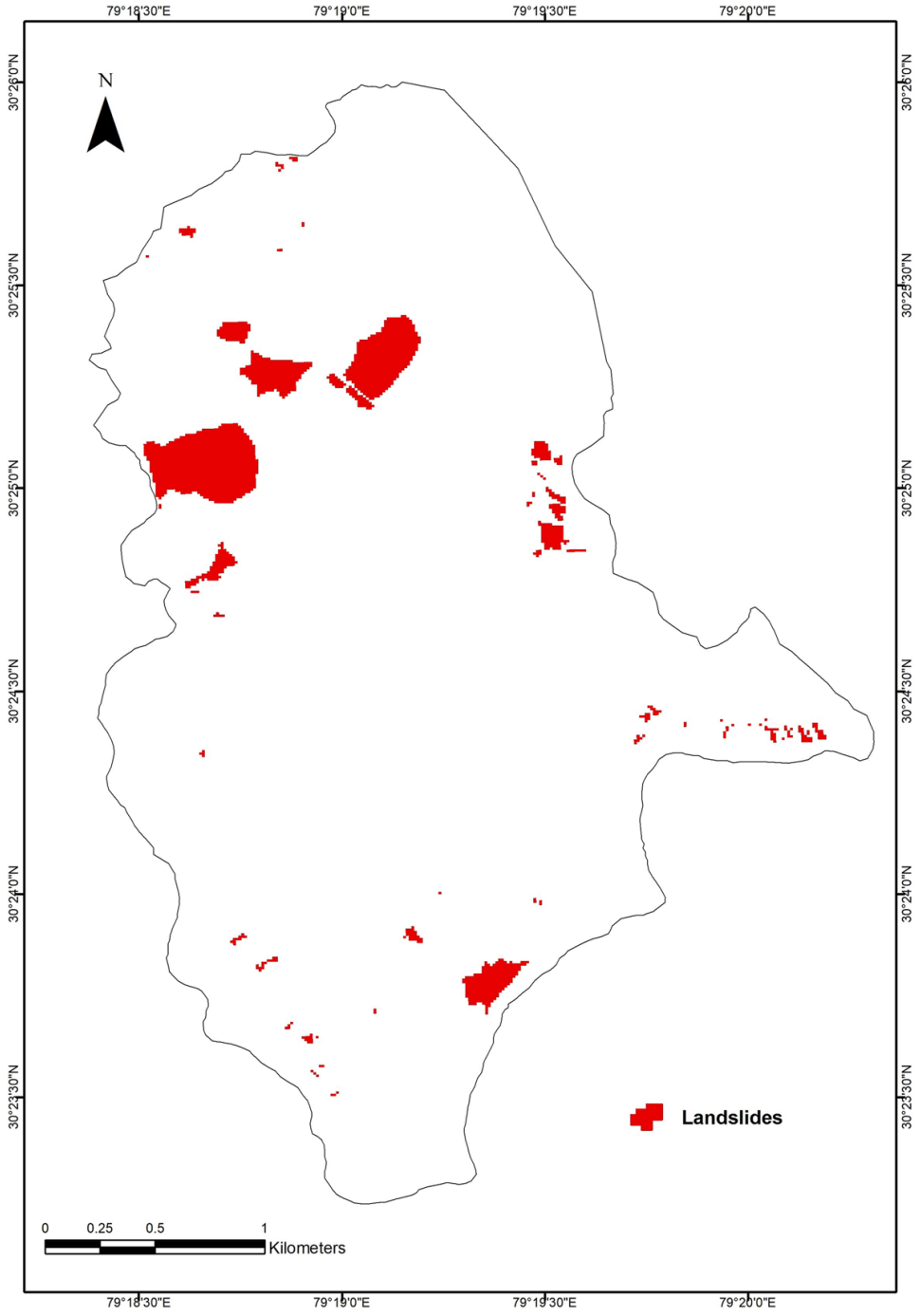


Fig. 4 Landslide distribution layer of the study area

4.4 Landslide risk assessment (LRA) in the study area

The input parameters such as landslide susceptibility zonation (LSZ), proximity (P_{rox}) of the buildings from the LSZ class and drainage channels and physical vulnerability (PV) of buildings to assess the landslide risk as obtained in the study area are discussed in the following sections.

Table 3 Ward-wise details of population, area and number of buildings in Gopeshwar Township. Source: Nagar Palika Parishad, Gopeshwar

Ward number	Ward name	Population	Area (km^2) (a)	Number of buildings (b)	Building density (per km^2) (b/a)
1	Gangol Gaon	1105	1.18	241	204
2	Gopeshwar Gaon	2173	3.28	592	181
3	Kund Colony	2074	0.76	516	679
4	Uppar Bazar	2397	0.3	596	784
5	Lower Bazar	2365	0.29	588	2028
6	Vasant Vihar	1853	0.69	452	655
7	Subhash Nagar	1850	0.7	469	670
8	Naigwaad	2466	1.04	613	589
9	Paduli-Papdiyana	2332	2.82	580	206
10	Kothiyal Sain	1847	1.72	459	267
11	Chamoli Kshetrapal	982	1.3	227	175
Total		21,444	14.08	5333	379



Fig. 5 Field picture showing buildings built in clusters haphazardly and without any town planning scheme

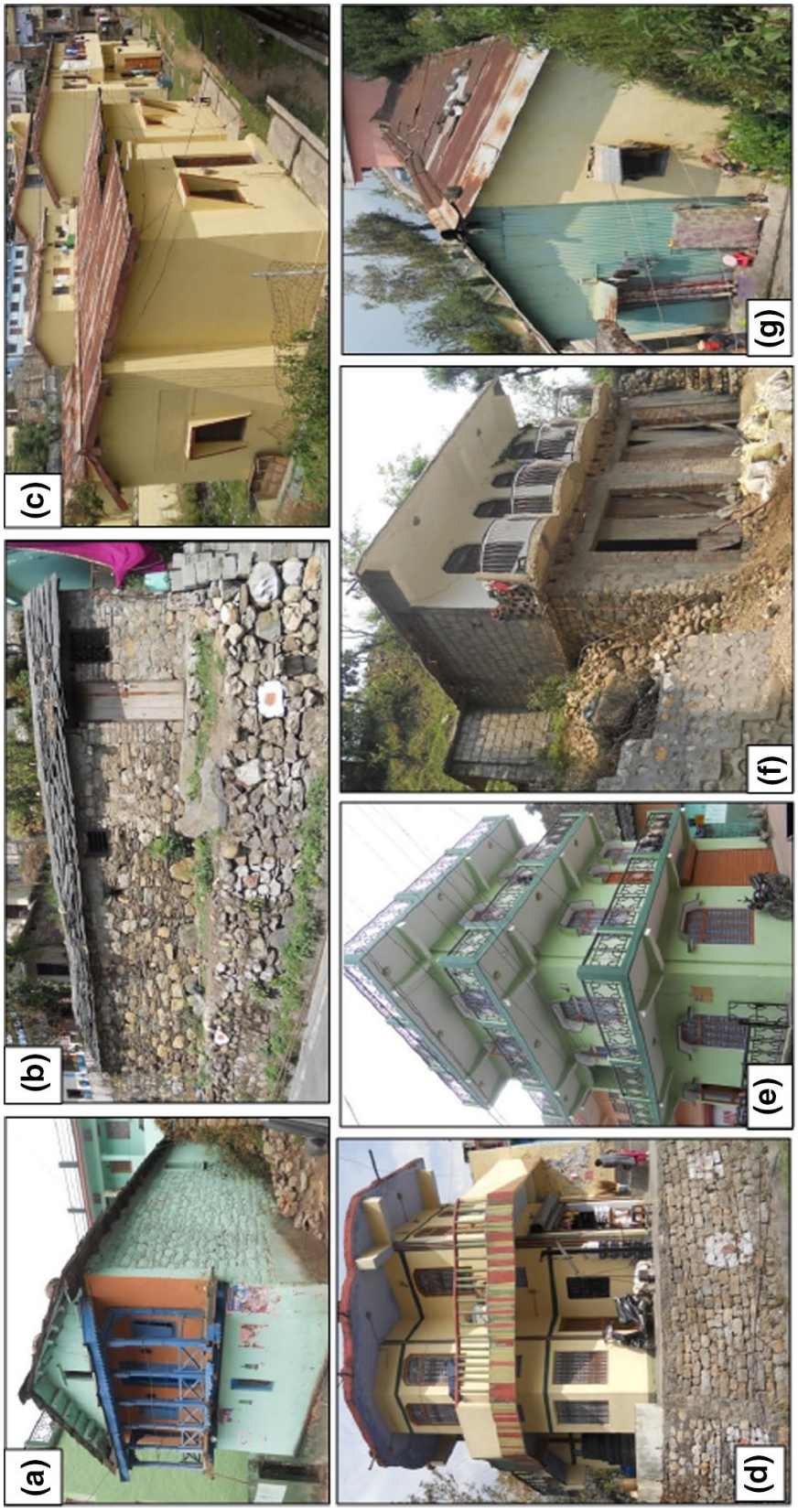


Fig. 6 Typical examples of buildings in study area showing **a** double storey stone masonry building with wooden pitched roof; **b** stone masonry building with thin layer of slate roof; **c** double and single storey buildings with pitched roof of GI sheets; **d** double storey irregular shaped building having flat reinforced brick roof; **e** triple storey RC building; **f** stone masonry typology on ground floor and brick construction on top; **g** partial construction of stone masonry and GI sheets

4.4.1 Landslide susceptibility zonation (LSZ) mapping

Since the fundamental necessity for LSZ studies is to determine the relative importance of thematic data layers and their categories for landslide occurrences in terms of weights and ratings, respectively, in the present case, seven thematic data layers (slope, aspect, drainage buffer, curvature, land use land cover, TWI and elevation) corresponding to different causative factors responsible for the occurrence of landslides along with the existing landslide distribution layer (Fig. 4) were carefully prepared in GIS. A brief description of the seven causative factors considered for the preparation of LSZ map of the study area is given as follows:

(i) *Land use land cover* (*LULC*) Eleven prominent LULC sub-categories in the study area that may have impacted the occurrence of landslide are identified and digitized. The sub-categories of LULC data layer considered are thick vegetation, moderate vegetation, sparse vegetation, agriculture land, barren land, built-up area, river sand and river.

(ii) *Slope (Sl)* Slope is the angle formed between any part of the earth's surface and a horizontal datum. Theoretically, as the slope angle increases, shear stress is escalated, thereby increasing the chance of failure. The slope map of the study area extracted through DEM varies from 0° to 90° and classified into five classes as 0°–15°, 15°–25°, 25°–35°, 35°–45° and > 45° according to classification given in other studies (Anbalagan 1992; Gupta and Joshi 1990; Dhakal et al. 2000; Kanungo et al. 2009).

(iii) *Aspect (As)* It can be designated as the second derivative of DEM (digital elevation model) and is related to the distribution and density of vegetation on mountainous slopes. The aspect map generated from the DEM shows slope directions from 0° to 360° with respect to the north. Further, as per the classification given in other studies (Sarkar and Kanungo 2004; Saha et al. 2005; Pham et al. 2017), it is grouped into nine categories as Flat, N, NE, E, SE, S, SW, W and NW.

(iv) *Elevation (El)* It is the deviation of maximum and minimum heights within terrains and is considered as an important causative factor, since elevation not only affects the slope material by weathering process but also controls runoff direction and rate of drainage density (Ahmed 2009). As per the classification given in other studies (Balamurugan et al. 2016), it is classified into five classes as 891–1124 m, 1124–1358 m, 1358–1592 m, 1592–1826 m and 1826–2062 m.

(v) *Curvature* (*Cu*) It presents the convergence or divergence of flow of surface water on hill slopes and are referred as convex (i.e. positive curvature), concave (i.e. negative curvature) and flat (i.e. zero curvature). Stocking (1972) stated that convex slopes are more stable in comparison with concave slopes because of uniform dispersal of runoff down the slope.

(vi) *Drainage Buffer (Db)* Erosional activity on hill slopes is generally associated with surface drainages and thereby initiates most of the landslides. The drainage lines extracted from DEM have been classified according to their orders on the basis of Strahler classification (Strahler 1964). The drainage lines thus obtained is buffered at an interval of 25 m to determine the influence of drainage in the occurrence of landslide activities. Therefore, drainage buffer has been classified into nine categories as 0–25 m, 25–50 m, 50–75 m, 75–100 m, 100–125 m, 125–150 m, 125–150 m, 150–175 m and 175–200 m.

(vii) *Topographic Wetness Index* (*TWI*) According to Moore et al. (1991), TWI is an indicator of the availability of soil moisture at a site and reflects the role of water in triggering instability in a given catchment (or area of interest). The TWI data layer has also been considered in other studies such as Guri et al. (2015), Sema et al. (2017) and Basu and Pal (2018). The TWI data layer has been classified into four categories in this study as: 0.68–3.9, 3.9–7.09, 7.09–10.3, 10.3–13.5.

As discussed in Sect. 3.3, the ratings for distinct categories of seven thematic data layers were obtained by using InfoVal technique. The landslide distribution layer and the seven thematic data layers were rasterized at 10 m × 10 m pixel size using spatial analysis tool of GIS. The total number of pixels in the study area is 90,559. The number of pixels occupied by individual category of all the thematic data layers was calculated in GIS. Then, individual thematic layer was integrated with the landslide layer to extract the number of landslide pixels in individual category of thematic layer. Using these derived data, the information value (denoted by Wi) for all the categories is obtained from Eq. (6) and is listed in Table 4. The categories of the seven thematic data layers having higher information values indicate greater contribution towards landslide occurrences in the study area.

As observed from Table 4, barren land, agricultural land and sparse vegetation have higher information value in comparison with other LULC class, implying more proneness to landslide activities. Higher information values have been observed in moderate to steep slopes than the gentle slopes. Sarkar and Gupta (2005) and Kanungo et al. (2006) have also reported this result in other parts of Indian Himalayas. The slope instability is indirectly influenced by slope aspect. Sinha et al. (1975) reported that south-facing slopes are more prone to erosional activity because of lesser vegetation density as compared to north-facing slopes. Some other works (Lin and Tung 2003; Sarkar et al. 2008) also concluded that south and south-east facing slopes are more prone to landslide activities. The present study also shows similar results that the south and south-east facing slopes have higher information values. In case of elevation, higher information value is obtained in category 1124–1358 m and 1358–1592 m. The landslide activities in the study area are more prominent at moderate elevations as compared to ridge tops and valley areas. Further, maximum information value is obtained for concave slopes indicating more susceptibility towards landslide activities as also reported by Stocking (1972). It is also observed from Table 4 that the maximum information value for drainage buffer is observed for 0–25 m zone, which indicates that the regions closer to drainage channels are more prone to landslides. In case of TWI layer, it has been observed in other studies (Sema et al. 2017; Wang et al. 2015) that more landslides activities have occurred in areas having lesser TWI values. On similar lines, in the present study maximum information value is exhibited by TWI categories 0.68–3.9 and 3.9–7.09, indicating their more proneness to landslide occurrences. Thus, out of 42 categories belonging to seven causative factors, it can be inferred that barren land, 35°–45° slope and southeast aspect categories seem to be the top three controlling factors for landslide occurrences in the study area.

As discussed in Sect. 3.3, the Landslide Susceptibility Index (LSI) map of the study area is prepared by integrating the information values corresponding to the categories of respective thematic layers for each pixel as per Eq. (7) using spatial analyst tool in GIS. The range of LSI values as obtained for the whole study area varied from 3.50 to − 12.0. In order to generate susceptibility zones, LSI values are segmented by employing success rate curve method. The observed mean (μ_0) and standard deviation (σ_0) from the probability distribution curve of the range of LSI values (3.50 to − 12.0) have been determined as − 2.29 and 3.12, respectively. The LSI values have been categorized into 5 distinct susceptibility

Table 4 Computed weights for classes of various thematic data layers based on landslide occurrences

Factors	Categories	Number of landslide pixel (Npix (*Si*))	Total number of pixel in thematic class (Npix (*Ni*))	*Wi* (Eq. 6)
Land use land cover (LULC)	Thick vegetation	20	883	−0.30
	Moderate vegetation	110	27,773	−1.49
	Sparse vegetation	275	11,120	0.17
	Agricultural land	372	13,190	0.30
	Barren land	1271	9259	1.72
	Built-up area	25	5763	−1.95
	River sand	2	11,219	−5.14
	Alaknanda River	0	11,352	0
Slope (Sl)	0°–15°	178	18,351	−1.14
	15°–25°	532	26,946	−0.43
	25°–35°	834	24,808	0.10
	35°–45°	479	5853	0.99
	> 45°	737	14,601	0.50
Aspect (As)	Flat (undefined)	1	1532	−3.84
	N (0–22.5, 337.5–360)	3	2789	−3.34
	NE (22.5–67.5)	10	1871	−1.74
	E (67.5–112.5)	152	14,991	−1.10
	SE (112.5–157.5)	374	18,093	0.46
	S (157.5–202.5)	532	13,577	0.25
	SW (202.5–247.5)	871	12,154	0.01
	W (247.5–292.5)	668	16,304	0.30
	NW (292.5–337.5)	149	9248	−0.64

Table 4 (continued)

Factors	Categories	Number of landslide pixel (Npix (*Si*))	Total number of pixel in thematic class (Npix (*Ni*))	*Wi* (Eq. 6)
Elevation (El)	891–1124 m	478	25,069	−0.47
	1124–1358 m	1039	27,130	0.23
	1358–1592 m	742	21,655	0.12
	1592–1826 m	346	10,570	0.07
	1826–2062 m	155	6135	−0.19
Curvature (Cu)	Convex	1284	42,349	−0.01
	Flat	163	6542	−0.20
	Concave	1313	41,668	0.03
Drainage Buffer (Db)	0–25 m	1572	45,767	0.12
	25–50 m	642	19,723	0.07
	50–75 m	332	13,358	−0.20
	75–100 m	147	6533	−0.30
	100–125 m	67	4323	−0.68
	125–150 m	0	778	0.00
	150–175 m	0	68	0.00
	175–200 m	0	9	0.00
Topographic Wetness Index (TWI)	0.68–3.9	850	24,713	0.12
	3.9–7.09	272	8327	0.07
	7.09–10.3	852	26,565	0.05
	10.3–13.5	786	30,954	−0.18

zones (i.e. VLS, LS, MS, HS and VHS) with boundaries at ($\mu_0 - 1.5\ m\sigma_0$), ($\mu_0 - 0.5\ m\sigma_0$), ($\mu_0 + 0.5\ m\sigma_0$) and ($\mu_0 + 1.5\ m\sigma_0$), where m is a positive, nonzero value. Thus, the cumulative percentage of landslide occurrences in different susceptibility zones ranging from VHS to VLS has been marked against the cumulative percentage of area of the susceptibility zones for LSZ maps for distinct values of m. Thus, five representative success rate curves corresponding to $m = 1.1$, 1, 0.9, 0.8 and 0.7 have been produced (Fig. 7) to check the suitability of the LSZ map, since greater percentage of landslides must happen in VHS zone in comparison with other susceptibility zones.

It has been observed that for 10% of the area of VHS zone, the curves corresponding to $m = 1.1$, 1, 0.9, 0.8 and 0.7 show the landslide occurrences of 56%, 56.2%, 55.3%, 52% and 50%, respectively. Hence, the highest success rate is indicated for the curve corresponding to m equal to 1 for the first 10% area of VHS zone. Further, the accuracies of the LSZ maps corresponding to different m values are assessed by area under the curve (AUC) method. In the present study, AUC is achieved graphically on the basis of the relationship between the LSZ areas and landslides in cumulative percentages (Fig. 7). The AUC values as achieved corresponding to $m = 1.1$, 1, 0.9, 0.8 and 0.7 are 0.971, 0.976, 0.963, 0.944, 0.896, respectively. Hence, it can be inferred that the maximum AUC of 97.6% is observed for the LSZ map corresponding to $m = 1$ and thus reveals to be the most suitable LSZ map for the study area.

Consequently, with $m = 1$, boundary of five different landslide susceptibility zones has been established with LSI values of −6.28, −3.71, −1.10 and 0.79. Thus, the LSZ map has been prepared with class intervals at −12 to −6.28, −6.28 to −3.71, −3.71 to −1.10, −1.10 to 0.79 and 0.79 to 3.46 which are designated as VLS, LS, MS, HS and VHS zones, respectively. The LSZ map thus prepared is shown in Fig. 8. The area covered by five different landslide susceptibility zones and the landslide distribution in different susceptibility

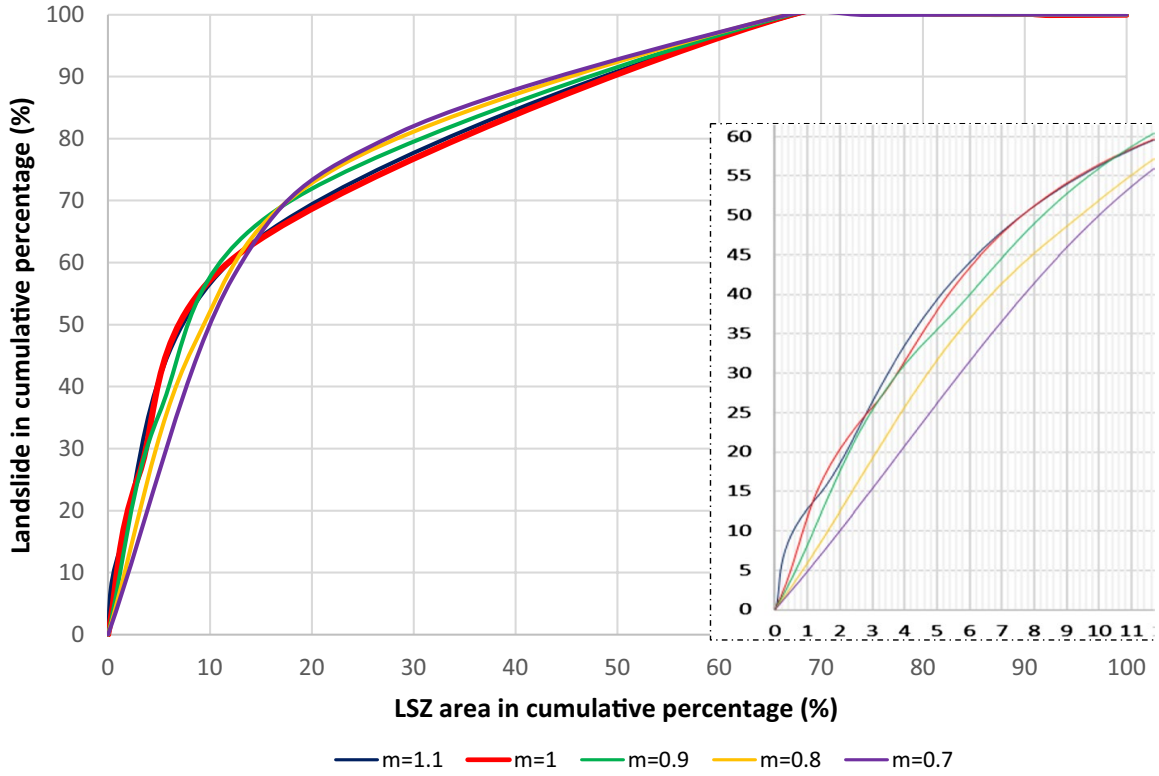


Fig. 7 Success rate curves for choosing the best segmentation in LSI values for LSZ classification

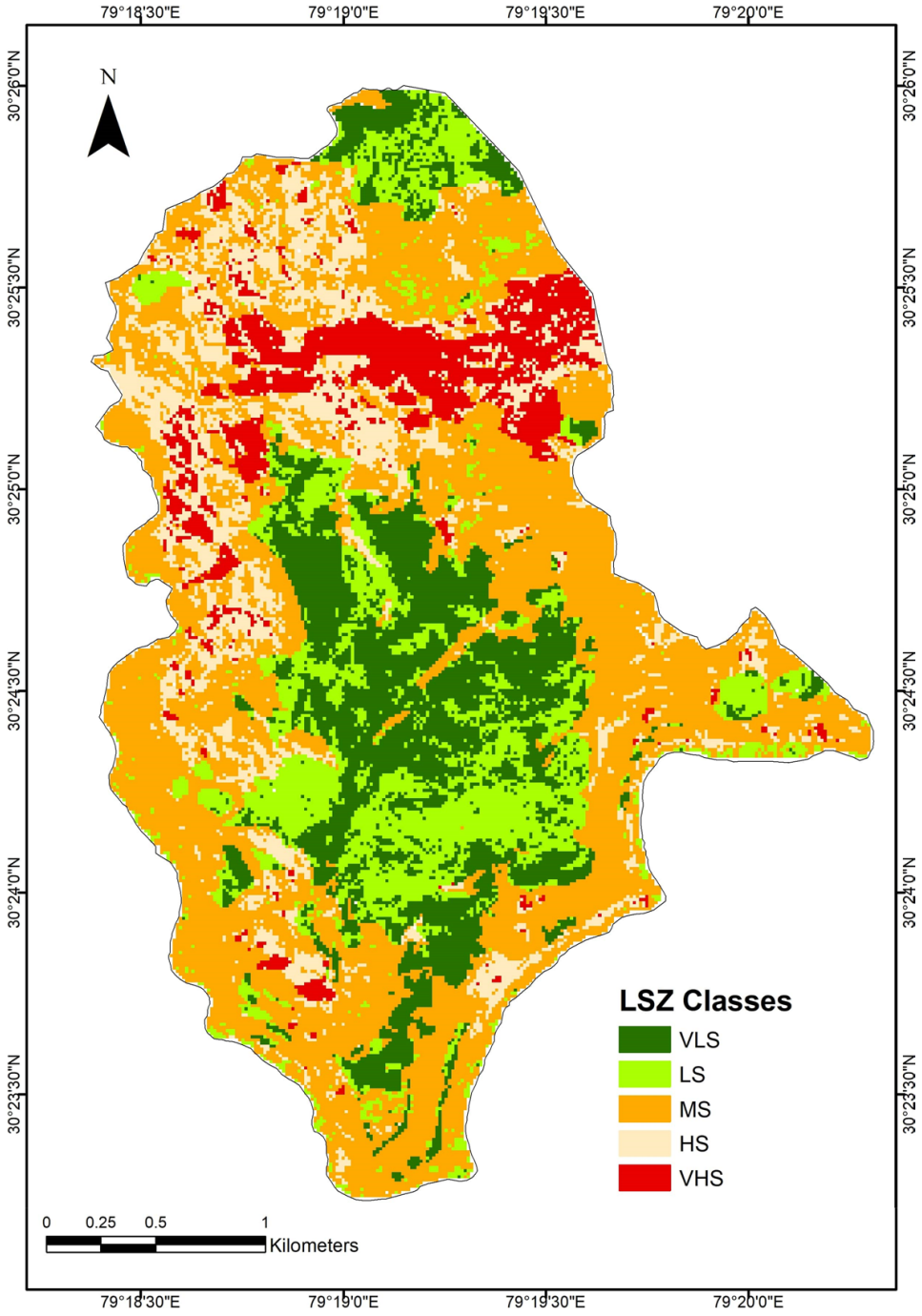


Fig. 8 Landslide susceptibility zonation (LSZ) map using segmentation of LSI values by success rate curve method in information value approach

Table 5 Distribution of existing landslides in different landslide susceptibility zones (LSZ)

Landslide susceptibility zones (LSZ)	Area (km^2)	Percent area (%) (*a*)	Landslide area per class (km^2)	Percent landslide area per class (%) (*b*)	Landslide density (*b*/*a*)
VLS	0.67	7.949	0.000	0.05	0.007
LS	2.15	25.675	0.000	0.09	0.004
MS	4.40	52.445	0.121	37.6	0.714
HS	0.89	10.604	0.118	36.6	3.456
VHS	0.28	3.328	0.083	25.7	7.733

zones have also been determined (Table 5). Visual investigation of LSZ map suggests that all five zones are distributed all over the study area.

From Table 5, it can be inferred that the VHS zone occupies 3.3% of the study area, whereas the HS, MS, LS and VLS zones occupy 10.6%, 52.45%, 25.7% and 7.9% of the study area, respectively. This shows that the susceptible zones are normally distributed area-wise, which ought to be the case. Further, the landslide occurrences in different susceptibility zones have also been correlated. It has been found that 25.7% of existing landslide area occurs in VHS zone, while 36.6%, 37.6%, 0.09% and 0.05% of landslide area fall in HS, MS, LS and VLS zones, respectively. Hence, it can be stated that 13.9% of HS and VHS areas together contain 62.3% of existing landslide areas. It is further observed from Table 5 that the landslide density increases from lower landslide susceptibility zones to higher susceptibility zones and is comparatively much higher for HS and VHS zones (i.e. 3.456 and 7.733, respectively).

4.4.2 Physical vulnerability assessment (PVA) of buildings

Physical vulnerability assessment (PVA) of the buildings in the study area is performed in accordance with the methodology published in Singh et al. (2019) by taking into account the resistance (R) of the buildings and the intensity (I) of the landslide.

4.4.2.1 Resistance of buildings in the study area During the field investigation, it was perceived that some of the wards in the study area were bigger in area than others. Hence, the cluster of buildings surveyed in each ward of the study area is not uniform. Maximum number of buildings was surveyed in ward number 6 (i.e. 75 buildings), followed by ward number 9 (i.e. 70 buildings), while the minimum being surveyed in ward number 11 (i.e. 17 buildings). However, a total of 45 buildings each were surveyed in ward numbers 1, 4, 7 and 8, and 50 buildings each were surveyed in ward numbers 2, 3, 5 and 10.

The resistance of the 542 surveyed buildings in the test site was calculated using the resistance factors assigned to different indicators for each of the buildings in the equation given in Singh et al. (2019) (refer to Eq. 5 of Sect. 3.5.2 in Singh et al. 2019) and is summarized in Table 6. Further, resistance (R) abilities of the buildings have been classified into three grades as: low resistance ($R \leq 0.5$), moderate resistance ($0.5 < R \leq 0.9$) and high resistance ($0.9 < R \leq 1$). The percentage distribution of resistance classes of all the surveyed buildings in the study area is displayed in Fig. 9 in ward-wise manner.

From Table 6 and Fig. 9, it can be heeded that none of the buildings in the test site exhibit high resistance. Low resistance of buildings with maximum and minimum percentages is observed in ward number 11 (Chamoli Kshetrapal) and ward number 6 (Vasant Vihar), respectively, in comparison with other wards in the study area. The maximum and minimum percentages for moderate resistance of buildings are observed in ward number 6 (Vasant Vihar) and ward number 11 (Chamoli Kshetrapal), respectively. Therefore, it can be inferred that the buildings in ward number 6 offer better resistance in comparison with the buildings in other wards. One of the prominent reasons being that ward number 6 has residents of higher income group and also in this ward the buildings are constructed in recent times.

4.4.2.2 Landslide intensity (I) In this study, the landslide intensity scenario with low intensity (LI), moderate intensity (MI), high intensity (HI) and very high intensity (VHI) as proposed by Singh et al. (2018, 2019) has been adopted. In the present context, different

Table 6 Resistance (R) of the buildings in the study area

No. of Buildings Surveyed	Ward Number										
	1	2	3	4	5	6	7	8	9	10	11
	Resistance (R) of the building										
1	0.58	0.50	0.70	0.61	0.48	0.59	0.64	0.57	0.54	0.54	0.49
2	0.47	0.62	0.69	0.52	0.52	0.75	0.58	0.62	0.55	0.53	0.55
3	0.53	0.57	0.50	0.54	0.58	0.60	0.60	0.64	0.56	0.73	0.55
4	0.50	0.63	0.50	0.58	0.54	0.61	0.70	0.54	0.49	0.50	0.49
5	0.48	0.47	0.69	0.72	0.59	0.60	0.67	0.50	0.56	0.70	0.58
6	0.57	0.64	0.59	0.57	0.49	0.57	0.50	0.70	0.61	0.50	0.49
7	0.49	0.55	0.61	0.57	0.48	0.62	0.50	0.67	0.50	0.49	0.49
8	0.50	0.65	0.54	0.57	0.50	0.55	0.48	0.48	0.50	0.50	0.55
9	0.53	0.61	0.50	0.64	0.50	0.55	0.48	0.49	0.50	0.43	0.55
10	0.64	0.65	0.56	0.50	0.62	0.61	0.48	0.49	0.49	0.50	0.58
11	0.60	0.64	0.72	0.55	0.62	0.58	0.61	0.50	0.50	0.54	0.49
12	0.56	0.47	0.72	0.58	0.58	0.58	0.68	0.50	0.50	0.50	0.54
13	0.44	0.61	0.59	0.64	0.44	0.57	0.66	0.50	0.50	0.50	0.49
14	0.47	0.56	0.49	0.63	0.47	0.59	0.60	0.57	0.63	0.55	0.58
15	0.59	0.68	0.49	0.50	0.50	0.60	0.50	0.66	0.60	0.50	0.49
16	0.64	0.69	0.50	0.60	0.50	0.58	0.72	0.62	0.47	0.64	0.49
17	0.67	0.69	0.59	0.50	0.50	0.70	0.48	0.62	0.50	0.60	0.49
18	0.50	0.47	0.64	0.58	0.62	0.68	0.50	0.62	0.64	0.62	-
19	0.48	0.45	0.64	0.60	0.58	0.70	0.50	0.62	0.62	0.64	-
20	0.48	0.45	0.69	0.46	0.59	0.70	0.48	0.62	0.61	0.58	-
21	0.55	0.47	0.69	0.50	0.62	0.50	0.48	0.50	0.58	0.65	-
22	0.54	0.63	0.69	0.54	0.59	0.50	0.61	0.50	0.65	0.69	-
23	0.50	0.64	0.69	0.57	0.62	0.55	0.53	0.61	0.61	0.65	-
24	0.65	0.57	0.49	0.57	0.50	0.58	0.50	0.57	0.58	0.50	-
25	0.48	0.6	0.72	0.50	0.59	0.58	0.53	0.67	0.57	0.50	-
26	0.50	0.64	0.60	0.55	0.64	0.50	0.50	0.48	0.63	0.58	-
27	0.57	0.58	0.60	0.55	0.64	0.60	0.53	0.50	0.63	0.60	-
28	0.48	0.55	0.60	0.55	0.64	0.57	0.60	0.54	0.50	0.50	-
29	0.50	0.45	0.58	0.50	0.59	0.54	0.56	0.60	0.50	0.55	-
30	0.59	0.63	0.56	0.50	0.67	0.54	0.64	0.59	0.63	0.62	-
31	0.53	0.47	0.56	0.63	0.54	0.62	0.70	0.64	0.50	0.50	-
32	0.55	0.47	0.50	0.50	0.59	0.55	0.67	0.54	0.50	0.60	-
33	0.48	0.47	0.50	0.60	0.49	0.55	0.50	0.50	0.49	0.58	-
34	0.47	0.50	0.60	0.50	0.48	0.61	0.50	0.70	0.49	0.62	-
35	0.50	0.57	0.69	0.58	0.50	0.58	0.48	0.67	0.50	0.50	-
36	0.57	0.55	0.64	0.60	0.50	0.58	0.48	0.48	0.58	0.54	-
37	0.50	0.55	0.69	0.46	0.62	0.57	0.48	0.49	0.63	0.50	-

Table 6 (continued)

38	0.54	0.61	0.59	0.50	0.62	0.59	0.61	0.49	0.60	0.73	-
39	0.55	0.47	0.59	0.50	0.58	0.60	0.68	0.50	0.47	0.50	-
40	0.44	0.50	0.69	0.57	0.44	0.58	0.66	0.50	0.50	0.50	-
41	0.48	0.56	0.69	0.57	0.47	0.70	0.60	0.67	0.64	0.58	-
42	0.51	0.55	0.49	0.50	0.50	0.68	0.50	0.48	0.62	0.60	-
43	0.56	0.47	0.72	0.50	0.59	0.50	0.72	0.50	0.61	0.50	-
44	0.47	0.57	0.60	0.64	0.62	0.55	0.48	0.54	0.58	0.55	-
45	0.59	0.54	0.60	0.64	0.50	0.58	0.50	0.60	0.47	0.62	-
46	-	0.50	0.60	-	0.59	0.58	-	-	0.49	0.50	-
47	-	0.55	0.70	-	0.64	0.50	-	-	0.64	0.60	-
48	-	0.47	0.49	-	0.64	0.60	-	-	0.62	0.58	-
49	-	0.57	0.58	-	0.64	0.57	-	-	0.61	0.49	-
50	-	0.55	0.56	-	0.59	0.54	-	-	0.58	0.50	-
51	-	-	-	-	-	0.54	-	-	0.65	-	-
52	-	-	-	-	-	0.62	-	-	0.61	-	-
53	-	-	-	-	-	0.55	-	-	0.58	-	-
54	-	-	-	-	-	0.55	-	-	0.57	-	-
55	-	-	-	-	-	0.55	-	-	0.63	-	-
56	-	-	-	-	-	0.55	-	-	0.63	-	-
57	-	-	-	-	-	0.61	-	-	0.50	-	-
58	-	-	-	-	-	0.58	-	-	0.50	-	-
59	-	-	-	-	-	0.58	-	-	0.63	-	-
60	-	-	-	-	-	0.57	-	-	0.56	-	-
61	-	-	-	-	-	0.59	-	-	0.49	-	-
62	-	-	-	-	-	0.60	-	-	0.56	-	-
63	-	-	-	-	-	0.58	-	-	0.61	-	-
64	-	-	-	-	-	0.70	-	-	0.50	-	-
65	-	-	-	-	-	0.68	-	-	0.50	-	-
66	-	-	-	-	-	0.55	-	-	0.50	-	-
67	-	-	-	-	-	0.58	-	-	0.50	-	-
68	-	-	-	-	-	0.58	-	-	0.50	-	-
69	-	-	-	-	-	0.50	-	-	0.56	-	-
70	-	-	-	-	-	0.60	-	-	0.49	-	-
71	-	-	-	-	-	0.57	-	-	-	-	-
72	-	-	-	-	-	0.54	-	-	-	-	-
73	-	-	-	-	-	0.54	-	-	-	-	-
74	-	-	-	-	-	0.62	-	-	-	-	-
75	-	-	-	-	-	0.55	-	-	-	-	-

The colour coding indicates the overall risk (R_o) where, green = risk class I; yellow = risk class II, red = risk class III (refer to Fig. 12)

landslide intensities relate to different LSZ classes (very low and low LSZ relates to low intensity (LI); moderate, high and very high LSZ relate to moderate (MI), high (HI) and very high (VHI) intensities, respectively). The landslide intensity lies on the scale of 0 and 1 (i.e. 0.2 for LI, 0.4 for MI, 0.8 for HI and 1.0 for VHI).

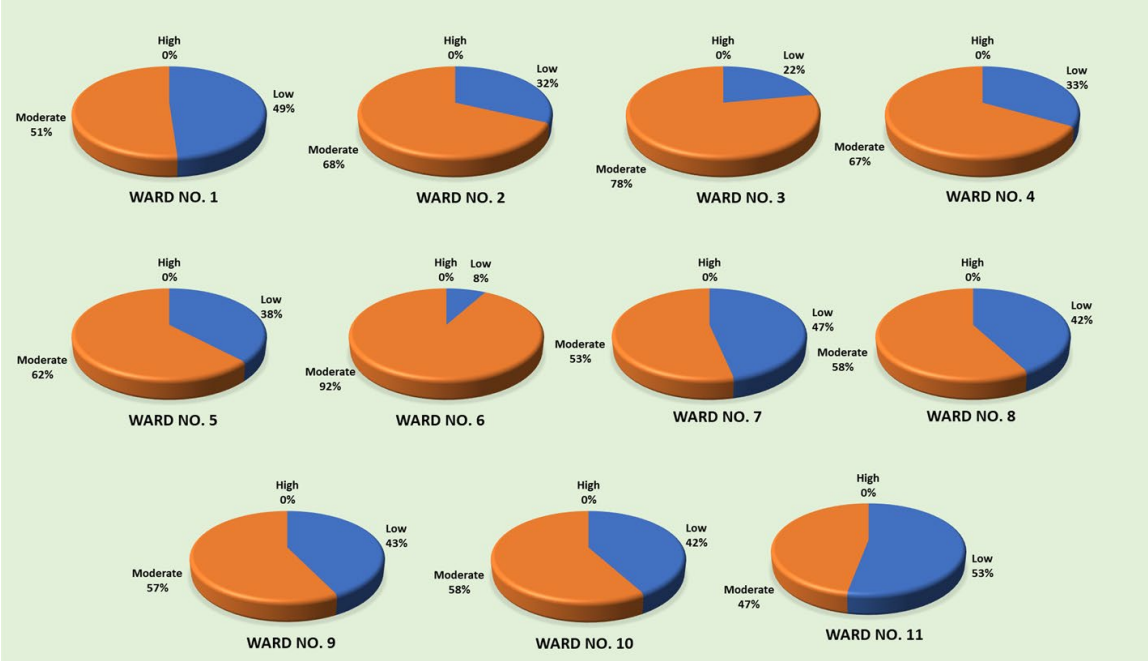


Fig. 9 Resistance class of the buildings (ward-wise) obtained in the study area

4.4.2.3 Quantification of PV of buildings Physical vulnerability (PV) of 542 buildings in the eleven wards of the study area has been estimated using Eq. (4), which is dependent on the resistance (R) ability of the buildings for different intensity (I) scenarios (i.e. LI, MI, HI and VHI). The PV thus obtained at low, moderate, high and very high intensities is referred as PV_{LI}, PV_{MI}, PV_{HI} and PV_{VHI}, respectively on a scale of 0 and 1, which is further classified into three categories, in order to estimate the degree of vulnerability as Class I ($PV \leq 0.5$: low damage), Class II ($0.5 < PV \leq 0.9$: moderate to severe damage) and Class III ($0.9 < PV \leq 1.0$: very severe to complete damage).

When the buildings on the test site were exposed to low intensity ($I = 0.2$), PV values of all the buildings fall in Class I (i.e. $PV \leq 0.5$) exhibiting low damage, as shown in Fig. 10a.

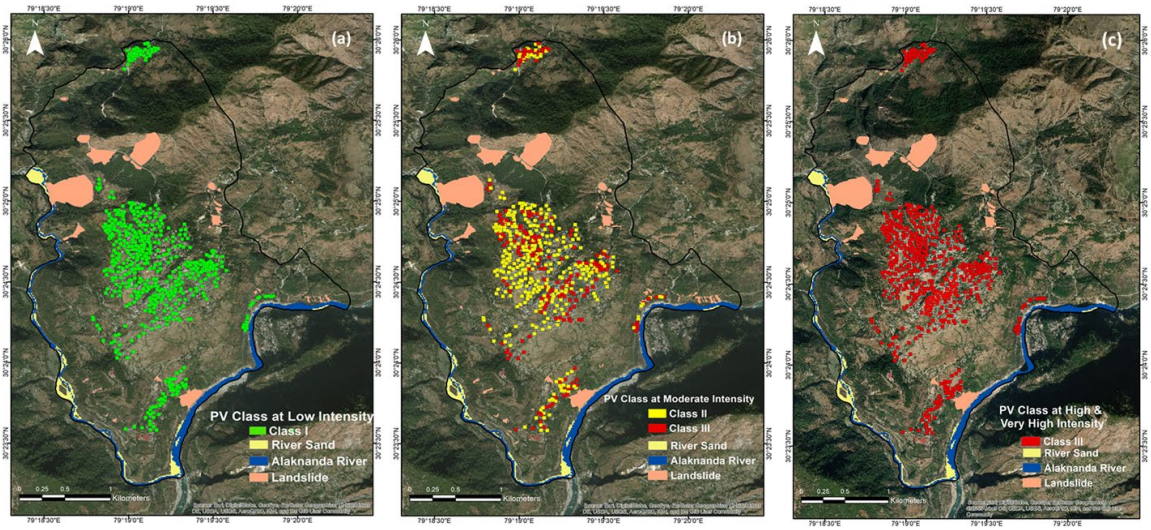


Fig. 10 Physical vulnerability of buildings at **a** low intensity (LI), **b** moderate intensity (MI) and **c** high intensity (HI) and very high intensity (VHI)

However, at moderate intensity ($I=0.4$), 64% of the buildings in the test site belong to Class II PV and 36% of the buildings exhibit Class III PV depicting moderate to severe damage and very severe to complete damage, respectively, as shown in Fig. 10b.

When the buildings in the study area were exposed to high intensity (i.e. $I=0.8$) and very high intensity (i.e. $I=1$), PV values of the buildings were obtained as '1' (i.e. Class III) depicting very severe to complete damage (Fig. 10c). Therefore, it can be concluded that the buildings in the study area cannot sustain high and very high intensity of landslide.

4.4.3 Quantification of LR of buildings

The overall risk (R_o) of the buildings in the study area is computed in two steps as discussed in Sects. 4.4.3.1 and 4.4.3.2.

4.4.3.1 Assessment of specific risk (R_s) of buildings at different landslide intensities relating to different landslide susceptibility zones The assessment of specific risks (R_s) to all the buildings (542 buildings) in the study area is based on PV of the buildings corresponding to different scenario-based landslide intensities (LI, MI, HI and VHI) and the proximity of the buildings from different LSZ classes and the drainage channels associated with them as discussed in Sect. 3. In the present case, very low and low LSZ classes correspond to low intensity (LI), whereas moderate, high and very high LSZ classes correspond to moderate intensity (MI), high intensity (HI) and very high intensity (VHI) of landslides, respectively. The four landslide intensity classes were extracted separately from the already prepared LSZ map (Fig. 8) using ArcGIS. Thereafter, the drainage channels associated with these landslide intensity classes are also extracted from the drainage map. Further, a particular landslide intensity class and associated drainage channels were considered to create multiple buffer zones around them at a distance of 0–50 m, 50–100 m, 100–150 m and $\geq$ 150 m using proximity tool in ArcGIS. Thus, four different buffer maps for four different landslide intensity classes were prepared. The GIS layer pertaining to the location of all the 542 buildings was then superimposed on each of the buffer maps.

Hence, depending upon the position of each building on each of the buffer maps corresponding to four different landslide intensity classes, a certain percentage of the original PV (as determined and discussed in Sect. 4.2) has been reduced according to the developed criteria (as discussed in Sect. 3.2) as given in Table 2. Thus, the specific risks to all the buildings at four different landslide intensities were determined using Eq. (2) as described in Sect. 3. The four different specific risks $(R_s)_i$ at LI, MI, HI and VHI are referred as $(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$, respectively. The values of $(R_s)_i$ for all the 542 buildings are further divided into three risk classes as risk class I ($(R_s)_i \leq 0.5$), risk class II ($0.5 < (R_s)_i \leq 0.9$) and risk class III ($0.9 < (R_s)_i \leq 1$) corresponding to low, moderate and high risk classes, respectively.

Therefore, the visual representation of the four specific risk maps as $(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$ indicating the risk class when exposed to four different intensities is made in Fig. 11a–d, respectively. A histogram chart showing the number of buildings falling under different risk classes (I, II and III) at varying landslide intensities (LI, MI, HI and VHI) is also given in Fig. 12. Hence, from Figs. 11a and 12, it can be stated that all the 542 buildings accounted for risk class I at low landslide intensity (LI) scenario. Under moderate landslide intensity (MI) scenario (Fig. 11b), out of the 542 buildings, 11%, 76% and 13% of the buildings (corresponding to 58, 412 and 72 buildings,

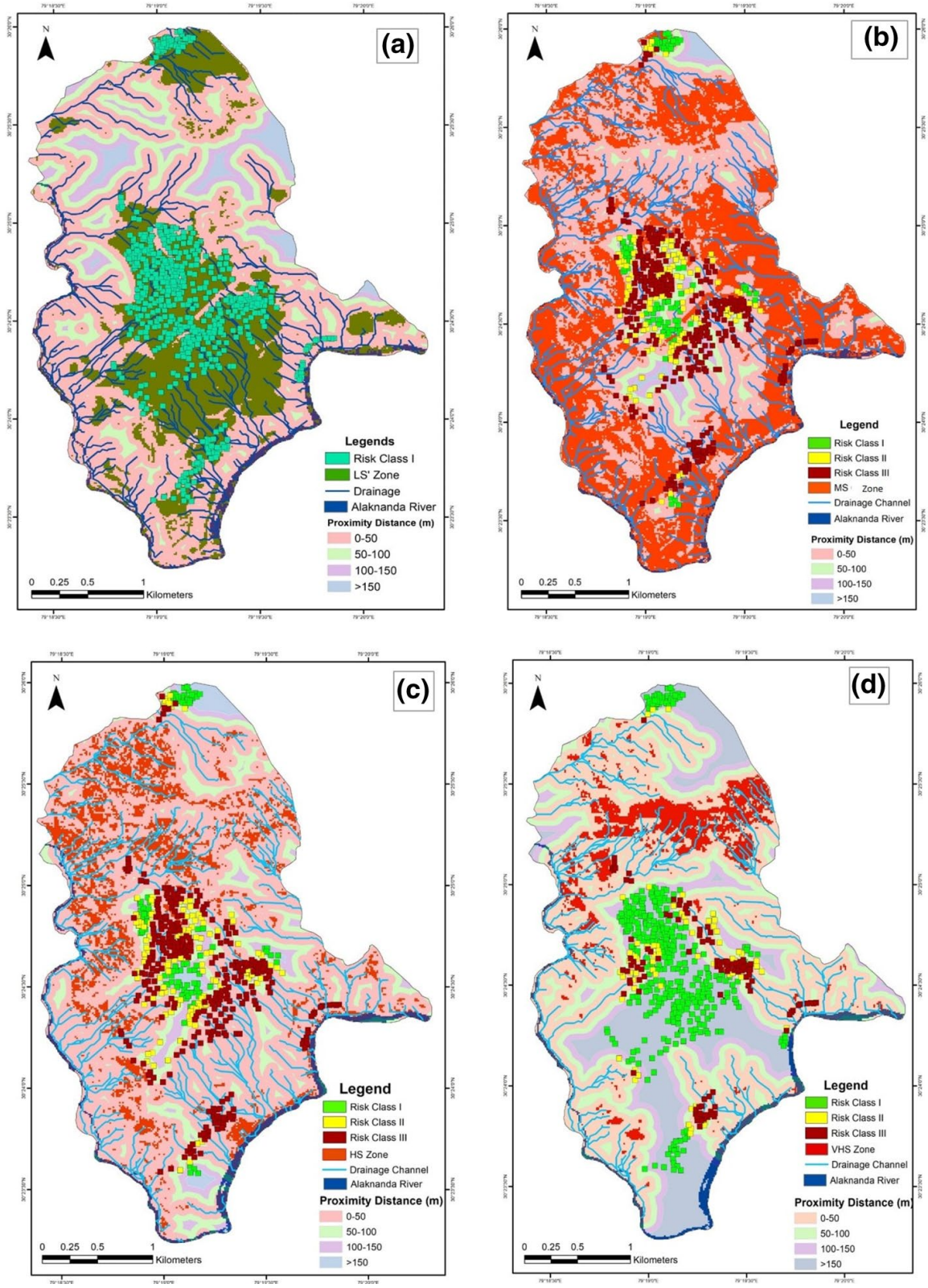


Fig. 11 Risk class distribution of the buildings at **a** $(Rs)_L$; **b** $(Rs)_M$; **c** $(Rs)_H$ and **d** $(Rs)_{VH}$

respectively) fall under risk classes I, II and III, respectively (Fig. 12). In case of high landslide intensity (HI) scenario (Fig. 11c), out of 542 buildings, 17%, 23% and 60% of the buildings (corresponding to 95, 123 and 324 buildings, respectively) fall under risk classes I, II and III, respectively (Fig. 12). Further at very high landslide intensity

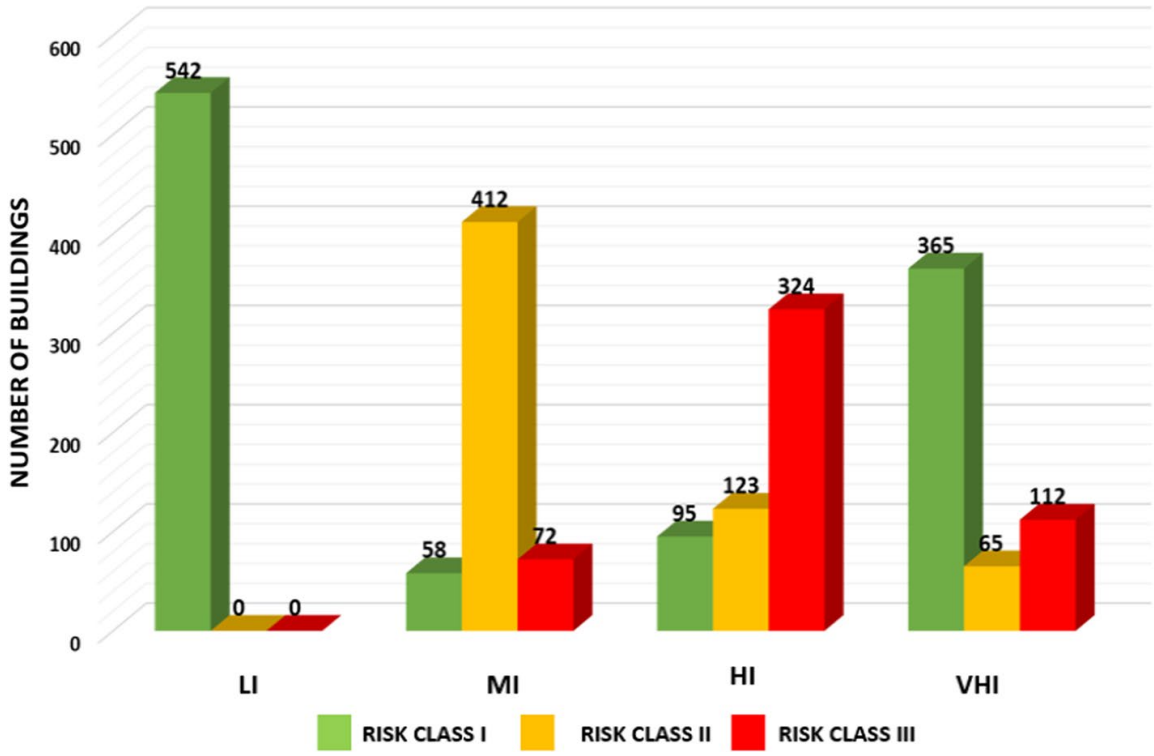


Fig. 12 Number of buildings falling under different risk classes (I, II and III) at varying landslide intensities (LI, MI, HI and VHI)

(VHI) scenario, 67%, 12% and 21% of the total 542 buildings (corresponding to 365, 65 and 112 buildings, respectively) fall under risk classes I, II and III, respectively (refer to Figs. 11d and 12).

Therefore, it can also be concluded that the buildings with certain degree of resistance ability (exhibiting low and moderate resistance classes as shown in Fig. 9) may fall into different degree of risks (i.e. risk classes I, II and III) when exposed to different intensity scenarios (i.e. LI, MI, HI and VHI) as shown in Fig. 11a–d.

4.4.3.2 Assessment of overall risk (R_o) of buildings It involves the integration of the four specific risk maps such as $(R_s)_L$, $(R_s)_M$, $(R_s)_H$ and $(R_s)_{VH}$ (Fig. 11a–d, respectively) on a GIS platform to generate an overall risk (R_o) map for surveyed buildings in the study area as discussed in Sect. 3. The overall risk scores $[(R_o)_{score}]$ of the buildings studied in the area were obtained from the arithmetic weighted overlay integration of the four specific risk maps using Eq. (3). After the arithmetic overlay integration of the four specific risk maps, all the studied buildings attained four-digit risk scores, which have further been classified into different overall risk classes as per the classification scheme given in Table 7. As per the classification scheme (Table 7), the overall risk class of each building in the overall risk map represents the worst or highest risk class (i.e. the highest numerical risk value of 1, 2 or 3 in the four digit risk score) of that particular building out of all the four specific risk classes obtained for four different intensity classes (as explained in Sect. 3). Thus, the highest numerical risk value in the overall risk score will indicate the risk class of that building. For example, if the overall risk score of a building is 1321, 1131, 3221, 3231 or 3321, that particular building will carry the highest numerical risk value of 3 corresponding to risk class III (i.e. high risk).

Table 7 Overall risk scores [$(R_o)_{score}$] of the buildings and their classification scheme into overall risk classes

S. no.	$(R_o)_{score}$ of the buildings obtained	Overall risk (R_o) class of buildings
1	1111	I
2	1121	II
3	1131	III
4	1211	II
5	1221	II
6	1231	III
7	1311	III
8	1321	III
9	1331	III
10	2111	II
11	2121	II
12	2211	II
13	2221	II
14	2231	III
15	2321	III
16	2331	III
17	3221	III
18	3231	III
19	3311	III
20	3321	III
21	3331	III

Following the classification scheme for overall risk (Table 7) as described above, the overall risk map for all 542 buildings considered in the study area has been prepared (Fig. 13). From the assessment, it is found that 50%, 23% and 27% of the total buildings lie in risk class I, II and III, respectively. The correlation between the overall risk classes (i.e. I, II and III) of the buildings of all the wards and their corresponding resistance abilities has also been established and is presented in Table 6 ward-wise with colour representations of risk classes (as per Fig. 13). This shows that the building resistance factor alone cannot decide the overall risk class, as the overall risk is an integration of landslide intensity, building resistance and the proximity of buildings to the landslide.

The ward-wise distribution of buildings in different overall risk classes (i.e. I, II and III) is presented in Fig. 14. It can be inferred from Fig. 14 and Table 6 that all the studied buildings in ward nos. 1, 4 and 7 (i.e. 45 buildings each) fall under low overall risk (i.e. risk class I). A maximum of 31 out of 50 buildings (i.e. 62%) in ward no. 2 falls under moderate overall risk (i.e. risk class II). Further, it is also found that 100% of buildings (i.e. 17 buildings) in ward no. 11 and 74% of buildings (i.e. 37 out of 50 buildings) in ward no. 10 fall under high overall risk (i.e. risk class III).

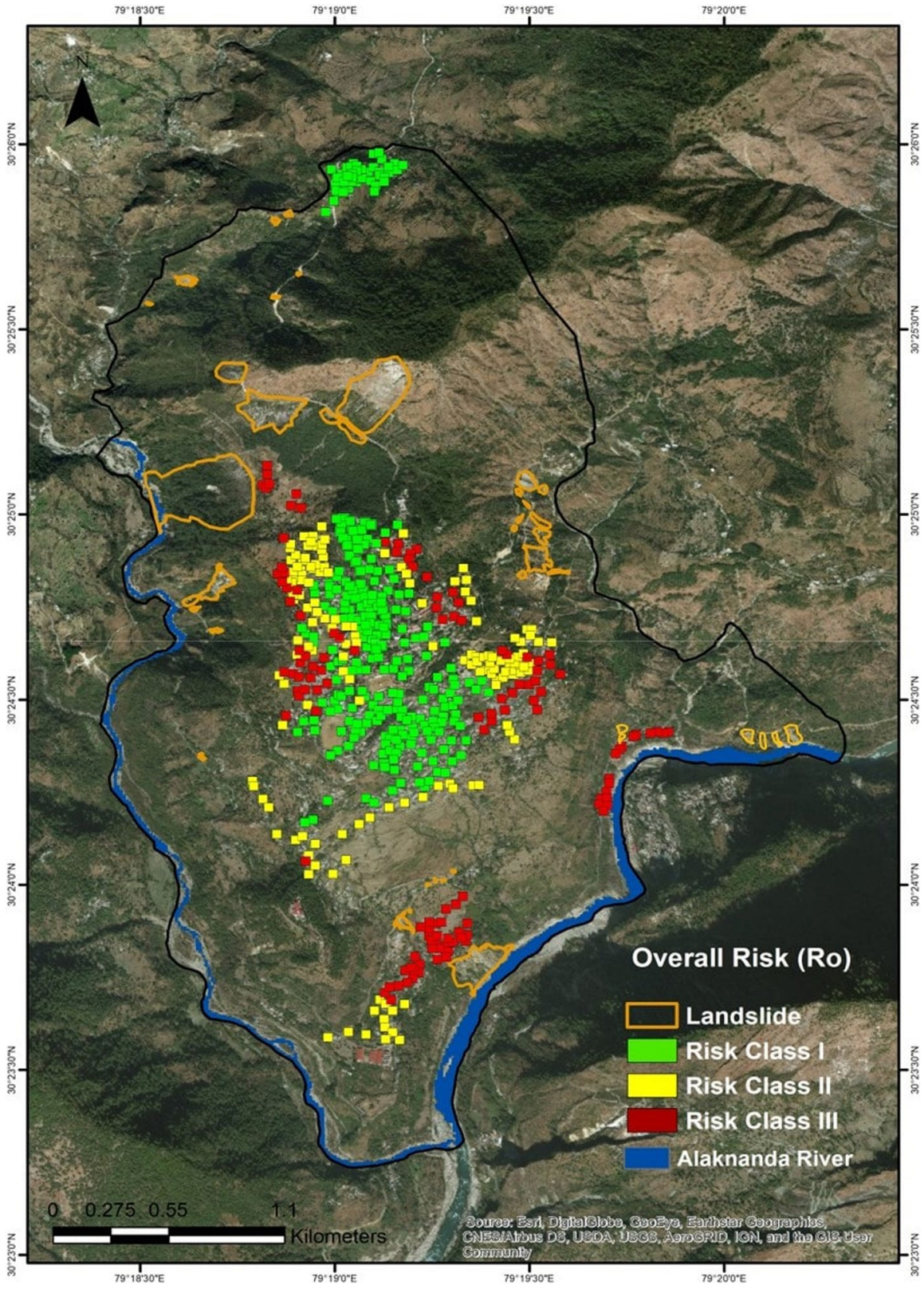


Fig. 13 Overall risk (R_o) map of the buildings in the study area showing the building distributed into three risk classes as risk class I (low risk), risk class II (moderate risk) and risk class III (high risk)

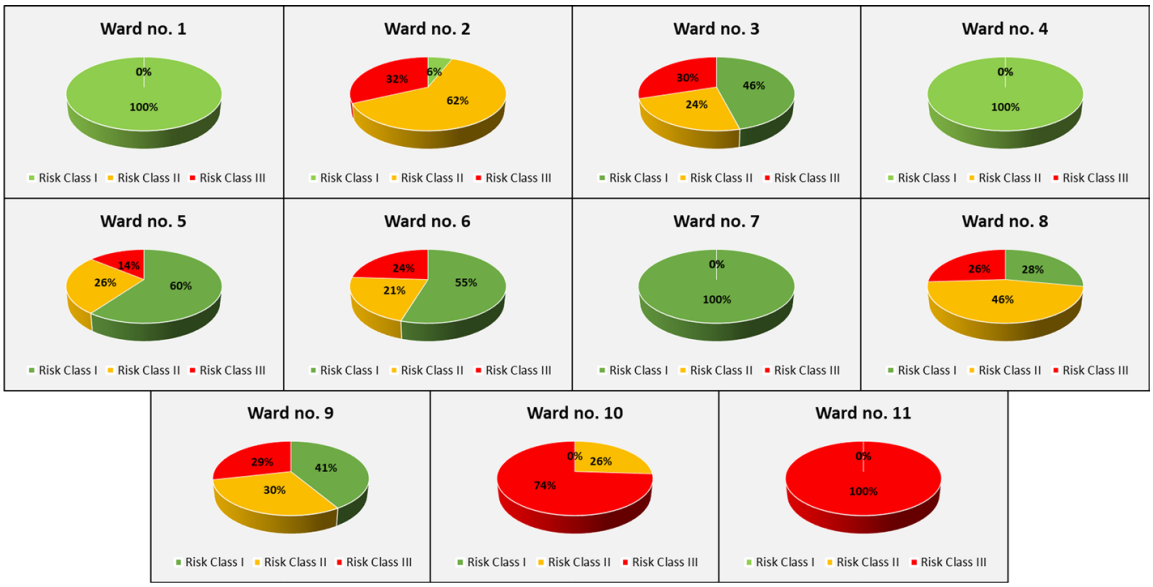


Fig. 14 Ward-wise distribution of buildings in overall risk (R_o) classes in the study area as risk class I (low risk), risk class II (moderate risk) and risk class III (high risk)

5 Results and discussions

Globally, a number of studies in context to the assessment of landslide vulnerability and risk have been conducted by various researchers and scientists (refer to the literature listed in Table 1). However, no unambiguous methodology exists till now to quantify landslide risk based on the integration of landslide susceptibility zonation, physical vulnerability and proximity to susceptible zones and associated drainage channels (Glade and Crozier 2005; Guillard-Gonçalves et al. 2016; Singh et al. 2017). It has also been observed that majority of the published literature on landslide vulnerability and risk assessment were from European countries. In context to India, no unified methodology exists till date that presents a stepwise procedure involving landslide susceptibility–vulnerability–risk assessment (Anbalagan and Singh 1996; Kanungo et al. 2008; Jaiswal et al. 2010; Das et al. 2011; Jaiswal et al. 2011; Ghosh et al. 2012; Martha et al. 2013; Sajinkumar et al. 2014; Thennavan et al. 2016; Singh et al. 2017, 2018, 2019). Moreover, landslide research in India has always been hindered by the deficiency of historical data. Therefore, keeping in view the need of a suitable framework for Indian scenarios, an integrated approach for landslide susceptibility–vulnerability–risk assessment has been proposed and executed in a test site in this study.

The developed integrated approach is tested on 542 buildings of different typologies of Gopeshwar Township, Chamoli District (Uttarakhand, India). The first step involved the preparation of LSZ map using information value (InfoVal) technique, based on various causative factors and existing landslide distribution in the study area. It has been found that 25.7% of landslide area falls under VHS zone (i.e. occupying 3.3% of the total study area), while 36.6%, 37.6%, 0.09% and 0.05% landslide areas fall under HS, MS, LS and VLS zones, respectively. It can further be stated that 13.9% of HS and VHS areas together contain 62.34% of existing landslide areas.

The second part of this research is the assessment of physical vulnerability (PV) of the buildings on the basis of resistance ability of the buildings and the intensity scenarios of the landslides in the study area. It was found that when the buildings in the study

area were exposed to low intensity scenario ($I = 0.2$) represented by very low and low landslide susceptible zones, PV of all the buildings belonged to Class I (when PV ≤ 0.5) exhibiting low damage. However, at moderate intensity ($I = 0.4$) represented by moderate landslide susceptible zone, 64% of the buildings in the study area belonged to Class II PV and 36% of the buildings exhibit Class III PV depicting moderate to severe damage and very severe to complete damage, respectively. When the buildings in the study area were exposed to high intensity (i.e. $I = 0.8$) and very high intensity, (i.e. $I = 1$) represented by high and very high landslide susceptible zones, respectively, PV of the buildings was obtained as '1' (i.e. Class III) depicting very severe to complete damage. Therefore, it can be concluded that the buildings in the study area cannot sustain high and very high intensity of landslides.

In the final stage of this research, LRA is represented as three risk classes I, II and III by integrating specific risks obtained from PV due to individual landslide intensity scenario (i.e. LI, MI, HI and VHI) and the proximity of the buildings from LSZ and associated drainage channels. At first, the specific risk of the buildings of all the wards related to different landslide intensity scenarios was analysed. It was found that all the 528 studied buildings with estimated resistance abilities in all the 11 wards of the hill township fell under risk class I at low intensity (LI) scenario (refer to Figs. 11a and 12). However, at moderate intensity (MI) scenario out of all these buildings, 58, 412 and 72 buildings were classified under risk class I, II and III, respectively (refer to Figs. 11b and 12). At high landslide intensity (HI) scenario, around 95, 123 and 324 buildings fell under risk class I, II and III, respectively (refer to Figs. 11c and 12). Further, at very high intensity (VHI) scenario, around 365, 65 and 112 buildings of the township were classified under risk class I, II and III, respectively (as shown in Figs. 11d and 12). Subsequently, the specific risk maps at four different landslide intensity scenarios were integrated spatially in GIS platform to develop an overall risk map for the studied buildings according to the developed classification scheme (refer to Table 7). The overall risk class of each building in the overall risk map represents the worst or highest risk class of that particular building out of all the four specific risk classes obtained for four different intensity scenarios. From the assessment, it was found that 50%, 23% and 27% of the total 528 studied buildings were classified under overall risk class I, II and III, respectively. In terms of ward-wise distribution, it was concluded that all the 11 buildings (i.e. 100%) of ward number 11 and a maximum of 74% of the buildings (i.e. 37 out of 50 buildings) of ward number 10 were categorized under high risk class III. However, a maximum number of buildings of ward nos. 1, 4 and 7 (i.e. 100% buildings) and of ward number 2 (i.e. 62% buildings) were classified as low-risk class I and moderate-risk class II, respectively.

Therefore, from this assessment, it can also be concluded that the landslide risk assessment of building infrastructures is the integration of various factors such as landslide intensity, resistance abilities of buildings and proximity of buildings to landslides and associated drainage channels. The earlier landslide risk assessment methodologies in India were based on the integration of landslide susceptibility zonation map and the land use land cover (LULC) map in a qualitative manner using risk ranking matrix (Anbalagan and Singh 1996) and in a semi-quantitative manner using danger pixel and fuzzy set concepts (Kanungo et al. 2008). The novelty of the present developed methodology for risk assessment of building infrastructures in hill township exposed to landslide hazards lies in its integrated assessment of susceptibility–vulnerability–risk in a step-wise quantitative manner.

6 Conclusions

A novel remote sensing and GIS-based quantitative technique for an integrated assessment of susceptibility–vulnerability–risk (i.e. LSZ–PV–LRA) of building infrastructures exposed to landslides has been developed. This is one of the unique techniques developed for the first time in this particular subject area that is very systematic and easy to adopt and can be replicated in other built-up areas of hilly regions. The step-wise execution of the developed methodology on a GIS platform includes integration of various causative factors for assessing LSZ, correlation of LSZ areas into landslide intensity scenarios, estimation of building resistance, assessment of PV by integrating resistance of buildings and landslide intensity, and finally assessment of landslide risk by integrating PV and proximity to LSZ class and associated drainage channels. Moreover, the methodology developed has been tested on 528 buildings of different typologies in Gopeshwar Township area of Chamoli District (Garhwal Himalayas, Uttarakhand, India) mostly dominated by debris flow types of potential failures.

The LSZ map prepared in the first phase of the present study will be useful for the future developmental planning in and around the study area. As observed, 13.9% of high susceptibility (HS) and very high susceptibility (VHS) areas together contain 62.34% of existing landslide areas. Hence, the zones falling under HS and VHS should be avoided for future infrastructural developments as ongoing landslide activities are evident or landslides are more likely or certain to occur in these zones. Under unavoidable circumstances in these zones, detailed geo-investigation is required for planning, designing and implementing appropriate control measures to reduce the risk due to landslides.

The assessment of physical vulnerability of buildings in the second phase depicted that at moderate intensity 64% and 36% of the buildings exhibit Class II and Class III PV depicting moderate to severe damage and very severe to complete damage, respectively (refer to Fig. 10b). Further, all the buildings in the study area belonged to PV class III depicting very severe to complete damage at high and very high landslide intensity corresponding to HS and VHS zones, respectively (refer to Fig. 10c). In this process of PV assessment of buildings, the resistance ability of each building and the parameters controlling the reduced resistance of the building could be easily assessed and identified. The practicing engineers and administrative authorities can use this building-specific information to plan and design a sustainable repair and retrofitting strategy. However, the collection of structural and non-structural information of buildings is time-consuming.

LRA in terms of overall risk is performed in the final phase by integrating specific risks to the buildings as obtained from PV due to different landslide intensity scenarios and the proximity of the buildings to LSZ and associated drainage channels. The proximity of buildings is a unique and realistic parameter introduced for overall risk assessment of the buildings in this research work. The specific risk (R_S) of all the 528 buildings fell under risk class I at low intensity scenario (refer to Figs. 11a and 12), and at moderate intensity (MI) scenario, 58, 412 and 72 buildings were classified under risk class I, II and III, respectively (refer to Figs. 11b and 12). At high landslide intensity (HI) scenario 95, 123 and 324 buildings fell under risk class I, II and III, respectively (refer to Figs. 11c and 12), whereas at very high intensity (VHI) scenario 365, 65 and 112 buildings were classified under risk class I, II and III, respectively (as shown in Figs. 11d and 12). Subsequently, by integrating these four specific risk maps at four different landslide intensity scenarios, an overall risk (R_o) map was produced (refer to Fig. 13). From such assessment, it was observed that 50%, 23% and 27% of the total

528 studied buildings were classified under overall risk class I, II and III, respectively (refer to Table 6). The 142 buildings classified under high risk (i.e. 16, 15, 7, 18, 12, 20, 37 and 17 number of buildings in Ward Nos. 2, 3, 5, 6, 8, 9, 10 and 11, respectively as shown in Table 6) need special attention by the engineers for their repair and retrofitting to improve their resistance abilities. In such way, this technique of landslide risk assessment will help planners and engineers to identify and highlight the hotspots of landslides along with building infrastructures, in order to prioritize the work and suggest a better protective and mitigation measure.

The major limitations in the present research can be enumerated as follows:

- The present approach of vulnerability assessment of buildings is field data-intensive and thereby time-consuming.
- Various resistance parameters have been assigned different weightages based on expert knowledge, which are subjective in nature. Further, the threefold classification scheme of vulnerability and risk is also subjective in nature. Statistical or other methods can be introduced to objectively define such class boundaries.
- The landslides in the study area may include different failure types/mechanisms. The proposed methodology can be refined for major types of landslides such as debris slides/flows, rock falls/topples, etc., in terms of assessment of their intensities.
- Landslide intensity being a dynamic parameter involves dynamic forces, which are not considered in the present study.

Nevertheless, this study may lead to some future prospects of research as follows:

- The study can be extended to debris flow and rock fall modelling with the help of numerical simulations through available software such as Flo-2D, Flo-3D, RAMMS, etc., based on field parameters for obtaining the detailed information on landslide volume, velocity, intensity and run-out effect, which can be integrated with the risk assessment process.
- Scope exists on development of physical vulnerability and risk assessment techniques for other infrastructures such as dams, bridges, roads networks, hydropower project sites, etc., at risk due to landslides.
- It could be interesting to consider risk assessment in terms of cost-economics of various elements at risk such as buildings, road networks, dams, bridges, hydropower project sites and other vital infrastructures in hilly terrains.
- Considering different types of hazards in Himalayan terrain (predominantly landslides and earthquakes), multi-hazard concept-based techniques for vulnerability and risk assessment of various infrastructures in hills can also be attempted.

Acknowledgements The third author wishes to thank the Director, CSIR-Central Building Research Institute (CBRI), Roorkee (Uttarakhand), India, for his kind permission to publish this work.

Compliance with ethical standards

Conflict of interest No potential conflict.

Code availability ArcGIS 10.3

References

Abdulwahid, W. M., & Pradhan, B. (2016). Landslide vulnerability and risk assessment for multi-hazard scenarios using airborne laser scanning data (LiDAR). *Landslides, 14*(3), 1057–1076.

Adiguzel, F., Cetin, M., Kaya, E., Simsek, M., Gungor, S., & Bozdogan Sert, E. (2019). Defining suitable areas for bioclimatic comfort for landscape planning and landscape management in Hatay, Turkey. *Theoretical and Applied Climatology, 139*(3–4), 1493–1503.

AGS (Australian Geomechanics Society Sub-Committee on Landslide Risk Management). (2000). Landslide risk management concepts and guidelines. *Australian Geomechanics Journal, 35,* 49–92.

AGS (Australian Geomechanics Society Sub-Committee on Landslide Risk Management). (2007). A national landslide risk management framework for Australia. *Australian Geomechanics Journal, 42*(1), 1–12.

Ahmed, S. (2009). Slope stability analysis using GIS and numerical modeling techniques. Unpublished MSc thesis, Vrije Universiteit, Brussel.

Anbalagan, R. (1992). Landslide hazard evaluation and zonation mapping in mountainous terrain. *Engineering Geology, 32,* 269–277.

Anbalagan, R., & Singh, B. (1996). Landslide hazard and risk assessment mapping of mountainous terrains—A case study from Kumaon Himalaya, India. *Engineering Geology, 43,* 237–246.

Balamurugan, G., Ramesh, V., & Touthang, M. (2016). Landslide susceptibility zonation mapping using frequency ratio and fuzzy gamma operator models in part of NH-39, Manipur, India. *Natural Hazards, 84,* 465–488.

Basu, T., & Pal, S. (2018). Identification of landslide susceptibility zones in Gish River basin, West Bengal, India. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards, 12*(1), 14–28.

Bednárik, M., Yilmaz, I., & Marschalko, M. (2012). Landslide hazard and risk assessment: A case study from the Hlohovec–Sered' landslide area in south-west Slovakia. *Natural Hazards, 64*(1), 547–575.

Bell, R., & Glade, T. (2004). Quantitative risk analysis for landslides—Examples from Bíldudalur, NW-Iceland. *Natural Hazards Earth System Sciences, 4,* 117–131.

BMTPC. (2019). *Landslide hazard zonation atlas of India*. Building Materials and Technology Promotion Council: Government of India and Anna University, Chennai.

Bogget, A. D., Mapplebeck, N. J., & Cullen, R. J. (2000). South Shores Cliffs, Whitehaven-geomorphological survey and emergency cliff stabilisation works. *Quarterly Journal of Engineering Geology and Hydrogeology, 33,* 213–226.

Cardinali, M., Reichenbach, P., Guzzetti, F., Ardizzone, F., Antonini, G., Galli, M., et al. (2002). A geomorphological approach to the estimation of landslide hazards and risks in Umbria, Central Italy. *Natural Hazards Earth System Sciences, 2,* 57–72.

Catani, F., Casagli, N., Ermini, L., Righini, G., & Menduni, G. (2005). Landslide hazard and risk mapping at catchment scale. *Landslides, 2,* 329–342.

Cetin, M., Adiguzel, F., Gungor, S., Kaya, E., & Sancar, M. C. (2019). Evaluation of thermal climatic region areas in terms of building density in urban management and planning for Burdur, Turkey. *Air Quality, Atmosphere and Health, 12*(9), 1103–1112.

Chau, K. T., Sze, Y. L., Fung, M. K., Wong, W. Y., Fong, E. L., & Chan, L. C. P. (2004). Landslide hazard analysis for Hong Kong using landslide inventory and GIS. *Computers & Geosciences, 30,* 429–443.

Chauhan, S., Sharma, M., Arora, M. K., & Gupta, N. K. (2010). Landslide susceptibility zonation through ratings derived from artificial neural network. *International Journal of Applied Earth Observation and Geoinformation, 12,* 340–350.

Chowdhury, R., & Flentje, P. (2003). Role of slope stability analysis in landslide risk management. *Bulletin of Engineering Geology and Environment, 62,* 41–46.

Chung, C. F., & Fabbri, A. G. (2003). Validation of spatial prediction models for landslide hazard mapping. *Natural Hazards, 30*(3), 451–472.

Corominas, J., Westen, C., Frattini, P., Cascini, L., Malet, J. P., Fotopoulou, S., et al. (2014). Recommendations for the quantitative analysis of landslide risk. *Bulletin of Engineering Geology and Environment, 73*(2), 209–263.

Dai, F. C., Lee, C. F., & Ngai, Y. Y. (2002). Landslide risk assessment and management: An overview. *Engineering Geology, 64*(1), 65–87.

Das, I., Kumar, G., Stein, A., Bagchi, A., & Dadhwal, V. K. (2011). Stochastic landslide vulnerability modeling in space and time in a part of the Northern Himalayas, India. *Environmental Monitoring and Assessment, 178,* 25–37.

Dhakal, A. S., Amada, T., & Aniya, M. (2000). Landslide hazard mapping and its evaluation using GIS: An investigation of sampling schemes for a grid-cell based quantitative method. *Photogrammetric Engineering & Remote Sensing, 66*(8), 981–989.

Fabera, M. H., & Stewart, M. G. (2003). Risk assessment for civil engineering facilities: Critical overview and discussion. *Reliability Engineering and System Safety, 80,* 173–184.

Fell, R. (1994). Landslide risk assessment and acceptable risk. *Canadian Geotechnical Journal*, *31*, 261–272.

Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., & Savage, W. Z. (2008). On behalf of the JTC-1 joint technical committee on landslides and engineered slopes: Guidelines for landslide susceptibility, hazard and risk zoning for land use planning. *Engineering Geology, 102,* 85–98.

Fell, R., & Hartford, D. (1997). Landslide risk management. In D. Cruden & R. Fell (Eds.), *Landslide risk assessment* (pp. 51–119). Honolulu: Balkema.

Froude, M. J., & Petley, D. N. (2018). Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards Earth System Sciences, 18,* 2161–2181.

Garcia, R. A. C., Oliveira, S. C., & Zêzere, J. L. (2016). Assessing population exposure for landslide risk analysis using dasymetric cartography. *Hazards Earth System Sciences, 16,* 2769–2782.

Ghosh, S., Van Westen, C. J., Carranza, E. J., Ghoshal, T. B., Sarkar, N. K., & Surendranath, M. (2009). A quantitative approach for improving the BIS (Indian) method of medium scale landslide susceptibility. *Journal Geological Society India, 74,* 625–638.

Ghosh, S., Van Westen, C. J., Carranza, E. J., & Jetten, V. G. (2012). Integrating spatial, temporal and magnitude probabilities for medium-scale landslide risk analysis in Darjeeling Himalayas, India. *Landslides, 9,* 371–384.

Glade, T., & Crozier, M. J. (2005). The nature of landslide hazard and impact. In T. Glade, M. G. Anderson, & M. J. Crozier (Eds.), *Landslide hazard and risk* (pp. 43–74). Chichester: Wiley.

Government of Uttarakhand, India. https://chamoli.gov.in/geography. Retrieved on March 31, 2019.

Guillard-Gonçalves, C., Zêzere, J. L., Pereira, S., & Garcia, R. A. C. (2016). Assessment of physical vulnerability of buildings and analysis of landslide risk at the municipal scale: Application to the Loures municipality, Portugal. *Natural Hazards Earth System Sciences, 16,* 311–331. https://doi.org/10.5194/nhess-16-311-2016.

Gupta, R. P., & Joshi, B. C. (1990). Landslide Hazard Zonation using the GIS Approach—A case Study from the Ramganga Catchment, Himalayas. *Engineering Geology, 28,* 119–131.

Guri, P. K., Champatiray, P. K., & Patel, R. C. (2015). Spatial prediction of landslide susceptibility in parts of Garhwal Himalaya, India, using the weight of evidence modelling. *Environmental Monitoring Assessment, 187,* 324. https://doi.org/10.1007/s10661-015-4535-1.

Guzzetti, F., Carrara, A., Cardinali, M., & Reichenbach, P. (1999). Landslide hazard evaluation: A review of current techniques and their application in a multi-scale study, Central Italy. *Geomorphology, 31*(1–4), 181–216.

Hughes, A., Hewlett, H., Samuels, P. G., Morris, M., Sayers, P., Moffat, I., et al. (2000). *Risk management for UK reservoirs. CIRIA C542*. London: Construction Industry Research and Information Association.

IS 1893-part 1. (2016). *Criteria for earthquake resistant design of structures*. New Delhi: Bureau of Indian Standards.

Jaiswal, P., & van Westen, C. J. (2009). Estimating temporal probability for landslide initiation along transportation routes based on rainfall threshold. *Geomorphology, 112,* 96–105.

Jaiswal, P., van Westen, C. J., & Jetten, V. (2010). Quantitative assessment of direct and indirect landslide risk along transportation lines in southern India. *Natural Hazards Earth System Sciences, 10,* 1253–1267.

Jaiswal, P., van Westen, C. J., & Jetten, V. (2011). Quantitative estimation of landslide risk from rapid debris slides on natural slopes in the Nilgiri hills, India. *Natural Hazards Earth System Sciences, 11,* 1723–1743.

Kannan, M., Saranathan, E., & Anbalagan, R. (2015). Comparative analysis in GIS-based landslide hazard zonation—A case study in Bodi-Bodimettu Ghat section, Theni District, Tamil Nadu, India. *Arabian Journal of Geosciences, 8,* 691–699.

Kanungo, D. P., Arora, M. K., Gupta, R. P., & Sarkar, S. (2008). Landslide risk assessment using concepts of danger pixels and fuzzy set theory in Darjeeling Himalayas. *Landslides, 5,* 407–416.

Kanungo, D. P., Arora, M. K., Sarkar, S., & Gupta, R. P. (2006). A comparative study of conventional, ANN black box, fuzzy and combined neural and fuzzy weighting procedures for landslide susceptibility zonation in Darjeeling Himalayas. *Engineering Geology, 85,* 347–366.

Kanungo, D. P., Arora, M., Sarkar, S., & Gupta, R. (2009). Landslide susceptibility zonation (LSZ) mapping—A review. *Journal of South Asia Disaster Studies, 2*(1), 81–105.

Kaya, E., Agca, M., Adiguzel, F., & Cetin, M. (2019). Spatial data analysis with R programming for environment. *Human and Ecological Risk Assessment: An International Journal*, *25*(6), 1521–1530.

Kelly, A. J., Clifton, A. W., Antunes, P. J., & Widger, R. A. (2005). Application of a landslide risk management system to the Saskatchewan highway network. In O. Hunger, R. Fell, R. Coutre, & E. Eberhardt (Eds.), *Landslide risk management* (pp. 571–580). Rotterdam: Balkema.

Koirala, N. P., & Watkins, A. T. (1988). Bulk appraisal of slopes in Hong Kong. In C. Bonnard (Ed.), *Landslides* (Vol. 2, pp. 1181–1186). Rotterdam: Balkema.

Lee, E. M. (1999). Coastal planning and management: The impact of the 1993 Holbeck Hall Landslide, Scarborough, East Midlands. *Geographer, 21,* 78–91.

Lee, E. M., & Clark, A. R. (2000). The use of archive records in landslide risk assessment: Historical landslide events on the Scarborough Coast, UK. In E. N. Bromhead, N. Dixon, & M. L. Ibsen (Eds.), *Landslides: In research, theory and practice* (pp. 904–910). London: Thomas Telford.

Lin, M. L., & Tung, C. C. (2003). A GIS-based Potential analysis of the landslides induced by the chichi Earthquake. *Engineering Geology, 71,* 63–77.

Martha, T., van Westen, C., Kerle, N., Jetten, V., & Vinod, K. (2013). Landslide hazard and risk assessment using semi-automatically created landslide inventories. *Geomorphology, 184,* 139–150.

Michael-Leiba, M., Baynes, F., & Scott, G. (1999). Quantitative landslides of Cairns. *Australian Geological Survey Organisation*, Department of Industry, Science & Resources, AGSO Records 25, 36, 40.

Moore, I. D., Grayson, R. B., & Ladson, A. R. (1991). Digital terrain modeling: A review of hydrological geomorphological and biological applications. *Hydrological Process, 5,* 3–30.

Oliveira, S. C., Zêzere, J. L., Catalão, J., & Nico, G. (2015). The contribution of PS InSAR interferometry to landslide hazard in weak rock-dominated areas. *Landslides, 12,* 703–719.

Palmer, J. S., Clark, A. R., Cliffe, D., & Eade, M. (2002). The management of risk on the Chalk Cliffs at Brighton, UK. In R. G. McInnes & J. Jakeways (Eds.), *Instability—planning and management* (pp. 355–362). London: Thomas Telford.

Pereira, S., Garcia, R. A. C., Zêzere, J. L., Oliveira, S. C., & Silva, M. (2016). Landslide quantitative risk analysis of buildings at the municipal scale based on a rainfall triggering scenario. *Geomatics, Natural Hazards and Risk, 8*(2), 624–648.

Pham, B. T., Tien Bui, D., Indra, P., & Dholakia, M. B. (2017). Hybrid integration of Multilayer Perceptron Neural Networks and machine learning ensembles for landslide susceptibility assessment at Himalayan area (India) using GIS. *CATENA, 149,* 52–63.

Rautela, P., & Lakhera, R. C. (2000). Landslide risk analysis between Giri and Tons Rivers in Himachal Himalaya (India). *International Journal of Applied Earth Observation and Geoinformation (JAG), 2*(3–4), 153–160.

Remondo, J., Bonachea, J., & Cendrero, A. (2005). A statistical approach to landslide risk modelling at basin scale: From landslide susceptibility to quantitative risk assessment. *Landslides, 2,* 321–328.

Royal Society. (1992). *Risk: Analysis, perception and management*. London: Report of a Royal Society Study Group, Royal Society.

SafeLand. (2010). *Revision 2, 7th framework programme*, Grant Agreement No.: 226479, SGI/ICG revision: 1—final.

Saha, A. K., Gupta, R. P., Sarkar, I., Arora, M. K., & Csaplovics, E. (2005). An approach for GIS-based statistical landslide susceptibility zonation—With a case study in the Himalayas. *Landslides, 2,* 61–69.

Sajinkumar, K. S., Anbazhagan, S., Rani, V. R., & Muraleedharan, C. (2014). A paradigm quantitative approach for a regional risk assessment and management in a few landslide prone hamlets along the wind ward slope of Western Ghats, India. *International Journal Disaster Risk Reduction, 7,* 142–153.

Sajwan, K. S., & Sushil, K. (2016). A geological appraisal of slope instability in Upper Alaknanda Valley, Uttarakhand Himalaya, India. *Journal of Geology and Geophysics, 5*(5), 1–7.

Saldivar-Sali, A., & Einstein, H. H. (2007). A landslide risk rating system for Baguio, Philippines. *Engineering Geology, 91,* 85–99.

Sandilands, N. M., Noble, M., & Findlay, J. W. (1998). Risk assessment strategies for dam based hydro schemes. In P. Tedd (Ed.), *The prospect for reservoirs in the 21st century*. London: Thomas Telford.

Sarkar, S., & Gupta, P. K. (2005). Techniques for landslide hazard zonation—Application to Srinagar-Rudraprayag area of Garhwal Himalaya. *Journal of Geological Society of India, 65,* 217–230.

Sarkar, S., & Kanungo, D. P. (2004). An integrated approach for landslide susceptibility mapping using remote sensing and GIS. *Photogrammetric Engineering & Remote Sensing, 70*(5), 617–625.

Sarkar, S., Kanungo, D. P., Patra, A. K., & Kumar, P. (2008). GIS based spatial data analysis for landslide susceptibility mapping. *Journal of Mountain Science, 5,* 52–62.

Sema, H. V., Guru, B., & Veerappan, R. (2017). Fuzzy gamma operator model for preparing landslide susceptibility zonation mapping in parts of Kohima Town, Nagaland, India. *Modeling. Earth Systems and Environment, 3,* 499–514.

Singh, A., Kanungo, D. P., & Pal, S. (2018). A modified approach for semi-quantitative estimation of physical vulnerability of buildings exposed to different landslide intensity scenarios. *Georisk, 13*(1), 66–81.

Singh, A., Kanungo, D. P., & Pal, S. (2019). Physical vulnerability assessment of buildings exposed to landslides in India. *Natural Hazards, 96*(2), 53–790.

Singh, A., Pal, S., Kanungo, D. P., & Pareek, N. (2017). An overview of recent developments in landslide vulnerability assessment- presentation of a new conceptual framework. In Matjaž Mikoš, Željko Arbanas, Yueping Yin, & Kyoji Sassa (Eds.), *Advancing culture of living with landslides* (Vol. 2, pp. 795–802)., Advances in landslide science New York: Springer.

Sinha, B. N., Varma, R. S., & Paul, D. K. (1975). Landslides in Darjeeling district (West Bengal) and adjacent areas. *Bulletins of the Geological Survey of India, Series-B, 36,* 1–45.

Soeters, R., & van Westen, C. J. (1996). Slope instability recognition, analysis and zonation. In A. K. Turner & R. L. Schuster (Eds.), *Landslides, investigation and mitigation, transportation research board, National Research Council, Special Report 24/7* (pp. 129–177). Washington DC: National Academy Press.

Stocking, M. (1972). Relief analysis and soil erosion in Rhodesia using multi-variate techniques. *Zeitschrift fur Geomorphologie NF, 16,* 432–443.

Strahler, A. (1964). Quantitative geomorphology of drainage basins and channel networks. In V. Chow (Ed.), *Handbook of applied hydrology* (pp. 439–476). New York: McGraw Hill.

Thennavan, E., Ganapathy, G. P., Sekaran, S. S. C., & Rajawat, A. S. (2016). Use of GIS in assessing building vulnerability for landslide hazard in The Nilgiris, Western Ghats, India. *Natural Hazards, 82*(2), 1031–1050.

Uttarakhand Tourism Development Board Department of Tourism, Government of Uttarakhand, India. https://uttarakhandtourism.gov.in/district-chamoli. Retrieved on March 31, 2019.

Uzielli, M., Catani, F., Tofani, V., & Casagli, N. (2015). Risk analysis for the Ancona landslide—II: Estimation of risk to buildings. *Landslides, 12,* 83–100.

Uzielli, M., Nadim, F., Lacasse, S., & Kaynia, A. M. (2008). A conceptual framework for quantitative estimation of physical vulnerability to landslides. *Engineering Geology, 102,* 251–256.

Van Dine, D. F., Jordan, P., & Boyer, D. C. (2002). An example of risk assessment from British Columbia, Canada. In R. G. McInnes & J. Jakeways (Eds.), *Instability—planning and management* (pp. 399–406). London: Thomas Telford.

Van Westen, C. J. (2004). Geo-information tools for landslide risk assessment, an overview of recent developments. In E. Lacerda & F. Sayão (Eds.), *Landslides: Evaluation and stabilization—glissement de terrain: Evaluation et Stabilisation: Proceedings of the 9th international symposium on landslides* (pp. 39–56). London: Balkema.

Van Westen, C. J., Castellanos, E., & Kuriakose, S. L. (2008). Spatial data for landslide susceptibility, hazard, and vulnerability assessment: An overview. *Engineering Geology, 102*(3), 112–131.

Van Westen, C. J., Rengers, N., Terlien, M. T., & Soeters, R. (1997). Prediction of the occurrence of slope instability phenomenal through GIS-based hazard zonation. *Geologische Rundschau, 86*(2), 404–414.

Van Westen, C. J., Van Asch, T. W., & Soeters, R. (2006). Landslide hazard and risk zonation—Why is it still so difficult? *Bulletin Engineering Geology Environment, 65*(2), 167–184.

Varnes, D. J. (1984). *Landslide hazard zonation: A review of principles and practice* (pp. 1–63). Paris: UNESCO Press.

Varnes, D., & The International Association of Engineering Geology Commission on Landslides and Other Mass Movements. (1984). *Landslide hazard zonation: A review of principles and practice*. Paris: UNESCO Press.

Vranken, L., Vantilt, G., VanDen Eeckhaut, M., Vandekerckhove, L., & Poesen, J. (2015). Landslide risk assessment in a densely populated hilly area. *Landslides, 12*(4), 787–798.

Wang, Q., Li, W., Chen, W., & Bai, H. (2015). GIS-based assessment of landslide susceptibility using certainty factor and index of entropy models for the Qianyang County of Baoji City, China. *Journal of Earth System Sciences, 124*(7), 1399–1415.

Wong, H. N. (2005). Landslide risk assessment for individual facilities. In O. Hunger, R. Fell, R. Couture, & E. Eberhardt (Eds.), *Landslide Risk Management* (pp. 237–296). Rotterdam: Balkema.

Zêzere, J. L., Garcia, R. A. C., Oliveira, S. C., & Reis, E. (2008). Probabilistic landslide risk analysis considering direct costs in the area north of Lisbon (Portugal). *Geomorphology, 94,* 467–495.

Zou, Q., Cui, P., Zeng, C., Tang, J., & Regmi, A. D. (2016). Dynamic process-based risk assessment of debris flow on a local scale. *Physical Geography, 37*(2), 132–152.

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

An overview of the opportunities and challenges in sustaining the energy industry in Afghanistan

Shambalid Ahady[1,*], *Nirendra* Dev[2], *Anubha* Mandal[2]

[1]Ph.D. Candidate, Department of Civil Engineering, Delhi Technological University (DTU), India
[2]Professor, Department of Civil Engineering, Delhi Technological University (DTU), India

Abstract. Energy access is not only crucial for economic growth but also important for any strategy to improve the health and social welfare of a nation. Afghanistan's energy industry is in poor condition due to many years of war and negligence. Despite international agencies' support and energy policies adopted in the last few years, Afghanistan has no universal access to power. Besides, the residences suffer from an irregular distribution of power supply. There is a growing gap between demand and supply, and the current predictions of demand do not show reality due to hindered economic growth. Afghanistan's domestic power transmission is limited, which must be extended for the country to enjoy a stable and sustainable energy supply. Sustainability and security of Afghanistan's power sector would rely on its ability to become self-reliant in power generation. Overall, the objective of this paper is to summarize the current energy status of Afghanistan and to identify energy opportunities for self-sufficiency and challenges in various aspects of energy sources. To meet energy demand, Afghanistan can develop its autochthonous hydrocarbon and renewable energy resources. By improving its domestic energy potential from natural resources, Afghanistan can fulfill its primary energy requirement. Further, along with policy formulation, appropriate and planned implementation of renewable energy policy, energy efficiency targets, and strategies, Afghanistan can reach energy self-sufficiency goals with socio-economic development.

1 Introduction

Asia has very significant influence on the global environment and energy trends. The energy choices made in the region have many social and health effects on nearly half of the world's population. Governmental decisions in Asia regarding energy production and consumption, energy conservation, and greenhouse gas affect energy resource depletion trends, global greenhouse gas emission levels and environmental conditions [1].

Afghanistan being a landlocked country, lies at the strategically important location. With a population of about 29.7 million (2017 update) and the total geographic area of 652,864 km^2, Afghanistan is ranked at 41st largest country in the world [2].

Four decades of conflict, civil war, foreign intervention, and political instability have constrained the country's development and have severely affected its' economic growth. However, the last decade has been characterized by remarkable progress after the security transition in 2014, but poverty increased to 54.5% [3]. Access to stable and affordable energy supplies leads to even economic growth. Energy access is crucial for any strategy to improve the health and social welfare of a nation.

During various conflicts, Afghanistan's energy infrastructure, generation, transmission, and distribution have been destroyed. The restoration and development of the nation's physical capital have been the priority of international development agencies since 2001. The progress is hindered due to the high level of damage and massive investment requirements [4]. Afghanistan is among the lowest per capita consumption of energy in the world, with fuelwood accounting for more than 85% of total energy consumption [5]. The per-capita annual power consumption in Afghanistan is 150 kWh, that is negligible to the global average of 2728 kWh. Due to the ever-growing gap between demand and supply, 85 percent of the local population in Afghanistan still does not have -continuous- access to electricity supply [6]. Energy access is one of the priorities for the Afghan government and efforts have been taken for modernization and expansion of the national electricity grid with the possibility of power trade with central and south Asia. However, the national grid, due to its limited infrastructure will not be able to serve the entire population of the country in near future [7]. In this context, the objective of present study is to review the energy status of Afghanistan for the promotion of energy self-sufficiency opportunities in the country, and

* Corresponding author: shambalidahady_phd2k18@dtu.ac.in

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

subsequently identify the challenges that might hinder the process of achieving - energy self-sufficiency.

2 Afghanistan's energy scenario

2.1 Energy demand

The demand for electricity is growing rapidly in Afghanistan due to factors like population growth, GDP increase, energy price changes, historical changes in energy intensity, per capita consumption, domestic energy conditions, and energy availability. There is limited data on Afghanistan's rural energy consumption. The rural Afghans are poor by most countries' standards, and over 20.4% of the rural people cannot meet the minimum level of dietary energy required to sustain a healthy life [8].

Afghanistan's power sector master plan reported that the net demand is expected to rise from 2,800 GWh in 2012 to 15,909 GWh in 2032, with an average growth rate of 9.8 percent per annum, as shown in Figure 1 [9]. The predictions indicated that Afghanistan would require new energy supplies to meet the energy demand even before its domestic energy supplies develop full accessibility [6]. It is asserted that if the energy supplies are efficiently and effectively managed, then the local resources might satisfy Afghanistan's mid-term primary energy requirements (2014-15 to 2024-25). Theoretically, the domestic energy resources might let Afghanistan become Energy self-sufficient shortly [6]. According to the report, identified hydrocarbon deposits in Afghanistan can meet up to 80 percent of its requirement for petroleum products [10].

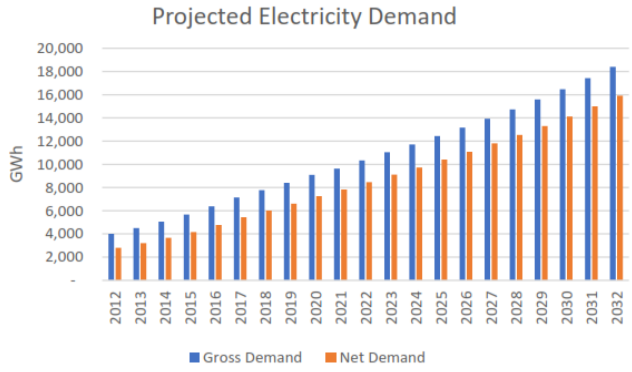


Fig 1. Afghanistan projected Electricity Demand from 2012 to 2032 [9]

2.2 Energy supply and potential

In Afghanistan, the current power supply system is deficient in various aspects, such as regional geographic coverage, flexibility, capability, and the cost of domestic supply [9]. Although the energy supply has almost tripled between 2006 and 2011, yet the energy access in Afghanistan is low. According to the Ministry of Energy and Water (MEW), only 30 percent of Afghanistan's population has access to energy. In Kabul the electrification is reaching 70-75 percent, but approximately 85% of the rural population does not have access to electricity required for daily needs [6]. The commercial electrical energy is supplied by kerosene, hydropower, and diesel. Hydropower and solar energy have the highest potential as a renewable source of energy, but the high initial cost is a significant hindrance [8]. The country relies heavily on electricity imports from neighboring countries, and only about 9% of the population has access to intermittent public power [5]. According to Da Afghanistan Breshna Sherkat (DABS), the domestic supply is 22.6 %, whereas 77.4 % electricity is purchased from neighboring countries. Uzbekistan supplies (35.3%), Tajikistan (30.7%), Iran (21.7%), and Turkmenistan (12.3%). Although Uzbekistan is the largest foreign supplier, transmission capacity is constrained [11].

Table 1: Electricity imports from neighboring countries [12]

	Apr 2013-Mar 2014	Apr 2014-Mar 2015	Apr 2015-Mar 2016
Tajikistan	947179	1137602	1179581.05
Uzbekistan	1392581	1242839	1356475.46
Turkmenistan	398586	427702	472958.96
Iran	839570	869143	831724.79
Total Imports	3577916	3677285	3840740

With the support of the International Development Organization, Afghan authorities have significantly improved access to electricity for the residence. However, people are still suffering from a deficiency of adequate and consistent supply of electricity, domestic power, and fossil fuel production. Afghanistan's grid-connected installed capacity, it is almost divided eventually between hydropower and thermal. Off-grid renewables (large hydropower, solar, wind, and biomass) offer a balanced amount of energy. Out of 623 MW installed domestic capacity, 312.5 MW is from thermal, 255.5 MW is from hydropower, and 55.0MW is from renewable energy as shown in Figure 2 [12].

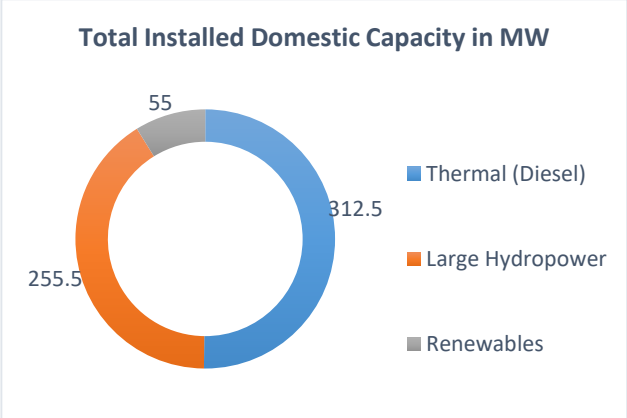


Fig 2. Sources of electricity in Afghanistan- Total Installed Domestic capacity in MW

The energy sector in Afghanistan is dominated by expensive fuel oil and diesel generation, which is costly. It can cost up to 35-40 U.S. cents per kWh and about 6 to 7 times higher the price of the electricity being imported from Central Asian countries [6]. Due to the lack of grid integration, the existing thermal generation capacity is costly to operate and increases dependence on different import sources though in recent years, the government has invested in renewable energy sources. The import bill of energy has increased 14 times from $16 millions to $224

millions from 2007 to 2015.The electric supply, i.e. domestic vs import has been shown in Figure 3.

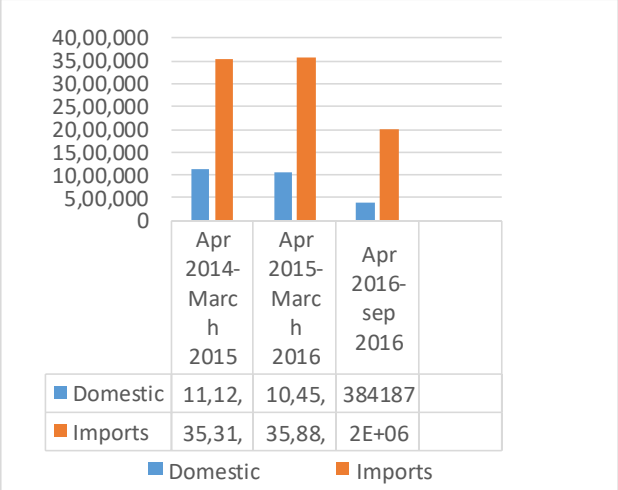


Fig 3. Electricity supply MWh (Domestic vs Imports)

Table 2 and Figure 4 gives the cost of supply, i.e. the average estimated unit price in Usc /kWh [8].

Table 2. Cost of Supply [8]

S/N	Type	Avgerage Unit Price (Usc /kWh)
1	Diesel (all provinces)	29.53
2	Thermal (NW Kabul)	27.115
3	Hydro, thermal and Diesel	6.473
4	Hydro and Diesel	5.19
5	Natural gas	2.8- 3.5
6	Imported	2.62
7	Hydro	2.29
8	Coal (1MW=4.5ton)	1ton=Afs 2.200

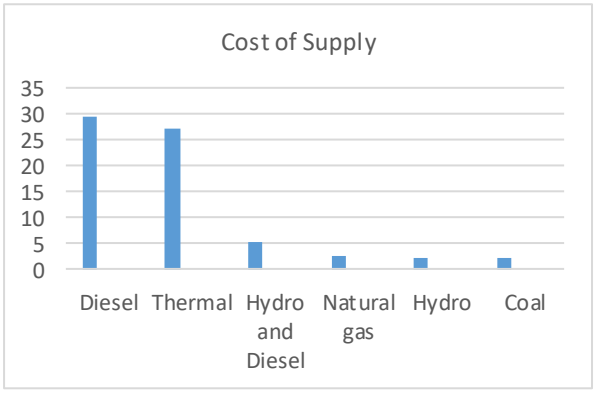


Fig 4. Cost of supply [13]

3 Transmission and distribution

The least developed part of the Afghan power system is the power transmission and distribution. The transmission system is independent grids or islands supported by various power systems and import sources. It has been noticed that between 2010 and 2015, number of household connections to the grid has increased by 60%, and the total number of connections has increased by 57%. Out of the total connections, 93% are households, and rest 7% are commercial customers and government agencies. Da Afghanistan Breshna Sherkat (DABS) is the only independent state-owned utility in the country owning all the infrastructure of the generation, distribution, and transmission in Afghanistan.

For operational purposes four major working groups link various supply sources to the grid for operational purposes, (i) NEPS (North East Power System) comprising of several small islands, connects 17 load centers which include Mazar-e-Sherif,Kabul, and Jalalabad with Uzbekistan and Tajikistan (at 110 kV, 220 kV, and 35 kV); (ii) SEPS (southeast Power System) comprising of Khandar and linking with Kajaki (110 kV) ; (iii) HS (Herat System) which links Iran with Turkmenistan (132 kV and 110 kV) ; and (iv) TS (Turkmenistan System) that links Andkhoy, Herat, Faryab, JawzJan, and Sar-e-Pul, districts (110 kV) [9].

Distribution networks in areas apart from NEPS also need to be rehabilitated as those were destroyed during conflict or have exhausted their economic life causing technical losses [5]. Hence, for the major urban centers, the main focus has been the rehabilitation and expansion of the distribution system. It is seen that each new connection is expensive and costs about $1,000 in distribution alone. Also, due to the weak distribution systems, there is a power wastage of up to 75% [14].

A focus of the development of energy sector efforts has been to upgrade Afghanistan's electricity grid from isolated islands into a national transmission grid. A unified grid will make the transmission of power generated from one part to another of the country. It will also ensure that imported power can be transmitted to the remote area of the country. Table 3 gives the existing and the planned transmission line lengths as per the source.

Table 3. Existing and planned transmission line lengths [13]

Line Voltage KV	Existing/ Km	Planned / km	Total/km
13.2	2	-	2
20	80	-	80
35	78	-	78
110	1,287	2,841	4,128
132	148	-	148
220	669	2,667	3,336
500	-	1,393	1,393
Total	**2,261**	**6,907**	**9,168**

4 Energy resources and consumers

Hydropower is one of the widely used renewable energy. Due to water resources availability and environmental benefits, hydropower plants have attracted particular attention. As discussed earlier, access to power is limited in Afghanistan though it has considerably improved in the past few years. There is a very high per capita consumption of fuelwood and charcoal due to high altitude and severe winter conditions, which has increased to more than 10 tons/year for households in some cases. The average households in the rural areas of Afghanistan frequently lack an adequate amount of fuel for cooking. In

rural areas, solid biomass is the primary fuel; on the other hand, wealthier families frequently use fossil fuel as a part of their energy portfolio. Utilization of solid fuel is related to major health issues and results in natural resource depletion. As per DABS, there was 8% increase in the number of connections by 2015, and there has been an approximately 4% year on year increase [15] [16]. Table 4 presents a number of DABS Electricity connections.

Table 4. Number of DABS Electricity connection [16]

Type of usage / Year	2013	2014	2015
Household	1,009,445	1,112,833	1,197,388
Commercial	77,980	82,467	87,694
Government	13,999	15,643	17,011
Total	1,101,424	1,210,943	1,302,093

Various factors determine choices for the energy system that includes technology, cost, accessibility, politics, demographics, convenience, and safety.

4.1 Renewable energy technologies

In Asia, renewable energy technologies viz. hydropower, solar, thermal, wind, and traditional biomass have been well-establish**.** According to the International Renewable Energy Agency (IRENA), the renewable energy power generation capacity in Asia increased from 387 550 MW in 2010 to 918 655 MW in 2017 [17]. Figure 5 depicts the total renewable energy capacity in Asia. Figure 5 comprises the renewable power generation capacity by the source in Asia during the years 2010 to 2017.

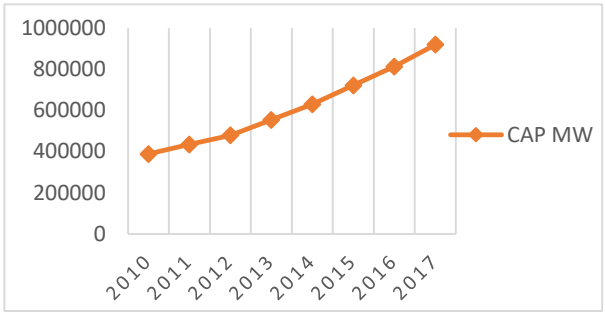


Fig 5. Total renewable energy capacity in Asia [17]

Renewable energy will be vital in the economic, social, and sustainable growth of Afghanistan. The renewable resources such as biomass, hydro, and solar are available in plenty, and their exploitation would help in improving the future supply gaps in terms of both economic and financial conditions. Figure 6 demonstrates the total renewable energy capacity and growth in Afghanistan from 2010 to 2017.

According to World Bank, Afghanistan's hydroelectric capacity 23,000 MW out of which about 87 % (20,000 MW) is located in the north-east on the Amu Darya, Panj, and Kokcha rivers. About 8 % (1,900 MW) is located to the east of Kabul, half of which is - on the Kunar river [9]. Installed Hydro projects of the capacity of up to 3 Megawatts. Table 5 shows the Hydro project's capacity [18].

Afghanistan's average solar-energy potential is about 6.5 kWh per m^2 per day, with approximately 300 days of sunshine per year. Higher values of solar potential is available in the southern areas of Kandahar, Farah, Herat, and Helmand provinces. Electricity generation is technologically feasible even in the northern provinces where average irradiance is only 4.5 kWh per m^2 per day [9].

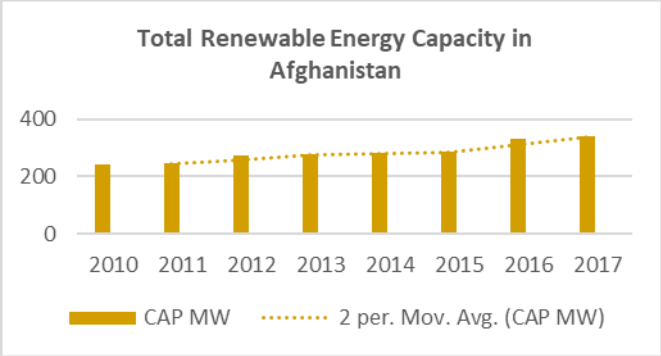


Fig 6. Total renewable energy capacity in Afghanistan, [17]

Table 5. Hydro Projects [16]

Type of Hydro project	Capacity (MW)
Small	<25000 kW
Micro	<2500 kW
Micro	<250 kW
Pico	<2.5 kW

The wind resources also have considerable potential in Afghanistan, but it is concentrated in the southwest near the Iran border. The nation's total wind energy capacity is 150,000 MW, whereas exploitable capacity is estimated to be 66,700 MW. Also, there are signs of the significant potential of geothermal and biomass, but further study of such sources of energy is required [9]. To meet the energy needs of citizens, biomass plays an essential role since fuel demand is in the form of fuelwood, charcoal, animal manure, and crop residues, mainly utilized by households. Geothermal Energy for power generation can be harvested at a low cost. Table 6 shows the potential for renewable energy resources in Afghanistan.

Table 6. Potential for Renewable Energy in Afghanistan [18]

Type of Energy	Potential
Solar	222000 Megawatts Average solar insolation of 6.5 kWh/m per day 300 days of sunlight
Hydropower	23000 Megawatts Potentials of large Dams, Mini and Micro potential each 600 Megawatts
Wind	67000 Megawatts 36000 Km 5 MW per Km, windy land
Biomass	4000 Megawatts Animal waste 840 Megatts 3090 Megawatts agriculture waste
Geo-Thermal	3000 – 3500 Megawatts 3 big possible regions 4-100 Megawatts

5 Energy policy

Afghanistan is rebuilding its energy sector. Since 2004, the Afghanistan government, along with the international

community, has been developing the National Development Strategy, and the energy sector has been in top priority. The Afghanistan National Development Strategy (ANDS) was adopted in 2008 to address the challenge of coordinating among various domestic agencies and their international partners for the transformation of the electricity sector [19].

The National Integrated Energy Policy adopted in 2016, provided broad guidelines on the scope for the energy sector in Afghanistan and made Ministry of Energy and Water (MEW) the apex agency for developing and implementing energy policy [20].

Ministry of Energy and Water (MEW) has also developed the Renewable Energy Policy, which envisages the mainstreaming of renewable energy projects in the development of the energy sector. The policy permits investors to generate electricity based on renewable energy resources at one or multiple locations and receive an equivalent amount for own use elsewhere on the grid after accounting for transmission losses and wheeling charges. The policy also deregulates small scale power production projects through renewable resources up to 100 kW [18]. Although the government has adopted energy policies and strategies for power generation and energy efficiency, yet there has been limited focus on energy conservation and has not implemented energy efficiency policy in a different sector. Additionally, cooking, and heating devices in households are based on either renewable energy sources or traditional biomass that can support the country's energy demand.

6 Challenges in reaching self-sufficiency

Every nation has a different path for achieving the energy requirements and following the transition to a sustainable energy sector. It involves a unique mixture of source opportunities and challenges related to access and efficiency. Afghanistan's energy challenges are somewhat different from other developing countries due to reasons noted earlier. Afghanistan is energy deficient nation, which can be seen from its domestic power generation capacity provided only 22 percent of country total consumption till 2015. Afghanistan is facing many challenges viz. economic and political problem due to insurgency, decreasing economic growth, and unending poverty. Also, Afghanistan's public electricity sector is facing problems related to workforce and technical capability [21].

- Afghanistan lacks policy frameworks that specify the provision of energy for sustainable development. There has been difficulty in implementing the existing policies and enforcing the laws because of poor infrastructure which poses many challenges in promoting energy efficiency.
- There is an irregular distribution in the country because of which the Afghan consumers are suffering. Presently, about 70 percent of consumers in the city of Kabul receive a continuous supply of electricity, while about two-third (67–75 percent) of the country's population do not get the regular power [22].
- The country has a limited capacity of transmission and distribution networks, that needs to be extended in order to have much energy security. The number of household connected to the electricity grid was 1,176,030 in March 2015 and has been increased by 11 percent in the past few years. Further the government's power extension policy has increased the transmission lines length from 2,261 km to 6,907 km [16].
- Accessibility to the water resources under the uncertainties of climatic - for future hydropower generation is a major apprehensions in the country's long-term energy planning [9].
- The absence of funding in gas production and transmission infrastructure limits the expansion of the gas reserves of 75 billion cubic meters to only 150 square kilometers area. Substantial renewable energy and fossil fuel resources are essential for development through private and public investment [23].
- Since several organizations are involved for the administration process, authorizing procedures and maintenance activities lead to coordination problems among the various authorities.
- There is limited public awareness and responsiveness of renewable energy technologies, social advantages, and environmental benefits along with the accessibility to inexpensive fuel energy resources.

7 Self-sufficiency opportunities

The self-sufficiency goal in the energy industry in any region can be achieved through the government's well laid - vision and objectives to produce enough energy locally. Demand stability -can be achieved through supplying even, secure, and high-quality energy services to Afghan citizen and by promoting renewable energy and energy conservation measures in different sectors.

7.1 Best use of autochthonous hydrocarbon and renewable energy resources

Afghanistan has plentiful fossil fuel and renewable resources that need to be exploited for meeting the energy requirements. The adoption of national and regional policies can play an essential role in supporting autochthonous hydrocarbon and renewable energy development and implementation. It will help the nation in identifying the priorities and ways to construct a sustainable and cost-effective energy market.

Renewable energy, including hydro, solar, wind, geothermal, biomass, and wood offers the greatest hope for Afghanistan. However, despite of its potential, hydropower is still untapped in many parts of Afghanistan. The country's annual renewable surface water resources are estimated at 57 billions $m^{3,}$ which is distributed across five river basins. Nonetheless, the water resources are unevenly distributed across the country or not equally accessible - throughout the year [24][25] There has been evidence that geothermal and biomass have considerable potential [9]. Providentially, the government, with the support of the international community, is setting specific targets for the renewable energy sector by supporting national and international companies to meet the country's energy demand and reach self-sufficiency in the near future.

7.2 Energy efficiency and conservation

Developed countries have proven that energy can be well-preserved by reducing the amount of wastage and losses, enhancing efficiency through technological advancements, and improved operation and maintenance [26]. The effective way to fill the energy demand gap in Afghanistan would prompt energy efficiency through rehabilitation, implementation of useful technologies, and construction of energy-efficient cities. This will decrease the energy consumption and lead to cost-saving after an off-set from the initial investment cost. Other than the financial benefits, energy-saving may contribute to environment protection through the reduction of emission of greenhouse gases. Energy conservation decreases the energy demands which in turn helps in increasing the quality of the environment, national security, personal financial security and cost savings. Therefore, the government needs to improve energy efficiency by focusing on application mechanisms in different sectors, and should follow energy efficiency policies more conscientious energy efficiency in the building sector will play a significant role in energy conservation since buildings consume a high level of energy. Along with other initiatives, enhancing energy efficiency is the cheapest, fastest and most environmental friendly way to meet a substantial amount of Afghanistan's energy requirements and will, therefore, reduce the need for investing in the energy supply.

8 Conclusion and recommendation

Currently, Afghanistan is not an Energy self-sufficient country. However, the Afghan government, along with the international communities is exploring the possibilities to achieve it.. -Afghanistan's energy status can be described as a prolonged 'energy poverty,' i.e., lack of access to modern energy services. Energy access is low in the country which has - considerably improved in the last few years. Renewable energy resources are abundant, such as hydro, wind and solar, but they are still untapped. The costs of exploitation of such resources have fallen in recent years, and hence can be explored by the government. Access to reasonable and uninterruptible electricity and primary energy supply is of great importance for the economic and social growth - of the Afghans. Further, In order to devise effective policy and planning, the Government of Afghanistan, donors, private sector actors, and civil society organizations need access to quality information and data about the current energy landscape in Afghanistan.

Electrifying is a long and challenging process that necessitates political stability, appropriate legislation and regulation, and international support. To provide electricity, the government must take up the opportunity to overcome the barriers by utilizing the ability to learn from the experience of other developing countries. The energy services should be reasonable, consistent, safe and environmental-friendly.

There are some strategies that can be taken for the betterment of the energy sector :

- Power generation strategy must focus on low cost energy generation, capacity utilization optimization, regulating the input cost, optimization of fuel mixture and its technology, and adoption of renewable energy sources.
- Proper implementation can help in improving the - reliability of the energy system of the country.
- The strategy should conserve and optimize the utilization of electricity with a focus on demand management, load management and up-gradation of the technology by providing energy-efficient equipment
- Regulation -should aim at -balancing the - interest of all stakeholders and making energy sector commercially viable.
- For the growth of the energy sector, finance-related strategies must be adopted.
- Local manufacturing of renewable energy product must be implemented at the central level.
- The government should provide an energy planning process to create competition and stability for domestic and foreign investors.
- Promote energy efficiency by proper management of utilities distribution systems.
- Government and private sector units along with community should work together for achieving the long-term energy efficiency initiatives benefitting its citizens.
- A large amount of energy is consumed in buildings; therefore energy efficiency strategies should strictly implement in buildings.

References

1. M. A. Schreurs, J. Balanowski, *Promoting socially and economically just energy transformations in Asia possibilities, challenges and the road ahead*. Hanoi, (2017).
2. CSO, *Afghanistan statistical yearbook 2016–2017*. Kabul : Central Statistics, (2017).
3. World Bank, *Afghanistan development brief, the implications of Afghanistan's 2014 security transition.* World Bank Report, (2019).
4. NEPA, *Second national communication.* Kabul: national environmental protection agency, 2-30, (2017).
5. ADB, Proposed multi-tranche financing facility Islamic republic of Afghanistan: energy supply improvement investment program. *Manila : report and recommendation of the president to the board of directors*, (2015).
6. D. Bochkarev, *Afghanistan reconnected linking energy supplies to consumers in Asia*. Eastwest Institute, (2014).
7. MEW, *Afghanistan investment opportunities in energy sector*. Kabul : Da Afghanistan Breshna Sherkat, (2017).
8. ANDS, Energy Sector Strategy (2007/08 – 2012/13). Afghanistan National Development Strategy Secretariat, (2008).
9. World Bank, *Afghanistan renewable energy development issues and options*. Washington, D.C, (2018).
10. World Bank, *Afganistan economic updated*, Washington D.C, (2013).

11. DABS, *Afghanistan energy information center*. Kabul : Da Afghanistan Breshna Sherkat, (2016).
12. ICE, inter-ministerial commission for energy, electricity imports (2016). Available on https://Sites.Google.Com/Site/Iceafghanistan/Electricity-Supply/Electricity-Imports)
13. ICE, *Inter-ministerial commission for energy transmission and distribution*, (2015). Available on Https://Sites.Google.Com/Site/Iceafghanistan/Transmission-And-Distribution-Development
14. MEW, *Draft power sector strategy for ANDS*. Kabul : ministry of energy and water, (2007).
15. World Bank, *Afghanistan overview*. Kabul: The World Bank. (2013).
16. ICE, *Number of DABS electricity connection. inter-ministerial commission for energy*, (2016). Available on Https://Sites.Google.Com/Site/Iceafghanistan/Electricity-Supply/Connections)
17. IRENA, *Renewable capacity statistics.* Masdar : Internaitonal renewable energy agency, (2018).
18. MEW, *Renewable energy policy*. Kabul: Ministry of Energy and water, (2015).
19. ADB, Power sector master plan, islamic republic of Afghanistan. Asian development bank. (2010) Available on: Https://Www.Adb.Org/Sites/Default/Files/Project-Document/76570/43497-012-Afg-Tacr.Pdf
20. MEW, *Ministry of Energy and water*. Kabul: Afghanistan national integrated energy policy,, (2016).
21. USAID, *Advisor to the secretariat of the inter-ministerial commission for Energy*, Kabul: United States aid for international development , (2010).
22. MEW. Call for expression of interest (EoI) for implementation of 100 Mw grid connected.Kabul : ministry of energy and water, Renewal, (2016).
23. ADB, Sector assessment, energy.Asian development bank, (2015).
24. NEPA, Second national communication. Kabul : national environmental protection agency, 2-30, (2017).
25. MEW, Afghanistan investment opportunities in energy sector, Kabul : DABS, (2017).
26. A. Al-Mofleh, S. Taib, M. A. Mujeebu, W. Salah, Analysis of sectoral energy conservation in Malysia. J. Energy, (34) 733-739, (2009).

2020 6th International Conference on Control, Automation and Robotics

Analog Inverse Filters Using OTAs

Jyoti Srivastava
Electrical Engineering Department
Delhi Technological University
Delhi, India
e-mail: jyotisrivastava_mt2k18@dtu.ac.in

Ram Bhagat
Electrical Engineering Department
Delhi Technological University
Delhi, India
e-mail: ram.bhagat@dtu.ac.in

Pragati Kumar
Electrical Engineering Department Delhi Technological University
Delhi, India
e-mail: pragatikumar@dtu.ac.in

***Abstract*—In this paper three new tunable voltage-mode analog inverse filter circuits using operational transconductance amplifiers (OTAs) and fractional capacitors (FCs) are presented. The proposed structures realize fractional order inverse high pass filter response, fractional order inverse band pass filter response and fractional order inverse band elimination filter response employing three single output OTAs and two FCs. The workability of these inverse filter configurations has been confirmed through Cadence Virtuoso simulation tool using 0.18µm CMOS technology parameters.**

Keywords-operational transconductance amplifiers; analog inverse filters; voltage mode; analog signal processing

I. INTRODUCTION

An analog inverse filter provides frequency selective characteristics which are inverse of the conventional analog filter. Inverse filters remove the consequences of distortion in the signal while processing it through system. Analog inverse filters have potential usages in communication systems, control systems, instrumentation and measurement systems. Over the years, many inverse filter circuits realized with different active elements have been proposed in open literature [1]-[20]. Among these reported inverse filter configurations, only few utilise FCs [1], [7], [8]. In [1], two fractional order inverse filter circuits employing operational amplifier were proposed, in which one of the circuits used one operational amplifier and five admittances whereas the other circuit employed one operational amplifier and seven admittances to realize different fractional order inverse filter responses, namely, fractional order inverse high pass response, fractional order inverse band pass response, and fractional order inverse low pass response. In [7], three fractional order inverse filters, namely, fractional order inverse low pass filter, fractional order inverse high pass filter, and fractional order inverse band pass filter employing two current feedback operational amplifiers (CFOAs) were presented. The fractional order inverse low pass filter circuit used four resistors and two FCs, fractional order inverse high pass filter circuit used two resistors, four FCs and fractional order inverse band pass filter circuit utilized three resistors and three FCs. Two operational transresistance amplifiers (OTRAs) and eight admittances based fractional order inverse filter was reported in [8]. The circuits presented therein can provide fractional order inverse low pass filter response, fractional order inverse high pass filter response, fractional order inverse band pass filter response and fractional order inverse notch filter response. From the above literature survey, it may be noted that there are very few inverse filter configurations reported in the open literature which use FCs. Also, the reported circuits generally use a very large number of passive elements and do not have any provision for control of different parameters of the realized circuits. Therefore, in the following, we present three new fractional order analog inverse filter circuits employing three OTAs and two FCs only. The maximum/minimum frequency (ω_m) and bandwidth of the presented fractional order inverse band pass filter and fractional order inverse band elimination filter can be tuned orthogonally by changing the transconductance of different OTAs. The high frequency gain of fractional order inverse high pass filter can be changed independently while its cut-off frequency (ω_h) and pole quality factor (Q) can be changed orthogonally. The performance of all the presented circuits has been evaluated using Cadence Virtuoso simulation tools with 0.18µm technology parameters.

II. PROPOSED FRACTIONAL ORDER VOLTAGE-MODE INVERSE FILTER CONFIGURATIONS

All the proposed fractional order inverse filter circuits employ three OTAs and two FCs. An OTA represented symbolically in Fig. 1 is characterized by the following terminal relationship between its input differential voltage and output current

978-1-7281-6139-6/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:19 UTC from IEEE Xplore. Restrictions apply.

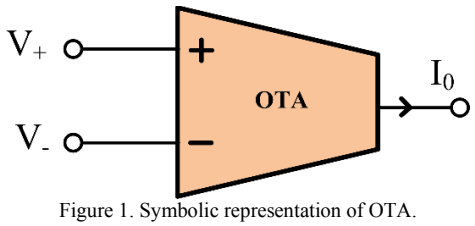


Figure 1. Symbolic representation of OTA.

$$I_{out} = g_m(V_+ - V_-) \tag{1}$$

The fractional order capacitor on the other hand is characterized by the following driving point impedance :

$$Z(s) = \frac{1}{Cs^{\alpha}}\Omega \tag{2}$$

The Fig. 2(a-c) depict the different fractional order inverse filter circuits proposed in this paper.

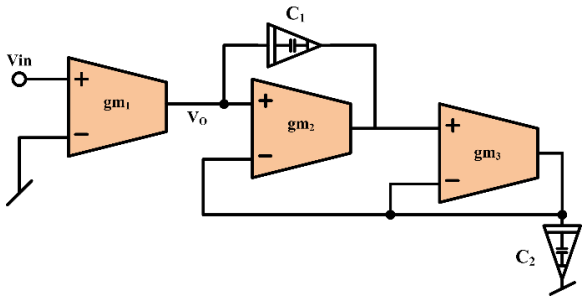


Figure 2(a). Fractional order inverse high pass filter.

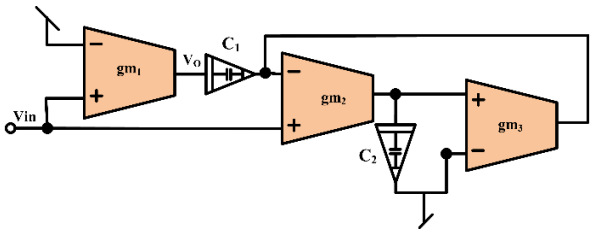


Figure 2(b). Fractional order inverse band pass filter.

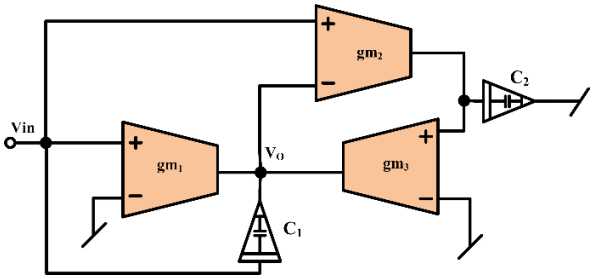


Figure 2(c). Fractional inverse band elimination filter.

Analyzing the circuits shown above using terminal characteristics of the OTAs and the fractional order capacitors given in equations (1)-(2) the transfer functions (3)-(5) for the various fractional order inverse filter circuits are obtained as:

(i) Fractional order inverse high pass filter

$$\frac{V_0(s)}{V_{IN}(s)} = \frac{s^{2\alpha} + a_0 s^{\alpha} + b_0}{s^{2\alpha} c_0} \tag{3}$$

$a_0 = g_3/C_2$, $b_0 = g_2g_3/C_1C_2$ and $c_0 = g_2/g_1$

(ii) Fractional order inverse band pass filter

$$\frac{V_0(s)}{V_{IN}(s)} = \frac{s^{2\alpha} + a_1 s^{\alpha} + b_1}{s^{\alpha} c_1} \tag{4}$$

$a_1 = g_2g_3/g_1C_2$, $b_1 = g_2g_3/C_1C_2$ and $c_1 = g_2g_3/g_1C_2$

(iii) Fractional order inverse band elimination filter

$$\frac{V_0(s)}{V_{IN}(s)} = \frac{s^{2\alpha} + a_2 s^{\alpha} + b_2}{s^{2\alpha} + c_2} \tag{5}$$

$a_2 = g_1/C_1$, $b_2 = g_2g_3/C_1C_2$ and $c_2 = g_2g_3/C_1C_2$.

The cut off frequency (ω_h) and maximum/minimum frequency (ω_m) of fractional order inverse high pass filter, fractional order inverse band pass filter/ fractional order inverse band elimination filter, correspond to these frequencies for their normal filter counterparts and may be found by solving the following non-linear equations [24]:

$$-\omega_h^{4\alpha} + 2a_0\cos\frac{\pi\alpha}{2}\omega_h^{3\alpha} + a_0{}^2(1+2\cos\pi\alpha)\omega_h^{2\alpha} + 2a_0{}^3\cos\frac{\pi\alpha}{2}\omega_h^{\alpha} + a_0{}^4 = 0 \tag{6}$$

$$(\omega_m^{2\alpha} - b_1)(\omega_m^{2\alpha} + b_1 + a_1\cos\frac{\pi\alpha}{2}) = 0 \tag{7}$$

$$(\omega_m^{2\alpha} - b_2)(\cos\frac{\pi\alpha}{2}\omega_m^{4\alpha} + a_2\omega_m^{3\alpha} + \{4b_2\cos\frac{\pi\alpha}{2} - 2b_2\cos\pi\alpha\cos\frac{\pi\alpha}{2}\}\omega_m^{2\alpha} + a_2b_2\omega_m^{\alpha} + b_2^2\cos\frac{\pi\alpha}{2}) = 0 \tag{8}$$

The theoretical value of maximum/minimum frequency of band pass and band elimination filter can be calculated from $\omega_m = (b_1)^{1/\alpha}$ and $\omega_m = (b_2)^{1/\alpha}$ respectively.

Table I shows the values of the $\omega_{m/h}$, H and Q for $\alpha = 1$. From Table I, it may be observed that H for fractional order inverse high pass filter can be varied by changing the value of g_1 while its ω_h and Q can be tuned orthogonally. On the other hand ω_m and BW of fractional order inverse band pass filter and fractional order inverse band elimination filter can be tuned orthogonally.

TABLE I. CUTOFF FREQUENCY, GAIN AND POLE QUALITY FACTOR FOR α=1.

	H	Frequency	BW	Q
Inverse high pass filter	$\frac{g_2}{g_1}$	$\omega_h = \sqrt{\frac{g_3g_2}{C_2C_1}}$	-	$\sqrt{\frac{g_2C_2}{C_1g_3}}$
Inverse band pass filter	1	$\omega_m = \sqrt{\frac{g_3g_2}{C_2C_1}}$	$\frac{g_3g_2}{g_1C_2}$	$g_1\sqrt{\frac{C_2}{g_2g_3C_1}}$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:19 UTC from IEEE Xplore. Restrictions apply.

Inverse band eliminatio n filter	1	$\omega_m = \sqrt{\frac{g_3 g_2}{C_2 C_1}}$	$\frac{g_1}{C_1}$	$\frac{1}{g_1}\sqrt{\frac{C_1 g_2 g_3}{C_2}}$

III. Simulation Results

The performance of the proposed circuits has been validated through Analog Design Environment (ADE) tool of Cadence Virtuoso with 0.18µm CMOS technology parameters. The CMOS OTA [23] used in this paper is shown in Fig. 3 and aspect ratio used are taken from [25]. The DC power supply of ±0.9V were used in the simulations.

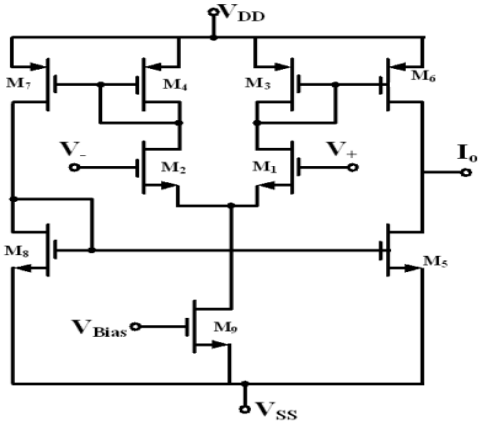
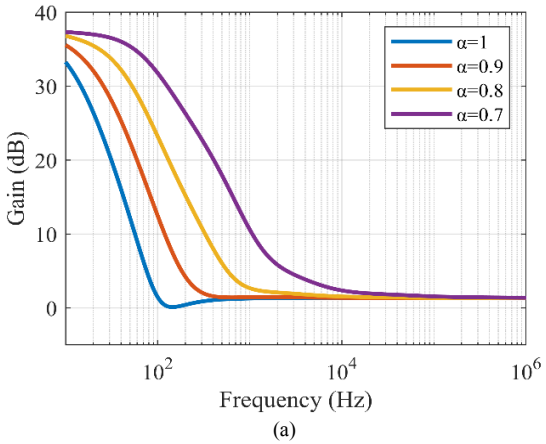
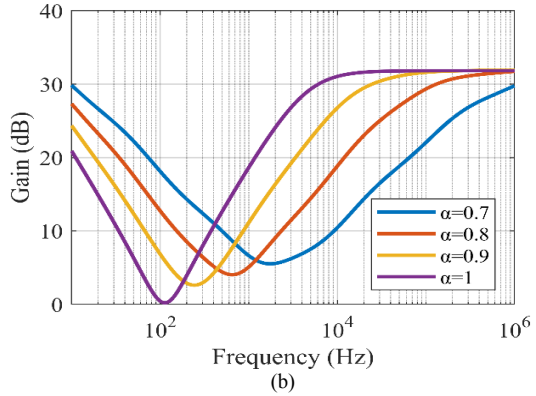


Figure 3. CMOS implementation of OTA.

We have used two identical valued fractional order capacitors of 0.382 µF (rad/sec)$^{(1-\alpha)}$ [1] designed using Valsa method [24]. Different values of α (0.7, 0.8, 0.9 and 1) have been used in simulations. The fractional order inverse filters were designed for a ω_h/ω_m of 113 Hz ($\alpha = 1$). The value of α was changed from 0.7 to 1.0 in step of 0.1. The frequency response of fractional order inverse high pass filter, fractional order inverse band pass filter and fractional order inverse band elimination filter have been displayed in Fig. 4(a)-(c) respectively.





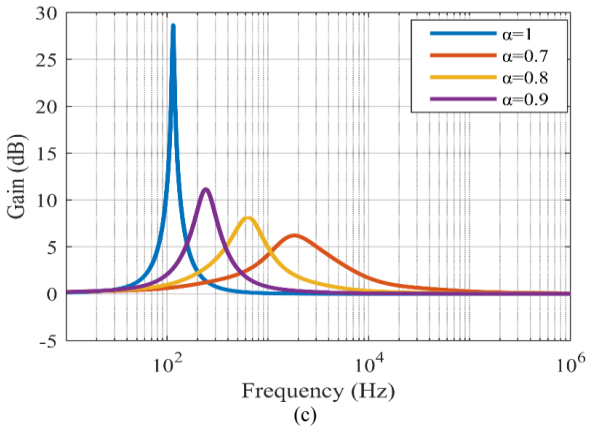


Figure 4. Frequency responses of (a) fractional order inverse high pass filter, (b) fractional order inverse band pass filter and (c) fractional order inverse band elimination filter.

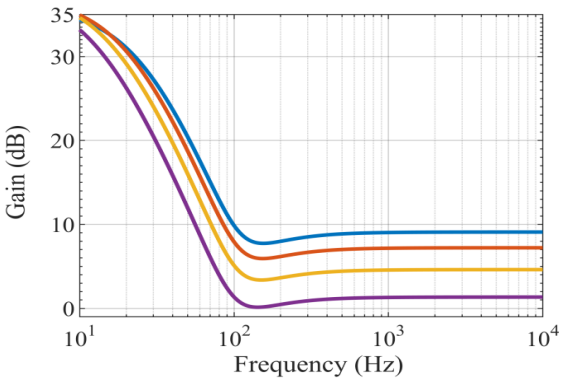


Figure 5. Tunability of gain for fractional order inverse high pass filter.

In Fig. 5, the electronic tunability of gain for fractional order inverse high pass filter has been shown by varying transconductance g_1 for 715.549µS, 516.164µS, 275.038µS and 152.162µS. The Q of the fractional order inverse band pass filter and inverse band elimination filter have been varied for constant frequency 130 Hz by varying transconductance g_1 (715.549µS, 516.164µS, 275.038µS and 152.162µS) and their respective results have been illustrated in Fig. 6 and Fig. 7.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:19 UTC from IEEE Xplore. Restrictions apply.

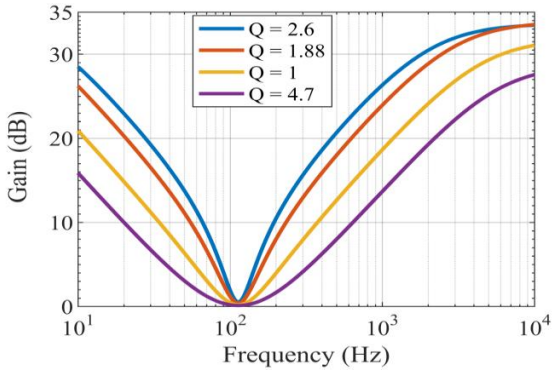


Figure 6. Tunability of Q for fractional order inverse band pass filter.

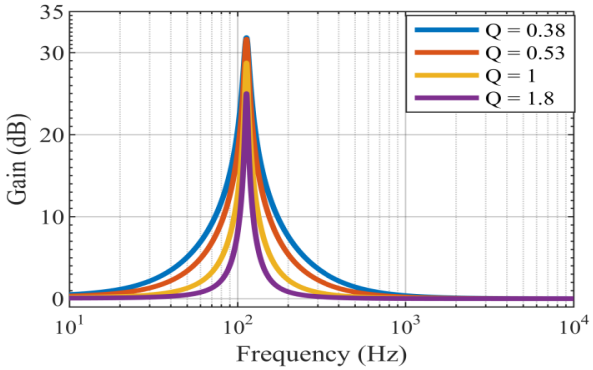


Figure 7. Tunability of Q for fractional order inverse band elimination filter.

Table II gives the values of the ω_h, ω_m of the presented fractional order inverse filters measured through simulations and calculated theoretically for different values of α.

TABLE II .CUT-OFF FREQUENCIES (IN HZ) OF FILTERS FOR DIFFERENT α

α	**1**	**0.9**	**0.8**	**0.7**
Inverse high pass (Simulated)	131.8	204.2	676	3715
Inverse high pass (Theoretical)	113.05	217.293	709.056	3349.88
Inverse band pass (Simulated)	112.2	234.4	660.7	1738
Inverse band pass (Theoretical)	113.05	234.479	583.63	1885.08
Inverse band elimination (Simulated)	114	239.9	631	1820
Inverse band elimination (Theoretical)	113.05	234.479	583.63	1885.08

IV. CONCLUSION

Three new VM analog inverse filter configurations employing three OTAs and two fractional capacitors have been presented. The proposed circuits are capable of providing fractional order inverse high pass response, fractional order inverse band pass response and fractional order inverse band elimination response. The presented fractional order inverse band pass filter and fractional order inverse band elimination filter circuits have orthogonal tunability of ω_m and bandwidth. The gain of the fractional order inverse high pass filter can be tuned independently while its ω_h and Q can be changed orthogonally. Simulation results have confirmed the working of these fractional order inverse filter circuits.

REFERENCES

[1] D. R. Bhaskar, M. Kumar and P. Kumar, "Fractional order inverse filters using operational amplifier," Analog Integer Circuits Signal Process, Vol. 97, no. 1, pp. 149-158, Oct. 2018.

[2] S. S. Gupta, D. R. Bhaskar, and R. Senani, A.K. Singh, "Inverse active filters employing CFOA," Elect. Eng. 1, pp 23-26, (2009).

[3] S. S. Gupta, D.R. Bhaskar, and R. Senani, "New analogue inverse filters realized with current feedback op-amp," Int. J. Electron.9, pp. 1103–1113, 2011.

[4] H. Y. Wang, S.H. Chang, T.Y. Yang, and P.Y. Tsai, "A novel multifunction CFOA based inverse filter", Circuits and Syst.2, pp. 14–17, 2011.

[5] V. N. Patil, and R. K. Sharma, "Novel inverse active filters employing CFOA", Int. J. for Scientific Research & Develop. 3, pp. 359–360, 2015.

[6] T. Tsukutani, Y. Kunugasa, and N. Yabuki, "CCII-Based Inverse Active Filters with Grounded Passive Components," J. Electr. Eng. 6, pp. 212-215, 2018.

[7] E. M. Hamed, L. A. Said, A. H. Madian, and A. G. Radwan. "On the Approximations of CFOA-Based Fractional-Order Inverse Filters," Circuits Systems and Signal Processing, pp.1-28, 2019.

[8] N. A. Khali, L. A. Said, A. G. Radwan, and A. M. Soliman. "Multifunction Fractional Inverse Filter Based on OTRA," in 2019 Novel Intelligent and Leading Emerging Sciences Conference (NILES), vol. 1, pp. 162-165. IEEE.

[9] R. Bhagat, D. R. Bhaskar, and P. Kumar,"Inverse Band Reject and All Pass Filter Structure Employing CMOS CDBAs," International Journal of Engineering Research & Technology (IJERT), pp. 39-44, 2019.

[10] K. Garg, R. Bhagat, and B. Jaint, "A novel multifunction modified CFOA based inverse filter," In 2012 Power Electronics (IICPE), IEEE 5th India International Conference., pp. 1-5.

[11] N. A. Shah, and M. F. Rather, "Realization of voltage-mode CCII based all pass filter and its inversion version," Indian J. Pure & Applied Physics, 44, pp. 269–271, 2006.

[12] B. Chipipop, and W. Surakampontorn, "Realization of current-mode FTFN-based inverse filter," Electron. Lett. 35, pp. 690–692, 1999.

[13] H. Y. Wang, and C.T. Lee, "Using nullors for realisation of current mode FTFN based inverse filters," Electron. Lett. Volume 35, Issue 22, pp. 1889-1890. 28 October 1999.

[14] A. K. Singh, A. Gupta, and R. Senani, "OTRA-based multi-function inverse filter configuration," Adv. in Elect. and Electron. Eng. 15, pp. 846-856, 2018.

[15] A. Pradhan, and R. K. Sharma, "Generation of OTRA-Based Inverse All Pass and Inverse Band Reject Filters," Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, pp. 1-11, 2019

[16] A. R. Nasir, and S. N. Ahmad, "A new current mode multifunction inverse filter using CDBA," International Journal of Computer Science and Information Security 11, no. 12, pp. 50-52, 2013.

[17] R. Pandey, N. Pandey, and T. Negi, V. Garg, "CDBA based universal inverse filter," ISRN Electronics, pp. 1-6, 2013.

[18] A. Sharma, A. Kumar, and P. Whig, "On performance of CDTA based novel analog inverse low pass filter using 0.35 µm CMOS parameter," International Journal of Science, Technology & Management 4, pp. 594–601, 2015.

[19] N. A. Shah, M. Quadri, and S. Z. Iqbal, "High output impedance current-mode all pass inverse filter using CDTA," Indian Journals of pure and applied physics vol. 46, pp. 893–896, 2008.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:19 UTC from IEEE Xplore. Restrictions apply.

[20] T. Tsukutani, Y. Sumi, and N. Yabuki, “Electronically tunable inverse active filters employing OTAs and grounded capacitors,” International Journal of Electronics Letters 4, pp. 166–176, April 2016.

[21] R. Verma, N. Pandey, and R. Pandey. “Electronically tunable fractional order filter,” Arabian Journal for Science and Engineering, 42(8), pp. 3409-3422, 2017.

[22] A. G. Radwan, A. S. Elwakil, and A. M Soliman, “On the generalization of second-order filters to the fractional-order domain,” Journal of Circuits, Systems, and Computers 18, pp. 361-386, April 2009.

[23] E.W. Greeneich, Analog Integrated Circuits, Chapman & Hall, 1941, pp. 192.

[24] J. Valsa, P. Dvorak, and M. Friedl, “Network Model of the CPE," Radioengineering, Vol. 20, pp. 619–626, Sep. 2011

[25] D. R. Bhaskar, A. Raj and P. Kumar, “Mixed Mode Universal Biquad Filter Using OTAs,” Journal of Circuits Systems and Computers, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:19 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Analysing the Effective Psychological State of Students using Facial Features

Shailender Kumar
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

Deepanshu Varshney
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
deepanshu.tcl@gmail.com

Gaurav Dhawan
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
grvdhwn8527@gmail.com

Harsh Jalutharia
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
harshjalutharia@gmail.com

***Abstract*— The psychological state of college students plays an important role in their academic performance. Negligence of one's mental and emotional health can lead to severe problems such as anxiety, stress, depression, etc. Quick detection and treatment of these problems are essential for the mental health of students. Depression detection in college students is very tough, largely due to the sheer population of students, along with the fact that they may not be aware of their mental problems. Furthermore, because of the stigma surrounding mental health, students may be concealing their psychological issues from everyone. An automatic depression detection system that can predict if students are dealing with depression is required. The proposed system uses video footage of college students to extract the facial features by use of the KLT Algorithm (Face-Detection/Tracking) and Gabor Filters to examine them with an intent to identify symptoms of depression. The final step is to classify the students as having depression or not by utilizing SVM based classifier. The system has been trained with datasets having images of faces along with the classification of whether they are contempt, disgusted, or happy. The level of these emotions in the extracted frames is analyzed to detect depression.**

I. INTRODUCTION

A person is considered to be Depressed when they experience a prolonged duration of negative thoughts, which instill them with a sense of being helpless and left out. Such people find it difficult to sail through daily life, social events cause anxiety to them. As a result, they begin giving up or limiting any social interactions [3]. Depression can ruin the thought process and make a person highly vulnerable. For students who need to be focused on building a career, depression can prove to be a plague.

A student in college is often made to face immense pressure and challenging assessments, which usually leave them feeling overwhelmed. The duty of self-management in terms of time, food, general lifestyle sometimes makes them feel burdened and homesick. Added to the above, lack of resources or failed relationships especially at such a tender age has led cases of Clinical Depression to shoot up.

One study was performed to find an association between the use of the internet by students and signs of depression. The study found that students suffering from depression had higher internet usage as compared to non-depressed students [3]. The rise in popularity of social media sites has led to teenagers feeling down by making them believe that they aren't living a successful or happier life as compared to their peers. Students can also become victims of cyber-bullying which may cause depression.

The stigma around mental illness and restrictive thinking is a major cause for affected students to refrain from treatment or even acceptance (WHO). Also, students are often fed with forced optimism which can further delay the diagnosis and hence treatment. Ultimately an illness, depression worsens with time. Therefore, early detection holds massive importance as far as recovery is concerned.

In the proposed system, the goal is to catalyze the detection of depression amongst students. With early detection, preliminary steps like meditation and counseling sessions have proven to be greatly effective in complete rectification of the illness before it gets out of hand.

The system works on snapshots captured at regular intervals from video footage of students. The underlying idea is to utilize facial muscle aesthetics or expressions to evaluate if an individual is depressed or not. The process involves sub-processes of face identification, expression analysis, and categorization as depressed or not. It is initially fed with pre-labeled data to recognize features associated with depression.

II. RELATED WORK

A. Expressions and Depression

An important characteristic of a depressed person is the kind of facial expressions they make. A range of studies back this association between the two. A specific system established for identifying such expressions involves discovering Action Units [12]. With entries of depressed individuals in the dataset, it was noticed that AU10(Disgust) and AU14(Contempt) occurred in abundance along with AU12(Smile) in minimal to very few cases. The Dataset used had information for both nondepressed and depressed individuals. The study concluded that Action Unit 14 seemed to be very effective in classification.

An important aspect of depression is that it hampers an individual's thought process making them feel constantly consumed and lethargic. Extending this observation to college students, often such students seem to be least concentrated and awake during lectures. If in a scenario we have a group of students, who exhibit somewhat similar expressions

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:13 UTC from IEEE Xplore. Restrictions apply.

(Positive/Negative). Speculation about the context of the situation can be made [14]. If the behaviour of students in class is associated with their facial expressions, a classification can be done on whether they are depressed or not [7].

Eye Movement is another important aspect. The posture of a person's head and characteristic eye movements of depressed individuals have been discussed in [9]. The use of features related to eyes has proven to be a better means of classifying depressed people. It evaluates multi-scale entropy (MSE) on the videos included in the dataset [4]. It is the measure of variation happening per pixel of a video. MSE values for individuals (non-depressed) who seem to be quite expressive were evaluated to be high as compared to depressed people.

B. Depression Analysis Datasets

Some datasets which deal with depression analysis are Pitt (created by the University of Pittsburgh), BlackDog (created by the BlackDog Institute), etc. One study [9] analyzed the aforementioned datasets to detect depression by combining them and training an SVM classifier. Their study found the best results when a higher variety of training data were used.

A study proposed another approach which performed facial geometry analysis for detecting depression [10]. Their work found that depression-related signs were low in frequency in videos of a smaller duration. Thus, detecting depression in longer videos improves efficiency. Data was collected by recording the subjects during a consultation with a professional. Dataset was created from diagnosis until the subject showed visible improvement [12][4]. One study found significant relations between vocal behavior and the facial features of depressed people [2].

In another study, subjects were provided with wearable devices to monitor the physical health, social interaction, and emotional behavior of the subjects to identify depression [5]. Researchers have also collected data by making subjects perform a simple task of distinguishing positive and negative facial expressions from various images [11]. One study found that for detecting depression, analyzing the entire video together gave better results than analyzing videos frame by frame [13]. This was done by first initializing the subject's face manually. Then, the face is tracked during the video using the KLT tracker (Kanade-Tomasi-Lucas). The tracker works by extracting curvature information from the image. This approach exhibited better accuracy because it can detect small movements in the face of the subject.

C. Methods for Extracting Features

When considering data that is in the form of video footage or audio recording/tracks, a measure called the Motion History Histogram (MHH) is used to signify slight changes happening in a person's face or voice characteristics [8]. An approach called the Gabor Wavelet Technique has been proposed [1] for evaluating the facial features of individuals. The discriminating attribute of filters used for feature extraction in this technique is that they are not considerate towards the Posture or Orientation. The features used are arithmetic average or standard deviation of around 40 Gabor filters [6]. Then SVM classifier is used for classification.

For facial features, the system utilizes the popular Kanade Lucas Tomasi Algorithm. The KLT algorithm's working is based on the change in the optical position of points from one frame to another [15]. The technique involves tracking of various distinct points. A moving object can be detected easily by shortlisting specific points from one frame and then locating the same points again in another frame. In the proposed system this method is used to track human face. For the very first frame, it utilizes the Harris Corners Detection method to find an initial set of corners to track. This is followed by the KLT algorithm tracking the optical flow of corners, this is done by evaluating the movement of pixels. KLT Algorithm is still considered to be widely used because of its automatic operation and its performance is competitive in terms of feature point quality.

III. PROPOSED SYSTEM

The use of facial expressions for the identification of a depressed person can be a challenging task. The properties of a depressed face can be easily mimicked by a flushed or sad face. Therefore, it becomes extremely crucial to draw a line of distinction to differentiate the two. A discriminating feature could be a notice of eyes pointing downwards, which signifies being let down or helpless in case of sadness. Another feature is a frown on the face, again in case of sadness. Other cases involve situations where a depressed person intentionally fakes being happy to hide their feelings or they may even appear expressionless.

The majority of data available for the proposed field contains data used for identifying depression amongst adults. Therefore, there is a potential for the establishment of a system that is dedicated to a more vulnerable class of humans; college students. As already mentioned, college students remain highly susceptible to mental illnesses for a variety of reasons.

The proposed system is designed to work on video footage of students. The training is done via features extracted from images showing disgusted, happy, and contempt faces. During testing, to maintain uniformity and to ease the task of feature extraction, the nature of videos is kept very similar. The footage is taken while students try answering a quiz given to them. This is followed by the extraction of features related to face and their normalization (to minimize redundancy) for the complete footage. The final step involves evaluating features for test data using an SVM classifier to conclude whether the student is depressed or not.

The underlying idea is to note the overall occurrence of 3 types of faces in the footages (disgust/happy/contempt). Based on the relative comparison of these 3 measures, a classification would be made to the extent of depression expected in a person. In cases where a measure of disgust/contempt is large, a conclusion can be made about being strongly depressed. Similarly, if the measure of disgust/contempt is moderate a conclusion of light/weak depression can be done and not depressed if a measure of disgust/contempt is negligible or comparatively very small.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:13 UTC from IEEE Xplore. Restrictions apply.

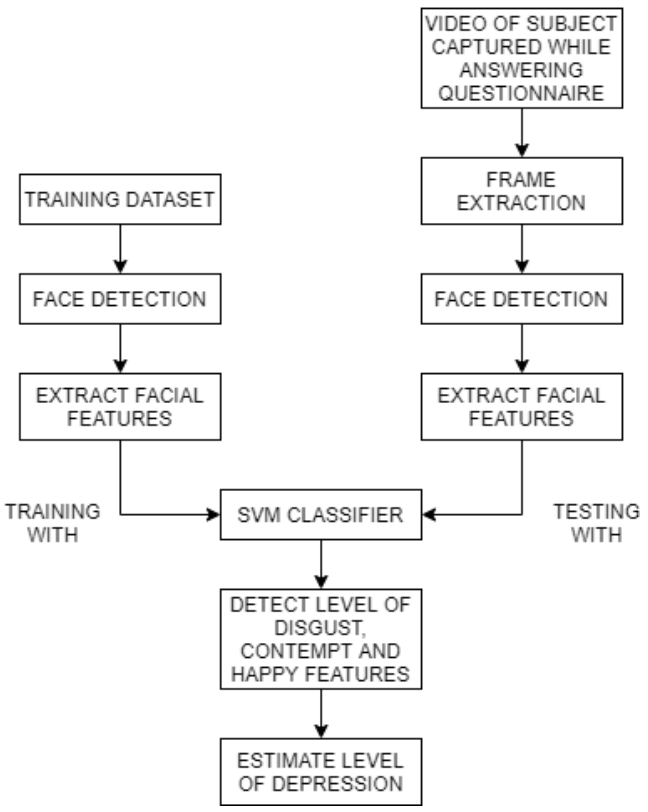


Fig. 1. Proposed System Architecture

The proposed system's architecture as shown in Fig. 1 work in 3 phases:

D. Training Phase

1) Input Dataset: The underlying concept is to detect the measure of negativity in the selected dataset. As stated earlier, the features related to depression had low counts of a smile (Action Unit 12). If a smile was present, it also shows the existence of disgust (Action Unit 10) and contempt (Action Unit 14) in hand with the obvious detection of a smile (Action Unit 12). To solve this, we can look for the existence of happy, disgust and contempt features, that is, we examine if the subject smiles frequently or smiles with contempt and disgust, or no smile entirely. To achieve this, we measure the disgust, contempt, and happy facial features in various images. Thus, the final dataset for training consists of disgust, contempt, and happy faces.

2) Face Detection: Detection of a face needs to be done before performing feature extraction. Detection of faces is performed using the Kanade-Lucas-Tomasi algorithm. This algorithm is proved to have high True-positive rates under different exposure settings. It can be used to effectively detect nose, eyes, mouth, etc. More than 100,000 features were extracted. Elimination of irrelevant features is done so that relevant features remain.

3) Extracting Facial Features: Gabor filter can be applied to find the relevant features from the faces. The different features can be extracted using a Gabor filter bank. The filters are of the orientation: 0, 22, 44, 67, 90, 112, 132, 155. The scale value starts from 4 and goes on as per the requirement.

4) Training using SVM: After the feature extraction process is completed, the support vector machine classifier is used to classify the features. The SVM classifier is an ML algorithm which is used for regression purposes and classification purposes. The SVM classifier is trained using disgust, contempt, and happy features. During testing, SVM can then classify the test data to find contempt, happy, and disgust faces.

E. Testing Phase

5) Obtaining video feed: The student is provided with two sets of questionnaires - one focusing on detecting depression and another unrelated to depression. The student's video is captured using the front camera while they answer the questionnaires. Two questionnaires are provided to avoid cases where students might conceal their expressions while answering the depression-related questionnaire. Hence, video feed recorded while the student answers the non-stress questionnaire might provide us with the student's authentic facial expression.

6) Extracting frames and Image processing: A standard video shot at 30 frames per second can generate 1800 frames in a minute. From the two videos recorded while the student answers the questionnaires, a frame can be extracted every couple of seconds for testing. These frames are then processed using the Kanade-Lucas-Tomasi Face detection algorithm to extract the face of the subject and are then saved for further processing. After face detection, the extraction of facial features can be done on the stored frames. This is done with the help of Gabor filters. This will provide us with a feature set containing facial features of the subject.

F. Classification Phase

Classification in the proposed model is done using an SVM classifier. It works by plotting the data points on to a multi-dimensional space to divide the data points into different categories. Next, the classifier identifies a separator to divide the categories in the feature space, after which transformation of the data takes place in a way that allows the separator to be modified into a hyper-plane that distinctly divides the different categories. This concludes the training phase following which the SVM can categorize testing data.

In the training part, the facial feature set obtained from the training data was fed into the SVM classifier. After extracting the facial features of the test subject, the feature set is fed into the classifier to detect the existence of different emotional features. The level of presence of the features can be used to indicate depression level as given:

- If only a small number of frames contain happy features, it can be inferred that the subject is either unhappy or is not expressing their emotional state. In case the video also contains disgust or contempt features, it may be an indication of a depressed state.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:13 UTC from IEEE Xplore. Restrictions apply.

- If a moderate number of frames contain happy features and the frames indicate the presence of disgust/contempt features, then the subject might have mild depression.
- If a large number of frames contain happy features and only a small number of frames contain disgust/contempt features, then the subject can be inferred to have no depression.

The depression level can be estimated based on the presence of features using Table 1.

Table 1. Estimating the depression level using facial features

Happy Features	Disgust and Contempt Features	Estimated Level of Depression
Low	Low	Sad/Not Expressive
Moderate	Low	Sad
High	Low	Not Depressed
Low	High	Severely Depressed
Moderate	High	Severely/Mildly Depressed
High	High	Severely Depressed (Hidden)

IV. EXPERIMENTATION AND RESULTS

G. Training Phase

7) Training Dataset Creation: The data provided as input, is an aggregation of images of 3 types of faces previously mentioned i.e., happy, disgusted, and contempt sourced from the internet and JAFFE Dataset. Initially, we consider a total of 120 pictures where the distribution amongst the above 3 types is equal. Fig. 2 contains some examples of images sourced from the internet that show disgust.

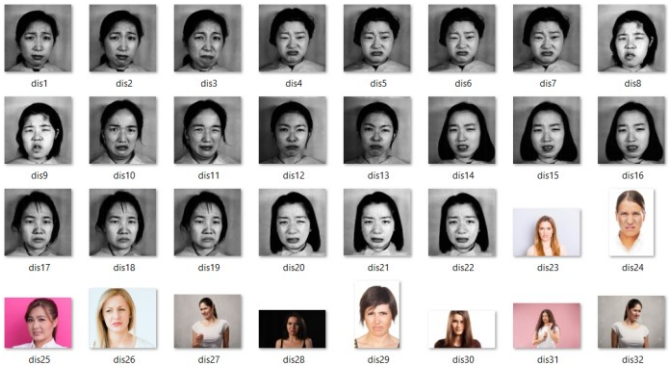


Fig. 2. Disgust Facial Expression

Similarly, Fig. 3 contains some examples of contempt faces.



Fig. 3. Contempt Facial Expression

Some examples of images showing happiness are shown in Fig. 4.



Fig. 4. Happy Facial Expression

2) Face Detection and Extracting Facial Features: Once ready with the input data, identification of faces is done using Kanade Lucas Tomasi Algorithm. A set of around 40 filters from the Gabor filter bank is prepared, each filter having different rotation and scale. The next step involves the use of this set of filters on faces present in input data.

Every filter is formed of a rotation and scale combination. In our case, 5 scales and 8 rotations are used so a total of 40 filters are available. Scales used in the system are 4.4, 4.3, 4.2, 4.1, and 4.0, similarly, rotations used are 0, 22, 45, 68, 90, 112, 135, and 155.

For all images, we apply this set of filters. This produces a set of 40 images with different scales and rotation filters applied to the same picture. This is followed by an amalgamation of the above images to one single image. This single image helps us extract features to eventually perform classification.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:13 UTC from IEEE Xplore. Restrictions apply.

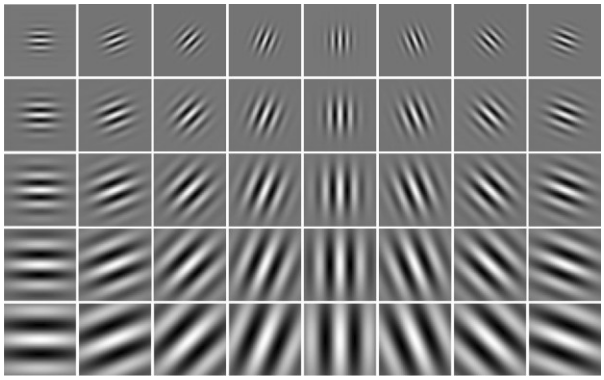


Fig. 5. Gabor Filters used in the system

3) Feature Set Formation: A separate feature vector is created for each image like the one in Fig. 6. The standard deviation and mean of the Gabor filtered image are used to create the feature set. For easy computation, the image is downsampled to 150x150 dimensions. This is followed by an evaluation of the weighted average of complete magnitude image, after this, a subtraction is done. At last, division utilizing the standard deviation of the image helps us produce a 62000x1 vector. This vector is termed as the initial image's feature vector.

	A	B	C
1	3.7015		
2	2.2451		
3	3.0887		
4	-0.6864		
5	1.0147		
6	-0.4538		
7	-0.8751		
8	0.2329		
9	1.1247		

featureVector

Fig. 6. Feature Vector created for a single image

The feature vector for the entire 40 happy images is constructed and combined to obtain the feature set as shown in Fig. 7(a) of dimension 62000x40. To reduce dimensionality, PCA is used to make a 40x40 matrix. Happy images are assigned the class value '0' as in Fig. 7(b).

	A	B	C
1	3.7015	4.0135	0.9057
2	2.2451	2.0521	4.6005
3	3.0887	3.6322	5.2851
4	-0.6864	-0.1274	2.6272
5	1.0147	2.8865	-0.3017
6	-0.4538	1.0725	0.2056
7	-0.8751	0.0918	-0.7632
8	0.2329	2.8081	-0.3632
9	1.1247	1.2113	4.5142

featureSetHappy

(a) Before applying PCA

	AM	AN	AO
1	-0.3713	0.8859	0
2	-0.0499	-0.6985	0
3	0.6182	0.8237	0
4	0.3555	-0.5036	0
5	0.9137	-0.7988	0
6	0.2397	0.9864	0
7	0.9778	0.5255	0
8	-0.1905	0.6315	0
9	0.9012	0.1441	0

featureSet0

(b) After applying PCA

Fig. 7. The feature set for all happy images

Similarly, the feature set for the entire 120 images is created. Values '1' and '2' have been assigned to disgust and contempt facial emotions respectively.

	AM	AN	AO
1	-0.3713	0.8859	0
2	-0.0499	-0.6985	0
3	0.6182	0.8237	0
4	0.3555	-0.5036	0
⋮			
39	-0.2632	0.6416	0
40	-0.9716	0.1143	0
41	0.9712	-0.1272	1
42	0.6982	0.5797	1
⋮			
79	0.2282	0.8212	1
80	0.8701	0.3786	1
81	-0.3839	-0.4242	2
82	0.8989	-0.4837	2
⋮			
117	0.4485	-0.2465	2
118	-0.6519	0.4172	2
119	-0.5493	-0.9323	2
120	-0.2275	0.6469	2

featureSetAll

Fig. 8. The feature set with class values happy-0, contempt-1 and disgust-2

H. Testing Phase

The test subject is made to answer a depression-related questionnaire and a video is captured. Image frames are extracted after every second for a 30fps video. These are saved at one location to form the testing database.

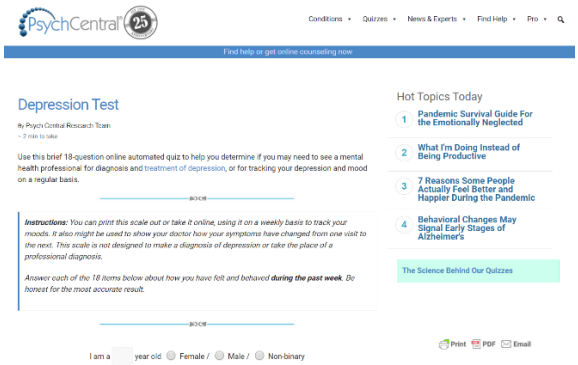
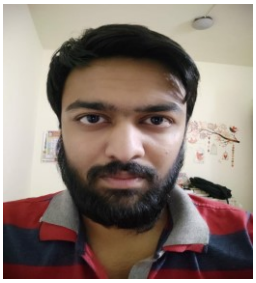
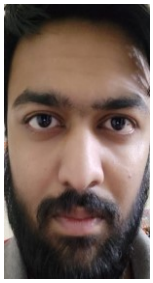


Fig. 9. Depression related questionnaire

Then extraction of the face using a Kanade-Lucas-Tomasi Face detection algorithm takes place. In the current scenario, the KLT algorithm is appropriate. The first 60 frames are taken for testing. Gabor features are extracted to create the testing feature set.



(a) Frame extracted



(b) Face detected

Fig. 10. Sample frame from captured video and extracted

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:13 UTC from IEEE Xplore. Restrictions apply.

The SVM classifier will be used to test the test feature set. In the case of high depression, the happy features will be less while the disgust and contempt feature will be more. If high contempt and disgust features are present and the moderate number of happy features are present, then the student can be classified as mildly depressed. If there are no signs of disgust or contempt features it is hard to predict whether the person has depression or not.

	A	B	C
1	0.7449	0.1753	0.4308
2	-0.7967	-0.7559	0.5701
3	0.2591	0.4324	-0.9398
4	-0.0691	0.4179	-0.3775
5	0.2015	0.5867	0.1183

testingFeatureSet

Fig. 11. Testing Feature set of dimensions 60x60

The above SVM classifier was also applied to an already known database - BlackDog. The database contains already classified videos of people with depression. It was found that the above classifier could predict the result with an accuracy of 62% (average recall).

V. CONCLUSION AND FUTURE WORK

A study has been executed on various facial expressions associated with depression and the methods of finding them. JAFFE database was used to create a training dataset of happy images. Kanade-Lucas-Tomasi algorithm was used for detecting faces. Gabor filter was used to extract various features from the images. The feature vector was used to create a feature set for training.

For the testing phase, the face of the student is recorded while he/she answers the questionnaires. The video is converted to frames and stored as a testing dataset. Classification is performed using an SVM classifier which predicts whether the subject is depressed or not. Depending upon the result, the student can be recommended to attend counseling.

Future work includes a collection of the properly curated dataset of videos of depressed students and increasing the accuracy of the model. Further improvements can be made by including various other facial features like the head pose and eye movement. Using vocal prosody analysis in conjunction with video analysis can improve the prediction accuracy.

REFERENCES

[1] Karthika, R., and Latha Parameswaran. "Study of Gabor wavelet for face recognition invariant to pose and orientation." In Proceedings of the International Conference on Soft Computing Systems, Springer, New Delhi, 2016.

[2] Cohn, Jeffrey F., Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. "Detecting depression from facial actions and vocal prosody." In Affective Computing and Intelligent Interaction and Workshops. ACII 2009. 3rd International Conference on, pp. 1-7. IEEE, 2009.

[3] Katikalapudi, Raghavendra, Sriram Chellappan, Frances Montgomery, Donald Wunsch, and Karl Lutzen. "Associating internet usage with depressive behavior among college students." IEEE Technology and Society Magazine 31, no. 4: 73-80. 2012

[4] Harati, Sahar, Andrea Crowell, Helen Mayberg, Jun Kong, and Shamim Nemati. "Discriminating clinical phases of recovery from major depressive disorder using the dynamics of facial expression." In Engineering in Medicine and Biology Society (EMBC), 38th Annual International Conference of the, pp. 2254-2257. IEEE, 2016.

[5] Tasnim, Mashrura, RifatShahriyar, NowshinNahar, and Hossain Mahmud. "Intelligent depression detection and support system: Statistical analysis, psychological review and design implication." In e-Health Networking, Applications and Services (Healthcom), 18th International Conference on, pp. 1-6. IEEE, 2016.

[6] Haghighat, Mohammad, Saman Zonouz, and Mohamed AbdelMottaleb. "CloudID: Trustworthy cloud-based and cross-enterprise biometric identification." Expert Systems with Applications 42, Vol no. 21, pp. 7905-7916, 2015.

[7] Sahla, K. S., and T. Senthil Kumar. "Classroom Teaching Assessment Based on Student Emotions." In the International Symposium on Intelligent Systems Technologies and Applications, pp. 475-486. Springer International Publishing, 2016.

[8] Meng, Hongying, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. "Depression recognition based on dynamic facial and vocal expression features using partial least square regression." In Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, pp. 21-30. ACM, 2013.

[9] Alghowinem, Sharifa, Roland Goecke, Jeffrey F. Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. "Cross-cultural detection of depression from nonverbal behaviour." In Automatic Face and Gesture Recognition (FG), 11th IEEE International Conference and Workshops on, vol. 1, pp. 1-8. IEEE, 2015.

[10] Pampouchidou, A., O. Simantiraki, C-M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias et al. "Facial geometry and speech analysis for depression detection." In Engineering in Medicine and Biology Society (EMBC), 39th Annual International Conference of the IEEE, pp. 1433-1436. IEEE, 2017.

[11] Pampouchidou, Anastasia, Kostas Marias, ManolisTsiknakis, P. Simos, Fan Yang, and FabriceMeriaudeau. "Designing a framework for assisting depression severity assessment from facial image analysis." In Signal and Image Processing Applications (ICSIPA), International Conference on, pp. 578-583. IEEE, 2015.

[12] Girard, Jeffrey M., Jeffrey F. Cohn, Mohammad H. Mahoor, Seyedmohammad Mavadati, and Dean P. Rosenwald. "Social risk and depression: Evidence from manual and automatic facial expression analysis." In Automatic Face and Gesture Recognition (FG), 10th IEEE International Conference and Workshops on, pp. 1-8. IEEE, 2013.

[13] Maddage, Namunu C., RajindaSenaratne, Lu-Shih Alex Low, Margaret Lech, and Nicholas Allen. "Video-based detection of the clinical depression in adolescents." In Engineering in Medicine and Biology Society, (EMBC). Annual International Conference of the IEEE, pp. 3723-3726. IEEE, 2009.

[14] Athira, S., R. Manjusha, and Latha Parameswaran. "Scene Understanding in Images." In the International Symposium on Intelligent Systems Technologies and Applications, pp. 261-271. Springer International Publishing, 2016.

[15] Ashraful Islam, Anin Naeem and Nazmul Hasan. "Comparision between Viola-Jones and KLT Algorithms and error correction of Viola-Jones Algorithm". In the International Journal of Computer Engineering and Applications, Vol 11, Issue V, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:35:13 UTC from IEEE Xplore. Restrictions apply.

This is an open access article published under an ACS AuthorChoice License, which permits copying and redistribution of the article or any adaptations for non-commercial purposes.

ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Au-Nanoplasmonics-Mediated Surface Plasmon-Enhanced GaN Nanostructured UV Photodetectors

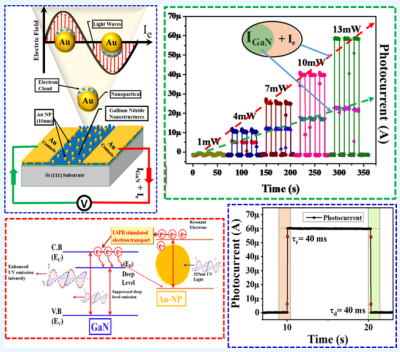
Lalit Goswami, Neha Aggarwal, Shibin Krishna, Manjri Singh, Pargam Vashishtha, Surinder Pal Singh, Sudhir Husale, Rajeshwari Pandey, and Govind Gupta*

Cite This: https://dx.doi.org/10.1021/acsomega.0c01239

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The nanoplasmonic impact of chemically synthesized Au nanoparticles (Au NPs) on the performance of GaN nanostructure-based ultraviolet (UV) photodetectors is analyzed. The devices with uniformly distributed Au NPs on GaN nanostructures (nanoislands and nanoflowers) prominently respond toward UV illumination (325 nm) in both self-powered as well as photoconductive modes of operation and have shown fast and stable time-correlated response with significant enhancement in the performance parameters. A comprehensive analysis of the device design, laser power, and bias-dependent responsivity and response time is presented. The fabricated Au NP/GaN nanoflower-based device yields the highest photoresponsivity of ∼ 380 mA/W, detectivity of ∼ 10^{10} jones, reduced noise equivalent power of ∼ 5.5 × 10^{-13} W $Hz^{-1/2}$, quantum efficiency of ∼ 145%, and fast response/recovery time of ∼40 ms. The report illustrates the mechanism where light interacts with the chemically synthesized nanoparticles guided by the surface plasmon to effectively enhance the device performance. It is observed that the Au NP-stimulated local surface plasmon resonance effect and reduced channel resistance contribute to the augmented performance of the devices. Further, the decoration of low-dimensional Au NPs on GaN nanostructures acts as a detection enhancer with a fast recovery time and paves the way toward the realization of energy-efficient optoelectronic device applications.



INTRODUCTION

Ultraviolet photodetector (UV PD) fabrication and development has been an area of intense research in the last few decades, and the ongoing research for finding new ways to enhance the UV detection capability of the detectors is very encouraging. The importance of UV PDs is universally recognized because of their diverse range of applications, such as in antimissile technology by plume detection, flame detection, calibration and monitoring of UV radiations, chemical as well as biological analyses, secure optical communications, environmental monitoring, and astronomical research.[1] For this, GaN semiconductors have demonstrated to be a potential backbone of durable device applications because of the numerous advantages offered by its significant inherent properties such as the wide and direct band gap, high thermal coefficient, high carrier mobility, higher resistance toward electric breakdown field, and high chemical stability.[2] The advancement in photodetection can be realized by using nanostructures (NSs), which offer additional advantages in detection applications by offering a large surface-to-volume ratio that can significantly increase the UV photon absorption rate. These advancements in UV PDs can lead to good photoconductivity and enhanced electrical conductivity.[3−5]

Recently, much attention has been paid for enhancing the performance of optoelectronic devices such as solar cells, light-emitting diodes (LEDs), and UV PDs via nanoplasmonic behavior of metal nanoparticles (NPs).[6−8] Some previous reports articulate that the use of nanoplasmonic behavior of metal NPs such as silver (Ag) NPs,[9] aluminum (Al) NPs,[6,10] platinum (Pt) NPs,[7] and gold (Au) NPs[8] has been realized as a remedy for the carrier enhancement seeker, GaN-based UV PDs. Thereby, metal NP-stimulated hot carriers are guided by local surface plasmon resonance (LSPR) for enhancing optical detector performance. Thus, LSPR-induced enhanced UV emission has been reported in Au NP-decorated ZnO NSs.[10,11] The Au NPs nowadays have splendidly established their identity in LSPR because of numerous inherent advanced physical properties such as nonlinear optics, magneto-plasmonics, optical trapping and optical activity,[12] high chemical and physical stabilities, ease of surface functionalization with biomolecules, and plasmon-oriented multitude of optical properties.[13,14] Besides, Au NPs are also known for access to easily polarizable hot carriers (hot resonant

Received: March 20, 2020
Accepted: June 2, 2020

Downloaded via DELHI TECHNOLOGICAL UNIV on June 19, 2020 at 06:06:15 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

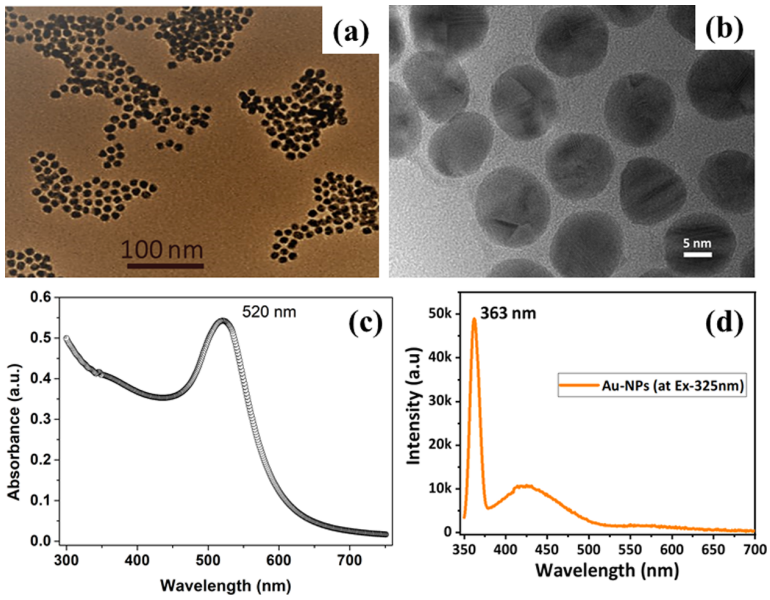


Figure 1. (a,b) HRTEM micrographs of NPs with lower and higher magnifications. (c) UV−vis absorbance spectra of Au NPs. (d) RT-PL spectra of Au NPs.

electrons). These polarizable conduction electrons interact with the electromagnetic field of the corresponding incident electromagnetic light wave and generate nonlinear optical phenomena.[15] Therefore, this is the first report where chemically synthesized monodispersed (~10 nm) Au NPs' nanoplasmonic impact has been realized on GaN-NS-based metal−semiconductor−metal (MSM) UV PDs as an UV detection enhancer. The study elaborates the nanoplasmonic effect, which leads to trapping of incident photons by the process of localized scattering and absorption as a consequence of incident photons and metal NP interaction, which increases the light and the material interaction time. This scattering stimulated an enhanced interaction to allow high photon absorptions, which generate more electron−hole pairs.[16]

EXPERIMENTAL SECTION

The Au NPs were synthesized in the presence of trisodium citrate via the Turkevich method[17] using gold(III) chloride hydrate ($HAuCl_4 \cdot 3H_2O$) precursor solution with trisodium citrate ($C_6H_5Na_3O_7 \cdot 2H_2O$) [for details, see Supporting Information (Figure S1)]. The as-synthesized Au NPs were characterized by ultraviolet−visible (UV−vis) spectroscopy (Agilent, Cary 5000i) and high-resolution transmission electron microscopy (HRTEM) (FEI, Tecnai F30 G2 S-TWIN). The HRTEM images of bare Au NPs elucidate spherelike shape with an average size of ~ 10 nm, and the UV−vis absorption spectrum shows the LSPR peak position at ~520 nm, which also confirms the size of Au NPs to be ~10 nm[18] (Figure 1). The UV−vis absorption data have also been used to calculate the number density of Au NPs to be 3.5×10^{12} using the calculation technique mentioned by Haiss et al.[19] The broadening of the UV−vis absorption spectrum is found directly in correlation with the concentration of Au NPs;[18] lower broadening and higher absorption postulates higher concentration of Au NPs.

Thus, it contributes more LSPR-generated hot carriers into the devices as attributed by room-temperature (RT) photoluminescence (PL) measurements of Au NPs (Figure 1d), which were carried out using a FLS980 D2D2 (Edinburg) system with a 325 nm excitation source (Hd-Cd Kimmon laser). The spectra exhibit a prominent emission peak at 363 nm and a secondary peak at ~430 nm. The secondary blue emission peak at 430 nm corresponding to the Au 6S-to-6P intraband transition is also attributed to the presence of isolated spherelike Au NPs of size < 20 nm.[20] These chemically synthesized Au NPs were spin-coated uniformly on GaN nanostructured surfaces, that is, on GaN nanoislands (GaN-NIs) and GaN nanoflowers (GaN-NFs) at a constant speed of 3000 rpm. Both the GaN nanostructured surfaces (GaN-NIs and GaN-NFs) were grown by using plasma-assisted molecular beam epitaxy (Riber Compact 21). The high-density GaN-NIs have been grown at the substrate temperature of 770 °C in the presence of N_2 plasma (1.5 sccm, 400 W) and Ga beam equivalent (BEP) of 6.5×10^{-7} Torr for 120 min.[21] These high density and vertically aligned GaN-NIs were grown on the Si(111) substrate with a distribution density of ~ 2×10^{10} cm^{-2}, average width of 50 ± 5 nm, and aspect ratio of 1.92 (width-to-height). For the growth of GaN-NFs, the buffer layer of AlN (thinness 30 nm) was deposited at a substrate temperature of 825 °C under stoichiometric conditions on Si(111). Further, the growth of the GaN film (80 nm) followed by GaN-NFs was performed at 730 °C at rf plasma power of 500 W with Ga BEP of 1×10^{-6} Torr.[22] Thus, the uniformly distributed GaN-NFs with the bottom tapered portion of ~200 nm to top higher opening part of ~300 ± 50 nm was performed on Si(111) with the AlN buffer layer. The high rms roughness of these nanostructured surfaces (5.8 nm for GaN-NIs and 9.8 nm for GaN-NFs) has been utilized to settle low-dimensional Au NPs. A schematic diagram elaborating the complete sample preparation process is demonstrated in Figure S1. The morphology and elemental mapping of uniformly distributed Au NPs decorated on GaN-NS surfaces were carried out using field-emission scanning electron microscopy (FESEM) and energy dispersive X-ray spectroscopy (EDS) techniques (Zeiss, Auriga) as shown in Figures S2 and S3. For the fabrication of UV PDs, Au-metal

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

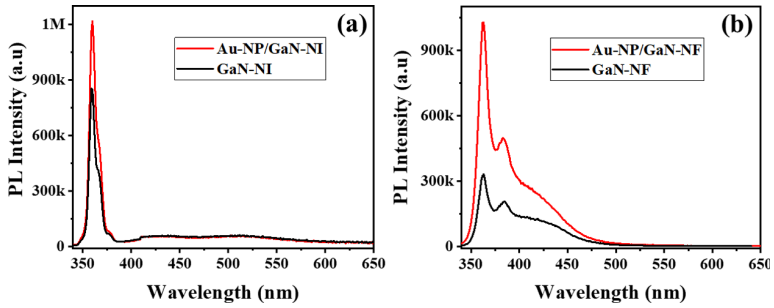


Figure 2. (a) RT-PL spectra of (a) bare GaN-NIs and Au NPs/GaN-NIs and (b) bare GaN-NFs and Au NPs/GaN-NFs.

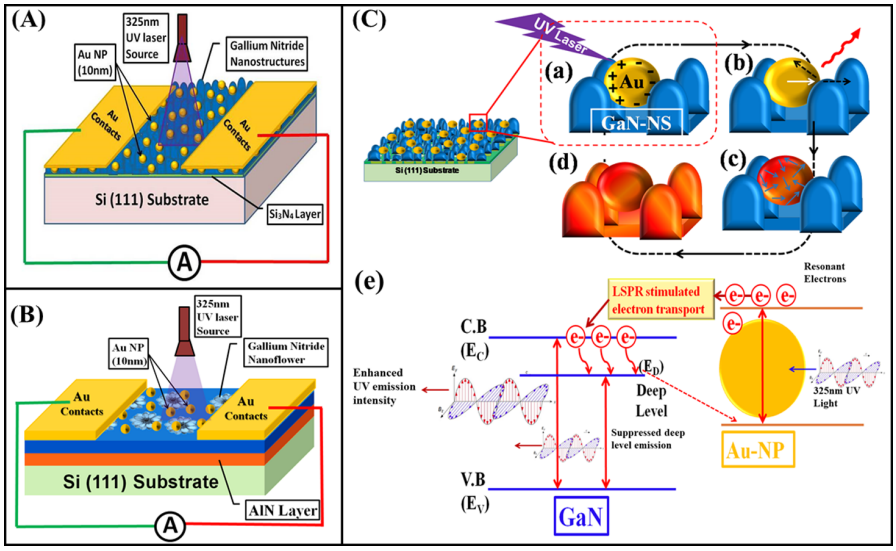


Figure 3. Schematic illustration of (A,B) both NP-decorated UV PDs and (C) their nanoplasmonic effect. (a) Local surface plasmons are excited by the electromagnetic wave of UV light propagating in free space. (b) Process of Landau damping. (c) Scattering and relaxation of LSPR-generated hot carriers. (d) Energy dissipation. (e) Energy band diagram of GaN-NSs and Au NP system showing local surface plasmon coupling (325 nm laser source excites GaN-NSs as well as Au NPs). The schematic demonstrates that GaN-generated carriers are getting merged with Au NPs' local surface plasmon-originated resonant hot electrons in the same band regime of GaN, i.e., ∼364 nm.

electrodes (200 nm thick) were deposited as a MSM structure by the thermal evaporation technique on bare and Au NP-coated GaN-NS samples, and the active area of the devices was fixed as 3×10^{-3} cm^2. The impact of Au NPs on the opto-electrical transport is analyzed by RT current−voltage and photoresponse measurements, which were carried out using a Cascade Microtech instrument (EPS150TRIAX with shield enclosure EPS-ACC-SE750).

RESULTS AND DISCUSSION

To analyze the enhancement in photon-generated hot carriers from Au NP-decorated GaN-NS devices, the RT-PL measurements were carried out. RT-PL spectra demonstrate UV emission at 364 nm and an additional peak of violet-blue emission for all samples (bare GaN-NIs, Au NPs/GaN-NIs, bare GaN-NFs, and Au NPs/GaN-NFs) when excited by the 325 nm UV laser as shown in Figure 2. The UV emission at 364 nm resembles the near-band edge emission (GaN band gap of 3.406 eV) and recognized the radiative recombination of excitons, although the presence of violet-blue emission (380−440 nm) is due to trapping states related to the optical states existing in between the band gap regime.[23,24] The spectra validate significantly enhanced UV emission for Au NP-decorated GaN-NIs (×1.4) and GaN-NFs (×3) as compared to their bare counterparts, as shown in Figure 2a,b, respectively. The coupling of Au NPs' localized surface plasmons at the interface with GaN-NSs postulates resonance interaction between the electromagnetic field of the incident UV light and the electron charge near the surface of Au NPs (emission at 363 nm as shown in Figure 1d) and contributes more electrons in-flow with the existing GaN channel.

The effect of nanoplasmonics on the opto-electrical behavior of GaN-NS-based devices can be realized because of the Au NP-stimulated LSPR effect, which can significantly enhance photocurrent in the fabricated device because of enhanced light absorption in the near-surface region.[25,26] As the synthesized metallic NPs having size dimensionality to the metal skin depth, thereby, the electric field of incident electromagnetic radiation of light having ease of penetration of Au NPs' surface. Thus, this surface penetration by UV radiation allow the polarization of Au NPs' conduction electrons.[27,28] The LSPR with uniformly distributed plasmon oscillation over the entire NP volume where the size of NPs is much smaller (∼10 nm) than the incident photon wavelength (325 nm) has nonpropagating excitations. The generated plasmons over the NPs have two possibilities, that is, either

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

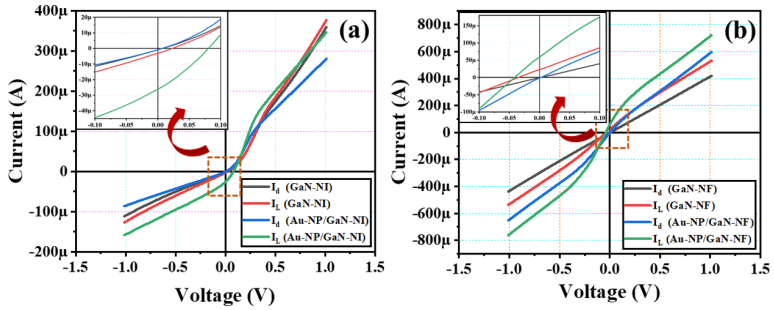


Figure 4. $I-V$ characteristics of (a) bare and Au NPs/GaN-NI- and (b) bare and Au NPs/GaN-NF-based devices under dark and UV illumination.

Table 1. Ideality Factors and SBH Variations at Two Au/GaN Interfaces on Each of Four Fabricated UV PDs

	GaN-NI (A)		Au NP/GaN-NI (B)		GaN-NF (C)		Au NP/GaN-NF (D)	
Au/GaN	1	2	1	2	1	2	1	2
n	n_{A1}	n_{A2}	n_{B1}	n_{B2}	n_{C1}	n_{C2}	n_{D1}	n_{D2}
	1.8	1.92	1.2	1.5	1.7	1.53	1.1	1.2
φ_b	φ_{bA1}	φ_{bA2}	φ_{bB1}	φ_{bB2}	φ_{bC1}	φ_{bC2}	φ_{bD1}	φ_{bD2}
	0.75	0.85	0.703	0.8	0.97	0.88	0.7	0.73

they got absorbed or scattered. The probability of absorption or scattering can be determined by σ_{Abs} and σ_{Sca}, respectively, which depend on the size, shape, and structure of the NPs.[29,30] The Au NPs of size < 25 nm have pure plasmon absorber properties due to the negligible scattering coefficient, that is, $\sigma_{Abs} \gg\gg \sigma_{Sca}$.[31] This absorbed plasmon-originated resonance or electron clouds are damped (through Landau damping) by these low-dimensional NPs in two ways: nonradiatively and radiatively through hot electron−hole pairs generation and re-emission of photons, respectively.[32] The distribution of these generated carriers depends upon factors such as particle size, plasmon energy, plasmon mode symmetry, state density of the material, and electronic structure.[33] The plasmon decay-generated carriers release their energy (via carrier relaxation) among many lower-energy electrons through their interscattering process, for example, Auger transition.[34] Figure 3A−C displays a complete schematic diagram (including the energy band diagram) of NP-decorated GaN-NS-based UV PDs having nanoplasmonics-generated LSPR excitation of Au NPs and their damping around the medium (GaN-NSs), carrier generation, and their scattering, relaxation, and distribution. These Au NPs' LSPR-injected extra carriers (resonant hot electrons) in addition to existing UV photon-generated carriers of GaN-NSs enhanced the cumulative photocurrent.[35]

To analyze the Au NPs nanoplasmonics-stimulated enhancement in the opto-electrical transport, self-powered (without bias) as well as photoconductive modes of operation were performed where the $I-V$ measurement for bare GaN-NIs, GaN-NFs, Au NPs/GaNIs, and Au NPs/GaN-NFs based MSM UV PD devices was performed under dark and UV illumination (325 nm) at RT, as shown in Figure 4. The data revealed that Au NPs/GaN-NS-based UV PDs yield enhanced photocurrent (I_{ph}) characteristics as compared to bare GaN-NS UV PD devices in the self-powered mode (0 V bias) as well as under variable bias conditions (from −1 to +1 V). At zero volt bias, the Au NPs/GaN-NIs UV PDs display very low dark current (I_d) ~ 22 nA and very high light current (I_L) ~ 25 μA; however, in the case of bare GaN-NIs UV PDs, these values are found to be ~12 nA (I_d) and ~2.1 μA (I_L), which elucidates ~12-fold enhancement in photocurrent ($I_{ph} = I_L - I_d$), as shown in the inset of Figure 4a. The analogous impact was observed in the case of the Au NP/GaN-NF-based device, where I_d and I_L ~ are found to be 97.4 nA and 60.3 μA, respectively, leading to I_{ph} ~ 60 μA, which is significantly higher than ~24 μA in the case of bare GaN-NF (I_d: 47.7 nA, I_L: 24.3 μA)-based UV PDs, as shown in the inset of Figure 4b. The self-powered behavior originated from the Au/GaN Schottky junction, and asymmetric Au/GaN interfaces triggered the inbuilt potential gradient as explained elsewhere.[36] The analysis reveals that, due to the presence of Au NPs, the fabricated Au NP/GaN-NI and Au NP/GaN-NF devices display enhanced self-powered characteristics as compared to their respective bare GaN-NS-based devices. Further, a significant variation in the photocurrent has also been observed for these devices (bare and Au NPs/GaN-NSs) under applied bias (−1 to +1 V). The Schottky barrier height (SBH) (φ_b) and the ideality factor (n) of the Au NPs/GaN-NSs as compared to their respective bare counterparts have been evaluated for forward as well as reverse bias by using standard thermionic emission theory,[37,38] as shown in Table 1.

The performance parameters such as detectivity (D), responsivity (R), noise equivalent power (NEP), and quantum efficiency (η) of the fabricated UV PDs were calculated by using the following relations.[39]

$$R = \frac{I_{ph}}{P_{inc} \times A_0} \tag{1}$$

$$\text{NEP} = \frac{\sqrt{2eI_d(I_d)}}{R} \tag{2}$$

$$D = \frac{R \times \sqrt{A_0}}{\text{NEP}} \tag{3}$$

$$\eta = \frac{R \times h \times c}{e \times \lambda} \tag{4}$$

where P_{inc} is the incident optical power illuminated on the active area of the device to generate corresponding carriers out of absorbed incident UV photons, A_0 is the active area of the fabricated device, h is the Planck's constant, c is the speed of

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

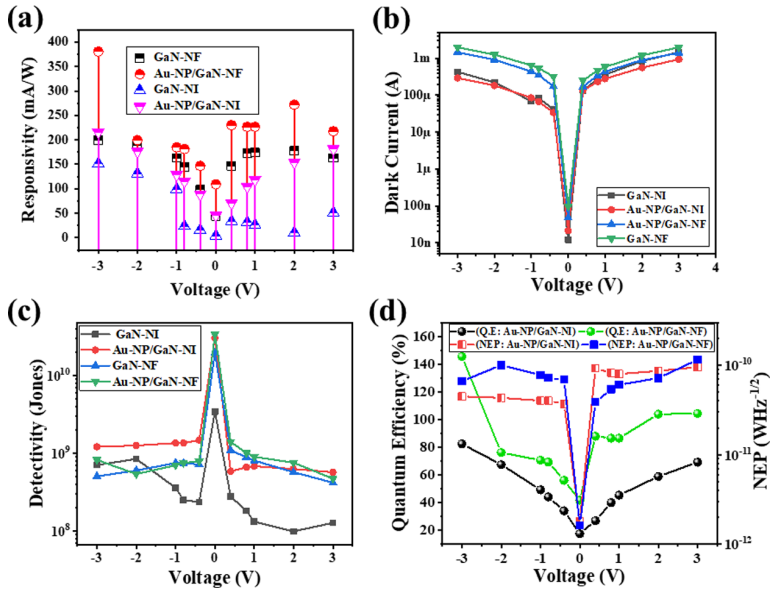


Figure 5. Detector performance parameters variation with respect to (a−d) applied bias voltages.

Table 2. Comparative Evaluation of Various Performance Parameters of Bare and Au NP-Incorporated GaN-NS-Based MSM UV PDs

	R (mA/W)		D (10^9 Jones)		NEP (10^{-11} W Hz$^{-1/2}$)		Q.E (%)	
fabricated detectors	0 V	−3 V	0 V	−3 V	0 V	−3 V	0 V	−3 V
GaN-NIs	3.89	151.28	3.43	0.708	1.59	7.73	1.48	58.73
GaN-NFs	43.95	199.36	19.5	0.507	0.281	10.08	16.8	76.2
Au NPs/GaN-NIs	45.99	215.67	30.5	1.22	0.179	4.49	17.6	82.4
Au NPs/GaN-NFs	109.1	380.60	33.8	0.826	0.162	6.63	41.7	145.5

light, e is the elementary charge, and λ is the incident photon wavelength of the illuminating source. The responsivity, detectivity, and NEP of all fabricated devices are calculated for zero and applied biases between −3 to +3 V. It was observed that the responsivity of the fabricated detectors in the self-powered mode is enhanced by ∼12× and ∼2.5× for Au NP-decorated GaN-NI and GaN-NF detectors as compared to their bare devices, respectively. The responsivity of these devices was analyzed at an applied bias of −3 to +3 V, where all fabricated detectors show a gradual increment in the value of "R" in either direction of bias polarity (Figure 5a). A significant enhancement in "R" from 151.28 mA/W (bare GaN-NIs) to 215.67 mA/W (Au NPs/GaN-NIs) and 199.36 mA/W (bare GaN-NFs) to 380.6 mA/W (Au NPs/GaN-NFs) at −3 V is determined (Table 2).

Furthermore, detectivity of the fabricated devices was also analyzed at self-powered as well as at variable bias conditions (±3 V). Au NPs' highly conductive behavior accelerated the reduction in the dark current (Figure 5b) and enhanced the responsivity, significantly reciprocating increased D values (Figure 5c and Table 2), that is, from 3.43 × 10^9 Jones (bare GaN-NIs) to 3.05 × 10^{10} Jones (Au NPs/GaN-NIs) and 1.95 × 10^{10} Jones (bare GaN-NFs) to 3.38 × 10^{10} Jones (Au NPs/GaN-NFs) in the self-powered mode. At higher applied bias, the increase in D values is limited due to the high dark current. The bias-dependent analysis also revealed the criticality of dark current, which can directly affect detectivity as well as the NEP of the detectors, as shown in Figure 5(c,d) and Table 2. The observation elucidates that unlike higher dark current values at higher bias voltage (−3 V), lower dark current at the self-powered mode allows the fabricated detectors to achieve higher D values and lower NEP, that is, 1.79 × 10^{-12} W Hz$^{-1/2}$ (Au NPs/GaN-NIs) and 1.62 × 10^{-12} W Hz$^{-1/2}$ (Au NPs/GaN-NFs) as compared to their bare counterparts (Table 2 and Figure 5d). The study divulges that the GaN-NS-based UV PDs with Au NPs demonstrates comparatively superior performance. The capability of these detectors have also been investigated by applied bias voltages and realized as an increment in the quantum efficiency, where they achieved their maximum values comparatively, that is, 82.4 and 145.5% by Au NPs/GaN-NIs and Au NPs/GaN-NFs, respectively, at −3 V, as shown in Figure 5d and Table 2.

The performance of all fabricated UV PDs with varying optical power is tested, which shows promising linear inclination in photocurrent values with increasing incident optical power "P_{opt}" (from 1 to 13 mW), as shown in Figure 6a,b for GaN-NI- and GaN-NF-based UV PDs, respectively. The results demonstrate increasing I_{ph} values from 1.02 μA (1 mW) to 2.87 μA (13 mW) (∼2.8× increment) for bare-GaN-NIs UV PDs and −1.05 μA (1 mW) to 22.6 μA (13 mW) revealing ∼22× increment in I_{ph} for bare-GaN-NF-based detectors. The Au NP-induced LSPR significantly enhances I_{ph} values with increasing power such as 2.46−26.2 μA (13 mW) and shows ∼11-fold enhancement in Au NPs/GaN-NIs, whereas I_{ph} displays ∼85-fold enhancement (from −0.68 μA at 1 mW to 58.5 μA at 13 mW) for Au NPs/GaN-NFs. Thus, Au NP hot carrier-induced enhancement is witnessed in the performance of the fabricated devices. Moreover, the NEP of

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

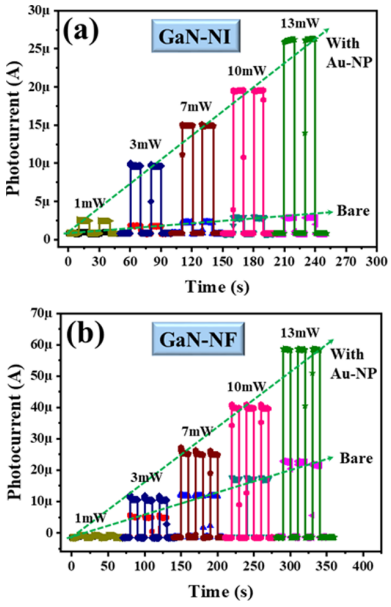


Figure 6. Power-dependent transient response of (a) bare and Au NP/GaN-NI and (b) bare and Au NP/GaN-NF detectors.

Au NP/GaN-NF-based detectors has also been investigated by varying the applied optical power (1 to 13 mW), where NEP values start decreasing from 1.072×10^{-10} W $Hz^{-1/2}$ (at 13 mW, −3 V) and reached its minimum value of 5.55×10^{-13} W $Hz^{-1/2}$ (at 1 mW, 0 V) because of inverse effect of increasing the responsivity values 21.61 mA/W (1 mW, 0 V) to 109.1 mA/W (13 mW, 0 V), as shown in Figure S4.

Next, the switching speed of the detector which elucidates the quality of a detector to detect illuminating photons has been evaluated. The switching speed of the fabricated UV PDs in terms of time-correlated response and recovery was evaluated from its transient response by UV light illumination turning on and turning off. Figure 7 shows the time-correlated photoresponse under self-powered mode (at 13 mW optical power) of all fabricated UV PDs, where the rise (response) time and decay (recovery) time curves are fitted with respect to the experimental data by the following equations[3]

$$I = I_0(1 - e^{-t/t_r}) \quad (5)$$

$$I = I_0(e^{-t/t_d}) \quad (6)$$

where I_0 is the maximum saturation photocurrent value at a particular time t and t_r and t_d stand for the rise and decay times, respectively. The rise time from 10 to 90% of the final value, that is, "t_r" under UV illumination and decay time from 90 to 10% of the final value "t_d" after switching off UV illumination have been measured from time correlated transient response of the fabricated devices.

The observation reveals that the fabricated detectors with Au NPs are benefited by LSPR-generated high-density charge carriers, wherein, the highly conducting behavior of Au NPs plays a significant role to reduce the defect-originated obstructions in the flow of charge carriers. This reduction in offered resistance to the photocurrent allows quick retardation of the photocurrent after switching off the UV light. Figure 7 shows the effect of quick retardation as the quick recovery time, that is, "t_d" of GaN-NI- (~63.6 ms) and GaN-NF-based detectors (~43.7 ms) is reduced to 40 ms of Au NP/GaN-NI- and Au NP/GaN-NF-based UV PDs. The rise and decay time measurement of the fabricated detectors is restricted to <40 ms because of system limitation.

Table 3 shows, upon comparing with previous reports on nanoplasmonics, that the fabricated GaN-NS UV PDs with Au NPs, herein, exhibit fast rise and decay times. Such a high switching speed is contributed by an additional potential force because of variation in SBHs, reduced channel resistance due to additional effect of Au NPs' higher conductivity, and a better crystalline quality, which were elaborately discussed

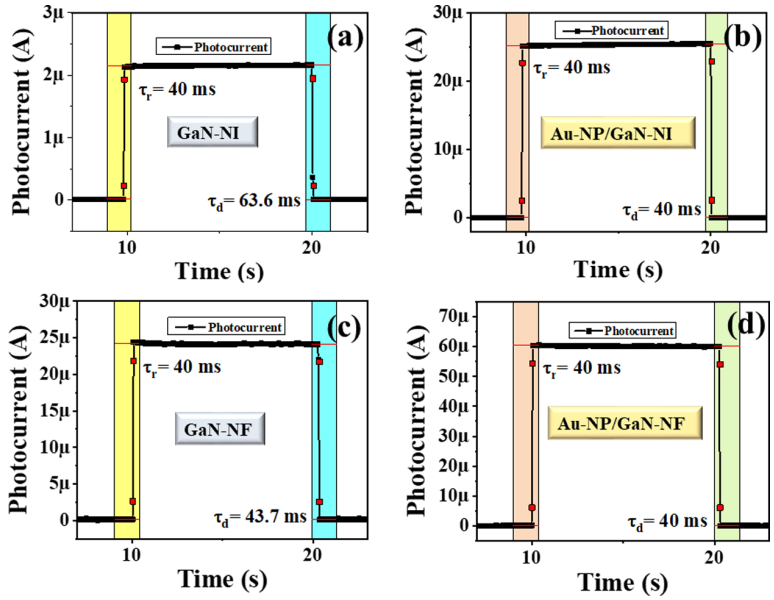


Figure 7. Time-correlated photoresponse rise time and decay time fitted curves. (a) Bare GaN-NIs, (b) Au NPs/GaN-NIs, (c) bare GaN-NFs, and (d) Au NPs/GaN-NFs MSM UV PDs under self-powered mode at 13 mW.

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

Table 3. Switching Speed Comparison with Various Devices Incorporating Hot Carrier Generated by Nanoplasmonic Effects

devices	response time (switching speed) rise time (ms)	decay time (ms)	references
Au NPs/GaN-NF/Si(111)	40	40	this work
Au NPs/GaN-NS/Si(111)	40	40	this work
Au NP/Ga-polar GaN	2900	6200	40
Au NP/ZnO/PET	10,300	14,200	41
Ag-NP/ZnO-NW	80	3270	42
Pt-NP/GaN-NS/Al_2O_3	1100	650	7
ZnO-NP/quartz	600	8000	43
ZnO-NP/quartz	22,000	11,000	44
ZnO-NP/glass	2000	12,000	45
TiO_2-NP/SiO_x	1000	900	46
TiO_2-nanofibers/SiO_2/Si	1500	7800	47

elsewhere,[21] and revealed a negligible defect band in PL and low stress via micro-Raman spectroscopy.

CONCLUSIONS

In conclusion, the present analysis postulates a significant impact of chemically synthesized, monodispersed Au NPs (∼10 nm) on the performance of GaN-NS-based UV PDs, wherein fabricated GaN detectors with Au NPs prominently respond toward UV in both self-powered as well as the photoconductive modes of operation and display fast and stable time-correlated photoresponse with significant enhancement in the performance parameters. The enhanced performance of the fabricated devices is due to nanoplasmonics-induced LSPR-stimulated hot resonant carriers. Moreover, the electronic transport properties affirmed the influence of Au NPs in lowering the ideality factor and SBH at the Au/GaN interfaces because of which the channel resistance of the fabricated detector is reduced. These distinctive barrier height-stimulated self-powered UV PDs have shown fast switching speed (rapid rise and decay time of 40 ms), enhanced responsivity, detectivity, higher quantum efficiency, and very low NEP. The best performance of the fabricated device with Au NPs/GaN-NFs yields photo responsivity of ∼ 380 mA/W, detectivity of ∼ 3.38×10^{10} jones, reduced NEP of ∼ 5.5×10^{-13} W $Hz^{-1/2}$, fast response/recovery time of ∼40 ms, and higher quantum efficiency of 145.5%. The decoration of low-dimensional Au NPs' nanoplasmonic behavior on GaN-NSs acts as a detection enhancer with a fast recovery time and paves the way toward the realization of energy-efficient optoelectronic device applications.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c01239.

Schematic diagram with a flow chart for chemical synthesis of Au NPs and process of their decoration on as-grown GaN-NSs (GaN-NIs and GaN-NFs), morphological information of Au NP-decorated GaN-NSs by FESEM and atomic force microscopy images, elemental information of Au NP-decorated GaN-NSs by elemental mapping and EDS spectra, and power-dependent NEP and responsivity graph of Au NP/GaN-NF UV PDs at self-powered mode (PDF)

AUTHOR INFORMATION

Corresponding Author

Govind Gupta − *CSIR-National Physical Laboratory, New Delhi 110012, India; Academy of Scientific and Innovative Research, Ghaziabad, Uttar Pradesh 201002, India;* orcid.org/0000-0001-5006-7623; Phone: +91-1145609503; Email: govind.npl@nic.in

Authors

Lalit Goswami − *Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi 110042, India; CSIR-National Physical Laboratory, New Delhi 110012, India*

Neha Aggarwal − *CSIR-National Physical Laboratory, New Delhi 110012, India; Academy of Scientific and Innovative Research, Ghaziabad, Uttar Pradesh 201002, India*

Shibin Krishna − *CSIR-National Physical Laboratory, New Delhi 110012, India; Academy of Scientific and Innovative Research, Ghaziabad, Uttar Pradesh 201002, India*

Manjri Singh − *CSIR-National Physical Laboratory, New Delhi 110012, India*

Pargam Vashishtha − *CSIR-National Physical Laboratory, New Delhi 110012, India; Academy of Scientific and Innovative Research, Ghaziabad, Uttar Pradesh 201002, India*

Surinder Pal Singh − *CSIR-National Physical Laboratory, New Delhi 110012, India*

Sudhir Husale − *CSIR-National Physical Laboratory, New Delhi 110012, India*

Rajeshwari Pandey − *Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi 110042, India*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c01239

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Director, CSIR-NPL, New Delhi, for his everlasting encouragement and support. The work is supported by the DST via grant DST/TM/CERI/C245 (C). One of the authors, N.A., acknowledges the CSIR for the financial support under CSIR-Research Associateship.

REFERENCES

(1) Monroy, E.; Omnes, F.; Calle, F. Wide-bandgap semiconductor ultraviolet photodetectors. *Semicond. Sci. Technol.* **2003**, *18*, R33−R51.

(2) Fu, N.; Li, E.; Cui, Z.; Ma, D.; Wang, W.; Zhang, Y.; Song, S.; Lin, J. The electronic properties of phosphorus-doped GaN nanowires from first-principle calculations. *J. Alloys Compd.* **2014**, *596*, 92−97.

(3) Yang, Q.; Wu, Y.; Liu, Y.; Pan, C.; Wang, Z. L. Features of the piezo-phototronic effect on optoelectronic devices based on wurtzite semiconductor nanowires. *Phys. Chem. Chem. Phys.* **2014**, *16*, 2790−2800.

(4) Tchoulfian, P.; Donatini, F.; Levy, F.; Dussaigne, A.; Ferret, P.; Pernot, J. Direct imaging of p-n junction in core-shell GaN wires. *Nano Lett.* **2014**, *14*, 3491−3498.

(5) Qian, F.; Wang, H.; Ling, Y.; Wang, G.; Thelen, M. P.; Li, Y. Photoenhanced Electrochemical Interaction between Shewanella and a Hematite Nanowire Photoanode. *Nano Lett.* **2014**, *14*, 3688−3693.

(6) Chang, S.; Chang, M.; Yang, Y. Enhanced Responsivity of GaN Metal−Semiconductor−Metal (MSM) Photodetectors on GaN Substrate. *IEEE, J. Phot.* **2017**, *9*, 1−7.

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

(7) Zhang, X.; Liu, Q.; Liu, B.; Yang, W.; Li, J.; Niu, P.; Jiang, X. Giant UV photoresponse of a GaN nanowire photodetector through effective Pt nanoparticle coupling. *J. Mater. Chem. C* **2017**, *5*, 4319.
(8) Shetty, A.; Sundar, K. J.; Roul, B.; Mukundan, S.; Chandan, G.; Mohan, L.; Ghosh, A.; Vinoy, K. J.; Krupanidhi, S. B. Plasmonic Enhancement of Photocurrent in GaN-Based UV Photodetectors. In *Emerging Electronics (ICEE); IEEE 2nd International Conference*, 2014.
(9) Li, D.; Sun, X.; Song, H.; Li, Z.; Chen, Y.; Jiang, H.; Miao, G. Realization of a High-Performance GaN UV Detector by Nanoplasmonic Enhancement. *Adv. Mater.* **2012**, *24*, 845−849.
(10) Khan, R.; Uthirakumar, P.; Bae, K.-B.; Leem, S.-J.; Lee, I.-H. Localized surface plasmon-enhanced photoluminescence of ZnO nanosheets by Au nanoparticles. *Mater. Lett.* **2016**, *163*, 8−11.
(11) Su, L.; Qin, N. A facile method for fabricating Au-nanoparticles-decorated ZnO nanorods with greatly enhanced near-band-edge emission. *Ceram. Int.* **2015**, *41*, 2673−2679.
(12) Amendola, V.; Pilot, R.; Frasconi, M.; Maragò, O. M.; Iatì, M. A. Surface plasmon resonance in gold nanoparticles: a review. *J. Phys.: Condens. Matter* **2017**, *29*, 203002.
(13) Amendola, V.; Meneghetti, M. Laser ablation synthesis in solution and size manipulation of noble metal nanoparticles. *Phys. Chem. Chem. Phys.* **2009**, *11*, 3805−3821.
(14) Daniel, M.-C.; Astruc, D. Gold Nanoparticles: Assembly, Supramolecular Chemistry, Quantum-Size-Related Properties and Applications toward Biology, Catalysis and Nanotechnology. *Chem. Rev.* **2004**, *104*, 293−346.
(15) Austin, L. A.; Kang, B.; El-Sayed, M. A. Probing molecular cell event dynamics at the single-cell level with targeted plasmonic gold nanoparticles: A review. *Nano Today* **2015**, *10*, 542−558.
(16) Bohren, C. F. *Absorption and Scattering of Light by Small Particles*, 2nd ed.; Huffman, D. R., Ed.; Wiley-Interscience: New York, 1998; Vol. *4*, p 93.
(17) Turkevich, J.; Stevenson, P. C.; Hillier, J. A study of the nucleation and growth processes in the synthesis of colloidal gold. *Discuss. Faraday Soc.* **1951**, *11*, 55−75.
(18) Zeng, S.; Yu, X.; Law, W.-C.; Zhang, Y.; Hu, R.; Dinh, X.-Q.; Ho, H.-P.; Yong, K.-T. Size dependence of Au NP-enhanced surface plasmon resonance based on differential phase measurement. *Sens. Actuators, B* **2013**, *176*, 1128−1133.
(19) Haiss, W.; Thanh, N. T. K.; Aveyard, J.; Fernig, D. G. Determination of Size and Concentration of Gold Nanoparticles from UV−Vis Spectra. *Anal.chem.* **2007**, *79*, 4215−4221.
(20) Devi, P. S.; Banerjee, S.; Chowdhury, S. R.; Kumar, G. S. Eggshell membrane: a natural biotemplate to synthesize fluorescent gold nanoparticles. *RSC Adv.* **2012**, *2*, 11578.
(21) Goswami, L.; Pandey, R.; Gupta, G. Epitaxial growth of GaN nanostructure by PA-MBE for UV detection application. *Appl. Surf. Sci.* **2018**, *449*, 186.
(22) Aggarwal, N.; Krishna, S.; Sharma, A.; Goswami, L.; Kumar, D.; Husale, S.; Gupta, G. A highly responsive self-driven UV photodetector using GaN nanoflowers. *Adv. Electron. Mater.* **2017**, *3*, 1700036.
(23) Khan, A.; Jadwisienczak, W. M.; Kordesch, M. E. From Zn microspheres to hollow ZnO microspheres: A simple route to the growth of large scale metallic Zn microspheres and hollow ZnO microspheres. *Phys E* **2006**, *33*, 331−335.
(24) Wan, Q.; Wang, T. H.; Zhao, J. C. Enhanced photocatalytic activity of ZnO nanotetrapods. *Appl. Phys. Lett.* **2005**, *87*, 083105.
(25) Hofmann, J.; Steinmann, W. Plasma resonance in the photoemission of silver. *Phys. Status Solidi* **1968**, *30*, K53−K56.
(26) Sipe, J. E.; Becher, J. Surface-plasmon-assisted photoemission. *J. Opt. Soc. Am.* **1981**, *71*, 1286−1288.
(27) Maier, S. A. *Plasmonics: Fundamentals and Applications*; Springer, 2007.
(28) Kreibig, U.; Vollmer, M. *Optical Properties of Metal Clusters*; Springer: Berlin, 1995.
(29) Amendola, V.; Pilot, R.; Frasconi, M.; Maragò, O. M.; Iatì, M. A. Surface plasmon resonance in gold nanoparticles: a review. *J. Phys.: Condens. Matter* **2017**, *29*, 203002.
(30) Austin, L. A.; Kang, B.; El-Sayed, M. A. Probing molecular cell event dynamics at the single-cell level with targeted plasmonic gold nanoparticles: A review. *Nano Today* **2015**, *10*, 542−558.
(31) Bharadwaj, P.; Deutsch, B.; Novotny, L. Optical antennas. *Adv. Opt. Photon.* **2009**, *1*, 438−483.
(32) Li, X.; Xiao, D.; Zhang, Z. Landau damping of quantum plasmons in metal nanostructures. *New J. Phys.* **2013**, *15*, 023011.
(33) Manjavacas, A.; Liu, J. G.; Kulkarni, V.; Nordlander, P. Plasmon-induced hot carriers in metallic nanoparticles. *ACS Nano* **2014**, *8*, 7630−7638.
(34) Watanabe, K.; Menzel, D.; Nilius, N.; Freund, H.-J. Photochemistry on metal nanoparticles. *Chem. Rev.* **2006**, *106*, 4301−4320.
(35) Chiang, T.-C.; Chiu, C.-Y.; Dai, T.-F.; Hung, Y.-J.; Hsu, H.-C. Surface-plasmon-enhanced band-edge emission and lasing behaviours of Au-decorated ZnO microstructures. *Opt. Mater. Express* **2017**, *7*, 313.
(36) Goswami, L.; Pandey, R.; Gupta, G. Ultra-Thin GaN nanostructures based self-powered ultraviolet photodetector via non-homogeneous Au-GaN interfaces. *Opt. Mater.* **2020**, *102*, 109820.
(37) Saron, K. M. A.; Hashim, M. R.; Naderi, N.; Allam, N. K. Enhanced light-sensing characteristics of nanostructured gallium nitride/silicon heterojunctions: Interface matters. *J. Appl. Phys.* **2013**, *114*, 134510.
(38) Dhar, S.; Majumder, T.; Mondal, S. P. Graphene Quantum Dot-Sensitized ZnO Nanorod/Polymer Schottky Junction UV Detector with Superior External Quantum Efficiency, Detectivity and Responsivity. *ACS Appl. Mater. Interfaces* **2016**, *8*, 31822.
(39) Alaie, Z.; Mohammad Nejad, S.; Yousefi, M. H. Recent advances in ultraviolet photodetectors. *Mat. Sci. Semicon. Proc.* **2015**, *29*, 16−55.
(40) Jia, R.; Zhao, D.; Gao, N.; Duo, L. Polarization Enhanced Charge Transfer: Dual-Band GaN-Based Plasmonic Photodetector. *Sci. Rep.* **2017**, *7*, 40483.
(41) Jin, Z.; Gao, L.; Zhou, Q.; Wang, J. High-performance flexible ultraviolet photoconductors based on solution-processed ultrathin ZnO/Au nanoparticle composite films. *Sci. Rep.* **2014**, *4*, 4268.
(42) Liu, Y.; Zhang, X.; Su, J.; Li, H.; Zhang, Q.; Gao, Y. Ag nanoparticles@ZnO nanowire composite arrays: an absorption enhanced UV photodetector. *Opt. Express* **2014**, *22*, 30148.
(43) Qin, L.; Shing, C. Metal−Semiconductor−Metal Ultraviolet Photodetectors Based on Zinc-Oxide Colloidal Nanoparticles. *IEEE Electron Device Lett.* **2011**, *32*, 51.
(44) Sawyer, S.; Qin, L.; Shing, C. Zinc oxide nanoparticles for ultraviolet photodetection. *Int. J. High-Speed Electr. Syst.* **2011**, *20*, 183−194.
(45) Khokhra, R.; Bharti, B.; Lee, H.-N.; Kumar, R. Visible and UV photo-detection in ZnO nanostructured thin films via simple tuning of the solution method. *Sci. Rep.* **2017**, *7*, 15032.
(46) Chakrabartty, S.; Mondal, A.; Sarkar, M. B.; Choudhuri, B.; Saha, A. K.; Bhattacharyya, A. TiO_2 Nanoparticles Arrays Ultraviolet-A Detector With Au Schottky Contact. *IEEE Photonics Technol. Lett.* **2014**, *26*, 1065−1068.
(47) Molina-Mendoza, A. J.; Moya, A.; Frisenda, R.; Svatek, S. A.; Gant, P.; Gonzalez-Abad, S.; Antolin, E.; Agraït, N.; Rubio-Bollinger, G.; Lara, D. P.; Vilatela, J. J.; Castellanos-Gomez, A. Highly responsive UV-photodetectors based on single electrospun TiO_2 nanofibres. *J. Mater. Chem. C* **2016**, *45*, 10707.

https://dx.doi.org/10.1021/acsomega.0c01239
ACS Omega XXXX, XXX, XXX−XXX

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

Autonomous Model Predictive Controlled Smart Inverter with Proactive Grid Fault Ride-Through Capability

Mitchell Easley, *Student Member, IEEE*, Sarthak Jain, *Member, IEEE*, Mohammad Shadmand, *Member, IEEE*, Haitham Abu-Rub, *Fellow, IEEE*

***Abstract*— This paper presents an autonomous model predictive controlled smart photovoltaic (PV) inverter with proactive grid fault-ride through capability. The proposed smart inverter control features decoupled active and reactive power. It can seamlessly switch between the anticipated modes of operation based on grid condition or grid operator command. The smart inverter autonomously adjusts its active and reactive power set-points according to the grid condition, it operates in maximum power point tracking (MPPT) and low voltage ride through (LVRT) modes in normal grid and faulty grid condition modes, respectively. The proposed novel autonomous model predictive control (AMPC) scheme is leveraged to enhance the operation of the smart inverter. The AMPC includes online weight factor auto-tuning and control objective normalization to eliminate the required trial-and-error weight factor design stage in conventional model predictive control. This feature is particularly beneficial to the dual-mode smart inverter operation. The performance of the proposed grid-tied smart inverter based on the AMPC is verified experimentally. The results demonstrate that the proposed AMPC-based smart inverter features robust grid fault detection, autonomous adjustment of active and reactive power set-points, seamless transition between modes of operation, and elimination of controller tuning effort.**

***Index Terms*—model predictive control, smart inverters, low voltage ride through, impedance source inverter**

I. Introduction

The interest of photovoltaic (PV) power generation has been steadily increasing, as evidenced by the substantial market growth from the last two decades [1]. To improve the commercial viability, the effects of high PV penetration of the grid must be addressed. Stability is a critical issue when considering a high PV-penetrated grid [2]. Voltage swelling can occur as a result of rapid increases in PV power generation/solar irradiance, referred to as *overloading* [3]. Further issues occur when the PV system transitions to islanded mode in the event of grid faults [1]. It is desirable to seek alternatives to disconnection during grid faults. The concept of low-voltage ride through (LVRT) is a proposed method of grid voltage support, where the grid-connected inverter will inject reactive power prior to disconnection upon sensing a dip in grid voltage. By applying such a control feature, the smart inverter can turn PV generation from a stability hazard to a stability asset. In recent years, there has been discussion to update grid codes pertaining to PV inverters to include this capability, and has been implemented in some grid codes [4-6]. However, adjusting the active and reactive power set-points to produce these advanced grid-support features makes for a more difficult control problem.

Creating the control scheme for grid-connected PV inverters is a complicated control problem in of itself, as cascaded control loops are generally required. This is because grid-connected PV strings are typically introduced to the grid through a two-stage converter scheme. That is, a DC/DC boost converter followed by a DC/AC inversion stage [7, 8]. The first stage performs the maximum power point tracking (MPPT) and boosts the PV voltage to meet the DC-bus voltage requirement, while the DC/AC inversion stage regulates DC-bus voltage and ensures PV-to-grid power flow. Converters which contain an impedance network, termed impedance source inverters or Z-source inverters (ZSI), have been proposed for such an application [9-11]. The arguments for these topologies are the possibility of removing the DC-DC converter or removing the need for a transformer on the AC side, potentially reducing the overall converter expenditure, improving the efficiency [12]. The permitted *shoot-through state* can adjust the input impedance of the converter, making it possible to perform a maximum power point tracking (MPPT) control, and can step up/down the DC source voltage to meet the desired DC link voltage level. The voltage-fed quasi impedance source inverter (qZSI) is generally considered more practical among the Z-source topologies for PV applications, as it permits continuous input current, positively influencing the lifetime of the PV string [13, 14]. Despite the implementation of a single-stage topology, dual-stage control structures are still proposed, on account of the contrast in dynamics at the AC and DC sides of the circuit. In [15], such a control structure is proposed, and even includes both a current-control loop for grid-connected applications and voltage-control for islanded mode. More traditional control schemes with cascaded structure such as these tend to suffer from slower dynamic response. Additionally, incorporating advanced, grid-supporting functionality such as LVRT capability is not straightforward.

Model predictive control (MPC) is a potential candidate to address the challenges in control of grid-tied PV inverters with advanced functionalities, as it has become a competitive control scheme [16-18]. It has also been recently implemented in industry [19]. With MPC, particularly the subset known as optimal switching sequence, the switching state is selected via minimizing a cost function, which can contain several control objectives of different natures. Thus, implementing adjustable

Mitchell Easley and Mohammad Shadmand are with the Department of Electrical and Computer Engineering at Kansas State University, Manhattan, KS (e-mail: mitch01@ksu.edu, mshadmand@ksu.edu. Sarthak Jain is with the Core Power Solutions Group in Renesas Electronics, Milipitas, CA (e-mail: jainsarthak2091@gmail.com). Haitham Abu-Rub is with the Department of Electrical and Computer Engineering at Texas A&M University at Qatar, Doha, Qatar (e-mail: haitham.abu-rub@qatar.tamu.edu)

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

active and reactive power set-points is actualized by adjusting references of the cost function, and thus will have no bearing on the control structure. However, what allows multi-objective MPC to boast its simplicity is what causes difficulty in the design stage. To adjust the *impact* of each control objective in the cost function, each penalty term is multiplied by a weight factor. This means implementing a multivariable cost function generally requires preliminary tuning. The optimization of the weight factors is a design burden tied to traditional MPC. Discussion in literature on how to design the weight factors is limited to trial and error techniques for finding an optimal set-point [20]. This also requires the user to define *optimal behavior* when designing the controller, which is not straightforward for multi-objective optimization. To be succinct, the optimal weight factor ratio of multi-objective MPC tends to be laborious to navigate and difficult to define. Additionally, a static weight factor ratio will likely not have the best achievable performance for all considerable scenarios i.e. MPPT versus LVRT operation modes and the transient periods therein. Such issues exist for those who attempt to implement previously proposed predictive control schemes for the qZSI, such as [21, 22]. The authors in [23] acknowledge the issue of static weight factors during transient conditions, and address the preliminary design of the weight factors for each mode must still be done, which is especially difficult for the implemented cost function which contains four penalty terms. In [24], a predictive control scheme for a four-leg qZSI is proposed, but is not considered for grid-connected applications. Predictive current control of the qZSI is also considered in [25], where the authors address the difficult cost function design by optimizing the objectives hierarchically, ultimately cutting the iterative computation in half. Still, the considered application is limited to a resistive load with an ideal DC source. In [26], the authors propose a predictive control for a battery-assisted qZSI, which reduces iterative computation by using the predictive control to decide if shoot-through should be implemented in the next-state. This concept is extended to a cascaded battery-assisted qZSI topology in [27], where the control is generalized to any number of qZSI converters (cells) in series. However, output current control is not implemented in the predictive control, and requires a proportional resonant controller and modulation scheme for output current tracking. The authors in [28] propose a time-averaged model of the qZSI dynamics to produce an optimal shoot-through duty ratio and modulation index. Unlike finite-set MPC, this control implements a modulator and thus does not include the fast-dynamic performance associated with direct computation of an optimal switching sequence. Additionally, the proposed control in [28] is for isolated qZSI (off-grid) with constant DC Source and R-L load.

In order to include the aforementioned control features while addressing the issues of traditional MPC, this paper proposes an autonomous model predictive control (AMPC) solution to improve the capabilities and robustness of grid-connected PV systems. Specifically, the active/reactive power of the smart inverter are decoupled, and the set-points are adjusted according to the grid condition. Three reactive power injection (RPI) strategies are shown, which adjust the active power reference [29]. The active power is adjusted using a flexible power point tracking technique, which can pull the PV module's voltage above its maximum power point to reduce power injection. The paper also proposes a method which removes any tuning of the weight factors by autonomously adjusting the weight factors according to the controller's tracking performance of each objective. This, in addition to the seamless transitions between the MPPT and LVRT modes of operation according to the grid's condition, make the proposed control considered autonomous. As a case study, the proposed control is implemented using a quasi-Z-source inverter (qZSI). The buck-boost capabilities of the inverter make it possible to use as a single-stage solution for grid-connected photovoltaics This paper is the first to consider an LVRT-enabled smart PV inverter for this topology via the proposed AMPC. This work is considered an extension of the work in [30], with additional analysis, experimental verification, and more advanced predictive control which features the auto-tuning weight factor technique. The tracking improvements from the auto-tuning feature are demonstrated experimentally, comparing to the original, static weight factor ratio in [30]. Additionally, this proposed control scheme implements an index variable, *s*, in the cost function. This variable is used to eliminate iterative computations associated with redundant next-state predictions. Without any degradation of tracking capabilities, five of the ten next-state prediction calculations are removed using this technique, enhancing the control's computational efficiency.

Beyond the Introduction, Section II explains the decoupled control scheme and its formulation. Section III presents the auto-tuning weight factor feature. Section IV includes several experiments to validate the performance of the proposed control scheme. Finally, the proposed system and its performance are summarized in the conclusion section.

II. System Description and Control Strategy

Fig. 1 shows the qZSI-based PV system along with the proposed control scheme. The control works in two modes: (*i*) normal operation mode when grid voltage is within 90%-110% [29] of its rating, in this mode the predictive MPPT technique is triggered to harvest the maximum power from PV; (*ii*) the LVRT mode during grid voltage sags of more than 10%, in this mode the qZSI will move from MPP operation to support grid as an ancillary service from PV inverter.

The qZSI, shown in Fig. 1, is analyzed for active states and shoot through state separately to develop the system model which are further required for designing the predictive controller. Considering full switching period to be T, active state time is taken as T_1 and shoot through time is taken as T_0. Table I shows the summary of dynamic equations of qZSI in both during active/non-shoot through state and shoot-through state. By applying volt-sec balance on L_1 and L_2 and equating the average voltage over a switching cycle to zero, the capacitors C_1 and C_2 voltages and dc-link voltage is given by:

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

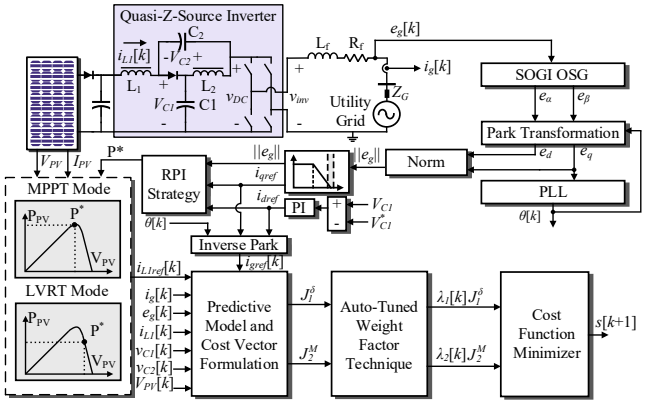


Fig. 1. Proposed power electronic interface for PV applications with LVRT capability.

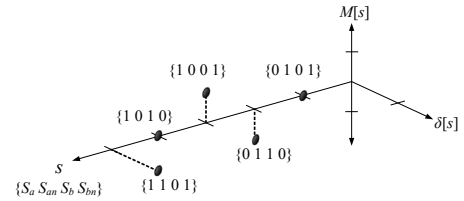


Fig. 2. Total optimization set. The switching states are defined and their relationship with M and δ are shown explicitly.

TABLE I: CURRENT REFERENCES IN LVRT MODE

	Const. Ave. P	Const. Peak Current	Const. I_d
I_d	I_N/v_g	$[1-\gamma^2(1-v_g)]^{1/2}I_N$	$2P/v_{g,rated}$
I_q	$\gamma(1-v_g)I_N$	$\gamma(1-v_g)I_N$	$\gamma(1-v_g)I_N$

$$V_{C1} = \frac{(1-D)\times V_{PV}}{1-2D} \tag{1}$$

$$V_{C2} = \frac{D\times V_{PV}}{1-2D} \tag{2}$$

$$V_{DC} = \frac{V_{PV}}{1-2D} \tag{3}$$

where D is the shoot through duty ratio, defined as T_0/T. Gernerally, qZSI control requires multiple system elements to be controlled simultaneously. For a qZSI, V_{C1}, i_{L1} and i_g are controlled, however, the proposed algorithm does not include V_{C1} in the cost function; a PI controller is used for the regulation of V_{C1} by appropriately generating the $i_{d,ref}$ (active power component of reference grid current) needs to be injected into the grid by qZSI which is achieved by MPC.

A. Quasi-Z-Source Inverter Modeling

By using the voltage-current relationship of an inductor and applying the Forward Euler method to the dynamic inductor current equations [13], the discretized model for i_{L1} is determined in (4):

$$i_{L1}[k+1] = i_{L1}[k] + T_S(L1)^{-1}\left(V_{PV}[k] - \delta V_{C1}[k] + (1-\delta)V_{C2}[k]\right) \quad \delta \in [0,1] \tag{4}$$

where $i_{L1}[k+1]$ is the predicted value of inductor current and δ is non-shoot through indicator as described earlier. Thus, the two-element sub-cost vector is defined as:

$$J_1^{\delta} = \left|i_{L1}[k+1] - i_{L1,ref}[k]\right| \times \frac{1}{i_{L1,RMS}[k]} \tag{5}$$

$i_{L1,RMS}[k]$ is used to normalize the cost and is discussed further in the following section.

The KVL equation for the grid side (6) is used for the discretization of i_g in (7) using Euler forward method:

$$v_{inv} = Ri_g + L\frac{di_g}{dt} + e_g \tag{6}$$

$$i_g[k+1] = i_g[k]\left(1 - \frac{R}{L}T_S\right) + \frac{T_S}{L}(v_{inv}[k] - e_g[k]) \tag{7}$$

$$v_{inv}[k] = M \times v_{DC}[k] = M \times \left(V_{C1}[k] + V_{C2}[k]\right) \quad M \in [-1,0,1] \tag{8}$$

Seeing that the output current prediction is dependent on the output voltage level M, the output current cost vector is defined:

$$J_2^M = \left|i_g[k+1] - i_{g,ref}[k]\right| \times \frac{1}{i_{g,RMS}[k]} \tag{9}$$

$i_{g,rms}[k]$ is a rolling RMS measurement and its formulation is discussed in section 3.B. Note that the cost vectors are dependent on the next-state predictions, which are dependent either on δ or M. Ultimately, these two variables are coupled, as shoot-through constrains the output voltage level. The true independent variable is switching state s shown in Fig. 2, but the cost vectors $J^{\delta}{}_1$ and $J^M{}_2$ are defined according to δ and M respectively instead of s to reduce the vector lengths, thus reducing iterative computation.

Finally, the sub-cost vectors, with their corresponding weight factors, are summed in the cost function vector:

$$s_{opt} = \arg\underset{s[k+1]}{\text{minimize}}(J[s])$$
$$J[s] = \lambda_1 J_1^{\delta} + \lambda_2 J_2^M \tag{10}$$

where s unites M, δ, and the gate signals associated with each switching state. Formulation of the time-variant weight factors λ_1 and λ_2 is discussed in section III. Although (10) requires a second pass through the sub-cost vectors to compute J, the next-state predictions do not need to be recomputed for cost function optimization. Control in both normal and LVRT modes are briefly explained in Sections II. B and II. C.

B. Normal Grid Mode

The system will operate in this mode when the grid voltage magnitude is within ±10% of the rated grid voltage. The MPPT algorithm, shown in Fig. 3, is used to generate the PV current reference or average input inductor current i_{L1} reference for operation at MPP. The detailed explanation of this algorithm is provided in sub-section 2. *D*. As illustrated in Fig. 1, the PI controller will generate the real power component

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

current reference i_{dref} to be injected by the qZSI to the grid. In this mode, reactive power component i_{qref} is kept at 0, making the qZSI work at unity power factor. The i_{dref} and i_{qref} are then converted to the phase current i_{gref} and fed to the MPC cost function along with the $i_{L1,ref}$ from the MPPT algorithm.

C. LVRT Mode

This mode of operation is triggered for grid voltage sag of more than 10%. SOGI based PLL [31] is used to detect the voltage magnitude as well as the phase for the control algorithm as shown in the control schematic of Fig. 1. The proposed system is tested for three reactive power injection strategies: constant average active power, constant peak current, and constant active current. The equations used to develop the current references for these three strategies are given in Table I. I_N is the nominal current rating of the inverter and the Euclidean distance between the reactive and active current components. Each term in the equations of Table I, except for I_N and P, is in per unit (p.u.). Note that all the RPI strategies respond equally to voltage sags, thus these RPI strategies vary only in how the active current injection is compensated. In these equations, γ represents the ramping rate of reactive power injection. More explicitly, the inverter will inject reactive power associated with the current rating when the grid drops to or below (1-1/γ) p.u.. A γ value of 2 p.u. has been suggested in literature, but these standards vary and can be adopted based on country of operation [29]. Specific implementation of these RPI strategies in cohesion with the MPPT is discussed in the next section.

D. MPPT with Flexible Power Point Tracking

The MPPT is achieved by creating the $i_{L1,ref}$ which is an input to the J^{δ}_{1} calculation. Although the input capacitance distinguishes I_{PV} from i_{L1}, their DC components are equal, thus $i_{L1,ref}$ is a suitable output of the MPPT. The flexible power point on the MPPT algorithm is what allows the LVRT mode and MPPT modes to be distinguished. Although the reactive current references from Table I are applied directly to the formulation of J^{M}_{2} using a lookup table, the active current component must be created indirectly, as the active current reference is developed by the PI controller which regulates V_{C1}. Specifically, to reach the active current references outlined in Table I during grid voltage sags, the power drawn from the PV string must be reduced (except for the constant average active power strategy). With a reduction in power sent across the DC bus, the active current reference will decrease to maintain the voltage across C1. Thus, each reactive power injection strategy holds a unique power reference to reach the active current specified in Table I. This is explained below:

Constant average active power: During voltage sags, the reference power is held constant at P_{MPP}. With a reduction in grid voltage, the active power injected decreases initially. This results in more energy being stored in C1, causing an initial voltage swell at the DC link. To return V_{C1} to the reference value, the PI increases i_{dref}. This, in addition to the i_{qref}, will substantially increase the output current amplitude.

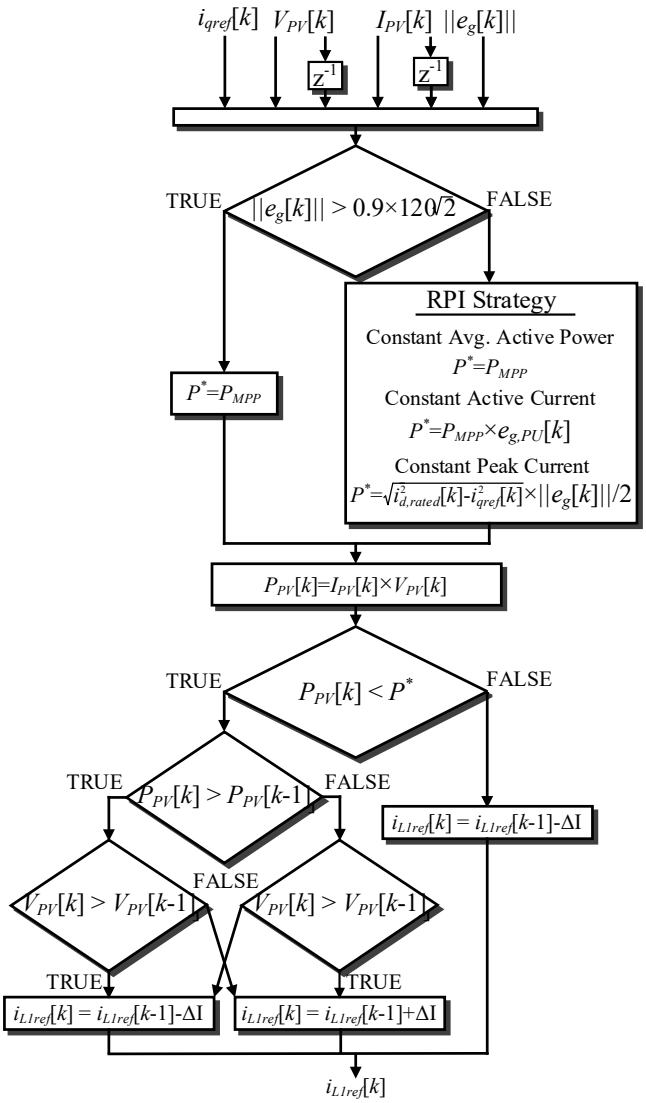


Fig. 3. RPI strategies and MPPT algorithm with flexible power point.

Constant active current: To maintain the active current component during grid voltage sags, V_{C1} should not experience a transient. This can be achieved by reducing the harvested PV power such that it is commensurate with the grid voltage sag. Thus, the power reference is defined as:

$$P^*[k] = P_{MPP} \times e_{g,PU}[k] \tag{11}$$

where the per-unit grid voltage $e_{g,PU}[k]$ is:

$$e_{g,PU}[k] = \frac{\sqrt{e_d[k]^2 + e_q[k]^2}}{120\sqrt{2}} \tag{12}$$

$e_d[k]$ and $e_q[k]$ are the decoupled components of the grid voltage. Although this power reduction ideally does not create a transient in the active current component, the non-zero response time of the MPPT to the grid voltage sag will create a

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

brief swell on the DC-link voltage and a corresponding i_{dref} swell. The settling time of this transient at the DC-link is dependent on the MPC sampling time and current step-size of the MPPT algorithm.

Constant peak current: To maintain the injected peak current that occurred prior to the grid voltage sag, the harvested PV power must decrease in concordance both with the grid voltage sag and the i_{qref}. For this strategy, P^* is constructed as:

$$P^*[k] = \frac{120\sqrt{2}e_{g,PU}[k]}{2}\sqrt{\left(\sqrt{2}i_{g,rated}\right)^2 - i_{q,ref}^2[k]} \quad (13)$$

Similar to the constant active current strategy, a voltage sag will initially create a swell on the DC link voltage and i_{dref}. However, the reduction in PV harvesting will be substantial enough such that C1 starts drawing out of its energy reserve, creating a brief sag in DC link voltage. Thus, the PI will reduce i_{dref} below its value before the swell to compensate.

A reference power term P^* is not used in traditional MPPT algorithms, as the reference power is the maximum extractable power. To adhere to the RPI strategies, a conditional statement is applied prior to the traditional MPPT, shown in Fig. 3. The extracted PV power is compared with P^*. If the extracted PV power exceeds P^*, the power point tracker immediately reduces the $i_{L1,ref}$. Reducing $i_{L1,ref}$ will pull the PV string to the right of the power-voltage characteristic curve. For a current-controlled MPPT, this side of the MPP is more stable, as the $|dP_{PV}/dI_{PV}|$ is lower.

III. Auto-Tuning of Weight Factors and Online Normalization

A. Auto-Tuning Algorithm

When considering multiple parameters in MPC, designing the cost function weight factors is a laborious task. These parameters, often with different units and orders of magnitude, are conglomerated into a single cost function subject to minimization, and finding a systematic approach for optimizing these weight factors is still a topic of discussion in literature [16, 32]. This section proposes an auto-tuning algorithm for the weight factor in the MPC cost function (10). The method of weight factor auto-tuning is based on an intuitive concept: poorly-tracked objectives should have increased weight in the cost function. This algorithm evaluates the tracking of each control objective, and makes the weight commensurate with the tracking error. The algorithm starts by collecting the minimized cost for each objective. Note this is distinct from the traditional optimization in MPC, as finite-set MPC typically only considers the overall cost function, or the sum of the cost terms. The two optimal sub-costs are defined as:

$$\zeta_1[k] = \min(J_1^{\delta}) \quad (14)$$

$$\zeta_2[k] = \min(J_2^{M}) \quad (15)$$

This auto-tuning procedure is first explained as a general procedure, then the implementation is described.

The auto-tuning weight factor procedure is explained for only one objective without loss of generality. Consider a threshold of acceptable objective error ε. In other words, if it is possible for the qZSI to keep a normalized control objective less than ε from its reference, then the objective is considered sufficiently tracked and no remedial action is necessary. This condition can be written concisely as:

$$\zeta_1[k] \le \varepsilon \Rightarrow \lambda_1[k] = \nu \quad (16)$$

where ν is an arbitrary value attached to both weight factors. Note that the actual value of ν is not important, as only the weight factor ratio will affect system performance. If the condition in (16) is not satisfied, then a larger value of the weight factor should be selected to place more emphasis on J^{δ}_1, promoting its reduction for the next sampling time. The evaluation of λ_1 when $\zeta_1[k]$ exceeds ε is as follows:

$$\begin{aligned} \zeta_1[k] &\le 2\varepsilon \Rightarrow \lambda_1 = 2\nu \\ \zeta_1[k] &\le 3\varepsilon \Rightarrow \lambda_1 = 3\nu \\ &\vdots \\ \zeta_1[k] &\le n_1\varepsilon \Rightarrow \lambda_1 = n_1\nu \quad n_1 \in \mathbb{N} \end{aligned} \quad (17)$$

Thus, each objective is weighted in J according to the controller's tracking performance.

Applying a conditional loop of indefinite size as is shown in (17) can be problematic for a controller operating at a fixed sampling rate. However, this concept can be applied to the controller in a definite time by recalling that the term ν is arbitrary. By setting ν equal to one, λ_1 can be computed directly:

$$\begin{aligned} (n_1 - 1)\varepsilon \le \zeta_1[k] \le n_1\varepsilon &\Rightarrow \lambda_1[k] = n_1 \\ roundup\left(\frac{\zeta_1[k]}{\varepsilon}\right) = n_1 &= \lambda_1[k] \\ \lambda_1[k] &= roundup\left(\frac{\zeta_1[k]}{\varepsilon}\right) \end{aligned} \quad (18)$$

where *roundup* refers to a function which rounds the inner argument up to the nearest integer. Thus, the indefinite loop has been removed while maintaining the original auto-tuning concept.

Although an adaptive and performance-based weight factor is desirable, next-state predictions can have substantial instantaneous error for lower sampling rates. Considering a moving window of tracking performance can improve stability by reducing the effect of high frequency error attributed to discrete predictions. The weight factors are computed using a moving average of the value computed in (19). The weight factors are stored in a vector of length N. Thus $\zeta[k]$ to $\zeta[k\text{-}N\text{+}1]$ contribute to the associated objective's weight factor $\lambda[k]$. Rather than placing the elements at each end from newest to oldest, a pointer variable cycles through the vector, replacing the variable it is pointing to with $\zeta[k]$. As long as the pointer follows a cycling pattern, $\zeta[k]$ will always replace $\zeta[k\text{-}N\text{+}1]$,

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

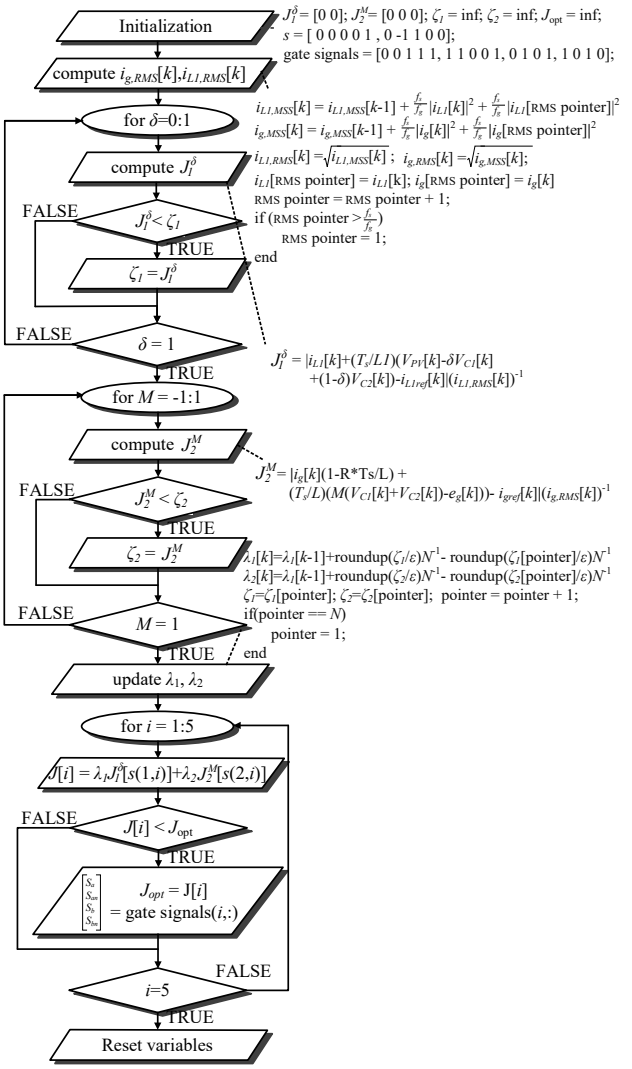


Fig. 4. Flowchart of the predictive model, auto-tuning technique, and cost-vector minimization.

regardless of the pointer's location in the vector. The moving average can be updated efficiently with the following equation:

$$\lambda_x[k] = \frac{1}{N}\sum_{i=0}^{N-1}\zeta_x[k-i] = \lambda_x[k-1] + \frac{\zeta_x[k]}{N} - \frac{\zeta_x[k-N+1]}{N} \qquad x \in [1,2] \tag{19}$$

Thus, (19) is used to update the weight factor for both objectives with the number of computations independent of the window size.

B. Adaptive Normalization of Control Objectives

Rather than hard-programming an expected operating point of the control objectives or dividing by the rated values, the objectives are normalized using previous measurement data. Specifically, the root-mean-square (RMS) is computed over a moving window. The RMS is computed efficiently by first updating the mean squared sum of each objective:

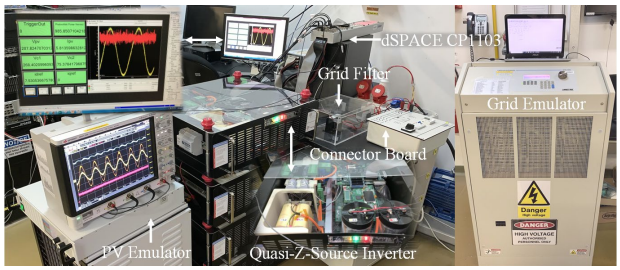


Fig. 5. Hardware prototype of single-phase grid-connected qZSI.

TABLE II: SYSTEM SPECIFICATIONS

Parameter	Value
Grid Frequency	60Hz
MPC Sampling Frequency	50kHz
MPPT Sampling Frequency	6.25kHz
PI controller Sampling Frequency	3.125kHz
PI controller proportional/integral gains	0.008/0.08
RMS window size	834
weight factor window size N	100
ε	1e-4
V_{MPP}/V_{OC}	200V/239V
I_{MPP}/I_{SC}	5A/5.4A
L_1/L_2	0.4μH
C_1/C_2	1.5mF
C_{PV}	200μF
L_f	2.5mH
R_f	200 mΩ
Ramping rate γ	4 p.u.
$i_{g,rated}$	8.33A_{RMS}
Grid Voltage	120 V_{RMS}
V^*_{C1}	275V

$$i_{x,MSS}[k] = \frac{f_g}{f_s}\sum_{i=0}^{\frac{f_s}{f_g}-1}\left|i_x[k-i]\right|^2 = i_{x,MSS}[k-1] + \frac{f_g\left|i_x[k]\right|^2}{f_s} - \frac{f_g\left|i_x[k-\frac{f_s}{f_g}+1]\right|^2}{f_s} \qquad x \in [L1, g] \tag{20}$$

where it is assumed the discrete sampling frequency f_s is an integer multiple of the grid frequency f_g. If this is not the case, then the term f_s/f_g can be rounded up to the nearest integer, with f_g/f_s replaced with the reciprocal of the rounded number. It should also be noted that f_s/f_g can be replaced with $\alpha f_s/f_g$, $\alpha \in \mathbb{N}$, while still considering practical constraints such as the controller's available RAM. Once the mean squared sum is updated, the root mean square is calculated:

$$i_{x,RMS}[k] = \sqrt{i_{MSS}[k]} \qquad x \in [L1, g] \tag{21}$$

Fig. 4 summarizes on iteration of the predictive control scheme.

IV. RESULTS AND DISCUSSION

The single-phase smart inverter using the qZSI topology is tested experimentally with the proposed control scheme, shown in Fig. 5. Table II includes several specifications of the prototype. The qZS is implemented with GB100XCP12-227 SiC IGBTs. The control algorithm is implemented on the dSPACE CP1103, with a sampling time of 20μs. The gate

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

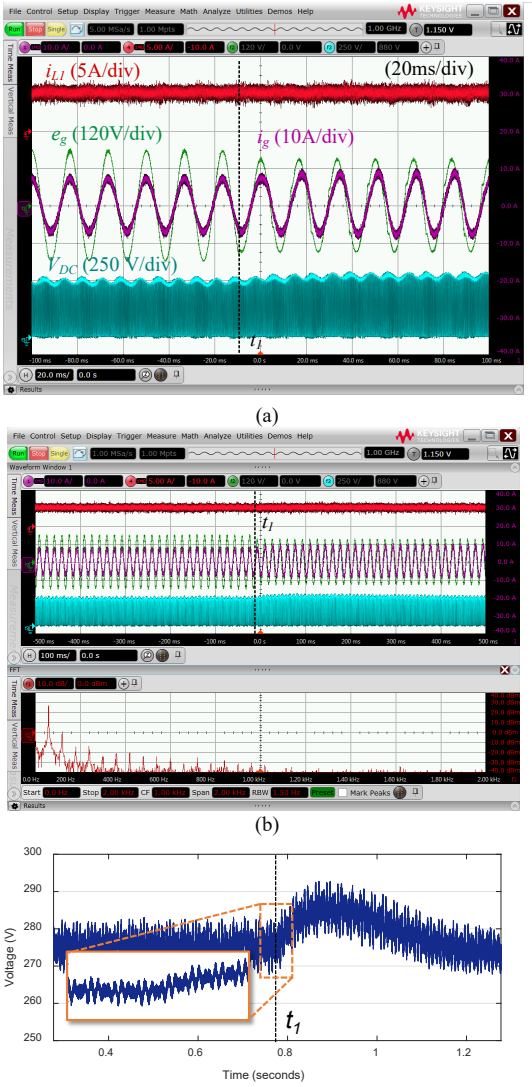


Fig. 6. Voltage sag using constant average active power strategy: (a) grid voltage and current, inductor L1 current, and dc-link voltage dynamic response, (b) extended view of dc-link voltage dynamics and FFT plot of injected grid current, (c) V_{C1} dynamic response.

signals from the CP1103 are applied directly to the qZSI gate driver from the digital I/O port, meaning no dead-time is implemented. This is generally risky, yet the impedance network prevents undesired current spikes during unintentional shoot through. The output of the qZSI is connected to a bidirectional grid emulator allowing grid voltage sags to be tested. The power supply used to emulate a 1kW PV string is the EA-PSI 9000 2U. The ramping rate γ is set to a large value of 4 p.u. to make the RPI easier to visualize on the oscilloscope.

The single-phase qZSI is interfaced with single-phase $120V_{RMS}$, 60Hz grid to verify the performance of the proposed system experimentally. The prototype is tested for the constant average active power strategy, constant active current, and constant peak current strategy. Then, the system is shown as the grid voltage recovers from a voltage sag using the constant average active power strategy.

A. LVRT Mode in Constant Average Active Power Strategy

Fig. 6 shows the system dynamics during a grid voltage sag, where the RMS grid voltage decreases from 120V to 100V at t_1. i_{L1} is shown to visualize changes in PV power harvesting. Using the constant average active power strategy, P^* is fixed at 1000W. Thus, no transient in i_{L1} is seen after t_1. The reactive power injection is easily identifiable with the grid current lagging the grid voltage. A sudden increase in grid current amplitude is noted, caused by the step-change in i_{qref}. With the reduction in grid voltage and unchanged active current amplitude, the power injection briefly drops. This initial reduction in power injection creates a swell at the DC link with the excessive energy stored on C1. Thus the PI controller begins increasing i_{dref}. The full DC link voltage is shown over an extended interval in Fig. 6b. As i_{dref} increases, the DC link voltage reaches steady-state. This settling time is dependent on the gains of the PI controller. The primary trade-off in designing the gains of the PI is settling time of the DC link voltage (and i_{dref}) during grid voltage transients and sub-harmonic content of the injected grid current. In other words, the settling time of the DC link voltage could be reduced by increasing the sampling frequency of the PI or increasing the integral gain. However, increasing the sensitivity of the PI controller would create substantial steady-state oscillation on i_{dref} and increase sub-harmonic distortion of the injected current. These oscillations are a result of voltage ripple on C1. An FFT of the injected grid current is also shown in Fig. 6b, using one second of measurements centered at t_1. Although the most substantial frequency content exists at harmonics of the grid frequency, the frequency content of the injected grid current is continuous. This is caused by the variable switching frequency inherent to finite-set MPC.

B. LVRT Mode in Constant Active Current Strategy

The grid voltage transient occurs using the constant active current strategy at t_2 in Fig. 7. Commensurate with the grid voltage sag, the power reference decreases to about 833W. This can be seen by the reduction in i_{L1} in Fig. 7a. The brief swell in DC link voltage still occurs at t_2, despite a reduction in power reference. This is expected, as the flexible power point tracking algorithm requires several milliseconds to reach steady-state. In this time, the difference in power harvesting and power injection is absorbed by C1, as discussed earlier. However, the eventual reduction in harvested power in addition to the initial increase in i_{dref} causes a brief sag in DC link voltage before settling. This is shown in Fig. 7b. By making the reduction in harvested power commensurate with the reduction in grid voltage, the active current reference is unchanged at steady-state. However, the reactive current reference will still create an increase in grid current amplitude. Note also there are distinct intervals in which shoot-through is not implemented, evidenced by V_{DC} not reaching zero volts. The MPC is driving down the input current by avoiding the

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

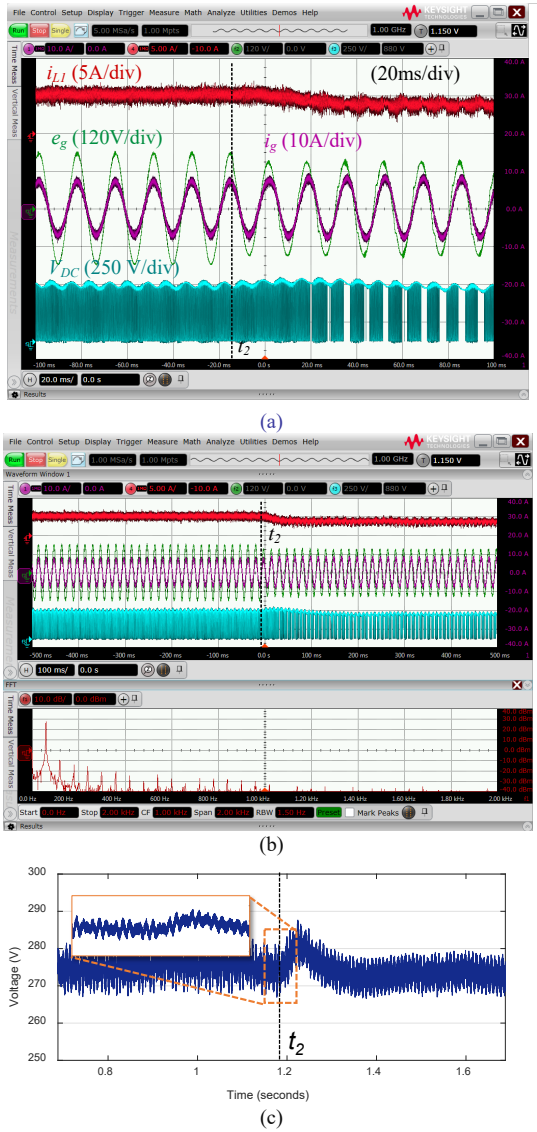


Fig. 7. Voltage sag using constant active current strategy: (a) grid voltage and current, inductor L1 current, and dc-link voltage dynamic response, (b) extended view of dc-link voltage dynamics and FFT plot of injected grid current , (c) V_{C1} dynamic response.

shoot-through state. This is intuitive when looking at the dynamic equation for v_{L1} in Table I. By placing the steady-state values of V_{PV} and V_{C1} into these equations, it is evident that shoot-through creates positive di_{L1}/dt, while the opposite is true for non-shoot through states. Thus, the shoot-through duty ratio is expected to decrease for reductions in harvested PV power. Fig. 7b also shows the FFT of the injected grid current amplitude. Although the oscilloscope does not provide a total harmonic distortion measurement to compare against the constant average active power strategy, a reduction in the subharmonic content of the injected current is observed, caused by a less substantial increase in grid current amplitude at t_2. This can be judged by observing the left-most side of the FFT plots prior to the 60Hz peak.

C. LVRT Mode in Constant Peak Current Strategy

The grid voltage sag is induced at t_3 using the constant peak current strategy, shown in Fig. 8. i_{qref} is set to 3.5A for the presented grid voltage sag and ramping rate. Based on (13), the power reference is reduced to about 795W. With the power reference close to that of the constant active current strategy, it is difficult to distinguish these two strategies. We expect a slightly more dramatic decrease in i_{L1}, in addition to a slightly larger grid voltage sag as the harvested power drops. Note that in Fig. 8b, the DC link voltage appears to settle at a peak voltage slightly below what it was before t_3. This is expected; reducing i_{L1} to reduce the harvested power creates a reduction in the shoot-through duty ratio D, as discussed previously. Additionally, a reduction in shoot-through duty ratio causes a reduction in the DC link's peak voltage, as evidenced by (3). The DC link voltage and V_{C1} have equal DC components. The steady-state reduction in D is expected to reduce the DC link voltage peak to maintain equality between the average DC link voltage and V_{C1}. Additionally, V_{C2} decreases for reductions in D, as shown in (2). Thus, the noted changes at the DC link voltage are required to regulate V_{C1} during the transient in harvested power.

Despite the RPI strategy being labelled as *constant peak current*, there is a notable increase in grid current amplitude after t_3. Although the power reference is reduced to maintain the peak current at steady state, this initial amplitude spike results from the differences in how the active and reactive current references are generated. i_{qref} is generated automatically, based on the severity of the grid voltage sag, whereas the i_{dref} relies on the cascaded settling times of the MPPT algorithm and the PI controller. This difference in time causes the initial spike in grid current amplitude. At steady-state the current amplitude is unchanged. Although the PI controller and MPPT algorithm can be altered to reduce this time, a non-zero interval of amplitude spike is inevitable, regardless of P^*. Fig. 8c shows the FFT of the grid current during this transient. The most notable difference in this FFT plot compared to those of the other two RPI strategies is the reduced low-frequency content, as a result of the less severe transient in grid current amplitude.

D. LVRT Mode to Normal Grid Condition

In Fig. 9a, the qZSI dynamics are shown for a recovery in grid voltage using the constant average active power strategy. The grid voltage increases from $100V_{RMS}$ to $120V_{RMS}$. The grid current immediately aligns with the grid voltage, as i_{qref} drops to zero. After t_4, Fig. 9b shows a brief sag in the DC link voltage. i_{dref} was raised during the grid voltage sag. As the grid voltage returned to normal condition, the increased active current injection pulled additional energy from C1, resulting in the sag. The DC link voltage returns to normal as the PI controller reaches V^*_{C1}.

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

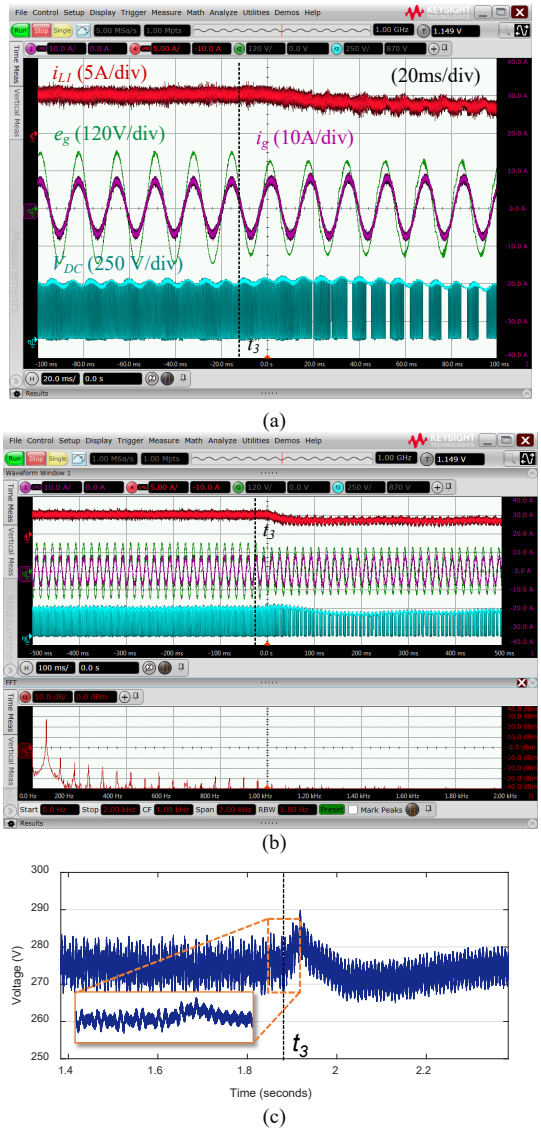


Fig. 8. Voltage sag using constant peak current strategy: (a) grid voltage and current, inductor L1 current, and dc-link voltage dynamic response, (b) extended view of dc-link voltage and FFT plot of injected grid current, (c) V_{C1} dynamic response.

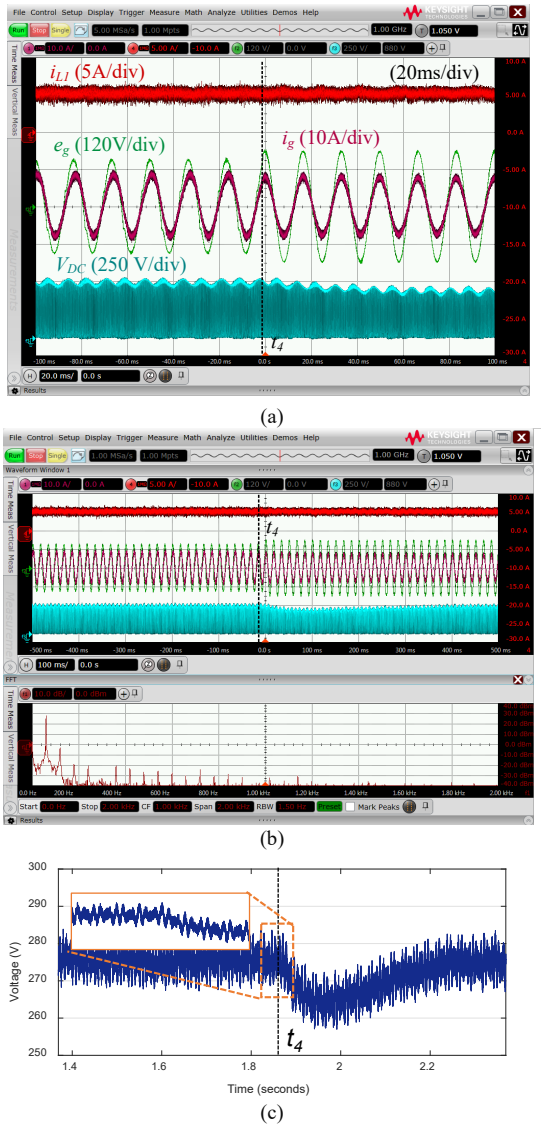


Fig. 9. Voltage recovery using constant average active power strategy: (a) grid voltage and current, inductor L1 current, and dc-link voltage dynamic response, (b) extended view of dc-link voltage and FFT plot of injected grid current, (c) V_{C1} dynamic response.

E. Solar Irradiance Transient

The qZSI response to a solar irradiance transient from 1000W/m^2 to 400W/m^2 at t_5 is shown in Fig. 10. A substantial dip in V_{PV} occurs after t_5, shown in Fig. 10a. This dip occurs from the rapid change in the I-V characteristic curve. Specifically, while the MPPT is adjusting its input current reference, the PV array is operating at its short-circuit current. Thus, this dip is proportional to the settling time of the MPPT algorithm. After dropping around 100V, it settles slightly below 200V at the new MPP. In Fig. 10b, we see the injected current drop substantially following t_5, dropping close to zero before settling around a 2.5 peak amplitude. This can be explained when seeing the dip in V_{C1} in Fig. 10c, as the energy injected to the AC side is not matching that which is being pulled from the PV array, thus energy is rapidly drawn from C1. The PI control responds by reducing the active current reference near to zero, until power can once again be harvested from the DC side. Once i_{L1} settles at the new MPP, power harvesting from the PV array continues, and V_{C1} begins to rise accordingly. In Fig. 10d, the weight factors λ_1 and λ_2, which respectively weight the input and output current objectives, have settled at a ratio near 2:3, until t_5. Once the solar irradiance transient occurs, the input current reference quickly drops from the MPPT, but is not immediately tracked, causing large spikes in λ_1. This causes i_{L1} to dominate the cost function J, allowing it to hold precedence for this transient. As the grid current decreases in magnitude, its normalized factor, determined by the rolling RMS from (20), is not immediately adjusted, causing the normalized tracking error to decrease.

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

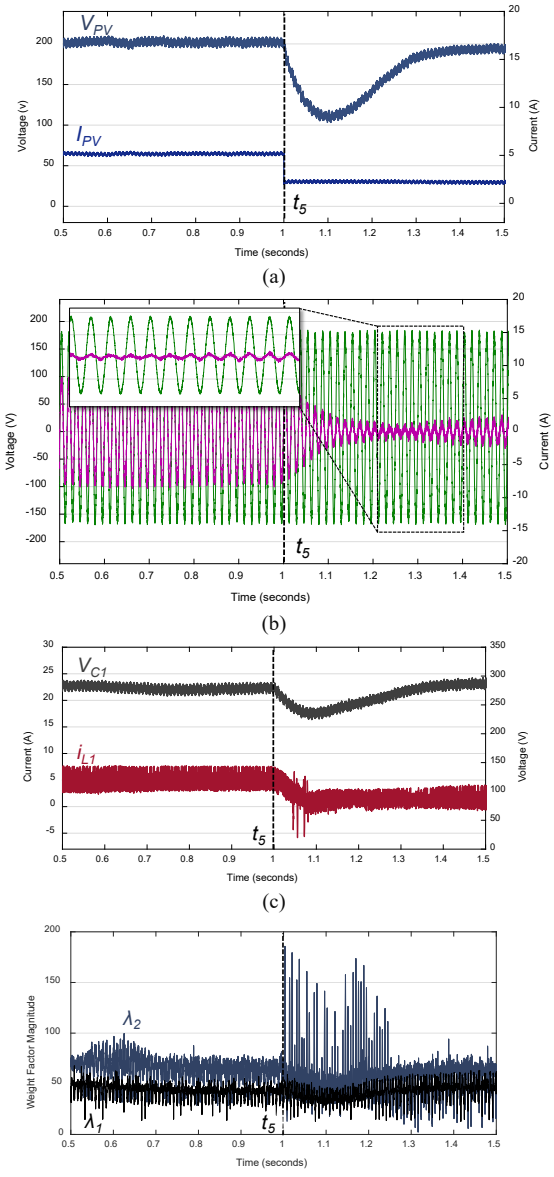


Fig. 10. Solar irradiance transient from 1000W/m^2 to 400W/m^2: (a) PV voltage and current, (b) grid voltage and current, (c) V_{C1} and i_{L1} inductor L1 current, and dc-link voltage dynamic response, (d) weight factor adaptation.

The weight factor λ_2 decreases accordingly, and both weight factors reach steady-state as their normalizing factors are adjusted according to the new magnitudes of i_{L1} and i_g.

F. Comparison with Static Weight Factors

Finally, the tracking performance of the auto-tuned weight factors is compared with the static weight factors implemented in [30]. Observing metrics such as transient performance (i.e. settling time during transient conditions) of the two control schemes does not directly reveal the performance of the MPC itself, only of the entire system, as transient performance is highly dependent on the reference values, which are merely inputs to the predictive control. Rather, the exact tracking performance is compared for both systems. To conduct this comparison, a term referred to as *tracking error* is defined as:

$$x_\varepsilon[k] = \left| x^*[k-1] - x[k] \right| \qquad x \in [i_{L1}, i_g] \tag{22}$$

Objective tracking error is decidedly the most apt metric, as this value demonstrates how close the controlled variables reached to their references in the next sampling period. To compare tracking errors of each control objective, the difference in tracking errors is shown for the two control schemes. For clarity, the measurements in Fig. 11 are defined as:

$$x_\Delta[k] = x_{\varepsilon,auto}[k] - x_{\varepsilon,static}[k] \qquad x \in [i_{L1}, i_g] \tag{23}$$

where $x_{\varepsilon,auto}$ and $x_{\varepsilon,static}$ are the tracking errors for the proposed control and the control of the static weight factor ratio implemented in [30]. Thus, a negative x_Δ would imply better tracking by the proposed, autonomously-tuned control. In Fig. 11, x_Δ is shown under normal condition at full solar irradiance, and for a step-change in the flexible power point tracking reference from 1kW to 0.5kW which occurs at t_6. While the magnitude is larger for $i_{L1,\Delta}$ than $i_{g,\Delta}$, both are negative on average. The results reveal better overall tracking for both objectives using the proposed control. After t_6, the difference decreases slightly for both values. This is due to a decrease in magnitude of both control objectives with the reduction in power harvesting from the PV array. Precisely, the average tracking error of the input current was reduced by nearly 10%, while the average output current tracking error was reduced by about 4% which shows significant improvement in tracking performance. The weight factor ratios for both control schemes are shown in Fig. 12. Like the autonomous weight factor tuning feature itself, the reason for better tracking is intuitive. The proposed control weights each objective according to normalized anticipated tracking error. Thus, an objective which is poorly tracked is quickly compensated via greater weight in the cost function. It is also worthwhile to mention that reaching to the weight factor in [30] involved substantial testing, whereas such a design stage is not required for the proposed control.

V. Conclusion

A smart inverter is actualized using the proposed autonomous model predictive control (AMPC) scheme, and is shown to seamlessly transition between low voltage ride through (LVRT) and maximum power point tracking (MPPT) modes according to the grid condition. The controller is able to retain stability during large transients in solar irradiance. Finally, when comparing to a carefully designed, yet static, weight factor ratio in [30], the auto-tuned weight factors exhibited consistently better tracking performance. Additionally, this auto-tuned weight factor in the cost function eliminates the trial-and-error design stage in conventional finite-set MPC. Experiments are provided to verify the system performance for three reactive power injection strategies over LVRT mode, while the reactive power injected adheres to recent grid codes such as the E. ON standard for the chosen reactive power injection (RPI) strategy. There remain areas for further discussion and analyses as future work. First, practical efficiency of the power stage creates a substantial disparity between the PV power reference and the actual power injected to the grid. The power reference of the flexible power point

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

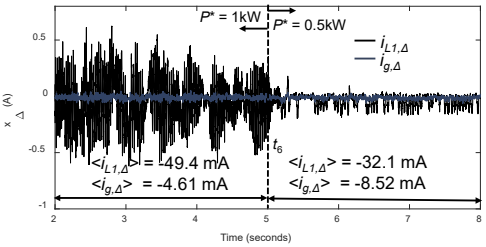


Fig. 11. Tracking error comparison of the proposed auto-tuned weight factors and the static weight factor ratio from [30].

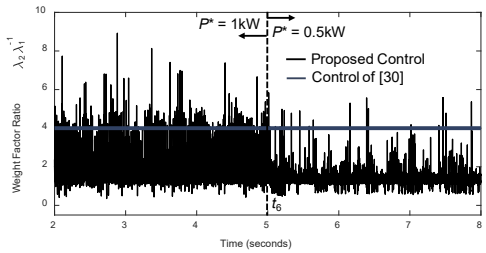


Fig. 12. Ratio of the proposed auto-tuned weight factors and the static weight factor ratio from [30].

tracking algorithm should be chosen to account for inefficiencies in the power conversion stage. Second, a brief increase in current amplitude was seen for the constant peak current RPI strategy. This was caused by the non-zero time to reduce the harvested PV power. The reactive current reference should be adaptive to align with the DC-side dynamics to avoid initial current spikes, rather than using a simple lookup table.

ACKNOWLEDGEMENT

This publication was made possible by grant # NPRP12S-0226-190158 from the Qatar National Research Fund (a member of the Qatar Foundation). The statement made herein are solely the responsibility of the authors.

REFERENCES

[1] Y. Yang, *et al*., "Wide-scale adoption of photovoltaic energy: Grid code modifications are explored in the distribution grid," *IEEE Industry Applications Magazine,* vol. 21, no. 5, pp. 21-31, 2015.

[2] X. Liang, "Emerging power quality challenges due to integration of renewable energy sources," *IEEE Transactions on Industry Applications,* vol. 53, no. 2, pp. 855-866, 2016.

[3] A. Sangwongwanich, *et al*. "Benchmarking of constant power generation strategies for single-phase grid-connected photovoltaic systems," *IEEE Transactions on Industry Applications,* vol. 54, no. 1, pp. 447-457, 2017.

[4] E. On, "Grid code high and extra high voltage," *E. On Netz Gmbh, Bayreuth, Germany,* 2006.

[5] S. Mukherjee, V. R. Chowdhury, P. Shamsi, and M. Ferdowsi, "Power-Angle Synchronization for Grid-Connected Converter With Fault Ride-Through Capability for Low-Voltage Grids," *IEEE Transactions on Energy Conversion,* vol. 33, no. 3, pp. 970-979, 2018.

[6] M. Dietmannsberger, F. Grumm, and D. Schulz, "Simultaneous implementation of LVRT capability and anti-islanding detection in three-phase inverters connected to low-voltage grids," *IEEE Transactions on Energy Conversion,* vol. 32, no. 2, pp. 505-515, 2017.

[7] Y. Liu, B. Ge, H. Abu-Rub, and H. Sun, "Hybrid pulsewidth modulated single-phase quasi-Z-source grid-tie photovoltaic power system," *IEEE Trans. Ind. Informat.,* vol. 12, no. 2, pp. 621-632, 2016.

[8] S. B. Kjaer, J. K. Pedersen, and F. Blaabjerg, "A review of single-phase grid-connected inverters for photovoltaic modules," *IEEE Trans. Ind. Appl.,* vol. 41, no. 5, pp. 1292-1306, 2005.

[9] D. Cao, S. Jiang, X. Yu, and F. Z. Peng, "Low-cost semi-Z-source inverter for single-phase photovoltaic systems," *IEEE Transactions on Power electronics,* vol. 26, no. 12, pp. 3514-3523, 2011.

[10] Y. Huang, M. Shen, F. Z. Peng, and J. Wang, "Z-Source Inverter for Residential Photovoltaic Systems," *IEEE transactions on power electronics,* vol. 21, no. 6, pp. 1776-1782, 2006.

[11] Y. Tang, J. Wei, and S. Xie, "Grid-tied photovoltaic system with series Z-source inverter," *IET Renewable Power Generation,* vol. 7, no. 3, pp. 275-283, 2013.

[12] Y. Liu, *et al*. "Z-Source/Quasi-Z-Source Inverters: Derived Networks, Modulations, Controls, and Emerging Applications to Photovoltaic Conversion," *IEEE Ind. Electron. Mag.,* vol. 8, no. 4, pp. 32-44, 2014.

[13] J. Anderson and F. Z. Peng, "Four quasi-Z-source inverters," in *39th Annual IEEE PESC*, Rhodes, Greece, pp. 2743-2749, 2008.

[14] Y. Liu, *et al.*, *Impedance source power electronic converters*. John Wiley & Sons, 2016.

[15] Y. Li, S. Jiang, J. G. Cintron-Rivera, and F. Z. Peng, "Modeling and control of quasi-Z-source inverter for distributed generation applications," *IEEE Transactions on Industrial Electronics,* vol. 60, no. 4, pp. 1532-1541, 2012.

[16] S. Vazquez, J. Rodriguez, M. Rivera, L. G. Franquelo, and M. Norambuena, "Model predictive control for power converters and drives: Advances and trends," *IEEE Transactions on Industrial Electronics,* vol. 64, no. 2, pp. 935-947, 2017.

[17] C. Garcia, *et al.* "Full predictive cascaded speed and current control of an induction machine," *IEEE Transactions on Energy Conversion,* vol. 31, no. 3, pp. 1059-1067, 2016.

[18] T. Jin, X. Shen, T. Su, and R. C. Flesch, "Model Predictive Voltage Control Based on Finite Control Set with Computation Time Delay Compensation for PV Systems," *IEEE Transactions on Energy Conversion,* vol. 34, no. 1, pp. 330-338, 2018.

[19] ABB, "ABB's Model Predictive Torque Control (MPTC) protects against gas supply interruptions," pp. [Online]. Available: https://new.abb.com/drives/media/abbs-model-predictive-torque-control-mptc-protects-against-gas-supply-interruptions, 2016.

[20] P. Cortés *et al.*, "Guidelines for weighting factors design in model predictive control of power converters and drives," in *IEEE ICIT*, Churchill, Victoria, Australia, pp. 1-7, 2009.

[21] M. Mosa, R. S. Balog, and H. Abu-Rub, "High-performance predictive control of quasi-impedance source inverter," *IEEE Transactions on Power Electronics,* vol. 32, no. 4, pp. 3251-3262, 2017.

[22] S. Jain, *et al.* "Distributed Predictive Control Scheme for Grid-Tied Cascaded Multilevel Impedance Source Inverter with LVRT Capability," in *IEEE ECCE*, 2018, pp. 6236-6242.

[23] S. Sajadian and R. Ahmadi, "ZSI for PV systems with LVRT capability," *IET Renewable Power Generation,* vol. 12, no. 11, pp. 1286-1294, 2018.

[24] S. Bayhan, H. Abu-Rub, and R. S. Balog, "Model predictive control of quasi-z-source four-leg inverter," *IEEE Transactions on Industrial Electronics,* vol. 63, no. 7, pp. 4506-4516, 2016.

[25] A. Bakeer, *et al*. "A powerful finite control set-model predictive control algorithm for quasi Z-source inverter," *IEEE Transactions on Industrial Informatics,* vol. 12, no. 4, pp. 1371-1379, 2016.

[26] A. Lashab, D. Sera, and J. M. Guerrero, "A low-computational high-performance model predictive control of single phase battery assisted Quasi Z-source PV inverters," in *2019 10th International Conference on Power Electronics and ECCE Asia*, 2019, pp. 1873-1878.

[27] A. Lashab, D. Sera, and J. M. Guerrero, "Model Predictive Control of Cascaded Multilevel Battery Assisted Quasi Z-Source PV Inverter with Reduced Computational Effort," in *2019 IEEE Energy Conversion Congress and Exposition (ECCE)*, 2019, pp. 6501-6507: IEEE.

[28] Y. Liu, *et al*. "A discrete-time average model-based predictive control for a quasi-Z-source inverter," *IEEE Transactions on Industrial Electronics,* vol. 65, no. 8, pp. 6044-6054, 2017.

[29] Y. Yang, *et al*, "Reactive power injection strategies for single-phase photovoltaic systems considering grid requirements," *IEEE Trans. Ind. Appl.,* vol. 50, no. 6, pp. 4065-4076, 2014.

[30] S. Jain, M. Easley, M. B. Shadmand, and B. Mirafzal, "Decoupled active and reactive power predictive control of impedance source microinverter with LVRT capability," in *2018 IEEE Power and Energy Conference at Illinois (PECI)*, 2018, pp. 1-6: IEEE.

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TEC.2020.2998501, IEEE Transactions on Energy Conversion

[31] M. Ciobotaru, R. Teodorescu, and F. Blaabjerg, "A new single-phase PLL structure based on second order generalized integrator," in *37th IEEE PESC*, Jeju, South Korea, pp. 1-6, 2006.
[32] J. Rodriguez and P. Cortés, "Weighting Factor Design," in *Predictive Control of Power Converters and Electrical Drives*Chichester, UK: John Wiley & Sons, Ltd., 2012, pp. 163-168.



Mitchell Easley (S'18) received the B.S. degree in electrical engineering from Kansas State University, Manhattan, KS, in 2018, where he is currently pursuing the M.S. degree with the Department of Electrical and Computer Engineering.

Since 2017, he has been a Research Associate with the Power Electronics and Autonomous Systems Research Laboratory, Kansas State University. His current research interests include model predictive control, wide bandgap-based power converters, and multilevel converters.



Sarthak Jain received the B.Tech. degree in electrical engineering from GGSIPU, Delhi, India, in 2014 and M.S. Degree in electrical engineering from Texas A&M University, College Station, TX in 2018.

He is currently an Application Engineer in Core Power Solutions Group, Renesas Electronics America, Milpitas, CA. He was a Research Associate at Delhi Technological University, Delhi, from 2014 to 2016, and at the Renewable Energy and Advanced Power Electronics Laboratory, College Station, from 2016 to 2017. His current research interests include PV grid tied inverters and dc/dc Point-of-load converters, resonant converters, wide band gap power devices.



Mohammad B. Shadmand (S'09–M'15) received the Ph.D. degree in electrical engineering from Texas A&M University, College Station, TX, USA, in 2015. From 2015 to 2016, he was an Instructor with the Department of Electrical and Computer Engineering, Texas A&M University. From 2016 to 2017, he was a Research Engineer with the Renewable Energy and Advanced Power Electronics Research Laboratory, College Station, TX, USA. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering and the Co-Director of the Power Electronics and Autonomous Systems (PEAS) Research Laboratory, Kansas State University, Manhattan, KS, USA. He has published more than 80 journal and conference papers. His current research interests include advanced model predictive control, grid-tied power electronics interfaces with advance functionalities, and control of smart microgrid systems.

Dr. Shadmand was awarded Michelle Munson Serban Simu Keystone Research Scholar, Kansas State University in 2017. He was awarded the 2019 IEEE Myron Zucker Faculty-Student Research Grant. He has served as Technical Program Co-Chair of the 2019 IEEE Smart Grid & Renewable Energy Conference. He serves as Associate Editor of IEEE Transactions on Industry Application and IET Renewable Power Generation.



Haitham Abu-Rub (M'99, SM'07, F'18) is a full professor holding two PhDs. He has worked at many universities in many countries including Poland, Palestine, USA, Germany, and Qatar. Since 2006, Dr. Abu-Rub has been with Texas A&M University at Qatar. Abu-Rub has served for five years as the chair of Electrical and Computer Engineering Program at the same university and currently has been serving as the managing director of Smart Grid Center. His main research interests include electric drives, power electronic converters, renewable energy and smart grid. Dr. Abu-Rub is the recipient of many national and international awards and recognitions. He is the recipient of the American Fulbright Scholarship, the German Alexander von Humboldt Fellowship, and many others. Abu-Rub has published more than 400 journal and conference papers, five books and six book chapters. He has supervised many research projects on smart grid, power electronics converters, and renewable energy systems.

0885-8969 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 08,2020 at 09:34:14 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Comparative Analysis of Deep Convolutional Generative Adversarial Network and Conditional Generative Adversarial Network using Hand Written Digits

Prabhat
Department of Information Technology
Delhi Technological University,
New Delhi, India
prabhat.saroha123@gmail.com

Nishant
Department of Information Technology
Delhi Technological University,
New Delhi, India
nishantnishu444@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Delhi Technological University,
New Delhi, India
dvishwakarma@gmail.com

***Abstract*-- Generative adversarial network framework has recently emerged as a promising generative modeling approach. It is composed or made up of a generative network and a discriminative network. There have been various types of adversarial networks present today. Among these types of networks, the most popular network is Deep Convolutional Generative Adversarial Network (DCGAN) for performing on the convolutional networks without using multilayer perceptrons. The multilayer layer perceptrons have the hidden layers because of this we have to more bind to extract the data with the parameters, because of this we also study the Conditional Generative Adversarial Network (CGAN) to add an extra label to the generator and the discriminator. We study the comparative analysis between these two popular networks to highlight the main differences and similarities using the handwritten image datasets.**

Keywords— Generative adversarial network (GAN), Deep convolutional generative adversarial network (DCGAN), Conditional generative adversarial network (CGAN).

I. INTRODUCTION

The Deep Learning method has to change in the day to day life. There is much process of parts of deep learning methods, neural networks. The faking document makes a higher risk in the part of this purposiveness it will not perform well if the idea of the faking mechanism is enhanced. The generative adversarial network performs on this regarding how to construct the well define algorithm in which the idea of the faking purposiveness is minimum. The deep convolutional network performs well on this idea how to generate the epochs in the middle of the data on images. It works on model distribution and performs on the data distribution. The conditional adversarial network makes the data on how to perform on the given labels on the generator and discriminator and hence the labels will perform on how to generate the data or discriminate the data on given parameters. MNIST handwritten digits make a well comparative analysis of how the DCGAN and CGAN perform on the same dataset. Now basically the TensorFlow is used with DCGAN and CGAN for the MNIST and it will make a better enhancement to perform on this data for the comparative analysis between these GAN's.

II. RELATED THEORY

The generative adversarial network makes the highly computable neural network, which performs on the two neural network models. These network models perform to each other to generate the dataset on bases of the training model of the generator and another network performs to classify or capture the data within the dataset. There has been a chance that most of the neural model network failed because when there is adding some small amount of noise in the data then there the whole network will be misclassifying the original data from the fake data hence it will predict correctly for wrong data [1]. Because of the existence of a most model of the network or the machine have only limited data set input hence there is no such adversarial network to classify our model of a large dataset. The generative adversarial network can take large datasets and classify the inputs through the neural networks. The DCGAN performance on the different datasets is well because it generates less fuzzy pixels and hence it will be used for complex data sets for generating purposes for human faces and more realistic images. While for the conditional GAN's it uses for the condition base networks for generation of the data [2]. For the face aging concept, it's used for the age synthesizing to evolve the face generation for particular criteria of the images [14]. The GAN network is a high performance and the whole concept of GAN works on unsupervised learning i.e., its work on non labeled data and that learn from its output, in this, we group the data according to similarities or a particular behavior. In some cases, this training network also works for supervised learning [15].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:39:13 UTC from IEEE Xplore. Restrictions apply.

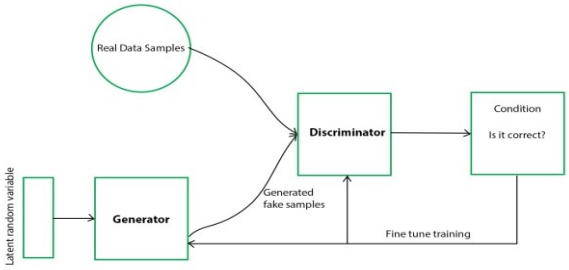


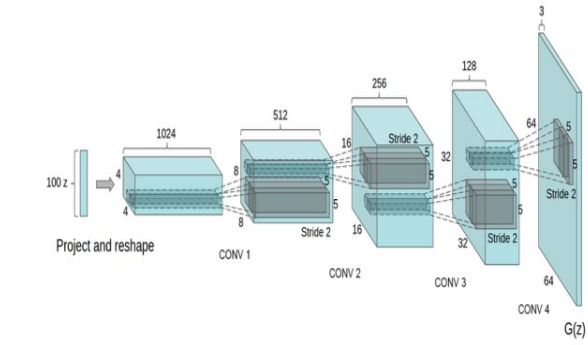
Fig 1. Generative adversarial network model [6]

III. MNIST DATASET

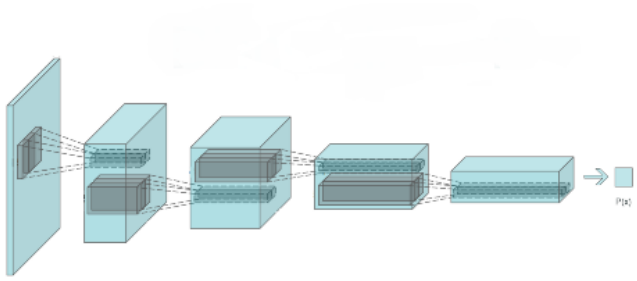
This dataset of handwritten digits consists of a 60k training sample set of buffer size having batch size 256 to compute on the shape of images of size 28x28x1. The data set has the centralized and separate with the square box of the images of the handwritten digits. MNIST has a large 10k testing dataset also. For the training of the generator we trained with the 50 epochs through the generator and hence it will generate the random noise for the discriminator to discriminate the handwritten digits from fake images [10].

IV. DCGAN

The deep convolutional network performs on ConvNets. It's composed of ConvNets in place of multi-layer perceptrons. The ConvNets square measure enforced while not pooling, which is replaced by convolutional stride. Also, the layers don't seem to be connected. The generator has the power to generate random noises to fool the discriminator that these noises have very much similar to the real dataset. In the DCGAN the deep network generates the dataset for discriminator which using the help to the process of generation. Because the layers are disconnected it tends to recreate the noises on its own. There have been many types of outputs such binary for real or fake (1 or 0), sometimes it is real-valued numbers in the form regression and there may be multiple categorical outputs such as MNIST or CIFAR-100. In the diagram, we have 100 input *z* as a Dimensional Vector for the deep network to generate or train the model in which we select the input randomly as *x1,x2,x3...x100* and generate the output in images to perform the network for Z_{out} by using the multilayer transpose of convolution. This Z_{out} is just random images [12]. The DCGAN has the second network which is called the discriminator to process how images are generated. The Generator has a training dataset that is used to generate the images and train the model of the discriminator to perform on the images to classify the images is real or fake. With the improvement of the generator to the training dataset, it will generate more realistic images. The discriminator has two purposes to use, first it determines to distinguish between the real images and fake images and give the output to the generator to create more random noises or realistic images. Both the network to perform simultaneously to improve the quality of the model of two CNN networks [5]. The Generator is to generate the images to decrease the gap between the real images and fake images while the discriminator to differentiate for better real and fake images.



(a)



(b)

Fig.2. (a) Generator model (b) Discriminator Model in DCGAN [5]

V. CGAN

The conditional generative adversarial network-based upon supervised learning. In this network, we feed the information as auxiliary as *y* to constitute the idea of a condition based approach to the discriminator and generator networks. The input noise *pz(z)* and *y* combined to make the joint hidden representation in which the *x* and *y* are the discriminative functions [3]. The network consists of a class label and data from other modalities. Discriminator and Generator both have the multilayer perceptrons to Rectified Linear Unit (ReLU) with activation function for hidden layers and sigmoid function for output layers. The Cost Function of CGAN is defined using Eq.1 [4].

$$\underbrace{min}_{G}\ \underbrace{max}_{D}\ V(D,G) = E_{x\sim pdata(x)}[\log D(\frac{x}{y})] + E_{z\sim pz(z)}\left[\log\left(1 - D(G(\frac{z}{y}))\right)\right] \quad (1)$$

Both Function *G* and *D* are non – linear mapping function. The multilayer perceptrons have used for network architecture, its make special because the extra condition has to add to the network. There is no control of the data to generated and hence also the labels to added to the discriminator to find the fake images. In the cost function *D(x|y)* and *G(z|y)* describe that we are generating or discriminate the images with the class label.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:39:13 UTC from IEEE Xplore. Restrictions apply.

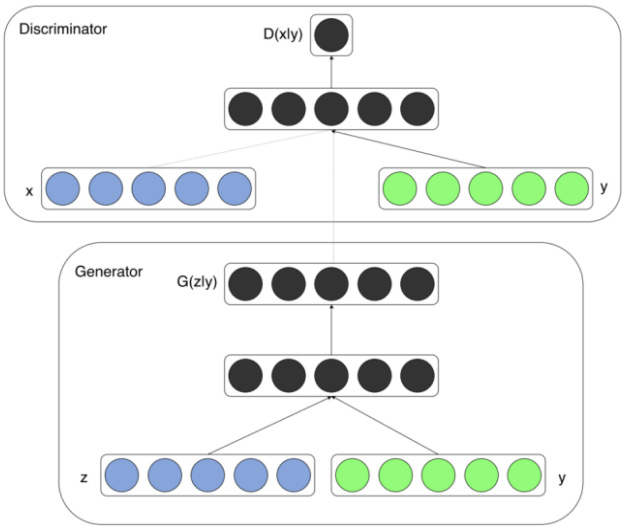


Fig. 3. The model architecture of CGAN [4]

The CGAN network makes the labeling adding in both the generator and discriminator to provide the corresponding images are generated and distinguish the images from real from fake images respectively. It makes the ideal with the multi-modal network to provide the input and output for mapping in the contextual information [4]. In Fig. 4, we have the conditional generative adversarial network in which the label act as an extension to the latent space z *to* generate and discriminate the images [2].

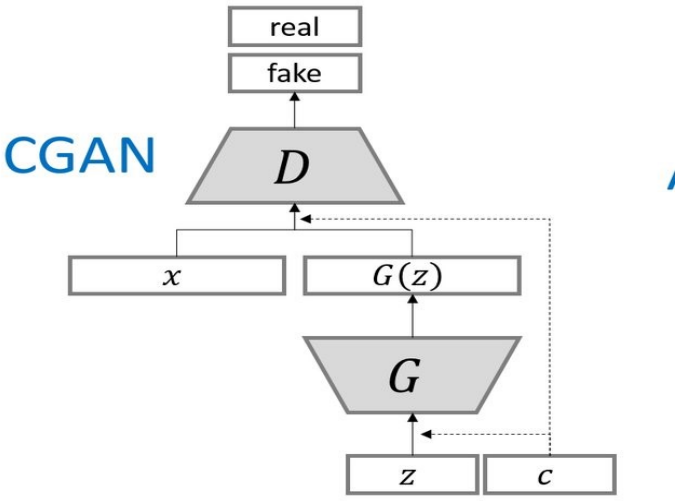


Fig. 4. Data flow of CGAN where *x* denotes the real images, *c is* a class label, *z* is the noise vector and *G* is the Generator, and *D* is the Discriminator [11].

VI. EXPERIMENTS

For the DCGAN we used the MNIST handwritten dataset for the generator and discriminator to train the model. The generator to use generates the handwritten digits to excise the MNIST dataset. The generator model for this purpose to train the 50 epochs. The discriminator to find the fake generation of images and hence simultaneously to train the model itself for the images. The generator model resembling itself to generate the images highly original. The generator used the Keras model for the generate the random noises of images to take upsample to desired the images of size 28x28x1.

Here is the loss function that is applied to the network for both generator and discriminator.

Discriminator loss function: It is used to measure how the discriminator distinguishes the fake images from real images. The discriminator update by ascending order of its stochastic function. The Discriminator loss [8] [9] is the sum of losses of the all real and fake batches and can be written as

$\log(D(x)) + \log(D(G(z)))$.

Generator loss function: This function is used to able to make complex for the discriminator. The generator has to obsess to generate the highly original images. Thus the loss function is used for the generator loss as [8] [9]

$\log(D(G(z)))$.

Here *D(G(z))* is the average discriminator output for all fake batches. Here the first number before *D* is updated as well as after *D* also updated. This number is started from 0 and converges to 0.5 as *G* gets better.

And *D(x)* is the average output of the discriminator or all real batches. This will be started from 1 and converges to 0.5.

TABLE I. PERFORMANCE ANALYSIS OF DCGAN

Criteria	**Analysis**	
Fuzziness	Less fuzzy	
Time	More time to train the model	
Testing	Only for real and fake images	
The loss function for discriminator	If it takes the real images it shows the higher value(close to 1) And if take fake images it shows the lower value(close to 0)	
Optimizers	One for each generator and discriminator	
Accuracy Vs 50 Epochs	For Real	For Fake
	0.553540	**0.693160**
Loss Vs 50 Epochs	Discriminator	Generator
	1.284056	**0.839429**



Fig. 5. MNIST dataset Epochs Generated [10]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:39:13 UTC from IEEE Xplore. Restrictions apply.

For the CGAN we also use the MNIST dataset of handwritten images of digits (0-9). In the CGAN we find defined the explicit *y* as the class labels for extra input in the encoder and the decoder. For the placeholder *x, y, z* we use a 1-hot vector for the label (0-9) as *y*. It is based on the unimodal. The CGAN has the conditional theory in which the data will be classified on the bases of feed information *y*. We have to merge the input for the generator and discriminator for the dataset of variables of *G (z, y)* and *D (X, y)* respectively.

TABLE II. PERFORMANCE ANALYSIS OF CGAN

Criteria	**Analysis**	
Fuzziness	Fuzzier	
Time	Same as for GAN	
Indexing	Consist one-hot vector for conditional variables for index	
Generator input	G(z, y)	
Discriminator input	D(X, y)	
Accuracy Vs 50 Epochs	For Real	For Fake
	0.688542	**0.523410**
Loss Vs 50 Epochs	Discriminator	Generator
	0.755254	**1.115031**

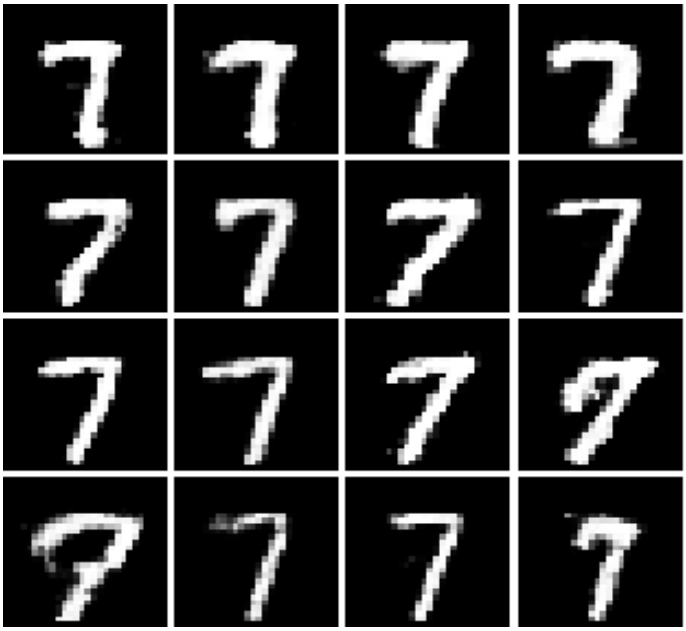


Fig. 6. Here is the '7' Digit generated Model [12]

VII. CONCLUSION

In these experiments on the image dataset, we find the DCGAN has the constraints based on the neural network and hence the discriminator used in this GAN is highly appreciable. While the CGAN has the supervised learning process in which the dataset has to feed *y* as class labels to use the generator to generate the highly rated images. In DCGAN case it is based on a convolutional neural network with constraints, while CGAN is used a multilayer perceptron structure. In DCGAN the *D* and *G* have to learn to hierarchy to scene the objects, while the CGAN has to maximize the function of *G* and minimize the function *D* [13]. In the above, the performance analysis of both DCGAN and CGAN concludes that in DCGAN the pixel is more clear for the images while in CGAN is the same as for GAN while if the GPU is not strong it will take for time for to trained the model for DCGAN. Optimization is compulsory for DCGAN because the losses in the generator and discriminator are available. Because of the additional parameters is added in the CGAN we have to merge the labels in the discriminator and generator function.

REFERENCES

[1] I. Goodfellow, P. Jean, M. Mehdi, X. Bing, W. David, O. Sherjil, C. Aaron, and B. Yoshua, "Generative adversarial nets." *In Advances in neural information processing systems*, pp. 2672-2680. 2014.

[2] P. Isola, Z. Jun-Yan, Z. Tinghui, and F. Alexei "Image-to-image translation with conditional adversarial networks." In Proceedings of the *IEEE conference on computer vision and pattern recognition,* pp. 1125-1134. 2017.

[3] H. Saifuddin "Comparative Study on Generative Adversarial Networks," in *arXiv:1801.04271v1 [cs.LG]*, 12 Jan, 2018.

[4] M. Mirza, O. Simon "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).

[5] A. Radford, M. Luke, and C. Soumith "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434 (2015).*

[6] R. Roy, "Generative Adversarial Network (GAN)," [Online]. Available: https://www.geeksforgeeks.org/generative-adversarial-network-gan/.

[7] M. Mehdi, O. Simon, "Conditional Generative Adversarial Nets," in *arXiv:1411.1784v1 [cs.LG]*, Montreal, QC H3C 3J7 ´ , San Francisco, CA 94103, 06 Nov , 2014.

[8] J. Gui, S. Zhenan, W. Yonggang, T. Dacheng, and Y. Jieping "A review on generative adversarial networks: Algorithms, theory, and applications." *arXiv preprint arXiv:2001.06937*, pp. 1 - 8, 20 Jan 2020.

[9] N. Inkawhich, "PyTorch," [Online]. Available: https://pytorch.org/tutorials/beginner/dcgan_faces_tutorial.html

[10] Tensorflow, "tensorflow.org," tensorflow, [Online]. Available: https://www.tensorflow.org/tutorials/generative/dcgan.

[11] A. Mino, and S. Gerasimos "LoGAN: Generating logos with a generative adversarial neural network conditioned on color." *In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 965-970, Netherland, 23 Oct 2018.

[12] J. Hui, "Jonathan Hui blog," Github, 05 march 2017. [Online]. Available: https://jhui.github.io/2017/03/05/Generative-adversarial-models/.

[13] J. Patel, S. Pandya, V. Shah "Review on generative adversarial networks" *International Journal of Technical Innovation in Modern Engineering*. vol . 4, no. 7. pp. 1230-1235, 2018

[14] G. Antipov, M. Baccouche, J.L. Dugelay "Face aging with conditional generative adversarial networks" *In IEEE*, 2017

[15] V. Premachandran and A.L. Yuille "Unsupervised Learning using Generative advasarial training and Clustring." *In Proceedings of the ICLR*,2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:39:13 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Controlling Mouse Motions Using Eye Tracking Using Computer Vision

Kshitij Meena
dept. of computer science and engineering
Delhi technological university
Delhi, *India*
kshitij.meena03@gmail.com

Manish Kumar
dept. of computer science and engineering
Delhi technological university
Delhi, India
mk131198@gmail.com

Mohit Jangra
dept. of computer science and engineering
Delhi technological university
Delhi, India
99mohitjangra@gmail.com

***Abstract*— This paper introduces an algorithm to carry out the functions of a mouse by providing a hands-free interaction between humans and computers. It provides an alternative to the traditional mouse computer. By using different expressions of a face using computer vision and matching it with already stored expression and execute actions as per the move. This algorithm will help physically disabled people to perform the functions of a mouse using their face and eye movement. It allows them to left-click, right-click, scroll up and down, move the cursor up, down, left, right. The system has a very basic need like webcam, NumPy, dlib, and a few other basic libraries.**

Keywords—computer vision, human-computer interface;

I. INTRODUCTION

The advent of computers in 1938 has changed the way we live our lives today at ease. They have gone through rigorous changes. Initially, their task was to solve mathematical problems and word processing now they have traveled a long way to making a necessary space in our lives with its uses ranging from professional to personal such as browsing on the internet, connecting the world through social media, controlling large machines in factories using the right and timely operations. Computers are readily usable to persons who have hands to control mouse and keyboard and eyes to see what is on the monitor. Now even blind can use the computer using the text to speech option which pronounces the content on the screen. But people who do not have hands cannot use it due to their inability to control mouse and keyword operations. In today's competitive environment everybody needs to be equipped with different sets of skills so as get a job. Knowledge and usage of computers being a necessity nowadays. So this causes them to live a miserable life. They have to beg and do pitiful activities to keep their lives going. By introducing them to technology and make them computer compatible will create a ray of hope for them to learn and do some work to their livelihood. Several efforts are made by professionals to assist the disabled in providing them a tool to interact with the computers using signals such as electroencephalography (EEG) from the brain, facial muscle signals (EMG), and electrooculogram (EOG). Also, methods like limbus, pupil, and eye/eyelid tracking, contact lens method, corneal, pupil reflection relationship have also been implemented. These methods require the use of attachments and electrodes means some physical connection is needed to the head, which makes them impractical. The method described in this paper does not use any of this expensive equipment. Also, even physical connection has to be made with the user through any equipment. It just needs a computer with a webcam making it easy and cheap and uncomplicated method. By taking real-time images of the user the program is designed to evaluate these images and identify the type of action to be performed by taking a match against the previously stored expressions to identify the operation to be performed and take parameter if needed in case of moving the cursor the change in the relative position of eyes concerning the central position tells about the direction to move the cursor. The implementation of this method would open a wide range of options for people who don't have hands. It will enable them to go to school and learn through the computer with the mouse being controlled by their head movement. It would also give them job opportunities in the service to earn them a job.

II. METHODOLOGY

The algorithm starts to detect a face of a person using a face detection algorithm. After detecting the face, the position of eyes and mouth is captured for controlling the mouse functions such as left-click, right-click, and cursor movement. This method doesn't require any special hardware and sensors. It is a hands-free system that is useful for handicapped. First, we have to activate the mouse control mode by opening the mouth. For moving the mouse cursor left or right we simply have to move our head in the intended direction of movement of a cursor. For enabling the scroll mode, we have to squint our eyes into the camera and on initiation, it will show scroll mode is on. Now you can move your head up for scrolling the up and similarly down to scroll the page down.

III. IMPLEMENTATION

The method for mouse movement by pupil and mouth movement was achieved using the following process:-

A. Face Detection

To represent a clean and accurate image, the user should face parallel to the webcam position as represented in Fig. 1.

The image of the person is captured by webcam pre-installed in the system using OpenCV and processed using python. Fig. 2 shows the captured image using a webcam.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:13 UTC from IEEE Xplore. Restrictions apply.

Dlib has a pre-trained dataset iBUG 300-W which has 68 coordinates that mapped to the face of a person. The facial detector detects key landmarks on the face of a person and tracks them. Fig. 3 shows 68 facial coordinates.

B. Eye Detection

For eye detection, we use Eye-Aspect-Ratio (EAR). It was used to observe if the person's eye is flickering or not in the video frame.

Each Eye is expressed as 6 coordinated (p1-p6), p1 is the coordinate of the left part of the eye and then p2-p6 is located accordingly when we travel in the clockwise direction as shown in (1).

$$EAR = \frac{(p2 - p6) + (p3 - p5)}{2(p1 - p4)}$$

p1-p6 *are points located on the eye*
p2-p6 and p3-p5 depicts the vertical distance between the eyes
p1-p4 depicts the horizontal distance between the eyes



Fig. 1 System set up for eye mouse

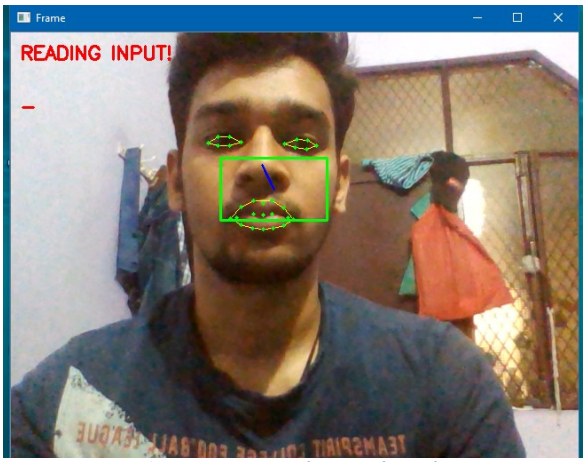


Fig. 2 Image captured using the webcam

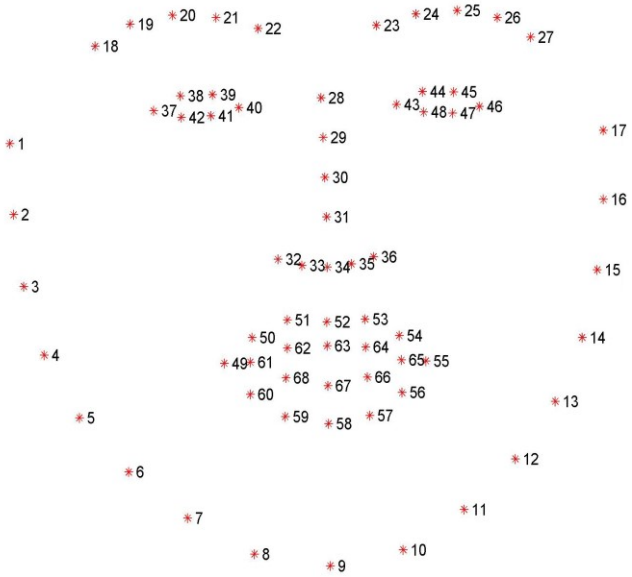


Fig. 3 68 Facial coordinate point

eye-aspect-ratio (EAR) is consistent until the time eye is open, but it tends to zero when a flickering is started as shown in Fig. 5.

C. Mouth Detection

Same as eye-aspect-ratio, we have mouth-aspect-ratio(MAR), which is used to identify whether the mouth is open or not as shown in Fig. 6

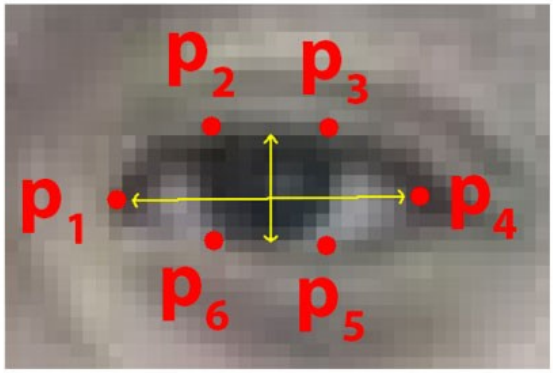


Fig. 4 The 6 facial landmark associated with eye

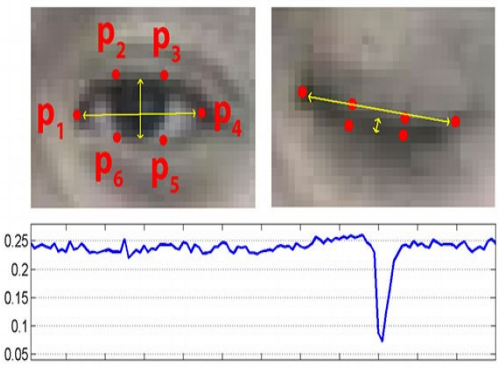
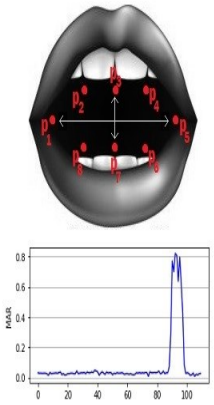


Fig. 5 Top-left: *Eye is open.* Top-right: *Eye is closed.* Bottom: *Plotting the Eye-Aspect-Ratio over time.*

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:13 UTC from IEEE Xplore. Restrictions apply.



$$MAR = \frac{\|p_2 - p_8\| + \|p_3 - p_7\| + \|p_4 - p_6\|}{2\|p_1 - p_5\|}$$

Fig. 6 Mouth-Aspect-Ratio

D. Algorithm

The algorithm used can be depicted by the flowchart shown in Fig. 7.

- The camera keeps looking for pictures at the rate of 30 frames per second. If it can detect a face then it checks whether the face control function is on or off.
- Then if it is enabled it captures frames for different instructions by the user through the movement of his face as per the directions given.
- If he wants to turn the scroll mode on he/she has to squint both eyes and then move his face up or down for scrolling.
- To simply move the cursor user will simply move his head.
- For click, he will squint one of his eyes as per the side of the mouse he wants to press.
- To disable the mouse control, he has to again open his mouse open.

E. Usage

The figure 8 explained the different modes and working of the algorithm. In this, we explained the action and their corresponding function we used in our algorithm.

- For activating/deactivate mouse control we have to open mouth continuously for some time.
- If the user blinks their left eye it will act as a left-click.
- If the user blinks their right eye it will act as a right-click.
- If the user squints his/her eye for some time it will activate/deactivate scrolling.
- Head will be used for cursor movements and scrolling up or down

IV. RESULTS AND DISCUSSION

After the implementation of the algorithm it was tested on a webpage and it showed correct results. The test started with the camera being turned on looking to detect a face. As soon it detects a face it marks the border of eyes and mouth to tell the user about the areas it will consider for an input.

First, we have to activate mouse control mode by opening mouth as shown in Fig 8 and it will show reading input. A rectangular box will be there to tell you about your movement

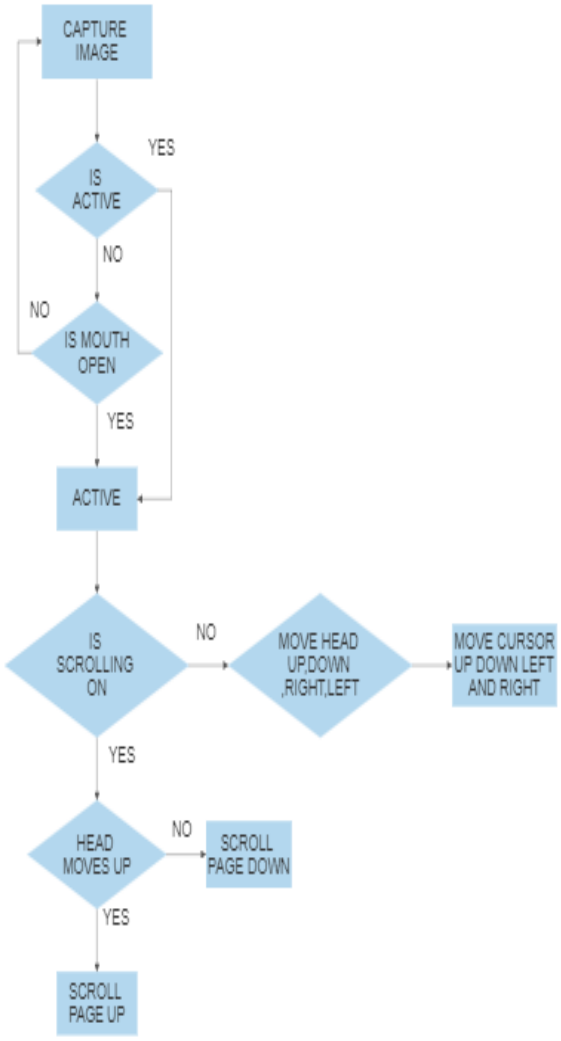


Fig. 7 Flowchart of the algorithm

from the initial position and will draw a line in the direction of movement which will tell about the length traveled from the epicenter.

For moving the mouse cursor left or right we simply have to move our head in the intended direction of movement of the cursor.

For turning the scroll mode on we have to squint our eyes into the camera and on initiation, it will show scroll mode is on. Now you can move your head up for scrolling the up and similarly down to scroll the page down.

For clicks, we have to blink the left eye and right eye for left-click and right-click respectively. Fig. 9 to Fig. 12 are shown for results.

- Fig. 9 depicts cursor movement towards the left direction when mouse control is activated by opening mouth for a certain time.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:13 UTC from IEEE Xplore. Restrictions apply.

- Fig. 10 depicts cursor movement in the right direction when mouse control is activated by opening mouth for a certain time.
- Fig. 11 depicts scrolling function towards a downward direction when mouse control is activated by opening mouth for a certain time and scrolling mode is also activated by squinting eyes.
- Fig. 12 depicts scrolling function towards upward direction when mouse control is activated by opening mouth for a certain time and scrolling mode is also activated by squinting eyes.

V. CONCLUSION

A system that facilitates contactless control of mouse is developed and tested. It has an easy and wide application. Easy mouse cursor movement using face and switching between modes using eyes to enable or disable scrolling. Blinking to click. It has a wide range of applications. It can be used for physically disabled people in their education as this would empower them to type instead of writing without hand. This will lead to job opportunities for disabled people. It can be used in vehicles to observe the laziness of the driver by detecting the drowsiness symptoms like yawning. It can also be used in gaming and the Internet of Things domain.

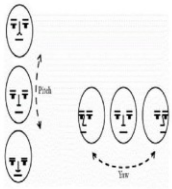
Action	Function
 Opening Mouth	Activate / Deactivate Mouse Control
 Right Eye Wink	Right Click
 Left Eye Wink	Left Click
 Squinting Eyes	Activate / Deactivate Scrolling
 Head Movements (Pitch and Yaw)	Scrolling / Cursor Movement

Fig. 8 Action and their respective functions

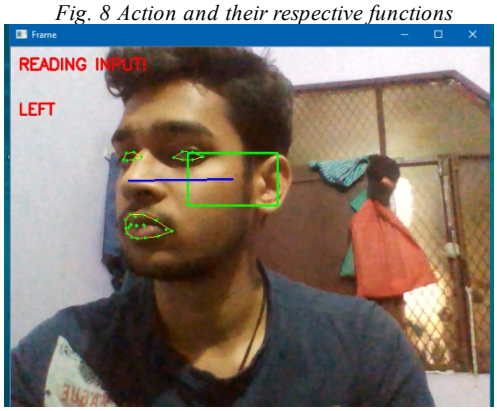


Fig. 9 Left Movement

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:13 UTC from IEEE Xplore. Restrictions apply.

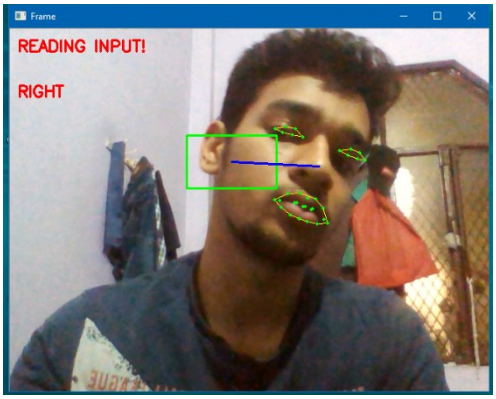


Fig. 10 Right Movement

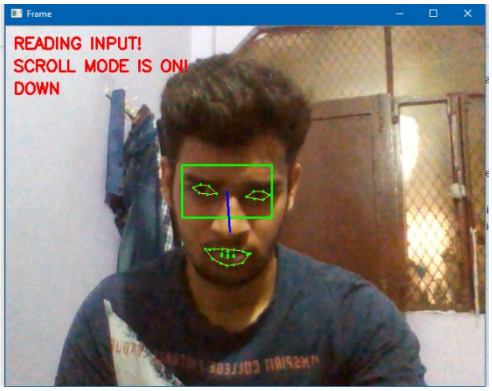


Fig. 11 Scroll Mode on with downward scrolling

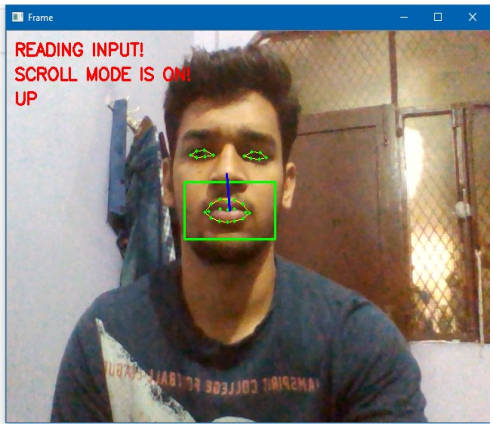


Fig. 12 Scroll Mode on with upward scrolling

REFERENCES

[1] Soukupová, Tereza, and Jan Cech. "Eye blink detection using facial landmarks." *21st computer vision winter workshop, Rimske Toplice, Slovenia*. 2016.

[2] Chin, Craig A., and Armando Barreto. "Enhanced hybrid electromyogram/eye gaze tracking cursor control system for hands-free computer interaction." *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006.

[3] Jacko, Julie A., ed. Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications. CRC press, 2012.

[4] Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. "Openface: an open source facial behavior analysis toolkit." *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016.

[5] Ballard, Philippe, and George C. Stockman. "Computer operation via face orientation." *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*. IEEE COMPUTER SOCIETY PRESS, 1992.

[6] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic. *300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge*. Proceedings of IEEE Int'l Conf. on Computer Vision (ICCV-W), 300 Faces in-the-Wild Challenge (300-W). Sydney, Australia, December 2013

[7] Balamurugan, P., J. Santhosh, and G. Arulkumaran. "HAND MOTION BASED MOUSE CURSOR CONTROL USING IMAGE PROCESSING TECHNIQUE." *Journal of Critical Reviews* 7.4 (2020): 181-185.

[8] Vasisht, Vinay S., Swaroop Joshi, and C. Gururaj. "Human computer interaction based eye controlled mouse." *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2019.

[9] Mehta, Sukrit, et al. "Real-Time Driver Drowsiness Detection System Using Eye Aspect Ratio and Eye Closure Ratio." *Available at SSRN 3356401* (2019).

[10] Pimplaskar, D., Nagmode, M. S., & Borkar, A. (2015). Real time eye blinking detection and tracking using opencv. *technology*, *13*(14), 15.

[11] Kalas, M. S. (2014). Real Time Face Detection and Tracking Using OPENCV. *international journal of soft computing and Artificial Intelligence*, *2*(1), 41-44.

[12] Chau, M., & Betke, M. (2005). *Real time eye tracking and blink detection with usb cameras*. Boston University Computer Science Department.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:13 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Convolutional Neural Network based Automated Attendance System by using Facial Recognition Domain

Shailender Kumar
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

Dhruv Kathpalia
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
kathpaliadhruv@gmail.com

Dipen Singh
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
dipen.singh.19@gmail.com

Mandeep Vats
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
mandeepvats98@gmail.com

***Abstract* - This project aims to recognize faces in an image, video, or via live camera using a deep learning-based Convolutional Neural Network model that is fast as well as accurate. Face recognition is a process of identifying faces in an image and has practical applications in a variety of domains, including information security, biometrics, access control, law enforcement, smart cards, and surveillance system. Deep Learning uses numerous layers to discover interpretations of data at different extraction levels. It has improved the landscape for performing research in facial recognition. The state-of-the-art implementation has been bettered by the introduction of deep learning in face recognition and has stimulated success in practical applications. Convolutional neural networks, a kind of deep neural network model has been proven to achieve success in the face recognition domain. For real-time systems, sampling must be done before using CNNs. On the other hand, complete images (all the pixel values) are passed as the input to Convolutional Neural Networks. The following steps: feature selection, feature extraction, and training are performed in each step. This might lead to the assumption, where convolutional neural network implementation has a chance to get complicated and time-consuming.**

Keywords – Face detection, face recognition, deep learning, convolutional neural networks

I. INTRODUCTION

Face recognition is a unique technique for performing authentication biometrically. It has broad applications in areas of finance, security, and military. Face recognition has gained a lot of interest in the last few years, which has led several researchers to work for developing new techniques and improve the existing ones. Its wide range of applications appeals to researchers and keeps them driven. Face recognition can be performed on a real-time video by considering it as a sequence of frames where each frame is considered to be a single image. Before facial recognition can be carried out, we must first ensure the presence of a face in the frame. This can be done by performing face detection. In this step, the model detects the face and separates it from the image for identification, eliminating redundant data that is not required for facial recognition. This reduces the number of pixels on which the model has to work on and hence increasing the overall efficiency.

However, facial recognition [12] also faces some problems, making it very hard to perform. Various factors like pose variation, facial hair, image illumination, image background, and facial expressions affect the image, and the outcome can differ based on these characteristics. In situations where the face is not visible or hidden from the camera, the face might not even be detected. Thus, the image used as input to the model could be in different conditions as opposed to the image, which is to be examined.

We are attempting to instill this technique in universities, so that common issues like higher time-consumption during attendance, students marking proxy attendance and mass bunks during lectures can be prevented. Marking attendance in classes is an overwhelming task for the professors as it is not only time consuming, but also the students tend to mark proxy attendance, which leads to inaccurate records of attendance. Manual attendance is certainly tough for the professors, as it makes it difficult to maintain a record of the students. The conventional ways often have their difficulties. The majority of these methods lack dependability. It leads to an increasing need for better methods of attendance. This research stresses on using facial recognition as a technique for marking attendance. Real-time automated attendance monitoring without wasting teacher's precious time is the main objective of this project. Not only does this method save time, but it is also more reliable than traditional methods.

In the next section, we discuss some of the facial recognition algorithms, assessing their pros and cons. In Section III, we present our proposed model and briefly explain how it works. The experimental results obtained after testing the model are given in Section IV. Section V concludes the paper with a summary.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:46 UTC from IEEE Xplore. Restrictions apply.

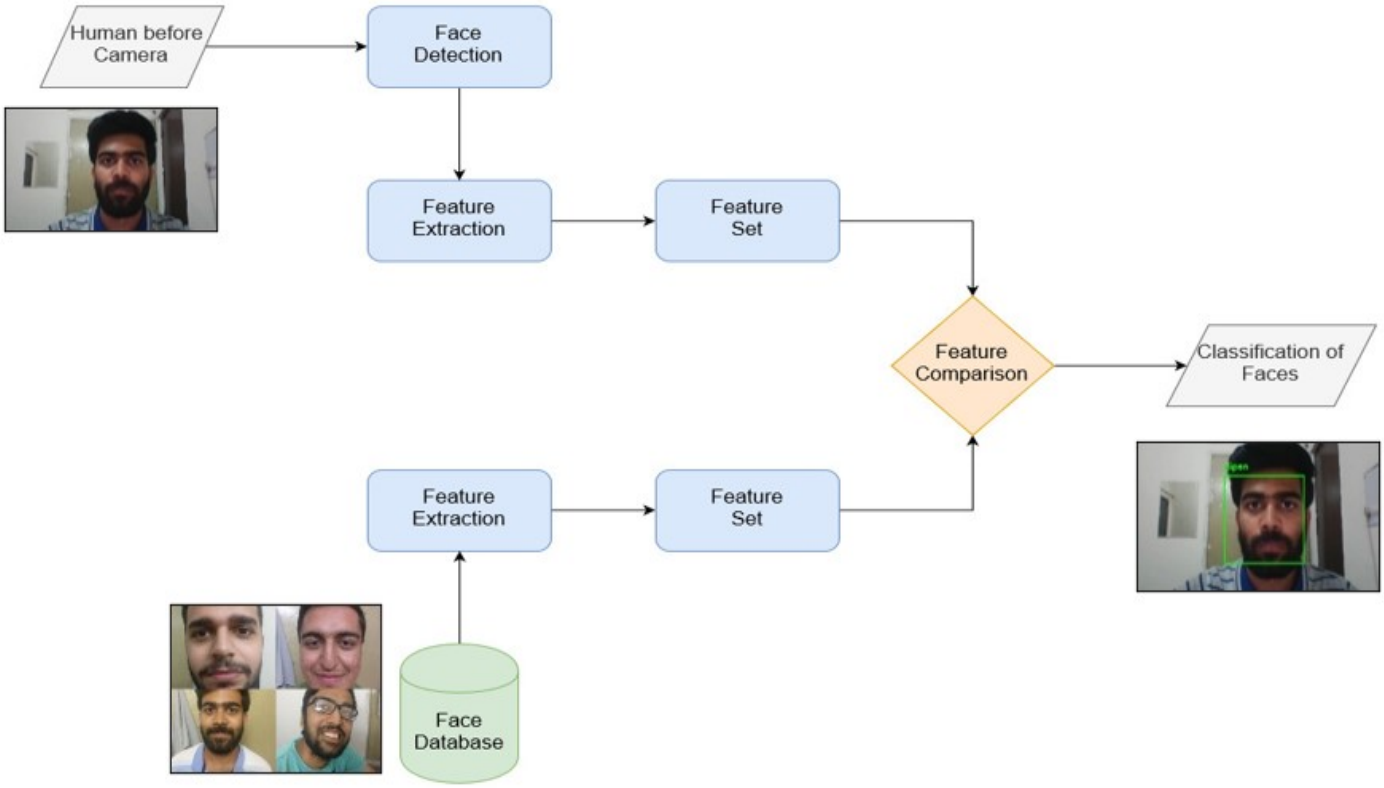


Figure 1: Flow-chart showing the working of the facial recognition aspect of our paper

II. RELATED WORKS

Dr. Priya Gupta et al. [1], obtained an accuracy of 97.05% using their proposed method. They performed feature extraction with the help of Haar Cascades, which were, in turn, fed forward to the network instead of raw pixel values. The complexity was greatly reduced as it led to a decrease in the redundant input features. It uses Deep Neural Networks, which makes the model very efficient in terms of using fewer resources and making it faster.

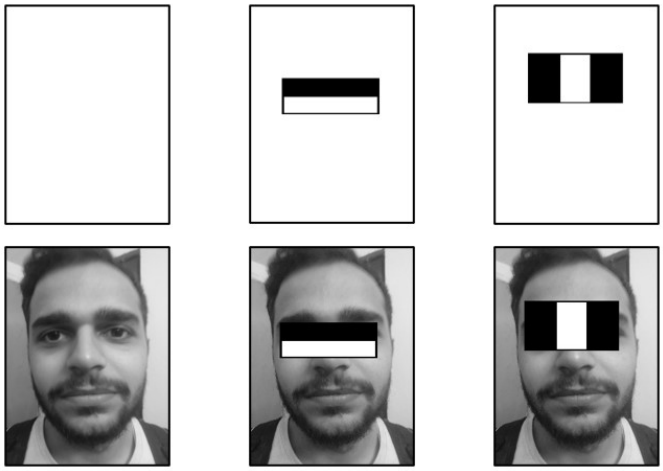


Figure 2: Haar Features applied to a Face

Haar cascades [10] use different features that are trained first with the help of a training set, which consists of both positive as well as negative images. After we have trained the classifier, it can be used to detect if the object is present in any test image or not. These have been immensely useful in tasks related to facial detection as they make the process much faster as it has considerably fewer computations as compared to other methods. Different filters can be used for features such as eyes, nose, mouth, etc. as shown in figure 1.

R. Rahim et al. [2], used the Fisher Linear Discriminant(FLD), which was discovered by Robert Fisher in 1936. It is a popular pattern recognition method, which has applications in the face and object recognition. It increases the accuracy of the classification by forming inter-class, and intra-class scatters. The algorithm can recognize the faces even with changes in some attributes of the faces, like expressions or wearing glasses.

Faizan Ahmed et al. [3], shows a comparative study on the different methods for performing facial detection and recognition. They achieved an accuracy of 96.7% using the Adaboost classifier [4] along with Haar features and an accuracy of 90.88% using Support Vector Machine(SVM) [5] classifier for Face detection application. For Facial Recognition, they used the following methods:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:46 UTC from IEEE Xplore. Restrictions apply.

i. Principal Component Analysis(PCA) [6] method which uses the concept of Eigenfaces, with an accuracy of 71.15%. The computations required in this method is much lower as compared to the other methods as it only considers the 2D face recognition problem. Thus the complexity is reduced by a substantial amount.
ii. Linear Discriminant Analysis(LDA) [7] method uses Fisherfaces, with an accuracy of 77.90%. It mainly focuses on reducing the number of features being applied on each face.
iii. Local Binary Pattern(LBP) [8] has an accuracy of 82.94%.
iv. Gabor Classifier [9], which takes into consideration the local features, has an accuracy of 92.35%. It is not designed specifically for face recognition, but its filters can recognize various prominent features in an object.

Convolutional Neural Networks [11] are a type of Neural Network that is mostly used in the field of image classification, particularly Face Recognition. Convolutional Neural Networks take an input image and tweak the weights of the network-based of the input image so that it can differentiate it from other images. This allows the network to learn and identify the important characteristics (that are essential for recognizing different faces) on its own. The need for human supervision is thus minimized- it can automatically differentiate the images into separate classes. Convolutional Neural Networks also reduce the need for pre-processing required to train the model, thus it utilizes less computation power. Due to these advantages, deep learning algorithms like convolutional neural networks have become the standard in facial recognition.

III. PROPOSED WORK

We have proposed a method for an automated attendance system using facial recognition. The system should be able to detect faces in each frame of a real-time video. Further, after recognizing the detected faces, it should be able to mark the attendance of students whose faces are recognized by the system. The crux of the system is that it marks the attendance of only those students who have attended more than sixty percent of the total class; the rest of the students are marked absent.

The proposed system requires a video camera in the class to be an initial requirement. The proposed system is designed to work on video footage of students. The underlying idea is to extract features of students' faces from the footage, and compare these features with those which are extracted from the training images used for training the model. If these features match, the student is marked present for that single frame.

- A video camera placed in the classroom would continually record the class and pass the input stream to the attendance system.
- At regular intervals, individual frames are analyzed by the model.
- All the faces of the students for a particular frame would be detected and, in turn, recognized by our model.
- The same process would be followed for each frame for real-time video analysis and facial recognition.
- At the end of a class, after all the frames are done, the model would mark the students as absent or present based on the model's results.
- The student must be recognized in at least 60 percent of the frames to be marked present.
- The database would update the information on marking the attendance of all the students for that particular date.
- The database also stores the attendance record for all dates. Therefore it will also make it easier to manage the attendance records.

Pre-processing

Our model is based on deep learning-based facial recognition. We have used the face_recognition library that is a pre-trained network and uses it to generate 128-d vectors from the training dataset. Pre-processing involves the following steps:

Step 1: Face Detection

- Convert the given image to grayscale.
- Apply Haar features to each image by dividing the image into smaller squares and detect the presence of different features such as edges, corners, etc.
- We obtain a basic structure of the image, representing the obtained features.
- We can compare this structure to a previously extracted pattern of a face. This helps us in identifying the different faces present in the image.

Step 2: Resolving the issue of projecting faces

- Although we have isolated faces, the computer mistakes a face looking in different directions as different faces.
- To resolve this issue, we need to alter the positioning of a face in an image.
- We take some points that lie on a face and using these points, we manage to detect a face, its boundary, the positions of eyes, nose, and lips from an input image.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:46 UTC from IEEE Xplore. Restrictions apply.

- Next, the image is rotated such that the transformed face in the rotated image is as close as possible to a perfectly centered face.
- Using this method, the computer does not categorize the projections of a face as different faces.

Step 3: Encoding Faces

- The faces that the model detected cannot be compared to each face in the database, as this method would require a lot of time.
- Instead, we have to take an approach that uses only a few metrics for each face. Metrics such as distance between the eyes or the shape of the nose aren't the most accurate in differentiating between faces.
- The metrics that are used are measurements that the network defines on its own, using Deep learning.
- Here the CNN will be trained to create 128 vectors for each face.
- Now, we take a triplet: 2 images of the same known person and one of a different person. The measurements for each image are generated.
- The network learns by adjusting so that the first two images are closer, and the measurements of the third image are farther.
- The neural network does this repeatedly for thousands of images so that for the same face, two different images give nearly the same 128 measurements.
- Huge amounts of input data and processing power is required to train a CNN using this process. However, once trained, the network can be used to generate accurate measurements, even for new faces.
- Hence, now the network (that is already trained) only requires our images to calculate the 128 measurements accurately.

Step 4: Finding a Match

- First, we detect the faces and match each of them to the encodings that we have from our training set.
- The Euclidean distance is calculated with each of the faces in our database:
    - The distance must be below some threshold value to predict that the face matched.
    - If the distance is higher than the threshold, then we consider the face to be unknown.

Processing the Input Data

The input data is the video stream from the video camera from the classroom. As the video stream is a collection of frames, we apply the face detection module to each of the frames. Once all the faces in a frame are detected, they are then compared with the generated pickle file from the training dataset. The faces that are matched are classified with the name of the student, whereas the faces that have no match are classified as unknown. The initial step of processing the data, i.e., to detect and recognize the faces in the input data, is shown in figure 3. The only step left is to mark the attendance of students whose faces are classified in the previous step.

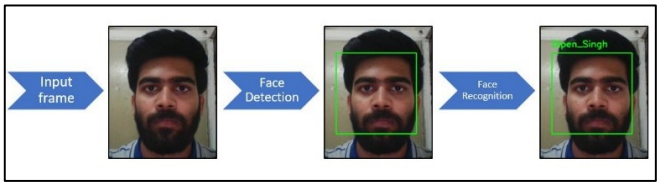


Figure 3: Steps involved in Classification of a face

The facial recognition is performed on each frame of the real-time video. So, each frame contributes to the whole attendance. Being present in only a single frame doesn't mean that the student would be marked present. This is due to the algorithm not being 100% accurate, as we have to consider situations where the students' faces are not visible. It is almost impossible to receive such high accuracy. Thus, we try and tackle the problem in a way that ensures that the students that are present in that class are only marked as present and not the ones that may be incorrectly recognized in a single frame. This also ensures that if a student's face isn't recognized in one or two frames due to any issue, they are still marked present.

The following steps are taken to ensure the above:

- We create a variable count that stores the count of the number of frames that each student's face has appeared in. We initialize this variable to 0 for each student.
- We also create a total count variable that stores the total number of frames.
- For each frame, we do the following:
    - Detect the faces in the frame.
    - Recognize the faces by matching each of them to the encodings that we have from our training set.
    - Increment the count corresponding to the student.
    - If a face is unknown, we skip that face
    - These steps are followed until we reach the last frame, i.e., the end of the class.
- Now we compare the count for each of the students to the total count. If the count for a student is greater than 60% of the total count, then we would mark the student as present in the database, else we would mark him/her as absent.
- This ensures that the student will only be marked present if he has attended at least 60% of the class.

Also, this helps in improving the accuracy of the system as even if a face is recognized incorrectly, its effect would almost be negligible on the final result.

We have proposed a method here in which the model has been trained and tested on our customized database that contains images of 16 students as subjects. To find an effective

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:46 UTC from IEEE Xplore. Restrictions apply.

number of images per person required for our model to provide the best result, we trained and tested the model with a different number of images per person. Starting from a single image of each person, we performed the testing with as many as 25 images per person. The images of every subject differ in their facial expressions, contrast, exposure, configuration, etc. We implemented the model on a 64-bit system using Python 3.6.9. For the pre-processing of input images, we have used the OpenCV package using Haar cascade and its frontal face feature.

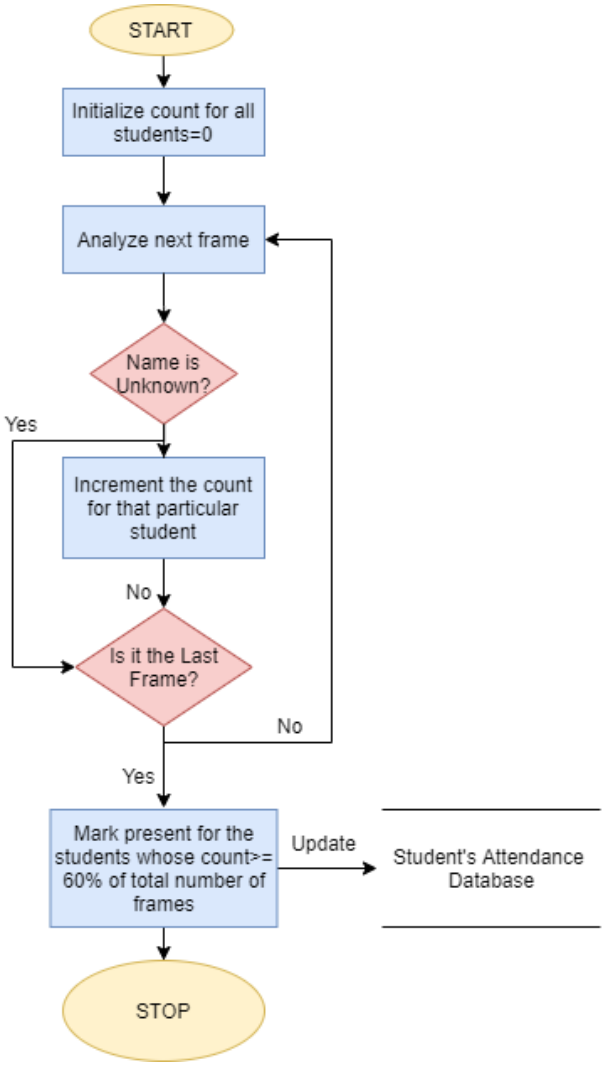


Figure 4: Flow Chart for Data Processing

IV. Experimental Results

All the images are then compressed, using a compression algorithm, to a suitable size that it contains all the important features required for recognition of faces and does not take too much time to extract the classifiers from the training images, converting them to 128-d vector and creating a pickle file. Further, the pickle file generated is used in the model to recognize the faces, instead of training the model every time the model is given an input.

We analyzed our model both quantitatively as well as quantitatively. We first assessed the number of faces the model was able to detect, i.e., quantitative analysis, when given real-time video input, and then measured the accuracy of the model by calculating the number of faces that were recognized correctly, i.e., qualitative analysis.

Also, the result of our framework, when stored in the database is sorted according to the dates on which the lecture has taken place. Initially, all the students for a lecture are marked absent and on recognizing their faces, and if they follow certain criteria mentioned above, the model updates their attendance in the database marking them present for the lecture. A snapshot of a part of a database, showing the names of only five students, is shown in Fig. 5. We have used WampServer for our MySQL database.

Date	Day	Dhruv_Kathpalia	Dipen_Singh	Mandeep_Vats	Dhruv_Ramdev	Deepanshu_Varshney
2019-10-07	Mon	A	A	A	P	A
2019-10-15	Tue	A	A	A	A	P
2019-10-23	Wed	P	P	A	A	A
2019-10-31	Thu	A	A	P	A	P
2019-11-08	Fri	P	P	P	A	P
2019-11-13	Wed	A	P	P	P	P
2019-11-20	Wed	P	P	P	P	P

Figure 5: A snapshot of the Attendance marked in the Database

Once we get the number of faces recognized in a frame, we store the list of such students. For each student, we maintain a count that contains the number of frames in which a student's face is present. And at the end of the session, if a student's face is present in 60 percent of the total number of frames, the student is marked present. Failing to fulfill the required criterion, the student is marked absent.

Table 1: Images used for Model Training and Testing

Test Case	Images Per Person in the Training set	Total students present	Students marked present in 4 classes (average)	Accuracy (%)
Case 1	1	13	7.25	55.77
Case 2	3	13	9	69.23
Case 3	5	13	10	76.92
Case 4	10	13	10.75	82.69
Case 5	15	13	11.5	88.46
Case 6	20	13	12	92.31
Case 7	25	13	12.5	96.15

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:46 UTC from IEEE Xplore. Restrictions apply.

On performing training and testing with a different number of images per person, we noticed that on increasing the number of images of a person, the number of important features in the pickle file from the compressed images also increases. This further increases the probability of the system comparing and deciding the detected face correctly with the features extracted from the pickle file.

In our case for automatic attendance based on facial recognition, the system admin can take multiple images of the student at the time of registration, which would be used for the training dataset for high accuracy. This makes the proposed attendance system reliable for marking attendance of students during lectures in the universities. The trend analysis of a different number of images per person on the accuracy of the system, as observed during the testing of the system is given in Fig. 6.

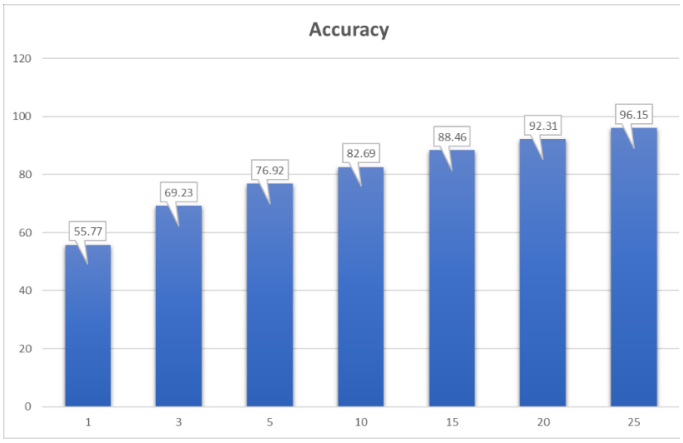


Figure 6: Graph depicting dependency on the accuracy of the model on the size of the training set

V. CONCLUSION

Using Convolutional Neural Network for facial recognition helps in reducing time and the processing power used as compared to other conventional methods. The model has great accuracy. For 25 images per subject, we achieve an accuracy of 96.15%. Although the accuracy is assumed to be very low for fewer images, it is compensated by the extra step that ensures that the student is marked present only if the number of frames their faces are identified is greater than the predefined threshold of 60%. This results in accuracy that is much higher than expected. This model can also be applied to online classes. During online lectures too, the conventional methods would waste precious time. They cannot be considered very reliable either, as anybody can log in as a student if they have the login credentials. Instead of manually taking attendance, which might be tedious for large groups of students, attendance will be taken automatically in the background. Facial recognition would ensure that attendance is reliable.

REFERENCES

[1] P. Gupta, J. Tripathi, M. Sharma, and N. Saxena, "Deep Neural Network for Human Face Recognition," *MECS*, p. 9, 2018.
[2] R. Rahim, "Research of Face Recognition with Fisher Linear Discriminant," *IOP Conf. Ser.*, 2018.
[3] A. Najam, Z. Ahmed and F. Ahmed, "Image-based Face Detection and Recognition," *IJCSI*, 2012.
[4] A. Tikare, S. Kadam, and K. T. Talele, "Efficient Face Detection using Adaboost," in *IJCA Proc on International Conference in Computational Intelligence*, 2012.
[5] B. McCane and I. Kukenys, "Support Vector Machines for Human Face Detection," in *Proceedings of the New Zealand Computer Science Research Student Conference*, 2008.
[6] M. A. Turk and A. P. Pentland, "Face recognition using Eigenfaces," in *Proceedings of the IEEE*, 1991.
[7] J. Lu, N. Plataniotis, and A. N. Venetsanopoulos, "Face recognition using LDA-based algorithms," in *IEEE Neural Networks Transaction*, 2003.
[8] T. Ahonen, A. Hadid, and M. Peitikainen, "Face recognition with local binary patterns," in *"In Proc. of European Conference of Computer Vision*, 2004.
[9] L. Wiskott, M. Fellous, N. Krger, and C. Malsburg, "Face recognition by elastic bunch graph matching," in *IEEE Trans*, 1997.
[10] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," *IEEE Computer Society Conference on Computer Vision*, 2001.
[11] Yamashita, R., Nishio, M., Do, R.K.G. *et al.* Convolutional neural networks: an overview and application in radiology. *Insights Imaging* **9,** 611–629 (2018).
[12] "A Summary of literature review: Face Recognition" by Kittikhun Meethongjan & Dzulkifli Mohamad, August 2007, Postgraduate Annual Research Seminar 2007
[13] King, Davis E. "Dlib-ml: A machine learning toolkit." *Journal of Machine Learning Research* 10.Jul (2009): 1755-1758.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:46 UTC from IEEE Xplore. Restrictions apply.

IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)
e-ISSN: 2278-1684,p-ISSN: 2320-334X, Volume 17, Issue 3 Ser. I (May - June 2020), PP 07-24
www.iosrjournals.org

Design and Analysis of a Parallel Link Manipulator for Heavy Lifting in Warehouses

Rohit Kumar, Shubham Kumar, Siddharth,Vikas Rastogi
Mechanical Engineering Department, Delhi Technological University
Professor Vikas Rastogi, Department of Mechanical Engineering, Delhi Technological University,
Rohit Kumar, Department of Mechanical Engineering, Delhi Technological University
Shubham Kumar, Department of Mechanical Engineering, Delhi Technological University
Siddharth, Department of Mechanical Engineering, Delhi Technological University

Abstract: *This paper is intended to put forth a new approach to robot design that is centered around Parallel Link Manipulator. Parallel Link Manipulator is a challenging but equally rewarding field of research. It can prove to be the remedy for the long-standing problem of low 'Lifting Weight to Robot Weight Ratio'. The very large size of actuators inserial link manipulators for heavy lifting is also a problem. Very heavy construction is required for heavy lifting purposes. In this paper we propose a new parallel link manipulator design to work as normal operating robot in warehouse and factories. We also present its forward and inverse kinematics with static structural analysis for testing the maximum safe lifting capacity of the robot. The main advantage of the proposed design is its great maximum lifting capacity to body weight ratio with ease in manipulability.*
Key Words: *Lifting Capacity to body weight ratio, Parallel Link Manipulators, Forward Kinematics, Inverse Kinematics, Static Structural Analysis, Manipulability.*

Date of Submission: 27-04-2020 Date of Acceptance: 10-05-2020

I. Introduction

The Science Fictions that we watched not so long ago and disparaged them as just a vague extrapolation of work done in robotics till then, does seem to have rubbed off positively on the mind of scientists. Tremendous amount of progress has been made in all the aspects of Robotics. Technologies like Artificial Intelligence and Machine learning etc. have given a shot in the arm to the Robotics Industry. Consequently, the pliability of robots in umpteen number of industries and offices has become a reality today.

Today a major untold problem of serial link manipulator is low lifting capacity with respect to body weight of the robot. Thus, very heavy construction is required for a serial link manipulator to lift heavy objects in warehouses. To get a detailed perspective on this ratio, it is worthwhile to look at the following analysis.

In Industries, a 1750 kg of Robot (IRB 660) has a weight lifting capacity of approximately 180 to 250 kg. The ratio that comes out is meagerly 0.1 to 0.15. This abysmally low ratio is a serious challenge to the applicability of robots in Heavy Industries, construction sites, etc

Through our study we have tried to address this issue. We need to design a robot system which can overcome this problem. Specifically, we need to design a robot that can be used in warehouse setting which has high maximum lifting capacity to body weight ratio. Thus, the robot should have simple, robust and light weight construction as compared to maximum load that can be lifted by robot.

Another major requirement for a robot to work in warehouse is ease of control and manipulability. This is a very important requirement in industry as most of the warehouses and factories have very low tolerance for error.

Most of the serial link manipulator also have the problem of compounding of required actuator strength when moving from end effector to base of the robot. It means that every successive actuator in the serial link manipulator should be powerful enough to lift the weight of entire robot body and object weight after the actuator till end-effector. Thus, the base actuator in a normal serial link manipulator need to be very powerful. Due to this cost of the large sizeactuatorsused in Serial Link Manipulator is a very much considerable factor.
In this paper we want to address most of these problems by presenting a new design for a robot that can be used in a warehouse or factory setting.

II. Problem Formulation

We formulate the task of building such a robot in four main phases

I. Design Phase
II. Forward Kinematics Phase
III. Inverse Kinematics Phase
IV. Structural Analysis Phase

In these phases we aim to design and analyse the robot which fulfils the desired requirements. The main objectives that we want to achieve are

I. A robot with considerably high lifting weight to body wight ratio as compared to a normal serial link 6 DOF manipulator.
II. A robot which should be considerably easy if not easiest to manipulate.
III. A robot which should be free from compounding of actuator power problem.

DESIGN

Today majority of the robot used in industry are Serial link Manipulators. They are most widely used because they are easy to manipulate, design and construct. But Serial link robots have a major drawback i.e. they require very heavy construction to lift heavy objects. The maximum lifting capacity to robot body weight ratio is usually very low (less than 1). On the other hand, Parallel link manipulators are very robust and rigid, but are also very complex to manipulate, design and manufacture. Usually Parallel link manipulators have very high maximum lifting capacity to body weight ratio (can be more than 1). They achieve this capability by parallelly using various link to distribute the load. Since we want to design a robot with high maximum lifting capacity to body weight ratio, we are required to design a parallel link manipulator.

Another major problem is manipulability. Parallel link manipulators are famous for their complex manipulability. A very complex parallel link manipulator has high chances of losing its accuracy and manipulability. This is not a very desirable characteristic of a robot that is being designed to work in warehouses and factories where tolerance for error is very low. So, the parallel link manipulator that we want to design should also be considerably easy and fast to manipulate.

Therefore, to meet the desired requirements, we propose the following design.

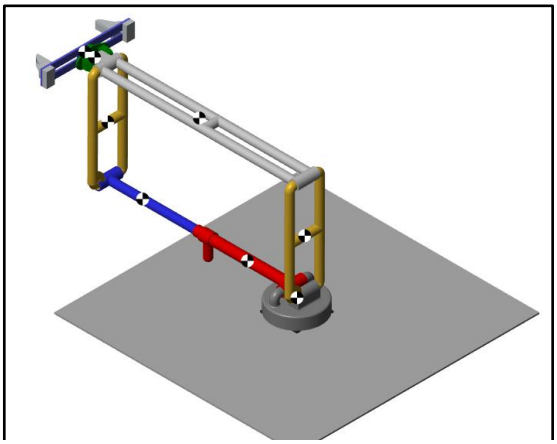


Figure 1: Isometric view of proposed robot design

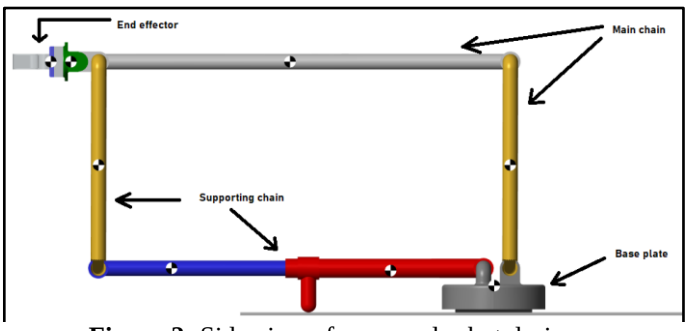


Figure 2: Side view of proposed robot design

The proposed design has following key features:

- The robot consists of one common "base plate" which consist of one revolute joint with grounded stationary base plate.
- The robot has one main manipulator chain which has two links. The vertical link in main chain is called as "standard link 1" while the longer horizontal link is termed as "main link". In the proposed design the main link is twice the length of standard link 1.
- The robot also has secondary supporting chain. It consists of one very large piston-cylinder link which is nearly of same size as that of main link in its maximum extended configuration. Another vertical link is known as "standard link 2". It is of exactly the same dimension of standard link 1.
- The "cylinder link" (in piston-cylinder link) which is attached to base plate has a rigid vertical extrusion towards ground. The extrusion provides robot with another contact to ground and also a very firm secondary support.
- The End-effector consists of two supporting plate and two gripper pads with friction pad for increased gripping strength. The first link plate is known as "roll plate" and is attached to main link and provides pitch motion to end-effector and also serves as base to connect to later "gripper plate". The joint between "roll plate" and "gripper plate" provides roll motion to end effector. The "gripper plate" has a very wide construction which helps robot to exercise strong grip over large boxes.

This robot is designed to work in two configurations. One is "parallelogram configuration" and the other is "extended configuration". The parallelogram configuration can be used to lift object which has height lower than height of the robot in rectangular configuration (a case in parallelogram configuration). While the extended configuration is used to lift much taller and higher objects. Both of the configurations are shown below.

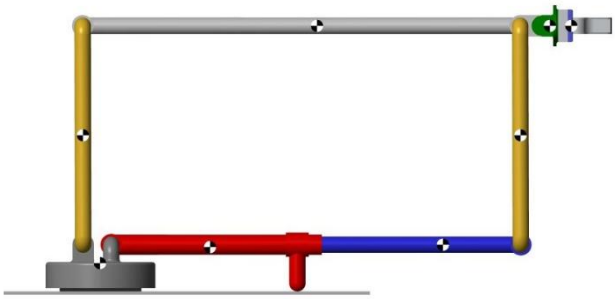


Figure 3: Parallelogram Configuration

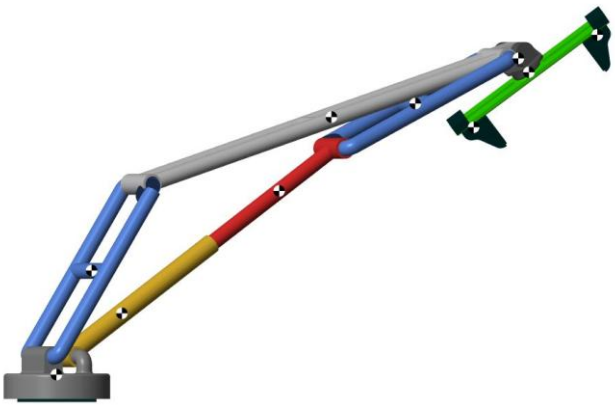


Figure 4: Extended Configuration

The secondary supporting chain acts as a support to main chain in both the configuration. The secondary chain also increases the maximum lifting capacity while not creating much difference in manipulability.

The robot has five degree of freedom. The degree of freedom is in X, Y, Z as translational DOF while robot also has roll and pitch rotational DOF. The robot is similar to a serial link manipulator. Due to which it is significantly easier to manipulate.

IV. Forward Kinematics

TheParallel link manipulators are known to have very complex forward kinematics. And our proposed robot is not very different in this respect. But with geometrical and analytical method formula for each unknown term in this robot can be derived. Therefore, we have used geometric and calculus to find formula for every

variable in this robot. The forward kinematics of this robot is derived for two separate chains individually. The Forward kinematic equation for supporting chain converges with main manipulator chain at "standard link 2" and "main link" revolute joint. Due to this arrangement we can find inverse kinematic of this robot using analytical and geometrical methods.

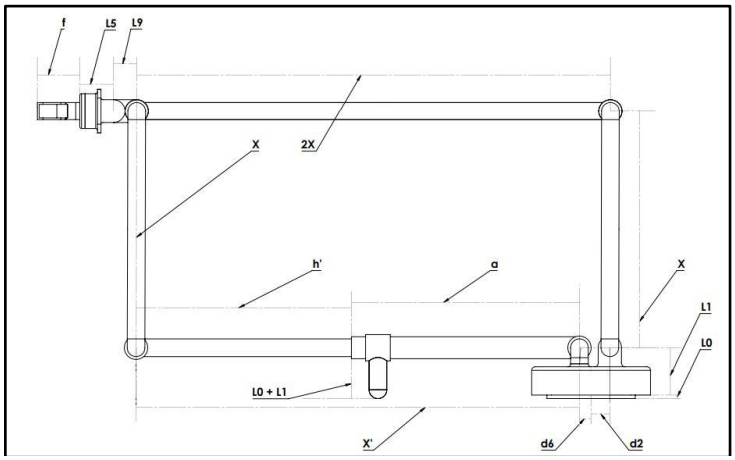


Figure 5: Parameter description of robot

For Main Chain and Manipulator

In main chain of the robot one or more frames are assigned to each link. The initial frame starts from Frame 0 i.e. ground frame and goes to tool frame via main chain. The description of these frames is given as follows:

Frame 0: Ground Stationary Frame

Frame 1: Attached to base of "Base Plate". L0 distance from Frame 0 in +z axis.

Frame 2: At "Base Plate" and "Standard link 1" revolute joint. Attached to "Base Plate" at "Base Plate" and "Standard link 1" junction axis.

Frame 3: At same point as of Frame 2 but can rotate about z axis with respect to frame 2. Attached to "Standard link 1".

Frame 4: At "Standard link 1" and "main link" revolute joint. Attached to "main link".

Frame 5: At "main link" and "roll plate" revolute joint. Attached to "roll plate".

Frame 9: At "main link" and "Standard link 2" revolute joint. Attached to "main link" and at 2*X distance from frame 4 along x axis.

Frame {Tool}: At L5 distance from Frame 4. On gripper plate. Gives revolute motion for roll DOF. Attached to "Gripper plate". Acts as final end-effector frame.

Denavit-Hartenberg Parameters for the Main chain and manipulator of the proposed robot is given in the table 1 below.

D-H Parameters for Main Chain and Manipulator

Table 1: Frame parameters of Main Chain

	α **(along axis)**	**a (along axis)**	**d (along axis)**	***θ*** **(along axis)**
Frame 0 → Frame 1	0	L0 (z axis)	0	q1 (z axis)
Frame 1 → Frame 2	+90 (x axis)	L1 (z axis)	-d2 (x axis)	0
Frame 2 → Frame 3	0	0	0	q3 (z axis)
Frame 3 → Frame 4	0	X (x axis)	0	q4 (z axis)
Frame 4 → Frame 5	0	2X + L4 (x axis)	0	q5 (z axis)
Frame 5 → Frame{Tool}	$\alpha 5$ (x axis)	L5 (x axis)	0	0
Frame 4 → Frame 9	0	2*x (x axis)	0	0

Therefore,

$${}^{0}[T]_{\{Tool\}} = {}^{0}[T]_{1} \times {}^{1}[T]_{2} \times {}^{2}[T]_{3} \times {}^{3}[T]_{4} \times {}^{4}[T]_{5} \times {}^{5}[T]_{\{Tool\}} \quad (4.1)$$

$$^{0}[T]_{\{Tool\}} = \begin{bmatrix} \omega(1,1) & \omega(1,2) & \omega(1,3) & x \\ \omega(2,1) & \omega(2,2) & \omega(2,3) & y \\ \omega(3,1) & \omega(3,2) & \omega(3,3) & z \\ 0 & 0 & 0 & 1 \end{bmatrix} (4.2)$$

The values of the above terms in matrix are:

$\omega(1,1) = cos(q3 + q4 + q5) * cos(q1)$
$\omega(1,2) = sin(\alpha5) * sin(q1) - sin(q3 + q4 + q5) * cos(\alpha5) * cos(q1)$
$\omega(1,3) = cos(\alpha5) * sin(q1) + sin(q3 + q4 + q5) * sin(\alpha5) * cos(q1)$
$\omega(2,1) = cos(q3 + q4 + q5) * sin(q1)$
$\omega(2,2) = -sin(\alpha5) * cos(q1) - sin(q3 + q4 + q5) * cos(\alpha5) * sin(q1)$
$\omega(2,3) = sin(q3 + q4 + q5) * sin(\alpha5) * sin(q1) - cos(\alpha5) * cos(q1)$
$\omega(3,1) = sin(q3 + q4 + q5)$
$\omega(3,2) = cos(q3 + q4 + q5) * cos(\alpha5)$
$\omega(3,3) = -cos(q3 + q4 + q5) * sin(\alpha5)$
$x = cos(q1) * [(cos(q3 + q4) * (L4 + 2 * x) - d2 + x * cos(q3)) + L5 * cos(q3 + q4 + q5)]$
$y = sin(q1) * [(cos(q3 + q4) * (L4 + 2 * x) - d2 + x * cos(q3)) + L5 * cos(q3 + q4 + q5)]$
$z = L0 + L1 + sin(q3 + q4) * (L4 + 2 * x) + x * sin(q3) + L5 * sin(q3 + q4 + q5)$

In main chain and manipulator part of this robot there is another matrix that is very important to find inverse kinematics for supporting chain. That matrix is $^{0}[T]_{5\text{-main}}$. It is represented as:

$^{0}[T]_{5\text{-main}} = {}^{0}[T]_{1} \times {}^{1}[T]_{2} \times {}^{2}[T]_{3} \times {}^{3}[T]_{4} \times {}^{4}[T]_{5}$ (4.3)

Therefore,

$$^{0}[T]_{5\text{-main}} = \begin{bmatrix} \omega5(1,1) & \omega5(1,2) & \omega5(1,3) & x5 \\ \omega5(2,1) & \omega5(2,2) & \omega5(2,3) & y5 \\ \omega5(3,1) & \omega5(3,2) & \omega5(3,3) & z5 \\ 0 & 0 & 0 & 1 \end{bmatrix} (4.4)$$

Also, $^{0}[T]_{\{Tool\}} = {}^{0}[T]_{5\text{-main}} \times {}^{5}[T]_{\{Tool\}}$ (4.5)

Where,

$$^{5}[T]_{\{Tool\}} = \begin{bmatrix} 1 & 0 & 0 & L5 \\ 0 & cos(\alpha5) & -sin(\alpha5) & 0 \\ 0 & sin(\alpha5) & cos(\alpha5) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} (4.6)$$

The value of terms in $^{0}[T]_{5\text{-main}}$matrix are given as:

$\omega5(1,1) = cos(q3 + q4 + q5) * cos(q1)$
$\omega5(1,2) = -sin(q3 + q4 + q5) * cos(q1)$
$\omega5(1,3) = sin(q1)$
$\omega5(2,1) = cos(q3 + q4 + q5) * sin(q1)$
$\omega5(2,2) = -sin(q3 + q4 + q5) * sin(q1)$
$\omega5(2,3) = -cos(q1)$
$\omega5(3,1) = sin(q3 + q4 + q5)$
$\omega5(3,2) = cos(q3 + q4 + q5)$
$\omega5(3,3) = 0$
$x5 = cos(q1) * [cos(q3 + q4) * (L4 + 2 * x) - d2 + x * cos(q3)]$
$y5 = sin(q1) * [cos(q3 + q4) * (L4 + 2 * x) - d2 + x * cos(q3)]$
$z5 = L0 + L1 + sin(q3 + q4) * (L4 + 2 * x) + x * sin(q3)$

For Supporting Chain

In this part of the robot one or more frames are assigned to each link starting from ground frame (Frame 0) to Frame 9 of the main chain that has been described earlier. This choice of end frame in this part of robot is essential, as it helps us in solving inverse kinematics for supporting chain. Each of the frame used in this part of robot is described as follows:

Frame 0: Ground Stationary Frame

Frame 1: Attached to base of "Base Plate". L0 distance from Frame 0 in +z axis.

Frame 6: At "Base Plate" and "Cylinder link" revolute joint. Attached to "Base Plate" at "Base Plate" and "Cylinder link" junction axis.

Frame 7: At same point as of Frame 6 but can rotate about z axis with respect to frame 6. Attached to "Cylinder Link".

Frame 8: At "Standard link 2" and "Piston link" revolute joint. Attached to "standard link 2" along x axis.

Frame 9: At "main link" and "Standard link 2" revolute joint. Attached to "main link" along x axis. Frame 9 is also 2*X distance from frame 4 along x axis.

Denavit-Hartenberg Parameters for the Supporting chain of the proposed robot is given in the table 2.

D-H Parameters for Supporting Chain

Table 2: Frame parameters of Supporting Chain

	α (along axis)	a (along axis)	d (along axis)	θ (along axis)
Frame 0 → Frame 1	0	L0 (z axis)	0	q1 (z axis)
Frame 1 → Frame 6	+90 (x axis)	L1 (z axis)	d6 (x axis)	0
Frame 6 → Frame 7	0	0	0	q7 (z axis)
Frame 7 → Frame 8	0	xn (x axis)	0	q8 (z axis)
Frame 8 → Frame 9	0	X (x axis)	0	q9 (z axis)

Therefore,

$$^{0}[T]_{4} = {}^{0}[T]_{1} \times {}^{1}[T]_{6} \times {}^{6}[T]_{7} \times {}^{7}[T]_{8} \times {}^{8}[T]_{9} \quad (5.1)$$

$$^{0}[T]_{4} = \begin{bmatrix} \omega s(1,1) & \omega s(1,2) & \omega s(1,3) & xs \\ \omega s(2,1) & \omega s(2,2) & \omega s(2,3) & ys \\ \omega s(3,1) & \omega s(3,2) & \omega s(3,3) & zs \\ 0 & 0 & 0 & 1 \end{bmatrix} (5.2)$$

The values of the above terms in matrix are:

$$\omega s(1,1) = cos(q7 + q8 + q9) * cos(q1)$$
$$\omega s(1,2) = -sin(q7 + q8 + q9) * cos(q1)$$
$$\omega s(1,3) = sin(q1)$$
$$\omega s(2,1) = cos(q7 + q8 + q9) * sin(q1)$$
$$\omega s(2,2) = -sin(q7 + q8 + q9) * sin(q1)$$
$$\omega s(2,3) = -cos(q1)$$
$$\omega s(3,1) = sin(q7 + q8 + q9)$$
$$\omega s(3,2) = cos(q7 + q8 + q9)$$
$$\omega s(3,3) = 0$$
$$xs = cos(q1) * (d6 + x * cos(q7 + q8) + xn * cos(q7))$$
$$ys = sin(q1) * (d6 + x * cos(q7 + q8) + xn * cos(q7))$$
$$zs = L0 + L1 + x * sin(q7 + q8) + xn * sin(q7)$$

V. Inverse Kinematics

The inverse kinematics is used to get required joint parameters for reaching an end effector configuration in cartesian space. So, we first need to figure out which variables in our forward kinematic matrices are constant and which one of them are free to manipulate. The list for variables and constant for the proposed robot design is given below.

I. For Main chain and manipulator

Table 3: Known and Unknown parameters of Main Chain

Known Constants	Unknown Variables
L0	q1
L1	q3
X	q4
L4	q5
L5	α5
d2	

II. For Supporting Chain

Table 4: Known and Unknown parameters of Supporting Chain

Known Constants	Unknown Variables
L0	q1
L1	xn (Prismatic joint)
X	q7 (in extended configuration)
d6	q8
q7 (in parallelogram configuration)	q9

During solving inverse kinematics for an orientation, the end effector configuration to be reached via end effector in cartesian space will also be known in the form of a 4x4 matrix. Let us assume a 4x4 orientation matrix that has to be reached via end effector.

$$[T]given = \begin{bmatrix} qxx & qxy & qxz & px \\ qyx & qyy & qyz & qy \\ qzx & qzy & qzz & rz \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5.3)$$

We assume every term in the above matrix is known and is given as an input to solve inverse kinematic equations. We will use these terms to derive solutions for each unknown variable that are being listed in above tables.

Solution for Main chain with manipulator

In this part the values of ***q1, q3, q4, q5, α5*** are needed by the robot to reach required end-effector position and orientation.

By comparing ^{0}**[T]**$_{\{Tool\}}$ and **[T]given**.

^{0}***[T]***$_{\{Tool\}}$ = ***[T]given***

By equating these two matrices we get twelve equations. From these twelve equations we get,

$$\boldsymbol{q1 = atan2(qy, px)} \quad (5.4)$$

$$\boldsymbol{\alpha 5 = atan2(-qzz, qzy)} \quad (5.5)$$

Now since the value of ***α5*** is known and the value of L5 is already known. Therefore, the value of ^{5}**[T]**$_{\{Tool\}}$is given by $\begin{bmatrix} 1 & 0 & 0 & L5 \\ 0 & cos(\alpha 5) & -sin(\alpha 5) & 0 \\ 0 & sin(\alpha 5) & cos(\alpha 5) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$. With the help of ^{5}**[T]**$_{\{Tool\}}$and equation ^{0}[T]$_{\{Tool\}}$ = ^{0}[T]$_{5\text{-main}}$ x ^{5}[T]$_{\{Tool\}}$ we have,

$[^{0}[T]_{5\text{-main}}]given$ = $[T]given$ x $[^{5}[T]_{\{Tool\}}]^{-1}$ (5.6)

Here ***[^{0}[T]$_{5\text{-main}}$]given*** gives the required orientation of main link frame 5. Now, since the value of ***[^{0}[T]$_{5\text{-main}}$]given*** is known for a required end-effector position and orientation. Let its value be,

$$[^{0}[T]_{5\text{-main}}]given = \begin{bmatrix} q5xx & q5xy & q5xz & p5x \\ q5yx & q5yy & q5yz & q5y \\ q5zx & q5zy & q5zz & r5z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5.7)$$

By comparing [^{0}***[T]***$_{5\text{-main}}$***]given*** and ^{0}**[T]**$_{5\text{-main}}$ we get

$$\mathbf{q3 = \sin^{-1}\left(\frac{d - a*sin(\epsilon)}{b}\right)} \quad (5.8)$$

$$\boldsymbol{q4 = \epsilon - q3} \quad (5.9)$$

$$\boldsymbol{q5 = \sin^{-1}(q5zx) - q3 - q4} \quad (5.10)$$

Here in equation 5.8, 5.9 and 5.10

$$a = L4 + 2 * X \quad (5.11)$$

$$b = X \quad (5.12)$$

$$c = \left(\frac{q5y}{\sin(q1)}\right) \quad or \quad c = \left(\frac{p5x}{\cos(q1)}\right) \quad (5.13)$$

$$d = r5z \quad (5.14)$$

$$\epsilon = \cos^{-1}\left(\frac{a^2 + c^2 + d^2 - b^2}{2 * a * \sqrt{(d^2 + c^2)}}\right) + atan2(d, c) \quad (5.15)$$

Solution for Supporting chain

This robot is being designed to work in two main configurations as already being described in Design section. They are distorted parallelogram configuration and supported extended configuration. In distorted parallelogram configuration angle q7 is zero. The main advantage of this configuration is secondary support from ground surface via "cylinder link" extrusion. Due to this extra support the robot can lift heavier weight, while at the same time being very stable. But the major drawback of this configuration is that it cannot be used above certain height (maximum reachable height = L0 + L1 + X + L9 + L5). One such typical distorted parallelogram configuration is shown in figure 6.

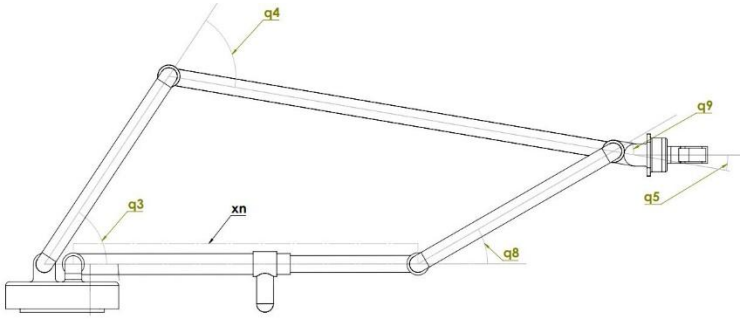


Figure 6: Parameters in Parallelogram (distorted) configuration

In this configuration since q7 = 0, therefore the required parameters to reach the end effector configuration is xn, q8, q9. Some of the parameters to reach end-effector configuration is already known from inverse kinematic solution of main chain and manipulator. Those parameters are q1, q3, q4, q5, α5.

Another main configuration in which this robot can work is supported extended configuration. In this configuration the "main link" of the robot is supported by "standard link 2" of the supporting chain. The mid beam of "standard link 2" becomes tangential to "main link" beams. In this configuration angle q9 becomes constant and can be very easily determined by geometry of the robot. This configuration is designed to help robot reach higher end-effector configurations and also increase the maximum liftable load which is one of the main aim of this project. An arbitrary supported extended configuration is shown as

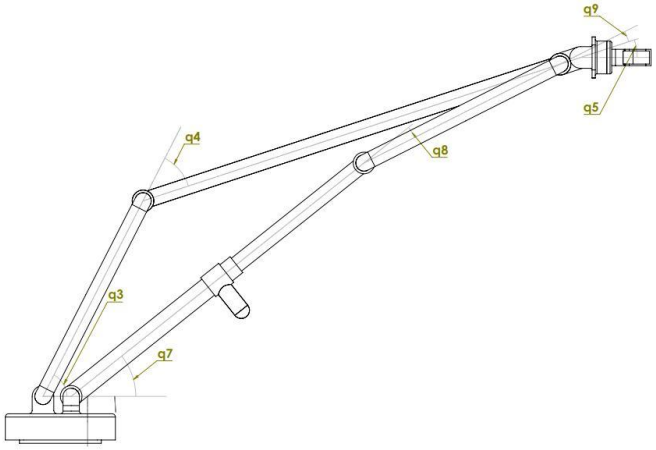


Figure 7: Parameters in Parallelogram extended configuration

In this configuration the required unknown parameters are q7, q8 and length xn (piston-cylinder length). While the known parameters form solution of main link and manipulator inverse kinematics are q1, q3, q4, q5, α5.

The inverse kinematic of the supporting chain is derived by geometric method. While deriving inverse kinematic of supporting chain parameters q1, q3, q4, q5 and α5 are taken as input with other known constants of the robot as described above in table 1 and 3. In this derivation parameters q7, q8, q9 and xn are needed to be derived using the listed known constants and derived parameters in main chain inverse kinematic section as described above. The inverse kinematic is needed to be derived for both the extended and parallelogram cases.

We need the values of frame 9 and frame 6 to find the inverse kinematics of secondary chain using geometry. The value of frame 9 with respect to ground can be found as

$$^{0}[T]_{9} = {}^{0}[T]_{1} \times {}^{1}[T]_{2} \times {}^{2}[T]_{3} \times {}^{3}[T]_{4} \times {}^{4}[T]_{9} \quad (5.16)$$

$$^{0}[T]_{9} = \begin{bmatrix} \omega9(1,1) & \omega9(1,2) & \omega9(1,3) & x9 \\ \omega9(2,1) & \omega9(2,2) & \omega9(2,3) & y9 \\ \omega9(3,1) & \omega9(3,2) & \omega9(3,3) & z9 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5.17)$$

The value of the above matrix is completely known in numeric terms. The values of q1, q3, q4, q5, α5 are already derived. And other known constants which are used to get $^{0}[T]_{9}$ are listed in table 1**.** The another required matrix is $^{0}[T]_{6}$**.** And it can be derived as

$$^{0}[T]_{6} = {}^{0}[T]_{1} \times {}^{1}[T]_{6}$$

$$^{0}[T]_{6} = \begin{bmatrix} \omega6(1,1) & \omega6(1,2) & \omega6(1,3) & x6 \\ \omega6(2,1) & \omega6(2,2) & \omega6(2,3) & y6 \\ \omega6(3,1) & \omega6(3,2) & \omega6(3,3) & z6 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (5.18)$$

Value of above matrix is also completely known as the value of q1 is the only variable required to derive $^{0}[T]_{6}$. The other constants used to derive $^{0}[T]_{6}$ are listed in table 2.

Distance between $^{0}[T]_{6}$ and $^{0}[T]_{9}$ is given as

$$(g, h, i) = (x9 - x6, y9 - y6, z9 - z6)$$

$$horizontal\ distance = k = \sqrt{g^2 + h^2}$$

$$vertical\ distance = i$$

i. Case of Distorted Parallelogram (q7 = 0)

In this configuration q7 is zero. Therefore, q8, q9 and xn are needed to be derived. By geometry of the proposed robot design

$$\boldsymbol{xn = \frac{2 * n * cos(\gamma) \pm \sqrt{(2 * n * \cos(\gamma))^2 - 4 * (n^2 - x^2)}}{2}} \quad (5.19)$$

In equation 5.19

$$\boldsymbol{xn \in [\left(\sqrt{5} - 1\right) * x - d6 - d2, 2 * x - d2 - d6]}$$

and

$$\boldsymbol{q8 = \cos^{-1}\left(\frac{n^2 - (xn)^2 - x^2}{2 * xn * x}\right)} \quad (5.20)$$

$$\boldsymbol{q9 = \varepsilon2 + \varepsilon3 + \varepsilon1} \quad (5.21)$$

In the equation 5.19, 5.20 and 5.21

$$m = k + d2 + d6 \quad (5.22)$$

$$n = \sqrt{k^2 + i^2} \quad (5.23)$$

$$u = \sqrt{m^2 + i^2} \quad (5.24)$$

$$\gamma = \cos^{-1}\left(\frac{k}{\sqrt{k^2 + i^2}}\right) = \sin^{-1}\left(\frac{i}{\sqrt{k^2 + i^2}}\right) \quad (5.25)$$

$$\varepsilon1 = \cos^{-1}\left(\frac{3 * x^2 + u^2}{4 * x * u}\right) \quad (5.26)$$

$$\varepsilon2 = \cos^{-1}\left(\frac{n^2 + u^2 - (d6 + d2)^2}{2 * u * n}\right) \quad (5.27)$$

$$\varepsilon3 = \cos^{-1}\left(\frac{n^2 + x^2 - (xn)^2}{2 * x * n}\right) \quad (5.28)$$

ii. Case of Extended configuration (q9 = known constant)

In this case q9 will be used as a known constant. While q7, q8, xn are needed to be derived. By geometry of the proposed robot design

$$\boldsymbol{xn = \sqrt{x^2 + n^2 - 2 * x * n * \cos(\varepsilon5)}} \quad (5.29)$$

Here $\boldsymbol{xn} \in [\,(\sqrt{\mathbf{5}}\ \mathbf{-1}) * \boldsymbol{x} - \boldsymbol{d6} - \boldsymbol{d2}, \mathbf{2} * \boldsymbol{x} - \boldsymbol{d2} - \boldsymbol{d6}]$

$$\boldsymbol{q7} = \boldsymbol{\gamma} + \boldsymbol{\varepsilon 4} \quad (5.30)$$

$$\boldsymbol{q8} = \boldsymbol{\varepsilon 4} + \boldsymbol{\varepsilon 5} \quad (5.31)$$

Here in the above equations

$$\gamma = \cos^{-1}\left(\frac{k}{\sqrt{k^2 + i^2}}\right) = \sin^{-1}\left(\frac{i}{\sqrt{k^2 + i^2}}\right) \quad (5.32)$$

$$\varphi = \cos^{-1}\left(\frac{m}{\sqrt{m^2 + i^2}}\right) - \varepsilon 1 \quad (5.33)$$

$$\varepsilon 5 = \gamma - q9 - \varphi \quad (5.34)$$

$$\varepsilon 4 = \cos^{-1}\left(\frac{(xn)^2 + n^2 - x^2}{2 * xn * n}\right) \quad (5.35)$$

k, i, m, n, ε1 from previous "distorted parallelogram case (case I)"

VI. Structural Analysis

In structural analysis two aspects of the robot are needed to be studied. First part is static structural analysis of robot in which structural strength of robot in various configurations are calculated. And in another partthe required acceleration or force required to manipulate the robot is being calculated. This can be very helpful in determining the required actuator strength of the robot.

For this analysis of the robot, we need to assign values to the constant parameters of the robot. The assumed values and specifications are listed below in table 5, table 6 and table 7.

Table 5: Assigned ParametersValues

Parameters	Assigned Values
L0	10 mm
L1	133 mm
X	670 mm
L4	66 mm
L5	95 mm
d2	53 mm
d6	33 mm
L0	66 mm

Table 6: Link Properties

Link	Weight (in Kg)	Density (Kg/m^3)	Material
Ground Plate	9.221	7850	AISI 4130 Steel
Base Plate	24.622	7850	AISI 4130 Steel
Standard Link 1 & 2	12.323	7850	AISI 4130 Steel
Main Link	22.124	7850	AISI 4130 Steel
Roll Plate	2.122	2700	Aluminium Alloy 6061-T6 (SS)
Gripper Plate	1.741	2700	Aluminium Alloy 6061-T6 (SS)
Gripper Jaw	1.208	2700	Aluminium Alloy 6061-T6 (SS) with Silicon Rubber friction Plate
Cylinder Link	8.370	7850	AISI 4130 Steel
Piston Link	5.748	7850	AISI 4130 Steel

Table 7: Material Properties for Structural Analysis

Material → **Properties** ↴	**AISI 4130 Steel**	**Aluminium Alloy 6061-T6 (SS)**	**Silicon Rubber**
Youngs Modulus (in GPa)	205	68.9	0.05
Poisson's Ratio	0.29	0.33	0.47
Tensile Yield strength (in MPa)	435	276	5.5
Tensile Ultimate Strength (in MPa)	670	310	---

With the help of table 6 we can figure out that the total weight of the robot is 101.03 kg.These parameters can be changed for further optimization of the design. Thus, our proposed robot can be scaled up or scaled down according to requirements.

Static Structural Analysis

Static structural analysis is the first analysis that is required to test the structural strength of the robot in both of the configurations. The robot was analysed in both the configuration with varying loads. In the analysis we have found a sweet spot for maximum lifting capacity of the robot. **And that turns out to be 150 kg**. Figure 8 and 9 shows application of 150 kg load in vertically downward direction on gripper jaws in both the cases on the gripper jaws.

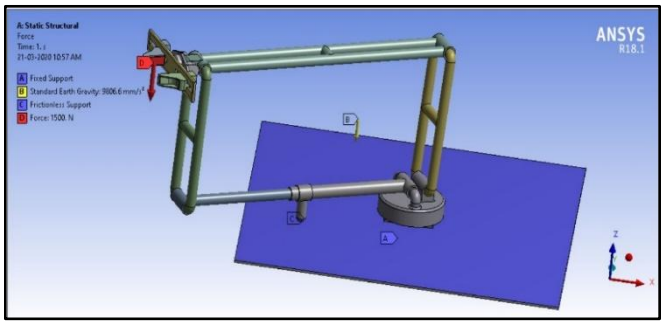


Figure 8: Application of load in Parallelogram Configuration

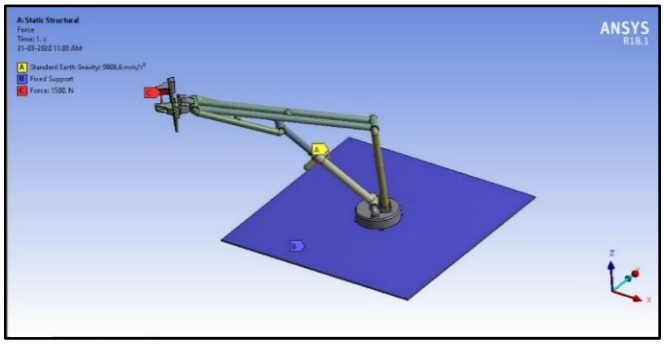


Figure 9: Application of load in Extended Configuration

The result of the above analysis is shown in following sub-sections.

a). Total Deformation

Maximum Deformation in parallelogram position is minimal approx. 4mm at gripper teeth.

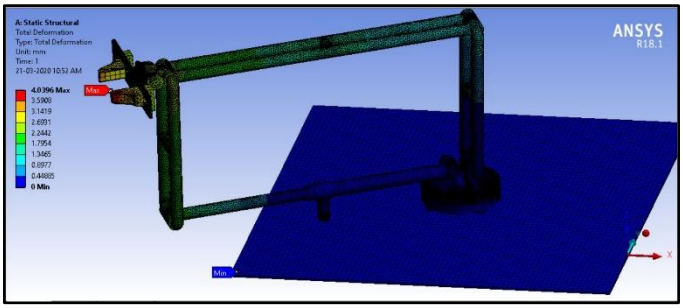


Figure 10

In second position i.e. when main link is supported on standard link 2. This is shown in figure 11.

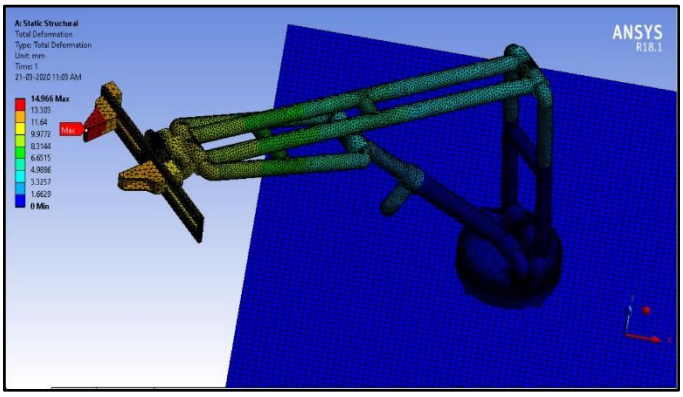


Figure 11

b). Equivalent Stresses

In both position 1 maximum stress generate at linkage of piston and cylinder i.e. 89.883 MPa which is far less than yield strength of material used i.e. AISI 4130 Steel (435 MPa). Maximum stress at this point due to support it provide to lift weight and act like fixed point of cantilever beam.

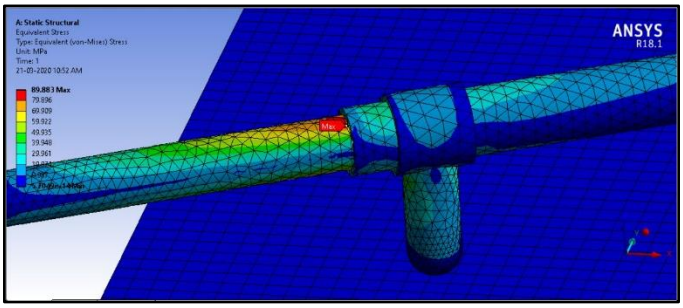


Figure 12

In second position maximum equivalent stress generated is at joint of baseplate and standard link i.e. 228.73 MPa. Majority of stress is compressive stress due to self-weight and load applied.

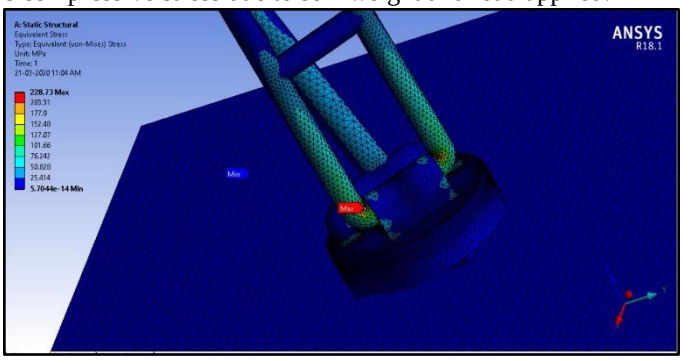


Figure 13

c). Maximum Principal Stresses

The result for maximum and minimum principal stresses are shown in image 14 and image 15 for both the configurations.

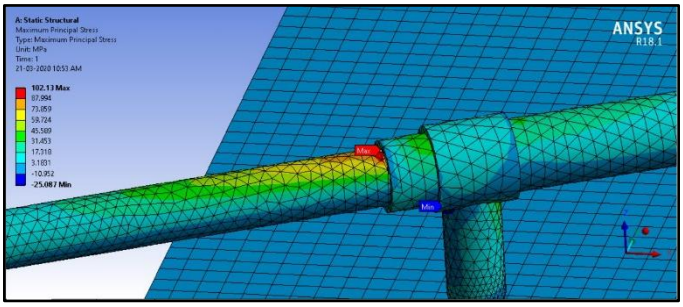


Figure 14

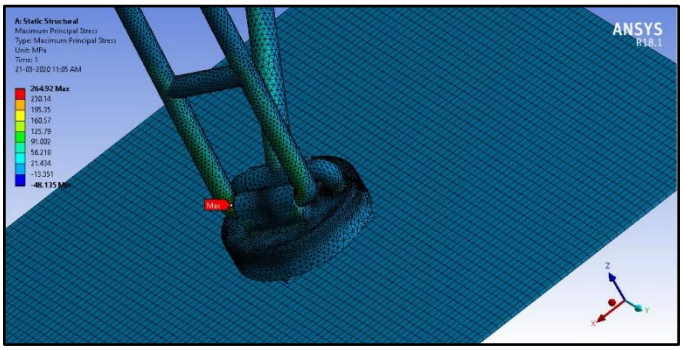


Figure 15

d). Factor of safety

The factor of safety for a link is the ratio of ultimate strength of material to the actual working stress or maximum permissible strength when in use. The factor of safety is calculated for the purpose to know whether structure is able to bear 150kg load without failure or not. The results of factor of safety looks quite promising in

this scenario. In parallelogram configuration the minimum factor of safety is 4.83 while in more bizarre extended configuration the minimum factor of safety is 1.90. The results are being shown in image 16 and image 17.

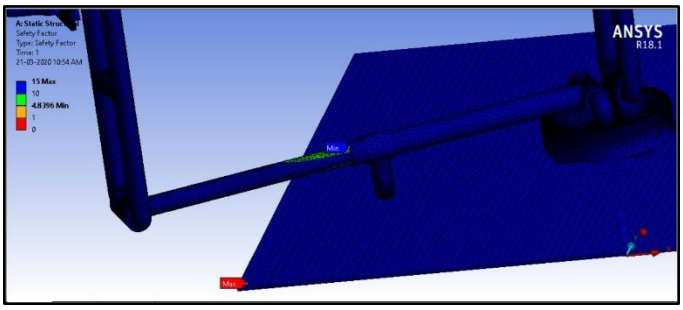


Figure 16

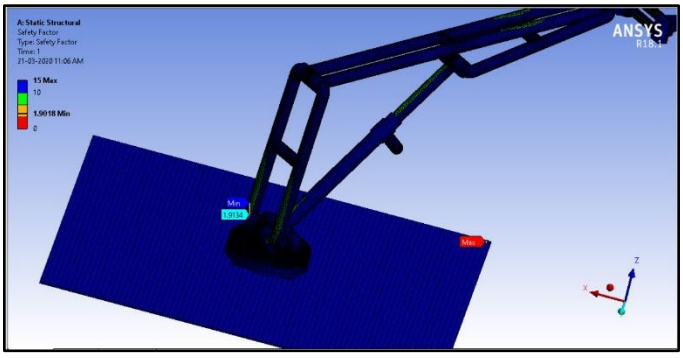


Figure 17

Actuator Motion Analysis

In this analysis we aim to study required velocity and acceleration of each joint to move from one end-effector configuration to another. During this analysis we calculated required velocity and acceleration for empty robot with no load in normal gravity of 9.80665 m/s^2. The initial configuration of the robot is parallelogram configuration and the final configuration is extended configuration. Both of the starting and final configuration with a mid-configuration is shown in image 18, 19 and 20. The entire movement takes 10 second to complete. In this analysis secondary support for the robot has been removed.This is achieved by removing extrusion from cylinder link. The weight of the cylinder link is adjusted to original weight by increasing its density in simulation environment. The support has been removed to calculate the actuator velocity, acceleration and force without any secondary support.Thus, the required actuator force to move the robot from one configuration to another will usually be less than the result of this analysis in zero load configuration.

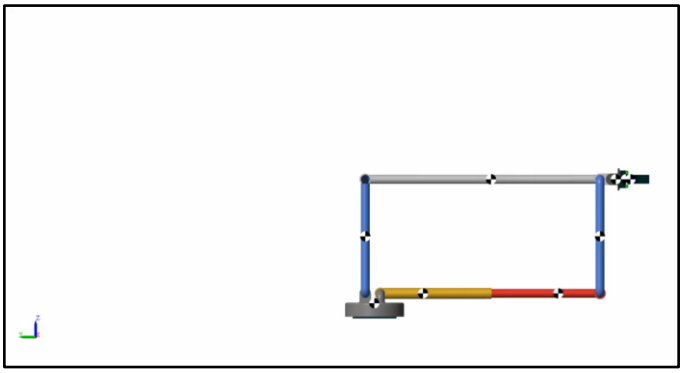


Figure 18: Starting Configuration at time 0.0 seconds

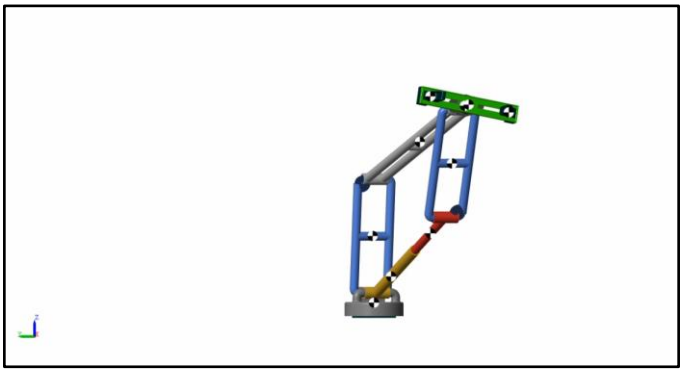


Figure 19: Mid Configuration at 5.00 seconds

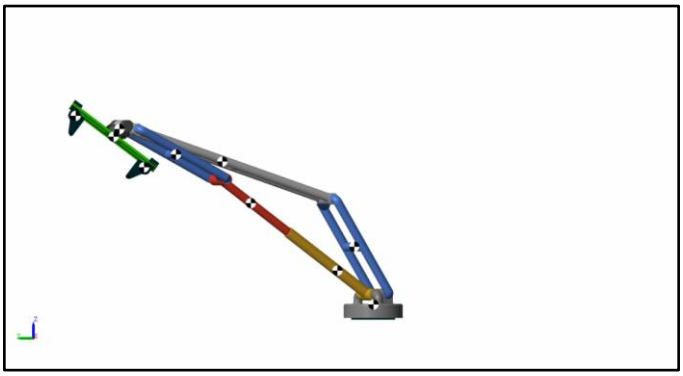


Figure 20: Final Configuration at 10.00 seconds

The velocity acceleration and position graph of the actuators over the entire timeline from zero to tenth seconds is shown in figure 21 to figure 29. The force and torque required for each actuator can be calculated by multiplying acceleration to manipulated weight (body weight with load on robot). Thus, following graphs can be used to select or design the actuators for the proposed robot in manufacturing phase. The robot has nine actuators excluding gripper jaws actuation system. Eight of these actuators are used to manipulate rotary joints. While only one very large hydraulic actuator (in figure 26) is used to manipulate prismatic hydraulic joint between piston and cylinder link in supporting chain.

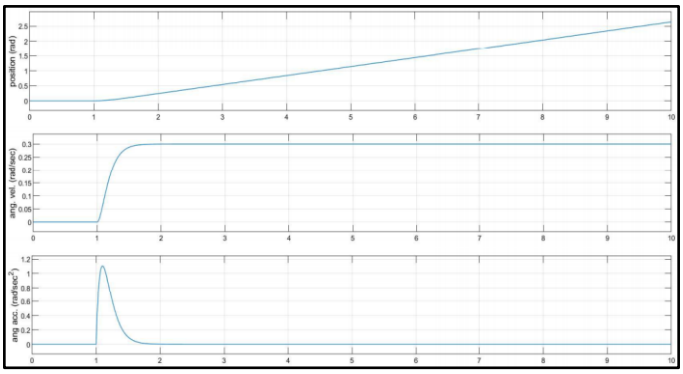


Figure 21: Ground and base plate Joint at frame 1

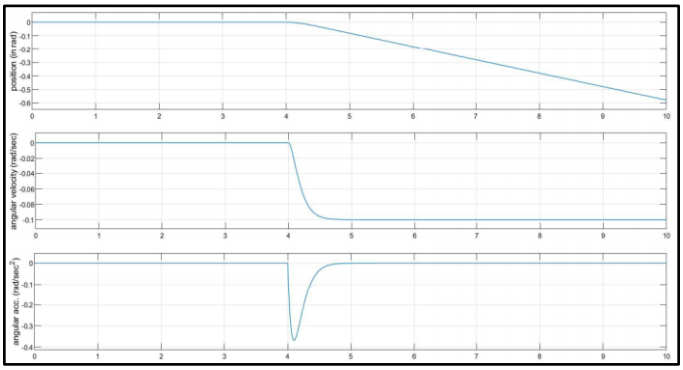


Figure 22: Base Plate and Standard Link 1 joint at frame 3

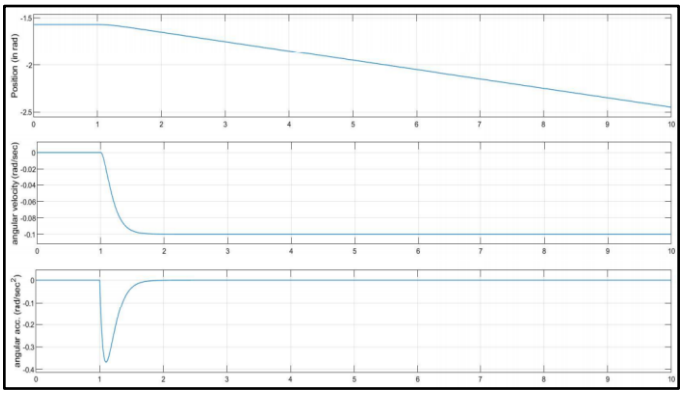


Figure 23: Standard link 1 and Main link joint at frame 4

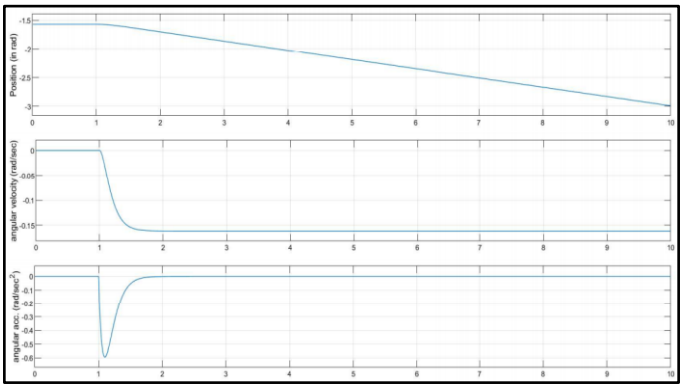


Figure 24: Main Link and Standard Link 2 joint at frame 9

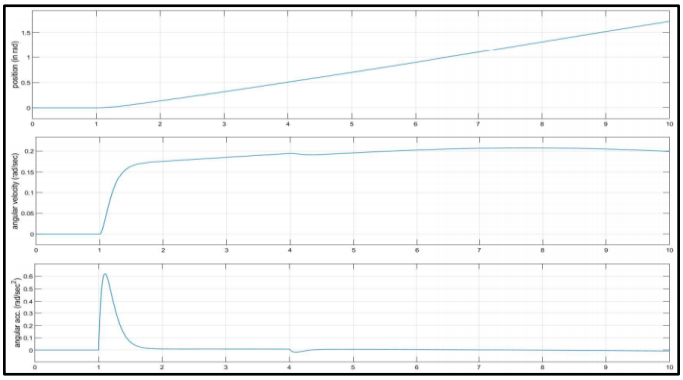


Figure 25: Standard Link 2 and Piston link joint at frame 8

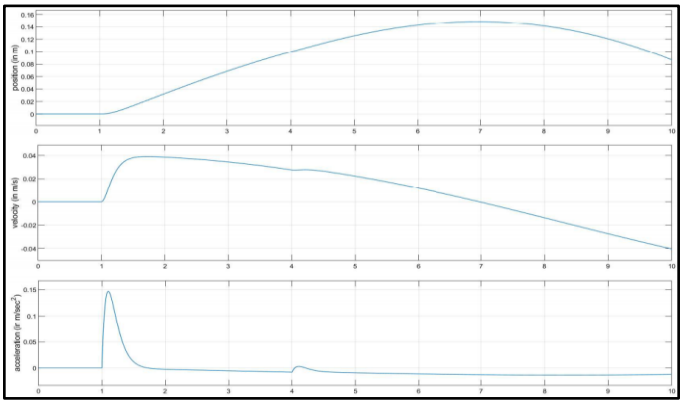


Figure 26: Hydraulic Joint between Piston and Cylinder Link

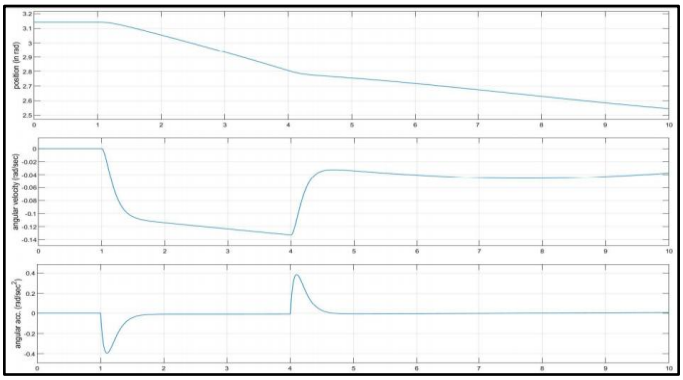


Figure 27: Cylinder Link and Base Plate joint at frame 7

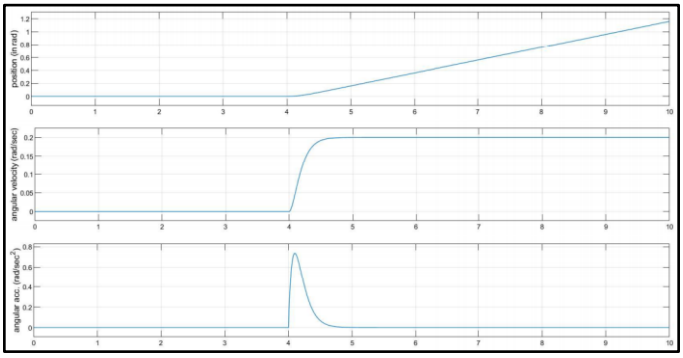


Figure 28: Main Link and Roll Plate joint at frame 5

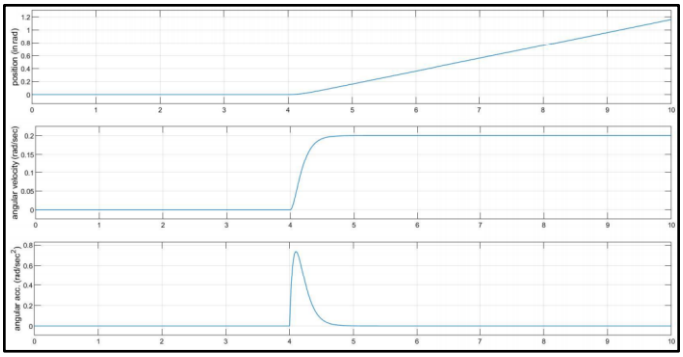


Figure 29: Roll Plate and Gripper Plate joint at frame {Tool}

VII. Conclusion

a) Through this robot design we have tried to counter many problems of normal serial link manipulators. Our proposed design is strong enough to lift 150 kg of weight while its own body is of 101.03 kg thus achieving a maximum lifting capacity to body weight ratio of more than one.

b) The proposed robot also has easy manipulation as shown in inverse kinematics section. The inverse kinematics of this robot can be achieved by analytical and geometrical methods. Which makes it very easy for a controller to control.

c) The required actuator strength for most of the actuators in this robot is also low. But the number of required actuators is more. Thus, one can say that a greater number of low strength actuator is required to manipulate this robot.

We want to present this robot as a future platform for development of parallel link manipulators. We call this proposed robot design as "Parallelo Arm Robot". Many future developments are required for this robot to become a reality. We as a group would be happy to see future work being done on this robot platform.

Refrences

[1]. Griffis, M. and Duffy, J. (1989), A forward displacement analysis of a class of stewart platforms. J. Robotic Syst., 6: 703-720. doi:10.1002/rob.4620060604

[2]. Zhang, X., Mills, J.K. & Cleghorn, W.L. Dynamic Modeling and Experimental Validation of a 3-PRR Parallel Manipulator with Flexible Intermediate Links. *J Intell Robot Syst* **50,** 323–340 (2007). https://doi.org/10.1007/s10846-007-9167-4

[3]. P. J. Parikh and S. S. Y. Lam, "A hybrid strategy to solve the forward kinematics problem in parallel manipulators," in IEEE Transactions on Robotics, vol. 21, no. 1, pp. 18-25, Feb. 2005.

[4]. Tsai LW. (1996) Kinematics of A Three-Dof Platform with Three Extensible Limbs. In: Lenarčič J., Parenti-Castelli V. (eds) Recent Advances in Robot Kinematics. Springer, Dordrecht

[5]. K. Cleary and T. Brooks, "Kinematic analysis of a novel 6-DOF parallel manipulator," [1993] Proceedings IEEE International Conference on Robotics and Automation, Atlanta, GA, USA, 1993, pp. 708-713 vol.1.

[6]. Siciliano, B. (1999). The Tricept robot: Inverse kinematics, manipulability analysis and closed-loop direct kinematics algorithm. *Robotica, 17*(4), 437-445. doi:10.1017/S0263574799001678

[7]. Angeles, J. (August 12, 2004). "The Qualitative Synthesis of Parallel Manipulators ." ASME. *J. Mech. Des.* July 2004; 126(4): 617–624. https://doi.org/10.1115/1.1667955

Rohit Kumar,etal. "Design and Analysis of a Parallel Link Manipulator for Heavy Lifting in Warehouses." *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE),* 17(3), 2020, pp. 07-24.

Design and Implementation of Biquad Filter for Shunt Compensation under Normal and Distorted Grid Conditions

Ankita Arora
Dept. of Electrical Engineering
Delhi Technological University
Delhi,India
arora.ankita09@gmail.com

Alka Singh
Dept. of Electrical Engineering
Delhi Technological University
Delhi,India
alkasingh.eed@gmail.com

***Abstract*— In this paper a Biquad Filter called Universal Filter is designed for reducing power quality-based problems in power distribution systems. Increased use of solid-state power electronic based devices and loads requires shunt active power compensation techniques for mitigation. The Biquad Active Filter is a second order filter offering all four basic filters viz high pass, band pass, low pass, notch filter. In this paper bandpass and low pass filter are of prime importance as they are used for estimating the peak of the load current. It can be tuned by varying the quality factor and center frequency. The compensation technique is designed in MATLAB environment and its performance is experimentally verified using dSPACE- 1104.**

Keywords— Biquad Active Filter, power quality, DSTATCOM, shunt compensation

I. INTRODUCTION

Power quality problems are becoming a major challenge these days due to increased use of solid-state power electronic based devices in power transmission and distribution systems. Use of such devices in power system results in less bulky and less costly networks which increase flexibility, reduce maintenance but at the same time impose problems in the grid connected systems. The power quality issues consist of injecting harmonics in the supply current leading to deterioration of power factor, voltage fluctuations and excessive heating of the equipment, increased losses etc [1].

Various filtering techniques consisting of passive and active filtering techniques have been devised but the passive filtering technique has become relatively less relevant these days due to their inability to produce satisfactory results under varying system conditions, large size, selecting the right sizes of filters is necessary to avoid causing resonance with the source [2].

The shunt power filter popularly called DSTATCOM modelled as Voltage Source Converter (VSC) with a capacitor connected in parallel is used as self-supporting dc bus to provide mitigation of PQ problems.

Many control techniques have been studied in literature such as d-q reference frame theory, Instantaneous Reactive power theory (IRPT), power balance theory and now days the complex neural networks and adaptive filter has been used for shunt power Compensation [3]. The neural network requires a lot of training data for giving good performance but still suffer from challenges of overfitting and saturation problems. These algorithms also do not work well in distorted grid conditions. Performance Realization of the DSTATCOM is dependent on the control algorithm that determines how precisely it extracts the component from the non-linear load current. Current algorithms in literature on shunt compensation nowadays are multiple fifth order Generalized Integrator Based Adaptive Control in [4], Kalman filter in [5], Adaptive Neuron based detection control in [6].

Here, the control scheme consists of a Biquad Active Filter used to generate the neutralizing current to cancel the harmonics under different loading conditions. The Biquad Active Filter is also called Universal Filter and the word Biquad is the short form of Biquadratic which means that its transfer function which is the ratio of numerator and denominator is quadratic in 's' [7-9]. It characterizes the features of all the four basic filters i.e. high pass, band pass, low pass and notch filter in one circuit Therefore it is also called Universal active filter. In this filter the band pass filter and low pass filter are of prime importance as they are used to compute the real power component of the load current. It is utilized for both as the PLL and for shunt compensation.

II. BIQUAD ACTIVE FILTER-DESIGN AND ANALYSIS

Biquad active filter is a Biquadatic Filter also called KHN Filter is an op_amp-based RC circuit shown in Fig. 1(a) that makes second order filter by cascading two integrator in a feedback loop. It is developed by Kerwin-Huelsman-Newcomb used to realize all the four fundamental filters namely high pass, low pass, band-pass and notch filter. The standard expression for current gain of the Biquad Active Filter (H(s)) [8] is given in standard form as

$$H(s) = \frac{ps^2 + qs + r}{s^2 + a_1 s + a_0}$$

$$\begin{cases} p = q = 0\,, low\ pass\ filter \\ q = r = 0, high\ pass\ filter \\ p = r = 0,\ \ band\ pass\ filter \\ q = -pa, r = pa_0, notch\ filter \end{cases}$$

The single line diagram in 's' domain is simulated in MATLAB 2017A in Fig. 1(b) is obtained from block diagram of Biquad Filter. Its analysis and derivation of the transfer function consisting of all four basic filters is discussed below.

978-1-7281-6664-3/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:08:19 UTC from IEEE Xplore. Restrictions apply.

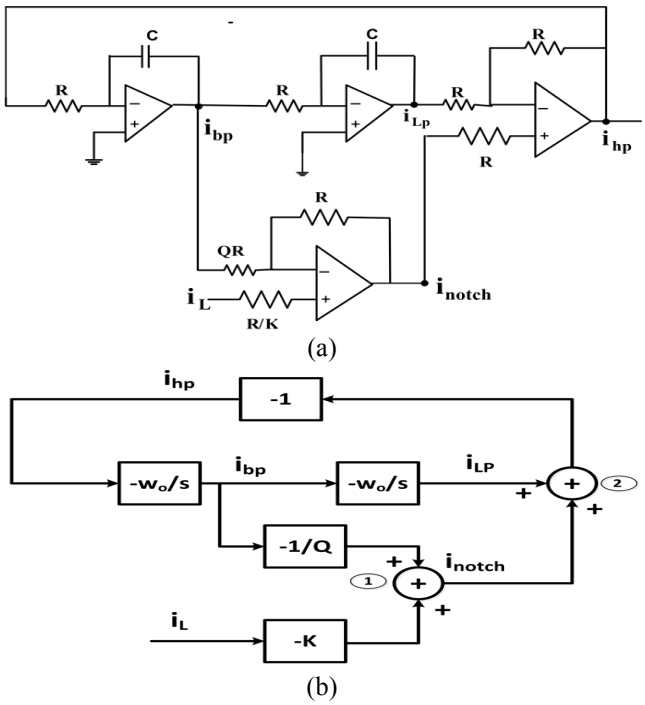


Fig. 1. (a) Biquad Filter Schematic Diagram (b) Single Line Diagram in 's' domain

Applying KCL at node 1 in Fig. 1(b) we get

$$-Ki_L - \frac{i_{bp}}{Q} = i_{notch} \tag{1}$$

$$i_{bp}\left(\frac{-w_o}{s}\right) = i_{LP} \tag{2}$$

$$i_{hp}\left(\frac{-w_o}{s}\right) = i_{bp} \tag{3}$$

Appling KCL at node 2

$$i_{Lp} + i_{notch} = -i_{hp} \tag{4}$$

For deriving transfer function of bandpass filter, we will put the value of i_{LP} from (2) in (4)

$$i_{Lp} + i_{notch} = -i_{hp} \tag{5}$$

$$i_{bp}\left(\frac{-w_o}{s}\right) + i_{notch} = \left(\frac{s}{w_o}\right) i_{bp} \tag{6}$$

$$i_{bp}\left(\frac{w_o^2+s^2}{sw_o}\right) = i_{notch} \tag{7}$$

Putting the value of i_{notch} in (1)

$$-Ki_L - \frac{i_{bp}}{Q} = i_{bp}\left(\frac{w_o^2+s^2}{sw_o}\right) \tag{8}$$

$$-Ki_L = \frac{i_{bp}}{Q} + i_{bp}\left(\frac{w_o^2+s^2}{sw_o}\right) \tag{9}$$

Rearranging (9) we get

$$\frac{i_{bp}}{i_L} = \frac{-Ksw_o}{s^2+\left(\frac{w_o}{Q}\right)s+w_O^2} \tag{10}$$

where i_L represents the input load current, i_{bp} represents the output current of band pass filter, w_o represents the natural frequency of oscillations and Q is the quality factor.

The bode diagram of the bandpass filter is illustrated in Fig. 2.

For low pass filter we get (11) by applying KCL at node 2

$$i_{Lp} + i_{notch} = -i_{hp} \tag{11}$$

$$i_{Lp} + i_{notch} = \left(\frac{-s}{w_o}\right) i_{bp} \tag{12}$$

$$i_{Lp} + i_{notch} = \left(\frac{-s}{w_o}\right)\left(\frac{-s}{w_o}\right) i_{Lp} \tag{13}$$

$$i_{Lp} + \left(\frac{s}{w_o}\right)\left(\frac{s}{w_o}\right) i_{Lp} = -i_{notch} \tag{14}$$

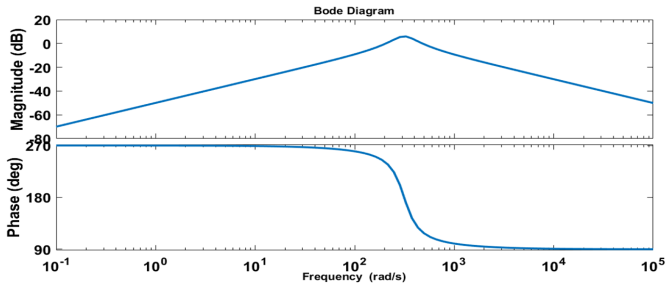


Fig. 2. Bode Plot of Bandpass Filter

Put the value of i_{notch} from (14) in (1), we get the transfer function of low pass filter.

$$-Ki_L - \frac{1}{Q}\left(i_{Lp}\left(\frac{-s}{W_o}\right)\right) = -\left[i_{Lp} + \left(\frac{s}{w_o}\right)\left(\frac{s}{w_o}\right) i_{Lp}\right] \tag{15}$$

$$-Ki_L + i_{Lp}\left(\frac{s}{W_oQ}\right) = -\left[i_{Lp} + \left(\frac{s}{w_o}\right)\left(\frac{s}{w_o}\right) i_{Lp}\right] \tag{16}$$

$$-Ki_L = -\left[i_{Lp} + \left(\frac{s}{w_o}\right)\left(\frac{s}{w_o}\right) i_{Lp} + \left(\frac{s}{W_oQ}\right)\right] \tag{17}$$

$$\frac{i_{Lp}}{i_L} = \frac{Kw_O^2}{s^2+\left(\frac{w_O}{Q}\right)s+W_O^2} \tag{18}$$

The bode diagram of the low pass filter is illustrated in Fig. 3 below.

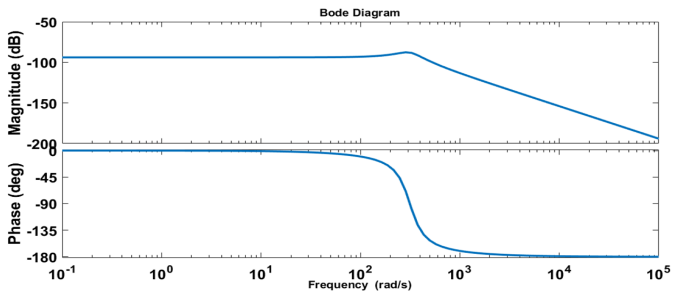


Fig. 3. Bode Plot of Low Pass Filter

For high pass filter we have (19) from node 2

$$i_{Lp} + i_{notch} = -i_{hp} \tag{19}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:08:19 UTC from IEEE Xplore. Restrictions apply.

$$i_{bp}\left(\frac{-w_o}{s}\right)+i_{hp}=-i_{notch} \tag{20}$$

$$i_{hp}\left(\frac{-w_o}{s}\right)^2+i_{hp}=-i_{notch} \tag{21}$$

Put the value of i_{notch} from (21) in (1), we get the transfer function of high pass filter

$$-Ki_L-\left(\frac{-w_o}{s}\right)\frac{i_{hp}}{Q}=i_{notch} \tag{22}$$

$$Ki_L=i_{hp}(1+\frac{w_o^2}{s^2}+\frac{w_o}{sQ}) \tag{23}$$

$$\frac{i_{hp}}{i_L}=\frac{Ks^2}{s^2+\left(\frac{w_o}{Q}\right)s+w_o^2} \tag{24}$$

where i_{hp} represents the output current of high pass filter The bode diagram of the high pass filter is illustrated in Fig. 4 below.

Similarly, for deriving the transfer function of Notch filter from node 2 we get,

$$i_{Lp}+i_{notch}=-i_{hp} \tag{25}$$

$$\left(\frac{w_o}{s}\right)(i_{Lp}+i_{notch})=i_{bp} \tag{26}$$

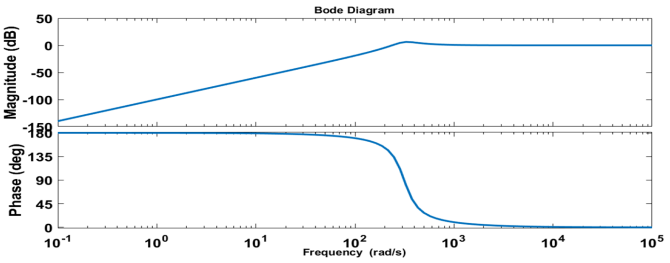


Fig. 4. Bode Plot diagram of High Pass Filter

$$\left(\frac{-w_o}{s}\right)i_{bp}+i_{notch}(\frac{-w_o}{s})=i_{bp} \tag{27}$$

$$i_{notch}=i_{bp}\frac{s^2+w_o^2}{sw_o} \tag{28}$$

Now put the value of i_{notch} from (28) in (1) we get

$$\frac{i_{notch}}{i_L}=\frac{-K(s^2+w_o^2)}{s^2+\left(\frac{w_o}{Q}\right)s+w_o^2} \tag{29}$$

where i_{notch} represents the output notch filter current. The bode diagram of the notch filter is illustrated in Fig. 5.

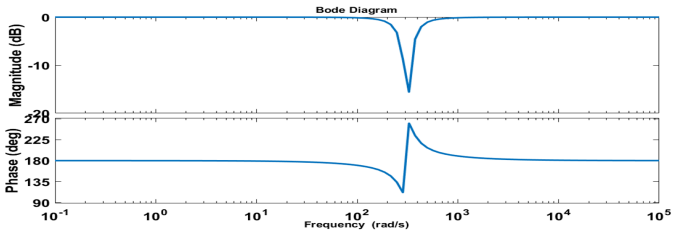


Fig. 5. Bode Plot diagram of Notch Filter

III. System Configuration

A single phase 110 V (rms) power distribution system connected to non-linear load is designed in our laboratory as demonstrated in Fig. 6. It injects harmonics in supply current which makes it non-sinusoidal/distorted. The Shunt active power filter-based algorithm called Biquad active Filter is used to cancel these harmonics in supply current.

IV. Biquad Active Filter Implementation

The implemented control technique based on Biquad Filtering scheme is illustrated in Fig. 7. The designing of the control technique is discussed below.

A. Generation of real power component of load current

The real power component of load current i_{fund} is calculated from the Biquad active filter by the equation shown below

$$i_{fund}=\sqrt{i_{bp}^2+i_{Lp}^2} \tag{30}$$

B. Estimation of switching losses

Here the comparison between the actual dc link voltage and the reference dc link voltage is done. The error obtained is provided to the PI controller which increases the current drawn from the grid to make the dc link voltage regulated to dc link reference value. This increased current is used to meet the switching losses (i_{loss}) of VSC.

$$i_{loss}=i_{loss}(k-1)+K_p\big(V_{dc}-V_{dc}(k-1)\big)+K_iV_{dc} \tag{31}$$

The i_{total} in (32) gives the total power supplied by the grid.

$$i_{total}=i_{loss}+i_{fund} \tag{32}$$

C. Reference current generation and driving gate pulses for DSTATCOM

In abnormal grids the unit templates ($\sin\theta$) is produced by the technique discussed below. The Biquad Active Filter is applied to the distorted source voltage. The root mean square of the output of the v_{bp} and v_{Lp} is taken to compute the fundamental magnitude.

$$v_m=\sqrt{v_{bp}^2+v_{Lp}^2} \tag{33}$$

The unit templates are calculated by

$$u_{ta}=\frac{v_{bp}}{v_m}=sin\theta \tag{34}$$

The reference current is generated by multiplying the unit template with the i_{total} which is compared with the actual grid current to drive the gate pulses for the DSTATCOM.

$$i_p^*=i_{total}*u_{ta} \tag{35}$$

V. Simulink Results

The simulink results shows the performance of the proposed controller under normal and abnormal grid conditions

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:08:19 UTC from IEEE Xplore. Restrictions apply.

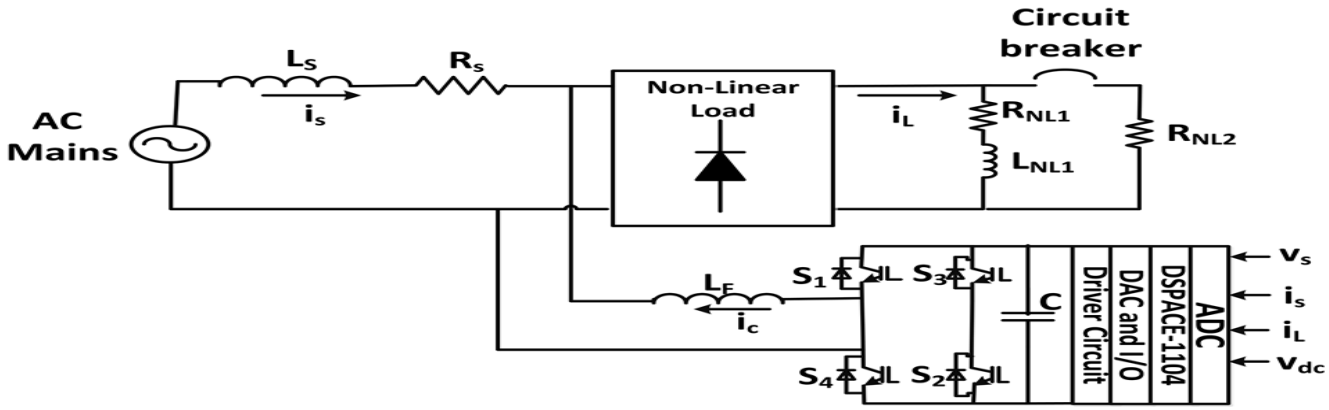


Fig. 6. Schematic Diagram of System

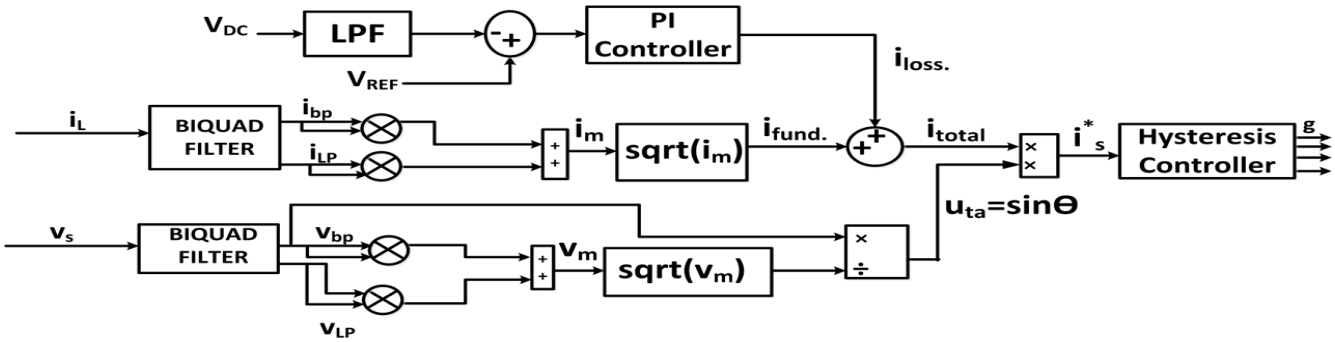


Fig. 7. Control Algorithm of Biquad Filtering Scheme

A. Performance analysis under normal and abnormal grid Conditions

The steady state and dynamic waveforms of grid voltage, current, load current, compensator current and dc voltage under closed loop conditions with DSTATCOM switched on is shown in Fig. 10 and Fig. 11. The grid current is found to be sinusoidal and in phase with the grid voltage under normal and distorted grid conditions.

The output of bandpass and low pass filter is observed in Fig. 8 and Fig. 9. The output of bandpass is found to be in phase with load current in shunt active power compensation and in phase with the distorted supply voltage in PLL whereas the output of low pass filter is 90º phase shifted with the output of bandpass filter in both shunt compensation and in PLL. The output of low pass filter is 90º phase shifted with the output of bandpass filter in both shunt compensation and in PLL.

The band pass and low pass filter output is used to calculate the real power component of the load current in shunt compensation and magnitude of supply voltage in PLL. PLL is used for generating the unit-in phase template for sinusoidal reference current.

The DC link voltage is regulated to 200 V reference by PI controller. Moreover, even under sudden load change from 0.2 to 0.3 secs, the system remains stable and the source current remains sinusoidal and balanced.

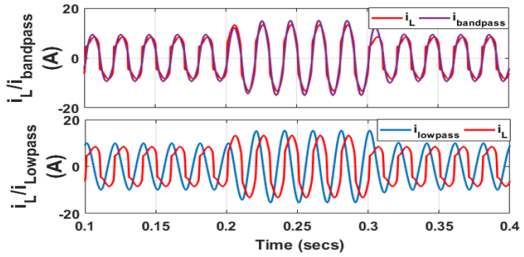


Fig. 8. Plot of Load Current, Bandpass current and low pass current

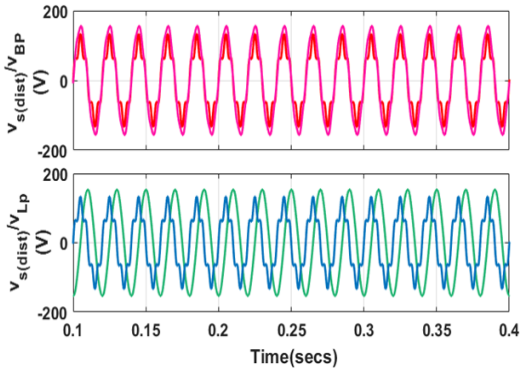


Fig. 9. Plot of distorted supply voltage, Bandpass voltage and low pass voltage

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:08:19 UTC from IEEE Xplore. Restrictions apply.

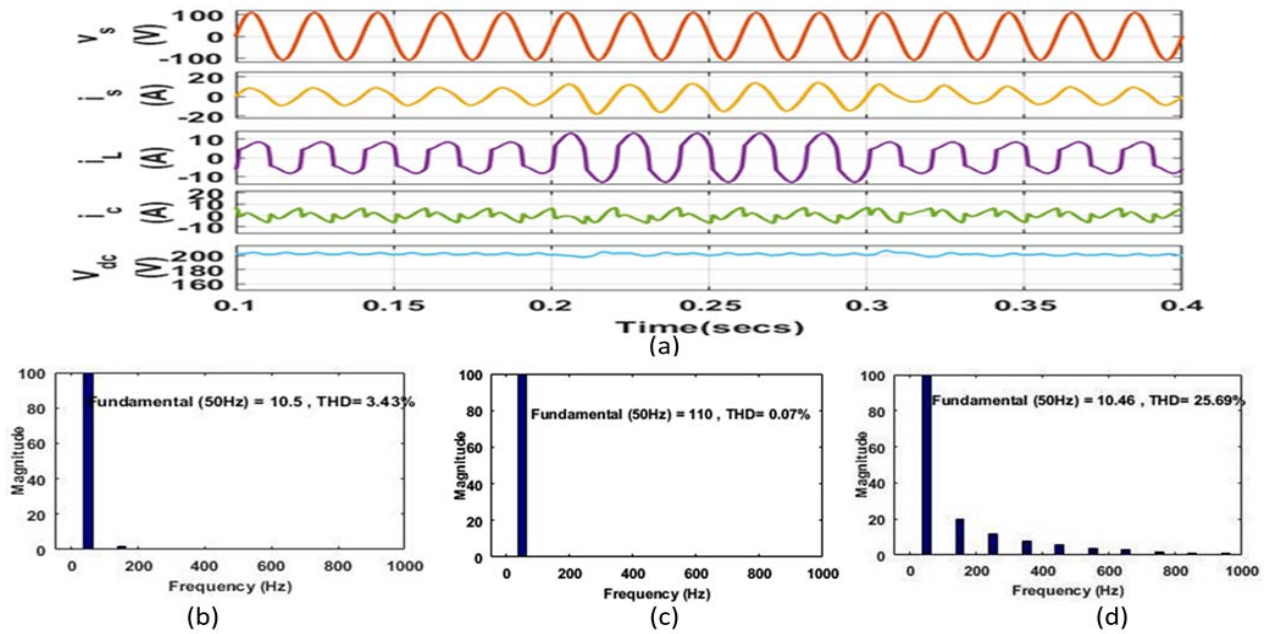


Fig. 10. (a) Simulation of one phase grid connected system using Biquad Active Algorithm feeding non-linear load with normal grid (b) THD in source current (c) THD in source voltage (d) THD in load current

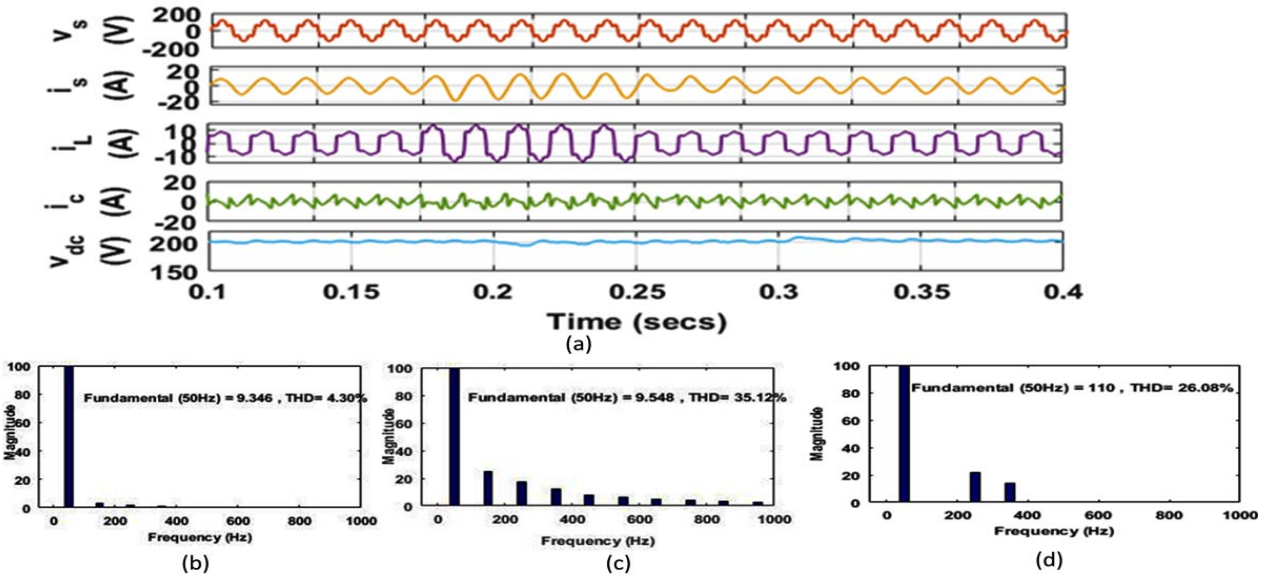


Fig. 11. (a) Simulation of one phase grid connected system using Biquad Active algorithm with distorted grid feeding non-linear load (b) Harmonics in source current (c) Harmonics in source voltage (d) Harmonics in load current

Fig. 10(b-d) illustrates the harmonic content in normal grid current, voltage and load current respectively. It is found to be 3.43% in source current,0.07% in source voltage and 25.69% in load current respectively.

Fig. 11(b-d) illustrates the THD in abnormal grid current, voltage and load current respectively. It is found to be 4.30% in source current, 35.12% in source voltage and 26.08% in load current respectively.

The simulation results show that the control algorithm gives satisfactory results.

VI. Experimental Results

The Simulink results obtained are justified through hardware results. The hardware results under both normal and abnormal grid conditions are discussed in this section.

A. Performance analysis under normal and abnormal grid conditions

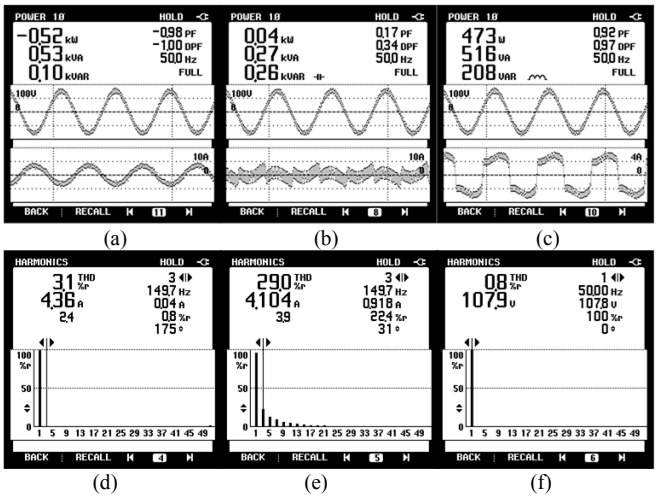
Fig. 12. illustrates the steady state power feeding non-linear load and THD analysis in normal grid. Fig. 12(a-c) illustrates the grid, compensator and load power respectively and Fig. 12(d-f) illustrates their harmonic analysis spectrum. We have found that the real power of 520 W delivered by the source is consumed by the load of 473 W and 40 W by the compensator. The THD in grid current is found to be 3.1%, 29% in load current and 0.8% in source voltage.

Fig. 13 illustrates the steady state power feeding non-linear load and THD analysis in abnormal grid conditions. Fig. 13(a-c) illustrates the illustrates the grid power, compensator power

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:08:19 UTC from IEEE Xplore. Restrictions apply.

and load power results. It has been found that the source delivers 459 W of source power out of which 415 W of power is consumed by the load and the remaining 46 W is absorbed by the compensator

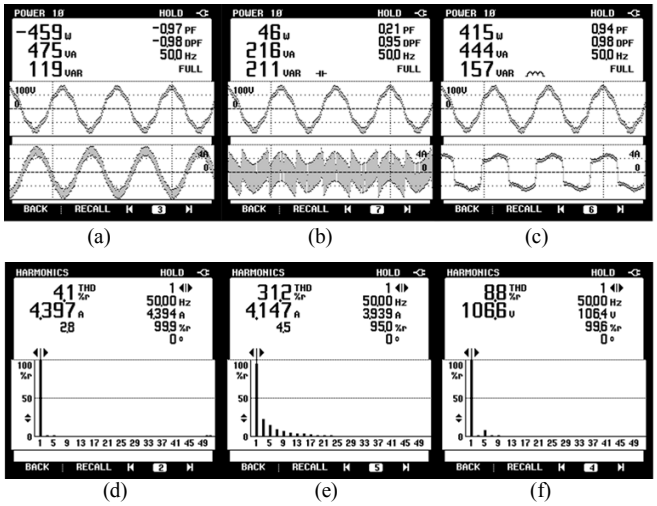
Fig. 13(d-f) illustrates the harmonic distortion factor in distorted grid conditions. It is found to have 4.1% THD in source current, 31.2% in load current and 8.8% THD in supply voltage.



(a) (b) (c)

(d) (e) (f)

Fig. 12. Steady state power and THD analysis under normal grid (a) Grid power (b) Compensator power (c) Load power (d) THD in grid current (e) THD in load current (f) THD in source voltage



(a) (b) (c)

(d) (e) (f)

Fig. 13. Steady state power and THD analysis in abnormal grid conditions (a) (a) Grid power (b) Compensator power (c) Load power (d) THD in grid current (e) THD in load current (f) THD in source voltage

VII. CONCLUSION

The performance realization of the designed Biquad Active Filter for shunt compensation is verified in both steady state and dynamic state conditions in MATLAB environment and on the hardware developed in the laboratory. It works quite well under both normal and abnormal grid conditions where THD in grid current is found to be less than 5% verifying IEEE-519 standard of harmonic distortion.

ACKNOWLEDGEMENT

The authors are thankful to the DST, SERB for the sponsored project (EMR/2016/001874).

REFERENCES

[1] A. Ghosh, G. Ledwich, Power quality enhancement using custom power devices, *Springer International Edition*, 2009.

[2] M.B. Latran, A.Teke, Y.Yoldaş, "Mitigation of power quality problems using distribution static synchronous compensator: a comprehensive review", *IET Power Electronics*, vol. 8, no. 7, pp. 1312-1328,2015.

[3] B. Singh, K. Al-Haddad, A. Chandra, "A review of active filters for power quality improvement", *IEEE Trans. Ind. Electron.*, vol. 46, no. 5, pp. 960-971, 1999.

[4] N. Kumar, I. Hussain, B. Singh, B.K. Panigrahi, "Implementation of Multilayer Fifth-Order Generalized Integrator-Based Adaptive Control for Grid Tied Solar PV Energy Conversion System", *IEEE Trans. on Industrial Informatics.*, vol. 14, no. 7, pp. 2857-2868, 2018.

[5] J. M. Kanieski, R. Cardoso, H. Pinheiro, H. A. Grundling, "Kalman filter-based control system for power quality conditioning devices", *IEEE Trans. Ind. Electron.*, vol. 60, no. 11, pp. 5214-5227, 2013.

[6] Y. Singh, I. Hussain, B. Singh, S. Mishra, "Single-phase solar grid-interfaced system with active filtering using adaptive linear combiner filter-based control scheme", *IET Gener. Transmiss. Distrib*., vol. 11, no. 8, pp. 1976-1984, 2017.

[7] P. E. Fleischer, J. Tow, "Design formulas for biquad active filters using three operational amplifiers", *Proc. IEEE*, vol. 61, pp. 662-663, May 1973.

[8] A.Goel, A.Vetteth, K.R. Rao, V. Sridhar, "Active Cancellation of Acoustic Noise Using a Self-Tuned Filter", *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol 51, no. 11, pp. 2148-2156, 2004.

[9] J.Tow, "Active RC filters—A state-space realization", *Proc. IEEE (Lett.)*, vol. 56, pp. 1137-1139, June 1968.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:08:19 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Design and Implementation of Microstrip Low Pass Filter at 2.4 GHz

Cheshta Kwatra
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
cheshta.kwatra@gmail.com

Kapil Gupta
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
kapil.bindal98@gmail.com

Kshitij Chhabra
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
kchhabra.1998@gmail.com

Dr. Priyanka Jain
Department of Electronics and Communication Engineering
Delhi Technological University
Delhi, India
priyankajain@dtu.ac.in

***Abstract*—Microwave low pass filter is an indispensable part of microwave circuits and wireless communication systems. This research work presents the design and simulation of low losses microwave low-pass filter of orders 3, 5, and 6 operating at 2.4 GHz cut-off frequency using the Insertion Loss Method. The design and simulation of the circuit have been carried out via Ansoft software. The outcomes predict the system is reliable and has the possibility of being transferred from lab potential experiments to industry applications. The designed filter is also fabricated on the FR-4 substrate with a dielectric constant =4.4 with thickness of the microstrip conductor 0.018mm and height = 1.6mm.**

Keywords—LPF, insertion loss, microstrip

I. INTRODUCTION

The use of UL-Frequencies in the domains of communication, wayfinding, surveillance and television increased sharply after World War II. These technological advancements fostered a curiosity in the research of microwave filters [1]-[3]. It is a two-port network which provides frequency selectivity in various systems and communication technologies operating at microwave frequencies (1 GHz and above). Since such components would attain a high value when designed using lumped impedance or coaxial lines, design using strip-lines is well suited.

Microstrip acts as a transmission line and is used to carry microwave signals when fabricated on a Printed Circuit Board (PCB) [4]. Micro-strips being compact and light is far less expensive than waveguide technology used traditionally. Micro-strips are used in high-speed digital PCB designs which require less crosstalk and radiation. However, these offer high losses and lower power handling capacity.

In this paper, a micro-strip based low-pass filter is proposed. Section II gives a brief about the implementation procedure which is followed by physical realization steps in Section III. The filter is then verified using HFSS simulation and results are presented in Section IV. Section V outlays results obtained by testing of the fabricated filter. Section VI presents a comparison between the designed filters of orders 3, 5 and 6. Section VII concludes the paper which is followed by the references.

II. FILTER DESIGN PROCEDURE

Classification of filters into various categories depends upon the locus of poles of insertion loss function and zeroes in the pass-band. "The various categories of microwave filter design include waveguide, dielectric resonators, coaxial lines, evanescent-mode filters, and printed circuit filters in micro-strip, strip-line, and suspended substrate" [5]-[6]. Micro-strip line filters provide fairly good results over an additional advantage of compact size and cheap cost. This section describes the details of the design procedure of microstrip filters, selection of optimal design method

There is a strong correlation between the designing of strip-line filters and lumped parameter filters, and strip-lines filters can be synthesized using the method for lumped parameter filters at much lower frequencies [7]-[8]. The pass-band and stop-band amplitude and phase characteristics of the filter can be controlled to a great degree using the insertion loss method [9]-[10] used in this paper. The insertion loss method also provides a systematic way to synthesize the required filter response [11]. The process for filter design in the insertion loss method is depicted by a flowchart as shown in Fig. 1.

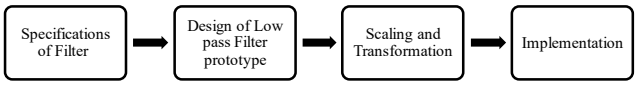


Fig 1. Insertion Loss Method

The first step is to select an appropriate transmission coefficient $|\Gamma(\omega)|^2$ or insertion loss ratio, P_{LR}:

$$"P_{LR} = \frac{\text{Power available from source}}{\text{Power delivered to load}} = \frac{P_{inc}}{P_{load}} = \frac{1}{1-|\Gamma(\omega)|^2}" \quad (1)$$

Next step is the selection of application specification transmission coefficient. A binomial response could be used for minimum insertion loss. For a sharp cut-off, Chebyshev response satisfies the requirement.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:15 UTC from IEEE Xplore. Restrictions apply.

"Butterworth or maximally flat: This response is also called the binomial response, and is the best in the sense that it provides the flattest possible pass-band response for a given filter order".

$$P_{LR} = 1 + \omega^4 \quad (2)$$

Chebyshev or equal ripple: If a Chebyshev polynomial specifies the Nth order LPF's insertion Loss as in 3

$$P_{LR} = 1 + k^2 T_N^2(\omega) \quad (3)$$

then a sharper cut-off will be observed, although ripples of amplitude $1 + k^2$ will be seen in the pass-band response.

The element values for any order filter are selected from the Nth order Butterworth and Chebyshev table for low pass prototype. A maximally flat low pass filter prototype is shown in the figure below (Fig. 2).

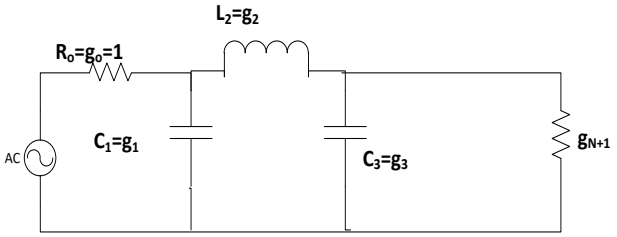


Fig. 2. Low Pass Filter Prototype

III. Physical Realization

Succeeding the design of a low pass filter, the subsequent step is to physically realize necessary compoents in the circuit. Several types of strip transmission line structures are presented in [12]-[14].

Since, the availability of the lumped elements - inductors and capacitors are limited only to a range of values.Richard's Transformation and Kuroda's Identities have been used in the proposed design to convert lumped parameter filter designs to distributed filters.

The lumped LPF- prototype has been set up using Richard's Transformation to substitute the "lumped L and C elements". For an inductance of L in a prototype filter, a short-circuited transmission line stub of length βL and characteristic impedance Z_o = L has been substituted using:

$$jX_L = j\Omega L = jZ_O \tan\beta L \quad (4)$$

"And the susceptance of a capacitor has been substituted with an open-circuited stub of length βL and characteristic impedance Y_o = 1/C using":

$$jB_C = \frac{j\Omega}{C} = jY_O \tan\beta L \quad (5)$$

For the prototype filter which was normalized to a cut-off frequency ω_c = 1 and admittance g_o = 1, the length of stub came out to be λ/8, using

$$\Omega = 1 = \tan\beta L \quad (6)$$

"where λ is the wavelength of the transmission line stub at cut-off frequency ω_c".

Then, the tricky or unrealizable elements were converted to elements having more malleable values and geometry using the Kuroda's Identities (Fig. 3).

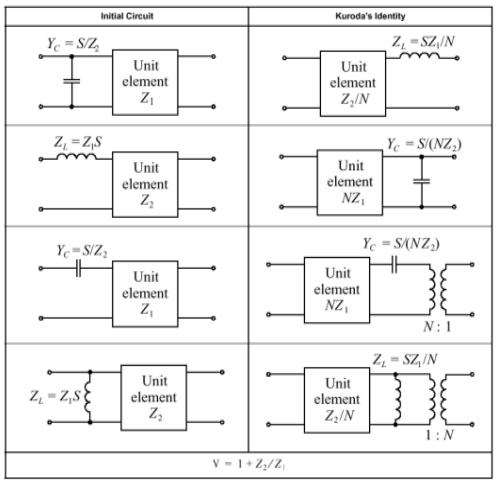


Fig. 3. Kuroda Identities

"Kuroda's four Identities use redundant transmission line to physically separate the line stubs and transform series stubs to shunt stubs or the other way round". This type of design is also known as the redundant filter synthesis since the filter response is not affected by the additional transmission line sections.

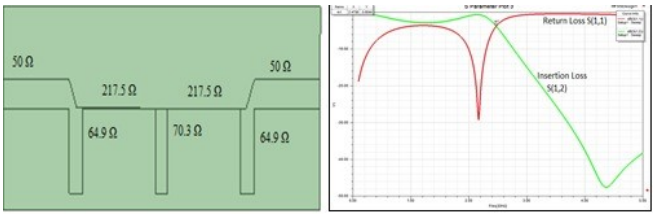
In the last step, the filter design has been scaled to the desired ω_c and Z_o (50Ω).

IV. Simulation And Results

The filter is designed on the FR-4 substrate with a dielectric constant (∈r) of 4.4 and height of substrate (h) equivalent to 1.6 mm. The electric length is given as 45 degrees and copper conductor thickness = 0.018 mm. The filter is designed with a microstrip transmission line having a stub on top of the substrate and a plain copper surface at the bottom side of the substrate. For the design and the simulation of the structure, HFSS software is used.

A. Simulation of 3rd order Butterworth Filter

The top view of the proposed filter is shown in Fig. 4 (a). Dimensions of the design are shown in the figure. A simulation was done to verify the filter's low pass characteristics by measuring its S parameters. The simulated response is shown in Fig. 4 (b). It may be observed that the passband region is at an average of 0dB while rejection band starts at a cut-off frequency of 2.3538 GHz.



Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:15 UTC from IEEE Xplore. Restrictions apply.

Fig. 4. (a) Top View of Filter (b) Simulated Response of Filter

B. *Simulation of 5th order Chebyshev Filter*

"The top view of the proposed filter is shown in Fig 5 (a). Dimensions of the design are shown in the figure". The simulation was done to verify the filter's low pass characteristics by measuring its S parameters. The simulated response is shown in Fig 5 (b). It may be observed that the passband region is at an average of 0dB while rejection band starts at a cut-off frequency of 2.3933 GHz with the stopband characteristics of less than 10dB (11dB), hence transferring more than 90% of the power.

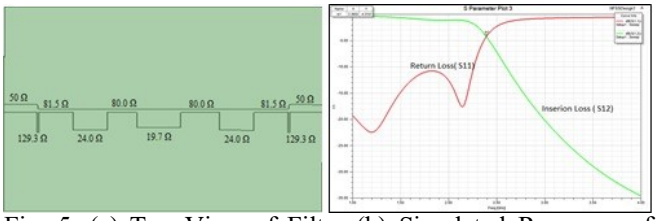


Fig. 5. (a) Top View of Filter (b) Simulated Response of Filter

C. *Simulation of 6th order Butterworth Filter*

"The top view of the proposed filter is shown in Fig 6 (a). Dimensions of the design are shown in the figure". The simulation was done to verify the filter's low pass characteristics by measuring its S parameters. The simulated response is shown in Fig 6 (b). It may be observed that the passband region is at an average of 0dB while rejection band starts at a cut-off frequency of 2.4245 GHz.

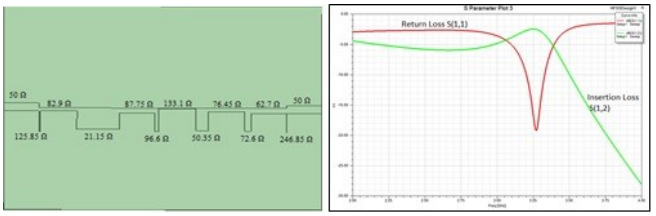


Fig. 6. (a) Top View of Filter (b) Simulated Response of Filter

V. Fabrication And Testing

The low pass filter is designed on a printed circuit board with copper as a perfect electric conductor (PEC) on an RF-4 substrate with a dielectric constant 4.4 and thickness equal to 0.018 mm. The layout of the transmission line is designed in schematic and PCB design software. The top and bottom surfaces are then exported separately as a Drawing Interchange Format or Drawing Exchange Format (.dxf) format. This is converted to a Gerber (.gbr) file to be fed into the ProtoMat Circuit Board Plotter. The fabrication is done using the plotter with great precision and reliability. SMA connectors are connected through soldering for providing input and output ports to the design.

Measurement and testing of the low pass filter were done by using Vector Network Analyzer (VNA). Experimental set up of the fabricated filter is shown in Fig.7.

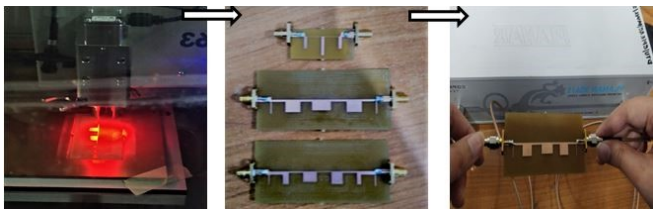


Fig. 7. Fabrication Process

A. *Result of 3rd order Butterworth Filter*

tested response withf 2.337 GHz cut-off frequency is shown in Fig.8.

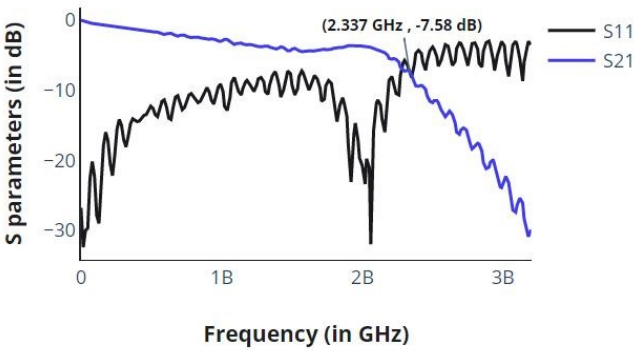


Fig. 8. Tested Response of Filter

B. *Result of 5th order Chebyshev Filter*

Tested response with cut-off frequency= 2.404 GHz. is shown in Fig.9.

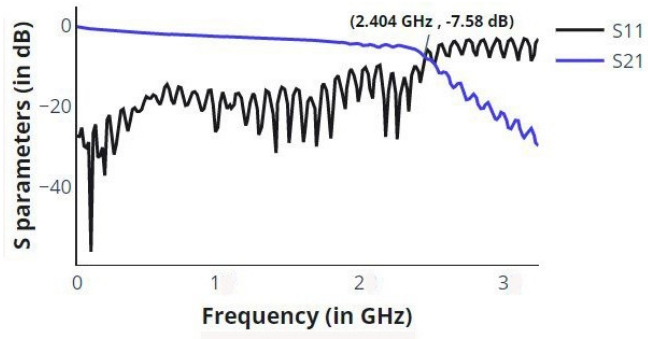


Fig. 9. Tested Response of Filter

C. *Result of 6th order Butterworth Filter*

Tested response with cut-off frequency =2.406 GHz. is shown in Fig.10.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:15 UTC from IEEE Xplore. Restrictions apply.

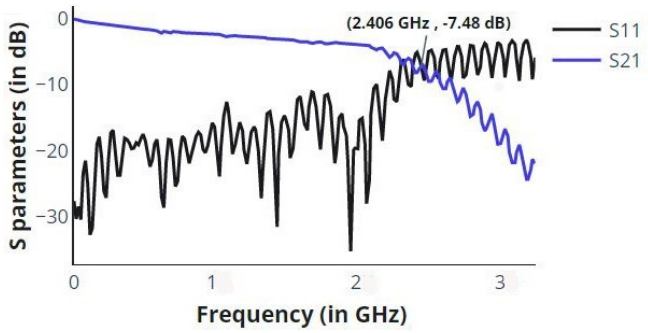


Fig. 10. Tested Response of Filter

VI. Comparison

In the table given below, it can be seen that the cut-off frequency approaches the desired value on increasing the order of the filter. Moreover, this analysis depicts that higher the order, lower is the return loss and sharper the roll-off factor of the low pass filter. Thus, 6th order Butterworth low pass filter gives us the most accurate result by transferring more than 99% of the power at 2.4 GHz with a roll-off factor of 19.55dB/GHz.

Table I : Comparison of Different Filters

Type of LPF	Order of LPF	Cut-off Frequency (GHz)	Roll off (dB/GHz)	Return Loss (dB)
Butterworth	3rd	2.370	19.55	30
Chebyshev	5th	2.404	20.72	40
Butterworth	6th	2.406	23.78	35

Conclusion

Low pass filters with the given specification could be commercially used at 2.4GHz frequency. The cut-off frequency can be controlled by changing the parameters of the design structure.

The filter was fabricated and tested for performance evaluation. The tested reponse evinces the proficiency of the system.

References

[1] W. P. Mason and R. A. Sykes, “The use of coaxial and balanced transmission lines in filters and wide band transformers for high radio frequencies:’ Bell Syst. Tech. J., vol. 16, pp. 275-302, 1937.

[2] R. Levy, S.B. Cohn , “A History of Microwave Filter Research, Design, and Development”, IEEE Transactions on Microwave theory and techniques, VOL. MTT-32, NO. 9,SEPTEMRER-1984.

[3] E. H. Bradley, “Design and Development of strip line filters” IEEE Transactions on Microwave Theory and Techniques, vol. 4(2), pp. 86–93.

[4] K.Rajasekaran, J.Jayalakshmi and T.Jayasankar, “Design and analysis of stepped impedance microstrip low pass filter using ADS Simulation Tool for wireless applications”, International Journal of Scientific and Research Publications, Volume 3, Issue 8, August 2013.

[5] R. Levy, R. V. Snyder and G. Matthaei, “Design of Microwave Filters” , IEEE Transactions on Microwave Theory and Techniques, vol. 50, no. 3, march 2002.

[6] Annapurna Das, Sisir K Das , “Microwave Engineering ”Tata McGraw-Hill New Delhi. Pp. 275-289.

[7] E. A. Guillemin, “Modern methods of network synthesis, ” Advances in Electronics, vol. 3, Academic Press, Inc., New York, N. Y. pp. 261-303; 1951.

[8] J. R. Whinnery, “Design of microwave filters” Proc. Sym. Mod. Network Synth., New York, pp. 296–3 11; April, 1952.

[9] G.L. Matthaei, “Design of Wide-Band (and Narrow-Band) Band-Pass Microwave Filters on the Insertion Loss Basis”, IRE Transactions on Microwave Theory and Techniques ,Volume: 8 , Issue: 6 , November 1960.

[10] E. G. Fubini and E. A. Guillemin, “Minimum Insertion Loss Filters”, Proceedings of the IRE ,Volume: 47,Issue: 1, Jan. 1959.

[11] M. Pozar, “Microwave Engineering “ 4 th Ed. Wiley, Ed. pp. 399–401.

[12] E. G. Fubini, W. E. Fromm, and H. S. Keen, “New techniques for high-Q strip components,” CONVENTION RECORD OF THE IRE, Part 8, p. 91; 1954.

[13] D. D. Greig and H. F. Englemann, “Microstrip-A new transmission technique for the kilomegacycle range” PROC. IRE, vol. 40, pp. 1644–1650; December, 1952.

[14] R. Levy and T. E. Rozzi, “Precise design of coaxial low-pass filters:’ IEEE Trans. Microwave Theory Tech., vol. MTT-16, pp.142-147, Mm-. 1968.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:36:15 UTC from IEEE Xplore. Restrictions apply.

Design and Testing of Frequency Adaptive Zero-Crossing Detector as a Synchronizing Technique

Hemant Saxena
Department of Electrical Engineering
Delhi Technological University
New Delhi, India
hsaxena531@gmail.com

Alka Singh
Department of Electrical Engineering
Delhi Technological University
New Delhi, India
alkasingh.dr@gmail.com

Jitendra Nath Rai
Department of Electrical Engineering
Delhi Technological University
New Delhi, India
jnraiphd1968@gmail.com

***Abstract*—In this paper, a simple and efficient technique has been proposed for estimating the synchronizing signals. The proposed method is based on the zero-crossing detection technique. The proposed method works very well even when with variation in the frequency of the input signal while the conventional Zero Crossing Detector (ZCD) fails. The proposed technique shows a very fast response in case of phase change, distorted signal and frequency change. The technique has been simulated in MATLAB/Simulink environment and also tested experimentally. The necessary simulation and experimental results have been shown for the verification of the proposed technique.**

Keywords—Zero Crossing Detector, Synchronising Signals, SOGI, Frequency Adaptive

I. INTRODUCTION

In the present day, power distribution networks have a number of different types of connected loads. Different non-linear loads such as computers, switch-mode power supplies (SMPS), modern lighting devices, modern small electronic equipment etc [1]. Form a large proportion of such loads in households as well as industries. These non-linear loads generate harmonics in the distribution network. Shunt active power filters (SAPF) are widely used in power distribution networks for power quality improvement and reactive power compensation [2], [3]. Effective working of SAPF requires a control technique for achieving proper load compensation. Another integral part of the control technique is the generation of synchronizing signals from the grid voltage for phase computation and for the generation of the reference current.

Different techniques have been widely reported in the research for the estimation of the synchronizing signals. The unit template method is one of the simplest methods to generate synchronizing signals from the grid voltage [4]. Although, this technique is the simplest and straight-forward, yet best results are achieved only under ideal grid conditions. This method works satisfactorily when the grid voltage experiences variation in magnitude in the form of sag and swell. It fails miserably under other cases such as frequency change, phase change and distortion in voltage.

Different types of Phase-Locked Loops (PLL) have also been used to estimate the synchronizing signals in the past and some common methods include Synchronous Reference Frame PLL (SRF-PLL) [5], Second Order Generalised Integrator PLL (SOGI-PLL) [6], and Enhanced Phase Locked Loop (EPLL) etc [7]. SRF-PLL is the most commonly used PLL and requires Park and Clarke transformations. The SRF-PLL works very well under normal conditions but its performance is not good under distorted grid conditions. SOGI PLL can be used to filter out the fundamental component of the grid voltage and hence SOGI-PLL gives better performance in case of distorted grid conditions. EPLL has a better double frequency cancellation capability. All these designed PLLs show a slight delayed response due to the complexity of the algorithm. The tracking delay observed can be of the order of a few cycles.

A zero-crossing detection technique can also be used to generate synchronizing signals. The conventional ZCD generates perfect synchronizing signals with a simple algorithm [8], [9]. However, the conventional ZCD has the disadvantage that it can generate synchronizing signals of only fixed frequency. When the grid frequency changes to a value other than the fundamental frequency, perfect synchronizing signals are not obtained. This paper discusses the design and experimental testing of the improved ZCD technique for synchronization. The proposed algorithm always generates the synchronizing signals of the same frequency as that of the input grid voltage. The main highlights of the proposed algorithm are:

1. The proposed technique generates synchronizing signals with almost negligible total harmonic distortion (THD).
2. It can detect phase change within ½ cycle.
3. The proposed technique is very simple in structure and easy to implement.
4. The proposed method is frequency adaptive

The paper has been organized into five sections. The first section discusses the need of synchronizing signals and also discusses the previous work on estimating the synchronizing signals. Section-II discusses the conventional ZCD for generating the synchronizing signals. Section-III discusses the frequency adaptive ZCD method. In Section-IV, the simulation and experimental results have been presented. In last, the conclusion has been discussed in Section-V.

II. ZERO CROSSING DETECTOR (ZCD)

The ZCD is based on the detection of zero crossings of the input signal. It is based on the simple "if" and "then" logic. If the voltage signal crosses the zero mark and is increasing to the positive-half cycle, then the ZCD generates the unit sine wave of positive-cycle. If the voltage signal crosses the zero and going towards the negative-half cycle then the ZCD generates the unit sine wave of only negative-cycle. After combining both the positive half cycle and

978-1-7281-6664-3/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:07:50 UTC from IEEE Xplore. Restrictions apply.

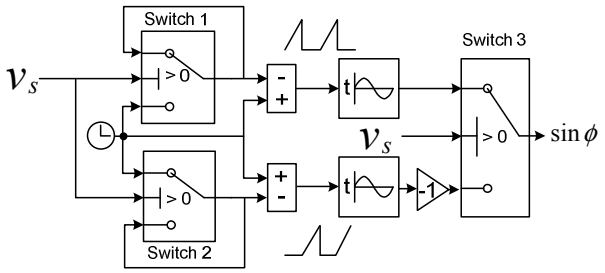


Fig 1 Structure of ZCD

negative half cycle, the generation of the complete sine wave of 50Hz frequency takes place. The generated wave is the frequency of 50Hz, in-phase with the voltage and free from any distortion. The structure of the ZCD has been shown in Fig 1. It doesn't depend on the magnitude of the input waveform. The conventional ZCD generates the sinusoidal waveforms of the frequency of 50Hz even when the frequency of the voltage signal deviates from the fundamental frequency (50Hz), but there will be some error at zero crossings. Hence, the conventional ZCD requires a frequency estimator, which can detect the frequency of the input signal and correspondingly generate the pure sine wave in phase with the input signal. The modification of the conventional ZCD to work under any variation in frequency is discussed in the next Section.

III. Frequency Adaptive ZCD

The grid-voltage is not always ideal. It may be contaminated with harmonics and noise signals. The distorted grid voltage can be represented as

$$v_s(t) = V_1 \sin(\omega t + \emptyset_1) + \delta_h(t) \quad (1)$$

where V_1 is the magnitude of the grid voltage, ω is the frequency, h is the harmonic number and $\delta_h(t)$ is harmonic in the grid voltage. The equation (1) can be discretized and written as

$$v_s(k) = V_1 \sin(\omega k T_s + \emptyset_1) + \delta_h(kT_s) \quad (2)$$

where T_s is the sampling time. The three-consecutive-sampling (3CSs) [10], [11] method is used and after applying the time delay distorted grid voltage can be expressed as

$$v_1(k) = v_s(k-\tau) = V_1 \sin(\omega k T_s - \omega\tau + \emptyset_1) + \delta_h(kT_s - \tau) \quad (3)$$

$$v_2(k) = v_s(k-2\tau) = V_1 \sin(\omega k T_s - 2\omega\tau + \emptyset_1) + \delta_h(kT_s - 2\tau) \quad (4)$$

Expanding equation (3) and (4) and written as

$$v_1(k) = \beta_1 V_1 \sin(\omega k T_s + \emptyset_1) - \beta_2 V_1 \cos(\omega k T_s + \emptyset_1) + \delta_h(kT_s - \tau) \quad (5)$$

$$v_2(k) = \beta_3 V_1 \sin(\omega k T_s + \emptyset_1) - \beta_4 V_1 \cos(\omega k T_s + \emptyset_1) + \delta_h(kT_s - 2\tau) \quad (6)$$

where $\beta_1 = cos\omega\tau$ (7a)

$\beta_2 = sin\omega\tau$ (7b)

$\beta_3 = cos2\omega\tau$ (7c)

$\beta_4 = sin2\omega\tau$ (7d)

Simplifying the above equation (7), the relationship obtained between $\beta_1, \beta_2, \beta_3$ and β_4 is

$$\beta_3 = 2\beta_1^2 - 1 \quad (8)$$

$$\beta_4 = 2\beta_1\beta_2 \quad (9)$$

From the equation (5) and (6),

$$\beta_2 V_1 \cos(\omega k T_s + \emptyset_1) = \beta_1 V_1 \sin(\omega k T_s + \emptyset_1) + \delta_h(kT_s - \tau) - v_1(k) \quad (10)$$

$$\beta_3 V_1 \sin(\omega k T_s + \emptyset_1) - \beta_4 V_1 \cos(\omega k T_s + \emptyset_1) + \delta_h(kT_s - 2\tau) - v_2(k) \quad (11)$$

Multiplying $2\beta_1$ in equation (10) and substituting in (8) and (11) gives the following result.

$$v_s(k) + v_2(k) = 2\beta_1 v_1(k) + \delta_h(kTs) - \delta_h(kTs - \tau) + \delta_h(kTs - 2\tau) = 2\beta_1 v_1(k) - \Delta(k) \quad (12)$$

In the ideal case, $\Delta(k)$ equals to zero. Equation (12) can be simplified as

$$v_s(k) + v_2(k) = 2\beta_1 v_1(k) \quad (13)$$

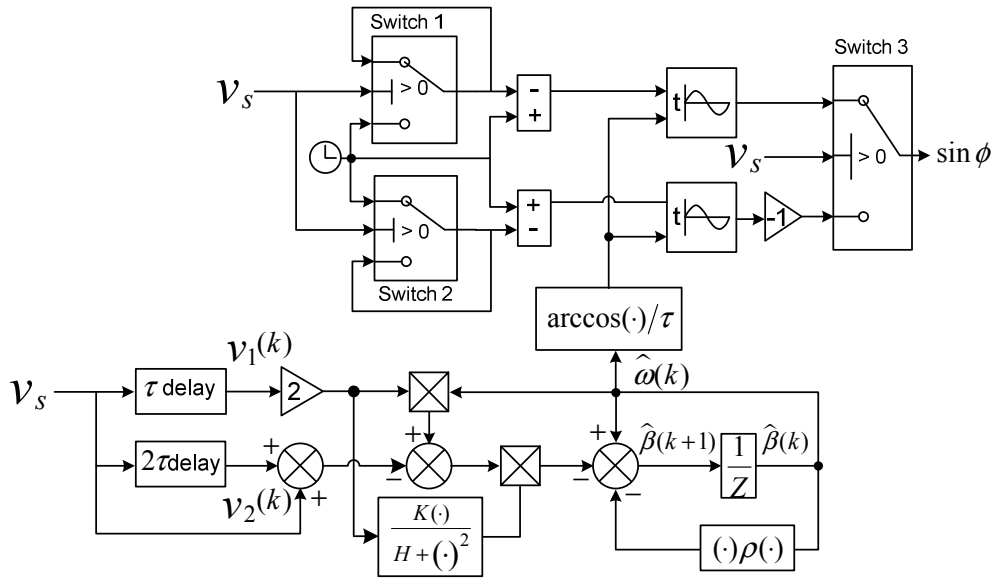


Fig 2. Proposed structure of adaptive-ZCD (A-ZCD)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:07:50 UTC from IEEE Xplore. Restrictions apply.

From equation (13), the frequency of the ideal grid voltage is estimated as

$$\omega = \frac{\arccos((v_s(k)+v_2(k)/2v_1(k))}{\tau} \quad (14)$$

For the estimation of the frequency of the grid voltage corrupted with harmonics or deviation from 50Hz, a robust error suppression is required. First, the error can be identified as

$$\varepsilon(k) = 2\hat{\beta}_1(k)v_1(k) - v_g(k) - v_2(k) \quad (15)$$

where $\hat{\beta}_1$ is the new estimation of β_1. From equation (12), equation (15) can be written as

$$\varepsilon(k) = 2\gamma_1(k)v_1(k) + \Delta(k) \quad (16)$$

Here, $\gamma_1 = \hat{\beta}_1 - \beta_1$ is the estimation error. For removing the effect of $\Delta(k)$, a robust law is used and is expressed as

$$\hat{\beta}_1(k+1) = \hat{\beta}_1(k) - \frac{2K\varepsilon(k)v_1(k)}{\alpha^2(k)} + g(k) \quad (17)$$

where $\alpha(k) = \sqrt{H + (2v_1(k))^2}$ (18)

H>0 and K>0 are designed values and g(k) is some modification and is estimated as

$$g(k) = -\rho(k)\hat{\beta}_1(k), \quad \rho(k) = \begin{cases}\rho_0 & |\hat{\beta}_1(k)| \geq \beta_1^* \\ 0 & |\hat{\beta}_1(k)| < \beta_1^*\end{cases} \quad (19)$$

where 0<K<1 and $0 < \rho_0 < (1-K)/2$, $\beta_1^* \geq \beta_{1_upper}$ and β_{1_upper} is the maximum possible value of $|\beta_1|$.

The maximum value of $|\beta_1|$ is 1, according to equation 7(a). In case of special disturbance condition, the value of $|\hat{\beta}_1(k)|$ may larger than 1. The higher value of H shows the slower convergence and lower transient value and vice versa. The value of H should be greater than zero and here the optimized value of H is taken as 100. Here, the value of τ is taken as T/4 as the design by 3C'S method [10]. T is the time period of the fundamental frequency of the grid. The value of K is taken as 0.03, as it provides the good settling time than compared with the lower value of K. According to the [11], the RFE achieves input-to-state stability.

Now, the from equation (7a) and equation (17), the new grid frequency ($\hat{\omega}$) of grid voltage is estimated as

$$\hat{\omega}(k) = \frac{\arccos(\hat{\beta}_1(k))}{\tau} \quad (20)$$

Now, the estimated frequency is given to the sine wave generator. The sine wave generator now generates a pure unit sine wave of the frequency of the input grid voltage signal. The proposed adaptive ZCD (A-ZCD) based synchronizing signal generator is shown in Fig 2.

IV. Results

The proposed method is implemented and simulated in MATLAB/Simulink environment. The working of the proposed technique is also checked experimentally.

A. Open-loop Simulation Results

For the simulation, the input signal of the voltage of 100V RMS and 50Hz frequency. The different test conditions of voltage sag, distorted voltage, change in frequency and phase angle deviation are taken and tested under these conditions. The proposed algorithm for generating the synchronizing signals is compared with the conventional ZCD and the SOGI-PLL. Fig. 3 show the open-loop simulation results for different grid voltage conditions

Fig. 3(a) shows the open-loop simulation results for variation in frequency from 50Hz to 52Hz at 0.3s. The conventional ZCD detects the zero crossings and gives the sine wave of 50Hz even though the frequency of the input signal is 52Hz. The SOGI-PLL detects frequency but it takes 5 to 6 cycles to obtain synchronization and obtain signals in phase with the input signal. The proposed method takes a maximum of 3 cycles to detect the exact frequency and gives pure unit sinusoid of 52Hz.

Fig. 3(b) shows the simulation results for the condition of change in phase angle of 30°. In this case, the conventional ZCD and the proposed A-ZCD give the almost same result and phase change has been detected within ½ cycles. While the SOGI-PLL takes almost 3 cycles for the detection of phase change.

Now the proposed A-ZCD is tested under the distorted grid voltage conditions. Under this condition, the simulation results have been shown in Fig. 3(c). The frequency

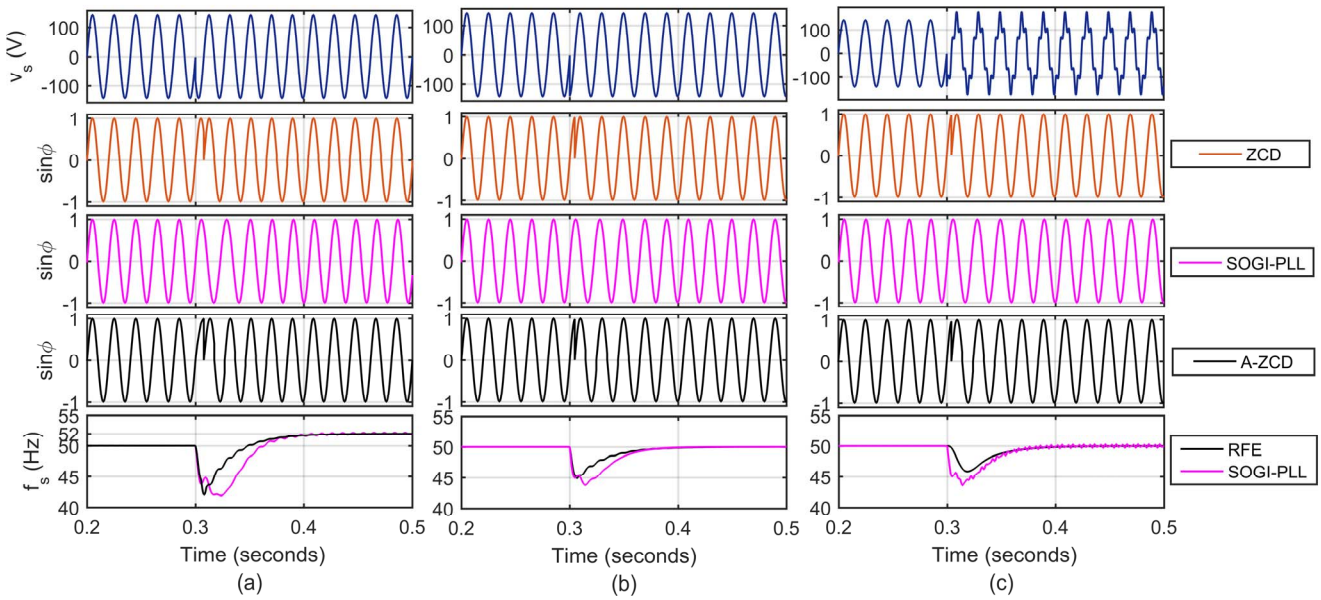


Fig. 3 Open-loop simulation results (a) Frequency change of 50Hz to 52Hz (b) Phase change of 30° (c) Distorted grid voltage

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:07:50 UTC from IEEE Xplore. Restrictions apply.

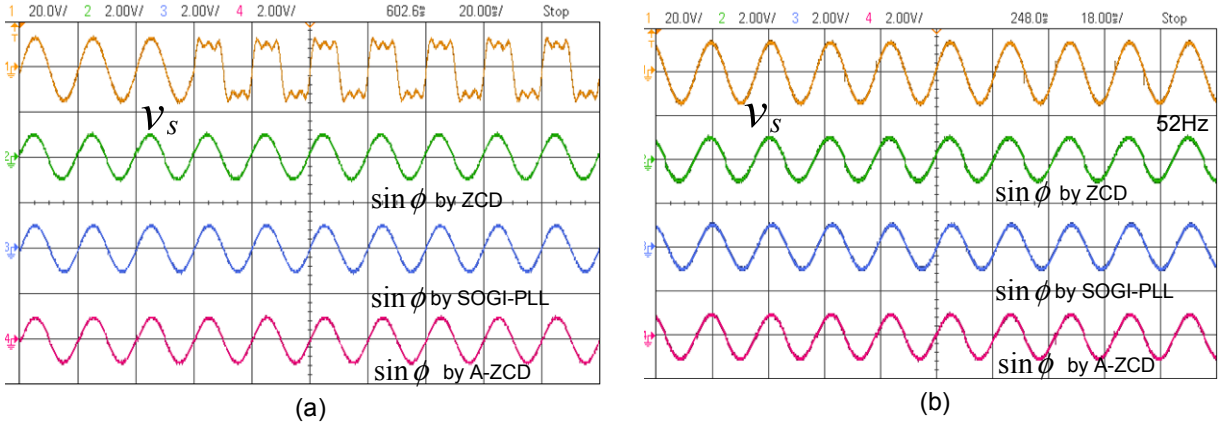


Fig. 4 Experimental comparison results of proposed A-ZCD with ZCD and SOGI-PLL (a) distorted grid voltage (b) grid voltage with frequency of 52Hz

estimated by the SOGI-PLL shows oscillations, so the unit synchronizing signal generated contains some harmonics. The proposed technique doesn't depend on the level of distortion hence it gives only pure sine wave. Hence, from these results, the proposed A-ZCD gives superior results than conventional ZCD and SOGI-PLL for frequency for the frequency deviation, phase change, and distorted grid voltage. The comparison table has been drawn according to the simulation performance and the nature of the algorithm. The cyclomatic metrics has been used to estimate the complexity of algorithms [12] and shown in Table 1.

TABLE I COMPARISON OF ZCD, SOGI-PLL AND A-ZCD

	Conventional ZCD	SOGI-PLL	A-ZCD
Type	Non-Adaptive	Adaptive	Adaptive
Transformation requirement ($\alpha\beta \rightarrow dq$)	No	Yes	No
PI controller requirement	No	Yes	No
Cyclomatic Complexity	4	6	6
Frequency Detection	No	fast	faster
Error at zero-crossing for frequency change	Yes	No	No
Phase change detection	<1/2 cycle	<3 cycles	<1/2 cycles

B. *Open-loop Experimental Results*

The proposed technique has been tested experimentally. For the generation of the voltage signal, Gwinstek make APS1102 programmable supply has been used. LEM make LA25-NP and LV25-p/sp2 current and voltage sensors have been used to sense the system parameters. The sensed signals are given to the dSpace1104 which processes the proposed technique. The outcomes of the A-ZCD have been analyzed by the Keysight DSOX-2014A is used. The experimental comparison results of proposed A-ZCD with conventional ZCD and SOGI-PLL have been shown in Fig 4. The sampling time for the hardware system is 50µs.

Fig 4(a) shows the results for the distorted grid voltage. In this case, the SOGI-PLL shows a little distortion in the synchronizing signal while the conventional ZCD and A-ZCD give almost similar results. Fig 4(b) shows the steady-state results for the grid voltage with a frequency of 52Hz. In this condition, the conventional ZCD shows a synchronizing signal of frequency 50Hz. While the SOGI-PLL and A-ZCD give synchronizing signals with 52Hz. So from the results, the A-ZCD gives good results in both the conditions. The SOGI-PLL gives good results in case of frequency change but it requires the αβ to dq transformation and also a PI controller.

C. *Application of A-ZCD for SAPF controller*

The developed algorithm has also been tested for the SAPF for the power quality improvement. The structure of the SAPF in single-phase system is shown in Fig. 5. A non-linear load has been connected to the single-phase ac source and an H-bridge inverter has been connected to it via interfacing inductor. The controller requires fundamental component (I_{fl}) of the load current (i_L), synchronizing signal and the DC-link voltage controller. For the DC-link voltage controller PI controller has been used and it's proportional (K_p) and integral gain (K_i) has been chosen optimally. The output of the PI controller gives the loss component of the inverter and is expressed as in the discrete form

$$I_{dc}(k) = I_{dc}(k-1) + K_p[V_{dce}(k) - V_{dce}(k-1)] + K_i V_{dce}(k) \quad (21)$$

$$V_{dce}(k) = V_{dc}^*(k) - V_{dc}(k) \quad (22)$$

For the estimation of fundamental load component (I_{fl}), SOGI has been used for its excellent filtering performance. SOGI first converts the input signal to the in-phase filtered

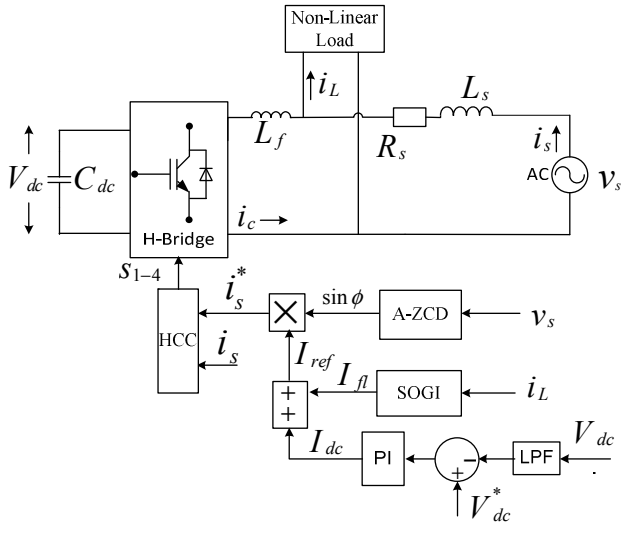


Fig. 5 Structure of single-phase SAPF

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:07:50 UTC from IEEE Xplore. Restrictions apply.

component and quadrature-phase filtered component. The reference current is estimated by the multiplying net fundamental component to the unit template of the voltage signal and expressed as in equation (23).

$$i_s^* = I_{ref} \times sin\emptyset = (I_{fl} + I_{dc})sin\emptyset \qquad (23)$$

Reference current (i_s^*) has been compared with the grid current (i_s) and given to the hysteresis current controller (HCC) to generate the appropriate gating signals for the inverter. The simulation results have been shown in Fig. 6 and Fig. 7 for the power quality improvement by SAPF. Fig. 6 shows the grid voltage, grid current, load current, inverter current and DC-link voltage. The grid voltage is undistorted

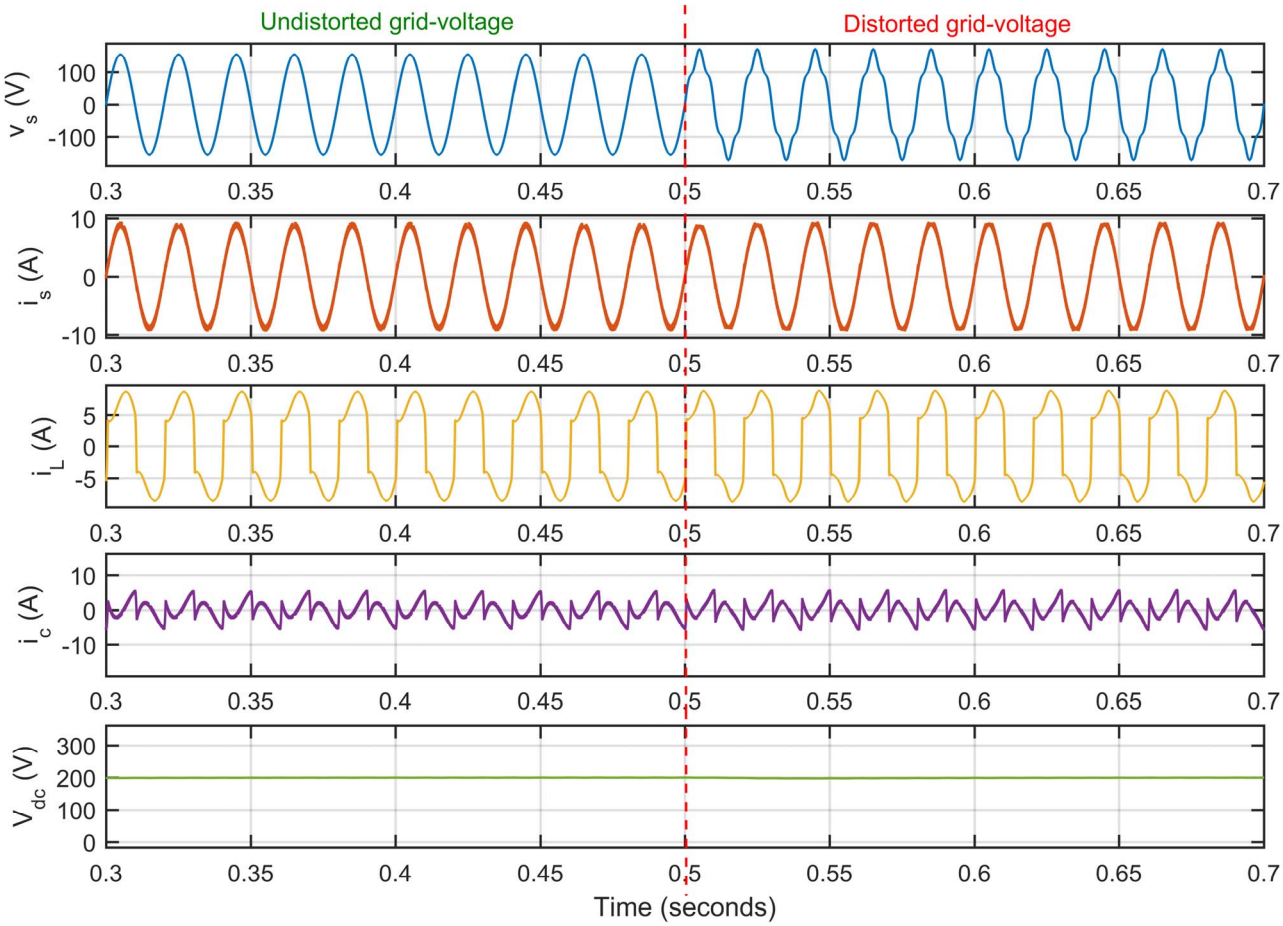


Fig. 6 Performance of A-ZCD in SAPF under undistorted and distorted grid voltage conditions

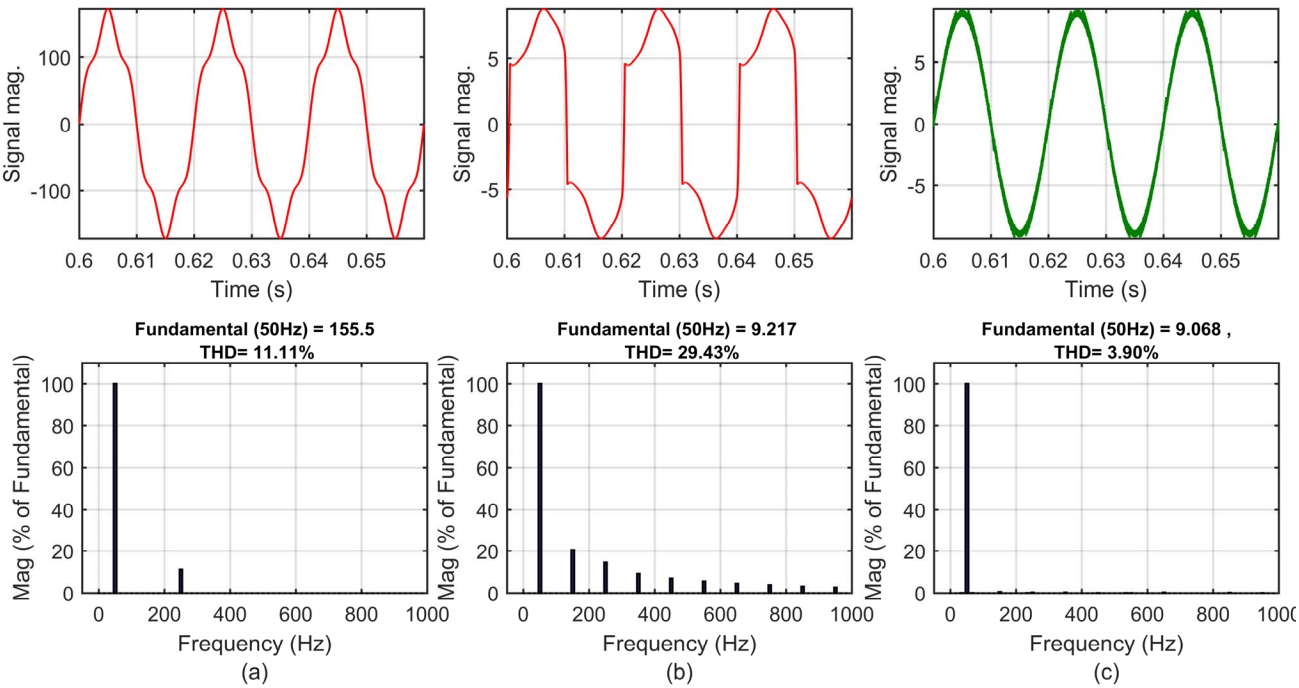


Fig. 7 THD analyses (a) Grid-voltage (v_s) (b) Load current (i_L) (C) Grid current (i_s)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:07:50 UTC from IEEE Xplore. Restrictions apply.

till t=0.5s and after t=0.5s the grid has some distortions in the waveform. In both cases the grid current is sinusoidal and the harmonic content is also removed. The DC-link voltage is constant at reference DC-link voltage in both distorted and undistorted grid voltage conditions. Fig. 7 shows the THD analyses of grid-voltage, load current and grid-current. The Load current THD is 29.43% and the grid voltage THD is 11.11%. After the compensation, grid current THD achieves only is 3.90% which follows the IEEE-519 standard. The presented simulation SAPF performance shows the effectiveness of the proposed A-ZCD technique.

V. Conclusion

In this paper, an adaptive ZCD technique has been proposed for the estimation of synchronizing signals. The conventional ZCD gives always the synchronizing signal of 50Hz even in case of a deviation of 50Hz. The proposed technique uses a robust frequency estimator and estimates frequency correctly and also generates synchronizing signals of the same frequency as of grid voltage. The simulation and experimental results have been shown for the effectiveness of the proposed technique. The proposed technique has also been effectively applied for the generation of reference currents for shunt compensator and its application in SAPF for power quality improvement.

Acknowledgment

The authors are thankful to the Department of Science and Technology, Science and Engineering Research Board for the sponsored project (EMR/2016/001874).

References

[1] R. C. Dugan, M. F. McGranaghan, S. Santoso, and H. W. Beaty, "Electrical Power Systems Quality," in *Electrical Power Systems Quality*, 2004, pp. 1–516.

[2] Q. Chen, R. Yuan, X. Deng, P. Guo, and Z. Xiao, "Shunt active power filter with enhanced dynamic performance using novel control strategy," *IET Power Electron.*, vol. 7, no. 12, pp. 3169–3181, 2014.

[3] J. Solsona, C. A. Busada, and S. Gomez Jorge, "Reduced order generalised integrator-based current controller applied to shunt active power filters," *IET Power Electronics*, vol. 7, no. 5. pp. 1083–1091, 2014.

[4] B. Singh and J. Solanki, "A comparison of control algorithms for DSTATCOM," *IEEE Trans. Ind. Electron.*, vol. 56, no. 7, pp. 2738–2745, 2009.

[5] S. Golestan *et al.*, "Conventional Synchronous Reference Frame Phase-Locked Loop is an Adaptive Complex Filter," *IEEE Trans. Power Electron.*, vol. 32, no. 3, pp. 1713–1719, 2016.

[6] H. Saxena, A. Singh, and J. N. Rai, "Design and performance analysis of generalised integrator-based controller for grid connected PV system," *Int. J. Electron.*, vol. 105, no. 7, pp. 1079–1096, 2018.

[7] Y. Singh, I. Hussain, B. Singh, and S. Mishra, "Real time implementation of EPLL with generalized filtering in single phase grid interfaced SPV system," in *2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering, UPCON 2016*, 2017, pp. 50–54.

[8] T. Patil and S. Ghorai, "Robust zero-crossing detection of distorted line voltage using line fitting," *2016 Int. Conf. Electr. Electron. Commun. Comput. Optim. Tech. ICEECCOT 2016*, vol. 92, no. 1, pp. 92–96, 2017.

[9] M. Azrik, K. H. Ahmed, S. J. Finney, and B. W. Williams, "Voltage synchronization scheme based on zero crossing detection for parallel connected inverters in AC microgrids," *IEEE Int. Symp. Ind. Electron.*, pp. 588–593, 2012.

[10] M. S. Reza, M. Ciobotaru, and V. G. Agelidis, "A robust frequency estimation technique based on three consecutive samples for single-phase systems," *IEEE J. Emerg. Sel. Top. Power Electron.*, vol. 2, no. 4, pp. 1049–1058, 2014.

[11] Z. Dai, M. Fan, H. Nie, J. Zhang, and J. Li, "A Robust Frequency Estimation Method for Aircraft Grids under Distorted Conditions," *IEEE Trans. Ind. Electron.*, vol. PP, no. c, pp. 1–1, 2019.

[12] T. J. Mccabe, "A Complexity Measure," *IEEE Trans. Softw. Eng.*, vol. SE-2, no. 4, pp. 308–320, 1976.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 19,2020 at 06:07:50 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Dispersion Compensation in a 64-Channel Dense Wavelength Division Multiplexing Network

Anshul Dagar
Deptt. of Electronics and Communication
Delhi Technological University
Delhi, India
anshuldagar004@gmail.com

Akhil Dixit
Deptt. of Electronics and Communication
Delhi Technological University
Delhi, India
akhil.dixit.ec1998@gmail.com

Akhil Kumar
Deptt. of Electronics and Communication
Delhi Technological University
Delhi, India
kumarakhil663@gmail.com

Dr. Gurjeet Kaur
Deptt. of Electronics and Communication
Delhi Technological University
Delhi, India
gurjeet_kaur@rediffmail.com

Abstract—In optical fiber communication, information is relayed from the transmitter and received at the receiver where optical signals are converted into electrical signals. In recent years the demand for bandwidth has increased enormously. To meet the demand of increasing bandwidth with higher data rates led to dense wavelength division multiplexing (DWDM), an advanced multiplexing technique in which multiple signals can be transmitted simultaneously for various wavelengths in the same channel. Pulse broadening due to dispersion results in signal quality deterioration. Dispersion should be minimized during long haul communication. This paper reviews a 64 channel optical communication DWDM network which is simulated and analyzed for transmission lengths up to 300km, and for Signal Powers from -2dBm to 6dBm using various compensating schemes, broadly using Dispersion Compensating Fibers and Fiber Bragg Grating types of compensation schemes. Modified Duobinary Return to Zero (MDRZ) encoding format is used in the transmitter. The software used for simulation purposes is Optisystem 16 and other software used are OriginPro8 and Visio. The simulated results of the maximum Q-factor and Minimum Bit Error Rate (BER) value indicate that a combination of symmetric compensating fibers and Fiber Bragg Grating (FBG) result in better performance at larger transmission distances when parameters mentioned above and eye diagrams are compared.

Index Terms—DWDM, dispersion compensation, fiber bragg grating(FBG), DCF, MDRZ, optisystem 16

I. INTRODUCTION

For long-distance transmission, both capacity and rate are required to meet the demand. The transport characteristic of an optical fiber is determined by attenuation, dispersion, and non-linear effects. High data rate in the optical transmission is achieved by three compensation techniques i.e. pre, post, and symmetric [1]- [5] using dispersion compensating fiber (DCF) at a rate of 40 Gbps using MDRZ encoding format and Fiber Bragg Grating. The capacity of the system is decreased by dispersion and loss in the transmitting path.

In long-distance transmission, the shape of the pulse spreads by the time it reaches the receiver. This spreading of the pulse is known as dispersion. Different types of dispersion are Chromatic dispersion, Material dispersion, and waveguide dispersion. Different wavelength signals have different speeds in the optical fiber which leads to chromatic dispersion. Different materials show different refractive index for different wavelengths in Material dispersion. Wave-guide dispersion occurs because of the pulse spreading due to different refractive indexes of the core and the cladding. The transmission lengths are limited by chromatic dispersion as follows:

$$L \leq 2\pi c/(16 \mid D \mid \lambda^2 B^2) \quad (1)$$

where L denotes Transmission length, c is the speed of light in vacuum, B is the bit rate, D is dispersion coefficient.

MDRZ has substantial tolerance for dispersion and lower linear tolerance as it has much narrower optical bandwidth than duobinary return to zero (DRZ) [7]. To increase the capacity of the system we have used the Dense Wavelength Division Multiplexing(DWDM) network [8]. Here dense signifies that the wavelength channels are very close to each other. In DWDM, multiple signals are combined and transmitted simultaneously at different wavelengths on the same fiber. Each fiber has several parallel channels and each channel using a different wavelength from the other. [11]- [12]

Jagjit Singh Malhotra, Manoj Kumar, et al. [13] worked on the performance enhancement strategy for a 32 Channel DWDM soliton link using Electronic Dispersion Compensation. Symmetrical type optical dispersion compensation scheme has been customed to reduce the accumulated dispersion within the fiber span using DCF to mitigate the four-wave mixing power by implementing certain calculated destructive interference between the beginning and ending distribution of the fiber optical link. Sumitpal Singh, Karamjeet Kaur [10] work on performance evaluation of the DWDM system with dispersion compensation was studied and compared to the older implementations. They analyzed the effect of fiber-optic dispersion on optical transmission systems. A comparative analysis of the performance of dispersion compensation in WDM network

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

using EDFA was performed on a 16 channel 25 Gbps system. Hodzic, A., Konrad [14] worked on implementing different modulation formats on a 40 GBPS WDM network. They utilized the use of carrier suppressed return to zero modulation scheme to obtain optimum high bit and data rate optical transmission. Hoshida, T., Vassilieva, et al. [15], research paper on the study of modulation formats to be employed for a 40GBPS long haul DWDM network was studied and comparison between NRZ and MDRZ was performed.

We have preferred single-mode fiber(SMF) over multi-mode fiber in our analysis as it shows less dispersion. SMF, when used in 1310 nm, shows dispersion which is close to zero, but in 1550 nm has the smallest loss. To receive good power signals at the receiving end, we need to use optical amplifiers in the transmission link. Hence we have used the Erbium-Doped Fiber Amplifier as EDFA can amplify multiple optical signals simultaneously and can be used with DWDM.

One of the advanced technology used for dispersion compensation is a fiber bragg grating technique in which the optical fibers of shorter transmission lengths reflect the backlight signal of a specific wavelength. FBG has zoned in the fiber, periodically spaced from each other that have refractive indices somewhat higher than the refractive index of the fiber core. This formation reflects a particular range of wavelengths while passing others. [5]- [6]

Dispersion shortens the distance that the signal travels inside the optical fiber and hence effectively blurs the signal. A Dispersion compensating fiber is employed to compensate and reduce the accumulated dispersion in an SMF. To handle this range of values, the Dispersion Compensating Module is set up using an exclusive type of dispersion compensating fiber situated inside a module having a negative value of dispersion coefficient to reduce dispersion, ranging from -35 to -305 ps/(nm*km). Patel, Gaurang H., et al. [17] proposed approach on the dispersion compensation in a WDM network and Pal, Raju and Sharma, Vishal [18] work on the comparison of dispersion Compensation Schemes for 10 GBPS at Various Optical Power Levels using Dispersion Compensated Fibers, was studied and reviewed. Belen Melian, Manuel Laguna, et al. [16], proposed a system that measures the dispersion compensation at various propagation distance and input power for nonlinearity analysis. Anu Sheetal, Ajay Sharma [19] worked upon the high capacity DWDM system. It was researched and its insight into the working of a fiber under Kerr's effect and resulting dispersion compensation was studied.

In all of the previous works, the number of channels and bandwidth speed was low. Jagjit Malhotra [13] stimulated a 32 channel DWDM at a low data rate. The analysis was limited to the use of compensating fibers. The number of users provided is low due to low channels. In Karamjeet Kaur [10], they analyzed dispersion compensating fiber at different lengths at a 16 channel transmitter. The number of users is quite less and the dispersion is low so accurate results can not be formed. Moreover, their analysis is limited to dispersion fiber analysis. In Patel [17], Vishal Sharma [18] their analysis was done at 10 GBPS transmitter. Dispersion at this speed is low and hence more practical results could not have been obtained. Moreover, their analysis was also limited to DCF.

We are proposing a 64 channel MDRZ encoded DWDM system in place of the basic NRZ/RZ/DRZ encoded DWDM systems for better performance in long haul communication and observing the various dispersion compensation schemes on it. MDRZ encoding results in substantial tolerance for dispersion and lower linear tolerance. 64 channel DWDM system helps us analyze communication for more users which can make it easy for generalization of the results for a higher number of channels/users as the results will be more precise for a higher number of channels than the standards of 4/8 or 16 channels.

II. SYSTEM DESIGN

While creating an optical system the aspects of system design and formation are :

1. The design of the MDRZ modulated transmitter.
2. The respective optical transmission span layouts for different dispersion compensation schemes.
3. The receiver.
4. A dedicated system for dispersion compensation schemes and FBG.
5. Various system parameters for MDRZ and optical transmitter.

With the complete analysis in hand, it will be really helpful to decide on what dispersion compensations to use on the modified duobinary return to zero formats suited for 2.56 Tbps data(64*40Gb/s) over an optical span of 300km with channel spacing of 100Hz as per the restrictions of Signal Powers and higher reaches of transmission distances.

Fig 1. provides an outlook of a 40 Gb/s modified duobinary return to zero (MDRZ) transmitter. The first step is to have an NRZ duo-binary signal, which is further modified, however here we have a time delay and subtract circuit rather than the time delay and add a circuit. The produced result is passed through the initial Mach Zender Modulator, whose output is concatenated with a secondary modulator which is driven by an oscillating electrical signal with forty gigahertz frequency. In DRZ signal case, we adjust the bits "1"s, once a small amount "0" appears, however within the MDRZ case, we tend to alternate the section between zero and zero for the bits "1" i.e. keeping the part of all the "zero" bits constant, a 180° part out of phase variation is done between all the consecutive "ones" that results in suppression of carrier of the DB signal. The input to the circuit is provided by CW laser and the output is given to an ideal mux. The Modified duobinary pulse obtained has a frequency response represented as :

$$S(f) = H_N(f) - exp(-j4\pi fT)H_N(f) \tag{2}$$

i.e.

$$S(f) = jH_N(f)sin(2\pi fT)exp(-j2\pi fT) \tag{3}$$

where

$$H_N(f) = 1/2Wrect(1/2W) \tag{4}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

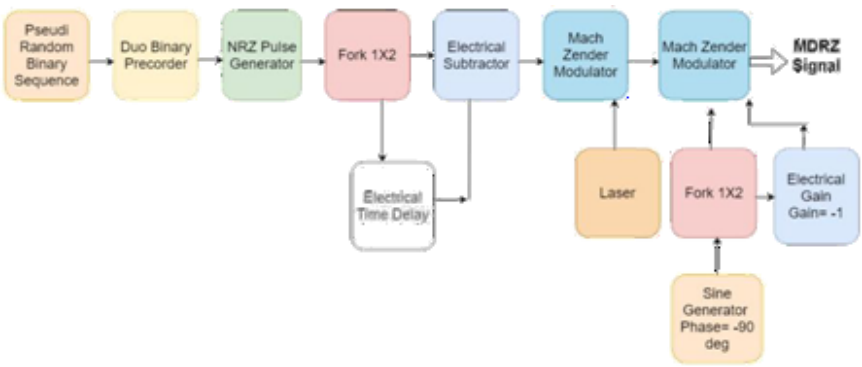


Fig. 1. MDRZ Encoder

$$s(t) = sinc(t/T) - sinc((t - 2T)/T) \quad (5)$$

2W symbols/sec being the Baud rate.
In the time frame, the pulse is represented as :

The multiplexed optical signal from the ideal mux is then transmitted through SMF. The model considers unifacial signal flow, dispersion effects. To suppress the negative effects of dispersion over the link, a dispersion compensating fiber(DCF) is utilized [9]. While extending the DCF there exist three scenarios where DCF can be extended to the fiber-

1. Incorporating before the single-mode fiber (SMF) known as a pre-compensation scheme.
2. Incorporating after the SMF known as a post compensation scheme.
3. Incorporating between the split SMF known as symmetrical compensation scheme

These configurations are further explored and also the better one to produce results is employed in our analysis. The transmitting length of the DCF is calculated by analyzing the dispersion accumulated in the SMF. The gain of Erbium-Doped Fiber Amplifier (EDFA) is set in such a manner as to control the attenuation when the signal propagates over both the fibers namely SMF and DCF. The gain is 7.5 dB for SMF and 10 dB for DCF. The proposed network with dispersion compensation via DCF is achieved as given in Fig. 3 (Pre Compensation).

On the receiver side we have the following components:
Demux, PIN diode, Filter, and 3R regenerator. On all the 64 output ports of the optical demultiplexer second-order Bessel bandpass filters are utilized that have characteristic of having 3 units providing an end frequency of 40 GHz. The electric output from PIN diode then passes through a fourth-order Bessel based electrical LPF whose cut-off frequency depends on the modulation format used. Further, a 3R binary regenerator is attached to the BER analyzer that generates results. The proposed network with dispersion compensation via FBG is achieved as given in Fig. 4(Post Configuration). Other than the represented configuration, there are other configurations namely Pre-Configuration, where FBG is placed to signal before passing through the multiplexer (MUX). Other one is a combi-national configuration where symmetric compensated fibers and FBG are used together.

This timing pulse is shown in Fig.2.

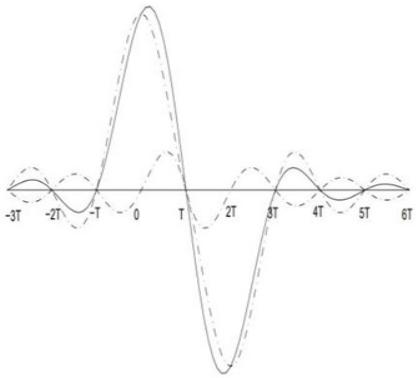


Table 1. shows the transmission simulation parameters used in the design of our DWDM network. Table 2. shows the fiber specifications for all compensation schemes.

III. RESULTS AND DISCUSSION

To compare the performance of different dispersion compensation schemes in the said 64-channel MDRZ encoded DWDM system, the maximum Q-factor, and minimum bit error rate(Min. BER) are the two important parameters considered. Q-factor represents the combination of optical and electrical Signal to Noise Ratio (SNR) for the optical communication system and provides a convenient measure of overall system quality. BER measurement, on the other hand, is a direct indication of system performance. Firstly, the transmission lengths are varied keeping the signal power constant and the two parameters mentioned above are studied. Then, the Signal Power is varied keeping the transmission length constant and again the two parameters mentioned above are studied. Hence, a two-fold analysis helps us understand the efficiency and system performance of the proposed system.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

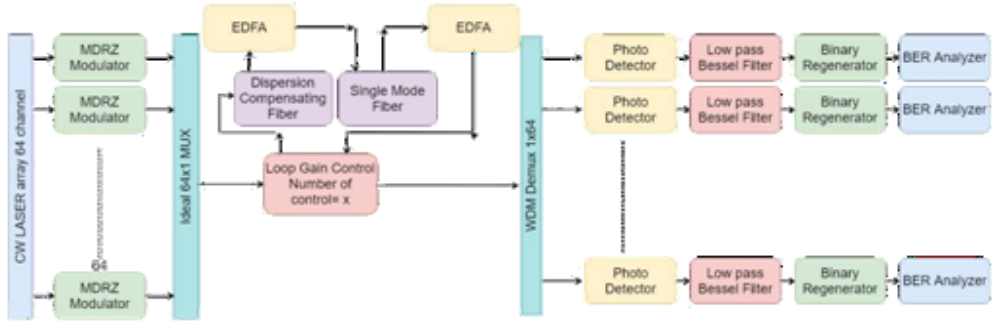


Fig. 3. 64-channel MDRZ encoded DWDM Network with DCF

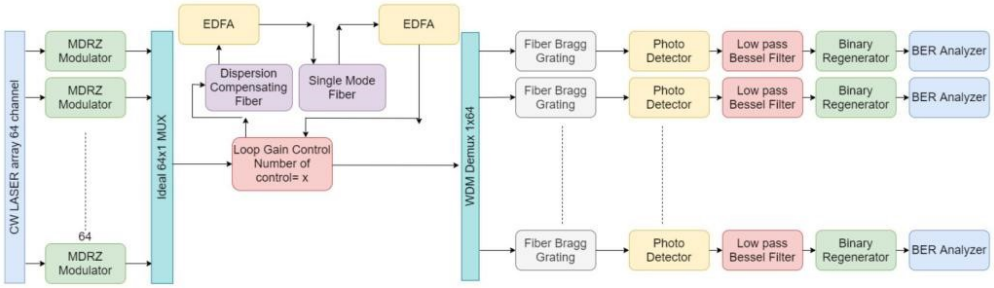


Fig. 4. 64-channel MDRZ encoded DWDM Network with Post-configured FBG

TABLE I
SIMULATION PARAMETERS

S.NO	SIMULATION PARAMETERS	MAGNITUDE
1.	Bit Rate	40Gbps
2.	Sequence Length	128
3.	Samples Rate	64
4.	Channel Spacing	100Hz
5.	Central Frequency(Main Channel)	1500nm
6.	Distance	60*N km
7.	PIN Responsivity	1 A/W
8.	Dark Current	0.1 nA
9.	LPF Cut-off Frequency	0.8*Bit Rate

TABLE II
FIBER PARAMETER SPECIFICATIONS

PROPERTY	SMF	DCF
Attenuation(dB/km)	0.2	0.5
Dispersion(ps/km-nm)	17	-85
Dispersion Slope(ps/km-nm^2)	0.08	-0.45
Effective core area(m^2)	80	30
Nonlinear refractive index	2.6	2.6

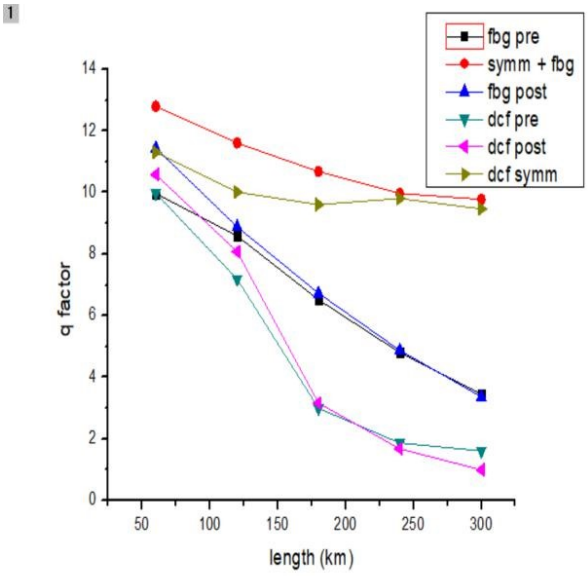


Fig. 5. Q-factor vs Transmission length

A. ANALYSIS OF SYSTEM FOR DIFFERENT DISPERSION SCHEMES AT CONSTANT POWER(0 dBm) FOR VARYING TRANSMISSION LENGTHS

Table 3. shows the obtained simulated values of the Q-factor of different dispersion compensation schemes for varying transmission length. From the said table, the graph between Q-factor and Transmission length(km) is plotted using ORIGIN-PRO 8 software.

From Fig 5., it can be seen that the Q-factor value of all the dispersion compensation schemes decrease as the transmission lengths increase from 60 to 300 km. Among the Dispersion Compensating Fiber schemes, the Symmetric configured DCF scheme outperforms the remaining Pre and Post Configured DCF. It is the second-best in performance among all the schemes studied. Among the FBG used schemes, the combination of symmetric configured fibers and FBG

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

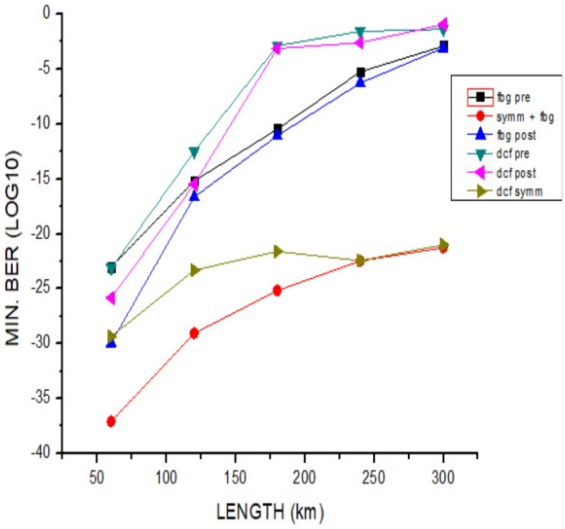


Fig. 6. Min. BER vs Transmission length

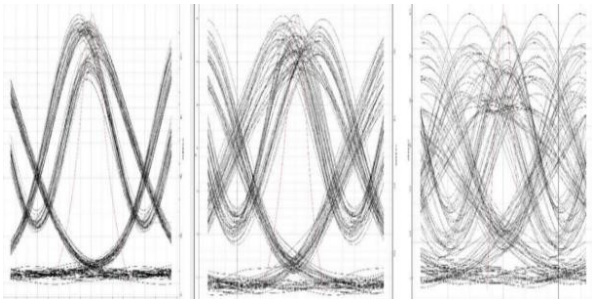


Fig. 7. POST CONFIGURED DCF for a) 60 b) 180 c)300 km respectively

outperforms the remaining Pre and Post configured FBG. The combination of Symmetric configured fibers and FBG in the proposed system outperforms all the schemes and is the best for long haul communications and Pre and Post Configured DCF schemes are least preferred.

Table 4. shows the obtained simulated values of the Minimum BER in the log10 scale of different dispersion compensation schemes for varying transmission lengths. From the said table, the graph between Min. BER and Transmission length (km) is plotted using ORIGINPRO 8 software. From Fig 6., it can be observed that the minimum BER values of all the dispersion compensation schemes become less negative as we increase the transmission length from 60 to 300km. More negative the BER value is better is the system performance. Here too the combination of symmetric configured fiber and FBG outperforms all the remaining schemes. The Symmetric con-figured DCF Scheme being second best at longer transmission lengths and POST configured FBG scheme being second best at the length of 60km. POST configured DCF scheme performs fairly decent at lower length but degrades as transmission length is increased. PRE configured DCF scheme performs the

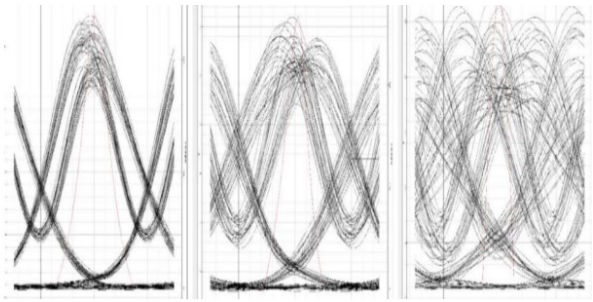


Fig. 8. PRE CONFIGURED DCF for a) 60 b) 180 c)300 km respectively

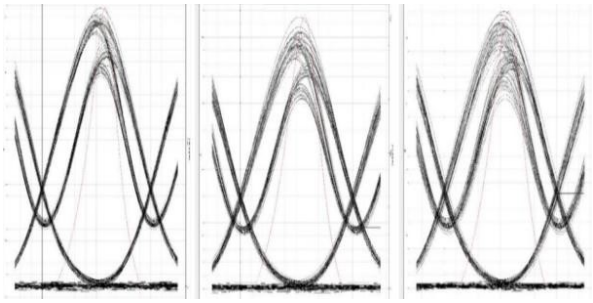


Fig. 9. SYMMETRIC CONFIGURED DCF for a) 60 b) 180 c)300 km respectively

worst and is least preferred for communication in the proposed system.

Fig. 7-12 show the Eye Diagrams of all six dispersion compensation Schemes for 60, 180, 300 km transmission lengths respectively. Now, maximum is the eye-opening, more is the eye closure and better is the performance. From the above results obtained, FBG schemes have better dispersion compensation performance than DCF schemes. However, the symmetrically configured DCF scheme outperforms the Pre and Post configured FBG compensation schemes. The combination of Symmetric configured fibers and FBG results in the best eye diagram and hence the best performance for long haul communications.

B. ANALYSIS OF SYSTEM FOR DIFFERENT DISPERSION SCHEMES AT CONSTANT TRANSMISSION LENGTHS(60km) FOR VARYING POWERS

Table 5. shows the obtained simulated values of the Q-factor of different dispersion compensation schemes for varying Signal Powers(dBm)

From the said table, the graph between Q-factor and Signal Power(dBm) is plotted using ORIGINPRO 8 software.

From Fig.13, an irregular pattern is observed in the value of Q-factor as the value of Signal Power increases from - 2 to 6 dBm. At -2dBm, Post configured FBG compensation scheme shows the best signal quality which it repeats at 0dBm and 2dBm. At 4dBm and 6dbm powers, the combination of symmetric configured fibers and FBG provides better compensation as per the performance. Use of symmetric compensated DCF scheme is the second-best option after the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

TABLE III
Q-FACTOR FOR TESTED DIPERSION COMPENSATION SCHEMES FOR CHANGING TRANSMISSION DISTANCES

Length	Q-factor	Q-factor	Q-factor	Q-factor	Q-factor	Q-factor
Km	FBG Pre	Symm+FBG	FBG Post	DCF Pre	DCF Post	DCF Symm
60	9.9651	12.792	11.4446	9.9808	10.5868	11.3006
120	8.58715	11.6092	8.88537	7.1894	8.07758	10.0114
180	6.52015	10.6783	6.72986	2.9917	3.1631	9.60376
240	4.80178	9.96673	4.87211	1.86247	1.6703	9.8057
300	3.45436	9.77563	3.36012	1.60723	0.9845	9.4632

TABLE IV
MINIMUM BER OF ALL TESTED DISPERSION COMPENSATION SCHEMES IN LOG10 SCALE FOR CHANGING TRANSMISSION DISTANCES

Length	MIN. BER	MIN. BER	MIN. BER	MIN. BER	MIN. BER	MIN. BER
Km	FBG Pre	Symm+FBG	FBG Post	DCF Pre	DCF Post	DCF Symm
60	-23.00363	-37.07353	-29.91648	-23.12583	-25.84217	-29.34248
120	-15.17645	-29.0567	-16.63946	-12.49998	-15.48565	-23.30644
180	-10.4712	-25.17409	-11.07297	-2.86525	-3.11358	-21.58554
240	-5.25717	-22.47513	-6.2599	-1.56293	-2.5905	-22.43492
300	-2.88784	-21.2389	-3.06164	-1.35937	-0.93262	-20.99217

TABLE V
Q-FACTOR OF ALL TESTED DISPERSION COMPENSATION SCHEMES FOR CHANGING SIGNAL POWERS

Signal Power	Q-factor	Q-factor	Q-factor	Q-factor	Q-factor	Q-factor
dBm	FBG Pre	Symm+FBG	FBG Post	DCF Pre	DCF Post	DCF Symm
-2	8.98885	10.4493	11.6171	9.44951	9.86233	11.0923
0	9.9651	11.4352	11.4446	10.1119	10.7201	11.3682
2	10.7840	12.4520	12.5542	9.84251	10.0698	11.4189
4	10.9382	12.9357	12.8167	10.7650	10.7437	11.4465
6	10.9383	13.2507	13.0097	10.6494	10.2356	12.2399

TABLE VI
MIN. BER OF ALL TESTED DISPERSION COMPENSATION SCHEMES IN LOG10 SCALE FOR CHANGING SIGNAL POWERS

Signal Power	Min. BER	Min. BER	Min. BER	Min. BER	Min. BER	Min. BER
dBm	FBG Pre	Symm+FBG	FBG Post	DCF Pre	DCF Post	DCF Symm
-2	-18.93642	-26.7375	-30.78489	-20.8966	-22.61292	-28.36668
0	-23.00363	-29.07353	-29.91648	-23.71911	-26.48965	-29.70859
2	-26.71494	-35.27011	-35.73765	-22.54666	-23.53989	-29.96832
4	-27.45001	-36.05212	-37.19406	-26.71088	-26.59702	-30.09858
6	-27.45103	-38.36648	-38.28461	-26.18197	-24.26478	-34.22813

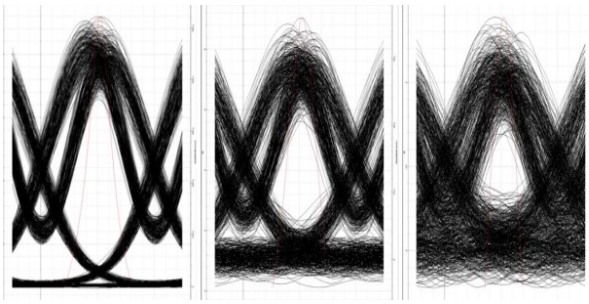


Fig. 10. PRE CONFIGURED FBG for a) 60 b) 180 c) 300km respectively

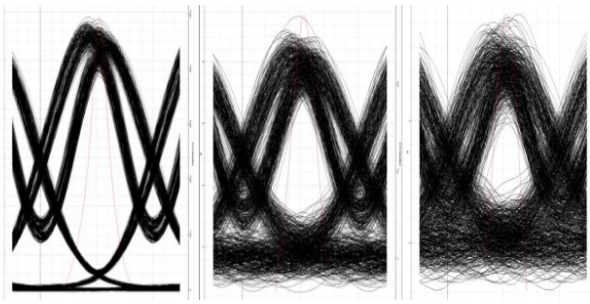


Fig. 11. POST CONFIGURED FBG for a) 60 b) 180 c)300 km respectively

ones specified before. While PRE configured DCF scheme and POST configured DCF schemes are least preferred, PRE configured FBG compensation scheme improves performance on increased powers of 2,4 and 6 dBm.

Table 6. shows the obtained simulated values of the minimum BER in the log10 scale of different dispersion compensation schemes for varying Signal Powers(dBm). From the said table, the graph between min. BER and signal power(dBm) is plotted

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

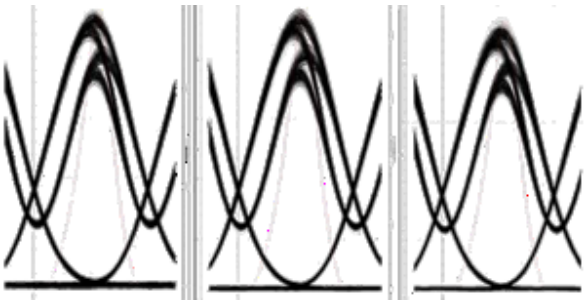


Fig. 12. SYMMETRIC CONFIGURED FIBERS + FBG for a) 60 b) 180 c)300 km respectively

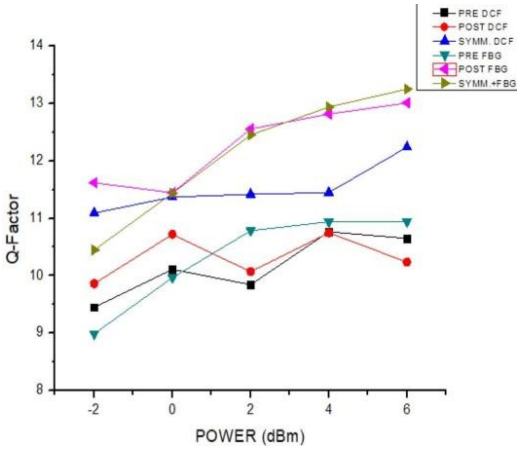


Fig. 13. Q-factor vs Signal Power

using ORIGINPRO 8 software.

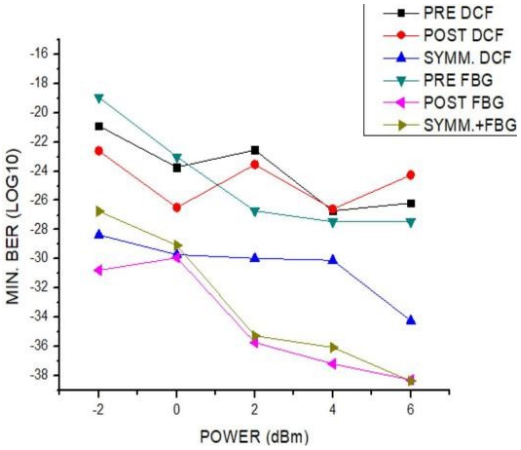


Fig. 14. Min. BER vs Signal Power

From Fig 14, it can be seen that the results obtained in the calculation of Min. BER vs signal power shows similar trends in performance as that of Q-factor against Signal Power.

IV. CONCLUSION

For the proposed 64-channel MDRZ encoded DWDM net-work, a two-fold analysis was conducted, i.e. analysis at constant signal power for varying transmission lengths and analysis at constant transmission length for varying signal powers and it is observed and recorded that the combination of symmetric compensated DCF and FBG compensation scheme results in the best performance in the case of varying transmission lengths at a constant signal power for the available basic SMF fibers whose parameters are mentioned in table 2.

Apart from it, the Use of Post configured FBG and Symmetric compensated DCF are also two schemes which perform relatively well for the purpose and rank next to the combinational scheme and the rest should not be even considered for long haul communications. For the second part of the analysis, we observe that if we consider constant transmission length and vary the signal powers, then Post-FBG configuration out-beats the performance of the scheme having a combination of FBG and DCF used together but only at lower powers. At higher power levels the combination outperforms again. Hence Post FBG scheme is a good option to use when we need smaller transmission lengths and have the flexibility to use varying signal powers (-2dBm-4dBm). At further higher powers of 6dBm, combination outperforms.

REFERENCES

[1] Mehtab Singh, "Different Dispersion Compensation Techniques in Fiber Optic Communication System : A Survey", International Journal of Advanced Research in Electronics and Communication Engineering, vol: 4, issue: 8, pp. 123-128, 2015

[2] R. Paxase, et al., "Dispersion post compensation using DCF at 10gbps", Global journal of computer science and technology, vol. 11, issue 3, version 1.0,2011.

[3] K. Khairi, et al., "Investigation on the Performance of Pre and Post Compensation Using Multi Channel CFBG Dispersion Compensators", IEEE International RF and Microwave Conference (RFM 2011), 12th - 14th December 2011.

[4] Ranjeet Kaur, Gaurav Mittal, " Comparative Study of Dispersion Compensating Techniques Pre, Post and Mix with 8 Channels", International Journal of Advanced Research in Electrical, Electronics and Instrumenta-tion Engineering, vol. 6, issue 7, 2017 75

[5] Gagandeep, Saxena et al, "Dispersion compensation using FBG and DCF in 120 Gbps WDM systems", International Journal of Engineering Science and Innovative Technology (IJESIT), vol. 3, 2014

[6] N.M. litchinitser, D.B. Patterson, "Analysis of Fiber Bragg Gratings for Dispersion Compensation in Reflective and Transmission Geometries", Journal of Lightwave Technology, vol. 15, no. 8, pp. 1303-1313, 1997

[7] Urvashi Bansal, Kamaldeep Kaur, "1.6 Tbps High Speed Long Reach DWDM System by incorporating Modified Duobinary Modulation Scheme", International Journal of Current Engineering and Technology, vol. 4, no. 6, 2014

[8] Yadav, Mulayam, " Design Performance of High Speed Optical Fiber WDM System with Optimally Placed DCF for Dispersion Compensation", International Journal of Computer Applications, vol. 122, 2015

[9] Kaur, Manpreet et al, "Dispersion Compensation with Dispersion Compensating Fibers", International Journal of Advanced Research in Com-puter and Communication Engineering, vol. 5, 2015

[10] Sumit Pal, Karamjit Kaur, "Analysis on Dispersion Compensation of DWDM System with DCF and Various Modulation Formats", International Journal Of scientific Engineering Research, vol. 4, 2013

[11] Muhammad Hammad Ashraf, "Simulative Design of DWDM System Using Different Dispersion Compensation Techniques", International Jour-nal of Scientific Engineering Research, vol. 8, issue 10, 2017

[12] M.Tosson1 , Walid S. El-Deeb et al, "Dispersion Compensation Techniques for DWDM Optical Networks", International Journal of Advanced Research in Computer and Communication Engineering vol. 4, Issue 8, 2015

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

[13] Malhotra, Jagjit Kumar, Manoj Sharma, Ajay, "Performance Enhancement Of 32 Channel Long Haul Dwdm Soliton Link Using Electronic Dispersion Compensation", International Journal of Electronics, Communication Instrumentation Engineering Research and Development (IJECIERD), vol. 2. pp. 11-16, 2012
[14] A. Hodzic, Konrad, "Alternative modulation formats in N/spl times/40 Gb/s WDM standard fiber RZ-transmission systems", Journal of Lightwave Technology, vol. 20, issue 4
[15] T. Hoshida, O. Vassilieva, K. Yamada, S. Choudhary, R. Pecqueur, H. Kuwahara, Optimal 40 Gb/s modulation formats for spectrally efficient long-haul DWDM systems, IEEE J. Lightwave Technol. 20 (12) (2002) 76
[16] Belen Melian , Manuel Laguna , Jose A. Moreno-Perez ," Capacity expansion of fiber optic networks with WDM systems: problem formulation and comparative analysis". Elseiver Computers Operations Research 31 (2004) 461–472
[17] Gaurang.H.Patel, Rohit.B.Patel, Sweta.J.Patel, "Dispersion compensation In WDM network using dispersion compensating fiber", Journal of Information Knowledge and Research In Electronics and Communication Engineering vol. 2,Issue 2
[18] Rajani, Rajani Pal, Raju Sharma Vishal. (2012). Comparison of Pre-, Post- and Symmetrical-dispersion Compensation Schemes for 10/15 GBPS using Different Modulation Formats at Various Optical Power Levels using Standard and Dispersion Compensated Fibers. International Journal of Computer Applications. 50. 6-13. 10.5120/7925-1199.
[19] Sheetal, Anu Sharma, Ajay Kaler, R. S. (2010). Simulation of high capacity 40Gb/s long haul DWDM system using different modulation formats and dispersion compensation schemes in the presence of Kerr's effect. Optik - International Journal for Light and Electron Optics. 121. 739-749. 10.1016/j.ijleo.2008.11.009.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:50 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Dynamic Traffic Management System Using Neural Network based IoT System

R.K. Yadav
Computer Engg. Deptt.
Delhi Technological University
Delhi, India
(rkyadav@dtu.ac.in)

Rishabh Jain
Computer Engg. Deptt.
Delhi Technological University
Delhi, India
(rishabhjain7894@gmail.com)

Sarvjit Yadav
Computer Engg. Deptt.
Delhi Technological University
Delhi, India
(sarvjityadav1997@gmail.com)

Satyam Bansal
Computer Engg. Deptt.
Delhi Technological University
Delhi, India
(satyambansal183@gmail.com)

***Abstract* - A concept to make a traffic signal management system with the help of genetic algorithms and the Internet of Things, to reduce the average waiting time for vehicles on traffic signals is proposed. Several scenarios with varying traffic density are simulated and a neural network to allot green light time to each road is trained by neuroevolution using genetic algorithms and self-adaptive genetic algorithms for minimising learning time, increasing accuracy and better generalisation. Based on results a model is proposed and performance data is analysed.**

Keywords - Internet of things, neuro-evolution, genetic algorithms, self-adaptation, simulation

I. INTRODUCTION

The general working of traffic lights requires more than just simple control and coordination of traffic light timers to make sure that vehicles move as quickly and safely as possible. With the increase in the number of vehicles in the cities, the old system of fixed intervals of traffic lights seems inappropriate and not feasible. Several different control systems are used to accomplish this, varying from simple clockwork mechanisms to sophisticated(smart) computerized control and coordination systems that self-adjust to minimize delay to people using the junction. Such computerized control and coordination systems or smart traffic management systems are required to be adaptable and easily trainable in varying environments. This paper presents one such model, trained by using self-adaptive evolutionary techniques on neural networks. A traffic light control system is proposed that can automatically analyze the traffic by using sensors and IoT and dynamically adjust the time for a particular traffic signal. The following sections present respectively – the model to which the self-adaptive evolutionary techniques were applied; data-based and graphical comparison of various aspects such as training time and performance of models using traditional parameter control versus models using self-adaptive parameter control.

II. LITERATURE REVIEW

With the advent of IoT, there have been many developments in the field since then. Many researchers have proposed various approaches to solve different problems in this domain.

Techniques used to get the data for simulation or analysis for traffic management systems generally involve the use of sensors and IoT. In most cases, traffic density[4] is taken as input from cameras that use digital image processing and sensor data. In some cases, the number of vehicles on a particular roadside and the importance of roads, determine the input of the system. Sometimes fire and smoke sensors deployed on the road detect fire, explosion, smoke, etc. There is also the use of Raspberry pi, IR sensors, and LCD. RFID[9] (Radio Frequency Identification) tags are used to prioritize emergency vehicles. Some researchers have used modules to describe the process such as:- Module 1 (Getting sensor data, the output is digital data), Module 2 (Perform Analysis, the output decision is taken), Module 3 (Send data on the cloud, the output is a new value for the signal countdown), Module 4 (Send data to the cloud, data is stored on the cloud), Module 5 (Display traffic data on the Web, the output is charts, graphs displayed on the Web).

Since there are many techniques to take the input, similarly there are many ways in which researchers did analysis using that data. Some use actual data (if available) and others use the simulation for analysis[3]. In some cases, an emulator is developed in Java[2] for representing dynamic traffic conditions. Machine learning algorithms such as Neural Networks[8], KNN (K-nearest neighbor)[1] have been used to find optimal waiting times or traffic signal intervals. A mobile application connected to the central server communicates with the rescue department in case of an emergency. In most cases, Web application gives information in graphical formats for future road planning. The analysis process includes data analysis, GIS (Geographic Information System), Information and Communication Technologies[11] (ICT), real-time analysis,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:44 UTC from IEEE Xplore. Restrictions apply.

VANET scenario (Vehicular Adhoc Networks), Machine to Machine communication[10] (M2M), fuzzy inference system, cellular automata, software like MATLAB[2], Python, MQTT (Message Queuing Telemetry Transport) Protocol, HTML, CSS, Javascript, Blynk application, GPS (Global Positioning System)[5][6].

The aim of most of the traffic management systems is generally the same. They include reducing congestion, reducing the waiting time of vehicles, reducing accidents, and fuel consumption, allowing emergency vehicles to reach the destination in minimum possible time, improving QoS[12] (Quality of Service).

III. METHODS AND MATERIALS

This research paper and the following methodology are based on some assumptions like the availability of sensor data with spatial and temporal times not greater than the maximum of any signal time in the crossroad. Practically, sensors may include heat sensors, proximity sensors, etc. Another possible way of detecting queue length and traffic density is through image processing and vehicle recognition techniques. Sensors can be linked to each other through gateways and gateways can be linked to the cloud.

This research uses the SUMO (Simulation of Urban MObility) as a road traffic simulation to run the analysis. It was used due to the limited availability of the data related to the movement of vehicles and the overall road traffic system in many areas. So no data can be considered as full and reliable for dynamic situations of the roads and traffic. SUMO traffic analysis gives a near to practical crossroad situation and data. The choice of SUMO is a result of following reasons: SUMO supports re-routing, traffic lights evaluation, simulation of vehicular communication, reliable random input to a particular path, high interoperability using XML data, random and grid network possible and many other advantages over other software.

The core decision making is done by a neural network that takes traffic length on each road as input and allots optimal green time such that average waiting time for vehicles is minimized for every scenario. Currently, neuroevolution is used for estimating the weights of the neural network, taking inspiration from the neuroevolution for making bots to play video games such as flappy bird, snakes and apples, etc. The motive behind using neuroevolution is to make the system able to be trained as soon as it is deployed and thereby adapt according to the environment (particular traffic trends on different roads), such models can be compared with models trained by using reinforcement learning also. A major advantage of using such models is no requirement of training using data or simply back-propagation of errors to estimate weights of the neural networks. Possible shortcomings such as fairly large training time of deep Q - learning(reinforcement learning) and difficulty in deciding the reward/cost brought us to use a more promising approach to train neural networks i.e. neuroevolution.

Our procedure is as described;

A. Neuro-evolution procedure

1. Start
2. Initialize population size, number of generations, mutation probability, retention probability, selection probability, step sigma, change factor, history length
3. Initialize starting population of neural network with random weight
4. For loop start.
5. For generation in range(no_of_gnerations):
6. Run simulation for each neural network
7. Sort neural networks based on fitness value(waiting time)
8. A select fittest neural network as a parent for the next generation
9. Mutate parents to explore search space.
10. Generate more networks for the next generation.
11. For loop end.
12. Combine parents and children to form the population of the next generation.

B. Simulation Procedure

1. Start
2. Declare probabilities of vehicle arrival in each road per second and simulation run time.
3. Initialize network of a crossroad in SUMO.
4. Run the simulation using an input neural network as the central decision-making system which will take input as the queue length of vehicles on each road.
5. Observe the average waiting time at the end of the simulation.
6. Close the simulation and return the average waiting time.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:44 UTC from IEEE Xplore. Restrictions apply.

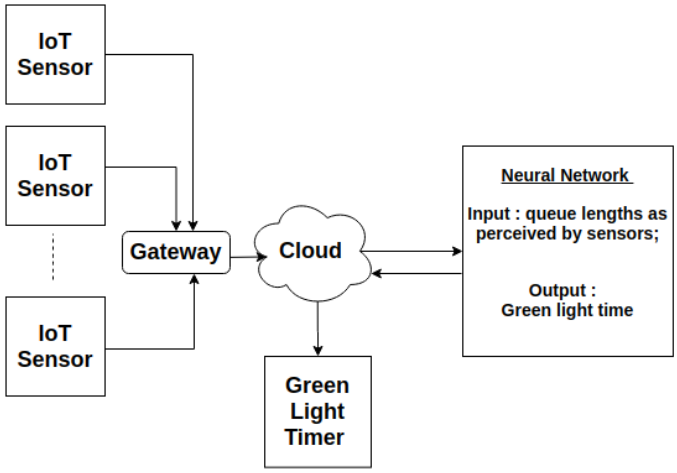


Fig 1. Functional Block Diagram of Proposed Model

Further, to use self-adaptive parameters, self-adaptivity means that some parameters of the evolutionary algorithm are changed during a run in a specific manner. This feature is inherent in modern evolution strategies.

Using a Gaussian or normal distribution with 0 mean and standard deviation σ to draw random numbers such that the property of normal distribution curve of decreasing

probability of picking a number of any magnitude as the number moves farther from mean as a function of σ is exploited.

Therefore the σ value is one such parameter of the algorithm that determines the extent to which current values x_i are altered by the mutation operator. Since it means the amount of change that happens to the current value in one mutation, σ is often called the mutation step size. The famous ⅕ success rule is used for achieving variable step size in mutations. According to this rule, the ratio of successful mutations (those in which a child has better fitness than it's the parent) to all mutations should be 1/5. Hence if the ratio is greater than ⅕ the step size should be increased to make larger steps to explore the search space, and in case the ratio is less than 1/5 then it should be decreased to control the randomness as the search is now moving further away from current region to a region of lesser fitness. The rule is executed at periodic intervals, for instance, after k iterations, each σ is reset by

$$\sigma = \begin{cases} \sigma/c & \text{if } p_s > 1/5, \\ \sigma * c & \text{if } p_s < 1/5, \\ \sigma & \text{if } p_s = 1/5 \end{cases} \quad (1)$$

where p_s is the relative frequency of successful mutations measured over several trials, and the parameter c is in the range $0.817 <= c < 1$.[7]

IV. DATA AND RESULTS

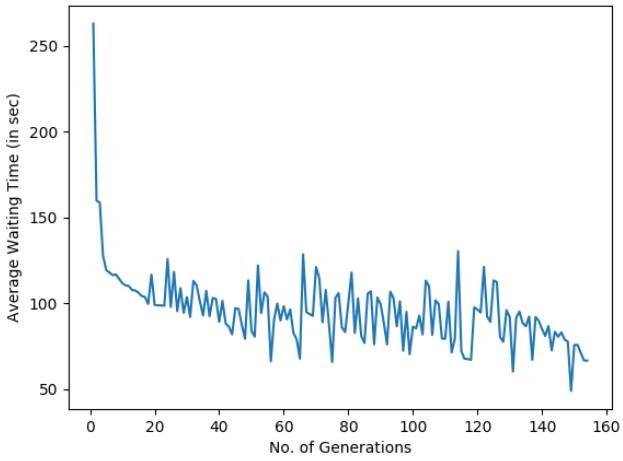


Fig 2. Average waiting time vs No. of generations (Whole population)

This Graph shows the average waiting time per vehicle of the whole population in every single generation. Here total traffic on all roads is 1.5 vehicle/sec. In 150 generations average waiting time reached about 40 sec/Vehicle.

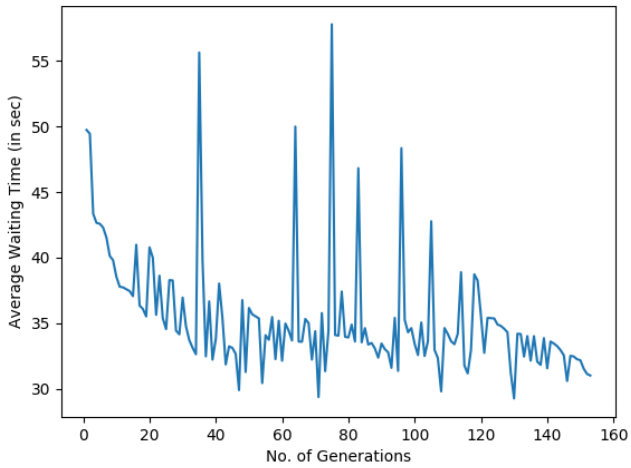


Fig 3. Average waiting time vs No. of generations (Generation best)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:44 UTC from IEEE Xplore. Restrictions apply.

This Graph shows the average waiting time per vehicle of the best neural network in every single generation. Here total traffic on all four roads is 1.5 vehicle/sec. After about 150 generations average waiting time reached about 30 sec/Vehicle.

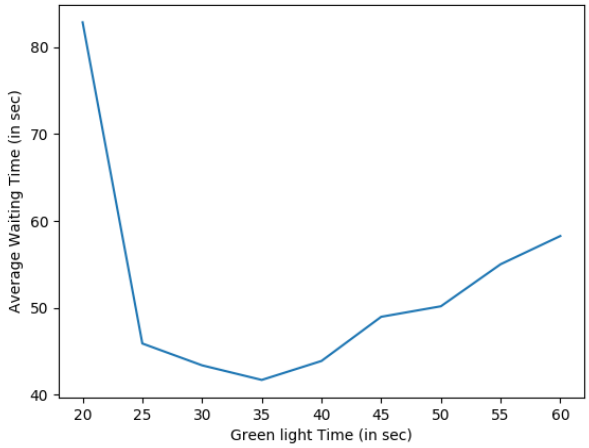


Fig 4. Average waiting time vs Green light time(sec)

This Graph shows the average waiting time per vehicle when a fixed amount of green light time is given in each road irrespective of traffic. Here total traffic on all four roads is 1.5 vehicle/sec. Among various possibilities least traffic is observed with green light time is 35 sec and the average waiting time per vehicle is around 42 sec.

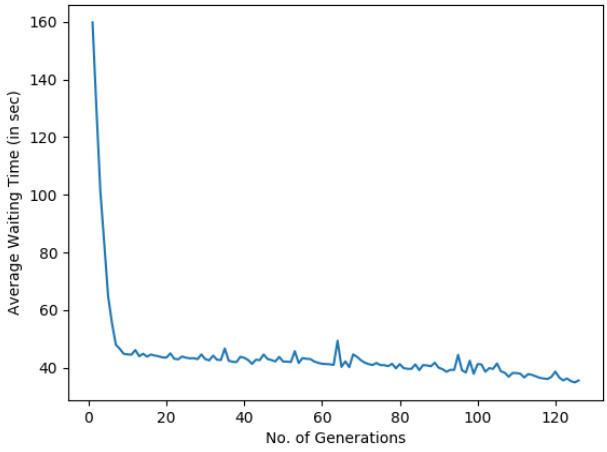


Fig 5. Average waiting time vs No. of generations (Whole population)

This Graph shows the average waiting time per vehicle of the whole population in every single generation. Here total traffic on all roads is 1.0 vehicle/sec. After about 130 generations average waiting time reached about 38 sec/Vehicle.

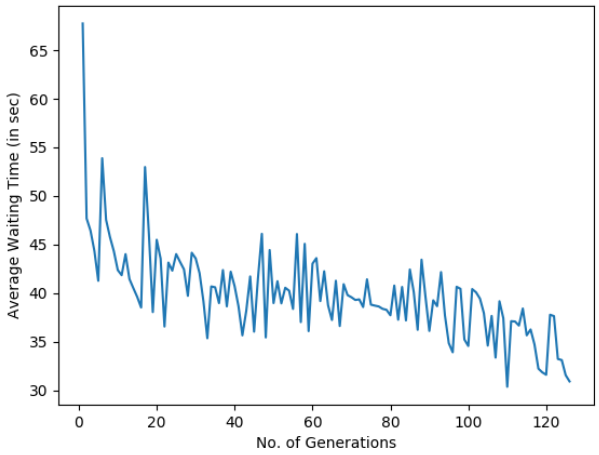


Fig 6. Average waiting time vs No. of generations (Generation best)

This Graph shows the average waiting time per vehicle of the best neural network in every single generation. Here total traffic on all roads is 1.0 vehicle/sec. After about 130 generations average waiting time reached about 30 sec/Vehicle.

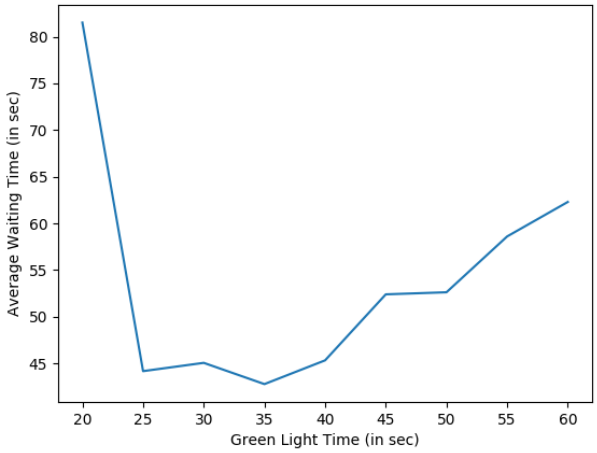


Fig 7. Average waiting time vs Green light time(sec)

This Graph shows the average waiting time per vehicle when a fixed amount of green light time is given in each road irrespective of traffic. Here total traffic on all four roads is 1.0 vehicle/sec. Among various possibilities least traffic is observed with green light time is 35 sec and the average waiting time per vehicle is around 42 sec.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:44 UTC from IEEE Xplore. Restrictions apply.

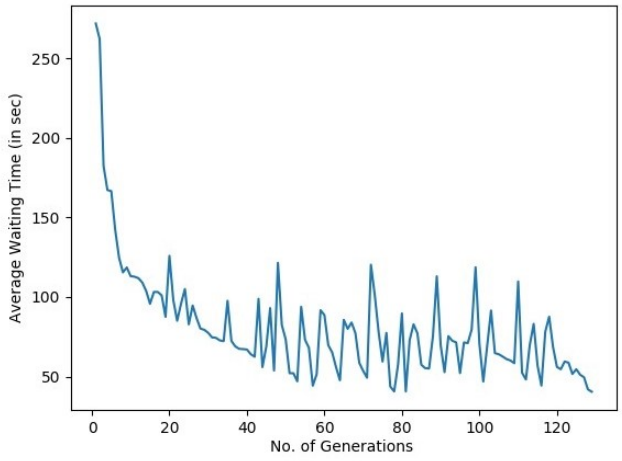


Fig 8. Average waiting time vs No. of generations(Whole population)

This Graph shows the average waiting time per vehicle of the whole population in every single generation. This result is after using self-adaptive mutation, with initial step size i.e. $\sigma = 0.5$, and change factor $c = 0.879$ and history window size = 12. Here total traffic on all four roads is 1.0 vehicle/sec. After about 130 generations average waiting time reached about 40 sec/Vehicle.

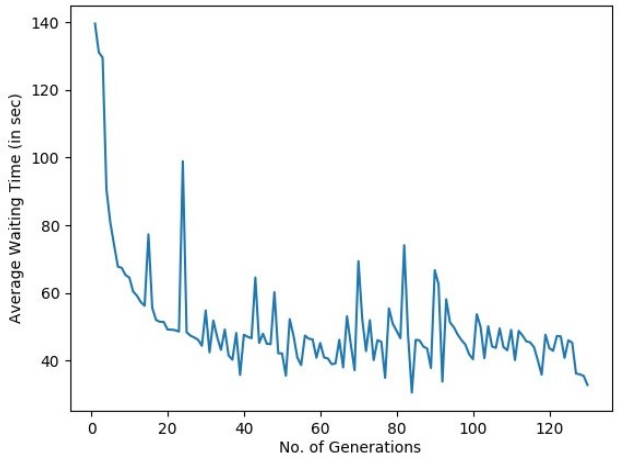


Fig 9. Average waiting time vs No. of generations(Generation best)

This Graph shows the average waiting time per vehicle of the best neural network in every single generation. This result is after using self-adaptive mutation, with initial step size i.e. $\sigma = 0.5$, and change factor $c = 0.879$ and history window size = 12. Here total traffic on all four roads is 1.0 vehicle/sec. After about 130 generations average waiting time reached to about 30 sec/Vehicle.

V. RESULT ANALYSIS

After training various models for the population size of 50 and around 100 generations models can be generated which are better than traditional models in which the same amount of green light time is given in each road irrespective of traffic and queue length. High fluctuations in average waiting time are observed between consecutive generations because of random traffic. Therefore to identify the performance of models in average cases simulate with every model 20-30 times and average its waiting time.

As traffic data is not available in many areas of the world due to high cost of installation and maintenance of sensors, many researchers avoided this field of research due to the lack of training and testing data for their models. Using SUMO allows us to simulate real-life traffic and when the genetic algorithm is applied, the result is practical, better than the naive approach of allocating equal signal times. This has been made possible due to our choice of the algorithm as Genetic algorithm which can fit solutions in less time, has a random mutation to guarantee a wide range of solutions, can be easily coded as compared to other algorithms that do the same job.

VI. CONCLUSION

After training for several generations that, our model reduce the average waiting time per vehicle to about two-third of waiting time observed for the static time allocation models traditionally used. As from figures 3, 6 & 9 which shows average waiting time per vehicle vs some generations of our best performing neural network in each generation. As generations, progress average waiting reduces to 28.64 sec/vehicle. Compare this with results of Figures 4 & 7 in which average waiting time/vehicle is calculated for various fixed green light times. In this case, the average waiting time is 42.56 sec/vehicle. Hence a reduction of 32.7% in waiting time is observed.

The queue lengths are taken as our inputs and assumed that taking snapshots of each road only at the time of green light switches is sufficient thereby reducing the network load. In the future, if the same model is developed for a network of roads, then the model is trained for efficiently handling occurrences of emergency vehicles and include the observation for temporal patterns (such as rush hours, congestion due to time-specific events such as office hours, etc).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:44 UTC from IEEE Xplore. Restrictions apply.

ACKNOWLEDGEMENT

We are immensely grateful to Delhi Technological University for providing every kind of support and opportunity, although any errors are our own and should not tarnish the reputation of this institution. The institution has provided us with all the means to nurture our talent and thus enabling us to work towards the completion of this research.

REFERENCES

1. S. Kumar Janahan, M. R.M. Veeramanickam, S. Arun, K. Narayanan, R. Anandan, S. Javed Parvez, "IoT based smart traffic signal monitoring system using vehicle counts", International Journal of Engineering and Technology, vol. 7, no. 221, pp. 309, 2018.
2. L. Singh, S. Tripathi, H. Arora, "Time optimization for traffic signal control using genetic algorithm", Int. J. Recent Trends Eng., vol. 2, no. 2, pp. 4-6, 2013.
3. Z. Li, M. Shahidehpour, S. Bahramirad, A. Khodaei, "Optimizing traffic signal settings in smart cities", IEEE Trans. Smart Grid, vol. 8, no. 5, pp. 2382-2393, 2017.
4. T. T. Thakur, A. Naik, S. Vatari, M. Gogate, "Real time traffic management using the Internet of Things", Proc. Int. Conf. Commun. Signal Process. (ICCSP), pp. 1950-1953, Apr. 2016.
5. P. Arunmozhi; P. Joseph William, "Automatic Ambulance Rescue System Using Shortest Path Finding Algorithm", International Journal of Science and Research (IJSR), vol. 3 Issue 5, pp. 635 - 638, May 2014.
6. M. S. Amin, J. Jalil, and M. B. I. Reaz, "Accident detection and reporting system using GPS, GPRS and GSM technology," in 2012 International Conference on Informatics, Electronics & Vision (ICIEV), Dhaka, Bangladesh, pp. 640-643, 2012.
7. A. E. Eiben, J. E. Smith. "Natural Computing Series: - Introduction to Evolutionary Computing". Springer Verlag Berlin Heidelberg, 2003, ISBN : 978-3-642-07285-7
8. Koresh, M. H. J. D. (2019). COMPUTER VISION BASED TRAFFIC SIGN SENSING FOR SMART TRANSPORT. Journal of Innovative Image Processing (JIIP), 1(01), 11-19
9. Kumar, R. Praveen, and S. Smys. "A novel report on architecture, protocols and applications in Internet of Things (IoT)." In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 1156-1161. IEEE, 2018.
10. L. Foschini, T. Taleb, A. Corradi, D. Bottazzi, "M2M-based metropolitan platform for IMS-enabled road traffic management in IoT", IEEE Commun. Mag., vol. 49, no. 11, pp. 50-57, Nov. 2011.
11. S. Misbahuddin, J. A. Zubairi, A. Saggaf, J. Basuni, S. A-Wadany, A. Al-Sofi, "IoT based dynamic road traffic management for smart cities", 2015 12th International Conference on High-Capacity Optical Networks and Enabling/Emerging Technologies (HONET), pp. 1-5, Dec. 2015.
12. Basim K. J. AI- Shammari, Nadia AI- Aboody, Hamed S. AI-Raweshidy, "Iot Traffic Management and Integration in the QoS Supported Network", IEEE Internet of Things Journal, vol. 5, no. 1, pp. 352-370, Feb. 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:44 UTC from IEEE Xplore. Restrictions apply.

Differential Equations and Dynamical Systems
https://doi.org/10.1007/s12591-020-00529-5

ORIGINAL RESEARCH

Dynamics of Predator–Prey System with Migrating Species and Disease in Prey Population

Charu Arora[1] · **Vivek Kumar**[2]

© Foundation for Scientific Research and Technological Innovation 2020

Abstract

In the proposed model, a prey-predator system with migration in both species along with the disease in prey population is considered. To study the dynamics among healthy predator, susceptible prey and infected prey, Holling type II functional response is incorporated. The model has been examined using local stability analysis around the existing equilibrium points and by considering migration in susceptible prey as the control parameter, bifurcation conditions are discussed in brief. Basic reproduction number is determined which affirms the disease to be endemic or not and global stability analysis has been implemented using geometric approach. At the end, numerical simulation to support theoretical results is exhibited.

Keywords Prey-predator system · Stability · Epidemics · Holling type II functional response

Introduction

The relationship between the predator and their prey has long been studied and is a very crucial component of study in ecology. The prey–predator interaction is prominent and significant area of research in applied mathematical modeling and population dynamics . The prey–predator relation is well studied in literature from many prospects. After the valuable contribution of Kermack and McKendrick in SIRS models, epidemiological models have drawn consideration of many biologists and ecologists [1–3]. Haque et al. investigated an eco epidemiological model where transmission of disease takes place in prey population only and stability analysis has been carried out along with the bifurcation study of the system [4]. The prey–predator system initiated by Cosner et al. [5] has been revised further from the biological aspect and the model is analyzed around the possible equilibria. Hopf bifurcation at the nonzero equilibrium point is carried out together with graphical

✉ Vivek Kumar
vivek_ag@iitkalumni.org

Charu Arora
charuarora099@gmail.com

[1] Department of Applied Sciences, Bharati Vidyapeeth's College of Engineering, New Delhi, India

[2] Department of Applied Mathematics, Delhi Technological University, Delhi, India

Published online: 02 June 2020

exhibition to confirm theoretical results in [6]. Much work has been done in literature on eco-epidemiological models with disease in prey only [7–11]. In [7], prey predator model is proposed with logistic growth rate of prey with the SIS parasite infection in the prey. The effect of environment toxic and disease on prey–predator system is studied in [7]. In [9], an eco-epidemiological model with disease in prey using Lotka Volterra prey-predator model is studied. It has been considered that both the susceptable and infected class of prey population are predated by predator population and the conditions for the feasibility of equilibrium for the system are attained. A prey–predator model with transmissible disease in the prey population is studied and system is analysed based on equilibrium, stability analysis (local and global) with bifurcation theory [10]. In [11], system is formulated with disease in the prey population and investigated from the view of nonnegativity, boundedness, equilibria, permanence, and global stability. At the end, numerical simulation supports the theoretical results.

Prey–predator models with the incorporation of delay have been recognised as a landmark in the study of prey–predator relationship. Efforts have made in implementing the deterministic and stochastic approach to examine a ratio dependent prey–predator model along with the inclusion of discrete time delay to interpret its outcomes in both analytical and numerical ways [12]. Much work has been done concerned with the delayed prey–predator system using stage structure in prey population where both mature and immature preys are predated [13]. Stage structure on both prey and predator population has been considered to study the dynamics of the system in [14].

In [15], Han et al. studied prey–predator relationship using the inclusion of SIS and SIR parasitic infection with the identification of thresholds. They determined the global stability results and it has been discussed that if disease prevails in prey population, then predator population also gets affected by the disease under certain preconditions. In [16, 17], a prey-predator model is proposed with infection in both populations to examine the feasibility of a transmittable disease spreading from prey to predator. Thereafter, the stability conditions are derived along with the calculation of basic reproduction number which depicts about the disease to be endemic or not in ecosystem.

Also, migration can be described as an important demographic process that occurs in all living beings [17] and the reason behind the migration varies in organisms such as climatic changes, to seek refuge, for food, shelter etc [18]. Therefore, migration factor can be considered as a significant aspect in the formation of prey predator mathematical models. This concept has been investigated from many ways in recent literature. In [19], migration and diffusion are considered to be important in the dynamics of spatial prey–predator interaction which is confirmed by both theoretical and numerical analysis.

Motivated by the study of Han and Zhien [15], Hsieh and Hsiao [16] and Shashi Kant et al. [17], we propose a prey-predator system with disease in prey population. The interactions between the healthy predator, susceptible prey and infected prey are studied using three-dimensional model along with the migration in both prey and predator population. The mortality rate of infected prey population is possible due to natural death and the death due to disease. In the paper, both the death factors are combined into the single exit rate d_1 along with the incorporation of migration factor. Extinction in susceptible prey population is demonstrated by the way of migration only and natural death rate has been ignored. The exit rate in infected prey population and predator population has been increased by considering both the migration term and death rate.

The present article is as follows: the next section deals with the formation of prey–predator model with disease in prey population. Section 3 presents the basic mathematical preliminaries of the model. Section 4 introduces the model in two subcases

as without disease and migration and secondly without disease only along with their conditions of stability and bifurcation results. Section 5, provides the dynamics of the main model and last section comprises of the discussion following the global stability analysis in Sect. 6 and numerical simulations in Sect. 7.

Model Formulation

The present model consists of two populations i.e., prey with total density $S(t) + I(t) = N(t)$ and predator with population $Y(t)$. The following assumptions are made for the mathematical model as:

(H1) Prey population grows logistically with intrinsic growth rate $r(r > 0)$ and carrying capacity $k(k > 0)$ in the absence of predators and disease.

(H2) It is assumed that disease occurs in prey population and in the presence of disease, prey density is divided into two classes i.e., susceptible $S(t)$ and infective $I(t)$. No disease in predator population is considered with its total density $Y(t)$. Hence, the model involves three populations as susceptible prey, infected prey and healthy predators.

(H3) The initial value of the prey density is $I(0) \geq 0$ due to the consideration of the fact that infective prey previously exist in the population and it has been presumed that no infective predators are present in the system. Moreover, this fact has been taken into consideration that predators do not get any disease/infection in predacity.

(H4) In [10, 20, 21], law of mass action for transmitting the disease has been assumed which states that the rate by which infection spreads is directly proportional to the product of the infectious persons I and susceptible persons S. Hence, it is inferred that spreading of disease takes place as per the law βSI where, β is the rate by which susceptible gets converted into infectives.

(H5) Migration among preys and predators has been incorporated in the model. m_1, m_2 and m_3 denote the migration coefficients in susceptible prey, infected prey and predators respectively.

(H6) Death rate of infected prey population will certainly be more as compared to susceptible prey population because the occurrence of disease weakens the infected prey class which makes it easy for predators to invade them as compared to susceptible population. Besides, susceptible prey class being more active is less likely to be caught by predators. Moreover, illness causes the infected prey population to die out soon. This leads to the higher and earlier death rate of infected prey class. Hence, the death rate of infected prey population is considered by d_1. The natural death rate of susceptible prey population would be low, therefore, it has been ignored in the proposed system and migration is considered to be a way by which susceptible prey population moves out of the system.

(H7) It has been assumed that infected prey recover with rate α.

(H8) Holling type II functional response has been implemented to depict the interaction among prey and predators Table 1.

Based on the assumptions as discussed above, the mathematical model takes the following form as:

Table 1 Details of the parameters in the system 2.1

Parameters and their definition			
Symbol	Definition	Symbol	Definition
r	Intrinsic growth rate	c_2	Conversion rate of I into Y
k	Carrying capacity	a	Saturation rate of predators
β	Rate of conversion from susceptible prey population to infective prey population	d_1	Death rate of infected preys
α	Recovery rate	d_2	Death rate of predators
p_1	Predation rate of susceptible prey	m_1	Migration coefficient in susceptible prey
p_2	Predation rate of infective prey	m_2	Migration coefficient in infected prey
c_1	Conversion rate of S into Y	m_3	Migration coefficient in predators

$$\begin{aligned}
\frac{dS(t)}{dt} &= rS\left(1-\frac{S+I}{k}\right)-\beta SI+\alpha I-\frac{p_1SY}{1+aS}-m_1S,\\
\frac{dI(t)}{dt} &= \beta SI-\frac{p_2IY}{1+aI}-\alpha I-m_2I-d_1I,\\
\frac{dY(t)}{dt} &= \frac{c_1p_1SY}{1+aS}+\frac{c_2p_2IY}{1+aI}-d_2Y-m_3Y,
\end{aligned} \tag{2.1}$$

with initial conditions:

$$S(0)>0, I(0)\geq 0, Y(0)>0. \tag{2.2}$$

Mathematical Preliminaries of the Model

In this section, positivity and boundedness for the system (2.1) are discussed as the variables S, I and Y represents the living species. Positivity guarantees that population cannot take negative values and population will survive always. Boundedness confirms the natural restrictions to the growth of population due to limited resources.

Positivity

Lemma 1 *Every solution of* (2.1) *with initial conditions* (2.2) *which exists in* $[0,+\infty[$, *remain positive for all* $t>0$.

Proof As discussed in [22, 23], the same procedure follows. The proposed model (2.1) with the imposed initial conditions (2.2) can be written in the matrix equation form,

$$\frac{dH}{dt}=G(H(t)),$$

where, $H(t)=(S,I,Y)^T, H(0)=(S(0),I(0),Y(0))^T\in R_+^3$ and

$$G(H(t)) = \begin{pmatrix} G_1(H(t)) \\ G_2(H(t)) \\ G_3(H(t)) \end{pmatrix} = \begin{pmatrix} rS\left(1 - \frac{S+I}{k}\right) - \beta SI + \alpha I - \frac{p_1 SY}{1+aS} - m_1 S \\ \beta SI - \frac{p_2 IY}{1+aI} - \alpha I - m_2 I - d_1 I \\ \frac{c_1 p_1 SY}{1+aS} + \frac{c_2 p_2 IY}{1+aI} - d_2 Y - m_3 Y \end{pmatrix},$$

where $G : R^3 \to R^3_+$ and $G \in C^\infty(R^3)$. It is clear that in system 2.1, $G_i(H_i) \mid_{H_i=0} \geq 0$, for $i = 1, 2, 3$. By [24], the solution of matrix equation with initial condition $H_0 \in R^3_+$, for instance $H(t) = H(t, H_0)$, is such that $G(t) \in R^3_+$ for all finite and positive time t. □

Boundedness

Lemma 2 *All the solutions of the system* (2.1) *with initial conditions* (2.2) *in* R^3_+ *are uniformly bounded.*

Proof **Case I.** $c_1 = c_2 = 1$.

Let $W = S + I + Y$, system (2.1) gives as,

$$\frac{dW}{dt} = rS\left(1 - \frac{S+I}{k}\right) - m_1 S - m_2 I - d_1 I - d_2 Y - m_3 Y.$$

Assuming a positive constant $\eta > 0$, it follows,

$$\frac{dW}{dt} + \eta W = S\left(r\left(1 - \frac{S+I}{k}\right) - m_1 + \eta\right) - (m_2 + d_1 - \eta)I - (m_3 + d_2 - \eta)Y.$$

On taking $\eta = \min\{m_2 + d_1, m_3 + d_2\}$, we get,

$$\frac{dW}{dt} + \eta W \leq S\left[r\left(1 - \frac{S+I}{k}\right) - m_1 + \eta\right].$$

The supremum of $S\left[r\left(1 - \frac{S+I}{k}\right) - m_1 + \eta\right]$ is $\frac{k(r+\eta)^2}{4r}$ as in [22]. So, we get the following inequality,

$$\frac{dW}{dt} + \eta W \leq \frac{k(r+\eta)^2}{4r} = \Omega > 0(say).$$

Now, by the application of differential inequality [25], we get,

$$0 \leq W \leq \frac{\Omega(1 - e^{-\eta t})}{\eta} + W(S(0), I(0), Y(0))e^{-\eta t}.$$

For $t \to \infty$, $0 < W < \frac{\Omega}{\eta} = \tilde{M}$.

Hence, the solutions of system (2.1) in R^3_+ is bounded in the region $U = \{(S, I, Y) \in R^3_+ : 0 < W < \frac{\Omega}{\eta} + \Phi, \Phi > 0, \Omega = \frac{k(r+\eta)^2}{4r}\}$.

Case II. $c_1 = c_2 \neq 1$.

In this case, we assume $W = c_1(S + I) + Y$ and the lemma can be proved similarly.

Case III. $c_1 \neq c_2$.

In this case, we assume $W = c_1S + c_2I + Y$ and the lemma can be proved likewise.

□

Dynamics of Model Without Disease

Model Without Migration and Disease

In this section, two assumptions are imposed on model (2.1) as:

1. No migration permitted.
2. No infection occurred.

So, the model (2.1) takes the form as:

$$\begin{aligned} \frac{dS(t)}{dt} &= rS\left(1-\frac{S}{k}\right) - \frac{p_1SY}{1+aS}, \\ \frac{dY(t)}{dt} &= \frac{c_1p_1SY}{1+aS} - d_2Y, \end{aligned} \tag{4.1}$$

with initial conditions:

$$S(0) > 0, Y(0) > 0. \tag{4.2}$$

System (4.1) possesses the following equilibrium points and the stability conditions are stated below:

1. The trivial equilibrium point $E_1(0,0)$ exist always and is an unstable saddle point.
2. The axial equilibrium point $E_2(k,0)$ exist always with local stability condition as $c_1p_1k < d_2(1+ak)$.
3. The interior equilibrium point $E_3(\tilde{S},\tilde{Y})$ where $\tilde{S} = \frac{d_2}{a+c_1p_1}$ and $\tilde{Y} = 1 + \frac{ad_2}{a+c_1p_1} - \frac{d_2}{(a+c_1p_1)k} - \frac{ad_2^2}{(a+c_1p_1)^2k}$ exists provided, $a + c_1p_1 + ad_2 > \frac{d_2}{k} + \frac{ad_2^2}{k(a+c_1p_1)}$ with local stability conditions as defined below:
   1. $r - \frac{2r\tilde{S}}{k} - \frac{p_1\tilde{Y}}{(1+a\tilde{S})^2} + \frac{c_1p_1\tilde{S}}{(1+a\tilde{S})} - d_2 < 0.$
   2. $\frac{c_1p_1\tilde{S}r}{1+a\tilde{S}} - d_2r - \frac{2r\tilde{S}^2c_1p_1}{(1+a\tilde{S})k} + \frac{2r\tilde{S}d_2}{k} + \frac{p_1\tilde{Y}d_2}{(1+a\tilde{S})^2} > 0.$

Model Without Disease

In this section, assumptions are imposed in model (2.1) as:

1. Migration permitted.
2. No infection occurred.

So, the model takes the form as:

$$\begin{cases} \frac{dS(t)}{dt} = rS\left(1-\frac{S}{k}\right) - \frac{p_1SY}{1+aS} - m_1S, \\ \frac{dY(t)}{dt} = \frac{c_1p_1SY}{1+aS} - d_2Y - m_3Y, \end{cases} \tag{4.3}$$

with initial conditions:

$$S(0) > 0, Y(0) > 0. \tag{4.4}$$

System (4.3) possesses the following equilibrium points and the local stability conditions along with bifurcation results as discussed in [4] are stated below:

1. The trivial equilibrium point $E_1^*(0,0)$ exists and the local stability condition becomes $r < m_1$.
2. The axial equilibrium point $E_2^*(S^*,0)$ where, $S^* = \frac{k(r-m_1)}{r}$ exists, provided, $r > m_1$ with local stability conditions as, $r < \frac{2rS^*}{k} + m_1$ and $\frac{c_1p_1S^*}{1+aS^*} < d_2 + m_3$.
3. The interior equilibrium point $E_3^*(\tilde{S},\tilde{Y})$ where, $\tilde{S} = \frac{d_2+m_3}{c_1p_1-d_2a-m_3a}$ and $\tilde{Y} = \frac{\left(r\left(1-\frac{\tilde{S}}{k}\right)-m_1\right)(1+a\tilde{S})}{p_1}$ exists, provided, $c_1p_1 > d_2a + m_3a$ and $r > \frac{r\tilde{S}}{k} + m_1$ and the variational matrix about the equilibrium point $E_3^*(\tilde{S},\tilde{Y})$ is given by,

$$J_{E_3^*(\widetilde{S},\widetilde{Y})} = \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix},$$

where,

$$\begin{aligned} v_{11} &= r - \frac{2r\widetilde{S}}{k} - \frac{p_1\widetilde{Y}}{(1+a\widetilde{S})^2} - m_1, \\ v_{12} &= \frac{-p_1\widetilde{S}}{(1+a\widetilde{S})}, \\ v_{21} &= \frac{c_1p_1\widetilde{Y}}{(1+a\widetilde{S})^2}, \\ v_{22} &= \frac{c_1p_1\widetilde{S}}{1+a\widetilde{S}} - d_2 - m_3. \end{aligned}$$

The characteristic equation based on the above matrix is given by,

$$\begin{aligned} v(\lambda;m_1) &= \lambda^2 + \lambda v_1(m_1) + v_0(m_1) = 0, \text{ where,} \\ v_1(m_1) &= -(v_{11} + v_{22}), v_0(m_1) = v_{11}v_{22} - v_{12}v_{21}. \end{aligned}$$

$E_3^*(\widetilde{S},\widetilde{Y})$ is locally asymptotically stable if,

1. $r + \frac{c_1p_1\tilde{S}}{1+a\tilde{S}} < \frac{2r\tilde{S}}{k} + \frac{p_1\tilde{Y}}{(1+a\tilde{S})^2} + m_1 + m_3 + d_2.$
2. $(r - \frac{2r\tilde{S}}{k} - \frac{p_1\tilde{Y}}{(1+a\tilde{S})^2} - m_1)(\frac{c_1p_1\tilde{S}}{1+a\tilde{S}} - d_2 - m_3) + \frac{c_1p_1^2\tilde{S}\tilde{Y}}{(1+a\tilde{S})^3} > 0.$

Theorem 1 *Let m_1, the migration parameter in susceptible prey population be the bifurcation parameter. Hopf bifurcation occurs at $m_1 = m_{10}$, if the following conditions are satisfied,*

1. $v_0(m_{10}) > 0,$
2. $v_1(m_{10}) = 0,$

3. $\frac{dv_1(m_{10})}{dm_1} \neq 0$ [26].

Proof By 1, $v_0(m_{10}) > 0$ (proved above).

By 2, $v_1(m_{10}) = 0$ gives,

$$m_{10} = r - \frac{2r\widetilde{S}}{k} - \frac{p_1\widetilde{Y}}{(1+a\widetilde{S})^2} - m_3 - d_2 + \frac{c_1p_1\widetilde{S}}{1+a\widetilde{S}}.$$

By 3, $\frac{dv_1}{dm_1} = 1 - \frac{1}{1+a\widetilde{S}} \neq 0$ for $m_1 = m_{10}$.

Theorem 2 *The equilibrium point* $E_1^*(0,0)$ *shows transcritical bifurcation* [26].

Proof When the migration parameter $m_1 > r$, the equilibrium point $E_1^*(0,0)$ is stable but when $m_1 < r$, $E_1^*(0,0)$ is unstable. In addition, $m_1 > r$ violates the stability condition of $E_2^*(S^*,0)$ which attains stability along with the condition $\frac{c_1p_1S^*}{1+aS^*} < d_2 + m_2$. Infact, $E_1^*(0,0)$ is stable when $E_2^*(S^*,0)$ and $E_3^*(\widetilde{S},\widetilde{Y})$ are unfeasible. In such situation, transcritical bifurcation occurs where there is an exchange in stability of equilibrium points. □

Dynamics of Model (2.1)

In this section, number of coupled conditions for local stability analysis of the proposed model are determined. Also, basic reproduction number R_0 to examine whether the disease will be endemic or not is calculated as in [27].

The model takes the following form:

$$\begin{cases} \frac{dS(t)}{dt} = rS\left(1 - \frac{S+I}{k}\right) - \beta SI + \alpha I - \frac{p_1SY}{1+aS} - m_1S, \\ \frac{dI(t)}{dt} = \beta SI - \frac{p_2IY}{1+aI} - \alpha I - m_2I - d_1I, \\ \frac{dY(t)}{dt} = \frac{c_1p_1SY}{1+aS} + \frac{c_2p_2IY}{1+aI} - d_2Y - m_3Y. \end{cases} \tag{5.1}$$

Based on above model, gain and loss matrices at the equilibrium point $E^{(6)}(S^{(6)},0,Y^{(6)})$ where, $S^{(6)} = \frac{d_2+m_3}{c_1p_1-d_2a-m_3a}$ and $Y^{(6)} = \frac{rS^{(6)}\left(1-\frac{S^{(6)}}{k}\right)-m_1S^{(6)}}{\frac{p_1S^{(6)}}{1+aS^{(6)}}}$ which is free from infection is considered and basic reproduction number takes the form as, $R_0 = \frac{\beta S^{(6)}}{\alpha+m_2+d_2}$ and disease is endemic if $R_0 > 1$.

Equilibria

The system (2.1) may have the equilibrium points as discussed below:

(i) $E^{(1)}(0,0,0)$.
(ii) $E^{(2)}(S^{(2)},0,0)$, where, $S^{(2)} = \frac{k(r-m_1)}{r}$.
(iii) $E^{(3)}(0,I^{(3)},0)$, $E^{(4)}(0,0,Y^{(4)})$.
(iv) $E^{(5)}(S^{(5)},I^{(5)},0)$, where, $S^{(5)} = \frac{\alpha+m_2+d_1}{\beta}$ and $I^{(5)} = \frac{rS^{(5)}-\frac{rS^{(5)^2}}{k}-m_1S^{(5)}}{\frac{rS^{(5)}}{k}+\beta S^{(5)}-\alpha}$.

(v) $E^{(6)}(S^{(6)}, 0, Y^{(6)})$, where, $S^{(6)} = \frac{d_2+m_3}{c_1p_1-d_2a-m_3a}$ and $Y^{(6)} = \frac{rS^{(6)}\left(1-\frac{S^{(6)}}{k}\right)-m_1S^{(6)}}{\frac{p_1S^{(6)}}{1+aS^{(6)}}}$.

(vi) $E^{(7)}(0, I^{(7)}, Y^{(7)})$.

(vii) $E^{(8)}(S^{(8)}, I^{(8)}, Y^{(8)})$, where, $S = S^{(8)}$ is a unique positive root of the equation given by,

$$-AFS^3 + (AD' - BF - CG)S^2 + (BD' - CE + GQ + AH)S + QE + BH = 0,$$

subject to the following conditions;

$$-18AF(AD' - BF - CG)(BD' - CE + GQ + AH)(QE + BH)$$
$$-4(AD' - BF - CG)^3(QE + BH) + (AD' - BF - CG)^2(BD' - CE + GQ + AH)^2$$
$$+4AF(BD' - CE + GQ + AH)^3 - 27A^2F^2(QE + BH)^2 < 0$$

and $(QE + BH) < 0$. where,

$$A = Ca + aB' - Qa^2, B = B' - 2Qa, C = c_1p_1, B' = c_2p_2, D' = \frac{-r'}{k} + r'a - am_1 - \beta'',$$
$$E = -\beta' - \frac{r'}{k} + a\alpha' + Ra, F = \frac{r'a}{k}, G = \frac{-r'a}{k} - a\beta' - a\beta'',$$
$$H = r' - m_1 + R, R = \alpha'' + m_2' + d_1', Q = d_2 + m_3,$$
$$d_1' = \frac{d_1}{p_2}, r' = \frac{r}{p_1}, \beta' = \frac{\beta}{p_1},$$
$$\alpha' = \frac{\alpha}{p_1}, m_1' = \frac{m_1}{p_1}, \beta'' = \frac{\beta}{p_2}, \alpha'' = \frac{\alpha}{p_2}.$$

$I^{(8)}$ is given by,

$$I^{(8)} = \frac{Q - CS^{(8)}}{B + S^{(8)}A},$$

and $Y^{(8)} = \frac{(\beta S^{(8)}-\alpha-m_2-d_1)(1+aI^{(8)})}{p_2}$.

Existence Conditions

The equilibrium points:

(i) $E^{(1)}(0, 0, 0)$ always exist.

(ii) $E^{(2)}(S^{(2)}, 0, 0)$ exists, if $r > m_1$.

(iii) $E^{(3)}(0, I^{(3)}, 0)$, $E^{(4)}(0, 0, Y^{(4)})$ do not exist.

(iv) $E^{(5)}(S^{(5)}, I^{(5)}, 0)$ exists, provided, either,

$$r > \frac{rL}{k} + m_1 \text{ and } \frac{rL}{k} + \beta L > \alpha$$

or

$$r < \frac{rL}{k} + m_1 \text{ and } \frac{rL}{k} + \beta L < \alpha \text{ where, } L = \frac{\alpha + m_2 + d_1}{\beta}.$$

(v) $E^{(6)}(S^{(6)}, 0, Y^{(6)})$ will exist if, $c_1p_1 > d_2a + m_3a$ and $r\left(1 - \frac{T}{k}\right) - m_1 > 0$, where, $T = \frac{d_2+m_3}{c_1p_1-d_2a-m_3a}$.

(vi) $E^{(7)}(0, I^{(7)}, Y^{(7)})$ does not exist.

(vii) $E^{(8)}(S^{(8)}, I^{(8)}, Y^{(8)})$ exists, provided, $\beta S^{(8)} > \alpha + m_2 + d_1$ and $Q > CS^{(8)}$ where, Q, C are defined above and $S^{(8)}$ is a unique positive root of the cubic equation,

$$-AFS^3 + (AD' - BF - CG)S^2 + (BD' - CE + GQ + AH)S + QE + BH = 0$$

subject to conditions,

$$-18AF(AD' - BF - CG)(BD' - CE + GQ + AH)(QE + BH)$$
$$-4(AD' - BF - CG)^3(QE + BH) + (AD' - BF - CG)^2(BD' - CE + GQ + AH)^2$$
$$+4AF(BD' - CE + GQ + AH)^3 - 27A^2F^2(QE + BH)^2 < 0 \text{ and } (QE + BH) < 0.$$

Stability Analysis

The variational matrix of the system (2.1) is given by;

$$J = \begin{pmatrix} r - \frac{2rs+rI}{k} - \beta I - \frac{p_1Y}{(1+aS)^2} - m_1 & \frac{-rS}{k} - \beta S + \alpha & \frac{-p_1S}{(1+aS)} \\ \beta I & \beta S - \frac{p_2Y}{(1+aI)^2} - \alpha - m_2 - d_1 & \frac{-p_2I}{(1+aI)} \\ \frac{c_1p_1Y}{(1+aS)^2} & \frac{c_2p_2Y}{(1+aI)^2} & \frac{c_1p_1S}{(1+aS)} + \frac{c_2p_2I}{(1+aI)} - d_2 - m_3 \end{pmatrix}.$$

Trivial Equilibrium ($E^{(1)}(0, 0, 0)$)

The variational matrix of the system (2.1) at $E^{(1)}(0, 0, 0)$ is given by;

$$J_{E^{(1)}(0,0,0)} = \begin{pmatrix} r - m_1 & \alpha & 0 \\ 0 & -\alpha - m_2 - d_1 & 0 \\ 0 & 0 & -d_2 - m_3 \end{pmatrix}.$$

The eigenvalues of the above upper triangular matrix are;

$$\begin{cases} r - m_1, \\ -\alpha - m_2 - d_1, \\ -d_2 - m_3. \end{cases}$$

So, $E^{(1)}(0, 0, 0)$ is a stable equilibrium if $r < m_1$ in order to have all negative eigenvalues.

Equilibrium ($E^{(2)}(S^{(2)}, 0, 0)$)

The variational matrix of the system (2.1) at $E^{(2)}(S^{(2)}, 0, 0)$ is given by;

$$J_{E^{(2)}(S^{(2)},0,0)} = \begin{pmatrix} r - \frac{2rS^{(2)}}{k} - m_1 & \frac{-rS^{(2)}}{k} - \beta S^{(2)} + \alpha & \frac{-p_1 S^{(2)}}{(1+aS^{(2)})} \\ 0 & \beta S^{(2)} - \alpha - m_2 - d_1 & 0 \\ 0 & 0 & \frac{c_1 p_1 S^{(2)}}{(1+aS^{(2)})} - d_2 - m_3 \end{pmatrix}.$$

The above matrix is an upper triangular matrix. So eigenvalues are given by;

$$\begin{cases} r - \frac{2rS^{(2)}}{k} - m_1, \\ \beta S^{(2)} - \alpha - m_2 - d_1, \\ \frac{c_1 p_1 S^{(2)}}{1+aS^{(2)}} - d_2 - m_3. \end{cases}$$

So, $E^{(2)}(S^{(2)}, 0, 0)$ is a stable equilibrium if the following conditions are satisfied;

$$\begin{cases} r < \frac{2rS^{(2)}}{k} + m_1, \\ \beta S^{(2)} < \alpha + m_2 + d_1, \\ \frac{c_1 p_1 S^{(2)}}{1+aS^{(2)}} < d_2 + m_3. \end{cases}$$

Based on these conditions, eigenvalues will have negative values.

Equilibrium ($E^{(5)}(S^{(5)}, I^{(5)}, 0)$)

The variational matrix of the system (2.1) at $E^{(5)}(S^{(5)}, I^{(5)}, 0)$ is given by;

$$J_{E^{(5)}(S^{(5},I^{(5)},0)} = \begin{pmatrix} r - \frac{2rS^{(5)}+rI^{(5)}}{k} - \beta I^{(5)} - m_1 & \frac{-rS^{(5)}}{k} - \beta S^{(5)} + \alpha & \frac{-p_1 S^{(5)}}{(1+aS^{(5)})} \\ \beta I^{(5)} & \beta S^{(5)} - \alpha - m_2 - d_1 & \frac{-p_2 I^{(5)}}{1+aI^{(5)}} \\ 0 & 0 & \frac{c_1 p_1 S^{(5)}}{(1+aS^{(5)})} + \frac{c_2 p_2 I^{(5)}}{(1+aI^{(5)})} - d_2 - m_3 \end{pmatrix}.$$

The corresponding characteristic equation is given by;

$$\begin{aligned} &\lambda^3 - \lambda^2\left(r - \frac{2rS^{(5)} + rI^{(5)}}{k} - \beta I^{(5)} - m_1 + \beta S^{(5)} - \alpha - m_2 - d_1\right. \\ &\left. + \frac{c_1 p_1 S^{(5)}}{(1 + aS^{(5)})} + \frac{c_2 p_2 I^{(5)}}{(1 + aI^{(5)})} - d_2 - m_3\right) + \lambda\left(\left(r - \frac{2rS^{(5)} + rI^{(5)}}{k} - \beta I^{(5)} - m_1\right)\right. \\ &\left.\left(\beta S^{(5)} - \alpha - m_2 - d_1\right) + \left(r - \frac{2rS^{(5)} + rI^{(5)}}{k} - \beta I^{(5)} - m_1\right)\left(\frac{c_1 p_1 S^{(5)}}{(1 + aS^{(5)})}\right.\right. \\ &\left.\left. + \frac{c_2 p_2 I^{(5)}}{(1 + aI^{(5)})} - d_2 - m_3\right) + (\beta S^{(5)} - \alpha - m_2 - d_1)\left(\frac{c_1 p_1 S^{(5)}}{(1 + aS^{(5)})}\right.\right. \\ &\left.\left. + \frac{c_2 p_2 I^{(5)}}{(1 + aI^{(5)})} - d_2 - m_3\right) - \left(\frac{-rS^{(5)}}{k} - \beta S^{(5)} + \alpha\right)(\beta I^{(5)})\right) \\ &+ \left(\frac{-rS^{(5)}}{k} - \beta S^{(5)} + \alpha\right)(\beta I^{(5)})\left(\frac{c_1 p_1 S^{(5)}}{(1 + aS^{(5)})} + \frac{c_2 p_2 I^{(5)}}{(1 + aI^{(5)})} - d_2 - m_3\right) \\ &- \left(r - \frac{2rS^{(5)} + rI^{(5)}}{k} - \beta I^{(5)} - m_1\right)(\beta S^{(5)} - \alpha - m_2 - d_1)\left(\frac{c_1 p_1 S^{(5)}}{(1 + aS^{(5)})}\right. \\ &\left. + \frac{c_2 p_2 I^{(5)}}{(1 + aI^{(5)})} - d_2 - m_3\right) = 0. \end{aligned}$$

Based on the above characteristic equation, eigenvalues are given by;

$$\begin{cases} r - \frac{2rS^{(5)}+rI^{(5)}}{k} - \beta I^{(5)} - m_1 + \beta S^{(5)} - \alpha - m_2 - d_1, \\ \frac{c_1 p_1 S^{(5)}}{(1+aS^{(5)})} + \frac{c_2 p_2 I^{(5)}}{(1+aI^{(5)})} - d_2 - m_3, \\ (\frac{-rS^{(5)}}{k} - \beta S^{(5)} + \alpha)(\beta I^{(5)}) - (r - \frac{2rS^{(5)}+rI^{(5)}}{k} - \beta I^{(5)} - m_1)(\beta S^{(5)} - \alpha - m_2 - d_1). \end{cases}$$

Hence, the equilibrium point $E^{(5)}(S^{(5)}, I^{(5)}, 0)$ is a stable equilibrium subject to the following conditions as;

$$\begin{cases} (r - \frac{2rS^{(5)}+rI^{(5)}}{k} - \beta I^{(5)} - m_1)(\beta S^{(5)} - \alpha - m_2 - d_1) > (\frac{-rS^{(5)}}{k} - \beta S^{(5)} + \alpha)(\beta I^{(5)}), \\ r - \frac{2rS^{(5)}+rI^{(5)}}{k} - \beta I^{(5)} - m_1 + \beta S^{(5)} - \alpha - m_2 - d_1 < 0, \\ \frac{c_1 p_1 S^{(5)}}{(1+aS^{(5)})} + \frac{c_2 p_2 I^{(5)}}{(1+aI^{(5)})} < d_2 + m_3. \end{cases}$$

Equilibrium ($E^{(6)}(S^{(6)}, 0, Y^{(6)})$)

The variational matrix of the system (2.1) at $E^{(6)}(S^{(6)}, 0, Y^{(6)})$ is given by,

$$J_{E^{(6)}(S^{(6)},0,Y^{(6)})} = \begin{pmatrix} r - \frac{2rS^{(6)}}{k} - \frac{-p_1 Y^{(6)}}{(1+aS^{(6)})^2} - m_1 & \frac{-rS^{(6)}}{k} - \beta S^{(6)} + \alpha & \frac{-p_1 S^{(6)}}{(1+aS^{(6)})} \\ 0 & \beta S^{(6)} - p_2 Y^{(6)} - \alpha - m_2 - d_1 & 0 \\ \frac{c_1 p_1 Y^{(6)}}{(1+aS^{(6)})^2} & c_2 p_2 Y^{(6)} & \frac{c_1 p_1 S^{(6)}}{(1+aS^{(6)})} - d_2 - m_3 \end{pmatrix}.$$

The characteristic equation is given by;

$$\begin{aligned} &\lambda^3 - \lambda^2(r - \frac{2rS^{(6)}}{k} - \frac{-p_1 Y^{(6)}}{(1+aS^{(6)})^2} - m_1 + \frac{c_1 p_1 S^{(6)}}{(1+aS^{(6)})} - d_2 \\ &- m_3 + \beta S^{(6)} - p_2 Y^{(6)} - \alpha - m_2 - d_1) + \lambda((r - \frac{2rS^{(6)}}{k} - \frac{-p_1 Y^{(6)}}{(1+aS^{(6)})^2} \\ &- m_1)(\beta S^{(6)} - p_2 Y^{(6)} - \alpha - m_2 - d_1) + (r - \frac{2rS^{(6)}}{k} - \frac{-p_1 Y^{(6)}}{(1+aS^{(6)})^2} \\ &- m_1)(\frac{c_1 p_1 S^{(6)}}{(1+aS^{(6)})} - d_2 - m_3) + (\beta S^{(6)} - p_2 Y^{(6)} - \alpha - m_2 - d_1)(\frac{c_1 p_1 S^{(6)}}{(1+aS^{(6)})} \\ &- d_2 - m_3) - (\frac{c_1 p_1 Y^{(6)}}{(1+aS^{(6)})^2})(\frac{-p_1 S^{(6)}}{(1+aS^{(6)})})) \\ &+ (\frac{c_1 p_1 Y^{(6)}}{(1+aS^{(6)})^2})(\frac{-p_1 S^{(6)}}{(1+aS^{(6)})})(\beta S^{(6)} - p_2 Y^{(6)} - \alpha - m_2 - d_1) - (r - \frac{2rS^{(6)}}{k} \\ &- \frac{-p_1 Y^{(6)}}{(1+aS^{(6)})^2} - m_1)(\beta S^{(6)} - p_2 Y^{(6)} - \alpha - m_2 - d_1)(\frac{c_1 p_1 S^{(6)}}{(1+aS^{(6)})} - d_2 - m_3). \end{aligned}$$

The equilibrium point $E^{(6)}(S^{(6)}, 0, Y^{(6)})$ is a stable equilibrium if the following conditions are satisfied:

$$\begin{cases} (r - \frac{2rS^{(6)}}{k} - m_3 - d_2 - m_1 + \frac{c_1p_1S^{(6)}}{(1+aS^{(6)})} - \frac{p_1Y^{(6)}}{(1+aS^{(6)})^2} < 0, \\ (r - \frac{2rS^{(6)}}{k} - \frac{p_1Y^{(6)}}{(1+aS^{(6)})^2} - m_1)(\frac{c_1p_1S^{(6)}}{1+aS^{(6)}} - d_2 - m_3) > \frac{-c_1(p_1)^2Y^{(6)}}{(1+aS^{(6)})^3}, \\ \beta S^{(6)} < p_2Y^{(6)} + \alpha + m_2 + d_1. \end{cases}$$

Equilibrium ($E^{(8)}(S^{(8)}, I^{(8)}, Y^{(8)})$)

The variational matrix of the system (2.1) at $E^{(8)}(S^{(8)}, I^{(8)}, Y^{(8)})$ is given by,

$$J_{E^{(8)}(S^{(8)},I^{(8)},Y^{(8)})} = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix},$$

where,

$$\begin{aligned} f_{11} &= r - \frac{2rS^{(8)} + rI^{(8)}}{k} - \beta I^{(8)} - \frac{p_1Y^{(8)}}{(1+aS^{(8)})^2} - m_1, \\ f_{12} &= \frac{-rS^{(8)}}{k} - \beta S^{(8)} + \alpha, \\ f_{13} &= \frac{-p_1S^{(8)}}{(1+aS^{(8)})}, \\ f_{21} &= \beta I^{(8)}, \\ f_{22} &= \beta S^{(8)} - \frac{p_2Y^{(8)}}{(1+aI^{(8)})^2} - \alpha - m_2 - d_1, \\ f_{23} &= \frac{-p_2I^{(8)}}{(1+aI^{(8)})}, \\ f_{31} &= \frac{c_1p_1Y^{(8)}}{(1+aS^{(8)})^2}, \\ f_{32} &= \frac{c_2p_2Y^{(8)}}{(1+aI^{(8)})^2}, \\ f_{33} &= \frac{c_1p_1S^{(8)}}{(1+aS^{(8)})} + \frac{c_2p_2I^{(8)}}{(1+aI^{(8)})} - d_2 - m_3. \end{aligned}$$

The characteristic equation of the above matrix is given by,

$$t(\lambda;m_1) = \lambda^3 + t_2(m_1)\lambda^2 + t_1(m_1)\lambda + t_0(m_1) = 0,$$

where,

$$\begin{aligned} t_2 &= (-f_{11} - f_{22} - f_{33}), t_1 = (f_{11}f_{22} + f_{11}f_{33} + f_{33}f_{22} - f_{32}f_{23} - f_{12}f_{22} - f_{31}f_{13}), \\ t_0 &= f_{11}f_{23}f_{32} - f_{11}f_{22}f_{33} + f_{12}f_{22}f_{33} + f_{31}f_{13}f_{22} - f_{12}f_{23}f_{31} - f_{13}f_{21}f_{32}. \end{aligned}$$

Using Routh Hurwitz criteria, stability conditions become as follows:

$$\begin{cases} -(f_{11}+f_{12}+f_{13})>0, \\ (-f_{11}f_{22}f_{33}+f_{11}f_{32}f_{23}+f_{12}f_{21}f_{33}-f_{12}f_{31}f_{23}-f_{13}f_{21}f_{32}+f_{13}f_{31}f_{22})>0, \\ (f_{11}+f_{12}+f_{13})(-f_{11}f_{22}-f_{11}f_{33}-f_{22}f_{33}+f_{32}f_{23}+f_{12}f_{21}+f_{13}f_{31})>(-f_{11}f_{22}f_{33}+f_{11}f_{32}f_{23}+ \\ f_{12}f_{21}f_{33}-f_{12}f_{31}f_{23}-f_{13}f_{21}f_{32}+f_{13}f_{31}f_{22}). \end{cases}$$

Theorem 3 *The equilibrium point* $E^{(8)}(S^{(8)}, I^{(8)}, Y^{(8)})$ *shows hopf bifurcation subject to the following conditions*;

i. $f_{11}f_{23}f_{32}-f_{11}f_{22}f_{33}+f_{12}f_{22}f_{33}+f_{31}f_{13}f_{22}-f_{12}f_{23}f_{31}-f_{13}f_{21}f_{32}\,|_{m_{11}}>0,$

ii. $(f_{11}f_{22}+f_{11}f_{33}+f_{33}f_{22}-f_{32}f_{23}-f_{12}f_{22}-f_{31}f_{13})\,|_{m_{11}}>0,$

where, $m_1=m_{11}$ is a root of $(-f_{11}-f_{22}-f_{33})(f_{11}f_{22}+f_{11}f_{33}+f_{33}f_{22}-f_{32}f_{23}-f_{12}f_{22}-f_{31}f_{13})$
$-f_{11}f_{23}f_{32}-f_{11}f_{22}f_{33}+f_{12}f_{22}f_{33}+f_{31}f_{13}f_{22}-f_{12}f_{23}f_{31}-f_{13}f_{21}f_{32}=0$

iii. The transversality condition at $m_1=m_{11}$ *holds* [26].

Proof Hopf bifurcation occurs at $m_1=m_{11}$ when the following conditions are satisfied as,

1. $t_0(m_{11})>0$ which is true (from the above stability condition).
2. $t_1(m_{11})>0$ which is true (by the analysis above).
3. The condition $D_2(m_{11})=t_1(m_{11})t_2(m_{11})-t_0(m_{11})=0$ gives,

$$m_{11}=\frac{-(A_1-2B_1)+\sqrt{(A_1-2B_1)^2-4B_1D_1}}{2B_1},\ \text{provided } B_1D_1<0,\ \text{where,}$$

$$A_1=-f_{22}f_{33}-f_{32}f_{23}+f_{13}f_{31}+f_{23}f_{32}-f_{22}f_{33}+f_{12}f_{22}+f_{12}f_{33}-f_{22}f_{22}-f_{33}f_{33}-2f_{22}f_{33},$$

$$B_1=f_{22}+f_{33},$$

$$D_1=(f_{11}+m_1)A_1-(f_{11}+m_1)^2B_1+f_{12}f_{22}f_{33}-f_{13}f_{21}f_{32}-f_{23}f_{32}f_{12}+f_{31}f_{22}f_{13}$$
$$-(-f_{12}+f_{22}+f_{33})(-f_{13}f_{31}-f_{23}f_{32}+f_{22}f_{33}-f_{12}f_{22}-f_{12}f_{33}).$$

4. Also, the following condition is required, $\frac{dD_2}{dm_1}=-A_1+2f_{11}(f_{22}+f_{33})\neq 0$, for $m_1=m_{11}$.

Global Stability

In this section, global stability of system (2.1) is discussed using geometric approach. We implement the method introduced by Li and Muldowney [28] in a concise manner and the method has also been practiced in [29].

Consider an autonomous n-dimensional system $f: D\to R^n, D\subset R^n$, an open set and $f\in C^1(D)$, where the dynamical system is defined by,

$$\frac{dx}{dt}=f(x). \tag{6.1}$$

The above defined system is globally stable under the following conditions listed as,

(H_1) The domain D is simply connected.
(H_1) System as defined above, has a unique positive equilibrium $x^*\in D$.
(H_3) There is a compact absorbing set $\Gamma\subset D$.

Definition 1 The unique positive equilibrium x^* of the system (6.1) attains global stability in domain D if it is locally stable and all trajectories in D converges to the interior equilibrium point x^*.

Definition 2 Let $J = (J_{ij})_n$ be the Jacobian matrix of the system (6.1) and $J^{[2]}$ be the second computed matrix with order

$$\binom{n}{2} \times \binom{n}{2},$$

i.e.

$$J = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

then, $J^{[2]}$ of the system is,

$$J^{[2]} = \begin{pmatrix} a_{11} + a_{22} & a_{23} & -a_{13} \\ a_{32} & a_{11} + a_{33} & a_{12} \\ -a_{31} & a_{21} & a_{22} + a_{33} \end{pmatrix}.$$

Let $V(x) \in C^1(D)$ be a

$$\binom{n}{2} \times \binom{n}{2}$$

matrix valued function. Consider a matrix P such that,

$$P = V_f V^{-1} + V J^{[2]} V^{-1},$$

where, the matrix V_f is represented by,

$$(V_{ij}(x))_f = \left(\frac{\partial V_{ij}}{\partial x}\right)^t . f(x) = \nabla V_{ij} . f(x).$$

Considering the *Lozinskiĭ* measure l of P with respect to a vector norm $|\,.\,|$ in R^N, $N = \binom{n}{2}$, then we have,

$$l(P) = \lim_{h \to 0^+} \frac{|\, W + hP \,| - 1}{h}. \tag{6.2}$$

Now, from [28], it is proved that if (H_1), (H_2) and (H_3) are satisfied along with the condition defined as,

$$\limsup \sup \frac{1}{t} \int_0^t l(P(x(s, x_0)))ds < 0, \tag{6.3}$$

then this ensures the absence of any orbits i.e. homoclinic orbits, hetroclinic cycles and periodic orbits, giving rise to a simple closed rectifiable curve in D, which is invariant for the dynamical system (6.1). It is known to be a robust Bendixson criterion.

For system (2.1), the jacobian matrix at (S, I, Y) is already defined in Sect. 5.3, so the corresponding second computed matrix is below given as,

$$J^{[2]} = \begin{pmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{pmatrix}.$$

where $q_{11} = r - \frac{2rs+rI}{k} - \beta I - \frac{p_1 Y}{(1+aS)^2} - m_1 + \beta S - \frac{p_2 Y}{(1+aI)^2} - \alpha - m_2 - d_1$, $q_{12} = \frac{-p_2 I}{(1+aI)}$, $q_{13} = \frac{p_1 S}{(1+aS)}$, $q_{21} = \frac{c_2 p_2 Y}{(1+aI)^2}$, $q_{22} = r - \frac{2rs+rI}{k} - \beta I - \frac{p_1 Y}{(1+aS)^2} - m_1 + \frac{c_1 p_1 S}{(1+aS)} + \frac{c_2 p_2 I}{(1+aI)} - d_2 - m_3$, $q_{23} = \frac{-rS}{K} - \beta S + \alpha$, $q_{31} = \frac{-c_1 p_1 Y}{(1+aS)^2}$, $q_{32} = \beta I$, $q_{33} = \beta S - \frac{p_2 Y}{(1+aI)^2} - \alpha - m_2 - d_1 + \frac{c_1 p_1 S}{(1+aS)} + \frac{c_2 p_2 I}{(1+aI)} - d_2 - m_3$.

Considering the matrix, $V(x) = diag(1, \frac{I}{Y}, \frac{I}{Y})$. It follows then,

$$V_f V^{-1} = -diag\left(0, \frac{\dot{I}}{I} - \frac{\dot{Y}}{Y}, \frac{\dot{I}}{I} - \frac{\dot{Y}}{Y}\right).$$

Now, with further simplification, we get,

$$P = V_f V^{-1} + V J^{[2]} V^{-1} = (P_{ij})_{2\times 2},$$

where,

$$\begin{aligned}
P_{11} &= r - \frac{2rs+rI}{k} - \beta I - \frac{p_1 Y}{(1+aS)^2} - m_1 + \beta S - \frac{p_2 Y}{(1+aI)^2} - \alpha - m_2 - d_1, \\
P_{12} &= \left[\frac{-p_2 Y}{(1+aI)}, \frac{-p_1 SY}{(1+aS)I}\right], \\
P_{21} &= \left[\frac{c_2 p_2 I}{(1+aI)^2}, \frac{-c_1 p_1 I}{(1+aS)^2}\right]^T, \\
P_{22} &= (\gamma_{ij}) \text{ is a } 2\times 2 \text{ matrix where,} \\
\gamma_{11} &= r - \frac{2rs+rI}{k} - \beta I - \frac{p_1 Y}{(1+aS)^2} - m_1 + \frac{c_2 p_2 I}{(1+aI)} + \frac{c_1 p_1 S}{(1+aS)} - d_2 \\
&\quad - m_3 + \frac{\dot{I}}{I} - \frac{\dot{Y}}{Y}, \gamma_{12} = \frac{-rS}{K} - \beta S + \alpha, \gamma_{21} = \beta I, \gamma_{22} = \beta S - \frac{p_2 Y}{(1+aI)^2} \\
&\quad - \alpha - m_2 - d_1 + \frac{c_2 p_2 I}{(1+aI)} + \frac{c_1 p_1 S}{(1+aS)} - d_2 - m_3 + \frac{\dot{I}}{I} - \frac{\dot{Y}}{Y}.
\end{aligned}$$

Define the norm in R^3 as $|(u, v, w)| = max\{|u|, |v + w|\}$; denote by l the *Lozinskiĭ* measure with respect to this norm. Then, it follows as,

$$l(P) \le sup\{g_1, g_2\} \equiv sup\{l_1(P_{11}) + |P_{12}|, l_1(P_{22}) + |P_{21}|\},$$

where $|P_{21}|, |P_{12}|$ are matrix norms with respect to the L^1 norm and l_1 denotes the *Lozinskiĭ* measure with respect to the L^1 norm. Now, $l_1(P_{11}) = r - \frac{2rs+rI}{k} - \beta I - \frac{p_1 Y}{(1+aS)^2} - m_1 + \beta S - \frac{p_2 Y}{(1+aI)^2} - \alpha - m_2 - d_1$, $|P_{12}| = max\left[\frac{p_2 Y}{(1+aI)}, \frac{p_1 SY}{(1+aS)I}\right]$, $|P_{21}| = \left[\frac{c_2 p_2 I}{(1+aI)^2} + \frac{c_1 p_1 I}{(1+aS)^2}\right)$, $l_1(P_{22}) = max\left[r - \frac{(2rS+rI)}{K} - \frac{p_1 Y}{(1+aS)^2} - m_1 + \frac{c_1 p_1 S}{(1+aS)} + \frac{c_2 p_2 I}{(1+aI)} - d_2 - m_3 + \frac{\dot{I}}{I} - \frac{\dot{Y}}{Y}, \frac{-rS}{K} - \frac{p_2 Y}{(1+aI)^2} - m_2 - d_1 + \frac{c_1 p_1 S}{(1+aS)} + \frac{c_2 p_2 I}{(1+aI)} - d_2 - m_3 + \frac{\dot{I}}{I} - \frac{\dot{Y}}{Y}\right]$. Now, using equations of system (2.1) and with further simplification, we get the following,

$$l_1(P_{11}) = r - \frac{2rs + rI}{k} - \beta I - \frac{p_1 Y}{(1+aS)^2} - m_1 + \frac{\dot{I}}{I} + \frac{p_2 aYI}{(1+aI)^2},$$
$$l_1(P_{22}) = max\Big[r - \frac{(2rS + rI)}{K} - \frac{p_1 Y}{(1+aS)^2} - m_1 + \frac{\dot{I}}{I}, \frac{-rS}{K} - \frac{p_2 Y}{(1+aI)^2} - m_2 - d_1 + \frac{\dot{I}}{I}\Big].$$

In view of system (2.1) as persistent, $\exists\ \gamma > 0$ and $\tau > 0$ such that $t > \tau$ and $S(t) \geq \gamma$, $I(t) \geq \gamma, Y(t) \geq \gamma$. Assuming that,

$$\widetilde{m} = \frac{-r + \frac{3r\gamma}{K} + \beta\gamma - m_1 - \frac{p_2 \widetilde{M}^2}{a\gamma^2} - \epsilon}{\widetilde{M} max(1, \frac{K}{\gamma})}, \bar{c} = min\{\frac{3r\gamma}{K^2}, \frac{r\gamma}{K^2} + \frac{m_2}{K} + \frac{d_1}{K} - \frac{\epsilon}{K}\},$$

where, ϵ is small and $\widetilde{M}$ is upper bound of system (2.1) as already discussed in lemma 2. So,

$$g_1 \leq r - \frac{3r\gamma}{K} - \beta\gamma + m_1 + \frac{p_2 \widetilde{M}^2}{a\gamma^2} + \widetilde{m}\widetilde{M} max(1, \frac{K}{\gamma}) + \frac{\dot{I}}{I} = \frac{\dot{I}}{I} - \epsilon.$$
$$g_2 \leq \frac{(c_1 p_1 + c_2 p_2)K}{(\alpha\gamma)} + max\Big\{r - \frac{2rS + rI}{K} - \frac{p_1 Y}{(1+aS)^2} - m_1, \frac{-rS}{K} - \frac{p_2 Y}{(1+aI)^2} - m_2 - d_1\Big\} + \frac{\dot{I}}{I}$$
$$\leq max\{\bar{c}K - \frac{3r\gamma}{K} - \epsilon, \bar{c}K - \frac{r\gamma}{K} - m_2 - d_1\} + \frac{\dot{I}}{I} \leq \frac{\dot{I}}{I} - \epsilon.$$

Therefore, we get,

$$\frac{1}{t}\int_0^t l(P)ds \leq \frac{1}{t}\int_0^\tau l(P)ds + \frac{1}{t}\log\frac{I(t)}{I(\tau)} - \epsilon\frac{t-\tau}{t},$$

which signifies equation (6.3) i.e.,

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t l(P)ds \leq \frac{-\epsilon}{2} < 0.$$

In view of the above assumptions, the unique equilibrium $E^{(8)}$ is globally stable.

Numerical Simulation

In this section, numerical simulation has been performed to assist the theoretical results by assuming the values of parameters.

We consider an example below as;

$$\begin{cases} \frac{dS(t)}{dt} = 1S\Big(1 - \frac{S+I}{1}\Big) - 0.4SI + 0.1I - \frac{0.01SY}{1+0.2S} - 0.1S, \\ \frac{dI(t)}{dt} = 0.4SI - \frac{0.01IY}{1+0.2I} - 0.1I - 0.1I - 0.1I, \\ \frac{dY(t)}{dt} = \frac{0.2SY}{1+0.2S} + \frac{0.02IY}{1+0.2I} - 0.2Y - 0.2Y. \end{cases}$$

The above system possesses the equilibrium points as (i)$E^{(1)}(0,0,0)$, (ii) $E^{(2)}(0.9,0,0)$ (iii) $E^{(5)}(0.75, 0.1184, 0)$

(iv) $E^{(8)}(0.89, 0.00108, 0.0000201)$. Migration parameters are varied by keeping other parameters constant and in this view, dynamics of the system (2.1) has been studied (Fig. 1).

Model Without Migration and Disease

By assuming the values of parameters as $m_1 = m_2 = 0$ with $I = 0$ in system (2.1). It becomes as:

$$\begin{cases} \frac{dS(t)}{dt} = \frac{1}{2}S\left(1 - \frac{S}{100}\right) - \frac{8SY}{1+\frac{1}{2}S}, \\ \frac{dY(t)}{dt} = \frac{16SY}{1+\frac{1}{2}S} - \frac{1}{4}Y. \end{cases}$$

The above system has all the three equilibrium points as (i)$E_1(0,0)$, (ii)$E_2(100,0)$, (iii)$E_3(\frac{2}{127}, \frac{25396}{403225})$. Using result in Sect. 4, it is observed that $E_1(0,0)$ is always unstable. Due to the condition $c_1p_1k - (d_2 + d_2ak) = \frac{51}{4} > 0$, $E_2(100,0)$ is unstable. Also, the condition $r - \frac{2r\tilde{S}}{k} - \frac{p_1\tilde{Y}}{(1+a\tilde{S})^2} + \frac{c_1p_1\tilde{S}}{(1+a\tilde{S})} - d_2 = 0.242 > 0$ exist, so the equilibrium point $E_3(\frac{2}{127}, \frac{25396}{403225})$ is locally unstable (Fig. 2).

Model with Migration but Without Disease

By assuming the values of parameters as $m_1 = \frac{1}{8}, m_2 = \frac{1}{8}$ with $I = 0$, system (2.1) becomes as:

$$\begin{cases} \frac{dS(t)}{dt} = \frac{1}{2}S\left(1 - \frac{S}{100}\right) - \frac{8SY}{1+\frac{1}{2}S} - \frac{1}{8}S, \\ \frac{dY(t)}{dt} = \frac{16SY}{1+\frac{1}{2}S} - \frac{1}{4}Y - \frac{1}{8}Y. \end{cases}$$

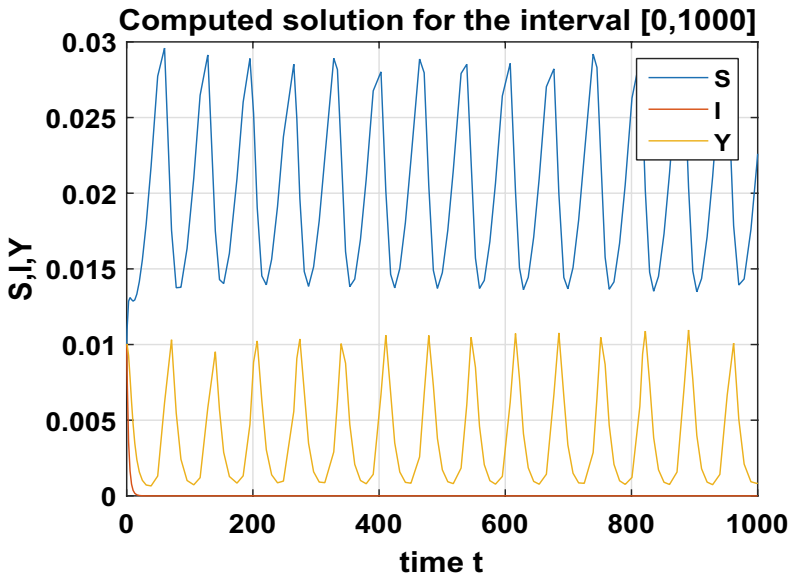


Fig. 1 Numerical solution for the example above showing E^8 is locally unstable

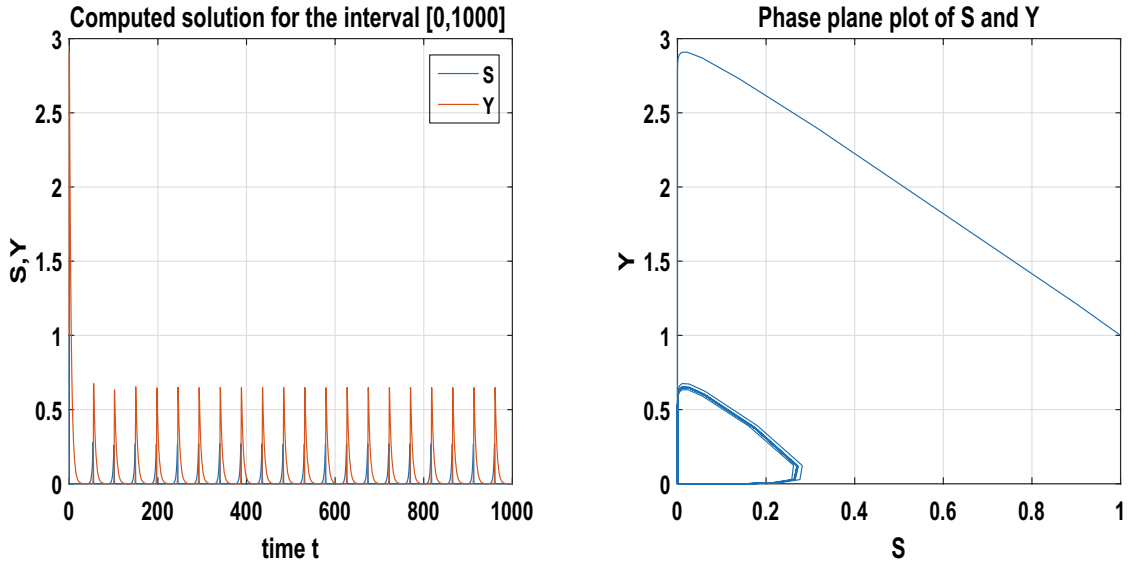


Fig. 2 Numerical solution for the model (2.1) without migration and disease ($m_1 = 0 = m_2, I = 0, r = \frac{1}{2}, k = 100, p_1 = 8, a = \frac{1}{2}, d_2 = \frac{1}{4}, c_1 = 2$)

The above system exhibits the existence of three equilibrium points as (1) $E_1^*(0,0)$, (2) $E_2^*(75,0)$, (3) $E_3(\frac{6}{253}, \frac{75876}{1600225})$. From Sect. 4, it is observed that $E_1(0,0)$ is unstable as the condition $r - m_1 < 0$ violates. The conditions $\frac{2rS^*}{k} + m_1 > r$ and $\frac{c_1 p_1 S^*}{1+aS^*} < d_2 + m_2$ are satisfied by taking the assumed parameters, so the equilibrium point $E_2(75,0)$ is stable. On substituting the values of parameters in condition, we get as, $r + \frac{c_1 p_1 \tilde{S}}{1+a\tilde{S}} = 0.874 \not< 0.870 = (\frac{2r\tilde{S}}{k} + \frac{p_1 \tilde{y}}{(1+a\tilde{S})^2} + m_1 + m_2 + d_2)$. So the equilibrium point $E_3(\frac{6}{253}, \frac{75876}{1600225})$ is locally unstable (Fig. 3).

Model with Migration and Disease ($m_1 > m_2$)

By assuming the values of parameters as $\alpha = \frac{1}{4}, \beta = \frac{1}{2}, c_1 = 2, c_2 = 2, m_2 = \frac{1}{16}, d_1 = \frac{1}{6}$(rest all parameters same). The system (2.1) becomes as:

$$\begin{cases} \frac{dS(t)}{dt} = \frac{1}{2}S\left(1 - \frac{S+I}{100}\right) - \frac{1}{2}SI + \frac{1}{4}I - \frac{8SY}{1+\frac{S}{2}} - \frac{1}{8}S, \\ \frac{dI(t)}{dt} = \frac{1}{2}SI - \frac{9IY}{1+\frac{I}{2}} - \frac{1}{4}I - \frac{1}{6}I - \frac{1}{16}I, \\ \frac{dY(t)}{dt} = \frac{16SY}{1+\frac{S}{2}} + \frac{18IY}{1+\frac{I}{2}} - \frac{1}{4}Y - \frac{1}{8}Y. \end{cases}$$

The above system shows the existence of the equilibrium point $E^{(1)}(0,0,0)$ but as the condition $r < m_1$ is not true, so $E^{(1)}(0,0,0)$ is unstable.

The equilibrium point $E^{(2)}(75,0,0)$ exists as the condition $r - m_1 = \frac{3}{8} > 0$ holds true and from the results in Sect. 5.3, the condition $\beta S^{(2)} < \alpha + m_2 + d_1$ violates, so it is an unstable equilibrium point.

The equilibrium point $E^{(5)}(\frac{23}{24}, \frac{40871}{26952}, 0)$ exist which is in accordance with the results defined in Sect. 5.2 i.e., $r > \frac{rL}{k} + m_1$ and $\frac{rL}{k} + \beta L > \alpha$ and from Sect. 5.3, it is an unstable equilibrium point. Further, it is observed that $E^{(6)}(\frac{6}{253}, 0, \frac{75876}{1600225})$ exists as the conditions

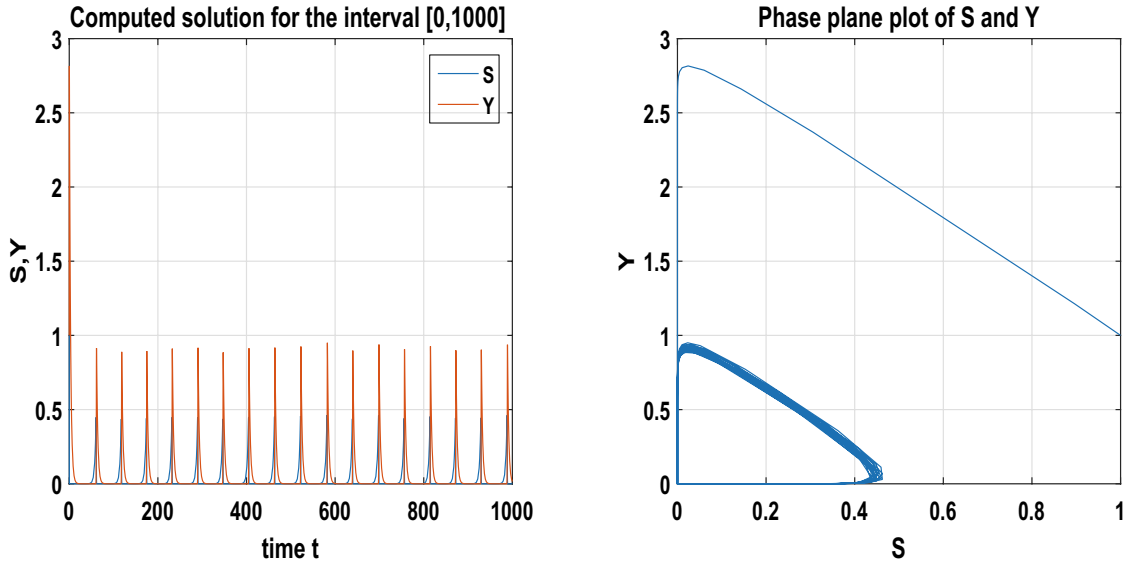


Fig. 3 Numerical solution for the model (2.1) with $m_1 = \frac{1}{8} = m_2, I = 0, r = \frac{1}{2}, k = 100, p_1 = 8, a = \frac{1}{2}, d_2 = \frac{1}{4}, c_1 = 2$

$c_1p_1 - d_2a - m_3a = \frac{253}{16} > 0$ and $r\left(1 - \frac{T}{k}\right) - m_1 = 0.37488 > 0$ hold true and is stable (from results in Sects. 5.2 and 5.3).

The interior equilibrium point $E^{(8)}$ does not exist (Fig. 4).

The basic reproduction number $R_0 = 0.027 < 1$ is found from the result as discussed in previous section. Hence, the disease will not be endemic.

Model with Migration and Disease ($m_1 < m_2$)

By assuming the values of parameters as $m_1 = \frac{3}{10}, m_2 = \frac{1}{2}$,(rest all parameters same as in case 7.3). So, the system (2.1) becomes as:

$$\begin{cases} \frac{dS(t)}{dt} = \frac{1}{2}S\left(1 - \frac{S+I}{100}\right) - \frac{1}{2}SI + \frac{1}{4}I - \frac{8SY}{1+\frac{S}{2}} - \frac{3}{10}S, \\ \frac{dI(t)}{dt} = \frac{1}{2}SI - \frac{9IY}{1+\frac{I}{2}} - \frac{1}{4}I - \frac{1}{6}I - \frac{1}{2}I, \\ \frac{dY(t)}{dt} = \frac{16SY}{1+\frac{S}{2}} + \frac{18IY}{1+\frac{I}{2}} - \frac{1}{4}Y - \frac{1}{8}Y. \end{cases}$$

The above system exhibits the existence of the equilibrium point $E^{(1)}(0, 0, 0)$ and as the condition $r > m_1$ is true, so $E^{(1)}(0, 0, 0)$ is a stable equilibrium point.

The equilibrium point $E^{(2)}(40, 0, 0)$ exists as the condition $r - m_1 = \frac{1}{5} > 0$ holds true and from the results in Sect. 5.3, the condition $\beta S^{(2)} < \alpha + m_2 + d_1$ does not validate, so it is an unstable equilibrium point.

The equilibrium point $E^{(5)}(\frac{11}{6}, \frac{2519}{4866}, 0)$ exists which is in accordance with the results defined in Sect. 5.2 i.e.,

$r > \frac{rL}{k} + m_1$ and $\frac{rL}{k} + \beta L > \alpha$ and from Sect. 5.3, it is an unstable equilibrium point as the condition $\frac{c_1p_1S^{(5)}}{1+aS^{(5)}} + \frac{c_2p_2I^{(5)}}{1+aI^{(5)}} < d_2 + m_3$ violates. Further, it is observed that $E^{(6)}(\frac{6}{253}, 0, \frac{40456}{1600225})$ exists as

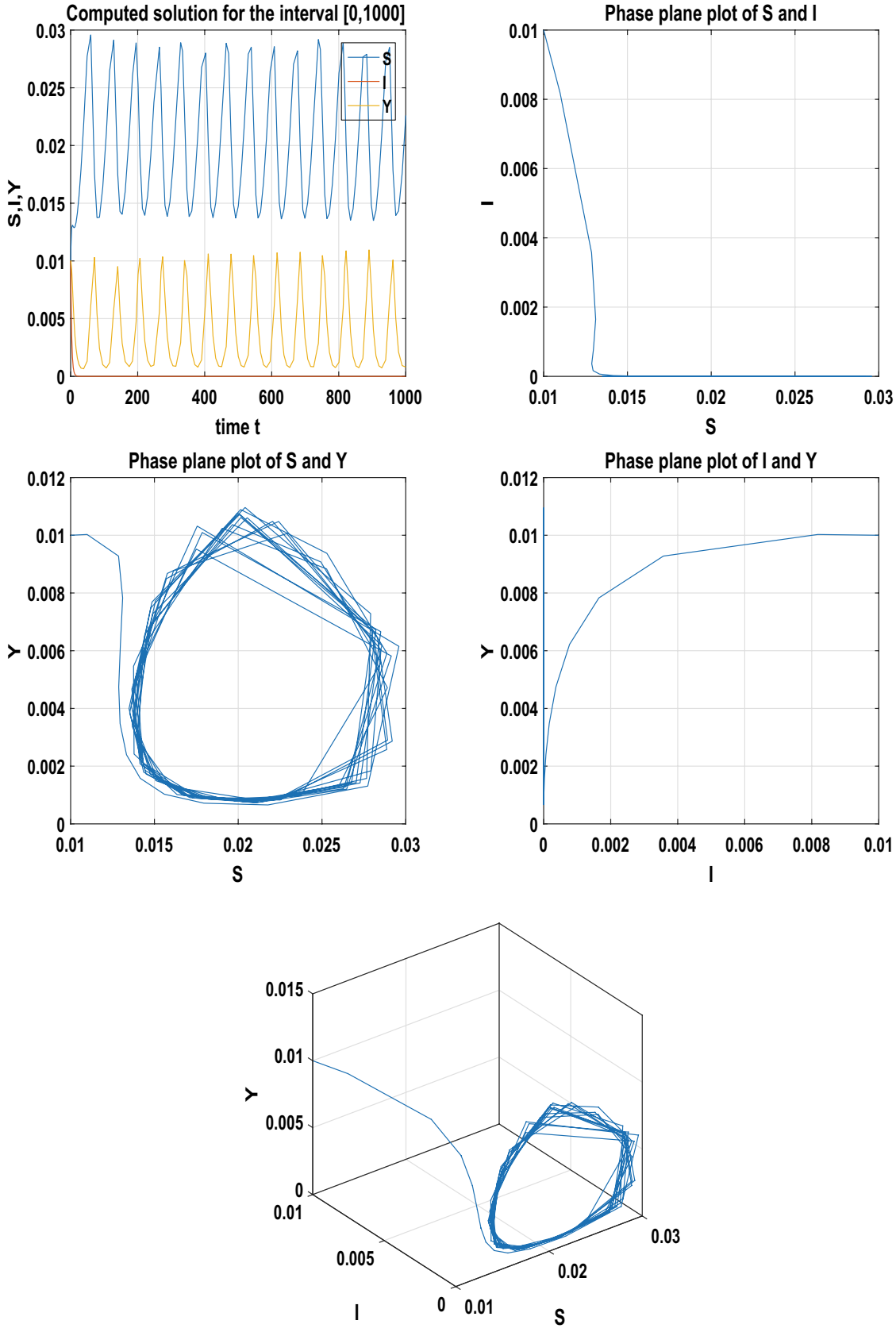


Fig. 4 Numerical solution for the model (2.1) with $m_1 = \frac{1}{8}, m_2 = \frac{1}{16}, r = \frac{1}{2}, k = 100, p_1 = 8, a = \frac{1}{2}, d_2 = \frac{1}{4}$, $c_1 = 2, c_2 = 2, \alpha = \frac{1}{4}, \beta = \frac{1}{2}$

the conditions $c_1p_1 - d_2a - m_3a > 0$ and $r\left(1 - \frac{T}{k}\right) - m_1 > 0$ hold true. It is a stable equilibrium point. (From the results in Sects. 5.2 and 5.3.)

The interior equilibrium point $E^{(8)}$ does not exist (Fig. 5).

The basic reproduction number i.e., $R_0 = 0.015 < 1$ and hence, the disease will not be endemic.

Discussions

It is implied from model (2.1) that migration and mortality are the two different conditions though they both indicates decease in population. Present study emphasis on the significance of migration in prey–predator dynamics with disease in prey population. It is observed that equilibrium points $E^{(3)}$ and $E^{(4)}$ do not exist which conveys the same meaning ecologically also as it is challenging for predators to live without preys. However, studies are available in literature where predators can survive without prey if the additional food conception is introduced in the system [30, 31]. The parameters m_1, m_2 and m_3 have the range [0, 1] and if, $m_1 = 0 = m_2 = m_3$, then extinction of population is feasible due to mortality only either by the way of natural death or death due to infection. If $m_1 = 1 = m_2 = m_3$, then the present study focusses on the loss of species followed through migration. Both these conditions are not pragmatic in approach [17]. In the present article, prey–predator relation is studied with disease in prey population where migration is allowed. Mathematically, the paper features migration as equivalent to the death rate only, but apparently it characterized differently.

The equilibrium points $E_3(\widetilde{S}, \widetilde{Y}), E_3^*(\widetilde{S}, \widetilde{Y})$ and $E^{(6)}(\widetilde{S}, 0, \widetilde{Y})$ as discussed in Sects. 4.1, 4.2 and 5.3 respectively, are the infection free states in system. Also, $E_2(k, 0), E_2^*(S^*, 0)$ and $E^{(2)}(S^{(2)}, 0, 0)$ as already determined in previous sections, are the equilibrium points involving only healthy prey and if they become stable at some point, then it will remain in that state for a longer time as no prey will be predated by predators. The equilibrium points $E_1(0, 0), E_1^*(0, 0)$ and $E^{(1)}(0, 0, 0)$ are always unstable as their stability would lead to extinction of population in ecosystem. These points are not important both mathematically and ecologically. The equilibrium points $E^{(3)}$ and $E^{(4)}$ do not exist as infected class of prey will die soon due to which predators cannot survive. Analogously, the equilibrium points $E^{(4)}(0, 0, {Y_1}^{(4)}, 0)$, $E^{(5)}(0, 0, 0, {Y_2}^{(5)})$ and $E^{(3)}(0, I^{(3)}, 0, 0)$ in [17] do not exist. It is to be noted that $E^{(7)}$ does not exist as the infected prey population existence is inconsequential for healthy predators survival unlike the equilibrium points $E^{(10)}(0, I^{(10)}, {Y_1}^{(10)}, {Y_2}^{(10)})$ and $E^{(9)}(0, 0, {Y_1}^{(9)}, {Y_2}^{(9)})$ in [17] which exist. The viable reason is the presence of infected predator class in [17] while in the present paper, only prey population is infected with disease and the possible ways for the extinction of predator population is through migration along with the natural death.

Further, the dynamics of the system may show variation when migration parameters are changed while keeping other parameters same. In Sect. 7.3, $m_1 > m_2$ i.e., migration rate of susceptible prey population is more as compared to infective ones and predator population being dependent more on susceptible prey population shows rapid oscillations. It is seen that infective prey population tends to zero over time, supplemented with increase in migration of susceptible prey population will ultimately leads to an unstable system. Also, from Fig. 4, it is observed that predator population density increases with increase in infective prey population and susceptible prey population exhibits rapid growth rate as

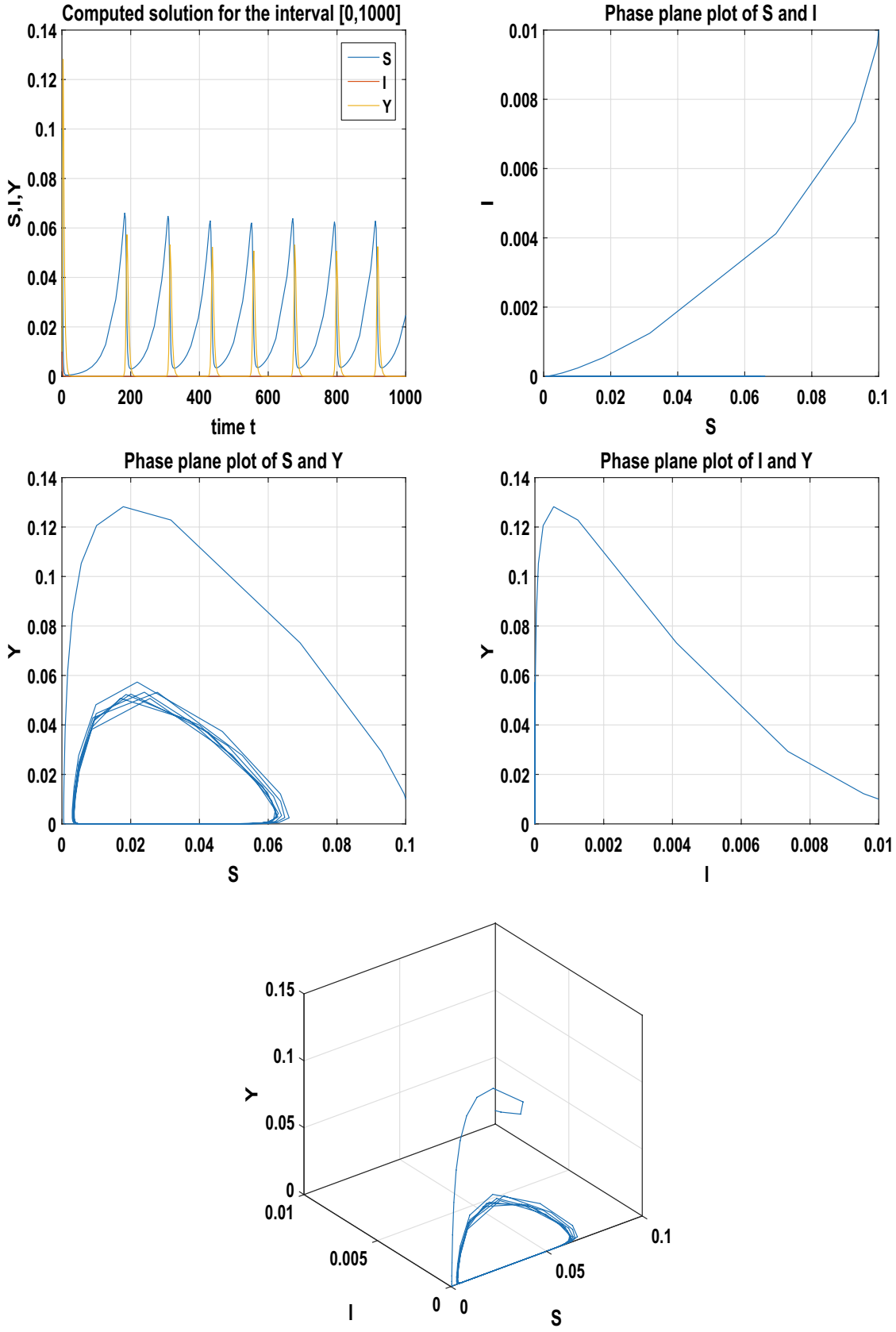


Fig. 5 Numerical solution for the model (2.1) with $m_1 = \frac{3}{10}, m_2 = \frac{1}{2}, r = \frac{1}{2}, k = 100, p_1 = 8, a = \frac{1}{2}, d_2 = \frac{1}{4}, c_1 = 2, c_2 = 2, \alpha = \frac{1}{4}, \beta = \frac{1}{2}$

the infected class of preys shows decline in its population. In Fig. 5, $m_1 < m_2$ i.e., migration rate of susceptible prey population is less as compared to infective prey population. It is observed from Fig. 5 that, in the initial phase, predator population shows a prompt increase in its growth even at the low population density of infective class of prey because of the presence of susceptible prey population. But gradually, predator population shows a decline in its growth even when there is an increase in the population of infective prey. The reason could be the migration in predator population which is substantially due to the fact that increase in the migration rate of infective class of prey leads to higher migration rate of predator class as well (Fig. 5).

In Fig. 6, the relationship between migration in susceptible prey population, infective preys and predators has been depicted. As the migration rate of susceptible prey population is increasing, infected prey population is decreasing due to the dependency of predators on infective prey population extensively. Generally, it can be asserted that infective preys are more easy to predate due to the lack of strength compared to susceptible preys. As a result, predators density rate will not decline much by the virtue of migration in susceptible prey density.

It has been seen that many factors effect the dynamical behaviour of prey–predator relationship and one of the most crucial component is the nonlinear functional response. Many biologists and mathematicians have experimented with different functional responses in the prey–predator interaction as ax(Holling type-I), $\frac{ax}{b+x}$(Holling type-II), $\frac{ax^2}{b+x^2}$(Holling type-III), $\frac{ax^2}{b+x+cx^2}$(Holling type-IV), ratio dependent response, Beddington DeAngelis type response etc. [32–37]. The dynamics of Holling type-II functional response is well studied in literature [38–40]. Motivated by the work of Chunyan Ji et al., Holling type-II functional response is incorporated in the system and some captivating dynamics is observed unlike Lotka-Volterra type linear functional response in [17].

Global stability of the system around the interior positive equilibrium point using geometric approach is captivating. Even though the geometric approach is not prominent as the Lyapunov method but the framing of the Lyapunov function is challenging. Besides, geometric approach is appropriate for systems with dimensions 4 or low since to evaluate the second compounded matrix is arduous. Hopf bifurcation conditions are also explained (in brief) for the system with no disease and the system with disease as in [41].

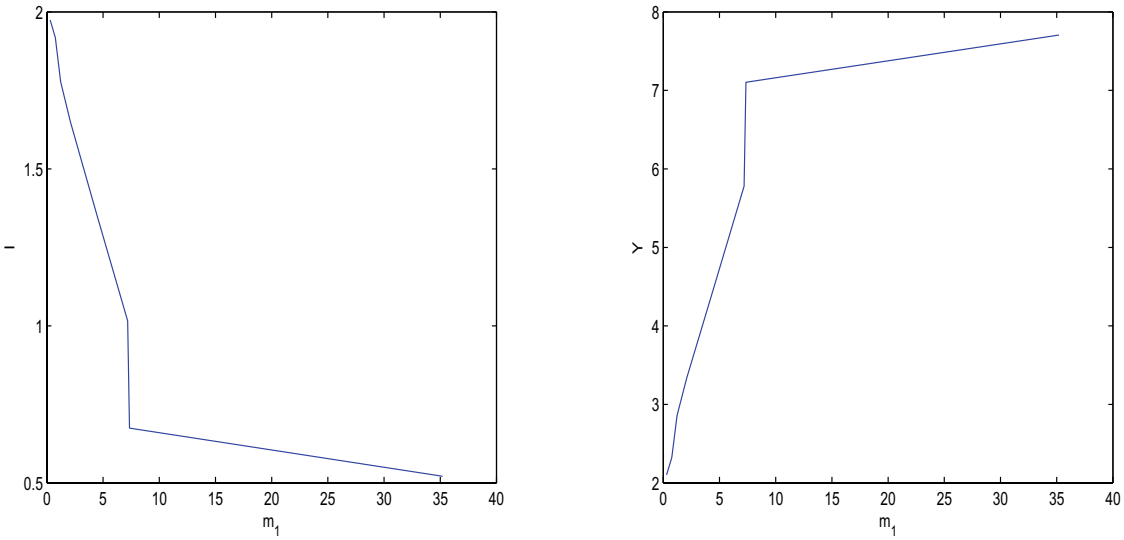


Fig. 6 Numerical simulation depicting the relationship of m_1 with Y and I

In general, the reproduction number is a threshold quantity that depends on various factors such as place, density etc. In the present paper, for the systems 7.2 and 7.3 reproduction number is significantly less than 1. The definition of R_0 is closely related with the stability of disease free equilibrium point i.e. if $R_0 < 1$, then equilibrium point is stable and unstable if $R_0 > 1$. Also, it has been observed that there is a formation of unstable limit cycle (Figs. 4 and 5). Rate of migration in the present paper (as in [17]) can be considered to follow different distributions to study the variation in the system dynamics. Also, efforts can be made in collecting data on migration of species for a further prominent analysis which is our future extent of study.

Acknowledgements Authors are thankful to anonymous reviewers for their very useful and valuable suggestions.

References

1. Hethcote, H.W.: A thousand and one epidemic models. In: Levin, S.A. (ed.) Frontiers in Mathematical Biology, Lecture Notes in Biomathematics, vol. 100, pp. 504–515. Springer, Berlin (1994)
2. Anderso, R.M., May, R.M.: Infectious Disease of Humans, Dynamics and Control. Oxford University Press, London (1991)
3. Xiao, Y., Chen, L.: Analysis of a three species eco-Eepidemiological model. J. Math. Anal. Appl. **258**, 733–754 (2001)
4. Haque, M., Zhen, J., Venturino, E.: An ecoepidemiological predator-prey model with standard disease incidence. Math. Methods Appl. Sci. **32**(7), 875–898 (2008). https://doi.org/10.1002/mma.1071
5. Cosner, C., DeAngelis, D.L., Ault, J.S., Olson, D.B.: Effects of spatial grouping on the functional response of predators. Theor. Popul. Biol. **56**, 65–75 (1999)
6. Rahman, S., Chakravarty, S.: A predator-prey model with disease in prey. Nonlinear Anal. Model. Control **18**(2), 191–209 (2013)
7. Sinha, S., Misra, O.P., Dhar, J.: Modelling a predator–prey system with infected prey in polluted environment. Appl. Math. Model. **34**, 1861–1872 (2010). https://doi.org/10.1016/j.apm.2009.10.003
8. Jana, S., Kar, T.K.: Modeling and analysis of a prey-predator system with disease in the prey. Chaos Solitons Fractals **47**, 42–53 (2013). https://doi.org/10.1016/j.chaos.2012.12.002
9. Johri, A., Trivedi, N., Sisodiya, A., Sing, B., Jain, S.: Study of a prey–predator model with diseased prey. Int. J. Contemp. Math. Sci. **7**(10), 489–498 (2012)
10. Greenhalgh, D., Haque, M.: A predator prey model with disease in the prey species only. Math. Methods Appl. Sci. **30**, 911–929 (2007). https://doi.org/10.1002/mma.815
11. Xiao, Y., Chen, L.: Analysis of a three species eco-epidemiological model. J. Math. Anal. Appl. **258**, 733–754 (2001). https://doi.org/10.1006/jmaa.2001.7514
12. Maiti, A., Jana, M.M., Samanta, G.P.: Deterministic and stochastic analysis of a ratio-dependent predator-prey system with delay. Nonlinear Anal. Model. Control **12**(3), 383–398 (2010). www.lana.lt/journal/26/Maiti.pdf. Accessed Dec 2018
13. Wang, F., Kuang, Y., Ding, C., Zhang, S.: Stability and bifurcation of a stage-structured predator-prey model with both discrete and distributed delays. Chaos Solitons Fractals **46**, 19–27 (2013). https://doi.org/10.1016/j.chaos.2012.10.003
14. Arora, C., Kumar, V., Kant, S.: Dynamics of a high-dimensional stage-structured prey–predator model. Int. J. Appl. Comput. Math. (2017). https://doi.org/10.1007/s40819-017-0363-z
15. Han, L., Zhien, M.A.: Four predator prey models with infectious diseases. Math. Comput. Model. **34**(7–8), 849–858 (2001). https://doi.org/10.1016/S0895-7177(01)00104-2
16. Hsieh, Y.H., Hsiao, C.K.: Predator-prey model with disease infection in both populations. Math. Med. Biol. **25**, 247–266 (2008). https://doi.org/10.1093/imammb/dqn017
17. Kant, S., Kumar, V.: Stability analysis of predator prey system with migrating prey and disease infection in both species. Appl. Math. Model. **42**, 509–539 (2017). https://doi.org/10.1016/j.apm.2016.10.003
18. Dingle, H., Drake, V.A.: What is migration? BioScience **57**(2), 113–121 (2007)
19. Sun, G.Q., Jin, Z., Liu, Q.X., Li, L.: Dynamical complexity of a spatial predator–prey model with migration. Ecol. Model. **219**(1–2), 248–255 (2008). https://doi.org/10.1016/j.ecolmodel.2008.08.009

20. Liu, M., Jin, Z., Haque, M.: An impulsive predator-prey model with communicable disease in the prey species only. Nonlinear Anal. Real World Appl. **10**(5), 3098–3111 (2009). https://doi.org/10.1016/j.nonrwa.2008.10.010
21. Jong, M.C.M.D.: How does infection depend on the population size? In: Mollison, D. (ed.) Epidemic Models, Their Structure and Relation in Data. Cambridge University Press, Cambridge (1994)
22. Subhas, K.: Dynamic behavior of a Beddington- DeAngelis type stage-structured predator-prey model. Appl. Math. Comp. **244**, 344–360 (2014). https://doi.org/10.1016/j.amc.2014.06.109
23. Sunita, G., Anuraj, S.: Control of chaos due to additional predator in the Hastings-Powell food chain model. J. Math. Anal. Appl. **385**, 423–438 (2012). https://doi.org/10.1016/j.jmaa.2011.06.047
24. Nagumo, M.: Uber die Lage der Integralkurven gew onlicher Differentialgeichungen. Proc. Phys. Math. Soc. Jpn. **24**, 555 (1942)
25. Birkhoff, G., Rota, G.C.: Ordinary Differential Equations. Ginn, Boston (1982)
26. Liu, W.: Criterion of Hopf bifurcations without using eigenvalues. J. Math. Anal. Appl. **182**(1), 250–256 (1994)
27. http://web.stanford.edu/jhj1/teachingdocs/Jones-on-R0.pdf. Accessed Dec 2018
28. Li, M.Y., Muldowney, J.S.: A geometric approach to global stability problems. SIAM J. Math. Anal. **27**(4), 1070–1083 (1996)
29. Haque, M., Zhen, J., Venturino, E.: An ecoepidemiological predator-prey model with standard disease incidence. Math. Methods Appl. Sci. **32**, 875–898 (2008)
30. Prasad, B.S.R.V., Banerjee, M., Srinivasu, P.D.N.: Dynamics of additional food provided predator-prey system with mutually interfering predators. Math. Biosci. **246**, 176–190 (2013). https://doi.org/10.1016/j.mbs.2013.08.013
31. Arora, C., Kumar, V.: Dynamics of one-prey and two-predator system highlighting the significance of additional food for predators with Beddington-DeAngelis functional response. Differ. Equ. Dyn. Syst. (2018). https://doi.org/10.1007/s12591-018-0442-6
32. Holling, C.S.: Some characteristics of simple types of predation and parasitism. Can. Entomol. **91**, 385–395 (1959)
33. Tian, X., Xu, R.: Global dynamics of a predator-prey system with Holling type II functional response. Nonlinear Anal. Model. Control **16**(2), 242–253 (2011). www.lana.lt/journal/41/NA16210.pdf. Accessed Dec 2018
34. Dawes, J.H., Souza, M.O.: A derivatation of Holling's type I, II and III functional responses in predator-prey system. J. Theor. Biol. **327**, 11–22 (2013). https://doi.org/10.1016/j.thbi.2013.02.017
35. Book, H.: A food chain system with Holling type IV functional response and impulsive perturbations. Comput. Math. Appl. **60**, 1152–1163 (2010)
36. Xu, R., Chaplain, M.A.J., Davidson, F.A.: Persistence and global stability of a ratio dependent prey model with stage-structure. Appl. Math. Comput. **158**(3), 729–744 (2004). https://doi.org/10.1016/j.amc.2003.10.012
37. Liu, S., Beretta, E.: A stage-structured predator-prey model of Beddington-DeAgelis type. SIAM J. Appl. Math. **66**(4), 1101–1129 (2006). https://doi.org/10.1137/050630003
38. Ji, C., Jiang, D., Shi, N.: Analysis of a predator–prey model with modified Leslie-Gower and Holling-type II schemes with stochastic perturbation. J. Math. Anal. Appl. **359**, 482–498 (2009). https://doi.org/10.1016/j.jmaa.2009.05.039
39. Chen, L., Chen, F., Chen, L.: Qualitative analysis of a predator-prey model with Holling type II functional response incorporating a constant prey refuge. Nonlinear Anal. Real World Appl. **11**, 246–252 (2010). https://doi.org/10.1016/j.nonrwa.2008.10.056
40. Liu, X., Chen, L.: Complex dynamics of Holling type II Lotka-Volterra predator–prey system with impulsive perturbations on the predator. Chaos Solitons and Fractals. **16**(2), 311–320 (2003). https://doi.org/10.1016/S0960-0779(02)00408-3
41. Upadhyay, R.K., Iyengar, S.R.K.: Introduction to Mathematical Modeling and Chaotic Dynamics. Chapman and Hall, London (2013)

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

Received: 20 March 2020 | Revised: 13 May 2020 | Accepted: 13 May 2020

DOI: 10.1002/pc.25666

Check for updates

Polymer COMPOSITES WILEY

REVIEW ARTICLE

Effect of reinforcement architecture on the macroscopic mechanical properties of fiberous polymer composites: A review

Furkan Ahmad[1] | **N Yuvaraj**[1] | **Pramendra K. Bajpai**[2]

[1]Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi, 110042, India

[2]MPAE Division, Netaji Subhas University of Technology, Dwarka, New Delhi, 110078, India

Correspondence
Furkan Ahmad, Department of Mechanical Engineering, Delhi Technological University, Shahbad Daulatpur, New Delhi 110042, India.
Email: furkanme13@gmail.com

Abstract

In order to curb the fuel consumption and corresponding CO_2 emission, defence, automobile, and aerospace industries are leaning towards the use of light weight-high stiffness engineered materials like fiber reinforced polymer composites (FRPCs). One of the major advantages of FRPCs is that various properties (stiffness, tensile strength, flexural strength, etc.) can be tailored according to requirements of the application. Architecture of the reinforcement used in FRPCs has been proved to affect these properties substantially. Composite materials have seen a lot of advancement in the field of reinforcement architectures starting from the plain-woven fabrics to advanced 3D braided/woven preforms. The architecture of reinforcement must be wisely selected to design the required properties of developed composites and to reduce the overall cost without compromising the performance. The problem addressed in the present study is the selection of reinforcement architecture while developing a composite material. 2D Plain woven reinforcements are better than other reinforcements in terms of in-plane properties but their out-of-plane properties are very poor. Therefore, advanced architectures like 3D woven fabrics, 5D Braided preforms, knitted preforms, and 3D needle punched fibers are being used in various high performance applications. The present article is an attempt to analyse the effect of various available architectures of reinforcement on the macroscopic mechanical properties of FRPC laminates and to propose a systematic way to select a reinforcement.

KEYWORDS

3D braided fibers, 3D knitted fibers, polymer composites, reinforcement architectures, textile architectures

1 | INTRODUCTION

Aerospace and defense industries are focusing on light weight and high strength materials for structural and non-structural components. Fiber reinforced polymer composites (FRPCs) are promising materials for aerospace and automotive industries. The consumption of FRPCs has increased by many folds over the last couple of decades.[1] There have been continuous efforts in the field of polymer composites to investigate and optimize the mechanical properties of developed materials either by experimentation or simulation.[2] Some of the researchers have used

 © 2020 Society of Plastics Engineers

hybridization of reinforcement to improve mechanical properties of the developed composites[3] while some other researchers have adopted different methods like fiber treatment[4,5] and optimization the processing conditions.[6–8] In the last few years, 3D woven textile preforms have gained much popularity as high performance reinforcements in the field of FRPCs. Advanced near-net-shape preforms has made it possible to reduce the wastage of material resources and overall cost of the developed products. These textile preforms are also being used to improve the properties of developed composites due to their higher drapability and better out-of-plane properties than 2D woven reinforcements. Various mechanical properties of FRPCs like toughness can be substantially increased by precisely controlling the architecture of reinforcement.[9] A promising method to control the architecture of reinforcement is to use the various available textile fabrication methods like braiding, knitting and various other types of weaving.

Figure 1 shows the classification of almost every form of reinforcement which is being used in polymer composite materials. Textile preforms are most popular form among all the available forms of reinforcement due to their ability to help in tailoring the mechanical properties according to the requirements of an application. Textile preforms are made by interlacing/interlocking of yarns in a particular fashion to achieve a particular architectural geometry. Due to favorable properties like low cost, light weight, engineered properties and easy processing of the textile preform reinforced polymer composites, automobile, aerospace, and marine industries are using them heavily to produce various structural and loading components. However, predicting and controlling the properties

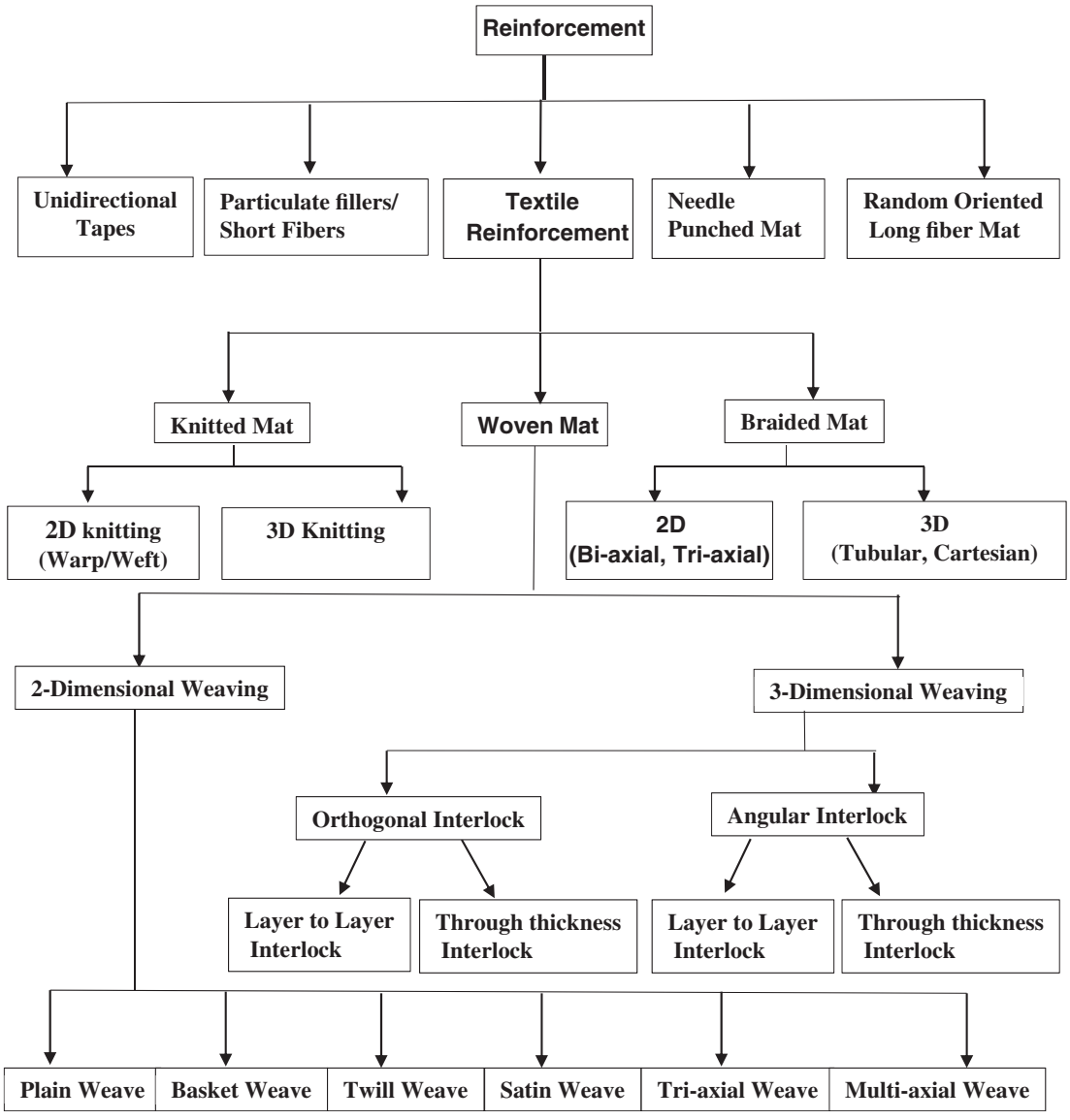


FIGURE 1 Commonly used reinforcement architectures in polymer composites

TABLE 1 Various textile architectures used by researchers in FRPCs

Sr. No.	Fiber	Fabric architectures	Prominent results	Refs.
1	Carbon fiber/PEI	Plain weave, twill, 4-harness satin	Twill weave architecture is better than plain and satin weave in terms of strength, modulus, and sliding wear resistance	[16]
2	Kevlar/epoxy	Plain weave, twill, basket, and 5 H-satin	Plain weave have the highest crimp due to highest cross-over points	[11]
3	Glass fiber/phenolic resin	8 Harness satin weave	Predicting out-of-plane properties with the help of Abaqus	[20]
4	Glass fiber/epoxy resin	Tri-axial woven fabrics	Tri-axial woven fabrics showed excellent shear resistance properties	[21]
5	Carbon fiber/epoxy resin	Two different twill weave fabrics	Higher crimp in fabric structure leads to shorter fatigue life	[22]
6	Flax fiber/epoxy resin	Plain woven fabric with different microstructural arrangement	Volume fraction is more influencing factor than microstructural arrangement of the reinforcement	[9]
7	Carbon fiber/epoxy resin	3D woven (angular and orthogonal interlocked)	Load bearing fibers must be free from waviness for higher strength of the composite	[23]
8	E-Glass fiber/epoxy resin	3D orthogonal woven	Composite with molded in hole showed higher strength than the composite with drilled hole	[24]
9	Carbon fiber/epoxy resin	3D woven preforms	Stiffness decreases nonlinearly with increasing crimp. 3D woven preforms possess higher crimp than 2D preforms	[25]
10	Carbon fiber/epoxy resin	2D tri-axially braided preforms	Decreased elastic properties with increasing braiding angle	[26]
11	Glass fiber/epoxy resin	Braided preforms	Strength of the composite decreased with increasing braiding angle	[27]
12	Glass fiber/epoxy resin	Woven and knitted fabrics (Milano and Rib knitted)	Hybridization affect the impact strength of the developed composites	[28]
13	Glass fiber/epoxy resin	Weft knitted fabric	Strength in wale direction is higher than course direction	[29]
14	Carbon/epoxy	Uni-directional, 5H-satin and 2 × 2 twill preforms	Shear strength of twill weave fabric reinforced composite was higher than that of others	[30]
15	Kenaf fiber/polyester resin	Plain, twill, basket, satin, leno, mock leno	Basket weave showed highest fiber content compared to other weave structures	[31]

of textile composites is not an easy task as a number of variables affect the properties in a number of ways. Complex spatial organization of fibers in every textile architecture makes it even harder to analyze the mechanical properties of textile composites in comparison to nontextile fiber reinforced composites.[10] Various researchers are continuously making efforts to investigate the effects of weave/textile architecture on the macroscopic properties of textile preform reinforced composites. Some mentionable entries are shown in Table 1.

There are three broad categories based on the length scale in which the changes can be introduced in FRPCs namely, microscopic, mesoscopic, and macroscopic scale.[11] Chemical treatment of fiber and spinning mechanism of natural fibers to make yarns are the changes at microscopic level. Microscopic changes can improve mechanical properties of developed composites such as chemical treatment improves the strength of FRPCs by improving the interfacial adhesion between the fiber and matrix interface.[12] Replacing the yarn or matrix introduces the changes at macroscopic level. Macroscopic changes such as replacing the normal yarn with hybrid yarn results in improved strength of developed composites if all other parameters are kept constant.[13] Another method to improve the mechanical properties of developed composites without even changing the fiber or matrix material is by changing the geometrical arrangement of fibers/reinforcement. These changes would be the changes at mesoscale level which lies in between the microscopic and macroscopic changes. The present article covers only the mesoscale changes which can be done in order to improve mechanical properties of developed

composites. Fibers/yarns can be organized into a number of geometrical arrangements due to their intrinsic property of flexibility. The diameter of fiber is another important parameter along with the stiffness of the fiber which controls the flexibility of fiber which in turn controls their ease of fabric conversion process. Starting from smallest form of reinforcement, particulates/fillers and short fibers are smallest forms of reinforcement which can be used in FRPCs. Tensile and flexural properties of particulates/fillers reinforced polymer composites are quite low while impact properties of such composites are found to be good.[13,14] Tensile properties of particulate/filler reinforced FPRCs can be improved without compromising the impact resistance by using short fibers instead of particulate/filler form. Long fibers can be used with some predefined orientation using unidirectional tapes or without any orientation using random orientation mats and also the controlling and predicting mechanical properties of developed composite is easier and more accurate with unidirectional tapes than random orientation mats. In all the textile preforms, plain weave is the simplest form of textile architecture obtained by simply interlacing the two sets of yarns in two orthogonal directions namely, warp and weft.[15] Plain weave fabric offers good stability of the textile preform but drapability is poor due to higher interlacing points in comparison to other 2D weave architectures like twill, basket, and satin.[16] Twill weave fabrics are more durable than plain weave due to lower interlacing points and close packing of yarns.[15] Basket weave is similar to the plain weave fabric but weaved with two or more than two yarns simultaneously. Basket weave also offers good stability of the fabric with less number of cross-over points per unit area compared to plain weave fabric. Another architecture, satin weave fabrics offers excellent draping characteristics and smooth appearance compared to other weave fabrics.[17] Some fabrics are prepared using three or more tows of fibers moving in different directions known as tri-axial woven fabrics and multi-axial fabrics respectively.

Owing to the demand of high performance reinforcement, 3D woven textile preforms are becoming very popular due to their better out-of-plane mechanical properties compared to 2D woven reinforcements.[18] Interlocking of the third dimensional fibers (binder yarns) can be at some angle making it angular interlocked textile preform or perpendicular to the other fibers making it orthogonal interlocked preforms. Further, interlocking of the third directional yarn can be through thickness or layer by layer.[19] The mechanical behavior of orthogonal interlocked 3D woven textile reinforced composite material is substantially different than angular interlocked preform reinforced composites due to the placement of third fiber. Braided, knitted, and needle punched fabrics are also being used as reinforcement in polymer composite materials to fabricate various components. The aim of the present article is to review the effect of various reinforcement architectures starting from particulate form to advance 3D woven textile preforms on the macroscopic mechanical properties of FRPCs. The present article aims to provide an overview of the effect of various basic and advanced reinforcement architectures on the macroscopic mechanical properties of FRPCs. Fabrication process of preforms is not considered in the present article. To the author's best of knowledge, no such review was found available till date which have focused on the changes in mechanical properties with respect to geometrical changes in wide range of available reinforcement.

2 | NONCONTINUOUS FIBER REINFORCEMENT

Reinforcement in the form of short fibers and particulates/fillers/flakes are noncontinuous type of reinforcement which are being used to fabricate specific polymer composite components. Technical fibers are generally of the order of few tens of microns in diameter and ranges from few mm to virtually continuous in length.[32] Along with flexibility, one of the important properties of fiber is that they possess high strength along the direction of length. So, length of the fiber is very critical parameter while using fibers as reinforcement. Various authors classified fibers into short fibers and long fibers according to length of the fiber. According to most of the authors, fibers having length from 3 to12.5 mm are considered to be short fibers.[33–35] At the same time, some authors used aspect ratio (length to diameter ratio) as a criteria to distinguish short fibers from long fibers. Though, aspect ratio seems to be a good parameter but it is not suitable parameter in case of nanotubes as those are considered as fillers. Shokrieh et al.[36] used multiwall nanotube with aspect ratio of the order of 1000 and tagged them as filler while Qi et al.[37] used fibers with aspect ratio of 15 and named them short fibers when the length and diameter are in micrometers. As the length of fiber decreases, the load taking capability of fibers decreases that is, there must be some minimum length of the fiber known as critical fiber length (L_c) which must be ensured to get the reinforcement effect out of the fibers used as reinforcement. Critical length of the fiber and corresponding critical aspect ratio can be used to distinguish short and long fibers. Incorporation of short fibers having aspect ratio lower than critical aspect ratio will not be able to produce the reinforcing effect in the composite. According to

bailie,[38] Critical length of the fiber/yarn can be found by using the following relation,

$$L_c = \frac{\sigma_f d}{2\tau_m} \quad (1)$$

Where σ_f is the ultimate tensile strength of fiber/yarn, τ_m is maximum shear strength of matrix, and d is the diameter of the fiber/yarn.

Particulate fillers were initially used to reduce the cost of low performance products by filling the compound but the modern market demands such fillers which can also improve some useful properties. Particulates are the particles of random sizes and shapes virtually having no length. Organic fillers like wood flour, nutshells, corncobs, rice hulls, and other low cost fillers like silica and talc are being used in plastics as extenders (increases the bulk of material). Metallic and synthetic fiber fillers (carbon fiber, graphite fiber, glass fiber) can be used to increase electrical-thermal conductivity, magnetic properties, and mechanical properties of the base polymer compound depending on the producers. Feldspar is another kind of filler used to improve the transparency of the developed products.[39] Gao et al.[40] used hybrid nano fillers to improve thermal conductivity of the epoxy based composites. Authors concluded that addition of 5 wt% of aluminium nitride (AlN)/graphene (GE) nano filler and 45 wt% of aluminium oxide micro filler can improve the thermal conductivity of the epoxy polymer by establishing a heat conducting chain/network. Pan et al.[41] used another combination of fillers (hexagonal boron nitride (hBN) platelets and aluminium nitride particles) to improve through-plane thermal conductivity of the PTFE (polytetrafluroethylene) composites. Optimized ratio of hybrid fillers for the highest thermal conductivity was found to be 1:2 (AlN to hBN) by volume. Mirmohseni and zavareh[42] studied the effect of nano-fillers on the mechanical properties of the epoxy composites and concluded that ABS, clay and TiO_2 nano-fillers improved impact strength and tensile strength by 168% and 64%, respectively. Katiyar et al.[43] used graphite and talc powder of 20 and 10 μm size, respectively, for the coating of epoxy based composites and investigated the wear behavior of coated composite. Addition of 15 wt % of graphite and 15 wt% of talc filler increased wear life by significantly reducing the specific wear rate. Authors concluded that if wt% of the fillers increases a certain limit, the bonding ability of filler with parent polymer decreases due to reduction in the available surface area for bonding with polymer. Some authors[44–46] have used natural bio-fillers (wood apple shell, coconut shell, rice husk, and sugar palm fillers) to improve mechanical properties of the developed polymer composites. Grit-O'cobs is another environment friendly and biodegradable filler derived from corn cob which is very hard and water absorbent in nature. Size and geometry of the particulate fillers is also an influencing factor affecting mechanical properties of the developed composites.[47] Surface modification of fillers can be used to improve the interfacial bonding of particulates with polymer. Coating of the particulates could also help in increasing their moisture resistance.[39]

3 | CONTINUOUS FIBER REINFORCEMENT

Continuous fibers offers great mechanical properties in the longitudinal direction. Thanks to the ultimate flexible nature of the fibers, a number of geometrical arrangements are possible in order to achieve the required architecture. Mechanical properties of developed composite material are highly influenced by the geometrical architecture of reinforcement as shown in Table 2. The unidirectional tapes are made up of continuous fibers, which offers great in-plane tensile properties in the longitudinal direction of fibers. At the same time, tape laminates show lower out-of-plane strength and stiffness compared to fabric reinforced textile composites.[48] Continuous fibers made textile preforms are consciously being used in aerospace industries to increase the damage resistance of the components.[49] Continuous fibers can be arranged in unidirectional, 2D woven, knitted, braided, and 3D woven textile architectures giving flexibility of applications.

3.1 | Uni-directional fibers

Uni-directional tapes are being used to produce a wide range of aerospace application products due to their ultimate strength in longitudinal direction. Simplicity of analysis of uni-directional FRPCs at both micro- and macrolevel helps in obtaining the desired properties. Stacking angle of unidirectional tapes control the mechanical properties of developed composites.[2,61] Longitudinal tensile properties of uni-directional FRPCs is better than other weaved FRPCs but shear strength of woven FRPCs dominates over uni-directional FRPCs. Tensile strength of uni-directional carbon fiber reinforced epoxy composite was found to be 2.4 times higher than that of plain woven carbon fiber reinforced epoxy composite while shear strength was 9 times lower in longitudinal direction.[62] At the same time, transverse tensile strength of Uni-directional FRPCs is very low in comparison to woven FRPCs.[63] Tendency of delamination

TABLE 2 Mechanical properties of commonly used architecturally different continuous fiber reinforced polymer composites

S. No.	Fiber/matrix	Reinforcement architecture		Tensile strength (MPa)	Tensile modulus (GPa)	Flexural strength (MPa)	Flexural modulus (GPa)	Refs.
1	Kenaf/poly lactic Acid	Random oriented		28.90	1.56	50.12	1.15	[50]
		Unidirectional		50.26	2.93	54.17	2.01	
		Bidirectional		30.25	3.41	71.42	2.85	
2	Flax/epoxy	Plain woven 1 × 1		64.28	5.95			[9]
		Plain woven 1 × 4		74.67	6.66			
		Plain woven 1 × 6		79.73	6.85			
		Plain woven 1 × 10		80.34	6.52			
3	Carbon/polyetherimide	Plain woven		535	73	589	40	[16]
		Twill weave		888	106	951	54	
		4-H satin weave		575	76	832	46	
4	Banana fiber/epoxy	Plain weave		23.25	0.42	60	6.13	[51]
		Twill weave		17.18	0.38	38.14	4.25	
		Basket weave		13.04	0.24	43.93	3.37	
5	PP*/PPE#	Plain weave			1.95		0.82	[52]
		Basket weave			2.21		1.09	
		Twill weave			2.45		1.25	
		Satin weave			3.06		1.58	
6	Jute/Polyester	Plain weave		30.42	1.18	33.77	1.39	[53]
		Basket weave		43.60	1.02	69.54	3.43	
		Herringbone		37.00	1.27	64.2	3.18	
		Intra-ply		44.32	1.31	72.8	3.12	
7	Carbon/epoxy	Orthogonal interlocked	W-1	1358.5	76.75	886.48	48.49	[23]
			W-2.1	1227.9	69.44	960.12	62.22	
			W-2.2	1103.1	78.67	805.26	55.04	
			W-2.3	1220.3	73.46	826.52	58.57	
		Angular interlocked	W-3	1276.2	80.49	1036.62	63.32	
			W-4	409.9	48.54	379.89	30.37	
8	Carbon/epoxy	5-H satin weave		686	63	85.3		[17]
9	Carbon/epoxy	2D woven						[54]
		3D orthogonal woven						
10	Carbon/vinyl ester resin	3D woven			72.6	44.3	1.0*	[55]
		2 × 2 twill			65.0	26.8	1.2*	
		Noncrimp (0)			54.5	30.4	1.4*	
		Noncrimp (90)			55.0	32.4	1.7*	
11	Carbon/epoxy	3D orthogonal woven			67.3			[56]
12	E-glass/vinyl ester resin	3D plain orthogonal weave		458-warp 502-weft	27.7 34.6			[57]
		3D normal layered interlock		496- warp 452- weft	29.2 27.2			
		3D offset layered interlock weave		531- warp 407- weft	30.8 25.8			
13	Glass/epoxy	45° braided		1000	20.2	428.2	12.6	[27]

TABLE 2 (Continued)

S. No.	Fiber/matrix	Reinforcement architecture		Tensile strength (MPa)	Tensile modulus (GPa)	Flexural strength (MPa)	Flexural modulus (GPa)	Refs.
		50° braided		850	13.3	299	6.8	
		65° braided		800	10.3	258	5.6	
14	Glass/epoxy	Weft knitted	Wale	62.83	5.35			[58]
			Course	35.5	4.37			
15	Aramid/polyester	Warp knitted	Wale	50.2	3.44			[59]
			Course	80.4	3.93			
16	Glass/polypropylene	Multilayered knitted		254	14.1			[60]
		Noncrimp		307	17.9			
		Plain woven		299	14.2			

*"lowest value of flexural modulus is selected from the reference"

propagation is higher in case of uni-directional FRPCs than woven FRPCs.[30] Handling of uni-directional fiber also pose problems in alignment during the fabrication process if prepregs are not used. Even uni-directional prepregs are not able to conform the complex shape due to poor drapability.[16] Although, uni-directional fiber reinforcement is a fabulous choice for some applications but their lower properties in transverse direction, poor shear strength, and lower drapability led to the use of continuous fibers in woven form.

3.2 | Woven architectures

3.2.1 | 2D woven

Weaving is the most common and easiest method of changing the geometrical architecture of reinforcement. Along with good mechanical properties, good drapability of the reinforcement must be ensured while selecting a woven architecture of reinforcement. Drapability of weaved fabric depends on the number of interlacing points per unit area of fabric. Drapability of the fabric can be increased by decreasing the number of interlacing points but on the other hand, structural stability of fabric reduces. As the interlacing points of a fabric increases, the crimp percentage also increases. Hence, lower crimp percentage must be ensured to achieve higher tensile strength of developed composites.[64] Bijwe et al.[16] studied the effect of three weave patterns namely, plain weave, twill weave, and 4-Harness satin weave on tribological properties of the carbon fiber reinforced PEI composites. Highest strength and modulus properties were obtained by twill weave fabric reinforced composites followed by satin and plain weave fabric reinforced composites in both tensile and flexural loading conditions. Plain weave fabric was found to be more stable architecture compared to twill and satin weave fabric due to higher interlacing points per unit area but the drapability of plain weave was found to be very poor compared to twill and satin weave fabric Yang et al.[65] developed mesoscale models of four different weave patterns namely, plain weave, basket weave, 4-Harness satin weave, and 2×2 twill weave using aramid yarns. Authors investigated the ballistic resistance of single layer and multi-layer fabrics. Authors concluded that interlacing points or cross-over points are most influencing factor controlling the ballistic response of the fabric. The plain woven fabric showed higher capacity to absorb energy during ballistic test but the response to impact load was slower than other three fabrics. Stress transfer was faster in case of satin weave fabric due to less number of cross-over points but the ballistic performance of satin weave fabric was lower than other fabrics due to easier displacement of yarns away from the direction of ballistic travel leading to easy penetration of projectile. The thickness of same number of layers was found to be higher for satin weave compared to plain weave.[66] Dynamic mechanical properties were also found to affected by the weave architecture. Rajesh and Pichaimani[67] studied dynamic mechanical behavior of jute/banana fiber reinforced in various architectures namely, plain, basket, twill, satin, huckaback and concluded that basket weave offers better dynamic properties in comparison to other architectures.

Erol et al.[11] investigated the mechanical response of four different weave architectures as shown in Figure 2 under uniaxial tensile and shear loading conditions. Authors concluded that the modulus of woven fabric can be significantly improved by choosing high modulus yarns while the shear resistance of the fabric can be

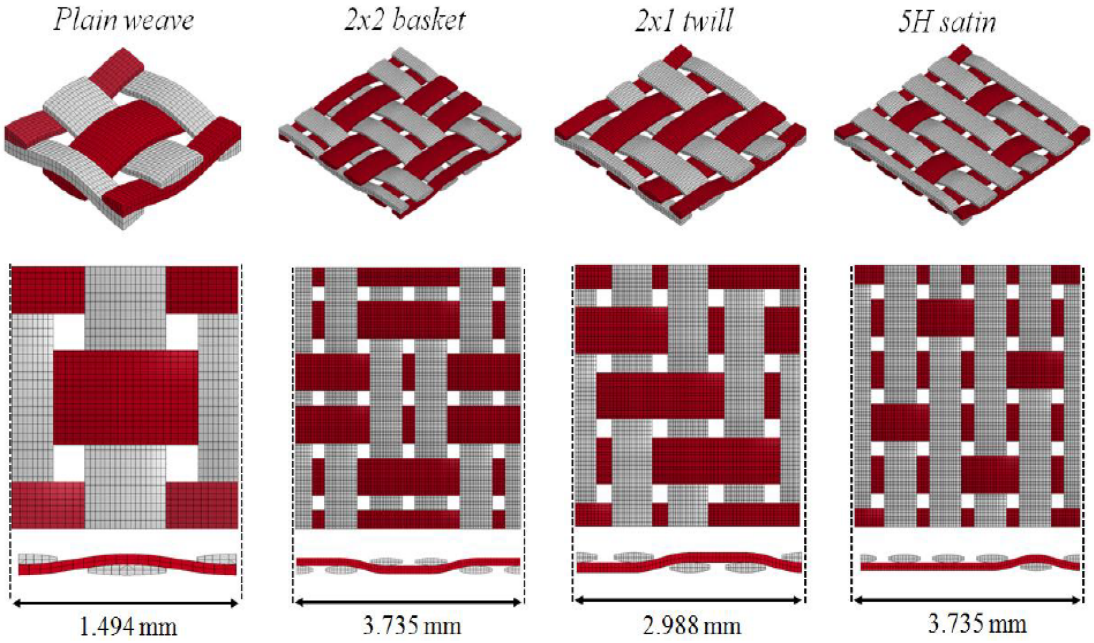


FIGURE 2 Custom MATLAB code generated mesoscale models of different 2D weave architectures (reprinted with permission from ref. [11]) [Color figure can be viewed at wileyonlinelibrary.com]

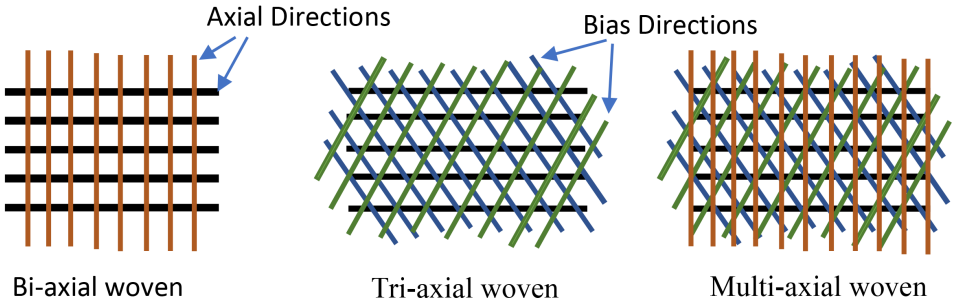


FIGURE 3 Axial and bias directions in 2D woven fabrics [Color figure can be viewed at wileyonlinelibrary.com]

altered more easily by changing the fabric architecture. Due to lower mechanical properties in the bias direction of bi-directional woven fabrics, the fibers are added in the bias direction as shown in Figure 3 to make fabric stiffer in bias direction. The higher isotropic behavior of tri-axial fabric may be obtained compared to bi-axial fabrics for both tensile and shear loading by adjusting the angle of bias tows. The most common angle of the bias tows for optimum results, as reported by various authors is $\pm 60^{\circ}$.[21,68,69] Fujita et al.[70] claimed that the fracture behavior of tri-axial fabric depend on the cross-over points and the angle of bias tows. Authors also showed superior properties of the tri-axial fabric reinforced composites compared to bi-axial plain weave fabric reinforced composites. Impact properties of natural fiber composites can be substantially increased by precisely controlling the architecture of the reinforcement but some authors[9] have also claimed that the organization of the fibers is less effective than the density of reinforcement to influence the toughness of developed composites. Architecture of layers within a laminate having same textile fabric reinforcement also influence the mechanical properties of developed composites. Axinte et al.[71] showed that increased 45° oriented layers improved the shear stiffness of the fibrous polymer composites.

3.2.2 | 3D woven

Introduction of the reinforcement in the thickness direction (known as "z" direction) of plain woven fabric can improve the impact strength and delamination resistance of 2D woven fabrics. These "z" direction fibers can be introduced by either through thickness or layer by layer methodology. Nawab et al.[72] experimentally showed that tensile response of layer by layer 3D textile preform was better than through thickness 3D preform as reinforcement. 3D woven textile reinforced composite materials possess higher structural rigidity with higher delamination resistance than layered laminates.[23] Dai

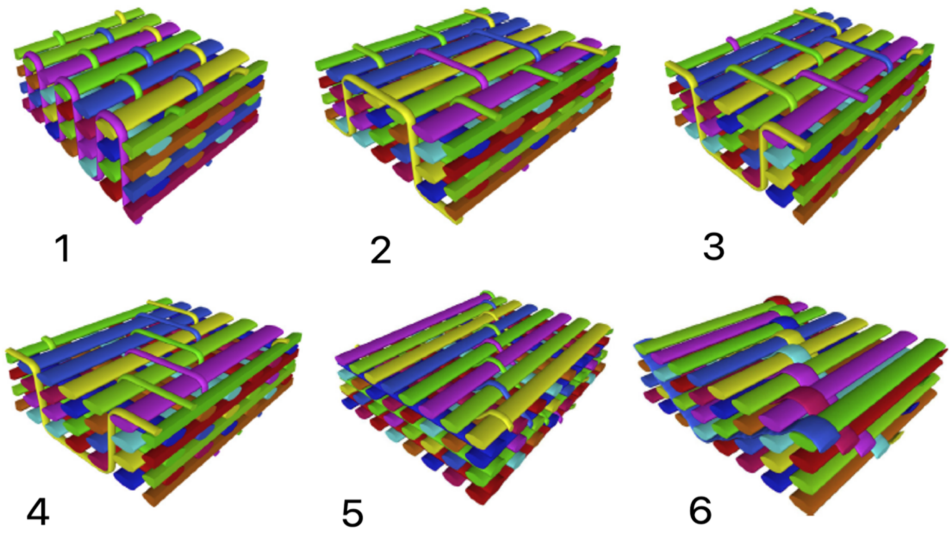


FIGURE 4 Orthogonal and angular interlocked 3D weave architectures used as reinforcement (reprinted with permission from ref. [23]) [Color figure can be viewed at wileyonlinelibrary.com]

et al.[23] developed six types of 3D woven preforms using orthogonal and angular interlock system as shown in Figure 4. Tensile modulus of the composite laminates decreased as the waviness of fiber tows increased. Waviness of fiber tows in all the developed preforms were compared with each other and orthogonal interlock preform was found to have maximum amount of waviness. Tensile modulus of the angular interlocked preform reinforced composite was found to be higher than that of orthogonally interlocked preform reinforced composites. The length of delamination crack can be effectively reduced by increasing the binding points of preform but at the same time crimp percentage of fabric also increases. Mishra[73] has found orthogonal interlocked perform better than angular interlock preform in terms of drapability. So, it is advisable to use angular interlocked preforms as reinforcement in highly complex shape products due to their higher drapability than orthogonal interlocked. Guenon et al.[54] compared the toughness of orthogonal interlocked 3D preforms reinforced composites with 2D laminates. The presence of the fibers in the through-thickness direction results in significant improvement of the interlaminar fracture toughness. Crack arresting, crack branching, and crack deviation phenomenon were present in 3D textile preform composites which allows the composite to dissipate more energy at the time of impact. Significant influence of the crimp was found on the longitudinal stiffness of the composites.[25] Although, in-plane tensile strength of 2D woven reinforced composite is greater than 3D textile reinforced composites for the same fiber loading but the failure strain of 3D textile reinforced composite is much greater than that of 2D woven reinforced composites. The multidirectional fiber in the 3D fabric improves the overall ductility of the composites. Charpy impact testing revealed that the impact strength of 3D textile reinforced composite is significantly higher than 2D fabric reinforced composites. The fibers in the third axis also known as *z*-binders are supposed to arrest the crack propagating in the other two direction.[74] Increasing the volume fraction of the *z*-binders increased the Young's modulus of the composite materials but fatigue life and strength was surprisingly decreased.[75] Complex near net shape products with higher delamination resistance can be easily manufactured by using 3D textile preforms as reinforcement in FRPCs. lower crimp percentage in preform must be ensured in order to achieve better mechanical properties. Delamination resistance, mechanical strength, and drapability of preform are highly influenced by number of binding points. Lower binding points results in lower crimp of fibers which in turn improves mechanical strength and drapability of preform but at the same time delamination resistance of preform decreases sharply.

3.3 | Knitted architectures

3.3.1 | 2D knitted

Compared to the interlacing technique of woven fabrics, inter-looping technique, required to fabricate knitted fabrics, is more complex. Straight yarns are bent to make loops and then all the loops are interlooped with each other.[76] Interlooping of the loops can be done in either of the two directions namely, warp and weft. Different knitting machines are used to knit in warp and weft directions. Rows of loops in the longitudinal and transverse direction is called “wale” and “course”, respectively as shown in Figure 5.[77] Interlooping of the yarn during

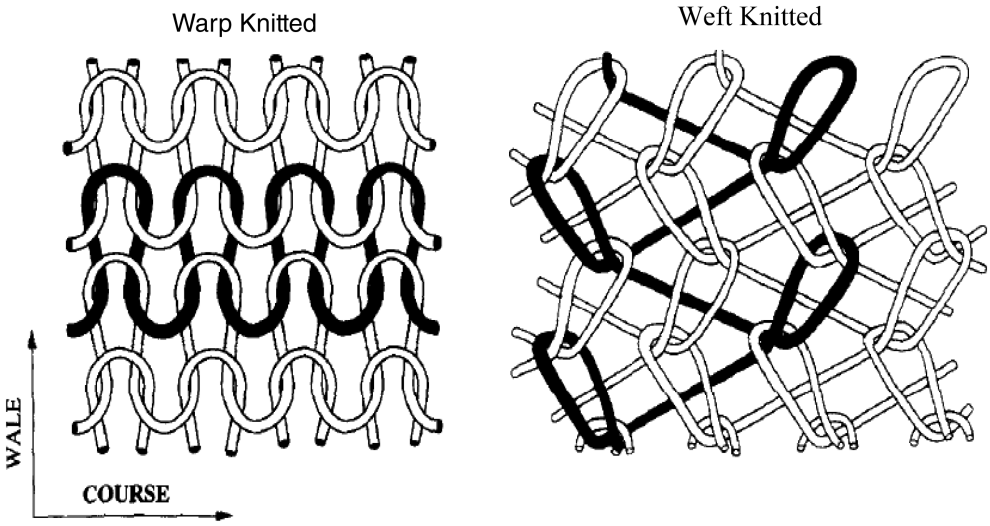


FIGURE 5 Knitted fabrics in warp and weft directions (reprinted with permission from ref. [77])

knitting process reduces the strength of fibers due to bending over a very small radius of the loop. Lau and Dias[78] studied the effect of knitting behavior of Kevlar and glass fibers. Authors found that the strength of the loop is influenced by the loop radii. Authors established that the loop strength increases with the increase in the loop radii. Brittle fibers like glass are more susceptible to lose their strength during the knitting process due to high degree of bending stresses developed during making loops.[79] In order to alleviate the problem of damages during the knitting process, more flexible yarns are recommended by researchers.[80]

Fibers exhibit greater strength when they are aligned in longitudinal direction which indicates that small radii of loops lowers the overall strength of the composites. Distribution of the strength in other directions due to loop formation and damages of fibers due to abrasion of fibers with knitting machine elements during the knitting process are two important reasons behind the reduced longitudinal strength of the knitted fiber reinforced composites.[81] The highly looped architecture of knitted reinforcement allows the fabric to deform easily and significantly on the application of external load. High deformability of the fabric makes them applicable to fabricate highly intricate and near net shape products.[80] In-plane properties of the 2D knitted fabric reinforced polymer composites are usually poor and anisotropic in nature.[80,81] Ashraf et al[81] compared the effect of knitted and woven fabric reinforcement while manufacturing Kevlar and glass fiber reinforced epoxy laminates. Authors found that the tensile strength of the laminates decreased as the percentage of knitted fabric reinforcement increased compared to the woven fabric. Authors concluded that the crimping of the yarns and looping were the major factors responsible for reduced strength of the knitted fabric reinforced composites. Authors also investigated the impact characteristics of the knitted fabric reinforced composites. Authors concluded that due to the flexibility of knitted fabric, the time of contact between the impactor and the composite increased. As a result of increased contact time, more energy was able to get transferred from impactor to material. Dynamic mechanical analysis was also performed by the same authors. Storage modulus of the knitted fabric reinforced composites was found be lower than woven fabric reinforced composites for same volume fraction of reinforcement due to the higher flexibility of knitted fabric than woven fabric. Chou et al[29] found that the impact strength of weft-knitted fabric reinforced composites is three times greater than that of plain woven fabric reinforced composites. Aktas et al[28] studied the effect of stacking sequence of woven and knitted fabric glass fiber reinforced epoxy composites. Authors suggested that for high rigidity of composites, the knitted fabric layers must be used in between the woven fabric layers. Authors concluded that the impact strength of composite laminates also depend on the stacking sequence of woven and knitted fabric reinforcement. Ramakrisha et al.[58] characterized the tensile properties of weft knitted glass fabric reinforced epoxy composites. Authors concluded that tensile properties of knitted fabric reinforcement can be increased by increasing the reinforcement content in the composite.

The fiber content of a knitted fabric composite can be increased without increasing the number of layers of reinforcement by increasing the knitting density. Authors also concluded that the tensile and flexural properties, in general, are better in the wale direction than course direction. The results of Leong et al.[80] also supports the above statement. Though, knitted fabric reinforced composites are economically inefficient but they can be used to produce complex shape products

due to their high drapability. Dexter et al.[82] produced four types of knitted reinforcement to manufacture the composite laminates. On the basis of experimental investigations, authors suggested that the multi-axial knitting process is suitable to obtain knitted fabric with quasi isotropic properties. Quasi isotropic fabrics are those fabrics, which contains approximately equal fiber content in all the directions. Chou et al.[29] fabricated four types of knitted structures namely, interlock knitted, full cardigan, Milano and rib knitted fabric by weft knitting process. Authors concluded that interlock fabric reinforced composites provides the highest fatigue resistance among all the four types of fabric reinforced composites but still lower than plain woven fabric reinforced composites.

De Carvalho et al[83] compared the mechanical properties of plain weave and weft knitted jute fabric reinforced polymer composites. The tensile properties of the plain woven fabric reinforced composites were found to be greater than that of knitted fabric reinforced composites irrespective of the fiber volume fraction. The behavior of plain weave fabric is directional dependent due to the orthotropy involved in the woven structure while knitted fabrics are more of isotropic nature. The impact strength of plain woven fabric reinforced composites was found to be lower than that of knitted fabric reinforced composites. The cracks formed in the woven fabric reinforced composites easily follow the direction of fibers to a long distance. The cracks developed in woven fabric reinforced composites does not detour which finally leads to the failure of component due to high crack propagation. In case of knitted fabric reinforced composites, highly looped structure of the fabric effectively arrests the crack propagation.

3.3.2 | 3D knitted

Guo et al.[84] classified the 3 D knitted fabrics into three categories namely, multi-axial warp knitted fabrics, space fabrics, and 3D knitted fabrics. Multi-axial knitted fabrics are those fabrics in which the yarns are placed at 0°, ± theta, and 90° where 0° represents the production direction. Two connected fabrics keeping some space between them is known as space fabrics. 3D knitted fabrics are more commonly known as near net shape fabrics. 3D knitted fabrics are being used in various engineering and medical application like jet engine vanes and medical prosthesis. The inferiority of the 2D knitted fabrics due to lower in-plane properties can be removed by adding one more knitting yarn in the third direction. A number of modifications and innovations like float stitches, split warp-knits, multiaxial multilayer warp-knits, fully fashioned preforms etc. have been reported by various authors to improve the out-of-plane strength of the knitted fabric reinforced composites.[80] Hufenbach et al.[60] compared the mechanical properties of glass fiber reinforced polypropylene composites. The reinforcement was used in three architectures namely, woven fabrics, noncrimp fabrics, and multi layered (3D) knitted fabrics. The stiffness of the multi layered 3D knitted fabric reinforced composite was found to be lower than that of other two types of composites due to high crimping and undulations of the fibers when tested in the longitudinal direction. Knitted fabric reinforced composites showed better tensile strength than other two composites when all the composites were tested in the transverse direction. Impact test of the developed composites revealed that the energy absorption capacity of the multi layered 3D knitted fabric reinforced composites was much higher than that of other composites. In comparison to the woven fabric reinforced composites, the energy absorption of 3D knitted fabric reinforced composite was increased by 35% at room temperature. Authors identified the special stitching system of multi layered 3D knitted fabric responsible for the improved energy absorption capacity of the concerned composites. Instead of lower in-plane mechanical strength, exceptional drapability due to low resistance to deformation of the knitted fabrics make them suitable for various complex shape products with moderate strength. High energy absorption capacity of highly looped knitted fabric increases the fracture toughness of developed composites. Knitted fabric is suitable for rapid production of complex shape products.

3.4 | Braided architectures

Braided fabrics are being used in various possible structural applications in aircraft industry like airlock developed by NASA.[85] Braiding is done by intertwinning two set of yarns onto each other in an alternate fashion at a biased angle ranging from 10 to 85.[85] For practical applications, this biased angle ranges from 45 to 75 in order to avoid the problem of locking of yarns.[86] Potluri and Nawaz[85] classified the braiding technique into two broad categories namely, 2D braiding and 3D braiding. Owing to the high degree of automation involved at the fabrication level of braided fabrics, industries are eagerly adopting and promoting the braided fabric reinforced composites.

3.4.1 | 2D braided

2D braided reinforcement can be bi-axial or tri-axial depending on the number of set of yarns used for braiding. Nearly isotropic properties of the composites can be

obtained with the help of tri-axial braids. Braiding can be used to make holes in composites without losing the fiber continuity.[49] The thickness of 2D braided composites can be increased by over-braiding the layers on top each other. Braid architectures, braiding angle, yarn size, yarn spacing, and crimp percentage of yarns are some parameters associated with braided fabrics which govern the properties of braided fabric composites. Longitudinal tensile modulus of 2D braided fabric reinforced composites was found to be approximately equal to that of uni-directional fiber reinforced composite laminates but the transverse modulus of 2D braided fabric reinforced composites was significantly lower than that of UD laminates.[86] Crimping of the yarns while braiding was found to be responsible for the reduced modulus. Falzon et al.[87] also found the tow waviness and fiber undulations of 2D braided fabric were responsible factors for the reduced strength of composites compared to the UD laminates. Braiding angle was also found to be an important parameter of the braided fabric controlling the in-plane properties of braided fabric reinforced composites. The possible range of braiding angle was found to be from around 20° to 60° in practice.[88] Figure 6 represents the effect of braiding angle on the braid architecture. In 2D braided fabrics, volume fraction of the fiber can increased without increasing the layer thickness by decreasing the yarn spacing. Yarn spacing was found to be minimum at 45° braid angle. Yarn spacing could be increased by increasing or decreasing the braid angle up to the locking angle. Locking angle is the angle at which the spacing between the fiber rovings becomes zero.[89] Tensile strength of the braided fabric reinforced composites decreases with increment in the braiding angle. Fiber undulations results in the development of stress concentration points which leads to the reduced compressive strength of the developed composites. Materials with high Poisson's ratio should be avoided in order to reduce the chances of failure in transverse direction. Dadkhah et al.[90] found that the Poisson's ratio for the tri-axial braids was very high when the loading was tensile and in axial tow direction. If the volume fraction of the axial tow increased, the tensile strength of the composite in axial direction was found to be increased. Fibers present in the axial tow bear the axial load while the fiber of braider tow sustain the load during bending loading.[91] Fiber distribution is more homogeneous in case of braided preforms compared to woven reinforcement. Falzon and Herszberg[87] three types of 2D tri-axial braided composites were developed using 30, 45, and 60 braiding angles. Longitudinal Young's modulus was higher for the composite with 30 braided fabric than that of other composites but the Poisson's ratio was also high which is not desirable. Lower braid angle results in higher longitudinal strength but lower transverse strength. Therefore, a compromise must be made on the basis of required strength of the component. Braiding process results in the reduced strength of fiber tows due to the physical damage of fibers during the braiding process. When compared to uni-directional tapes, the braided composites showed up to 30% lower tensile strength in axial direction. Mechanical properties of 2D tri-axially braided composites are amenable to braided fabric related parameters like braid angle and fiber volume content of axial tow.[92] Effect of yarn size on the mechanical properties of braided fabric reinforced composites was found to be virtually negligible. Braid angle strongly influences the mechanical properties of composite laminates. Increasing the braid angle from 45 to

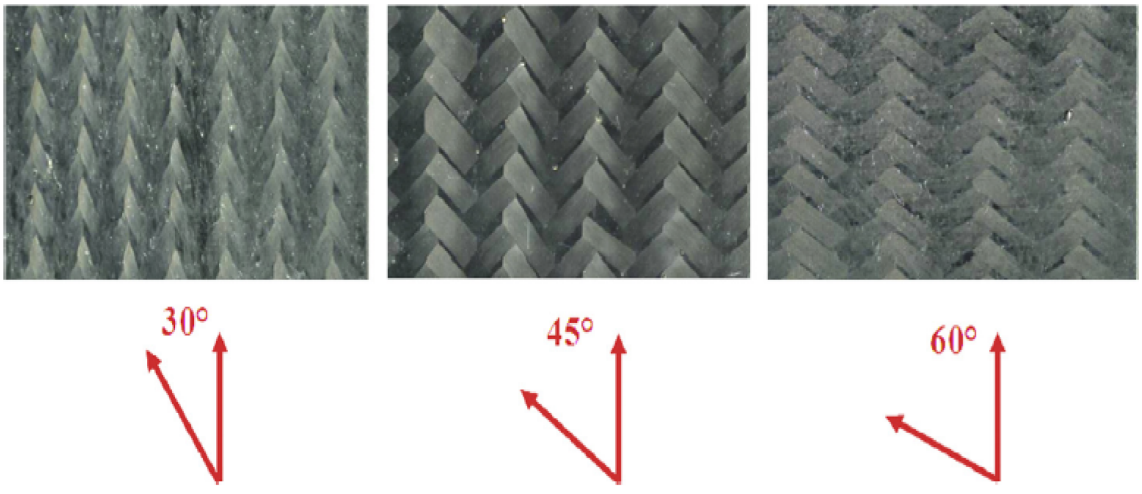


FIGURE 6 Tri-axial braided preforms with tows at different angles (reprinted with permission from ref. [26]) [Color figure can be viewed at wileyonlinelibrary.com]

75 resulted in decrement of longitudinal modulus by 4% but the transverse modulus was increased by 144% and the shear modulus was decreased by 30%.

3.4.2 | 3D braided

Poor out-of-plane properties of laminated composites results in interlaminar failure of the component. This difficulty of the laminated structures can be addressed by reinforcing the laminate in the out-of-plane direction. Unidirectional prepregs and 3D braided preforms were used by some authors as reinforcement separately to fabricate and compare the properties of the developed composites.[93] Additional directional yarns can also be added in 3D fabrics according to the required directional strength of the composites. At least three directional tows must be there to make 3D braided fabrics. The number of tows can be increase to make 4-directional, 5-directional, and 6-directional 3D braided fabrics. Yuanyuan et al.[94] fabricated the 3-dimensional 5-directional (3D5d) and 3-dimentional 6-directional (3D6d) braided fabric reinforced epoxy based composites as shown in Figure 7. Authors concluded that the tensile strength and modulus in axial direction of composite depend on the modulus and volume fraction of the fiber placed in axial direction. Similar to the 2D braided fabrics, the braiding angle controls the mechanical properties of the 3D braided fabrics along with some other parameters like pitch length and yarn tension.[95,96] Pitch length must be consciously selected as the larger pitch length tends to decrease the fiber volume fraction of a yarn which in turn reduces the tensile strength of composite while the shorter pitch length increases the braiding angle which also reduces the tensile strength in axial direction. The tensile stiffness of the braided yarn was found to increase with increase in the pitch length due to lower crimping of the yarn for higher pitch length but at the time, stiffness in other directions may reduce significantly due to absence of fibers in those directions. Wan et al.[97] developed 3-dimensional 4-directional (3D4d) braided fabric reinforced epoxy polymer composites using three types of reinforcement namely, carbon fiber, Kevlar fiber, and carbon-Kevlar hybrid fibers for the use of bone fixation. The effect of hybridization was found to be positive in case of impact strength due to the introduction of ductile Kevlar fiber into the carbon fiber yarn. Superior crashworthiness properties of a product can be achieved using braided preforms in FRPCs. Braided preforms are less sensitive to the notches hence reduces the crack propagation probability. In-plane strength and stiffness of braided preform reinforced composites is generally lower than 2D laminates and also the automation cost is very high due to highly complex machinery to fabricate the 3D braided preforms.

4 | NEEDLE PUNCHED MATS

Needle punched mats are nonwoven layers bonded mechanically in thickness direction used as reinforcement in the FRPCs. These mats are different from randomly oriented fiber mats in terms of high entanglement of fibers with each other in through thickness direction of preform. Though, nonwoven preforms are easy to manufacture but they show relatively lower interlaminar strength while needle punched preforms, being nonwoven, show higher interlaminar strength. Punching needles orient the fibers in the direction of thickness with the help of needle barbs

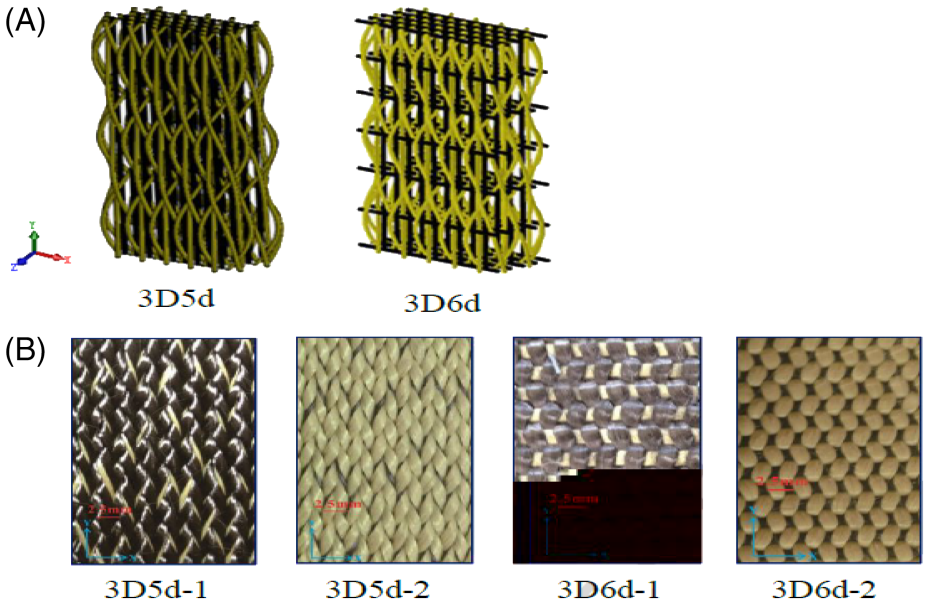


FIGURE 7 A, Multi-directional 3D braided fabrics, B, local surface texture of hybrid preforms (reprinted with permission from ref. [94]) [Color figure can be viewed at wileyonlinelibrary.com]

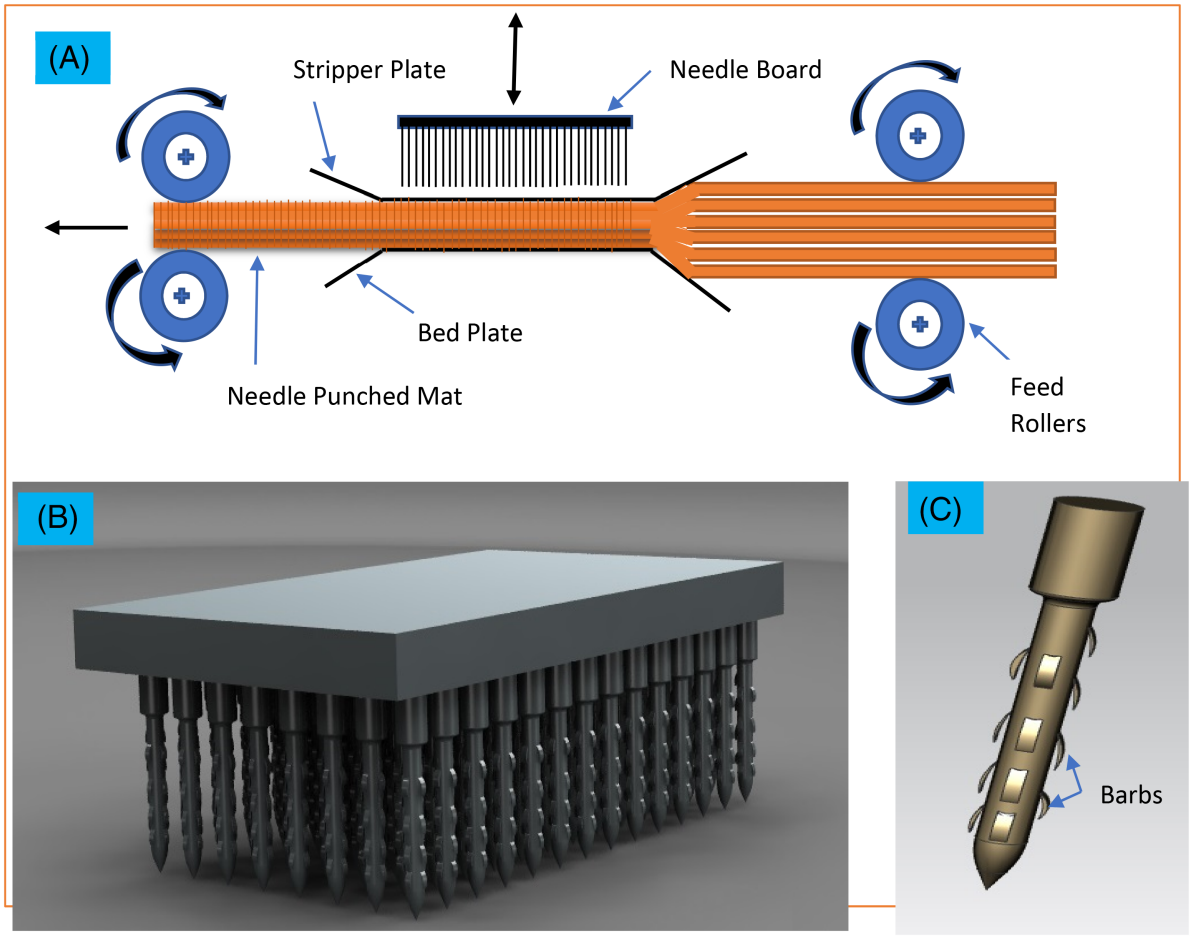


FIGURE 8 A, Schematic of needle punching process, B, enlarged 3D view of needle board, C, enlarged view of single needle [Color figure can be viewed at wileyonlinelibrary.com]

which improves the interlaminar strength of the needle punched nonwoven preform.[98] Barbs are designed in such a way that they leave the fiber at the intended place while leaving the out of the preform as shown in Figure 8. Diameter of working part of punching needle is generally of the order of 0.5 mm.

Automobile, aerospace, and civil engineers are using needle punched preforms in FRPCs due to their ease of machining property.[99] Needle punching process decreases the mean fiber length which in turn would definitely affect the mechanical properties of developed composites. Mean fiber length and distribution of fibers majorly depend on the punching density, that is, the penetrations per unit area. Kang and lee[100] studied the effect of punching density on mean fiber length and mechanical properties concluding that mean fiber length of E-glass fiber web was decreased by 30% when the punching density increased to 90 penetrations/cm^2. Tensile strength also decreased with significantly at high punching density. As the punching density increases, the fiber entanglement increases but at the same time fiber breakage also increases. Thus, brittle fibers result in reduction of tensile strength with increment in punching density due to high breakage of fibers during needle punching. Lee and kang[101] showed that interlaminar fracture toughness of nonwoven preform improved using needle punching process but at the same time other mechanical and wear properties decreased due to higher fiber breakage. Authors suggested the use of ductile fibers to avoid fiber breakage. Punching depth, which is the distance between needle tip and lower surface of lower layer when the needle has fully penetrated into the preforms is also vital factor controlling the intensity of entanglement.[102] Composite materials reinforced with highly needle punched nonwoven mat showed more isotropic properties than simple nonwoven mat reinforced composites. Homogeneity in fiber distribution throughout the thickness of preform can be achieved using 3-D needle punching process.[103] Authors also concluded that the mechanical properties were decreased in case of both 3-D and 2-D needle punching processes. Although, needle punched preforms are inferior to most of their

counterparts in terms of tensile properties, they offers some freedom of material selection to suit an application with more isotropic properties along with high interlaminar toughness. Some authors have shown the improvement in the toughness and bonding strength of the fiber/matrix interface of needle punched preform reinforced composites by adding some fillers.[104,105]

5 | RECOMMENDATION TABLE

Selection of a reinforcement architecture for a particular application is a tedious task as a number of technical architectures are available. One architecture cannot fulfil the demand of all applications such as flexural strength of twill weave architecture is good but the drapability is quite low compared to other available architectures. Table 3 is a recommendation table for the selection of reinforcement architecture depending on various parameters like strength, drapability, and so on.

6 | FUTURE SCOPE AND CHALLENGES

Reinforced polymer composites are excellent materials which have demonstrated a great potential for crucial lightweight applications. Although a lot of research have been conducted in the field of polymer composite area, there are still various vital technical challenges which needs to addressed. Primary challenge is the accurate performance prediction of developed composites addressing every inhomogeneity involved the architecture of the reinforcement. Fibers/yarns are generally assumed to be straight while performing simulation of reinforced polymer composites by most of the commercially available finite element analysis codes while practically fibers/yarns are constantly changing their direction in an architecture. So, a study is needed to formulate a mathematical model to consider the effect of interlacing and interloping points on the weaved and knitted architectures, respectively. Thickness of a yarn also affects every mechanical and aesthetic aspect of architecture which needs to addressed experimentally. The effect of braiding yarn diameter needs to addressed quantitatively as only the qualitative effects are available in the literature. The factors affecting the drapability of the available architectures and the surface finish of the composite structures needs to be find out. An important issue to be addressed is the absence of analytical techniques to predict and distinguish the performance of available architectures.

TABLE 3 Hands-on recommendation table for the selection of reinforcement architectures

Reinforcement	In-plane strength	Out-of-plane strength	Flexural Strength	Drapability	In-shape product formability	Isotropicity	Surface finish	Ease of manufacturing	Delamination resistance	Crack propagation resistance
Noncontinuous	**1**	4	**1**	**1**	6	**10**	**1**	9	6	4
Plain woven 2D	8	3	8	3	3	3	8	8	3	3
Twill woven 2D	**10**	**1**	**10**	4	4	2	9	7	2	2
Satin woven 2D	9	2	9	6	5	**1**	**10**	6	**1**	**1**
Woven 3D	5	**10**	7	2	8	4	6	3	**10**	6
Knitted 2D	6	5	3	8	**1**	5	4	4	5	7
Knitted 3D	3	8	5	**10**	9	9	2	**1**	8	9
Braided 2D	7	6	4	7	2	6	5	5	4	8
Braided 3D	4	9	6	9	**10**	7	3	2	9	**10**
Needle punched	2	7	2	5	7	8	7	**10**	7	5

Note: {Interpretation of numbers → **10**—highly recommended; **1**—lowest recommended}. The above table of recommendation is formed by collecting information from the available literature[9,16,23,27,56,57] but a complete comparison of all the architectures in a single article is not available. Hence, the above table is an approximate projection formed by the author by keeping various factor in mind which can affect the performance of architecture.

7 | CONCLUSIONS

The article has presented systematic review of various types of reinforcement architectures used in FRPCs, their associated features and their effect on the performance of developed composites. The review revealed that 2D plain woven architecture offers higher in-plane mechanical properties compared to 3D woven architecture. Among all the 2D woven, satin weave architecture is preferred for exceptional surface finish with better drapability which depends on the number of interlacing points per unit area. 3D preforms are preferred as reinforcement for development of complex near net shape products. These preforms also offer better delamination resistance in comparison to 2D reinforcement. Two other widely used fabrics as reinforcement are knitted fabric and braided preforms which offer better fracture toughness and less probability of crack propagation, respectively, which can help in designing materials for armoured vehicles. Size and geometry of filler influence mechanical properties of filler reinforced polymer composites significantly. In case of short fibers, critical length of fibers must be ensured to obtain reinforcement effect in the developed composite.

ORCID

Furkan Ahmad https://orcid.org/0000-0002-1211-4094

REFERENCES

[1] K. L. Pickering, M. G. A. Efendy, T. M. Le, *Compos. Part A Appl. Sci. Manuf.* **2016**, *83*, 98.
[2] F. Ahmad, P. K. Bajpai, *Def. Technol.* **2018**, *14*, 278.
[3] V. Chaudhary, P. K. Bajpai, S. Maheshwari, *J. Nat. Fibers* **2018**, *15*, 80.
[4] M. Sood, G. Dwivedi, *Egypt. J. Pet.* **2018**, *27*, 775.
[5] H. A. Sharifah, P. A. Martin, *Compos. Sci. Technol.* **2004**, *64*, 1219.
[6] P. K. Bajpai, F. Ahmad, V. Chaudhary, Processing and characterization of bio-composites. in *Handbook of Ecomaterials* (Eds: L. Martinez, O. Kharissova, B. Kharisov), Springer, Cham **2017**. https://doi.org/10.1007/978-3-319-48281-1_98-1.
[7] A. Memon, A. Nakai, *Energy Procedia* **2013**, *34*, 818.
[8] S. Misri, S. M. Sapuan, Z. Leman, M. R. Ishak, *Mater. Des.* **2015**, *65*, 953.
[9] Q. Liu, M. Hughes, *Compos. Part A Appl. Sci. Manuf.* **2008**, *39*, 1644.
[10] P. Tan, L. Tong, G. P. Steven, *Compos. Part A Appl. Sci. Manuf.* **1997**, *28*, 903.
[11] O. Erol, B. M. Powers, M. Keefe, *Compos. Part A Appl. Sci. Manuf.* **2017**, *101*, 554.
[12] J. Rout, M. Misra, S. S. Tripathy, S. K. Nayak, A. K. Mohanty, *Compos. Sci. Technol.* **2001**, *61*, 1303.
[13] M. S. Saharudin, A. Jumahat, A. Z. A. Kahar, S. Ahmad, *Appl. Mech. Mater.* **2013**, *393*, 88.
[14] S. Rahmanian, K. S. Thean, A. R. Suraya, M. A. Shazed, M. A. M. Salleh, H. M. Yusoff, *Mater. Des.* **2013**, *43*, 10.
[15] X. Gong, X. Chen, Y. Zhou, Advanced weaving technologies for high-performance fabrics. in *High Performance Textiles and Their Applications*, 1st ed. (Eds: J. Mcloughlin, T. Sabir), Woodhead Publishing, Royston Road, Duxford, CB22 4QH, United Kingdom **2017**, p. 75.
[16] J. Bijwe, R. Rattan, *Wear* **2007**, *263*, 984.
[17] Y. Liang, H. Wang, C. Soutis, T. Lowe, R. Cernik, *Polym. Test.* **2015**, *41*, 82.
[18] A. P. Mouritz, M. K. Bannister, P. J. Falzon, K. H. Leong, *Compos. Part A Appl. Sci. Manuf.* **1999**, *30*, 1445.
[19] M. N. Saleh, C. Soutis, Recent advancements in mechanical characterisation of 3D woven composites. in *Mech. Adv. Mater. Mod. Process*, Vol. 3 **2017**, p. 12. https://doi.org/10.1186/s40759-017-0027-z.
[20] R. S. Choudhry, K. A. Khan, S. Z. Khan, M. A. Khan, A. Hassan, *J. Compos. Mater.* **2016**, *51*, 705.
[21] R. Mishra, *Text. Res. J.* **2013**, *83*, 1836.
[22] J. Xu, S. V. Lomov, I. Verpoest, S. Daggumati, W. V. Paepegem, J. Degrieck, *Compos. Struct.* **2016**, *135*, 306.
[23] S. Dai, P. R. Cunningham, S. Marshall, C. Silva, *Compos. Part A Appl. Sci. Manuf.* **2015**, *69*, 195.
[24] F. Xu, L. Sun, L. Zhu, S. Yang, D. Hui, Y. Qiu, *Compos. Part B* **2018**, *133*, 193.
[25] F. Stig, S. Hallstrom, *Compos. Struct.* **2013**, *95*, 114.
[26] Z. T. Kier, A. Salvi, G. Theis, A. M. Waas, K. Shahwan, *Compos. Part B* **2015**, *68*, 288.
[27] B. Dauda, S. O. Oyadiji, P. Potluri, *Appl. Compos. Mater.* **2009**, *16*, 15.
[28] A. Aktas, M. Aktas, F. Turan, *Compos. Struct.* **2013**, *103*, 119.
[29] S. Chou, H. C. Chen, C. C. Lai, *Compos. Sci. Technol.* **1992**, *45*, 283.
[30] Y. Liang, H. Wang, X. Gu, *Polym. Test.* **2013**, *32*, 594.
[31] M. Y. Yuhazri, M. H. Amirhafizan, A. Abdullah, H. Sihombing, A. B. Saarah, O. M. Fadzol The Effect of Various Weave Designs on Mechanical Behavior of Lamina Intraply Composite Made from Kenaf Fiber Yarn *IOP Conference Series: Materials Science and Engineering* **2016**, *160*, 012021. http://dx.doi.org/10.1088/1757-899x/160/1/012021.
[32] A. R. Bunsell, *The Textile Institute Book Series*, 2nd ed. Great Abington, Cambridge, CB21 6AH, UK **2018**.
[33] Y. Elmogahzy, R. Farag, Tensile properties of cotton fibers: importance, research, and limitations. in *Handbook of Properties of Textile and Technical Fibres* (Ed: A. R. Bunsell), Woodhead Publishing, Duxford, United Kingdom **2018**, p. 223.
[34] M. Ericson, L. Berglund, *Compos. Sci. Technol.* **1992**, *43*, 269.
[35] M. C. Faudree, Y. Nishi, M. Gruskiewicz, *Mater. Trans.* **2014**, *55*, 1292.
[36] M. M. Shokrieh, A. Saeedi, M. Chitsazzadeh, *Mater. Des.* **2014**, *56*, 274.
[37] L. Qi, W. Tian, J. Zhou, *Compos. Struct.* **2015**, *131*, 843.
[38] S. Panthapulakkal, L. Raghunanan, M. Sain, B. KC, J. Tjong, in *Green Composites*, 2nd ed. (Eds: C. Baillie, R. Jayasinghe), Woodhead Publishing, Great Abington, Cambridge, CB21 6AH, UK **2017**, p. 39.
[39] J. Murphy, in *Additives for Plastics Handbook*, 2nd ed. (Ed: J. Murphy Kidlington, Oxford) **2001**, p. 19.
[40] Z. Gao, L. Zhao, *Mater. Des.* **2015**, *66*, 176.
[41] C. Pan, K. Kou, Y. Zhang, Z. Li, G. Wu, *Compos. Part B* **2018**, *153*, 1.
[42] A. Mirmohseni, S. Zavareh, *Mater. Des.* **2010**, *31*, 2699.
[43] J. K. Katiyar, S. K. Sinha, A. Kumar, *Wear* **2016**, *362*, 199.

[44] S. Ojha, G. Raghavendra, S. K. Acharya, *Polym. Compos.* **2014**, *35*, 180.
[45] I. Z. Luna, K. C. Dam, A. M. S. Chowdhury, M. A. Gafur, R. A. Khan, *Int. J. Compos. Mater.* **2015**, *5*, 162.
[46] B. Rashid, Z. Leman, M. Jawaid, M. J. Ghazali, M. R. Ishak, *Int. J. Precis. Eng. Manuf.* **2016**, *17*, 1001.
[47] C. Liu, M. Chen, D. Zhou, D. Wu, W. Yu, *J. Nanomater* **2017**, *2017*, 6375135. https://doi.org/10.1155/2017/6375135.
[48] R. A. Naik, *J. Compos. Mater.* **1995**, *29*, 2334.
[49] R. A. Naik. NASA Contract. Rep. **1994**, 194930.
[50] A. Manral, P. K. Bajpai, *Polym. Compos.* **2019**, *41*, 691. https://doi.org/10.1002/pc.25399.
[51] N. Venkateshwaran, A. Elayaperumal, R. H. ArwinRaj, *J. Polym. Environ.* **2012**, *20*, 565.
[52] S. Houshyar, R. A. Shanks, A. Hodzic, *Macromol. Mater. Eng.* **2005**, *290*, 45.
[53] M. Rajesh, S. P. Singh, J. Pitchaimani, *J. Ind. Text.* **2018**, *47*, 938.
[54] V. A. Guenon, T. W. Chou, J. W. G. Jr, *J. Mater. Sci.* **1989**, *24*, 4168.
[55] F. Stig, S. Hallstrom, *Compos. Sci. Technol.* **2009**, *69*, 1686.
[56] Z. Hu, R. Karki, *J. Compos. Mater.* **2015**, *49*, 1513.
[57] P. J. Callus, A. P. Mouritz, M. K. Bannister, K. H. Leong, *Compos. Part A Appl. Sci. Manuf.* **1999**, *30*, 1277.
[58] S. Ramakrihna, N. K. Cuong, H. Hamada, *J. Reinf. Plast. Compos.* **1997**, *16*, 946.
[59] W. L. Wu, M. Kotaki, A. Fujita, *J. Reinf. Plast. Compos.* **1993**, *12*, 1096.
[60] W. Hufenbach, R. Bohm, M. Thieme, A. Winkler, E. Mader, J. Rausch, M. Schade, *Mater. Des.* **2011**, *32*, 1468.
[61] S. Mortazavian, A. Fatemi, *Compos. Part B* **2015**, *72*, 116.
[62] S. Eksi, K. Genel, *Acta Phys. Pol. A* **2017**, *132*, 879.
[63] K. Benzarti, L. Cangemi, F. D. Maso, *Compos. Part A Appl. Sci. Manuf.* **2001**, *32*, 197.
[64] M. P. Saiman, M. S. Wahab, M. U. Wahit, in *Proceedings of the International Colloquium in Textile Engineering, Fashion, Apparel and Design* (Eds: M. R. Ahmad, M. F. Yahya), Springer, Singapore **2014**. https://doi.org/10.1007/978-981-287-011-7_5.
[65] C. C. Yang, T. Ngo, P. Tran, *Mater. Des.* **2015**, *85*, 282.
[66] M. V. Hosur, U. K. Vaidya, C. Ulven, S. Jeelani, *Compos. Struct.* **2004**, *64*, 455.
[67] M. Rajesh, J. Pitchaimani, *J. Reinf. Plast. Compos.* **2015**, *35*, 228.
[68] A. B. H. Kueh, *Int. J. Mech. Sci.* **2013**, *66*, 45.
[69] M. L. Dano, G. Gendron, A. Picard, *Mech. Compos. Mater. Struct.* **2000**, *7*, 207.
[70] A. Fujita, H. Hamada, Z. Maekawa, *J. Compos. Mater.* **1993**, *27*, 1428.
[71] A. Axinte, N. Taranu, L. Bejan, *Appl. Compos. Mater.* **2016**, *23*, 987.
[72] Y. Nawab, M. Kashif, M. A. Asghar, A. Asghar, M. Umair, K. Shaker, M. Zeeshan, *Appl. Compos. Mater.* **2018**, *25*, 747.
[73] R. Mishra, *Polym. Compos.* **2014**, *37*, 472.
[74] S. Chou, H. C. Chen, H. E. Chen, *Compos. Sci. Technol.* **1992**, *45*, 23.
[75] S. Rudov-Clark, A. P. Mouritz, *Compos. Part A Appl. Sci. Manuf.* **2008**, *39*, 1018.
[76] S. C. Ray, in *Fundamentals and Advances in Knitting Technology*, 1st ed. (Ed: S. C. Ray), Woodhead Publishing, Daryaganj, New Delhi-110002, India **2011**.
[77] S. Ramakrishna, *Compos. Sci. Technol.* **1997**, *57*, 1.
[78] K. W. Lau, T. Dias, *J. Text. Inst.* **2009**, *85*, 173.
[79] S. Chou, C. J. Wu, *J. Reinf. Plast. Compos.* **1992**, *11*, 1239.
[80] K. H. Leong, S. Ramakrishna, Z. M. Huang, G. A. Bibo, *Compos. Part A Appl. Sci. Manuf.* **2000**, *31*, 197.
[81] W. Ashraf, Y. Nawab, M. Umair, K. Shaker, M. Karahan, *J. Text Inst.* **2017**, *108*, 1510.
[82] H. B. Dexter, G. H. Hasko, *Compos. Sci. Technol.* **1996**, *56*, 367.
[83] L. H. De-Carvalho, J. M. F. Cavalcante, J. R. M. D'Almeida, *Polym-Plast. Technol.* **2006**, *45*, 791.
[84] Z. Guo. *Specialist Yarn and Fabric Structures* (Ed: R. H. Gong), Woodhead Publishing, **2011**.
[85] P. Potluri, S. Nawaz. *Specialist Yarn and Fabric Structures* (Ed: R. H. Gong), Woodhead Publishing, **2011**.
[86] P. J. Minguet, C. K. Gunther. NASA Contract. Rep. **1994**, 4610.
[87] P. J. Falzon, I. Herszberg, *Compos. Sci. Technol.* **1998**, *58*, 253.
[88] K. Birkefeld, M. Roder, T. V. Reden, M. Bulat, K. Drechsler, *Appl. Compos. Mater.* **2012**, *19*, 259.
[89] K. M. Charlebois, R. Boukhili, O. Zebdi, F. Trochu, *J. Reinf. Plast. Compos.* **2005**, *24*, 1539.
[90] M. S. Dadkhah, J. G. Flintoff, T. Kniveton, B. N. Cox, *Composites* **1995**, *26*, 561.
[91] E. Zywicz, T. Nguyen, *Compos. Sci. Technol.* **2000**, *60*, 2989.
[92] R. A. Naik, P. G. Ifju, J. E. Masters, *J. Compos. Mater.* **1993**, *28*, 656.
[93] H. Ahn, W. R. Yu, *Compos. Struct.* **2016**, *160*, 1105.
[94] Z. Yuanyuan, S. Ying, L. Jialu, L. Limin, C. Li, L. Junling, T. Shuquan, *Mater. Des.* **2016**, *116*, 246.
[95] X. Sun, C. Sun, *Compos. Struct.* **2004**, *65*, 485.
[96] X. Gao, A. Siddique, B. Sun, B. Gu, *Appl. Compos. Mater.* **2019**, *26*, 1261.
[97] Y. Z. Wan, Y. L. Wang, F. He, Y. Huang, H. J. Jiang, *Compos. Part A Appl. Sci. Manuf.* **2007**, *38*, 495.
[98] T. J. Kang, K. H. Jung, J. K. Park, J. R. Youn, *Polym. Polym. Compos.* **2002**, *10*, 521.
[99] R. Yan, S. Y. Huang, C. H. Huang, C. T. Hsieh, C. W. Lou, J. H. Lin, *J. Compos. Mater.* **2017**, *51*, 1045.
[100] T. J. Kang, S. H. Lee, *J. Compos. Mater.* **1999**, *33*, 2116.
[101] S. H. Lee, T. J. Kang, *J. Compos. Mater.* **2000**, *34*, 816.
[102] S. Sengupta, P. Ray, P. K. Majumdar, *Indian J. Fiber Text.* **2008**, *33*, 411.
[103] X. Chen, L. Chen, C. Zhang, L. Song, D. Zhang, *Compos. Part A Appl. Sci. Manuf.* **2016**, *85*, 12.
[104] Y. Cai, S. Fan, X. Yin, L. Zhang, L. Cheng, Y. Wang, *Ceram. Int.* **2014**, *40*, 399.
[105] Y. Cai, S. Fan, H. Liu, L. Zhang, L. Cheng, J. Jiang, B. Dong, *Mater. Sci. Eng. A* **2010**, *527*, 539.

How to cite this article: Ahmad F, Yuvaraj N, Bajpai PK. Effect of reinforcement architecture on the macroscopic mechanical properties of fiberous polymer composites: A review. *Polymer Composites*. 2020;1–17. https://doi.org/10.1002/pc.25666

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Efficient Power Theft Detection using Smart Meter Data in Advanced Metering Infrastructure

Salil Manocha
Computer Science Engineering
Delhi Technological University
New Delhi, India.
Salilmanocha34@gmail.com

Vivek Bansal
Computer Science Engineering
Delhi Technological University
New Delhi, India.
Vivekbansal2910@gmail.com

Ishan Kaushal
Computer Science Engineering
Delhi Technological University
New Delhi, India.
Ishankaushal25@gmail.com

Dr. Aruna Bhat, Associate Professor
Department of Computer Science Engineering, Delhi Technological University
New Delhi, India.
Aruna.bhat@dtu.ac.in

Abstract: Since its discovery, electricity is the basic necessity of human life. It has helped to boost the industrial revolution and any economy, especially the developing ones are highly dependent on a regular and cheap supply of power. But power theft is a major threat to electric companies, it is catastrophic to power distributors and is a reason for financial damage. Regular Manual Inspection of all customers is a solution but it is very time and cost consuming and also leaves scope for corruption. So, AMI (Advanced Metering Infrastructure) has been introduced, which acts as two-way communication between the supplying company and customer via a smart meter which automates the measurement of electric energy usage. But this new technique brings in the possibility of new kinds of thefts and the consumption data compiled from this AMI has the potential to help build a system for fraud detection. Here, a machine learning-based detection methodology is presented, which first applies GMM clustering to categorize data into consumption classes. Then we improve the working of an ANN-based Classification model using a softmax layer and ensembling. The robust dataset used here helps to train the model efficiently and it achieved an astonishing accuracy of 92%. Results indicate the surpassing performance of this method in comparison to other existing detectors.

***Keywords*: AMI, ANN, Ensembling, GMM, Power Theft Detection, Softmax.**

I. INTRODUCTION

Power theft and its illegal consumption results in immense economic loss for many countries such as India ($6 B/year) and the United States ($5 B/year) [1], [2]. The worst-hit are the developing economies. In [3], it is also evaluated that power utilities are losing about 21 Billion Dollars globally, as power theft every year. These days companies are introducing a new Smart Grid called AMI (Advanced Metering Infrastructure), which involves the installation of a smart meter at the customer's end. This acts as two-way communication between supplier and consumer. All real-time readings can be viewed at the supplier's end, hence the conventional practice of bribing the reading agent (to manipulate manual readings) and meter tampering goes out of the equation. Yet, a new category of thefts like cyber-interventions has appeared that targets the smart meters [4].

In this paper, we exploit the potential of consumption data and develop an efficient system to identify fraudulent customers. This will help to save the cost incurred in regular manual inspections of all meters since our system will point out a possible defaulter and inspection would be done only at his site.

The rest of our research document has been formulated as: Section II gives a concise Literature Survey of the extant research in this field. Section II discusses a network and threat model. Energy consumption data and our proposed methodology are discussed in section IV. Corresponding study results are discussed in section V and also comparative study and the conclusion is drawn in section VI.

II. LITERATURE REVIEW

The energy losses have been categorized into 2 types [5]: NTL (nontechnical loss) and TL (technical losses). The technical losses come with the transmission of power, which occurs due to heat generated in the power transmission and other components like the transformers and wires [6]; the difference between the net loss and technical loss gives NTL, which is primarily caused by power theft.

The power theft occurs mostly through line tapping, meter breaking, or meter reading tampering (physical attacks) and cyber interventions [7].

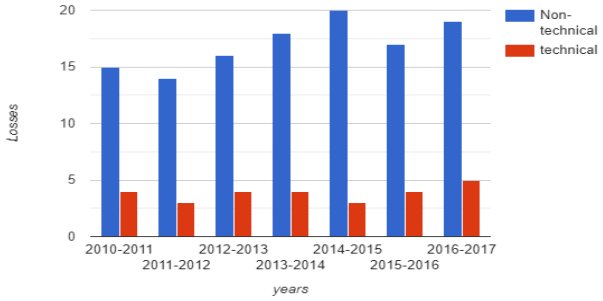


Fig. 2. The difference in technical and non-technical losses

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:20 UTC from IEEE Xplore. Restrictions apply.

Several automatic fraud detection models have been proposed up to the date. They can be classified mainly into machine learning/deep learning-based, special hardware installation, and game theory-based. [8] involves the installation of a device called the FRTU (feeder remote terminal unit), which helps tighten the analysis zone of power theft check-in a locality. But the budget of these units is an issue here, so [8] also proposes a new dynamic programming-based technique that installs a minimum number of FRTUs satisfying detection rate constraint by evaluating every contender solution in a bottom-up method with help of an innovative pruning technique.

Several methodologies have been worked upon using machine/deep learning techniques. In [9], a top-down model uses a DT (Decision Tree) and SVM (Support Vector Machine), it is a two-level model. Dataset was first processed on Decision Tree and then given for training to the SVM classifier. Deep Learning techniques have also been used in a variety of ways. In [10] the author uses Artificial Neural Networks (ANN) for fraud detection in AMI by building a model on consumption patterns of genuine profiled customers and then comparing the actual test data with predicted consumption. If the difference is more than a certain threshold, a warning is issued. [11] further ties to bring in an improvement by exploiting the time-sensitive nature of the customer's power usage, hence implementing a gated recurrent neural network unit (GRU-RNN), working on a similar dataset as [10]. [12] uses a slightly different approach, the 1-D electricity data is convolved by a convolutional neural network (CNN) layer and this produces a 2-D consumption pattern, this helps in extracting features from the dataset. These convolution features are fed as input to the recurrent neural network (RNN) which further has some hidden layers and each cell in this RNN is a Bi-LSTM cell (Bi-Directional Long Short-Term Memory). The output of this model used in [12] are detected values of power theft. [13] also uses a similar approach on a customized IEEE bus test-feeder, by first using a CNN and then LSTM layer and later implementing Transfer Learning by using an autoencoder (Spatial and Temporal Deep Learning). [14] uses a different approach for classification of consumers into consumption profiles, this is done by first clustering the data using a C-means-based fuzzy algorithm and then using a fuzzy classification where potential fraudsters with irregular consumption patterns are identified. An improvement model for SVM-based Loss Detection is presented in [15]. In [16], the author implements three models i.e. Boolean logic, fuzzy logic, and SVM on real-world data, and presents their comparison, this research shows its efficiency on extensive and unbalanced real-world datasets.

There are some other approaches too, like [17] uses the random matrix theory (RMT) to conduct a study with the IEEE bus system and readings gathered from the electricity utility of a place to study the efficiency of their method.

The contributions of this research project are described as:

1. A network model and a threat model for the power grid are discussed.
2. A description of Dataset is given.
3. Working and training of the model have been explained.
4. Its Application for Fraud Detection is discussed.
5. The results are discussed.

As [18] states that clustering combined with classification is a technique to improve the accuracy of the model, we build our model upon this principle. The complete dataset is fed to the clustering algorithm, and it forms 3 clusters (Industrial, Commercial, Residential) based on the total power consumption of the consumer over some time. This will add a new feature to our dataset which will be further used for building the classification model. This distinction is useful because different groups may exhibit different usage patterns, e.g. during the night a retail shop(commercial) will have almost nil consumption, while in a house consumption during night time will be more (e.g. use of heating systems during winters).

III. THREAT MODEL

Power Theft techniques are examined in [19]. [2] also discusses these by presenting a network model (Fig. 1) and a threat model. It is the responsibility of distribution substation to analyze if the readings coming from the smart meters of that area are genuine or fraudulent, therefore security measures have to be undertaken at this substation. When the amount of billed power differs considerably from the supplied power (after taking into account the technical losses),

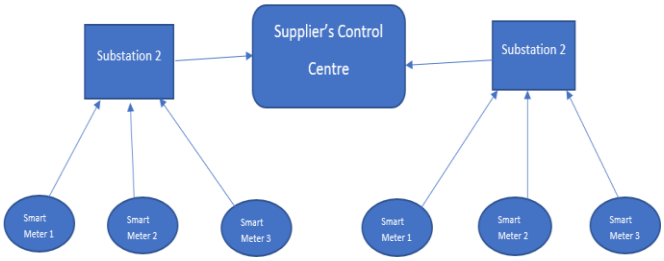


Fig. 1. Network Model

1) Primary Attackers: Electricity customers, who may indulge in line-hooking or meter tampering.
2) Secondary Attackers: Professional hackers, who use other techniques to manipulate meters.

Power Theft [5] is segregated into 3 classes of attacks:

i) Physical: Line-hooking (bye-passing the actual meter to remove loads). Meter tampering (using magnets to disturb the working of the meter) is another type.
ii) Cyber: Cyber-attacks are executed via transmission /communication line between the power supplying organization and the smart meter by interrupting the measurements by network congestion and remote network exploits.
iii) Data: Can be done through Physical and Cyber-attacks, mainly targeting power consumption readings.

IV. PROPOSED METHODOLOGY

A. Dataset

The energy consumption dataset of 3000 meters used in this research project has been taken from OpenEI's data repository. In our proposed methodology, we analyze the data of consumers over 28 days. The file had consumption data of consumers from various backgrounds.

Every day, it has hourly readings also. It is structured as in table I.

For all customers, aggregated readings for a day are obtained.

After clustering, the data of fraudulent customers are added.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:20 UTC from IEEE Xplore. Restrictions apply.

This bogus data is generated as a result of attack simulation on genuine data.

Table I: Smart Meter Dataset

Date/Time	Electricity:Facility [kWh]	Fans: Electricity [kWh]	Cooling:Electricity [kWh]	Heating: Electricity [kWh]	InteriorLights:Electricity [kWh]	InteriorEquipment:Electricity [kWh]	Gas: Facility [kWh]	Heating: Gas [kWh]	InteriorEquipment:Gas [kWh]	Water Heater:WaterSystems:Gas [kWh]
01/01 16:00 :00	157.8693	0.857363	55.84186	0	43.51334	52.53974	4.85426	0	36.25695	17.33396
01/01 17:00 :00	132.9779	0.422348	25.9926	0	43.51334	52.53974	4.85426	0	4.83426	0.02

For every sample in the dataset, the following types of bogus data are generated:

1. On some random day, each attribute is assigned a value 0.
2. On some random day, the value of each attribute is multiplied by a random value less than 0.5.
3. On some random day, the value of all attributes is replaced by the mean of that column over the month.

Table II: Trustworthiness of Consumers (0-Fraudulent, 1-Genuine)

Type	Original Data	Modified Data	Output
0	D_i	D_i(genuine)	1
1	D_j	D_j ✕ 0	0
2	D_k	D_k ✕ Random(0-0.5)	0
3	D_x	Mean($D_1 \ldots D_{28}$)	0

The data with these three classes (1,2,3) is added to the original dataset.

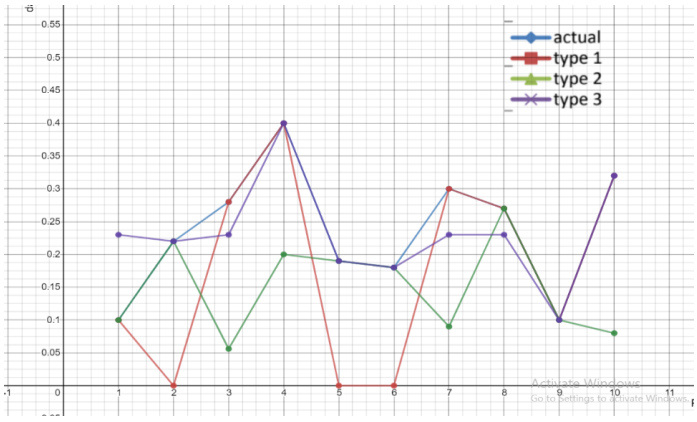


Fig. 3. Sample customer data for 10 days with simulated fraudulent data (Power-day)

But there is a problem of class imbalance in the dataset. To reduce this, [20] discusses an algorithm called Synthetic Minority Oversampling Technique (SMOTE) to make the number of fraudulent and genuine samples comparable in the dataset by minority class oversampling and this helps to avoid our model's biasing towards the majority class. In a standard SMOTE algorithm:

1. In set S, N number of K-nearest neighbors (belonging to minority class) are listed for a random sample (x_1) picked from minority class.
2. For each y in S, the following equation generates a new sample (z):

$$z = x_1 + random(0,1) \times |x_1 - y| \quad (1)$$

B. Building the Model

We have used a supervised learning methodology to build our model.

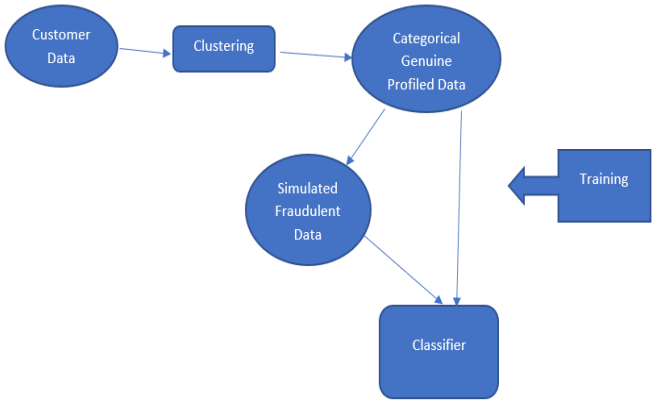


Fig. 4. Proposed Model

In the first step, the Gaussian mixture model clustering algorithm is used to tag the original genuine data into 3 classes.

Clustering Algorithm: Here we use an unsupervised technique called GMM (Gaussian mixture models), it doesn't represent clusters as exclusive circles but rather as probability distributions and they can cluster with any elliptical shape. This function describes Gaussian distribution:

$$G(y;\ M, \rho^2) = \frac{1}{\sqrt{2\pi\rho^2}} e^{-\frac{(y-M)^2}{2\rho^2}} \quad (2)$$

y is the observation, ρ = standard deviation, M = mean and ρ^2 = variance in the given distribution.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:20 UTC from IEEE Xplore. Restrictions apply.

We need to explicitly define the total number (3 here) of clusters that we want. GMM assumes that a limited mixture of Gaussians generates data. GMMs attempt to learn parameters from data and this is done by using Expectation-Maximization (EM) to find the maximum likelihood. Hence it uses Bayes Theorem [24] to calculate the probability of a given observation D_i to belong to each cluster k. This EM can be explained in two phases [24]:

1. In Estimation step, we calculate the likelihood (L_k) of each observation D_i belonging to cluster k, using:

$$L_k = \frac{G(D_i \mid \mu_k, \sigma_k^2)\Theta_k}{\Sigma_{k=1}^{K} G(D_i \mid \mu_k, \sigma_k^2)\,\Theta_k} \quad (3)$$

   Θ, our consideration that a sample was drawn from one of the modeling Gaussians.

2. In the Maximization step, we calculate our learning parameters as:

$$\mu_k = \frac{\Sigma L_k y}{\Sigma L_k},$$

$$\sigma_k^2 = \frac{\Sigma L_k (D_i - \mu_k)^2}{\Sigma L_k},$$

$$\Theta_k = \frac{1}{n} \Sigma L_k \quad (4)$$

   n is the total number of observations. For each cluster, $\mu_k, \sigma_k^2, \Theta_k$ are updated [26].

The clusters in GMM appear as [24]:



Fig. 5. Clusters with GMM

Python has direct library support to implement this clustering algorithm in sklearn.
Then fraudulent data is added. After feature selection, the final dataset is fed to the classification system for training our model.

Classification Algorithm: To make the classification process more robust, we build an ensemble deep neural network classifier with a softmax output layer, as discussed in [21].

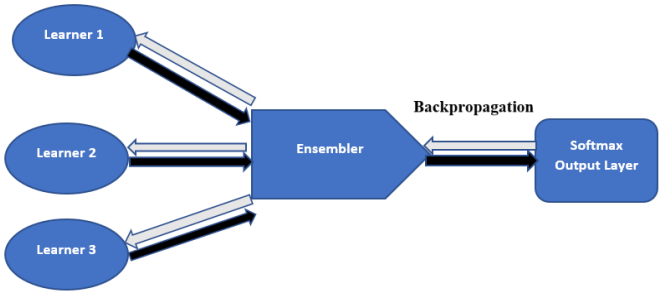


Fig. 6. Classification Architecture

The softmax layer[23] works as follows:

1. The ensemble model's output is a vector of probabilities, each probability describes the belongingness of some test sample to each output class.
2. Softmax does the work of aggregation to give a final probability.
3. The loss function used here is the usual cross-entropy loss function used in logistic regression.
4. The class conditional probability using Bayes Rule are [26]:

$$P(Class|Data) = P(Data|Class)P(Class)/P(Data),$$

   where P(Class) is called the former probabilities (assumed equal here) and

$$P(Data) = \sum_{classes} P(Data|Class)P(Class) \quad (5)$$

This technique outperforms traditional regression because it provides statistical significance to features or parameters. As discussed above, we used Softmax Activation with ensembled neural network classification.
For training, the neural networks have input dimension size the same as the number of features in the dataset, and batch size during training is equal to the number of days under consideration for a customer (28 here). If training for the whole year, the accuracy of the model will increase further. The dataset that we have used here is perfect for the job since it segregates the energy consumption in different activities and if this consumption is analyzed over a year, then neural network understands the pattern of usage as per season(the season is an important factor to determine power usage in various activities e.g. heating systems, etc.), hence is more accurate in predicting abnormalities.
Once the trained model is ready, it can be used for fraud detection.

C. Fraud Detection System

After the training part is over, our model is ready for deployment. It will run checks on the consumption data of every customer collected over a month. Again, after dropping the insignificant features, this new data is tested by the classifier and it decides whether that customer is genuine or fraudulent. If the user is found to be fraudulent, then manual onsite inspections are carried out. The performance parameters are discussed in section V.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:20 UTC from IEEE Xplore. Restrictions apply.

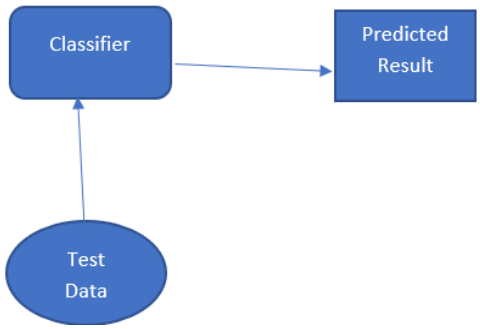


Fig. 7. Fraud Detection

V. RESULT

The dataset was split in a train to a test ratio of 70:30. The proposed technique performs very well and the performance parameters are given in Table III.

Table- III: Performance Measures

Accuracy	Recall	Precision
0.92	0.93	0.89

Since we have two output classes (Genuine and Fraud) we use a 2 ✕ 2 confusion matrix. The accuracy, recall, and precision for the model are evaluated using the confusion matrix as [22]:

$$Accuracy = \frac{(\text{True Positive} + \text{True Negative})}{(\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative})} \quad (6)$$

$$Recall = \text{True Positive} / (\text{True Positive} + \text{False Negative}) \quad (7)$$

$$Precision = \text{True Positive} / (\text{True Positive} + \text{False Positive}) \quad (8)$$

Where,
True Positives = number of records rightly classified as genuine
True Negatives = number of records rightly classified as fraudulent
False Negatives = number of records misclassified as fraudulent
False Positives = number of records misclassified as genuine

The dataset used in this study was tested on some other methodologies too. The comparative results are shown in Table IV.

Table- IV: Result Comparison

Model	Accuracy
SVM	0.82
Logistic Regression	0.65
Random Forest	0.80

VI. CONCLUSION

As we can interpret from the table, our technique significantly outperforms other methods by giving an accuracy of 92%.
This deep learning-based model can outperform conventional techniques since it does an automatic feature extraction from the dataset, unlike others that do feature extraction manually [25].
The same is evident from the ROC curves plotted corresponding to all listed techniques.
As we can interpret, the area under the curve (AUC) for the proposed methodology is maximum, hence it performs best.

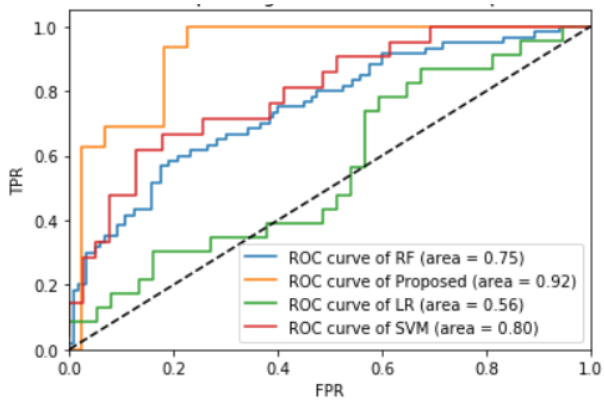


Fig. 8: Receiver operating characteristics for comparison

The scope of further improvement lies in the availability of real-time data of fraudulent consumers. With the proper balancing of two classes in the dataset, it will improve the overall performance of the classifier, hence reducing the cost and effort incurred during manual inspections.

Acknowledgment

The explanations given here are exclusively the responsibility of the writers. The writers would like to thank the OpenEI Data Center for granting us access to the Smart Meter Data of diverse consumers from the United States.

REFERENCES

1. P. Jokar, N. Arianpoo, and V. C. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, 2016.
2. S. K. Singh, R. Bose, and A. Joshi, "Entropy-based electricity theft detection in AMI network," *IET Cyber-Physical Systems: Theory & Applications*, 2017.
3. T. B. Smith, "Electricity theft: a comparative analysis," Energy Policy, vol. 32, no. 1, pp. 2067–2076, 2004.
4. M. Nabil, M. Ismail, M. Mahmoud, M. Shahin, K. Qaraqe, E. Serpedin, "Deep recurrent electricity theft detection in AMI networks with random tuning of hyper-parameters", *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, pp. 740-745, Aug. 2018.
5. S. S. S. R. Depuru, L. Wang, and V. Devabhaktuni, "Electricity theft: overview, issues, prevention and a smart meter based approach to control theft," Energy Policy, vol. 39, no. 2, pp. 1007–1015, 2011.
6. J. P. Navani, N. K. Sharma, and S. Sapra, "Technical and non-technical losses in power system and its economic consequence in Indian economy," *International Journal of Electronics and Computer Science Engineering*, vol. 1, no. 2, pp. 757–761, 2012.
7. S. McLaughlin, B. Holbert, A. Fawaz, R. Berthier, and S. Zonouz, "A multi-sensor energy theft detection framework for advanced metering infrastructures," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 7, pp. 1319–1330, 2013.
8. Y. Zhou, X. Chen, A.Y. Zomaya, L. Wang, S. Hu, "A dynamic programming algorithm for leveraging probabilistic detection of energy theft

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:20 UTC from IEEE Xplore. Restrictions apply.

in smart home", *IEEE Trans. Emerging Topics in Computing*, vol. 3, no. 4, pp. 502-513, Dec. 2015.

9. A. Jindal, A. Dua, K. Kaur et al., "Decision tree and SVM-based data analytics for theft detection in smart grid", *IEEE Trans. Ind. Inf.*, vol. 12, no. 3, pp. 1005-1016, 2016.
10. A Korba, "Energy Fraud Detection in Advanced Metering Infrastructure AMI", ICSENT '18, December 26–28, 2018, Hammamet, Tunisia.
11. M. Nabil, M. Ismail, M. Mahmoud, M. Shahin, K. Qaraqe, E. Serpedin, "Deep recurrent electricity theft detection in AMI networks with random tuning of hyper-parameters", *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, pp. 740-745, Aug. 2018.
12. P Chandel, T Thakur "Smart Meter Data Analysis for Electricity Theft Detection using Neural Networks", *Advances in Science, Technology and Engineering Systems Journal,* Vol. 4, No. 4, 161-168 (2019).
13. R. Bhat, R. Trevizan, R. Sengupta, X. Li, A. Bretas, "Identifying nontechnical power loss via spatial and temporal deep learning", *IEEE International Conference on Machine Learning Applications*, pp. 272-279, Dec. 2016.
14. E.W.S. dos Angelos, O.R. Saavedra, O.A.C. Cortés et al., "Detection and identification of abnormalities in customer consumptions in power distribution systems", *IEEE Trans. Power Deliv.*, vol. 26, no. 4, pp. 2436-2442, 2011.
15. J. Nagi, K.S. Yap, S.K. Tiong et al., "Improving SVM-based nontechnical loss detection in power utility using the Fuzzy inference system", *IEEE Trans. Power Deliv.*, vol. 26, no. 2, pp. 1284-1285, 2011.
16. P Glauner, A. Boechat, L. Dolberg et al., "Large-scale detection of non-technical losses in imbalanced data sets", *Innovative Smart Grid Technologies Conference (ISGT) 2016 IEEE Power & Energy Society*, 2016.
17. F. Xiao, Q. Ai, "Electricity theft detection in smart grid using random matrix theory", *IET Gen. Transmiss. Distrib.*, vol. 12, no. 2, pp. 371-378, 2017.
18. Yaswanth Kumar Alapati, Korrapati Sindhu, "Combining Clustering with Classification: A Technique to Improve Classification Accuracy", *International Journal of Computer Science Engineering (IJCSE)*.
19. McLaughlin, S., Holbert, B., Fawaz, A., et al.: 'A multi-sensor energy theft detection framework for advanced metering infrastructures', *IEEE J. Sel. Areas Commun*, 2013, 31, (7), pp. 1319–1330.
20. Shuan Li,Yinghua Han, Xu Yao,Song Yingchen, Jinkuan Wang, Qiang Zhao, "Electricity Theft Detection in Power Grids with Deep Learning and Random Forests", Available: https://www.hindawi.com/journals/jece/2019/4136874/.
21. Haihan Lan 2017, *The Softmax Function, Neural Net Outputs as Probabilities, and Ensemble Classifiers,* Medium, viewed 4 April 2020, Available: https://towardsdatascience.com/the-softmax-function-neural-net-outputs-as-probabilities-and-ensemble-classifiers-9bd94d75932 .
22. Koo Ping Shung 2018, *Accuracy, Precision, Recall or F1 ?,* Medium, viewed on 3 April 2020, Available: https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9
23. Bridle J.S. (1990) *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*. In: Soulié F.F., Hérault J. (eds) Neurocomputing. NATO ASI Series (Series F: Computer and Systems Sciences), vol 68. Springer, Berlin, Heidelberg.
24. Thalles Silva 2019, *How to code Gaussian Mixture Models in Python from scratch in Python,* Medium, viewed 3 April 2020, Available: https://towardsdatascience.com/how-to-code-gaussian-mixture-models-from-scratch-in-python-9e7975df5252 .
25. Bashar, Abul. "SURVEY ON EVOLVING DEEP LEARNING NEURAL NETWORK ARCHITECTURES." *Journal of Artificial Intelligence* 1, no. 02 (2019): 73-82.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:20 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

End-to-End Neural Network for Shadow Detection using Combined Color Spaces

Shailender Kumar
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

Dhruv Ramdev
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
dhruv_bt2k16@dtu.ac.in

Himanshu Aggarwal
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
himanshuagg_bt2k16@dtu.ac.in

Himanshu Choudhary
Department of Computer Science and Engineering
Delhi Technological University
Delhi, India
himanshu19_bt2k16@dtu.ac.in

***Abstract*—Over the years, many techniques for shadow detection have been proposed. However, most of them are computationally expensive and cannot be used in further applications. In this research, a new technique for shadow detection is presented, which is computationally efficient and can be used for further computer vision tasks. Our method uses images converted into different color spaces as input to a framework based on deep learning to detect shadow features from single images in an end to end manner. Our framework observes given image completely to learn and preserve the shadow features in various global and local contexts simultaneously. Our proposed method is trained and assessed on shadow detection benchmark dataset (SBU Shadow Dataset) and performs better than the state-of-the-art convolutional methods for shadow detection as it improves shadow error rate (SER) by 9% while also pointing out some irregularities in the dataset.**

***Index Terms*—Computer vision, deep learning, convolutional neural networks, color spaces**

I. INTRODUCTION

Shadow is an important characteristic of an object and is present at all places, where there is light and an opaque object. Shadow detection and manipulation is an elementary issue in various computer vision applications. Shadows are used in tasks related to shape and size of an object, movement of light source and illumination conditions. Shadows also have applications in augmented and virtual reality. On the other hand, shadows can interfere and cause issues in various image processing tasks, aerial imaging and object tracking [1], [2]. They significantly reduce the performance of many recognition, segmentation and reconstruction tasks.

Intelligent living beings can easily understand the difference between an object and a shadow. However, computer applications require a systematic approach to detect shadows and distinguish them from real objects so that further computer vision and image processing applications can run without any performance issue. This makes shadow detection a fundamental issue in computer vision domain.

Dealing with shadows for computer vision tasks is a well-researched subject and many novel techniques have been put forward over the years. These methods can be categorized into given two classes: non learning-based techniques [3]–[7], [7]-[11] and learning-based techniques [12]–[17]. The first category [3]-[7], [7]-[11] requires the understanding of all the properties



(a) (b)

Fig. 1. Shadow Detection Examples: (a) Input (b) Output. The figure shows some examples of the shadow masks obtained from our proposed method along with the input image

associated with shadows and necessitates an accurate labelling of non-shadow and shadow regions in an image, which is not possible in the real-world scenario as most images do not contain the details about any obstructing objects or the light source or any other essential property of the object. The second category [12]–[17] of shadow detection methods is based on deep learning frameworks. Majority of the current state of the art approaches for detecting shadows belong to the second category and rely on neural networks for predicting the shadow regions.

In this research, we introduce a novel approach for detecting shadows based on encoder-decoder convolution neural network architecture, which is widely accepted as the best framework for semantic segmentation tasks. In our method, we convert images to different color spaces and then feed

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:56 UTC from IEEE Xplore. Restrictions apply.

them to the architecture for training. The network learns to estimate the shadow regions in input images and the network performance is then evaluated using the BER measure. Comparing the performance of the model on different color spaces gives us the best result that can be used in further computer vision applications. Since our model has been trained to predict shadow regions in all color spaces, it is independent of the image quality or any other limitation of the standard RGB color space model used in existing models. Therefore, our network architecture is suitable for any computer vision or robotic applications and requires only minimal pre-processing. No other prior information or parameter tuning is required thus making our method faster than the current methods. Our prime contributions in this research are outlined below:

- First, converting input images into various color spaces.
- Second, constructing a novel network architecture for detecting shadows based on encoder-decoder architecture. We then train the model using weighted BCE dice loss that balances the BER measure for shadow and no shadow areas.
- At last, we assess the model on the benchmark SBU Shadow dataset [13] and compare the performance with different color spaces and against pre-existing methods. The result reveals that our model performs best when used with LAB color space and the performance is equally good as any when compared to already existing models.

This paper is organised in 5 sections. The remaining sections elaborate on the related and recent studies on shadow detection (Section II), Section III, elaborates on our work methodology and architecture of our network. In section IV, we present a comparative study of our model's performance against existing state of the art methods over SBU dataset. We summarize our method and results in Section V.

II. Related Work

Given the importance of detecting shadows in natural images, shadow detection is a well-researched topic. Many methods are available for detecting shadows in images including both learning-based and non-learning-based methods. Earlier works [9]–[11] focused on physical models of color and illumination for detecting shadows in natural images. These methods take advantage of different illumination components between the no shadow and shadow pixels. However, the methods using such an approach tend to yield acceptable results only for the images that have a wide dynamic range [4], [15]. Some methods [3]–[7] learn the shadow features in images using handcrafted features that are based on shadow images that have been annotated carefully. Such methods [3]–[7] first segment image regions by shadow features like color, texture, edge and then using probabilities, classify them as either non-shadow region or shadow region through the use of decision trees, SVM and other classifiers. Vicente et al. [8] devised a novel method that partitions the image into multiple disjoint regions and shadow probability of each region is computed using Least-Squares SVM.

Recently, many learning-based methods [12]–[17] are being used for detecting shadows particularly after the introduction of modern methods based on deep learning. CNN have been particularly found extremely useful in tasks involving segmentation of images. Many recent methods have used various CNN architectures for detecting shadows and outperforming traditional approaches. Khan et al. [12] presented a framework that learned shadow features using several CNNs along the boundaries and at super-pixel level. CRF (Conditional Random Field) was then used to locate shadow pixels. Vicente et al. [13] in their CNN based method – Stacked CNN, used a prior CNN to learn semantic features and use these features to train another CNN to accurately detect shadow regions. They [13] also presented the shadow detection dataset – SBU Shadow dataset that is now used as a benchmark dataset to detect shadows. A similar method is used in Hosseinzadeh et al. [14] in which prior shadow map generated using handcrafted features are used along with a patch level CNN framework to predict regions of shadows in images. Nguyen et al. [15] were one of the first to use Generative Adversarial Network (GAN) framework to detect shadow in single images. They [15] presented a Conditional GAN (scGAN), which both generator and discriminator networks incorporate a sensitivity parameter to efficiently train the network and predict the shadow masks in given images. Their work stressed on the importance of global features in predicting shadow masks. Mohajerani et al. [16] presented Context-Preserver CNN (CP-Net) that uses both local and global context features to predict the shadow masks for images. CP-Net [16] is computationally inexpensive compared to other methods. Recently, Le et al. [17] proposed an AD Net framework based on GANs which used detection network to detect shadows in both input samples and hard-to-predict samples generated by attenuation network.

In our work, we put forward a method for predicting shadow masks. Our proposed method uses both local and global shadow features for efficiently detecting shadow regions while also incorporating advantages offered by different color space models. The next section elaborates on our proposed method and training strategy.

III. Methodology

A. Proposed Color Spaces

For correct identification of shadow by intelligent systems, it is recommended to take advantage of color and camera calibration. While dealing with single image shadow detection, this is not possible. This presents problems in detection of hard shadows or being able to distinguish dark objects from shadows. Recent shadow detection techniques like CP-Net face similar problems where they are unable to distinguish shadows from darker albedo regions. We propose to use a combination of multiple color spaces as an alternative to multiple images as an input to our classifier model. This provides different perspectives for our classification neural network to categorize the pixel as either a non-shadow or shadow pixel. For example, for the RGB color space, the RGB values of the shadow compared to its surrounding whereas the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:56 UTC from IEEE Xplore. Restrictions apply.

HSV colors offer improved separation of chromaticity and intensity between the shadow and surrounding. Using multiple colors spaces offer a unique advantage to the neural network to associate the different values from different color spaces to obtain supposedly better results. We choose RGB, LAB, HSV, YCbCr and their various combinations as input to the model. To combine various color spaces we increased the input size of our classification model to 256x256x6 or 256x256x9 from 256x256x3 depending upon the number of images given as the input.

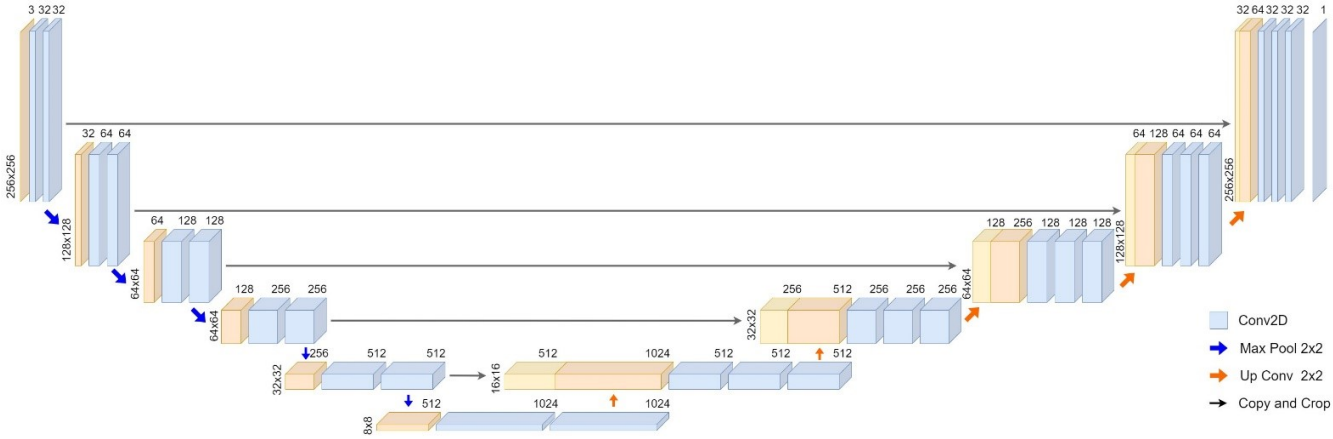


Fig. 2. Architecture of the proposed network

B. Proposed Model Architecture

We propose a customized, highly iterative shadow segmentation neural network built upon the existing U-Net model [18], an end to end CNN architecture for quick and accurate segmentation of images. The input for our model are three channelled images of size 256×256×3, and output are binary segmentation maps of size 256×256×1, in which the black and white pixels corresponds to the non-shadow and shadow areas respectively. The architecture has two paths, the downsampling (encoding) path and the corresponding symmetric upsampling (decoding) path. Both the downsampling and the upsampling path consist of 5 convolution blocks each, with 1 bridge block connecting the two paths.

In the downsampling path, input to each block is followed by two convolution layers with the size of the kernels (filters) as 3x3, and the convolution layers output is fed to a batch normalization layer [19] before feeding it to the ReLU activation function. Batch Normalization forces the input to each layer of the network to be of uniform distribution, leading to overall faster training time with an added advantage of introducing a bit of regularization. A max-pooling layer with 2×2 filters and 2×2 stride follows the output from the activation function. Starting with convolution layers of 32 filters, the number of filters doubles with each consecutive block in the downsampling (encoding) path. The context features of the input image are encoded via the convolution layers in the down- sampling path and the spatial size of the accumulated features is then reduced by the max-pooling layer.

Bridge block connects both upsampling and downsampling paths and consists of two convolution layers with the size of the kernels (filters) as 3x3 along with batch normalization. The most dominant features are accumulated in bridge block and these features (contexts) are decoded step-by-step to the full input image resolution along the upsampling path.

Input to each block in the upsampling path is followed by an upsampling (transposed convolution) layer, which is then concatenated with the output of the corresponding downsampling block with the same height for precise localization, followed by three convolution layers with 3×3 sized kernels (filters) along with batch normalization. Starting with convolution layers of 512 kernels (filters), the number of filters halves with each consecutive block in the upsampling (decoding) path.

The final layer in the architecture consists of a sigmoid activation function that classifies each pixel of the input as a non-shadow or a shadow pixel.

C. Training Strategy

The input for our model are three channelled images of size 256×256×3, therefore before training, all the training input images are resized to 256×256×3 and all the ground-truth binary segmentation masks are resized to 256×256×1. The activation for all the convolution layers in the network is ReLU [20] except for the last layer which uses sigmoid function to extract probabilities of all pixels as either a non-shadow or a shadow pixel.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:56 UTC from IEEE Xplore. Restrictions apply.

With batch size of 32, and 0.001 as the initial learning rate, our network uses RMSProp optimizer [21] and minimizes a custom loss function which is a combination of weighted binary cross-entropy loss (WBCE) and weighted DICE loss (WDL). The inputs to entropy loss are converted with the help of sigmoid function. Both the loss functions are augmented with custom weight parameter w_0. w_0 allows the loss function to trade-off the number of false positives or false negatives by up- or down-weighting it's value.



Fig. 3. Shadow Detection Results: Input (a),(d), Ground truths (b),(e), Predictions (d),(f). Some sample results from our proposed method where the predicted shadow masks are shown along with the original input image and corresponding ground truths available in the dataset.

$$\text{Loss} = \text{WBCE} + \text{WDL} \quad (1)$$

$$\text{WBCE}(x, \hat{x}) = -(w_0 x \log(\hat{x}) + (1 - x)\log(1 - \hat{x})) \quad (2)$$

$$\text{WDL}(x, \hat{x}) = 1 - \frac{2w_0^2 x\hat{x} + 1}{w_0^2 x + w_0^2 \hat{x} + 1} \quad (3)$$

where, x denotes the predicted values and $\hat{x}$ denotes the ground truth.

After training is completed, we obtain a set of weights that are adequate for predicting a binary shadow segmentation mask for any given image in which the white-colored pixels represent the shadow areas and the black colored pixels represent the non-shadow areas.

IV. Experimental Results

To thoroughly evaluate the model, we trained the model on SBU dataset and then compared our method's performance with pre-existing methods.

A. Dataset

We train the proposed model on the largest publicly accessible annotated shadow dataset- SBU shadow dataset [13]. The SBU dataset [13] has 3676 available images for training and 409 images available for evaluating the performance of the model. The dataset [13] contains images with a number of backgrounds including roads, mountains, vehicles, animals, beaches, houses and also including many landscapes, close range and aerial shots.

B. Evaluation Metric

Our method is based on pixel-level labelling and thus, requires an evaluation metric to take into consideration the skewed distribution of pixels in shadow and non-shadow class in the output mask. To effectively assess the performance of our technique, we need a balanced metric which is unbiased to the skewed distribution. We use balanced error rate or BER metric for assessing the performance of different methods. BER metric is mean of shadow error rate and non-shadow error rate. The mathematical expression for BER metric is:

$$BER = \frac{1}{2}\left(\frac{FN}{TP + FN} + \frac{FP}{TN + FP}\right) \quad (4)$$

where, FN is number of False Negatives, FP is number of False Positive, TP is number of True Positives and TN is number of True Negatives.

We have used the BER measure as our evaluation metric, since it is less affected by skewed output classes than mean pixel accuracy as natural images have a greater number of non-shadow pixels than number of shadow pixels. In addition to BER, we also use Non-shadow and shadow error rate measures in model training, evaluation and comparison with other models. Non-shadow and shadow error rate are calculated using PER (Per Pixel Error Rate):

$$SER = \frac{FN}{TP+FN} \quad (5)$$

$$NSER = \frac{FP}{TN+FP} \quad (6)$$

C. Results and Comparison with Other Methods

Experimental results on SBU dataset for our model are summarized in Table 1 while Table 2 compares our model's performance with existing models [13], [15]–[17] based on BER metric. Table 1 gives a glimpse of the experimental results of our model with different color spaces. Our method performs the best with LAB color space with BER=6.47. The best BER for RGB is also close with a BER value of 6.6. Combination of various color spaces followed the individual color spaces with

TABLE I
Performance of Model on SBU Dataset with Different Color Spaces

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:56 UTC from IEEE Xplore. Restrictions apply.

Color Space	Shadow Error Rate	Non-Shadow Error Rate	BER
LAB	3.93	9.01	**6.47**
HSV	3.76	9.31	6.54
RGB	**3.60**	9.61	6.60
RGB+HSV	5.13	**8.10**	6.61
RGB+HSV+LAB	4.45	9.13	6.79
RGB+LAB	3.81	10.05	6.93
MIX	4.57	10.5	7.54
YCbCr	3.54	12.07	7.81

TABLE II
COMPARISON WITH EXISTING SHADOW DETECTION NETWORKS

Methods	Shadow Error Rate	Non-Shadow Error Rate	BER
Stacked-CNN [13]	9.6	12.5	11.0
scGAN [15]	7.8	10.4	9.1
CP-Net [16]	9.0	**5.2**	7.1
AD-Net [17]	5.3	5.5	**5.4**
Our Method (Best)	**3.93**	**9.01**	**6.47**

minimal loss of accuracy. Another important point to note is the shadow error rate of our model in all sorts of combinations of color spaces is better than the error-rate of any of the pre-existing convolutional methods.

As inferred from results from Table 2, our model shows promising results in comparison with other existing methods. Results show that are our model outperforms CP-Net [16] by 9%, scGAN [15] by 29% and Stacked CNN [13] by 41% whereas, the performance is quite close to AD-Net [17] model architecture. Moreover, the shadow-error rate for our model is best among all models.

Our proposed model architecture outperforms every state of the art method in detecting shadow pixels- AD-Net [17] by nearly 26%, CP-Net [16] by 56%, scGAN [15] by 49% and Stacked-CNN [13] by 59%. This shows our model detects shadows better than any other state of the art model [17]. Although our model gives best results in shadow error rate, the non-shadow error is not that promising when compared with AD-Net [17] and CP-Net [16]. This may be attributed to the reasons mentioned in analysis later in this section. Some examples of shadow masks predicted by our method from the SBU dataset are shown in Fig. (3).

D. Analysis

As mentioned earlier our model outperformed all but one state of the art method in BER metric measure. However, it was the best in shadow error rate among all models. The main problem with our model was with the non-shadow error rate. But on careful analysis this can be attributed to some anomalies in SBU dataset [13] itself.

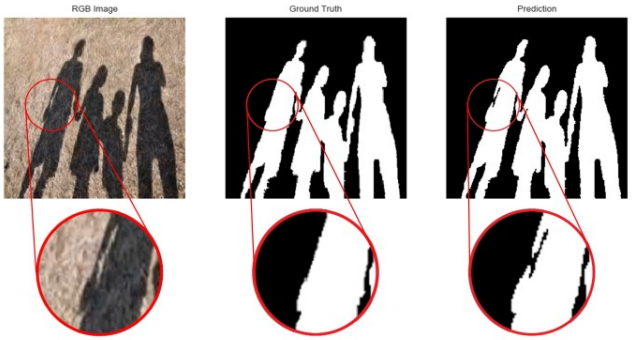


Fig. 4. Analysis: Predictions better than Ground Truth. On zooming into the given image, it can be clearly seen that the corresponding ground truth available in the dataset does not properly annotate the shadow pixels of the arm, however, our proposed method accurately recognizes the shadow pixels of the arm.

On careful examination of the example in Fig. (4), we can find one possible explanation for the non-shadow error rate in our model being high. For us, humans, it is obvious that the mask predicted by our model gives a more accurate measure of shadow in the image, but since the loss is measured by keeping ground truth as the base, a certain amount of pixels of our model will be regarded as false positives thereby increasing the non-shadow error rate.

However, all of the error cannot be attributed to the dataset anomaly only. But this case also highlights the fact that our model can also detect the shadows without any supervision thereby, increasing its real-world applications in computer vision and other areas.

E. Limitations

Although, the metric score for the proposed model is quite high and close to state of the art score but the model does produce some anomalies in its predictions. Such for the image in Fig. (5), the model produces a shadow mask as given. This problem arises when parts of the object are dark-colored and the model identifies them as the part of the shadow. Such regions are called albedo regions and almost all the state of the art models suffer from this anomaly. However, such anomalies are not easily found in the real world due to proper illumination conditions in our environment. But, for robotic applications such anomalies must be removed which will be our main goal in our future work.



Fig. 5. Failure Cases: Image with darker albedo regions. Our proposed method tends to predict inaccurate results when dark regions are present in the image along with the shadow.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:56 UTC from IEEE Xplore. Restrictions apply.

V. CONCLUSION

In this paper, we present a novel method to detect shadows in images, which in future can be exploited by various images processing and robotic applications. Our method uses traditional color spaces along with a neural net architecture for feature extractor using encoder-decoder architecture. Our key idea is to analyse the images to extract the features that can be attributed as being shadow features in different color space models. Our approach only requires minimal pre/post-processing making it computationally efficient. Experimental results on our model demonstrate significant improvement over pre-existing models in terms of detecting shadow pixels while also showing the short-comings of the dataset where our model predicts shadow masks better than the ground truth in the dataset itself. We plan to improve non-shadow error rate as well as focus on computationally efficient methods by which shadows can be removed from the images using the shadow masks predicted by our model in our future work.

REFERENCES

[1] C. Aeschliman, J. Park, and A. C. Kak, "Tracking vehicles through shadows and occlusions in wide-area aerial video," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 50, no. 1, pp. 429–444, 2014.

[2] W. Min, M. Fan, X. Guo, and Q. Han, "A new approach to track multiple vehicles with the combination of robust detection and two classifiers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 174–186, 2017.

[3] J. Zhu, K. G. Samuel, S. Z. Masood, and M. F. Tappen, "Learning to recognize shadows in monochromatic natural images," in *2010 IEEE Computer Society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 223–230.

[4] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan, "Detecting ground shadows in outdoor consumer photographs," in *European conference on computer vision*. Springer, 2010, pp. 322–335.

[5] R. Guo, Q. Dai, and D. Hoiem, "Single-image shadow detection and removal using paired regions," in *CVPR 2011*. IEEE, 2011, pp. 2033–2040.

[6] X. Huang, G. Hua, J. Tumblin, and L. Williams, "What characterizes a shadow boundary under the sun and sky?" in *2011 International Conference On Computer Vision*. IEEE, 2011, pp. 898–905.

[7] Y. Vicente, F. Tomas, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3388–3396.

[8] T. F. Y. Vicente, M. Hoai, and D. Samaras, "Leave-one-out kernel optimization for shadow detection and removal," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 682–695, 2017.

[9] G. D. Finlayson, S. D. Hordley, C. Lu, and M. S. Drew, "On the removal of shadows from images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 1, pp. 59–68, 2005.

[10] G. D. Finlayson, M. S. Drew, and C. Lu, "Entropy minimization for shadow removal," *International Journal of Computer Vision*, vol. 85, no. 1, pp. 35–57, 2009.

[11] J. Tian, X. Qi, L. Qu, and Y. Tang, "New spectrum ratio properties and features for shadow detection," *Pattern Recognition*, vol. 51, pp. 85–96, 2016.

[12] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic shadow detection and removal from a single image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 431–446, 2015.

[13] T. F. Y. Vicente, L. Hou, C.-P. Yu, M. Hoai, and D. Samaras, "Largescale training of shadow detectors with noisily-annotated shadow examples," in *European Conference on Computer Vision*. Springer, 2016, pp. 816–832.

[14] S. Hosseinzadeh, M. Shakeri, and H. Zhang, "Fast shadow detection from a single image using a patched convolutional neural network," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3124–3129.

[15] V. Nguyen, Y. Vicente, F. Tomas, M. Zhao, M. Hoai, and D. Samaras, "Shadow detection with conditional generative adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4510–4518.

[16] S. Mohajerani and P. Saeedi, "Cpnet: A context preserver convolutional neural network for detecting shadows in single rgb images," in *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2018, pp. 1–5.

[17] H. Le, Y. Vicente, F. Tomas, V. Nguyen, M. Hoai, and D. Samaras, "A+ d net: Training a shadow detector with adversarial shadow attenuation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 662–678.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[21] T. Tieleman and G. Hinton, "Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," *Technical Report.*, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:29:56 UTC from IEEE Xplore. Restrictions apply.

Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-020-01001-x

ORIGINAL ARTICLE

Exploiting bad-smells and object-oriented characteristics to prioritize classes for refactoring

Ruchika Malhotra[1] · **Priya Singh**[1]

Received: 25 August 2019 / Revised: 20 March 2020
© The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2020

Abstract Bad-smell indicates code-design flaws and poor software-quality that weaken software design and inversely affects software development. It also works as a catalyst for bugs and failures in the software system. Refactoring methods are used by software practitioners as corrective actions for bad-smells. The problem relies in the fact that there are over seventy refactoring methods available in literature and multiple refactoring methods can be used to nullify the effect of a particular bad-smell. So, it becomes very difficult to apply refactoring on complete source-code and almost impossible if software size is dramatically large. Thus, there arises a need for prioritizing classes in some way. This study aims at applying refactoring solution to only severely affected classes to improve the overall software quality. We proposed a framework that detects a small subset of classes from the entire source-code instantly require refactoring. This prioritization of classes is based on two factors-severity of the presence of bad-smells and object-oriented characteristics. The approach is evaluated on eight open-source Java software systems using ten most common bad-smells and six widely known C&K metrics. Both these factors help in calculating a new metric Quality Depreciation Index Rule (QDIR) that exposes those classes that are highly affected by bad-smells and demand an immediate refactoring solution. Results of the empirical study indicate that QDIR is an effective metric to remove bad-smells in an environment of stringent time constraints and limited cost making the maintenance of software system easier and effective with enhanced software quality.

Keywords Bad-smells · Refactoring methods · Object-oriented characteristics · Software quality · Software maintenance

1 Introduction

Software quality, in simple words, is defined as the degree to which a given software system, product or process adheres to the desired requirements imposed on it and meets user's expectations. By quantifying the software quality, a better estimate of how a software system adheres to specified requirement and user expectations can be made. Many metrics have been proposed by researchers in the past to measure the software quality like MOOD's metrics (Abreu and Carapuça 1994; Chidamber and Kemerer 1994; Lorenz and Kidd 1994) to name a few. In the current study, we have captured object-oriented characteristics of the chosen software systems that cover aspects of size, abstraction level, complexity, inheritance, cohesion, and coupling for each class by means of C&K (Chidamber and Kemerer 1994) metrics.

According to Fowler et al. (1999), a bad-smell is a poor internal-structure in the source code's design which does not stop it from executing but weakens its design and serves as a ground for bugs thereby degrading the overall software quality and increases the overhead of software maintenance. Fowler et al. (1999) has given a catalog of 22 code bad-smells that marks flaws in the design and if not

✉ Ruchika Malhotra
ruchikamalhotra2004@yahoo.com

Priya Singh
priya23294@gmail.com

[1] Department of Software Engineering, Delhi Technological University, Bawana, Delhi 110042, India

Published online: 18 June 2020

handled, can cause serious damages. Refactoring is regarded as a solution to bad-smells and is defined as a technique that refines the internal structure of the source code without compromising its functionality. It tends to reduce the source code complexity and improve readability, maintainability and overall software quality. Fowler et al. (1999) has identified over 70 refactoring methods specific to the design flaws existing in the code. Refactoring is complex, error-prone and tiresome in nature i.e. why it is not possible to refactor all available classes in the software and there is a need to prioritize them and apply refactoring to only highly severe classes having a large number of bad-smells. It helps in achieving better software quality within budget in timeliness manner.

By applying different refactoring methods, different impact can be seen on the software-quality. Some may improve, others may degrade and yet another may have no effect on certain quality attributes. Many a studies determine the upshot of refactoring on software quality attributes (Higo et al. 2008; Fontana et al. 2012). Also, a lot of effort is made by the researchers to apply refactoring methods to remove one or more bad-smells from the source-code (Al Dallal 2012). But, this paper addresses the issue by taking into consideration both the object-oriented characteristics and severity of presence of bad-smells in classes to prioritize them and hence achieve better software quality saving significant time and cost.

To conduct this empirical study, eight open-source medium-large sized software systems from various domains are selected for generality and wider-acceptability of the results. Object-oriented characteristics are captured by C&K metric-suite (Chidamber and Kemerer 1994). Ten most probable bad-smells are identified for each class. Then, the combined effect of both the object-oriented characteristics and severity of presence of bad-smell is considered to calculate a new metric QDIR (Quality Depreciation Index Rule) proposed by Malhotra et al. (2015) to prioritize classes for the application of refactoring methods. Following research questions are addressed in the current research:

RQ1 What percentage of classes is poorly affected by prevailing bad-smells?

RQ2 Which bad-smell predominantly harms classes in the selected software systems?

RQ3 Is Quality of Depreciation Rule (QDIR) useful in providing treatment to critically affected class?

RQ4 Do use of Quality of Depreciation Rule (QDIR) leads to a reduction in Maintenance Effort?

The rest of the paper is organized as follows: Sect. 2 reviews the related research work adopted by the study. Section 3 explains the research methodology used in this empirical study. Section 4 throws light on the results of the empirical work following which Sect. 5 explains the probable threats to validity and Sect. 6 finally concludes the empirical work and gives future work directions to researchers.

2 Related work

There are a wide number of research articles that discussed bad-smells and refactoring or effect of refactoring on object-oriented characteristics. Ouni et al. (2015) proposed a technique on the basis of chemical reaction optimizations to expose appropriate refactoring methods that maximize the no. of fixed riskiest bad-smells. Results indicate that their meta-heuristic approach is better than other popular meta-heuristics techniques. Ouni et al. (2017) also provided an approach to provide recommendations to fix prevailing bad-smells in source code and improve software quality in terms of quality attributes. While Charalampidou et al. (2017) identified possible portions of code where extract method refactoring can be applied on the basis of functional relevance criterion, Han et al. (2015) identified move method refactoring opportunities that can be simultaneously applied based on the concept of maximal independent set. Nongpong (2015) proposed a metric to detect feature envy bad-smell and apply required refactoring operation to it. Higo et al. (2008) generated couple of metrics to recognize the way in which code-clones are being refactored and came up with a tool—Aries based on their approach. Fontana et al. (2012) considered three smells—Data Class, Duplicate Code and God Class and suggested recommendations on which design debt should be first paid, as per the code smells identified. They suggested removing Duplicate Code prior to any other bad-smell and Data Class and stated that God Class bad-smells are dependent on the domain. Al Dallal (2012) used univariate logistic regression technique that empirically investigates the capabilities of 25 metrics (based on the software size, cohesion prevailing within the modules and coupling across the modules) to detect Extract-Subclass Refactoring opportunity. The results reflected a strong correlation between some of the studied metrics and recommendation of whether Extract Sub-Class refactoring is required or not. Malhotra et al. (2015) proposed a novel approach in order to prioritize the classes that demands instant treatment in terms of refactoring by giving equal importance to both design characteristics and bad-smells present in the source code. They developed a new metric to prioritize classes based on the severity of bad-smells present and evaluated it on a single, middle-sized, open source system written in Java language considering C&K metric

suite and four code bad-smells. The preliminary results show that by applying refactoring methods to only 10% of highly affected classes, 47% improvement can be achieved by saving quite a large amount of effort and cost.

This study adopts the idea of work done by Malhotra et al. (2015). In the present study, ten bad-smells have been captured for eight software systems using C&K metrics. The reason for selecting these ten bad-smells is that they are amongst the most common bad-smells under study (Vicente et al. 2018) and have available tool-support. The aim of the study is to prioritize the classes and provide refactoring methods as solutions to bad-smells to only those classes having both high severity of presence of bad-smells and poor object-oriented characteristics.

3 Research methodology

This section describes the background of the research and methodology followed in the current empirical research has been discussed elaborately. Many researchers have tried to prioritize prevailing bad-smells in the software systems for refactoring. Sae-Lim et al. (2016) collected a set of issues from developer's context using an issue-tracking system to prioritize bad smells for their removal. This could help the developers to prioritize various existing bad-smells based on their own relevant contexts. For the four selected software systems, their technique provided better ranking of bad-smells at class-level but failed for two of these systems at method-level. Also, their results did not have any statistical significance. Fontana et al. (2015) developed a tool JCodeOdour to detect bad-smells and calculate Intensity Index, an indicator of extent of bad-smell in order to prioritize removal of bad-smells. They considered six bad-smells and evaluated their technique on a large data-set reflecting satisfactory results but their experiment did not reach a proper experimental validation and lacked reliability. Choudhary and Singh (2016) provided an approach to identify class rankings based on existing bad-smells and estimated maximum bad-smell correction given minimum refactoring solution to it. They first removed the classes that are least refactored and found out the architecture relevance among the remaining classes and finally gave class rank to refactoring highly prone classes. Their approach resulted in saving considerable refactoring effort but required replication under different systems and languages for reliability. Sae-Lim et al. (2016) appropriately pointed out lack of studies with empirical evidences for prioritizing bad-smells. They worked with developers to identify factors responsible for prioritizing and refactoring the bad-smells. They identified that Module Importance and Task Relevance are the most crucial factors for prioritizing bad-smells from a developer's perspective. Oizumi et al. (2018) conducted an experiment to decipher whether a developer can point out problems in design flaws when working with stinky code. Their results showed that about 36% of the developers pointed out more design flaws in the source code when explicit reasoning of multiple bad-smells is there in contrast to single bad-smell. They identified lack of bad-smell prioritization tools and methods and found out bad-smell removal hideous and time-taking and developed a tool—Organic, that can help identify bad-smells. Eposhi et al. (2019) in contrary to most of the Java based studies, worked on two C# systems to identify if refactored portions contain more or less signs of design-flaws as compared to non-refactored portions of the system. They also investigated the effect of refactoring on two design related symptoms—density and diversity. Their results indicate that bad-smells along with internal quality attributes can be used to find design problems. But, the density and diversity of few symptoms were less in refactored classes. Vicente et al. (2018) came up with a systematic review on bad-smell and refactoring. They identified prioritization of bad-smells as one of the issue that is least prevalent in the research. Their research also indicated that many researchers agree that refactoring improved maintenance effort but there were a few studies that disagree with it and the results are contradictory. In the present study, we considered ten bad-smells and evaluated our approach on a wider dataset comprising of eight java based software systems having different sizes and domains. Evaluating results on more than one open-source software system with different domain, expertize, size, number of classes etc. helps us in achieving more general results. Figure 1 shows the research events that are being followed across the current empirical study. In the first place, we calculate QDIR (Quality Depreciation Index Rule) proposed by Malhotra et al. (2015) by means of severity of bad-smell present in the classes and its C&K metrics. Then, we refactor critical classes–classes that have greater QDIR and require refactoring to the earliest. As a result of this, we get an improved software-system with better which is rich in its design. This section is sub-divided into six sections. Section 3.1 details about the selected C&K metrics to capture object-oriented characteristics of the software

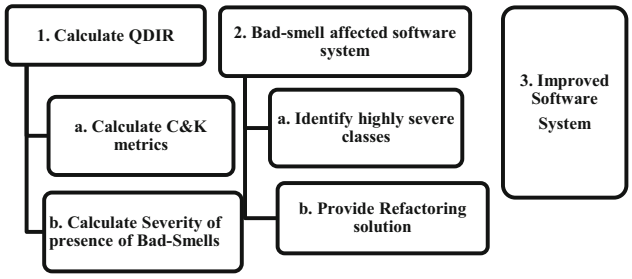


Fig. 1 Flow of research events

system. Section 3.2 explains process of empirical data collection. Section 3.3 describes the selected bad smells along with the way of detecting way. Section 3.4 discusses the refactoring methods that are selected as remedy to bad smells present in source code and Sect. 3.5 explains the way of evaluating QDIR for each class by which we can assign priorities to them. Finally, in Sect. 3.6, all the classes are prioritized based on their QDIR values into one of the four types (Critical, Bad, Mild and Low) where Critical value for a class means that it requires instant refactoring whereas Low value for a class means it require no or very little refactoring.

3.1 Object-oriented characteristic calculation

Software metric is a characteristic of the software product, process, or resource (Pressman and Maxim 2015). Object-oriented characteristics for given software are captured by software metrics. These software-metrics are indicators of object-oriented features like inheritance, coupling, abstraction level, cohesion etc. We have selected widely known (Chidamber and Kemerer 1994) that estimates the structural quality of the systems. This is because C&K metrics are being empirically calculated and properly tested by many researchers (Dagpinar et al. 2003; Elish and Elish 2009). It consists of a set of 6 metrics namely Response for a Class (RFC), Depth of Inheritance Tree (DIT), Lack of Cohesion on Methods (LCOM), Number of Children (NOC), Coupling between Objects (CBO) and Weighted Method per Class (WMC). It is worthwhile mentioning that DIT and NOC capture inheritance, CBO and RFC capture the coupling, LCOM captures cohesion and WMC captures complexity (size). These metrics have been computed using Understand metric tool (https://scitools.com/features) and are then compared to their corresponding threshold values given by Shatnawi et al. (2010). The details of C&K metrics along with their corresponding thresholds are listed in Table 1. The classes that have calculated values of metrics higher than their respective threshold values are treated as that of compromised quality.

Table 1 C&K metrics and their thresholds

C&K metric	Threshold value
CBO (coupling between objects)	13
DIT (depth of inheritance tree)	7
LCOM (lack of cohesion on methods)	7
NOC (number of children)	1
RFC (response for a class)	44
WMC (weighted methods per class)	24

3.2 Empirical data collection

This empirical study uses eight open-source software systems that are developed using java language. They are downloaded from sourceforge.net (https://sourceforge.net). Table 2 shows the selected software systems along with their size details. The analysis and review of the software systems are done using Eclipse (https://www.eclipse.org) tool. The descriptive statistics of the object oriented metrics for each of the eight selected software systems is computed. Table 3 represents the maximum, mean, minimum and standard deviation values for C&K metrics for all the eight selected software systems. The interesting observations and revelations from the descriptive statistics of C&K metrics of software systems in Table 3:

1. Inheritance in the software systems is limited as mean and mean values for DIT are very less. Maximum level of inheritance present is 2.
2. Value of LCOM for classes is quite large. Therefore, cohesion in classes is less. Maximum value of LCOM for almost all systems is 100.
3. Coupling between classes is high as mean and median of CBO is large.
4. Complexity is high among classes as WMC and RFC are quite large.

3.3 Bad-smell detection

Bad-Smells are poor design structures violating design principles that make the system complex and difficult to manage. Often people confuse them with bugs but they are not bugs and unlike bugs, they do not stop a program from executing. Instead, they if not handled properly can weaken a program and attract bugs or even much serious problems. (Fowler et al. 1999) appropriately describes a bad-smell as "a surface indication th—at usually corresponds to a deeper problem in the system". Fowler had provided a catalog of 22 bad-smells. In this study, 10 different bad-smells are

Table 2 Details of selected software systems

Dataset used	No. of classes	Lines of code
Frogger	20	100,500
Jedit-5.5.0	1336	319,404
JGraphX-3.9.3	572	70,216
Jsettlers-1.1.20	255	54,226
jVLT-1.3.3	409	23,858
jHotDraw-7.0.6	1068	97,553
Xerces-2.11.0	976	141,609
ArtOfIllusion-3.0.3	893	119,203

Table 3 Descriptive statistics of C&K metrics of selected software systems

Software system	Metric value	Min	Max	Mean	Standard deviation
Frogger	CBO	1	13	5.4	3.69
	DIT	0	2	1.15	0.93
	LCOM	0	100	27.91	39.16
	NOC	0	10	0.5	2.24
	RFC	16	33	18.8	3.79
	WMC	1	17	4.8	3.68
jedit5.5.0	CBO	0	189	9.9	12.23
	DIT	0	3	0.23	0.49
	LCOM	0	100	27.46	34.18
	NOC	0	38	0.22	1.68
	RFC	0	987	134.26	16.08
	WMC	0	351	6.1	16.08
JGraphX-3.9.3	CBO	0	80	7.94	9.54
	DIT	0	3	0.19	0.47
	LCOM	0	100	27.91	39.16
	NOC	0	13	0.21	1.08
	RFC	12	924	70.41	182.72
	WMC	0	924	70.41	182.72
Jsettlers 1.1.20	CBO	0	173	9.98	19.95
	DIT	0	3	0.57	0.67
	LCOM	0	100	44.25	31.41
	NOC	0	84	0.48	5.3
	RFC	13	752	117.56	204.21
	WMC	0	159	11.31	19.49
jVLT-1.3.3	CBO	0	126	9.99	12.63
	DIT	0	4	0.45	0.72
	LCOM	0	100	39.88	34.95
	NOC	0	16	0.34	1.41
	RFC	0	905	156.27	282.03
	WMC	0	52	4.88	5.38
JHotDraw 7.0.6	CBO	0	98	11.15	11.32
	DIT	0	6	0.56	1.07
	LCOM	0	100	32.46	37.94
	NOC	0	28	0.33	1.56
	RFC	0	871	85.56	183.93
	WMC	0	80	6.65	9.25
Xerces 2.11.10	CBO	0	115	9.18	11.66
	DIT	0	5	1.02	1.6
	LCOM	0	100	50.41	38.48
	NOC	0	52	0.41	2.36
	RFC	12	849	61.5	88.91
	WMC	0	129	10.22	14.12
ArtOfIllusion 3.0.3	CBO	0	134	10.78	12.66
	DIT	0	5	0.45	0.69
	LCOM	0	100	33.72	35.07
	NOC	23	55	0.26	2
	RFC	0	880	33.73	67.32
	WMC	0	108	8.15	11.03

selected and they are identified for each class of all the studied software systems. Table 4 lists all the selected bad-smells along with their refactoring solutions. There are various bad-smell detection tools present online like PMD, inFusion, JDeodorant, JSpirit, CCFinder, Robusta and TrueRefactor. Fernandes et al. (2016) have provided a case study of bad-smell detection tools. They identified that out of 89 bad-smell detection tools identified in literature, only 29 were available for download and 56 listed tools were designed for Java followed by 16, 15 and 6 tools for C, C++ and Python respectively. Also, Robusta is quite good for exception handling bad-smells among other options like SpotBugs and PMD etc. (Coelho 2018). The selection of common bad-smells and respective detection tools was not easy. Three eclipse plugins for java namely—Jdeodorant (http://jdeodorant.com), JSpirit (http://sites.google.com/site/santiagoavidal/projects/jsirit) and Robusta metric tool for metric collection (https://marketplace.eclipse.org/content/robusta-eclipse-plugin) are used to detect bad-smells from all the classes of the selected software systems. The reason behind selecting these ten bad-smells and the tools is that they are very common among software practitioners and have available tool support unlike many other bad-smells (Coelho 2018; Fernandes et al. 2016; Hamid et al. 2013).

3.4 Refactoring

As defined by Fowler et al. (1999), Refactoring is a popular and effective technique that improves a software's internal structure making it much easier to read, understand, and manage without changing how a software its behavior functionally.

Below is the description of refactoring methods used to remove selected bad-smells as given in Table 4:

1. *Move Method* If a method from class A uses some method from the class B excessively, place the method from class B in class A. Now, either completely remove this method from a second class or provide it refers to its new location if required.
2. *Extract Class* For a given class that has so many responsibilities to handle, move all the desired fields and methods in a newly created class.
3. *Extract Method* In a long method where some part of the code can be grouped together, move that part of the code in a newly created method and make a reference for this new method in the old existing method.
4. *Replace Type Code with State/Strategy* If there is coded-type which negatively impacts the functionality but child-class can't be used to avoid it, replace type code with a state object. Also, plug-in another state-object if it is required to replace a field-value with its type-code.
5. *Replace Inheritance with Delegation* Put base-class' object in a new field and delegate methods to base-class' objects instead of having a generalization relationship between class and get-rid-of inheritance.

Table 4 Selected bad-smells and their respective refactoring solutions

Bad-smell captured	Description of bad-smell	Refactoring method
Long Method	It is a method that has got so much to do instead of performing a single responsibility well	Extract Method
God Class	It indicates a class in a system that is quite large and majorly controls all of its intelligence/workings	Extract Class
Feature Envy	It occurs when a method relies on another object's data more than that of its own	Move Method
Type Checking	It occurs when multiple methods are assigned the responsibility to perform a single functionality instead of the need for only one method	Replace Type Code with State/Strategy
Nested Try Statement	It is a chain of a large number of try statements just like an if-else ladder	Extract Method
Empty Catch Block	It is a catch block without body showing negligence of the developer in handling the respective error condition	Re-throw with Exception
Shotgun Surgery	It occurs when non class members extensively uses method of a given class	Move Method
Refused Parent Bequest	It happens when a child class either can't inherit base class' data or it uses a small subset of methods defined in its base class and fields of its base class	Replace Inheritance with Delegation
Brain Method	It is a method where the major functionality of a class resides	Move Method
Intensive Coupling	It occurs when a client-method tends to communicate too much with one or more classes it depends upon	Move Method

6. *Re-throw with Exception* Re-throw an exception in catch block in spite of keeping it empty or printing an error statement.

3.5 Quality Depreciation Index Rule (QDIR)

Quality Depreciation Index Rule (QDIR) helps us in prioritizing the classes in order of severity of presence of bad-smells. Equal importance is given to both bad-smells and C&K metrics as both are equally important factors in determining the design flaws in the classes. As a result, equal weight is given to BoB (Base of Bad-Smell) and BoM (Base of Metric) i.e. measurable effects of design metrics and presence of bad-smells. Detailed rationale in giving equal weightage to these two factors is provided by Malhotra et al. (2015). It can be calculated as mentioned in following steps:

(1) *Calculation of Base of Bad-Smell (BoB)* We have detected ten different bad-smells in each class and given equal Smell Weightage (SW) to all the bad-smell which is 0.10. Our intent in giving equal SW of 0.10 to each bad-smell is that we want to prioritize refactoring of those classes which are affected by maximum number of bad-smells. This research works on the idea of 80:20 principal that suggests that 80% of the overall quality of the system can be improved by providing refactoring to 20% classes in the system. In the preliminary study, only first four bad-smells given in Table 4 were selected. BoB for a class is calculated by taking an average of SW of all the detected bad-smells in that class to that of the ten selected bad-smells as shown in Eq. (1).

$$Base\ of\ Bad\ Smell\ (BoB) = \frac{1}{10}\sum_{i=1}^{10} SW_i \tag{1}$$

(2) *Calculation of Base of Metric (BoM)* All the selected six Chidamber and Kemerer (C&K) metrics are first identified for each class of all the software systems under study as discussed in Sect. 3 and then compared to their corresponding threshold values. This computed value is then divided by its corresponding threshold value to arrive at Metric Value (MV) as shown in Eq. (2).

$$Metric\ Value = \frac{Calculated\ Metric}{Thresold\ Metric} \tag{2}$$

After this, BoM is calculated by capturing average values of MV for all six selected metrics as given in Eq. (3).

$$Base\ of\ Metric = \frac{1}{6}\sum_{j=1}^{6} MV_j \tag{3}$$

(3) *Calculation of QDIR* Finally, QDIR is calculated using (1) and (3) as given in Eq. (4).

$$Quality\ Depreciation\ Index\ Rule\ (QDIR) = \frac{1}{2}BoB + \frac{1}{2}BoM. \tag{4}$$

3.5.1 Illustration by Example

Take for example, one of the classes VFSBrowser of the project jedit5.5.0.

To calculate QDIR, first we need to find BoM and BoB. BoM for any class can again be found out by getting average of the MVs of that class. MVs for the class VFSBrowser for LCOM, DIT, CBO, NOC, RFC and WMC are 94, 0, 114, 0, 759 and 62 respectively.

$$Base\ of\ Metric\ (BoM) = (94+0+114+0+759+62)/6 = 7.00$$

VFSBrowser is affected by six bad-smells-God Class, Type Checking, Long Method, Feature Envy, Dummy Handler and Exception thrown from finally block. So,

$$Base\ of\ Bad\ Smell\ (BoM) = (0.1+0.1+0.1+0.1+0.1+0.1)/10 = 0.06$$

QDIR is now calculated by giving equal weightage to both BoM and BoB.

$$Quality\ Depreciation\ Index\ Rule\ (QDIR) = (7.00+0.06)/2 = 3.53$$

The value of QDIR $3.53 > 1.84$ (see Table 7). Hence, the class VFSBrowser falls under Critical category and it is required to prioritize applying refactoring on this class as soon as possible.

The approach that is followed for calculation of QDIR is illustrated in Fig. 2.

3.6 Prioritization of classes

This study aims to save maintenance cost and time by refactoring only highly severe classes. QDIR metric assigns priority to classes so that refactoring to only a small subset of software system's classes can be provided. This can be achieved by arranging classes in non-increasing order of their QDIR values and selecting classes for different priority levels according to Table 5. Also, a higher value of QDIR for a class indicates critical flaws in the design and higher priority is assigned to it whereas a class having a lower value of QDIR signifies better design and assigns less priority to it.

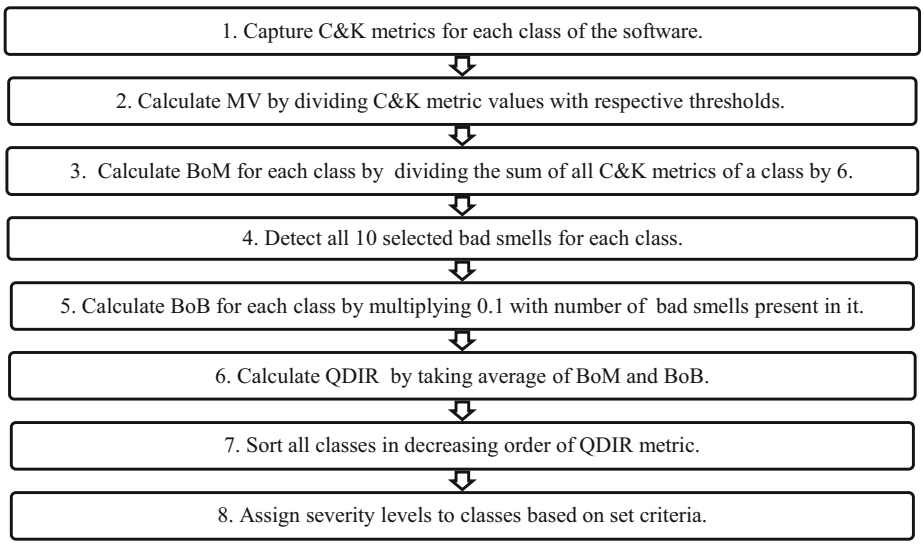


Fig. 2 Algorithm to prioritize classes based on QDIR

Table 5 Severity levels of classes

Severity level	Description	Range (%)
Critical	Classes that are highly severe and require refactoring at the earliest.	10
High	Classes that are comparatively less severe and require refactoring but not immediately.	25
Mild	Classes that require refactoring but not instantly.	25
Low	Classes that are least severe having no harmful bad-smells.	40

4 Results and discussion

This section is dedicated to discussing the results and answering the research questions that were raised before conduction the empirical research.

RQ1 What percentage of classes is poorly affected by prevailing bad-smells?

Figure 3 represents the division of presence of bad-smells in selected software systems. It is found that on an average, 30% of the total number of classes in the software-system are poorly affected by bad-smells. As a considerable amount of classes have a bad influence of the bad-smells, removing them from the software-system becomes really important for its better design and lesser possibilities of errors. This revelation supports our attempt to remove bad-smells from the software systems by means of their prioritization considering the bad-smells' severity.

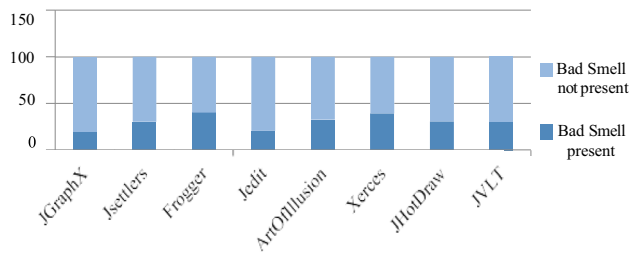


Fig. 3 Percentages of bad-smells in software systems

RQ2 Which bad-smell predominantly harms classes in the selected software systems?

Figure 4 depicts the presence of ten selected bad-smells across eight studied software systems. As bad-smells are present in all the software systems in a significant amount, immediate remedial actions are required.

(1) Bad-Smells in decreasing order of their presence across selected software systems: Long Method, Brain Method, Refused Parent Bequest, God Class, Intensive Coupling, Shotgun Surgery, Empty Catch Block, Type Checking, Feature Envy and Nested Try Block.

(2) It can be noted that Long Method is present in a maximum number of classes. Apart from it, Brain Method and Refused Parent Bequest are also highly present. These bad-smells require immediate attention and need to be removed as early as possible.

(3) Nested Try Block and Feature Envy are present in the least number of classes among the selected eight software systems so they can be given least importance.

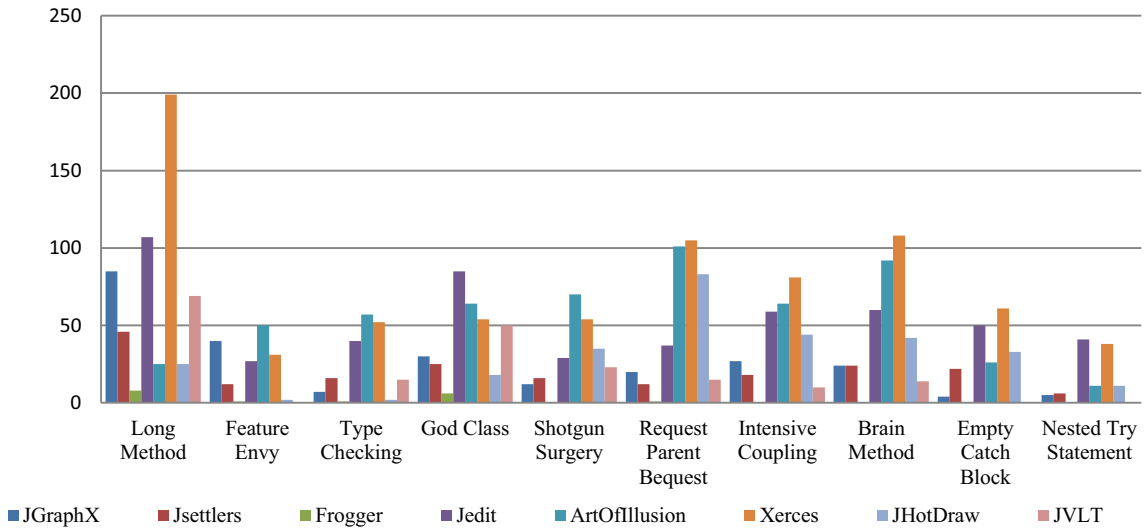


Fig. 4 Divisions of bad-smells across selected software systems

(4) It can also be observed that degree of presence of bad-smell increases as complexity and size of a software system increases, Xerces-2.11.10k and ArtofIllusion-3.0.3 account for the largest amount of presence of bad-smells which corresponds to 39.04% and 32.59% respectively.

RQ3 Is of Quality of Depreciation Rule (QDIR) useful in providing treatment to critically affected class?

Once bad-smells are detected and C&K metrics are captured, QDIR is calculated. Both bad-smells and C&K metrics are assigned equal importance for calculation of QDIR; i.e. BoB and BoM are given equal weights in the formula for calculating QDIR. A higher value of QDIR metric for a class is an indicator of high severity and reflects the need for instant refactoring method application to that class. On the other hand, a lower value of QDIR metric means less severity and no sooner or completely no need of refactoring. Correspondingly, four severity levels have been identified (see Table 5) after arranging QDIR values for a system in decreasing order. It is not possible to arrive at a common range for eight systems to fit percentage requirement of 10, 25, 25 and 40 as per Table 5 because all of them are having a completely different domain and the difference in sizes. Table 6 provides ranges of four severity levels and distribution of classes in each those levels. Based on the results, following observations can be made:

1. QDIR helps in reducing the number of refactoring operations and restricting it to only a small portion of classes by putting highly severe classes into Critical Level which is only 10% of the total available classes (see Table 6).
2. On an average, classes with QDIR metric value ≥ 1.75 fall under Critical severity level (see Table 7).

RQ4 Do use of Quality of Depreciation Rule (QDIR) leads to a reduction in Maintenance Effort?

QDIR classifies classes in one of the four severity-levels so that refactoring methods can be applied to only critically affected classes to save Maintenance-Effort and time of the developer. Table 8 represents the Effort Estimation (EE) and Effort Saved (ES) for all the selected eight software systems and Table 9 provides average Effort Estimation (EE) and Effort Saved (ES) for Critical and High severity levels. CE and HE in Tables 8 and 9 refer to Critically Effected and Highly Effected classes respectively. Let, for example Jedit-5.5.0 software system to calculate EE and ES for CE and HE severity levels. Jedit-5.5.0 has a total of 1336 classes, where 135 and 334 are classes which are CE and HE respectively. So, $\mathrm{EE(CE)} = \frac{135}{1336} \times 100 = 10.10\%$, $\mathrm{EE(CE)} = \frac{134}{1336} \times 100 = 25\%$. And, $\mathrm{ES(CE)} = 100 - 10.10 = 89.90\%$ $\mathrm{ES(HE)} = 100 - 25 = 75\%$.

Based on the results, following observations can be made:

1. It can be observed that on an average, EE (CE) and EE (HE) classes are 9.20% and 25.45% respectively (see Table 8).
2. Also, on an average, ES (CE) and ES (HE) is 90.79% and 74.56% respectively (see Table 9).

Table 6 Range of different severity levels for selected software systems and division of classes into them

Software system	Severity level	Range	No. of classes
Frogger	Critical	$\geq$ 1.28	1
	High	1.27–1.06	6
	Mild	1.05–1.0	9
	Low	$\leq$ 1.0	4
Jedit-5.5.0	Critical	$\geq$ 1.84	135
	High	0.83–1.83	334
	Mild	0.11–0.82	332
	Low	$\leq$ 0.10	535
JGraphX-3.9.3	Critical	$\geq$ 1.50	56
	High	0.59–0.63	143
	Mild	0.62–0.16	140
	Low	$\leq$ 0.15	233
JSettlers-1.1.20	Critical	$\geq$ 2.11	25
	High	0.95–2.10	63
	Mild	0.65–0.94	60
	Low	$\leq$ 0.64	107
jVLT-1.3.3	Critical	$\geq$ 2.45	38
	High	1.05–2.44	99
	Mild	0.50–1.04	108
	Low	$\leq$ 0.49	154
JHotDraw-7.0.6	Critical	$\geq$ 1.70	108
	High	0.95–1.69	266
	Mild	0.14–0.94	264
	Low	$\leq$ 0.13	430
Xerces-2.11.10	Critical	$\geq$ 1.70	95
	High	1.15–1.69	240
	Mild	0.66–1.14	247
	Low	$\leq$ 0.65	394
ArtOfIllusion-3.0.3	Critical	$\geq$ 1.40	88
	High	0.85–1.39	224
	Mild	0.11–0.84	229
	Low	$\leq$ 0.10	352

Table 7 Average value of QDIR for Critically Effected (CE) classes

Severity level	Average QDIR value
Critical	$\geq$ 1.75

5 Threats to validity

By capturing object oriented characteristics and identifying bad-smells in the classes upon eight selected software systems, we tried to arrive at the best possible generalized results that are unbiased and true to our knowledge. But,

Table 8 Effort estimation of critically and highly effected classes

Software system	EE (CE)	ES (CE)	EE (HE)	ES (HE)
Frogger	5	95	30	70
JEdit-5.5.0	10.10	89.90	25	75
JSettlers-1.1.20	9.80	90.20	24.70	75.30
JGraphX-3.9.3	9.79	90.21	25	75
JVLT-1.3.3	9.29	90.71	24.20	75.80
JHotDraw-7.06	10.11	89.89	25	75
Xerces-2.11.10	9.73	90.27	24.59	75.41
ArtOfIllusion-3.0.3	9.85	90.15	25.08	74.92

Table 9 Average effort estimation of critically and highly effected classes

EE (CE)	ES (CE)	EE (HE)	ES (HE)
9.20	90.79	25.45	74.56

the results of this empirical study would not be complete lacking the discussion of construct, internal and external validity that are quite evident across empirical studies. Construct validity is how much accurately the object-oriented characteristics are calculated from the selected software systems. To reduce this threat to the maximum, we have used a professional industrial tool-Understand to calculate these object-oriented characteristics. Internal validity can be described as the extent to which conclusions can be arrived at about the consequences of certain debts prevailing in design of the software system. The key factors that can go against the internal validity are size of the software systems and transcription errors while evaluating various results for the systems. We have tried to minimize this threat quite successfully by investigating ten bad smells and their effect on software quality over eight software systems with medium-large sizes and reviewing the results multiple times. The results drawn from the study are consistent and unbiased in nature. External validity can be regarded as the extent to which the results of an empirical study can be obtained in generalized form. To reduce this threat, we have selected a wider set of medium to large sized open-source software systems written in Java language across different domains. For even more generality across different object-oriented languages, we are further planning to repeat this empirical work on languages like C#, C++ and Python so that developers working on languages other than Java can also be benefitted from the results of this empirical study.

6 Conclusions and future work

In this empirical study, we have focused on prioritizing classes in the software-system so that refactoring can be applied to only a smaller subset of severely affected classes instead of applying it on the whole set of available classes in order to effectively exploit limited time and budget. For this, we assigned one of the four severity levels—Critical, High, Mild and Low to provide refactoring to classes belonging to Critical severity level. Critical, High, Mild and Low severity levels are assigned 10, 25, 25 and 40% of classes from the entire range of classes. QDIR is calculated by giving equal weightage to the presence of prevailing bad-smells and C&K characteristics for all the classes. Ranges for four severity levels is then decided and classes falling in critical severity level are given instant refactoring solution. Also, C&K metrics are used to capture object-oriented characteristics of classes present in selected software systems and bad-smells are identified in every class for these software systems using available tool-support.

The results of the empirical study reveal that 30% classes, on an average, are poorly affected by bad-smells with Long Method and Brain Method being present in a maximum number of classes and Feature Envy and Nested Try Block being present in the least number of classes. Classes with QDIR $\geq$ 1.75 falls under Critical severity and require instant refactoring. This helps in focusing on bad-smells that are highly prominent and giving less attention to bad-smells that are hardly present. Also, giving refactoring solution to 10% and 25% of critically & highly affected classes, an average of 90.79% and 74.56% of effort can be saved resulting in improved software quality under cost and time constraints.

The current empirical study used a set of eight medium-large java based systems having different domains for the generality of results. But for overall generality and wider acceptability of results across different languages as well, a future direction would be analysing the results on other object-oriented languages like C++, Python, and C#. Also, there are over 70 bad-smells that are present in literature and this research captured ten common bad-smells whereas there are other bad-smells that can also be explored by researchers. We finally hope that significant quality of the software-system can be improved by including refactoring as a solution to only highly severe classes that have instant refactoring requirements saving considerable maintenance effort and time.

References

Abreu FB, Carapuça R (1994) Object-oriented software engineering: measuring and controlling the development process. In: 4th international conference of software quality. https://doi.org/10.5281/zenodo.1217609

Al Dallal J (2012) Constructing models for predicting extract subclass refactoring opportunities using object-oriented quality metrics. Inf Softw Technol 54(10):1125–1141. https://doi.org/10.1016/j.infsof.2012.04.004

Charalampidou S, Ampatzoglou A, Chatzigeorgiou A, Gkortzis A, Avgeriou P (2017) Identifying extract method refactoring opportunities based on functional relevance. IEEE Trans Softw Eng 43(10):954–974. https://doi.org/10.1109/TSE.2016.2645572

Chidamber SR, Kemerer CF (1994) A metrics suite for object oriented design. IEEE Trans Softw Eng 20(6):476–493. https://doi.org/10.1109/32.295895

Choudhary A, Singh P (2016) Minimizing refactoring effort through prioritization of classes based on historical, architectural and code smell information. In: Proceedings of the 1st international workshop on technical debt analytics (TDA 2016)

Coelho R (2018) Towards a catalogue of Java exception handling bad smells and refactorings. In: Proceedings of the 25th conference on pattern languages of programs (PLoP'18). The Hillside Group, USA, no 7, pp 1–17

Dagpinar M, Jahnke JH, Canada BC (2003) Predicting maintainability with object-oriented metrics—an empirical comparison. In: Proceeding of the 10th working conference on reverse engineering, 2003 (WCRE 2003), Victoria, British Columbia, Canada, pp 155–164. https://doi.org/10.1109/WCRE.2003.1287246

Elish MO, Elish KO (2009) Application of TreeNet in predicting object-oriented software maintainability: a comparative study. In: Proceedings of the 2009 European conference on software maintenance and reengineering (CSMR'09). IEEE Computer Society, USA, pp 69–78. https://doi.org/10.1109/CSMR.2009.57

Eposhi A, Oizumi W, Garcia A, Sousa L, Oliveira R, Oliveira A (2019) Removal of design problems through refactorings: are we looking at the right symptoms? In: 2019 IEEE/ACM 27th international conference on program comprehension (ICPC), pp 148–153. https://doi.org/10.1109/ICPC.2019.00032

Fernandes E, Oliveira J, Vale G, Paiva T, Figueiredo E (2016) A review-based comparative study of bad smell detection tools, (Dcc). In: Proceedings of the 20th international conference on evaluation and assessment in software engineering (EASE'16). Association for Computing Machinery, New York, NY, USA, no 18, pp 1–12. https://doi.org/10.1145/2915970.2915984

Fontana FA, Ferme V, Spinelli S (2012) Investigating the impact of code smells debt on quality code evaluation. In: Proceedings of the 3rd international workshop on managing technical debt, (MTD'20), Zurich, pp 15–22. https://doi.org/10.1109/MTD.2012.6225993

Fontana FA, Ferme V, Zanoni M, Roveda R (2015) Towards a prioritization of code debt: a code smell intensity index. In: Proceedings of the 7th international workshop on managing technical debt (MTD), Bremen, pp 16–24. https://doi.org/10.1109/MTD.2015.7332620

Fowler M, Beck K, Brant J, Opdyke W, Roberts D (1999) Refactoring improving the design of existing code. Addison-Wesley, Boston

Hamid A, Ilyas M, Hummayun M, Nawaz A (2013) A comparative study on code smell detection tools. Int J Adv Sci Technol. https://doi.org/10.14257/ijast.2013.60.03

Han AR, Bae DH, Cha S (2015) An efficient approach to identify multiple and independent Move Method refactoring candidates. Inf Softw Technol. https://doi.org/10.1016/j.infsof.2014.10.007

Higo Y, Kusumoto S, Inoue K (2008) A metric-based approach to identifying refactoring opportunities for merging code clones in a Java software system. J Softw Maint Evol 20(6):435–461. https://doi.org/10.1002/smr.394

Lorenz M, Kidd J (1994) Object-oriented software metrics: a practical guide. PTR Prentice Hall, Upper Saddle River

Malhotra R, Chug A, Khosla P (2015) Prioritization of classes for refactoring: a step towards improvement in software quality. In: *Proceedings of the third international symposium on women in computing and informatics (WCI'15)*. Association for Computing Machinery, New York, NY, USA, pp 228–234. https://doi.org/10.1145/2791405.2791463

Nongpong K (2015) Feature envy factor: a metric for automatic feature envy detection. In: Proceedings of the 7th international conference on knowledge and smart technology, (KST'15). Chonburi, pp 7–12. https://doi.org/10.1109/KST.2015.7051460

Oizumi W, Sousa L, Oliveira A, Garcia A, Agbachi AB, Oliveira R, Lucena C (2018) On the identification of design problems in stinky code: experiences and tool support. J Braz Comput Soc 24(13):1–30. https://doi.org/10.1186/s13173-018-0078-y

Ouni A, Kessentini M, Bechikh S, Sahraoui H (2015) Prioritizing code-smells correction tasks using chemical reaction optimization. Softw Qual J 23:323–361. https://doi.org/10.1007/s11219-014-9233-7

Ouni A, Kessentini M, Cinnéide MÓ, Sahraoui H, Deb K, Inoue K (2017) MORE: a multi-objective refactoring recommendation approach to introducing design patterns and fixing code smells. J Softw Evol Process. https://doi.org/10.1002/smr.1843

Pressman RS, Maxim BR (2015) Software engineering: a practitioner's approach, 8th edn. ACM SIGSOFT software engineering notes. https://doi.org/10.1145/1226816.1226822

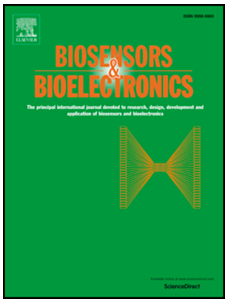
Sae-Lim N, Hayashi S, Saeki M (2016) Context-based code smells prioritization for prefactoring. In: Proceedings of the XP2017 scientific workshops (XP'17). Association for Computing Machinery, New York, NY, USA, no 3, pp 1–5. https://doi.org/10.1109/ICPC.2016.7503705

Shatnawi R, Li W, Swain J, Newman T (2010) Finding software metrics threshold values using ROC curves. J Softw Maint Evol 22(1):1–16. https://doi.org/10.1002/smr.404

Vicente E, Sobrinho DP, De Lucia A, Maia MDA (2018) A systematic literature review on bad smells—5 W's: which, when, what, who, where. IEEE Trans Softw Eng. https://doi.org/10.1109/TSE.2018.2880977

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

Journal Pre-proof



Exploring *Providencia rettgeri* for application to eco-friendly paper based microbial fuel cell

Sharda Nara, Rahul Kandpal, Veeru Jaiswal, Shine Augustine, Srishti Wahie, Jai Gopal Sharma, Ryusuke Takeuchi, Shigeori Takenaka, Bansi Dhar Malhotra

PII: S0956-5663(20)30318-3

DOI: https://doi.org/10.1016/j.bios.2020.112323

Reference: BIOS 112323

To appear in: *Biosensors and Bioelectronics*

Received Date: 3 April 2020

Revised Date: 21 May 2020

Accepted Date: 22 May 2020

Please cite this article as: Nara, S., Kandpal, R., Jaiswal, V., Augustine, S., Wahie, S., Sharma, J.G., Takeuchi, R., Takenaka, S., Malhotra, B.D., Exploring *Providencia rettgeri* for application to eco-friendly paper based microbial fuel cell, *Biosensors and Bioelectronics* (2020), doi: https://doi.org/10.1016/j.bios.2020.112323.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2020 Published by Elsevier B.V.

Journal Pre-proof

CRediT author statement

Sharda Nara: Conceptualization, Methodology, Investigation, Writing- Original draft preparation. Rahul Kandpal: Conceptualization, Methodology, Investigation. Veeru Jaiswal: Visualization, Investigation. Shine Augustine: Visualization, Validation Srishti Wahie: Visualization, Investigation. Jai Gopal Sharma: Supervision, Ryusuke Takeuchi: Visualization, Investigation. Shigeori Takenaka: Supervision. Bansi Dhar Malhotra: Supervision, Visualization, Writing- Reviewing and Editing

Journal Pre-proof

Journal Pre-proof

Exploring *Providencia rettgeri* for application to eco-friendly paper based microbial fuel cell

Sharda Nara,[a*] Rahul Kandpal,[a*] Veeru Jaiswal,[a] Shine Augustine,[a] Srishti Wahie,[a] Jai Gopal Sharma,[a] Ryusuke Takeuchi[b], Shigeori Takenaka,[b] Bansi Dhar Malhotra[a$]

[a] Nanobioelectronics Laboratory, Department of Biotechnology, Delhi Technological University, Shahbad Daulatpur, Delhi, India

[b] Research Center for Biomicrosensing Technology (RCBT), Kyushu Institute of Technology, Fukuoka 804-8550 Japan

[*]These authors contributed equally to this work.

[$]Corresponding author: Prof. Bansi D Malhotra

E-mail: bansi.malhotra@gmail.com

Journal Pre-proof

Journal Pre-proof

Abstract

We report results of the studies relating to improved stability (40 days) of small sized microbial fuel cell (MFC) fabricated using agarose embedded paper-based proton exchange membrane. A fermentative bacterium *Providencia rettgeri* was isolated from rotten potato slurry and identified by 16S rRNA sequencing. The electroactivity of the bacteria was monitored *via* chronoamperometric and cyclic voltammetric studies using a three-electrode system which indicated the presence of bacterial redox mediator. The Matrix-Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) and UV-Vis absorption spectroscopy provided the evidence that *Providencia rettgeri* synthesized folate (vitamin B_9) during fermentation that was found to act for the first time as a redox mediator in an MFC. The paper based designed MFC fed with *Providencia rettgeri* yielded open circuit voltage of 787.9 mV with power and current density of 5.02 W/m^3 and 11.26 A/m^3, respectively when measured across 10 kΩ. The microbial re-chargeable battery comprising of an assembly of parallelly aligned four units of MFCs when connected in series generated 1.5 V that was used for powering a red-light emitting device (LED).

Keywords: Microbial fuel cell; Electroactive bacteria; *Providencia rettgeri*; Fermentative bacteria; Folate

Journal Pre-proof

Journal Pre-proof

1. Introduction:

There is increased interest towards the development of microbial fuel cells (MFCs) based on organic waste management due to exponential increase in human population and effluents from food processing units (Xin et al., 2019). With increasing urbanization in developing countries, potato-based food processing industries have been growing rapidly, generating organic residue of nearly 10 tons/day (dos Santos et al., 2012). India alone followed by China and Russia, produce 34,391 thousand tons of potatoes in a year, creating organic waste of nearly 12-20 % of the total production of the potatoes (Singh and Rana, 2014). In this context, MFCs have emerged as a promising solution for potato waste management with conversion of stored chemical energy into electrical energy *via* microbial electrochemical reactions, representing an environment friendly and cost-effective alternative to ecosystem balance (Xia et al., 2019).

It has been reported that harvesting of energy from the waste wherein the biomass is mainly generated *via* photosynthesis such as potato waste depends on the efficiency of the microbial community which predominantly comprises of the electroactive bacteria like *Geobacter, Shewanella, Klebsiella, Bacillus subtilis* and fermentative bacteria like *Bacteroids, Veillonella, Enterococcus* etc (Mayer 1998; Jia et al., 2013; Tian et al., 2017). Compared to electroactive bacteria, it has been found that in starch fuelled MFC, fermentative bacteria like *Bacteroids* can be utilized to significantly contribute to electricity generation (Jia et al., 2013). However, the role of fermentative bacteria and the released compound(s) have not yet been un-ravelled and is still a subject matter of research.

Another important aspects on which the efficiency of electricity generation depends are the materials (Perveen et al., 2018; ul Haque et al., 2018; Haque et al., 2019) and design of the MFC (Logan et al., 2006; Shahi et al., 2017). Nowadays, paper based MFCs, have drawn

Journal Pre-proof

much attention due to their ease of fabrication, fast response, low cost, portability, eco-friendly and disposable nature (Xu et al., 2016; Chouler et al., 2018). However, the short shelf life of such an MFC is currently a handicap towards continuous electricity generation. We have constructed an MFC with a new type of proton exchange membrane (PEM) that can be used to continuously provide high voltage for about 40 days with manual feeding at every seventh day and hence it can be called as rechargeable MFC (Jiang et al., 2020). Conventionally used PEM eg. salt bridge is known to yield low power output due to its high internal resistance (Shahi et al., 2017) whereas use of Nafion membrane, is known to result in increased cost (Fraiwan et al., 2013). On the other hand, a major problem currently associated with the paper-based PEM is the rapid loss of electricity resulting in significant current decrease after about 15 minutes. This has been ascribed to the infusion of catholyte and anolyte due to paper mediated fluid permeation (Fraiwan et al., 2013; Lee et al., 2016). In the present study, the agarose embedded paper was used as a PEM to resolve the problems associated with the use of paper and salt bridge. Graphite coated sheets were employed as anode/cathode and the paper chambers sealed with hydrophobic coating of the wax used as anodic/cathodic chambers. Cultures of four different bacterial isolates (*Klebsiella pneumoniae*, *Providencia rettgeri*, *Bacillus subtilis 168_1 and Bacillus subtilis 168_2*) from rotten potato slurry were employed as anolyte.

The *Providencia rettgeri* has been identified for the first time as a new fermentative bacterial species that can be used to generate electricity.

2. Materials and Methods:

2.1 Isolation of bacteria and 16S rRNA Sequencing:

The four morphologically distinct bacterial colonies grown on nutrient agar plates were selected and identified using 16S rRNA gene cloning, sequencing and analysis using MEGA blast (see Supplementary Information).

2.2 Construction of MFC:

2.2.1 Assembly of MFC:

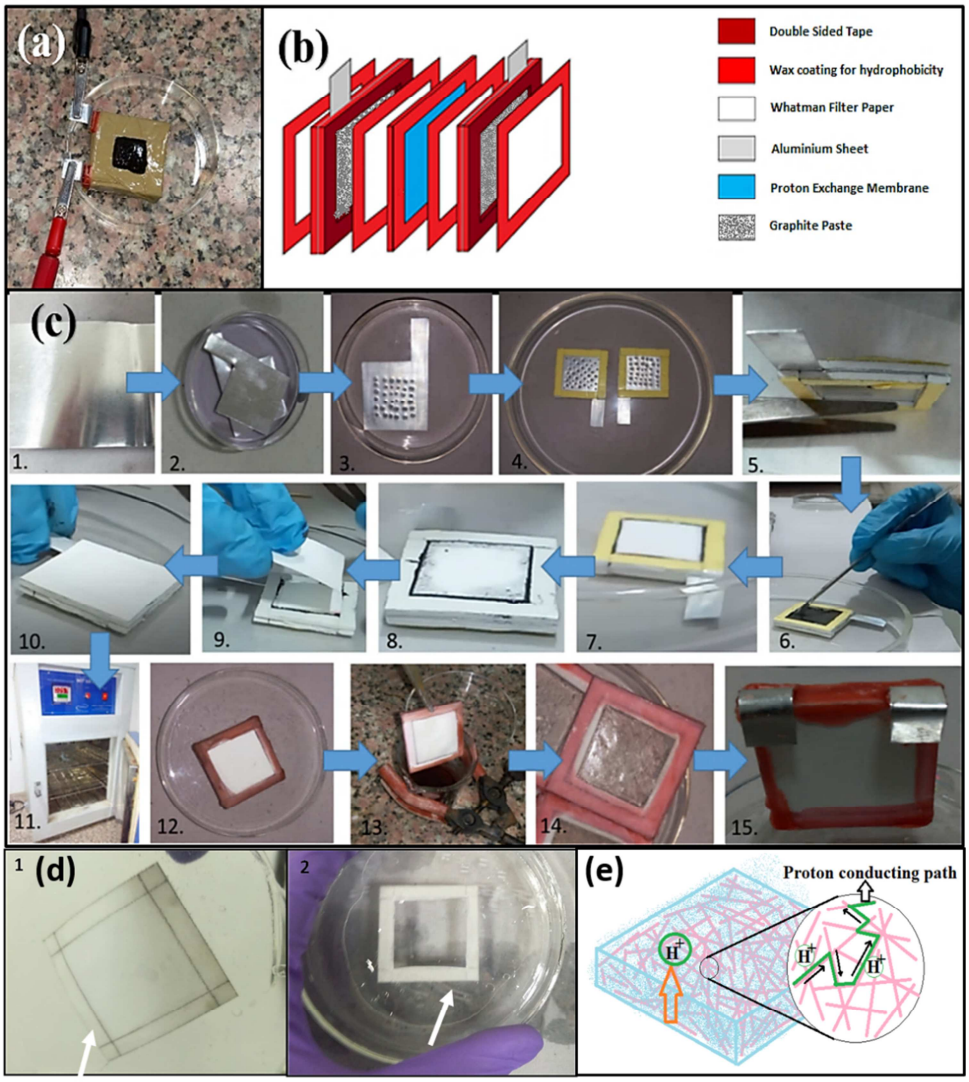


Figure 1. Photograph of fully assembled MFC (a), schematic of different layers used in the MFC (b), pictorial representation of various steps involved in fabrication of small size (3×3 cm^2) MFC (c), agarose embedded paper-based PEM (d1), transparent agarose-based PEM (d2) and schematic representation of proton conducting pathway (e).

A new design was employed to assemble the double chamber MFC. The laboratory grade aluminum (Al) sheet (thickness: 0.44 mm) was cut into the desired size (4×4 cm^2). The

Journal Pre-proof

aluminum sheet was perforated to make the contact of anolyte and catholyte with PEM. The double-sided tape was applied around each side of the Al piece covering 1 cm width of the four sides. As a result, a well (3×3 cm^2) was created that was subsequently filled with graphite paste which in turn was covered with Whatman filter paper as shown in Figure 1(a-c). The preparation of cathode was similarly carried out. Both the electrodes were dried in a hot air oven (Chem Tech, India) at 80°C for 7 days. To prevent leakage, all sides of the assembled fuel cell were sealed with molten wax. The fabricated fuel cell had different vertically stacked layers on both sides, having total thickness of 10 mm. The first outermost layer encompassed the paper. The second layer was made of graphite deposited at the base of the perforated Al sheet, graphite served as a scaffold for inhabitation of microbes and it also acted as an anode. The Al sheet acted as electrons collector and its perforated surface provided the passage for protons towards catholyte via PEM. The third layer made of PEM was sandwiched between anodic and cathodic surface. The cathodic chamber was similarly fabricated (Figure 1). The working area of the fabricated MFC was 3×3 cm^2.

Two different MFCs were configured based on two distinct PEMs. The configured MFC was sterilized by UV irradiation for about 1 h prior to addition of the bacterial anolyte.

2.2.2 Fabrication of PEM:

Two different types of PEMs were constructed. In the first design, PEM was prepared using 5% agarose (HiMedia, India) and 2.5% KCl (SRL, India). The required amount of agarose and salt were mixed and heated until the solution became completely transparent. Subsequently, 2 ml of the prepared solution was applied between cathodic and anodic chamber that allowed it to be sandwiched between the two chambers. Another PEM was prepared with additional use of Whatman filter paper grade 1 (Merck, India) having pore size of 11 µm. 2 ml of solution containing agarose (5%) and KCl (2.5%) was poured on the back

Journal Pre-proof

side of glass petrie dish after which the Whatman filter paper was applied and agarose solution was again poured onto upper surface of the paper (Figure 1d,e). The agarose embedding was accomplished by pressing it against another petrie dish and the temperature was simultaneously lowered by circulating water around the PEM.

2.2.3 Operation of MFC:

The MFCs were fed with 2 ml of broth containing four identified bacteria having bacterial density of 10^6 $CFUml^{-1}$ as anolytes while a phosphate buffered saline ferrocyanide-ferricyanide solution (50 mM, pH 7.0) was used as the catholyte. The MFC fed with 2 ml of broth without bacterial culture was taken as a control, other conditions were kept same. The volume of the anolyte/catholyte (2 ml) was taken according to the capacity of the prepared MFC. The current and voltage were continuously monitored using a digital multimeter (Keithley, Tektronix, USA). The power and current densities were estimated using anode surface area as 3×3 cm^2 and anodic chamber volume as 2 ml. The polarization curve of the whole MFC was recorded at varying external load (1000 kΩ to 10 kΩ) for the MFC. The voltage was measured with step-wise increase in external resistance after a steady state condition for 10 minutes while the current was measured using Ohms law. The internal resistance was obtained by the slope of the polarization curve.

2.3 Electron microscopy:

The morphology of the bacterial biofilm was investigated by scanning electron microscopy (SEM, Zeiss EVO 18, USA) following the standard protocol of sample preparation.

2.4 Set-up of fed-batch reactor:

The sodium acetate-based growth medium was fed in a fed-batch reactor, 1ml of the *Providencia rettgeri* broth was added as inoculum per 20 ml of the growth media (see Supplementary Information).

Journal Pre-proof

148

149 **2.5 Chronoamperometry and Cyclic voltammetry:**

150 Autolab Potentiostat (Metrohm, Netherlands) was used to conduct the chronoamperometric
151 study at a constant potential (0.2 V) with reference to Ag/AgCl, the value of current was
152 recorded every 60s. The cyclic voltammetric (CV) measurements were conducted at day 2, 4,
153 6, 8, 10, 20 at a scan rate of 1 mVs^{-1}. After each CV measurement, chronoamperometric
154 measurements were recorded. The total charge transferred was evaluated by integrating the
155 area under curve obtained in the chronoamperometric study.

156 **2.6 UV-Vis spectroscopy:**

157 The UV-Vis spectroscopic studies were conducted using a UV/Visible spectrophotometer,
158 Lambda 950, (Perkin Elmer, USA) to evaluate the presence of redox mediators in anodic
159 solution. The anodic solution was centrifuged at 8000 rpm at 4°C for 20 minutes. Then, the
160 supernatant was collected and further purified by solvent extraction method (Kumar et al.,
161 2009). Briefly, chloroform and methanol were used in a 2:1 ratio (volume) and an equal
162 amount of the centrifuged supernatant was added dropwise. The solution was mixed and left
163 undisturbed until the separation of its phases (aqueous and organic) occurred. The purified
164 organic phase was collected gently using an auto pipette and filter sterilized using 0.2 µm
165 filters.

166 **2.7 MALDI-TOF mass spectrometry:**

167 Mass spectrometric measurements were carried out by BRUKER Microflex (Bruker
168 Daltonics, Billerica, MA) mass spectrometer using 2,5-dihydroxybenzoic acid (DHB) as
169 matrix. The data was obtained in the range, m/z=0 to 100. 1 µL of the sample was placed at
170 the parafilm after which 1µL of DHB was added and mixed with a pipette and the mixture
171 was placed onto MSP 96 target ground steel plate. After drying under reduced pressure, the
172 plate was used to set the instrument for the MALDI-TOF MS analysis. The data was

collected from 0 to 800 m/z at laser frequency of 60 Hz, with 500 laser shots collected at random-walk shot pattern equipped with a N_2 laser operated at 337 nm. The data was analysed using the Compass 1.3 software suite (FlexControl 3.3, and FlexAnalysis 3.3, Bruker Daltonics). The resulting mass spectrum was binned at 0.5 m/z increments and manually inspected for masses of interest. Masses were visualized using false colouring on an optical image of the original sample plate.

3. Results and discussion:

3.1 16S rRNA sequencing:

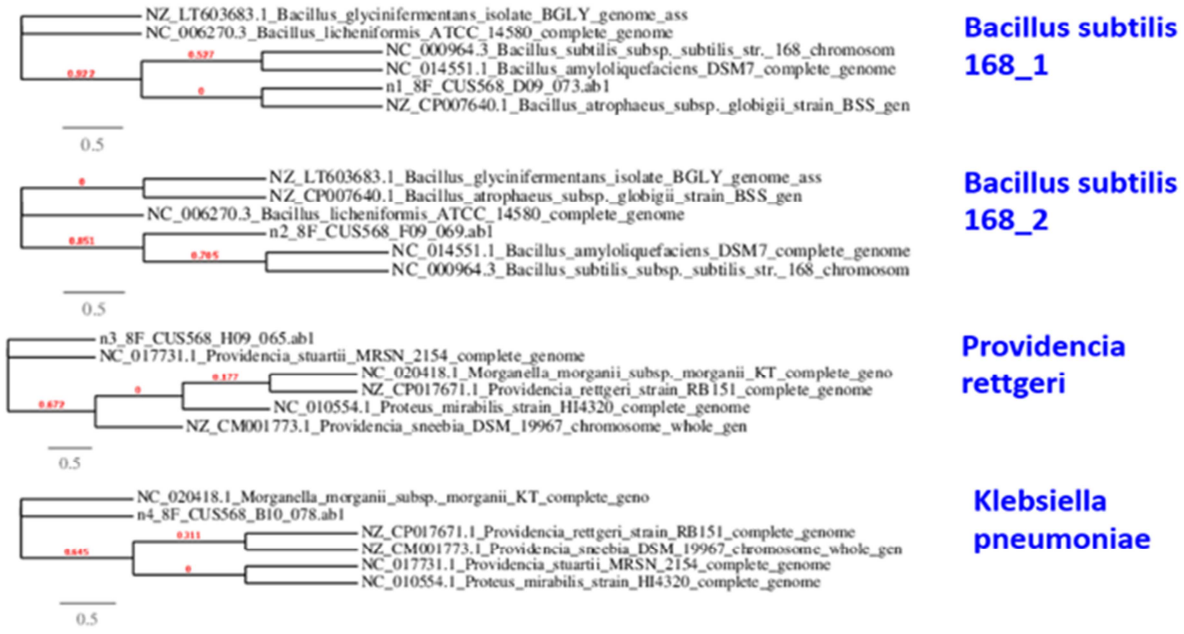


Figure 2. Phylogenetic tree of four bacterial isolates constructed using neighbor joining method of 16S rRNA gene sequencing. Phylogeny confirmed the identification of *Providencia rettgeri*, *Klebsiella pneumoniae* and two different strains of *Bacillus subtilis 168*.

All the four types of isolated bacteria (Figure S.1, Figure S.2) belonged to *Enterobacteriaceae* family as revealed by phylogenetic analysis (Figure 2). Amplification of the extracted DNA sequence indicated the presence of phylogenetic groups belonging to the *Klebsiella pneumoniae*, *Providencia rettgeri* and *Bacillus subtilis 168* (Table S.1 to Table

S.4). Two bacterial species identified as *Bacillus subtilis 168* had different morphology of colonies when grown on NB agar plates (Figure S.1(i)) revealing that two of the bacterial isolates belonged to two different strains of *Bacillus subtilis 168*, hereafter designated as *Bacillus subtilis 168_1* (Figure S.1(i1) and *Bacillus subtilis 168_2* (Figure S.1(i2)).

3.2 Chronoamperometric biofilm growth of *Providencia rettgeri*:

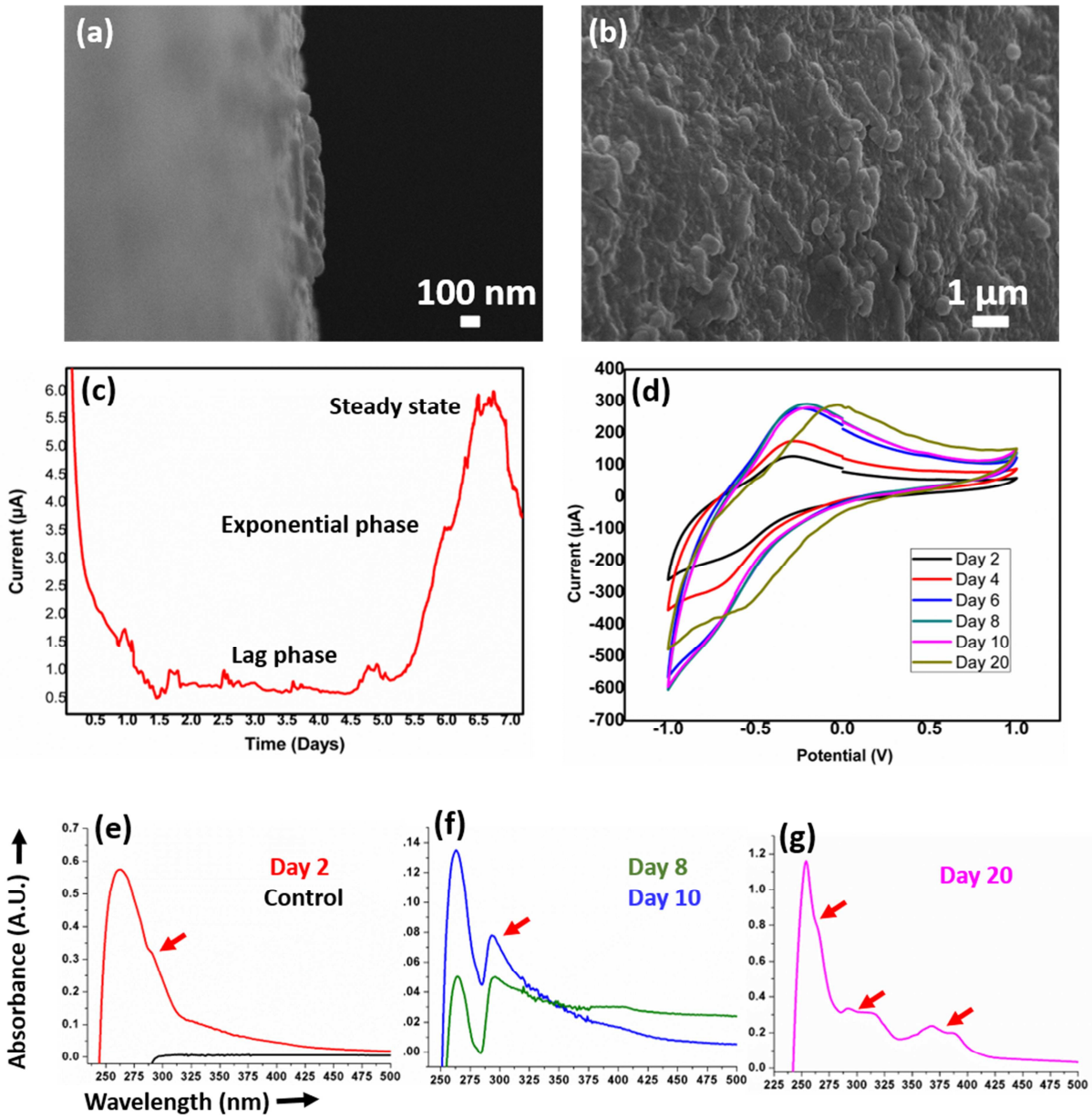


Figure 3. SEM micrographs (a,b) depicting biofilm formation on electrode (graphite) surface at day 7. Chronoamperometric results of fed-batch reactor-based biofilm growth, the oxidative current reaches a steady state after three media change (c). Cyclic voltammogram

Journal Pre-proof

of the *Providencia rettgeri* indicating the shift of redox potential at 20th day, observed peak shift may shows the changes in redox active substances (d). UV-Vis spectrum at different growth phase of the bacteria indicating release of mediator in the media (e-g).

The results of chronoamperometric studies revealed that the magnitude of current gradually increased to 6 µA after about one week of incubation, with current density of 3 µAcm^{-2} (Figure 3c).

There are three known modes of bacterial electron transfer viz. a) direct membrane associated redox active proteins (Uria et al., 2017), b) conductive pili called nanowires as reported in *Geobacter species* (Holmes et al., 2016), and/or c) transfer of electrons *via* redox shuttle mediators e.g. phenazines, flavins as studied in the *Pseudomonas aeruginosa* (Rabaey et al., 2005; Jo et al., 2017) and *Shewanella oneidensis* (Kotloski and Gralnick, 2013), respectively. The mechanism involving direct electron transfer from the bacterial outer membrane proteins (cytochromes) to the anode surface perhaps did not form the biofilm and instead the monolayer of bacteria participated in the transfer of electrons (Torres et al., 2010). Another type of electron transfer mechanism has been reported to involve conductive extracellular matrix secreted by bacteria wherein bacterial extracellular extension involving nanopilli or nanofibers efficiently transfer the electrons that are known to exhibit the highest current density as estimated by the Nernest and Monod equations (Torres et al., 2010). *Providencia rettgeri* adhered and formed the biofilm at the electrode surface. However, it may be noted that we did not observe any nanowire formation (Figure 3 a,b). Thus, the possibility of electron transfer by direct cellular protein and/or nanopilli was almost negligible. The mechanism based on electron transport via mediators could perhaps be ascribed to the shuttling mechanism wherein electrons can be transported by soluble electron mediators (in reduced form) to the anode surface following discharge of electrons to the surface and moving back again to bacteria in their oxidized form (Von Canstein et al., 2008). In this mode of electron transfer, current density was constrained due to diffusion of the redox mediators and their periodic use for gain and loss of electrons by bacteria and anode surface, respectively (Torres et al., 2010). The current density in this study was determined to be as 3 µAcm^{-2} indicating the presence of weak redox mediator-based shuttling mechanism in *Providencia rettgeri.* Consequently, current density did not reach the highest value.

3.3 Cyclic voltammetry of *Providencia rettgeri* growth:

Journal Pre-proof

Cyclic voltammograms recorded as a function of time revealed that enhancement in current was due to bacterial growth and slight shift of the oxidation peak from -0.31 to -0.21V was observed from day 2 to day 10 (Figure 3d). Interestingly, the major shift of the redox peak was observed at day 20 with the oxidation peak seen at -0.032V (Figure 3d). This shift in the redox potential can be attributed to different reduced/oxidized state of redox mediator involved in electrons shuttling mechanism (Liu et al., 2013). Herein, it appears that folate molecules acted as soluble redox shuttles. We believe that along with the inter-molecular electron transfer, these molecules perhaps also participated in intramolecular electron transfer to stabilize themselves in the solution resulting in the evolution of different intermediate states and perhaps also different conformations of the molecules. Thus, the shift in redox potential is possibly attributed to the observed changes in the folate molecules while transferring the electrons from bacteria to anodic surface.

3.4 UV-Vis spectroscopy studies:

The UV-Vis spectra of the supernatant extracted from *Providencia rettgeri* culture grown in three-electrode setup regulated with a potentiostat on day 2, revealed the absorption peak at 263 nm with a small bulge around 293 nm (Figure 3e) corresponding to the absorption of folate molecules (Matias et al., 2014). On day 8 and 10, two distinct absorption peaks were observed at 263 and 293 nm (Figure 3f) whereas an additional peak was found at 365 nm on day 20 (Figure 3g). The complete disappearance of peaks was not found indicating that conversion of the molecular conformation was a reversible reaction. The appearance of multiple peaks with increase in bacterial incubation, further indicated the presence of redox mediators at different oxidized or reduced states (Liu et al., 2013; Matias et al., 2014). It may also be possible that the stereoscopic structure of the molecule got modified so that electron

transfer to the anodic surface became thermodynamic feasible as reported in case of phenothiazine redox mediators (Liu et al., 2013).

3.5 MALDI-TOF mass spectrometry:

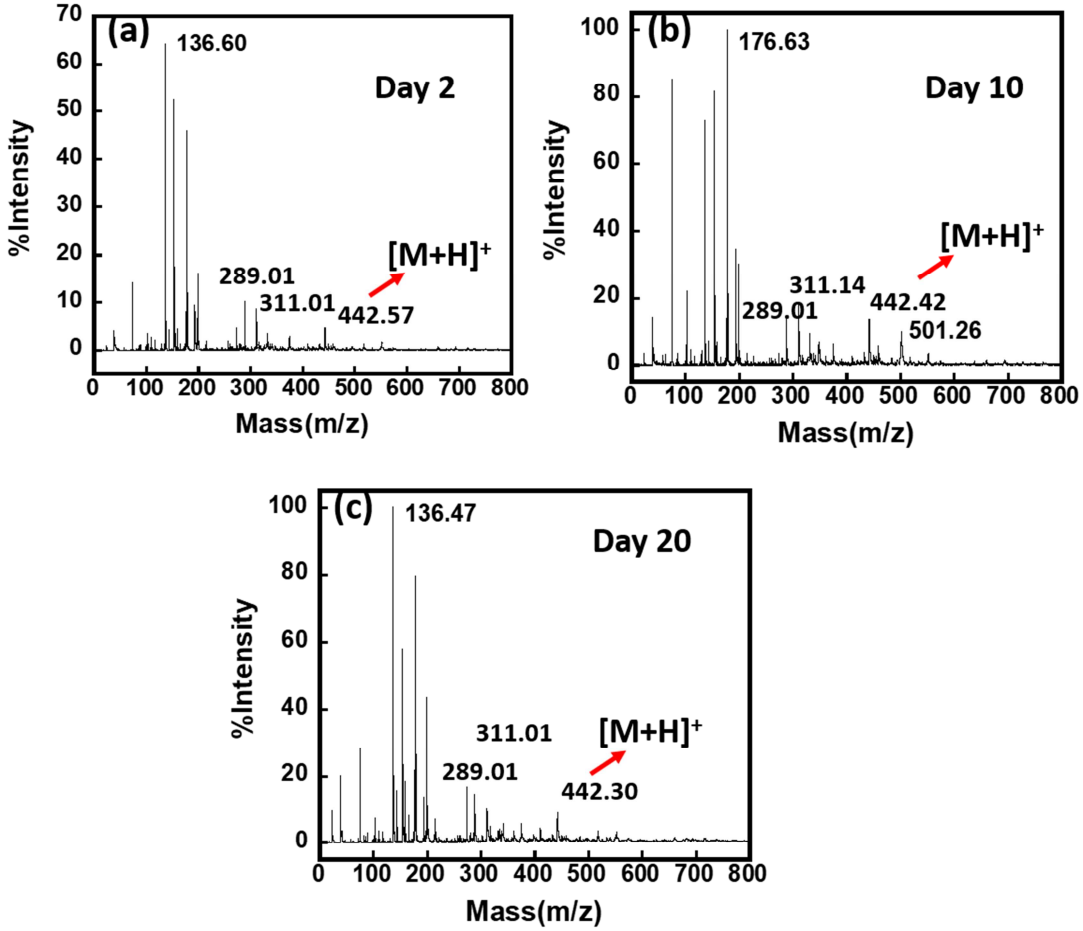


Figure 4. Mass spectra of sample extract of fed batch reactor bacterial culture of day 2 (a), day 10 (b) and day 20 (c) with DHB as matrix. In all samples (a, b and c), mass spectral peaks (m/z) appeared at 442 indicating the molecular ion mass peak of folate $(M+H)^+$.

MALDI-TOF spectra of 2, 10 and 20-day old bacterial culture samples were carried out. MALDI-TOF spectra (Figure 4) demonstrated that the molecular ion mass peak at m/z= 442.57, 442.42 and 442.30, respectively corresponded to the presence of protonated molecular ionic species $[M+H]^+$ of folate (vitamin B_9) (Cha and Kim, 2003).

Journal Pre-proof

The folate molecules had an aroyl group attached to the molecule of glutamic acid (Neta et al., 2014). The mass peaks seen at 311.01 (Figure 4a), 311.14 (Figure 4b) and 311.01 (Figure 4c) and 289.01 (Figure 4a,b,c) were due to characteristic fragmentation peaks of folate (Bhandari et al., 2013; Lu et al., 2017) while the peaks seen at 136.60 (Figure 4a), 176.63 (Figure 4b) and 136.47 (Figure 4c) were due to matrix ions (DHB) (Becker et al., 2014). Thus, the appearance of mass peak around 442 in MALDI-TOF spectrum indicated that folate molecules secreted by *Providencia rettgeri* in the acetate media. It may be noted that the folate has recently been identified as a redox regulator in human cells with the capability of scavenging free radicals (Lai et al., 2017). In the present study, folate molecules perhaps acted as an electron acceptor in the absence of oxygen and intramolecular electron transfer (Kim et al., 1991) may perhaps be responsible for the heterogenous electron transfer in the different growth phase.

There are two possibilities, either in the process of adapting for the anaerobic condition, *Providencia rettgeri* consumes acetate through anaerobic oxidation and has secreted folate that acted as a mediator for electron transfer or synthesized folate accumulated within the bacterial cell that was released due to bacterial cell lysis upon the death of the bacteria that acted as a redox mediator.

The high fermentation rate (0.03 ppm/min) (see Figure S.3 and Supplementary Information) revealed that bacteria was capable of rapidly degrading the substrate. However, the low (0.03 %) coulombic efficiency (CE) (Table S.5, equation S.1) showed that generated electrons could not be efficiently transferred to the anodic surface. The *Providencia rettgeri* biofilm (Figure 3 a,b) perhaps hindered electron transfer to the anodic surface resulting in low CE that could be increased using the co-culture of anode respiring bacteria like *Geobacter sulfurreducens* (Nevin et al., 2008).

Moreover, study of the single bacterial species may significantly decipher the electro-kinetics occurring in MFC using kitchen waste wherein a microbial consortium along with fermentative bacteria degraded the organic waste into simple substance (Uria et al., 2017).

3.6 Characteristics of *Providencia rettgeri* based MFCs:

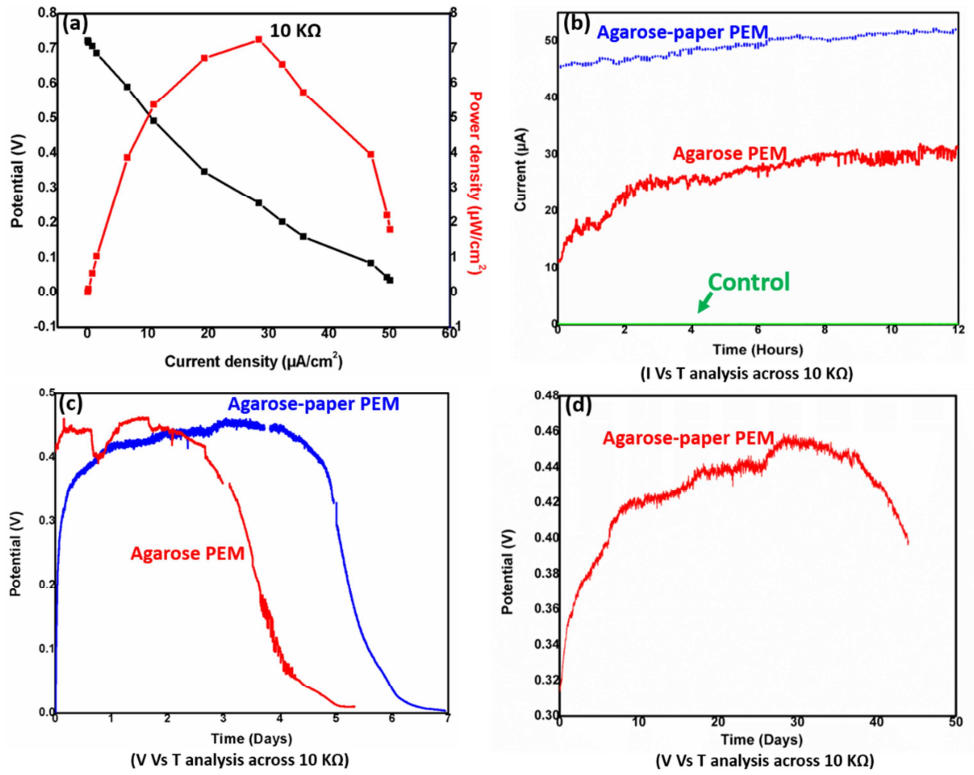


Figure 5. Polarization curve (a) indicating power output of the MFC as a function of current. Current and Voltage output of MFCs (b,c) fabricated using two different PEMs indicating high current and voltage stability for MFC with agarose embedded paper-based PEM (agarose-paper PEM) as relative to agarose-based PEM (agarose PEM) while no measurable current was observed in control (b).Voltage Vs time graph of MFC across 10 KΩ (d), showing stable voltage output for 40 days with the inoculation of fresh anolyte and catholyte.

3.6.1 Polarization curve:

The maximum power attained in the configured MFC was evaluated by the polarization curve. The electrical output could be tuned with varying external resistance, and finally when

Journal Pre-proof

the applied external resistance became equal to the internal resistance of the MFC leading to maximum power output (Cheng et al., 2006).

The power density was found to be maximum at the external resistance of 10 KΩ (Figure 5a). Subsequently, all current and voltage measurements were conducted using external resistance of 10 KΩ.

3.6.2 Performance of MFC depending on agarose membrane and agarose embedded paper-based PEM:

The bacterial culture was perhaps entirely responsible for the current generation as negligible current was observed in case of control in comparison to the MFC fed with bacterial culture (Figure 5b). The designed agarose embedded paper based-MFC exhibited the voltage output for 40 days (Figure 5d). Herein, the instantaneous value of the current and voltage were obtained (Figure 5b-d) while the traditional MFCs required prolonged bacterial incubation for bacterial acclimation at the anode surface which stretched the start-up time. The wicking ability of the paper perhaps facilitated the formation of a hydrophilic reservoir that helped in faster absorption of anolyte and endorsed the bacterial attachment on the graphite surface (Fraiwan et al., 2013). The average current output (50 µA, Figure 5b) obtained in agarose embedded paper-based PEM was nearly two folds higher as compared to the agarose-based PEM (27 µA, Figure 5b). Moreover, average voltage output (0.43 V) in the agarose embedded paper-based PEM could be retained for longer time period (Figure 5c,d; Figure S.5) while the voltage started to drop after 3 days in agarose-based PEM (Figure 5c). Interestingly, the current (59 μAcm^{-2}) and power (27 μWcm^{-2}) output were found to be higher than previously reported paper based MFCs (1.2 μWcm^{-2}) (Fraiwan and Choi 2016) as the use of only paper perhaps allowed the backflow of protons while the salt bridge in agarose embedded paper-based PEM provided the favourable condition of proton transfer (Figure S.4, Figure S.5). It appeared that the combination of paper and agarose perhaps yielded a paradigm shift in the case of low cost MFC.

Journal Pre-proof

The high internal resistance offered by the paper was compensated with the wicking ability of the incorporated paper. The diffusion of electron acceptor from paper-based PEM is known to result in decreased coulombic efficiency (Fraiwan et al., 2013; Lee et al., 2016) of the system while the agarose embedded-paper based PEM perhaps prevented the diffusion of the electron acceptor. The present configuration can be operated in the batch mode (Fed-batch mode). To obtain the enhanced overall system voltage, multiple MFCs units could be stacked in series (Aelterman et al., 2006; Fraiwan and Choi 2016).

3.6.3 Performance of bacterial isolates (*Providencia rettgeri, Klebsiella pneumoniae* and two different strains of *Bacillus subtilis 168 (*Type 1 and Type 2) based MFCs:

The comparison of voltage and current output obtained due to different bacterial isolates are summarised in Table 1.

Table 1. Characteristics of the fabricated MFC fed with different cultures.

	Mixed Culture	***Bacillus subtilis 168* (type 1 strain)**	***Bacillus subtilis 168* (type 2 strain)**	***Klebsiella pneumoniae***	***Providencia rettgeri***
Peak voltage	483.7 mV	428.4 mV	407.2 mV	409.1 mV	430.6 mV
Peak current	53.6 µA	37.88 µA	36.32 µA	37.39 µA	50.36 µA
Volumetric current density	14.89 Am^{-3}	10.57 Am^{-3}	10.43 Am^{-3}	10.50 Am^{-3}	11.26 Am^{3}
Volumetric power density	6.84 Wm^{-3}	4.55 Wm^{-3}	4.49 Wm^{-3}	4.49 Wm^{-3}	5.02 Wm^{-3}
OCV	794.3 mV	536.7 mV	528.9 mV	579.7 mV	787.9 mV

It can be seen that the MFC fed with pure isolates of *Providencia rettgeri* provided OCV value of 787.9 mV was higher than the OCV value obtained in the pure isolates of the *Bacillus subtilis 168* and *Klebsiella pneumoniae*, confirming the electrochemical potential of *Providencia rettgeri*.

Journal Pre-proof

3.6.4 Powering the LED by *Providencia rettgeri* based MFCs:

MFCs used for the experiments were assembled to realize the application as proof of the concept. Multiple units of MFCs fed with *Providencia rettgeri* were arranged as a mini power source. Four MFCs were arranged in parallel to lower down the internal resistance while the four such sets were arranged in series to generate potential (1.5 V) that power the red-light emitting device (LED) for 4 days (Video S1, Figure S.6). The designed MFC was operated in fed batch mode, periodically the spent media and catholyte were replaced by fresh media and catholyte in anodic and cathodic chambers, respectively to recharge the microbial battery. The whole assembly exhibited stability for approximately 40 days with continuous voltage output of 1.5 V.

4. Conclusions:

The *Providencia rettgeri*, a fermenting bacterium, has been identified as a new electrochemically active bacterial species that transferred the metabolically derived electrons to the electrode surface indirectly through folate that perhaps acted as redox mediator. The use of graphite paste coating as anode surface and incorporation of agarose embedded paper-based PEM, enhanced the MFC stability to about 40 days. The fabricated MFC is a promising potential for utilization as a power source for miniaturised bioelectronic devices (Nasar and Perveen, 2019) including biosensors (Pandey et al., 2018; Singh et al., 2020) and for application as water quality monitors.

Notes

The authors declare no competing financial interest.

Acknowledgements:

The authors thank Prof. Yogesh Singh, Vice Chancellor, Delhi Technological University for providing necessary research facility. Sharda Nara acknowledges University Grant

Commission, Govt. of India for providing Dr. D.S. Kothari Postdoctoral fellowship (F. 4-2/2006 (BSR)/BL/16-17/0144). Bansi Dhar Malhotra thanks the Science & Engineering Research Board (SERB), Government of India for the award of a Distinguished Fellowship (SB/DF/011/2019).

References:

Aelterman, P., Rabaey, K., Pham, H.T., Boon, N., Verstraete, W., 2006. Environ. Sci. Technol. 40(10), 3388-3394.
Becker, L., Carré, V., Poutaraud, A., Merdinoglu, D., Chaimbault, P., 2014. Molecules 19(7), 10587-10600.
Bhandari, D., Kertesz, V., Van Berkel, G.J., 2013. Rapid Commun. Mass Spectrom. 27(1), 163-168.
Cha, S.-W., Kim, H.-J., 2003. Bull. Korean Chem. Soc. 24(9), 1308-1312.
Cheng, S., Liu, H., Logan, B.E., 2006. Environ. Sci. Technol. 40(7), 2426-2432.
Chouler, J., Cruz-Izquierdo, Á., Rengaraj, S., Scott, J.L., Di Lorenzo, M., 2018. Biosens. Bioelectron. 102, 49-56.
dos Santos, T.C., Gomes, D.P.P., Bonomo, R.C.F., Franco, M., 2012. Food Chem. 133(4), 1299-1304.
Fraiwan, A., Choi, S., 2016. Biosens. Bioelectron. 83, 27-32.
Fraiwan, A., Mukherjee, S., Sundermier, S., Lee, H.-S., Choi, S., 2013. Biosens. Bioelectron. 49, 410-414.
Haque, S.U., Nasar, A., Asiri, A.M., 2019. Int. J. Hydrogen Energy 44(14), 7308-7319.
Holmes, D.E., Dang, Y., Walker, D.J., Lovley, D.R., 2016. Microbial genomics 2(8).
Jia, J., Tang, Y., Liu, B., Wu, D., Ren, N., Xing, D., 2013. Bioresour. Technol. 144, 94-99.
Jiang, M., Xu, T., Chen, S., 2020. Appl. Energy 260, 114336.
Jo, J., Cortez, K.L., Cornell, W.C., Price-Whelan, A., Dietrich, L.E., 2017. Elife 6, e30205.
Kim, S.T., Heelis, P.F., Okamura, T., Hirata, Y., Mataga, N., Sancar, A., 1991. Biochemistry 30(47), 11262-11270.
Kotloski, N.J., Gralnick, J.A., 2013. MBio 4(1), e00553-00512.
Kumar, T., Singh, M., Purohit, H., Kalia, V., 2009. J. Appl. Microbiol. 106(6), 2017-2023.
Lai, K.-G., Chen, C.-F., Ho, C.-T., Liu, J.-J., Liu, T.-Z., Chern, C.-L., 2017. Tumor Biology 39(6), 1010428317702649.
Lee, S.H., Ban, J.Y., Oh, C.-H., Park, H.-K., Choi, S., 2016. Scientific reports 6, 28588.
Liu, X.-W., Sun, X.-F., Chen, J.-J., Huang, Y.-X., Xie, J.-F., Li, W.-W., Sheng, G.-P., Zhang, Y.-Y., Zhao, F., Lu, R., 2013. Scientific reports 3, 1616.
Logan, B.E., Hamelers, B., Rozendal, R., Schröder, U., Keller, J., Freguia, S., Aelterman, P., Verstraete, W., Rabaey, K., 2006. Environ. Sci. Technol. 40(17), 5181-5192.
Lu, Q., Hu, Y., Chen, J., Jin, S., 2017. Anal. Chem. 89(16), 8238-8243.
Matias, R., Ribeiro, P., Sarraguça, M., Lopes, J., 2014. Anal. Methods 6(9), 3065-3071.
Mayer, F., 1998. Polym. Degrad. Stab. 59(1-3), 231-235.
Nasar, A., Perveen, R., 2019. Int. J. Hydrogen Energy. 44(29), 14467-15702.
Neta, P., Farahani, M., Simón□Manso, Y., Liang, Y., Yang, X., Stein, S.E., 2014. Rapid Commun. Mass Spectrom. 28(23), 2645-2660.

Journal Pre-proof

Nevin, K.P., Richter, H., Covalla, S., Johnson, J., Woodard, T., Orloff, A., Jia, H., Zhang, M., Lovley, D., 2008. Environ. Microbiol. 10(10), 2505-2514.
Pandey, C.M., Augustine, S., Kumar, S., Kumar, S., Nara, S., Srivastava, S., Malhotra, B.D., 2018. Biotechnology journal 13(1), 1700047.
Perveen, R., Nasar, A., Asiri, A.M., 2018. Int. J. Biol. Macromol. 106, 755-762.
Rabaey, K., Boon, N., Höfte, M., Verstraete, W., 2005. Environ. Sci. Technol. 39(9), 3401-3408.
Shahi, A., Rai, B., Singh, R., 2017. Resource-Efficient Technologies 3(1), 78-81.
Singh, B., Rana, R.K., 2014. History of potato and its emerging problems in India. Souvenir-National Seminar on Emerging Problems of Potato, pp. 1-2.
Singh, N., Ali, M.A., Rai, P., Ghori, I., Sharma, A., Malhotra, B., John, R., 2020. Lab on a Chip. 20(4),760-777.
Tian, Y., Mei, X., Liang, Q., Wu, D., Ren, N., Xing, D., 2017. RSC Advances 7(14), 8376-8380.
Torres, C.I., Marcus, A.K., Lee, H.-S., Parameswaran, P., Krajmalnik-Brown, R., Rittmann, B.E., 2010. FEMS microbiology reviews 34(1), 3-17.
ul Haque, S., Nasar, A., Asiri, A.M., 2018. Chem. Phys. Lett. 692, 277-284.
Uria, N., Ferrera, I., Mas, J., 2017. BMC microbiology 17(1), 208.
Von Canstein, H., Ogawa, J., Shimizu, S., Lloyd, J.R., 2008. Appl. Environ. Microbiol. 74(3), 615-623.
Xia, T., Zhang, X., Wang, H., Zhang, Y., Gao, Y., Bian, C., Wang, X., Xu, P., 2019. Bioresour. Technol. 284, 72-79.
Xin, X., Chen, B.-Y., Hong, J., 2019. Bioresour. Technol. 289,121652.
Xu, Z., Liu, Y., Williams, I., Li, Y., Qian, F., Zhang, H., Cai, D., Wang, L., Li, B., 2016. Biosens. Bioelectron. 85, 232-239.

379

Journal Pre-proof

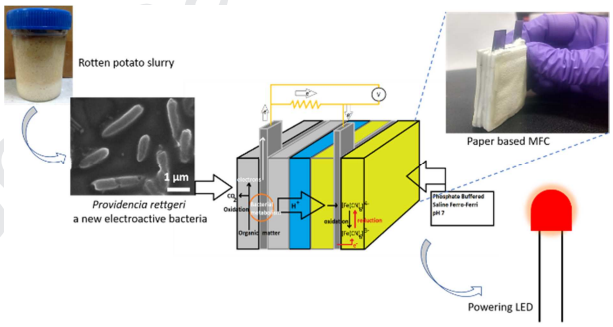
Journal Pre-proof

Highlights

- *Providencia rettgeri* identified as a new fermenting bacterial species that generates the electricity.
- Folate (vitamin B_9) observed to release by *Providencia rettgeri* during fermentation which possibly act as redox mediator.
- *Providencia rettgeri* fed Microbial fuel cell (MFC) shows an open circuit voltage of 787.9 mV with the power and current density of 5.02 Wm^{-3} and 11.26 Am^{-3} respectively.
- Environment friendly paper based MFC with agarose embedded paper-based proton exchange membrane, shows stability for 40 days.

Journal Pre-proof

Table of Content



Isolation and characterization of *Providencia rettgeri* as a new electroactive bacterium for application to paper based Microbial Fuel Cell.

Journal Pre-proof

Declaration of interests

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

'Declarations of interest: none'

Journal Pre-proof

J. Inst. Eng. India Ser. B
https://doi.org/10.1007/s40031-020-00450-z

ORIGINAL CONTRIBUTION

FACTS Devices Impact on Congestion Mitigation of Power System

Jitender Kumar[1] · **Narendra Kumar**[1]

Received: 13 October 2018 / Accepted: 27 May 2020
© The Institution of Engineers (India) 2020

Abstract The voltage magnitude and reactive power control of a large power system are computed on the basis of its dynamic voltage controllers, which may overcome the limitations of a conventional controller of the power system. The approach of the voltage controller is dynamic in nature and capable of controlling voltage of the bus to the specified level and eliminates such circumstances which may lead the system to become uncontrollable. The static series compensation devices are capable of providing additional flexibility in the operation of power transfer on transmission lines. The previously presented paper discussed multiple applications of FACTS devices either in series or shunt connected. The FACTS devices, i.e., TCSC and SSSC, are incorporated in IEEE-14 bus test system for the analysis of their performance in a relation of reactive power and voltage magnitude. The steady-state modelling of these devices has been carried out by deriving the static equations. The WLS technique is used for measuring system parameters on individual buses. The techniques help in reducing error and improving accuracy in measuring of system parameters. The weak buses on which the voltage profile is poor have been identified. The FACTS devices as mentioned above are located on these buses. The voltage profile on these buses improves remarkably. The reactive power variation on these buses has been computed and graphically plotted. The size of the FACTS devices and their allocation has been determined on the basis of voltage magnitude and reactive power required on those buses during loading conditions. The IEEE 14 bus test system is designed on MATLAB/SIMULINK environment.

Keywords
Thyristor controlled series compensator (TCSC) · Static synchronous series compensator (SSSC) · Flexible AC transmission systems (FACTS) devices · Congestion mitigation · WLS technique

Introduction

The FACTS devices are static devices [8] which help in enhancement of the power system controllability, small-scale system stability, power transfer capability, etc. This perception initially pioneered by Hingorani [17]. The advancement of the FACTS devices in the era of power system is the most interesting area of research. The flexible AC transmission systems help in reduction of the power flow through overloaded transmission lines, improve system loadability, reduce losses over transmission lines, improve system efficiency, upgrade system reliability in addition to stability, enhance system security, etc. So that power transmission over an existing line is improved to compensate for higher industrial load and deregulate the power supply. To compensate requirement and demand for power will need to explore new ideas to minimize losses and maximize power transfer capability on existing transmission lines. Also, they concentrate on the lower and upper limit of the system in respect of network stability and reliability. The power electronic devices are more efficient with lower switching losses, so that they can provide eminent performance in control on power system. The FACTS controllers [19] help in practically regulating power flows in maintaining transmission voltages within

✉ Jitender Kumar
jitender3k@gmail.com

[1] Department of Electrical Engineering, Delhi Technological University, New Delhi, India

Published online: 05 June 2020

the specified limit and also mitigate the dynamic disturbances developed due to load variation in system. The TCSC and SSSC [21] are the robust and effective devices for power system stability. Reference [2] is discussing the problem associated with power system such as power losses and voltage regulation. They explore TCSC and Statcom process for mitigation of control of power flow in congested lines. They suggested technique benefits in reference of Indian grid network. The connecting lines between western grid and northern grid allow to transfer large power. They tested the network on IEEE 14 bus system.

The authors [4] have discussed the effect of series and shunt controller performance in respect of available transfer capability (ATC) of transmission grids. They explore the effect of TCSC, SVC and UPFC on single-area ATC and multi-area ATC. For analysing the performance and operation of power system, they have used particle swarm optimisation technique on IEEE test systems. The multiple combination of multi-type controller is analysed with their viability and technical merits on single-area ATC and multi-area ATC. The optimum allocations [5] of FACTS device have been decided on the basis of genetic algorithm optimization method. The optimum size and location of FACTS device are decided on annual daily load profile as a whole. The optimisation is obtained by graph-based panellized objective function used on mixed integer research heuristic optimization technique. The TCSC and SVC are employed for improving system performance in respect of voltage profile and other line parameters.

The author [7] suggested a nonlinear discrete time model with optimal stabilizing controller for islanded medium sized grid and micro-grid. The discrete time Hamilton Jacobian Isaacs optimal control technique is employed. The paper discusses on power quality in distribution system with renewable energy sources. The stabilizing controller is introduced with FACTS device to mitigate the grids oscillation. The successive approximation technique with neural network is employed for grid cost function. The system is implemented on grid tie inverter and synchronous generator in weak and islanded grids. The cumulative gravitational search algorithm [14] is introduced in industrial zone. The active and passive mass interactions are considered in CGSA. The system adopted for research is 72 bus southern grid system and 16 bus real-time industrial zone.

The power system quality [15] is analysed by using adaptive neuro fuzzy inference system. The system uses an advanced version of TCSC and SVC. The system is designed for a situation, when one generator is failed and other lines are overloaded. The adaptive neuro fuzzy inference system (ANFIS) is implemented to identify the value of capacitor connected in FACTS. The system is tested on IEEE 14 and 30 bus system. The market-based [16] power dispatch system is optimised with FACTS device. The paper focussed on positive externality problem in market design with active transmission participation. They adopted a sensitivity-based method to estimate market value of FACTS. The marginal value of FACTS device suggests impedance adjustment as well as financial incentive for FACTS owner. The authors suggested inclusion of FACTS price leads to improvement in significant social welfare in competitive power markets.

The authors [8] have proposed multiple FACTS controllers for the maximum power enhancement. The researchers have concentrated on transient energy flow during dynamic system damping [1]. The authors discussed several types of disturbances associated with the power system and how they can be minimized by the use of FACTS devices. The authors [6] explore the use of mixed type FACTS devices for improvement of loadability by use of GUI-based genetic algorithm. A dynamic model [19, 24] has been designed to verify the dynamic performance of FACTS devices. This paper helps in analysing/identifying the apparent power of FACTS devices in multiple combinations for the enhancement of power transfer capability. The researcher also focussed on AC–DC source coupling for improvement of active power transfer between buses. The process of implementation of multi-type and multiple FACTS devices are discussed on the IEEE Test Bus system. In [3, 21] FACTS devices are considered for rectifying the situation of India in relation to power consumption and power generation mismatch over the existing power transmission lines. The authors [9] discussed on the basis of linear and nonlinear analysis for dynamics of the power system. The FACTS devices like SSSC, STATCOM, and UPFC are used for damp-out of inter-area oscillations and design of supplementary damping controller in the time-domain simulation.

In the present paper, the steady-state analysis of the IEEE test power system has been conducted incorporating two FACTS devices, namely static synchronous series compensator (SSSC) [9] and thyristor controlled series capacitor (TCSC) [20]. Such FACTS devices are most popularly applicable to compensate transmission losses, reactive power, improve power transmission capability within the specified voltage profile. The series compensation basically provides control over the line impedance within a specified limit. In this case, the line impedance is varied from capacitive to inductive by adjusting the firing angle of the devices used, so that these devices provide control over real power with voltage magnitude. The static synchronous series compensator (SSSC) and thyristor controlled series capacitor (TCSC) are connected in series with the transmission line which helps in controlling line impedance and enhances power handling capability

[20, 21] of transmission lines. But the basic difference between TCSC and SSSC is that the TCSC provides a variable capacitor performance, but SSSC in addition is provided with a voltage source inverter. The SSSC can exercise control over reactive as well as real power [23].

We have used weighted least square (WLS) technique for improvement in parameter measurement. The WLS technique provides accurate and fast online measurement of the parameters of the FACTS devices which facilitates fast computation of the voltage profile. In this paper, FACTS devices are tested on the model of the IEEE 14 Bus System analysed using MATLAB/SIMULINK. The parameters analysed during the entire process are voltage magnitude and reactive power with and without the implementation of the proposed FACTS devices on a power system.

Modelling of Facts Devices

Thyristor Controlled Series Compensator

The TCSC [10, 19] is a power electronic device which is a combination of parallel connected inductor and capacitor with bidirectional control of power electronic component (Fig. 1). The combination of such devices will lead to control of a variable reactance operation from inductive to capacitive reactance. The device is dynamic and continuous in nature due to that they can quickly transform from inductive to capacitive or vice-versa, as per the requirement of the power system. It also compensates the voltage profile of the power system. A simplified form of TCSC [1, 10] behaviour is a variable inductor connected in parallel with a fixed capacitor. The maximum carrying current and voltage limit depends upon the thyristor valve rating, capacitor banks capacity and the reactor rated to compensate the required voltage profile and reactive power.

The effective reactance values of TCSC [2, 14] are determined by the implementation of KCL and KVL method. The changing reactance of X_{TCSC} represents the equivalent reactance of series–parallel combination of L & C, while an operation is either negative (capacitive) or positive (inductive) regions. The region will be defined on the basis of firing angle (α_{TCSC}) of TCSC's thyristor valve [4, 15] and can be calculated by

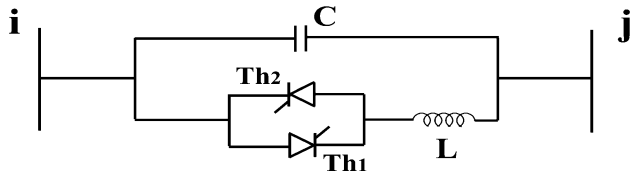


Fig. 1 TCSC

$$Z_{TCSC} = jX_{TCSC} \tag{1}$$

$$\begin{aligned} X_{TCSC} = &- x_C + \frac{x_C + x_{LC}}{\pi}[2(\pi - \alpha_{TCSC}) + \sin\{2(\pi - \alpha_{TCSC})\}] \\ &- \left[\frac{4x_{LC}^2}{\pi x_L}\right]\cos^2(\pi - \alpha_{TCSC})[\omega\tan\{\omega(\pi - \alpha_{TCSC})\} \\ &- \tan(\pi - \alpha_{TCSC})] \end{aligned} \tag{2}$$

where

$$\omega = \sqrt{\frac{x_C}{x_L}} \tag{3}$$

The x_L and x_C are used to indicate the fundamental harmonic reactance of TCSC's inductor and capacitor. The x_{LC} can be represented as

$$x_{LC} = \frac{x_C x_L}{x_C - x_L} \tag{4}$$

So that TCSC admittance be

$$y_{TCSC} = \frac{1}{Z_{TCSC}} = g_{TCSC} + jb_{TCSC} \tag{5}$$

where g_{TCSC} and b_{TCSC} are TCSC's conductance and susceptance, respectively, and here g_{TCSC} is zero.

In the power system, the flow of real power and reactive power [16, 21] within two buses is derived as

$$p_i = v_i \sum_{j=1}^{nb} v_j(g_{busij}\cos\delta_{ij} + b_{busij}\sin\delta_{ij}) \tag{6}$$

$$q_i = v_i \sum_{j=1}^{nb} v_j(g_{busij}\sin\delta_{ij} - b_{busij}\cos\delta_{ij}) \tag{7}$$

Linearised Newton's equation of the compensated transmission lines is given as in equation and TCSC reactance is taken as the state variable for matrix

$$\begin{bmatrix} \Delta P_i \\ \Delta P_j \\ \Delta Q_i \\ \Delta Q_j \\ \Delta P_{ij} \end{bmatrix} = \begin{bmatrix} \frac{\partial P_i}{\partial \delta_i} & \frac{\partial P_i}{\partial \delta_j} & \frac{\partial P_i}{\partial V_i} & \frac{\partial P_i}{\partial V_j} & \frac{\partial P_i}{\partial X_C} \\ \frac{\partial P_j}{\partial \delta_i} & \frac{\partial P_j}{\partial \delta_j} & \frac{\partial P_j}{\partial V_i} & \frac{\partial P_j}{\partial V_j} & \frac{\partial P_j}{\partial X_C} \\ \frac{\partial Q_i}{\partial \delta_i} & \frac{\partial Q_i}{\partial \delta_j} & \frac{\partial Q_i}{\partial V_i} & \frac{\partial Q_i}{\partial V_j} & \frac{\partial Q_i}{\partial X_C} \\ \frac{\partial Q_j}{\partial \delta_i} & \frac{\partial Q_j}{\partial \delta_j} & \frac{\partial Q_j}{\partial V_i} & \frac{\partial Q_j}{\partial V_j} & \frac{\partial Q_j}{\partial X_C} \\ \frac{\partial P_{ij}}{\partial \delta_i} & \frac{\partial P_{ij}}{\partial \delta_j} & \frac{\partial P_{ij}}{\partial V_i} & \frac{\partial P_{ij}}{\partial V_j} & \frac{\partial P_{ij}}{\partial X_C} \end{bmatrix} \begin{bmatrix} \Delta\delta_i \\ \Delta\delta_j \\ \Delta V_i \\ \Delta V_j \\ \Delta X_C \end{bmatrix} \tag{8}$$

where $\Delta P_{ij} = P_{ijreg} - P_{ijcal.}$ is the mismatch of real power flow in the series reactance.

$\Delta X_C = \Delta X_C^i - \Delta X_C^{i-1}$ is an incremental alteration in series reactance.

The state variable of TCSC (i.e., X_C of the series controller) is rationalized at the end of each iterative step

$$\Delta X_C^i = \Delta X_C^{i-1} + \left(\frac{\Delta X_C}{X_C}\right)^i \Delta X_C^{i-1} \tag{9}$$

The up-gradation of load flow equation implemented in the MATLAB program as the changes take place in X_C.

Static Synchronous Series Compensator

The static synchronous series compensator [13] is widely known as SSSC [5, 24], will act as a static phase shifter which injects quadrature voltage to one sideline end voltages and helps in regulating their active power flow (Fig. 2). The SSSC consists of its own reactive power [7, 18] provision as a capacitor in comparison with phase shifter required an additional reactive power source. So that SSSC is more capable to compensate not only voltage parameter but also control over apparent power compensation to power system.

The SSSC voltage angle and magnitude are adjusted by any suitable method so that it satisfies real and reactive power flow throughout the operation of SSSC on transmission line [1, 13].

$$L\frac{\mathrm{d}i}{\mathrm{d}t} + Ri = v_s - v_C - v_R \tag{10}$$

$$i = i_1 + ji_2, v_s = v_{s1} + jv_{s2} \tag{11}$$

$$v_C = v_{C1} + jv_{C2}, v_R = v_{R1} + jv_{R2} \tag{12}$$

The transforming of above equation into 1, 2 to *D–Q* components which related to

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} i_D \\ i_Q \end{bmatrix} \tag{13}$$

where $\theta = \omega_0 t + \theta_0$. In this case, no loss of generalization is assuming at $\theta_0 = 0$. Similarly transformation as given above applies to the variables v_{s1}, v_{s2}andv_{sD}, v_{sQ} and so on

$$(i_2 + ji_1) = (i_Q + ji_D)e^{j\omega_0 t} = Ie^{j\omega_0 t} \tag{14}$$

$$L\frac{\mathrm{d}(Ie^{j\omega_0 t})}{\mathrm{d}t} + R\left(Ie^{j\omega_0 t}\right) = (V_s - V_C - V_R)e^{j\omega_0 t} \tag{15}$$

On simplifying the equation, we get

$$(R + j\omega_0 L)I = (V_s - V_C - V_R) \tag{16}$$

The power flow [3, 13] between buses is as follows

$$S_{ij} = V_i I_{ij}^* \tag{17}$$

$$P_{ij} = V_i^2 G_{ii} - V_i V_j\left(G_{ij}\cos\theta_{ij} + B_{ij}\sin\theta_{ij}\right) - V_i V_{CR}(G_{ij}\cos(\theta_i - \delta_{CR}) + B_{ij}\sin(\theta_i - \delta_{CR})) \tag{18}$$

$$Q_{ij} = -V_i^2 B_{ii} - V_i V_j\left(G_{ij}\sin\theta_{ij} - B_{ij}\cos\theta_{ij}\right) - V_i V_{CR}\left(G_{ij}\sin(\theta_i - \delta_{CR}\right) + B_{ij}\cos(\theta_i - \delta_{CR})) \tag{19}$$

$$P_{ji} = V_j^2 G_{jj} - V_i V_j\left(G_{ij}\cos\theta_{ji} + B_{ij}\sin\theta_{ji}\right) + V_j V_{CR}(G_{ij}\cos(\theta_j - \delta_{CR}) + B_{ij}\sin(\theta_j - \delta_{CR})) \tag{20}$$

$$Q_{ji} = -V_j^2 B_{jj} - V_i V_j\left(G_{ij}\sin\theta_{ji} - B_{ij}\cos\theta_{ji}\right) + V_j V_{CR}\left(G_{ij}\sin(\theta_j - \delta_{CR}\right) - B_{ij}\cos(\theta_j - \delta_{CR})) \tag{21}$$

So that the power flow constraints [22] are given as

$$\mathrm{Re}\left(V_{CR}I_{ji}^*\right) = 0 \tag{22}$$

WLS State Evaluation Technique

Consider that the measurement set on an electrical network system is specified as a vector z:

$$z = h(x) + e \tag{23}$$

where

- hi (x) matrix is the nonlinear rationale related with measurement of i on the vector state x.
- x^T is the vector state of electrical network system [11].
- e_T is the error size in state vector.

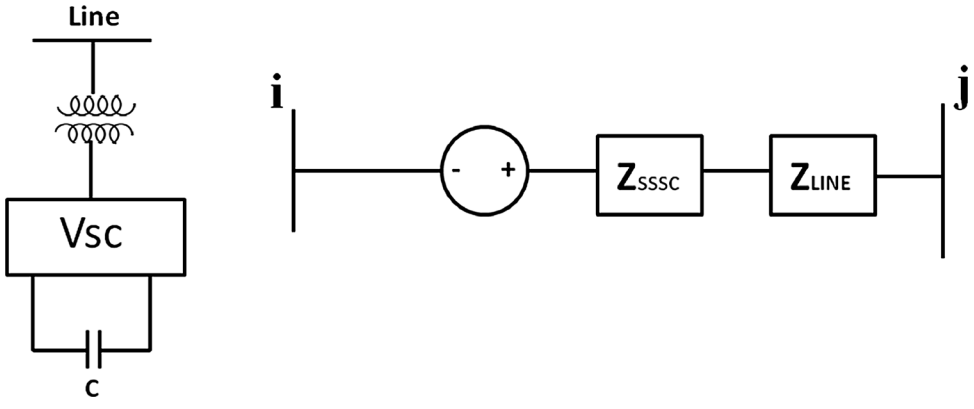


Fig. 2 SSSC

The full weighted least square (WLS) estimator [13, 17] will help to diminish the subsequent objective function:

$$J(x) = \sum_{i=1}^{m} \frac{(z_i - h_i(x))^2}{R_{ii}} = [z - h(x)]^T R^{-1}[z - h(x)] \quad (24)$$

At a smallest significance of an objective function, the first-order optimum condition is to be contented. It may be articulated [11, 12] as follows:

$$g(x) = \frac{\partial J(x)}{\partial x} = -H^T(x)R^{-1}[z - h(x)] = 0 \quad (25)$$

The nonlinear function $g(x)$ of Taylor series can be extended for the vector state x^k by disregarding the higher terms [12, 17] as

$$g(x) = g(x^k) + G(x^k)(x - x^k) + \ldots = 0 \quad (26)$$

The Gauss–Seidel technique is exercise to resolve the above equations. Where k is the iteration indicator, x^k is the vector state at iteration k and $G(x)$ is known as the gain matrix and expressed as:

$$G(x) = \frac{\partial g(x^k)}{\partial x} = H^T(x^k)R^{-1}H(x^k) \quad (27)$$

$$g(x^k) = -H^T(x^k)R^{-1}[z - h(x^k)] \quad (28)$$

Normally, the gain matrix is a sparse matrix and decomposed into triangular factors. At each iteration k, the gain matrix is solved by means of forward/backward replacement, where $\Delta x^{k+1} = x^{k+1} - x^k$ and

$$\begin{aligned}[G(x^k)]\Delta x^{k+1} &= H^T(x^k)R^{-1}\left[z - h(x^k)\right] \\ &= H^T(x^k) \cdot R^{-1}\Delta z^k \end{aligned} \quad (29)$$

These iterations are going on till the maximum variable difference satisfies the condition, ' Max $|\Delta x^k| < \varepsilon'$.

The percentage variation in parameter is calculated by using the following relations:

$$\%\text{Age variation} = \frac{\text{values after FACTS} - \text{values before FACTS}}{\text{values after FACTS}} \times 100\% \quad (30)$$

WLS Algorithm Along with FACTS devices

1. Study input data of the power system;
2. When FACTS device implements to node k, then analyse modified admittance matrix (Y bus)
3. Merge both power equations and network equation,
4. The inclusion of facts devices with traditional Jacobian matrix then these parameters increase their Jacobian matrix dimensions.
5. Then WLS state estimation technique pursues from Eqs. 23–30 (Fig. 3).

Simulation Result and Discussion

The tabular information associated with IEEE 14 Bus system [10] on standard loading conditions is as shown in Table 1.

Figure 4 presents the IEEE 14 bus test system. The buses from 9 to 14, on the basis of load flow study have been identified as the weak buses on the basis of low voltage profile. The FACTS devices TCSC and SSSC are considered located at these buses.

Realization of TCSC

The TCSC placement in IEEE 14 bus test system and its effect in respect to variation on the parameter with respect to base values are shown in Tables 2 and 3. The realisation of TCSC on different location over the IEEE 14 Bus system reveals its effect on the voltage profile and reactive power compensation. The assignment in placement of TCSC at a particular location at the power system is based on the bus voltage either varied frequently or falls below a specific limit while implementing practical load on buses. The target of TCSC realisation in the power system is to build and set up the system parameter as 1.0 p.u.

The realisations of an individual unit of TCSC at various locations in the power system, for analysis of its impact on the performance of the system parameter, are tested in the power system. The TCSC is implemented on different locations of buses where the variation in bus parameter is wide or close to the specific limit as per Table 1 which shows, e.g., 9–10; 10–11; 12–13 and 13–14. The variation of voltage magnitude and reactive power in different locations is shown in Figs. 5, 6, 7, 8, 9, 10, 11, 12, 13 and 14.

Realization of SSSC

The SSSC placement in IEEE 14 bus test system and its effect in respect to variation in the parameter with respect to the base values are shown in Tables 4 and 5. The realisation of SSSC on different location over the IEEE 14 Bus system reveals its effect on the voltage profile and reactive power compensation. The assignment in placement of SSSC at a particular location in the power system is based on the bus voltage either varied frequently or falls below a specific limit while implementing practical load on buses. The targets of SSSC controller for realisation in the power system are to build and set up the system parameter as 1.0 p.u.

The realisations of an individual unit of SSSC at various locations in the power system, for analysis of its impact on the performance of the system parameter, are to be

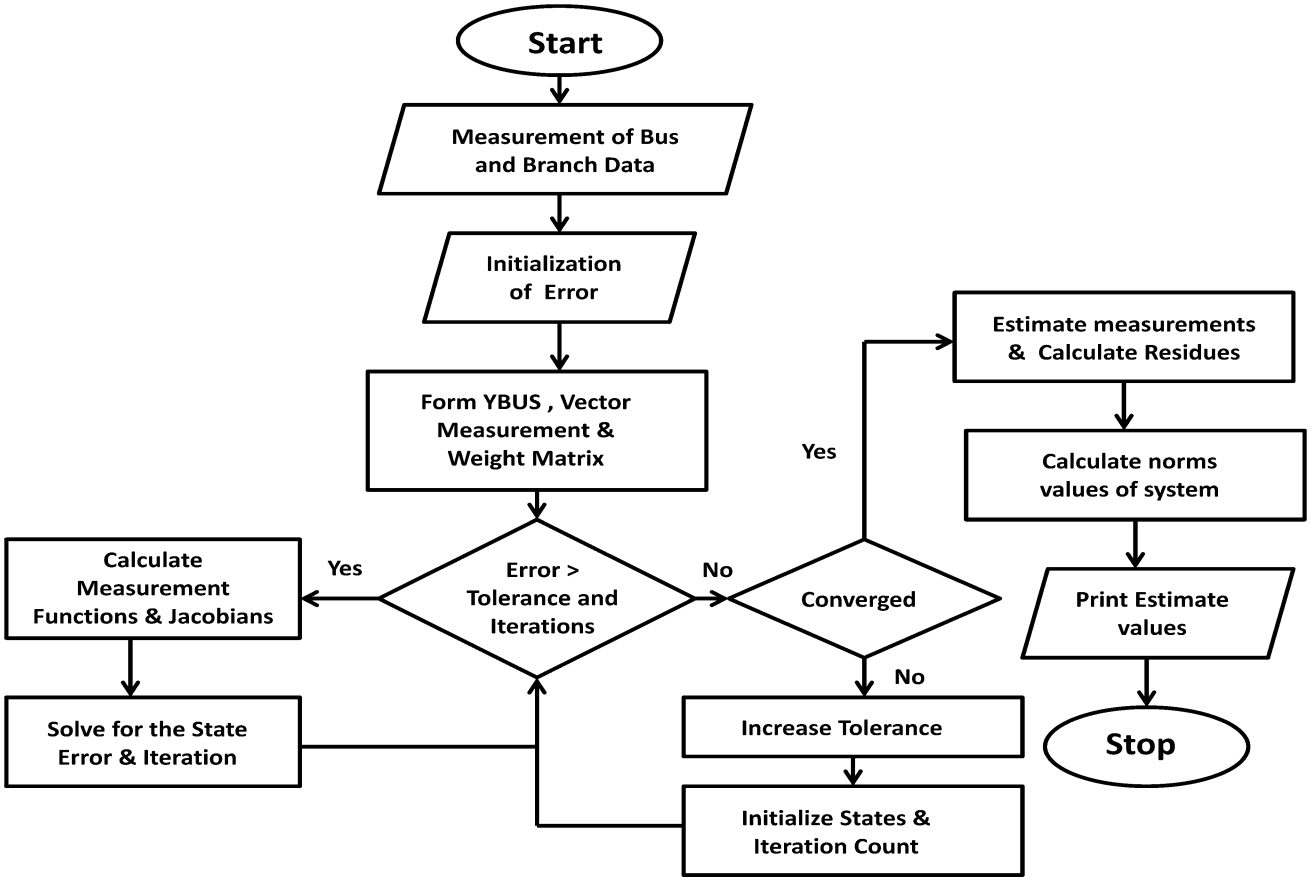


Fig. 3 Flow chart

Table 1 Base values IEEE 14 Bus system

Bus no	Bus voltage (pu)	Voltage angle (deg.)(pu)	*P* WATT (pu)	*Q* VAR (pu)
1	1.7012	1.0394	0.0123	0.0803
2	1.4328	0.9615	− 0.015	0.0153
3	1.1047	0.8351	0.0141	0.0057
4	1.0110	0.7908	− 0.002	− 0.005
5	0.9979	0.7843	− 0.049	− 0.044
6	1.3807	0.9440	0.0096	0.0734
7	1.0881	0.8275	− 0.001	− 0.003
8	1.6916	1.0369	− 0.002	− 0.032
9	0.9584	0.7642	0.0172	0.0114
10	0.9464	0.7579	0.0000	0.0000
11	0.9696	0.7700	− 0.003	0.006
12	0.9696	0.7700	0.002	0.0067
13	0.9659	0.7680	0.0127	0.0165
14	0.9439	0.7565	− 0.000	0.001

identified in the power system. The SSSC is implemented on different locations of buses where the variation in bus parameter is wide or close to the specific limit as per Table 1 shows, e.g., 9–10; 10–11; 12–13 and 13–14. The variation of voltage magnitude and reactive power in different locations are shown in Figs. 15, 16, 17, 18, 19, 20, 21, 22, 23 and 24.

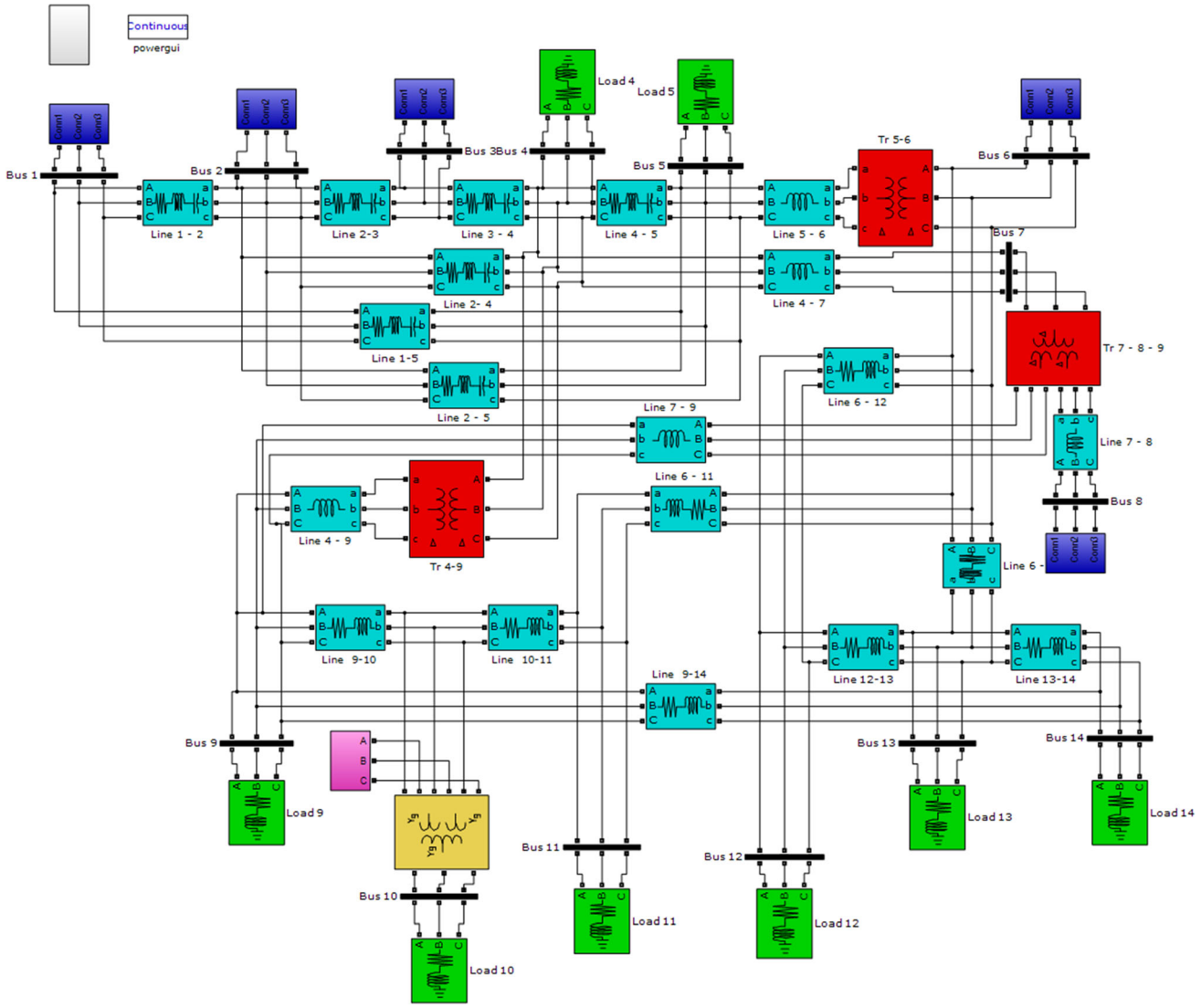


Fig. 4 IEEE 14 Bus System with TCSC

Table 2 Variations in average voltage with TCSC

Bus	Average voltage with or without TCSC				
	Without	9–10	10–11	12–13	13–14
1	0.95844	1.00454	–	–	–
2	0.94644	1.02363	1.00377	–	–
3	0.96958	–	1.04171	–	–
4	0.96961	–	–	1.00105	–
5	0.96587	–	–	1.01950	1.00364
6	0.94387	–	–		1.05274

Table 3 Variations in reactive power with TCSC

Bus	Reactive power with or without TCSC				
	Without	9–10	10–11	12–13	13–14
1	1.14E−02	1.36E−02	–	–	–
2	8.56E−06	8.98E−06	1.03E−05	–	–
3	5.70E−03	–	− 8.47E−03	–	–
4	6.71E−03	–	–	7.16E−03	–
5	1.65E−02	–	–	4.26E−03	1.78E−02
6	3.92E−04	–	–	–	− 6.11E−04

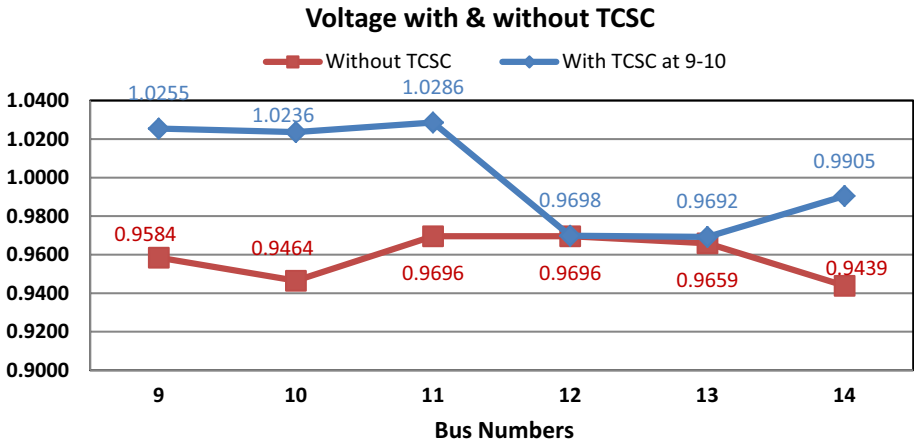


Fig. 5 Performance of voltage with TCSC at Bus 9–10

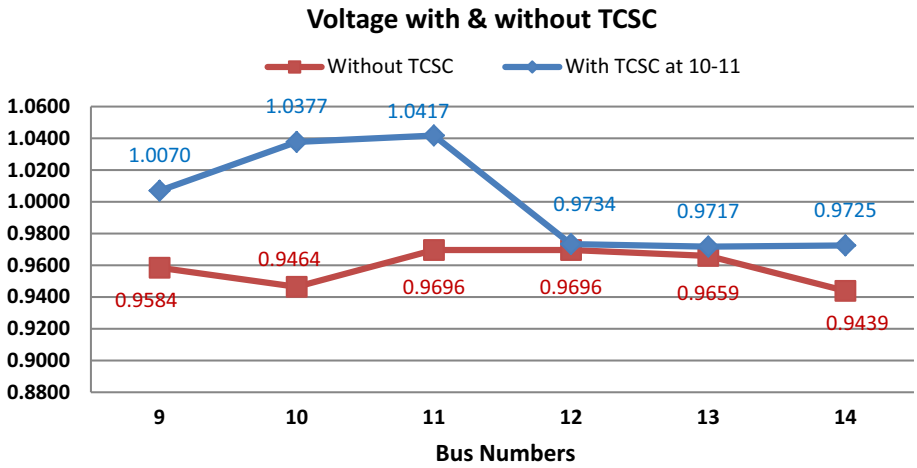


Fig. 6 Performance of voltage with TCSC at Bus 10–11

Conclusion

The TCSC and SSSC are located at different location identified on the basis of weak voltages on the buses of IEEE 14 bus test system. The FACTS devices are located independently in IEEE test system at different location to obtain the unity or specific value of bus voltage. The allocation of size and location of devices will help in sinking load congestion on buses of the power system. The relative cost of the whole system will increase due to FACTS device realization in power system. The cost of a large unit of FACTS devices is uneconomical on the basis of economic power transmission or distribution for the power system. For the optimisation of the power system, the optimal size and allocation of FACTS devices at the desired location are critical decision for power system

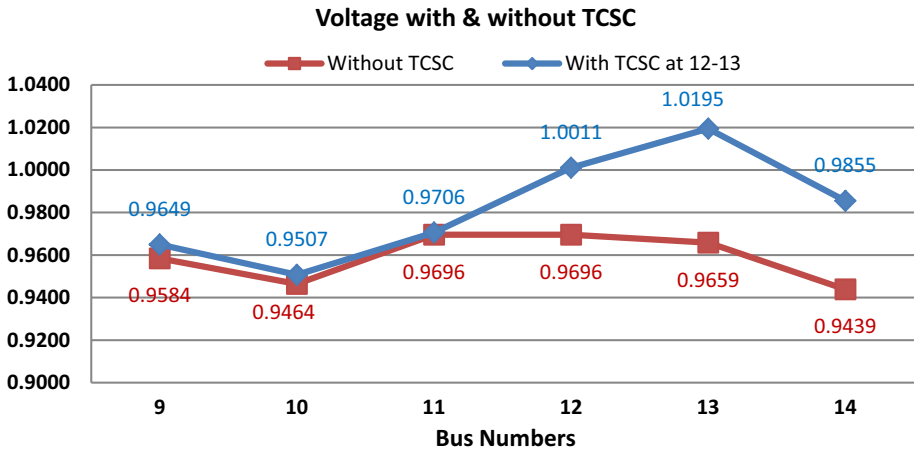


Fig. 7 Performance of voltage with TCSC at Bus 12–13

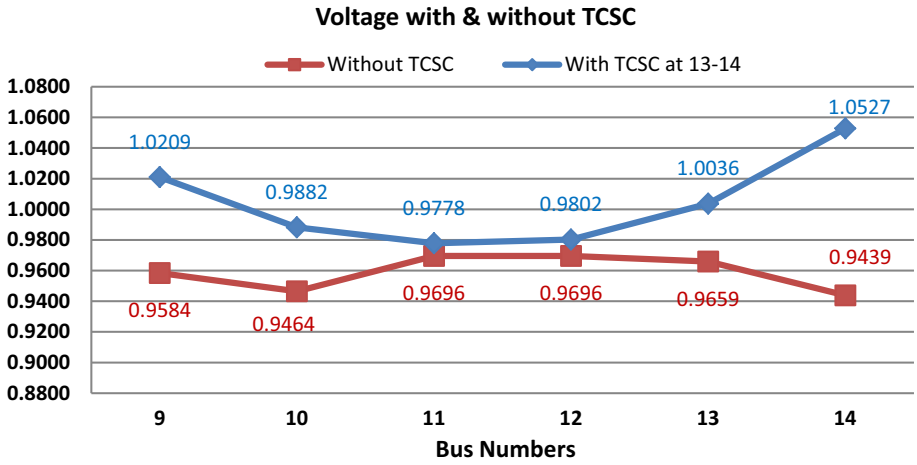


Fig. 8 Performance of voltage with TCSC at Bus 13–14

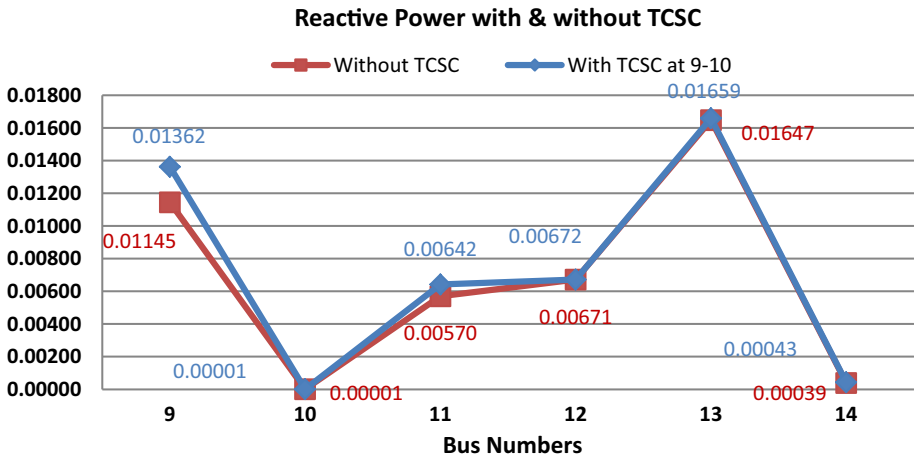


Fig. 9 Performance of reactive power with TCSC at Bus 9–10

economy. It is identified that while the placement of FACTS devices at a different location will help in improving the system performance. The allocation of TCSC at a different location on the IEEE Bus System will lead to improvement in voltage profile from 1.5 to 4.9% with respect to without FACTS devices. The reactive power constraint of the power system in IEEE System is varied from 13.0 to 48.2% as per the different location

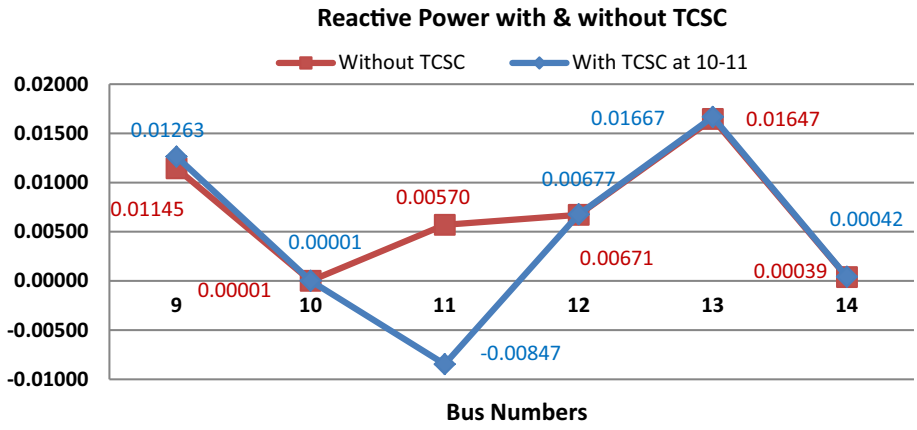


Fig. 10 Performance of reactive power with TCSC at Bus 10–11

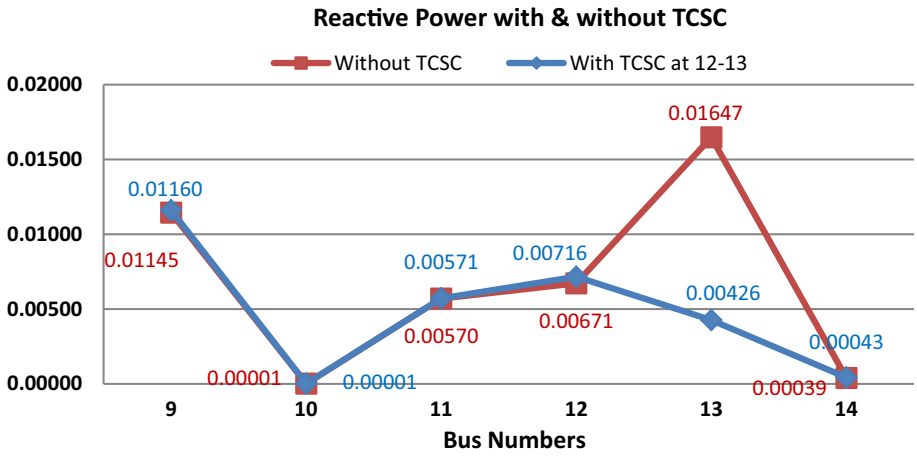


Fig. 11 Performance of reactive power with TCSC at Bus 12–13

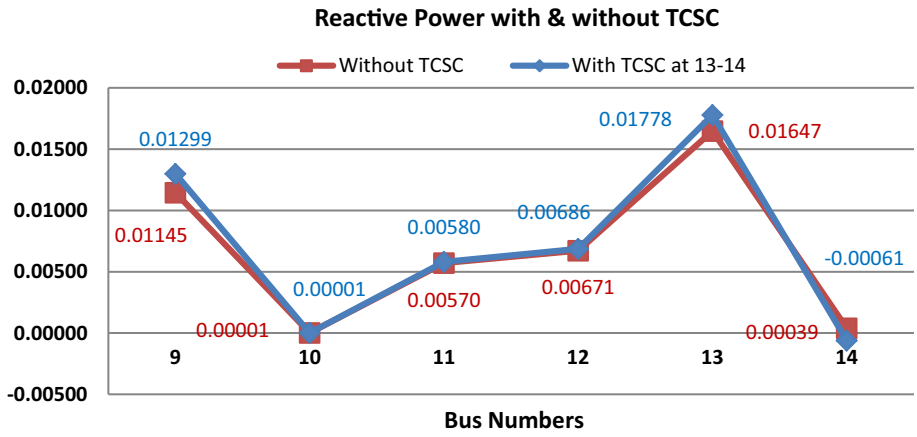


Fig. 12 Performance of reactive power with TCSC at Bus 13–14

identified on the basis of operation of the Basic IEEE 14 bus test system.

Similarly, the SSSC implementation on IEEE bus test system is arranged on the basis of voltage profile and Reactive Power profile of Buses as per the values identified by the normal operation of the power system under loading conditions. The variation in system parameters is identified during placing of SSSC to maintain the bus parameters as per the specified limit. The voltage profile of the power system varies on implementation of SSSC from 1.2 to 4.4%

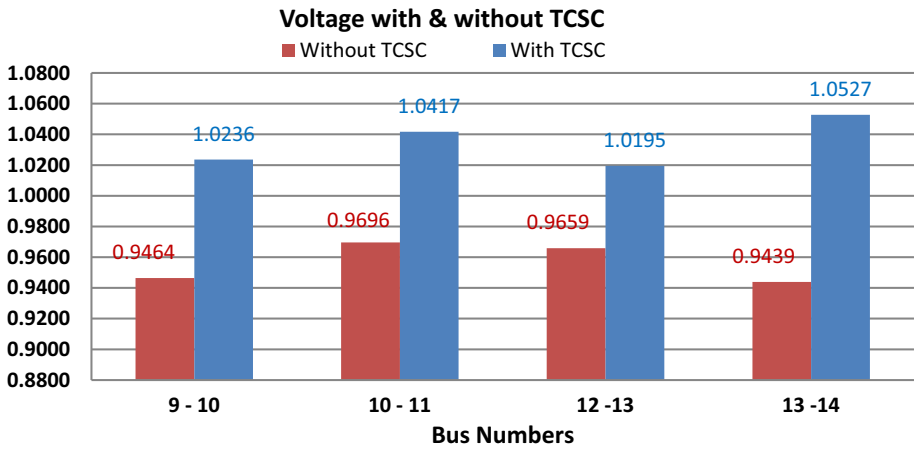


Fig. 13 Performance of voltage with TCSC

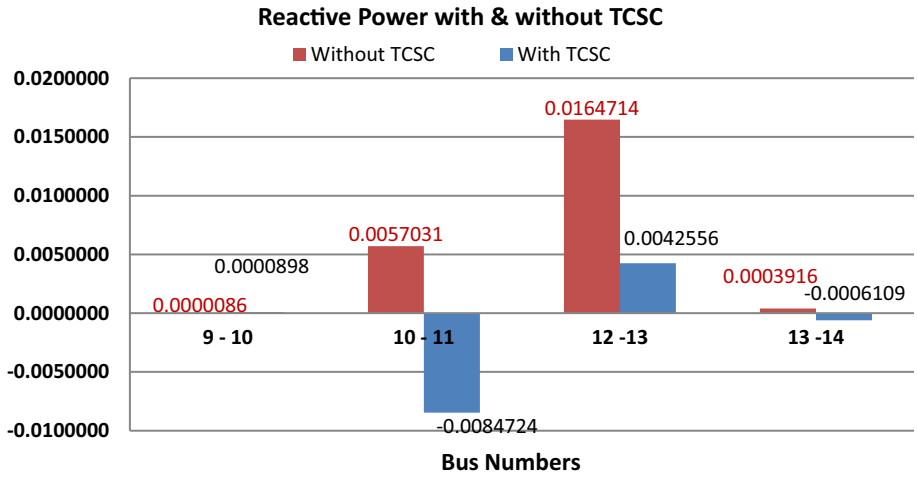


Fig. 14 Performance of reactive power with TCSC

Table 4 Variations in average voltage with SSSC

Bus	Average voltage with or without SSSC				
	Without	9–10	10–11	12–13	13–14
1	0.95844	0.93779	–	–	–
2	0.94644	1.01555	0.94376	–	–
3	0.96958	–	1.05453	–	–
4	0.96961	–	–	0.96697	–
5	0.96587	–	–	1.03233	0.95742
6	0.94387	–	–	–	1.00187

with respect to without FACTS device. The Reactive Power flow between buses will be varied by SSSC implementation at a different location from 4.9 to 74.2% with respect to the case without FACTS device. Finally, it is concluded that the SSSC implementation in IEEE is more effective FACTS device as compared to TCSC in respect of Reactive Power compensation. The SSSC help in the improvement of the Voltage profile of the system but not as much TCSC does. On the basis of results obtained on IEEE 14 bus test system with FACTS devices will help in improvement of voltage profile and reactive power compensation of the test system during the loaded condition. So

Table 5 Variations in reactive power with SSSC

Bus	Reactive power with or without SSSC				
	Without	9–10	10–11	12–13	13–14
1	1.14E−02	1.32E−02	–	–	–
2	8.56E−06	3.31E−05	1.18E−05	–	–
3	5.70E−03	–	− 7.44E−04	–	–
4	6.71E−03	–	–	7.37E−03	–
5	1.65E−02	–	–	8.53E−03	1.65E−02
6	3.92E−04	–	–	–	− 4.89E−05

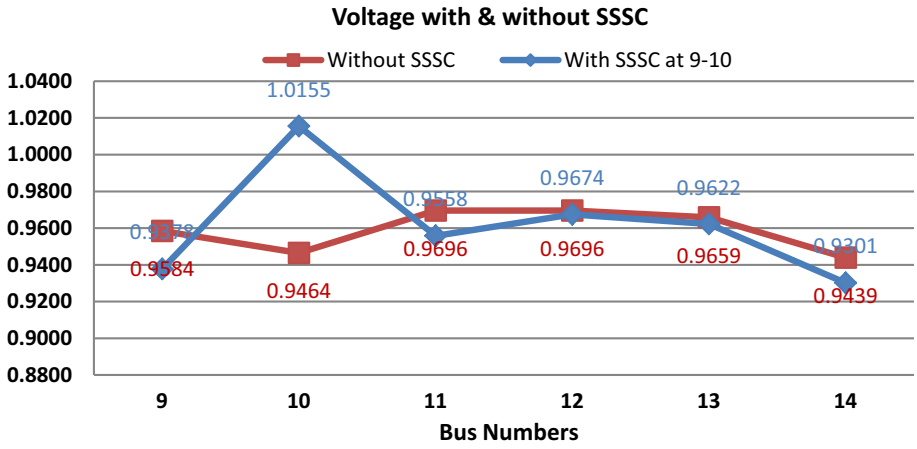


Fig. 15 Performance of voltage with SSSC at Bus 9–10

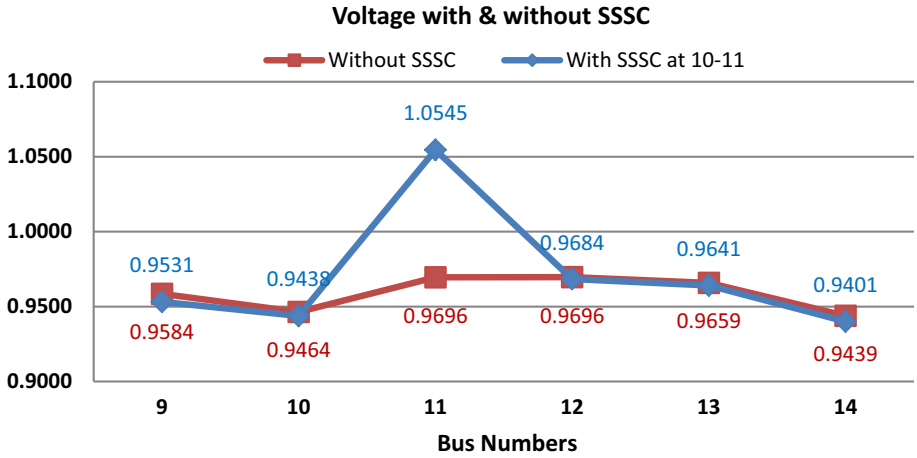


Fig. 16 Performance of voltage with SSSC at Bus 10–11

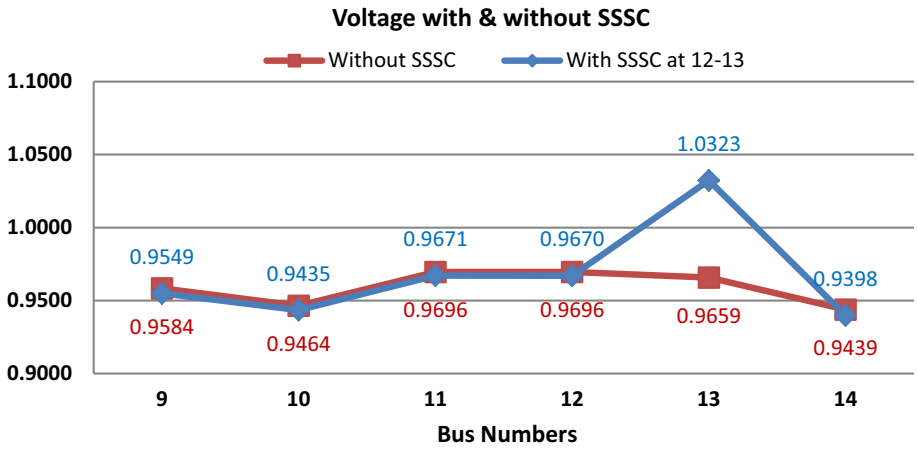


Fig. 17 Performance of voltage with SSSC at Bus 12–13

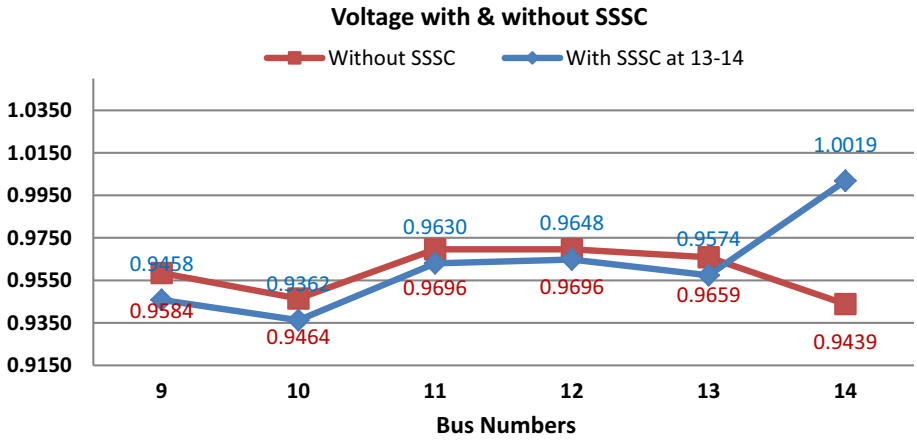


Fig. 18 Performance of voltage with SSSC at Bus 13–14

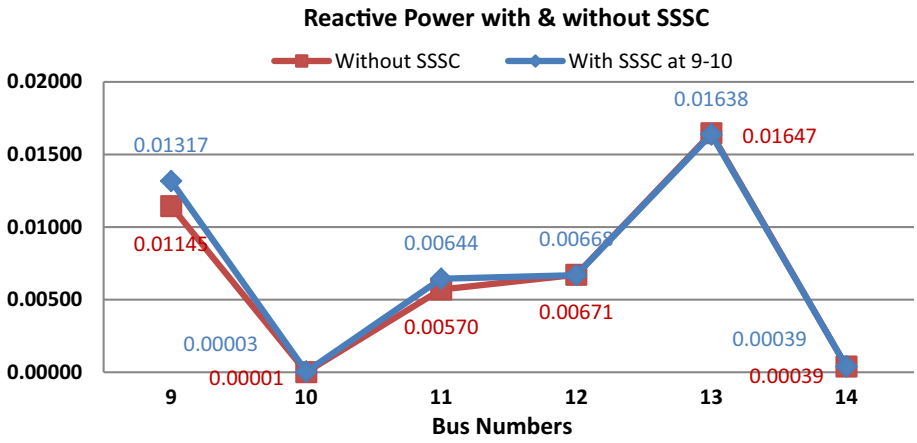


Fig. 19 Performance of reactive power with SSSC at Bus 9–10

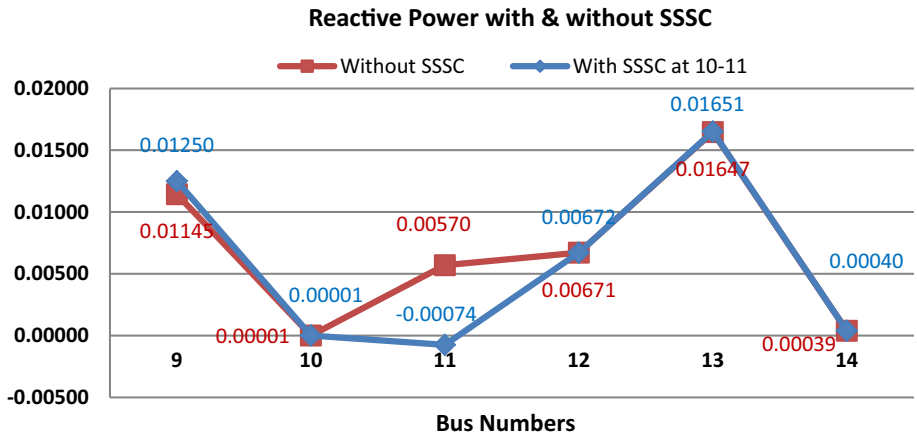


Fig. 20 Performance of reactive power with SSSC at Bus 10–11

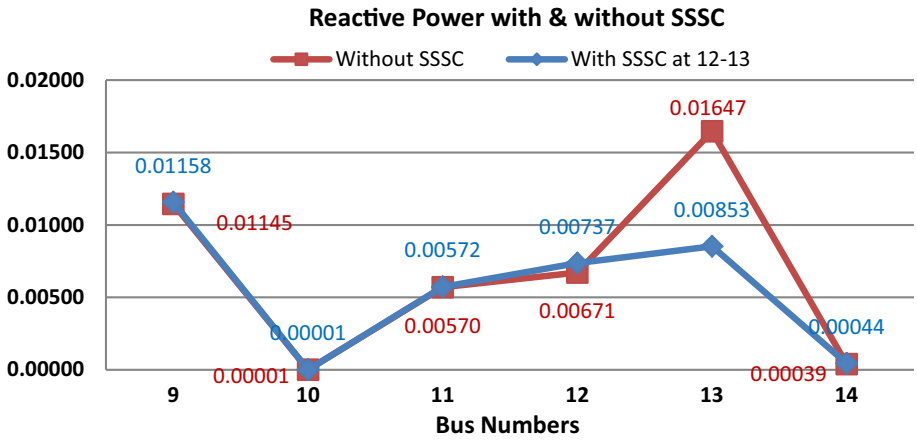


Fig. 21 Performance of reactive power with SSSC at Bus 12–13

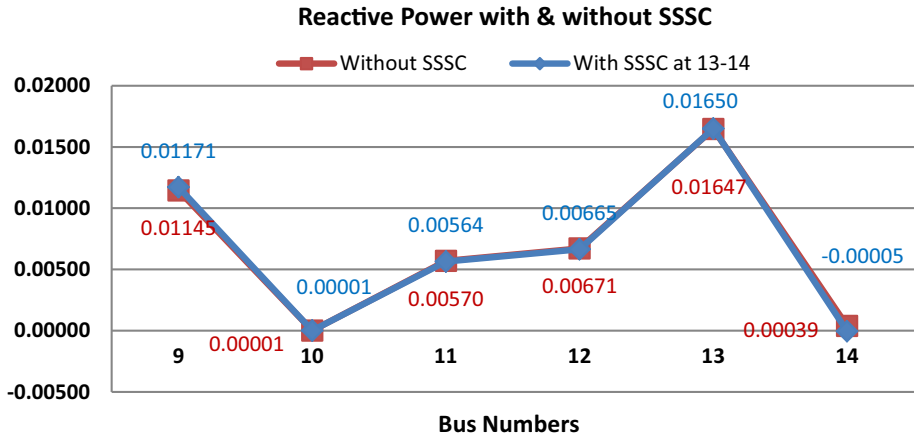


Fig. 22 Performance of reactive power with SSSC at Bus 13–14

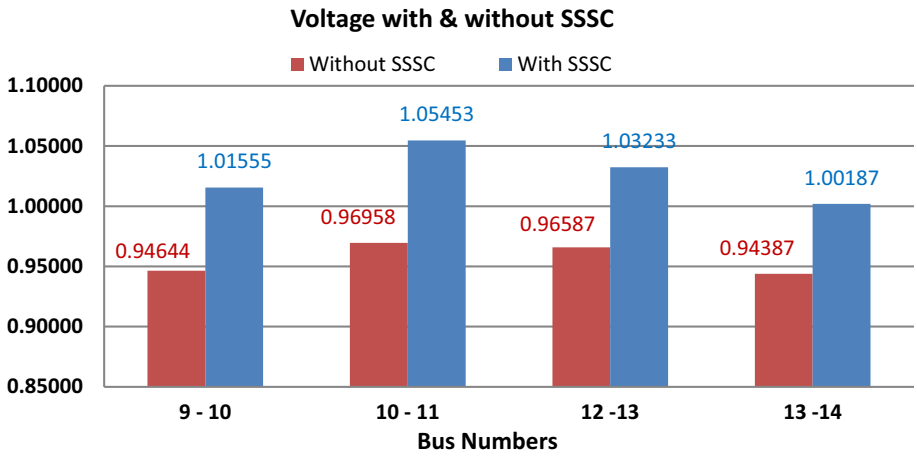


Fig. 23 Performance of voltage with SSSC

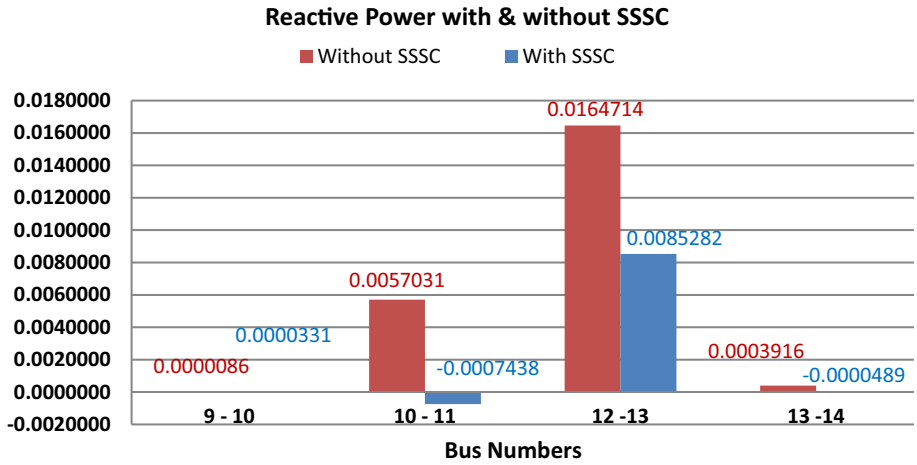


Fig. 24 Performance of reactive power with SSSC

that FACTS devices will offer an enhanced load congestion relief to the power system.

References

1. S. Amara, H.A. Hsan, Power system stability improvement by FACTS devices: aA comparison between STATCOM, SSSC and UPFC. 2012 First international conference on renewable energies and vehicular technology (REVET), 26–28 March 2012, pp. 360–365 **(ISBN: 978-1-4673-1168-7)**
2. A.S. Siddiqui, M.T. Khan, F. Iqbal, Determination of optimal location of TCSC and STATCOM for congestion management in deregulated power system. Int. J. Syst. Assur. Eng. Manag. **8**, 110–117 (2017)
3. Bone Gorazd and Mihalic Rafael, "Modelling the SSSC Device in the Power Flow Problem as Power Injections Regulated According to First Order Sensitivity" Engineering and Industry Series Volume Deregulated Electricity Market Issues in South Eastern Europe, DOI: 10.22618/TP.EI.20151.192013, pp. 80 – 85, 2015.
4. B.V. Manikandan, S.C. Raja, P. Venkatesh, Available transfer capability enhancement with FACTS devices in the deregulated electricity market. J. Electr. Eng. Technol. **6**(1), 14–24 (2011)
5. C. Ersavas, E. Karatepe, Optimum allocation of FACTS devices under load uncertainty based on penalty functions with genetic algorithm. Electr. Eng. **99**(1), 73–84 (2017)
6. E. Ghahremani, I. Kamwa, Optimal placement of multiple-type FACTS devices to maximize power system loadability using a generic graphical user interface. IEEE Trans. Power Syst. **28**(2), 764–778 (2013)
7. H. Nazaripouya, S. Mehraeen, Modeling and nonlinear optimal control of weak/islanded grids using FACTS device in a game theoretic approach. IEEE Trans. Control Syst. Technol. **24**(1), 158–171 (2016)
8. J. Park, G. Jang, K.M. Son, Modeling and control of VSI type FACTS controllers for power system dynamic stability using the current injection method. Int. J. Control Autom. Syst. **6**(4), 495–505 (2008)
9. X. Jiang, J.H. Chow, A.-A. Edris, B. Fardanesh, E. Uzunovic, Transfer path stability enhancement by voltage-sourced converter-based FACTS controllers. IEEE Trans. Power Deliv. **25**(2), 1019–1025 (2010)
10. J. Kumar, N. Kumar, Application of FACTS devices for congestion mitigation in power system. Int. J. Manag. Technol. Eng. **9**(2), 1588–1601 (2019)
11. J. Kumar, N. Kumar, Determination and allocation of optimal size of TCSC and STATCOM for obstruction alleviation in power system. 2nd IEEE international conference on power electronics, intelligent control and energy systems (ICPEICES-2018), Delhi Technological University, New Delhi, India, PRINT

©2018 IEEE, Oct 22–24, pp. 1–8, 2018 **(ISBN: 978-1-5386-6625-8/18/$31.00)**
12. J. Kumar, J.N. Rai, N. Hasan, Use of phasor measurement unit (PMU) for large scale power system state estimation. 2012 IEEE Fifth India international conference on power electronics, IICPE2012, Delhi Technological University, New Delhi, India, Dec 6–8, pp. 1–5, 2012 **(ISSN: 2160-3162, PRINT ISBN : 978-1-4673-0931-8)**
13. L. Garg, S.K. Agarwal, V. Kumar, Enhancement of transient stability of multimachine system using TCSC, SSSC and UPFC, In *Intelligent communication control and devices advances in intelligent systems and computing* ed. by L. Garg, S.K. Agarwal, V. Kumar. (Springer, Berlin, 2018), pp. 381–391
14. M. Packiasudha, S. Suja, J. Jerome, A new cumulative gravitational search algorithm for optimal placement of FACT device to minimize system loss in the deregulated electrical power environment. Int. J. Electr. Power Energy Syst. **84**, 34–46 (2017)
15. M.R. Sayed, M.A.M. Hassan, A.A. Hassan, Power system quality improvement using flexible AC transmission systems based on adaptive neuro-fuzzy inference system. WSEAS Trans. Power Syst. **8**(2), 65–73 (2013)
16. M. Sahraei-ardakani, S.A. Blumsack, Market-based operation of series FACTS devices. IEEE Trans. Power Syst. **31**(5), 3702–3714 (2016)
17. N.G. Hingorani, L. Gyugyi, *Understanding FACTS concepts and technology of flexible AC transmission systems* (IEEE Press, New York, 1999)
18. N.M. Kirby, P.E. Marken, M. Paradis, P. Wang, I. Plowright, H.-B. Moon, D. Ingemansson, R. Mendis, B. Mehraban, Extending their lifetimes: keeping HVDC and FACTS installations in service longer. IEEE Power Energy Mag. **14**(2), 57–65 (2016)
19. N.P. Padhy, M.A.A. Moamen, Power flow control and solutions with multiple and multi-type FACTS devices. Electr. Power Syst. Res. **74**(3), 341–351 (2005)
20. R. Agrawal, S.K. Bharadwaj, A. Bodhe, D.P. Kothari, B. Deshmukh, *Optimal location of TCSC using evolutionary optimization techniques* (Springer, Berlin, 2018), pp. 389–397
21. S.K. Singh, L. Phunchok, Y.R. Sood, Voltage profile and power flow enhancement with FACTS controllers. Int. J. Eng. Res. Technol **1**(5), 1–5 (2012)
22. T. Deb, A.S. Siddiqui, Application of multiple FACTS devices of similar type for congestion mitigation. Int. J. Syst. Assur. Eng. Manag. **7**(3), 387–397 (2016)
23. R. Tiwari, N.R. Niazi, V. Gupta, Optimal location of FACTS devices for improving performance of the power systems. Power and Energy Society General Meeting, 2012 IEEE, (22–26 July 2012), pp. 1–8 **(ISSN :1944-9925)**
24. Y. Xiao, Y.H. Song, C.-C. Liu, Y.Z. Sun, Available transfer capability enhancement using FACTS devices. IEEE Trans. Power Syst. **18**(1), 305–312 (2003)

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

PAPER

Formation of Self-Assembled Monolayer and Characterization of AA7075-T6/B_4C Nano-ceramic surface composite using Friction Stir Processing

To cite this article: Ravi Butola *et al* 2020 *Surf. Topogr.: Metrol. Prop.* **8** 025030

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.

This content was downloaded from IP address 14.139.251.100 on 08/06/2020 at 10:36

Surface Topography: Metrology and Properties

PAPER

Formation of Self-Assembled Monolayer and Characterization of AA7075-T6/B_4C Nano-ceramic surface composite using Friction Stir Processing

RECEIVED
28 January 2020

REVISED
15 May 2020

ACCEPTED FOR PUBLICATION
27 May 2020

PUBLISHED
4 June 2020

Ravi Butola, Qasim Murtaza and Ranganath M Singari

Department of Mechanical, Production & Industrial and Automobile Engineering, Delhi Technological University, Delhi 11042, India

E-mail: ranganath@dce.ac.in **and** qasimmurtaza@dce.ac.in

Keywords: self assemble monolayer, surface modification, Friction stir processing, surface composite, nanoparticles

Abstract

In the present study, fabrication of surface composite using AA7075-T6 as a matrix material and boron carbide nanoparticles as reinforcement through Friction stir processing (FSP) has been done. FSP technique has been widely utilized for surface modification and the formation of composite material. The B_4C nanoparticles size (<30 nm) as reinforcement were padded in dimension of 2 mm width and 1.5 mm depth groove of AA7075-T6 plate as a matrix material. The single-pass process executed using a square tool pin with tool rotation and traverse speed of 1000 rpm and 40 mm min^{-1} respectively. This research aims to observe and process Self-Assembled Monolayer (SAM), investigate the effect of the B_4C nano-ceramic particles on the AA7075-T6 and its mechanical properties of the nano-ceramic surface composite. Frictional and wear analysis investigations under various physical conditions have highlighted surface durability characteristics of the metal matrix composite of AA7075-T6. Microstructure results along with fractography-image highlight the homogeneous distribution of boron carbide nano-ceramic particle. Tensile test, Microhardness, microstructure, Field Emission Scanning-Electron Microscope (FESEM), and X-Ray Diffraction (XRD) analyzed the fabricated (Al + B_4C) nano-ceramic surface composite. The fabricated nano-ceramic surface composite could be utilized in lightweight applications such as aerospace, marine, defence, and automotive industry.

1. Introduction

Aluminum-based metal-matrix surface composite reinforced with B_4C ceramic particles are developed for new materials, which are widely used in lightweight applications like aerospace, marine, defence, and automotive industry, due to their high strength to weight ratio and wear resistance. Surface composite (SC) are replacing conventional metallic alloys in present industries. The nano-ceramic particles reinforcement materials are applied for the development of new material. AA5083 surface composites reinforced with micro and nano-size ceramic particles B_4C using FSP were administered. Its mechanical properties evaluated through the tensile test and microhardness and then the result was compared to the base metal [1]. Aluminum/B_4C Surface composite with nano-ceramic particles reinforcement have been widely used as a neutron absorber in nuclear power plant, bulletproof vest, armor tank (shielding technology) and transportation application, etc due to their excellent thermal and chemical stability, low density and high hardness [2–4]. Various researchers through different technique manufactured Al-B_4C surface composite in the last few years using mechanical alloying [5], Stir casting [6–8], Casting [9–11], Microwave sintering and hot extrusion processes [12], Powder metallurgy technique [13], Squeeze casting [14], Electromagnetic stir casting [15]. Several researchers have reported the application of acetone and methanol in surface modification of Al-alloy using Friction stir processing (FSP) [16–18], whereas in the current method we are using Diethyl ether (DEE) as a linking agent [19].

In the current method, we have modified the surface using B_4C nano-ceramic particles, which is linked through the Self Assembled Monolayer (SAM) on the surface of Al-alloy. SAM is widely being used

© 2020 IOP Publishing Ltd

as a linking process on various surfaces [20]. The advantage of using the Self-Assembled Monolayer (SAM) technique is the optimization of adhesion property. The surface modification using nano-ceramic materials enhance the surface properties of Al alloy. Torbjorn Fangstrom *et al* [21] presented the reaction between Al alloy and Di-methyl ether (DME) by the comparative examination of density functional theory and EPR results. Setup of stationary points by discussing the reaction between Al alloy and Di-methyl ether was located by the help of density functional theory at B3LYP and BP86 levels. Both methods provide qualitatively the same results but the barrier differed slightly because of the methods. Duhua Wang *et al* [22] investigated the Self-assembled monolayer, multilayer thin films along with corrosion resistance by using four different alkyl-saline on Al-alloy 2024-T3 base metal. From the results, it has seen that the self-assembling of the alkyl-saline compounds was altered by the properties of the complex surface of the substrates.

Formation of AA7075-T6/B_4C nano surface composite using FSP was administered. Subsequently, the various surface composite was fabricated with substrates of Al, [23–26] Cu, [27, 28] Mg, [29–31] and reinforcements like B_4C, SiC, TiC, Al_2O_3, etc. Numerous researchers have performed different nano-particles reinforcement procedures, which influence chemical stability and microstructural modification of surface composite using FSP. Different methods have been reported for applying of reinforcement particles which initiate by preparation of grooves on the matrix metal and filled by reinforcement particles [32], Matrix Hole Drilling followed up by particle reinforcement [33–36], particle reinforcement with a volatile liquid mixture (methanol or acetone) applied over the matrix metal [37, 38]. The selection of the nano-ceramic reinforcement particles depends on the application of the component and compatibility of nano reinforcement and matrix metal. Authoritatively B_4C offers excellent thermal and chemical stability, neutron absorber, low density and high hardness [2–4]. Hence, researchers have not reported the surface modification of AA7075-T6 as a substrate through the FSP technique using B_4C nano-ceramic reinforcement mixed with Di-ethyl ether (DEE).

The main objective of the current research is to create a homogeneous layer of B_4C nano-ceramic particles using SAM technique into AA7075-T6 as substrate and Di-ethyl ether (DEE) as a linking agent, altering the surface through FSP technique.

In figure 1(a) Groove on the BM (Base Metal) surface followed by (b) After sonication of BM and dipping in ethanol for 24 h (c) Formation of SAM on the groove of BM (d) Fabrication of surface composite using FSP technique (e) Friction stir processed plate (f) Sample cut out from the processed plate for testing.

2. Experimental methods

The matrix metal used in this research is Aluminium alloy 7075-T6 plate of 200 mm length 80 mm width, 6 mm thickness, and B_4C nanoparticles of 99.9% purity with the particle size of (<30 nm) was used as reinforcement materials purchase from Intelligent Materials Pvt. Ltd(Nanoshel). The chemical composition and mechanical properties of matrix metal are shown in table 1. The experiment was conducted in two stages. In the first stage, the formation of the Self-Assemble Monolayer (SAM) on the surface of matrix materials was performed. In the second stage, the capping pass was done with a pin-less tool to cover the self-assembled monolayer, which is bind with each groove of the AA7075-T6 plate and the tool profile and rpm were adjusted as per process parameter given in table 2.

2.1. Formation of self-assemble monolayer (SAM)

To get the uniformity and homogeneous layer of B_4C nanoparticles onto the AA7075-T6 substrate a self-assembled monolayer of B_4C nano-ceramics was prepared. Before starting the formation of the self-assembled monolayer, process on the matrix materials the first step of the AA7075-T6 plate surface had to be made clean by following the sonication technique. Primary sonication was done with a 5% chemical detergent solution in distilled water. The sonicator was switched on for periods of 30 min and left to soak for a 15-min interval, repeated twice. The secondary sonication was performed by only using distilled water with similar cycles of sonication and soaking repeated 3 times. The sonicated plates were immediately dipped into ethanol and leftover for 24 h in an airtight container to prevent any further reaction or accumulation of moisture on its surface as shown in figure 2(a). B_4C nano-ceramic powder is uncreative to Di-ethyl ether (DEE) solvent; hence, it requires constant stirring to form a colloidal solution. According to the dimension of groove, we have selected that 0.5 molar concentration of B_4C nano-ceramic particle was being used to do for the uniformity and homogeneity of B_4C nano-ceramic particle on the Aluminium substrate.

Using a micropipette solution dropped onto the specimen and following the smearing technique, spread across the plate using a glass slide. The active –OH– bonds on a sample from ethanol were prepared to receipt the uniform layer of (B_4C-Diethyl ether) solution and form a temporary bond onto the substrate. The sample was then kept on a hot plate to dry inside a chemical hood with temperature ranging from 75 °C–200 °C for 10 to 15 min as shown in figure 2(c). Keeping least contact with moisture or any other agent the sample plates immediately placed in an airtight container for carrying out the further process. The B_4C nano-particle blended with Di-ethyl ether (DEE) and filled in the groove of 160 mm length, 2 mm width

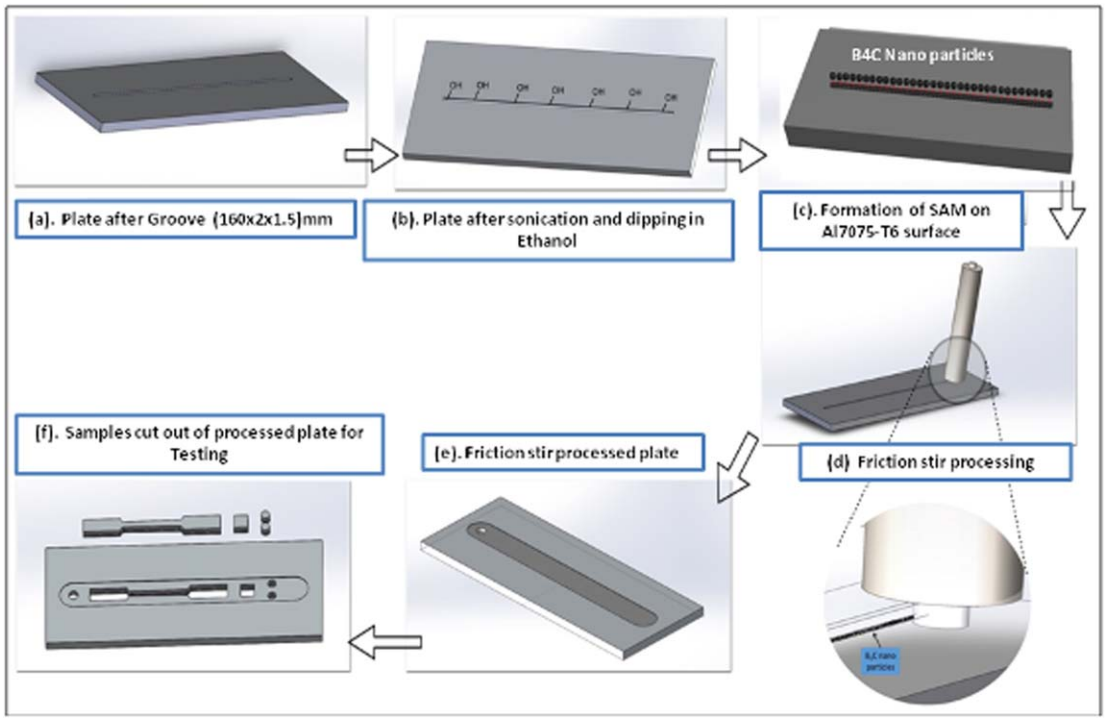


Figure 1. Schematic representation of Nano-ceramic surface composite using Friction stir processing.

Table 1. Mechanical properties and Chemical composition of AA7075-T6.

Alloy	Zn	Mg	Cu	Fe	Si	Mn	Ti	Cr	Al
Wt%	2.273	2.4	1.3	0.18	0.06	0.04	0.08	0.18	REST
Modulus of elasticity E (GPA)		Tensile strength (MPa)			Yield strength (MPa)			Elongation (%)	
71.0		544			502			21.5	

Table 2. Process parameters of FSP.

Process parameters	Value taken
Shoulder profile of Tool	Flat
Shoulder dia of Too	19.95 mm
Pin profile of Tool	Square
Pin length of Tool	3.5 mm
Plunge depth of Tool	0.20 mm
Tilt angle of Tool	2° constant
Tool rotational speed	1000 rpm
Type of reinforcement	B_4C (<30 nm)

and 1.5 mm depth on the matrix metal shown in figure 2(b).

2.2. Frictions stir processing

The experimental tests were carried out using FSW/FSP setup. AA7075-T6 matrix materials were cut according to the standard dimension of the FSW/FSP set up. In the matrix, material dimension of (200 × 80 × 6) mm a groove of 160 mm length, 2 mm width and 1.5 mm depth was processed using H13 tool steel heat-treated with 60–65 HRC hardness. H-13 Tool Steel is the most widely used work steel in hot work die and cold work applications. The tool pin was both the side of the tool with a total length of 117 mm. The diameter of the tool shoulder was 19.95 mm and the length of the square pin was 3 mm. The collar diameter is taken as 24 mm and the collar thickness is taken as 10 mm. The square tool pin profile must have a lesser amount of flashes on the matrix and reinforced material should be well disseminated by the tool pin to form the good processing zone. The effect of the pin profile tool on ceramic particle distribution is thoroughly analyzed and investigated. As mentioned by Mahmoud *et al* in reports highlight, SiC particles are distributed more homogeneously in aluminum, where they have used square probe in place of circular or triangular probe tool [39]. Vijay and Murugan used Al/TiB2 composite samples with probes having square, hexagon or octagon pin geometry and have concluded that a straight square pin achieves the best mechanical properties, which signifies a uniform TiB2 distribution [40]. Square tool pin profile has been found good trade-off among various reports on the properties of surface composite fabricated using friction stir processing [41]. Square pin profile was found to be most suitable due to high pulsating action generated. The pinless tool M2 grade H13 tool steel manufactured by

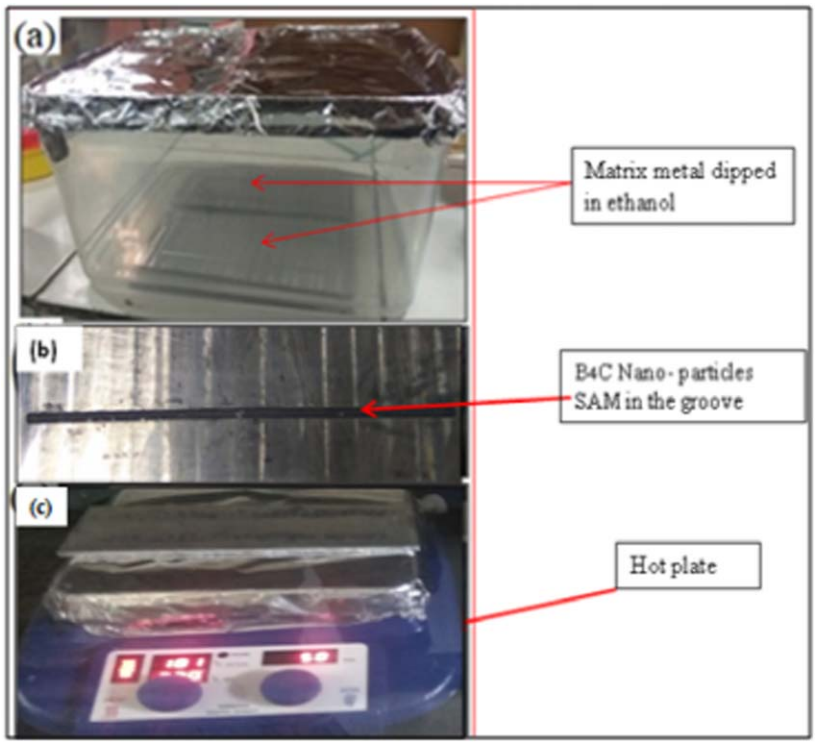


Figure 2. Formation of SAM process (a) Matrix metal dipped in ethanol for 24 h(b) filled B_4C nano-particles in the groove (c) Samples kept on hot plate for the evaporation of the solvent and proper adsorption of nano ceramics particles to the Base Metal (BM).

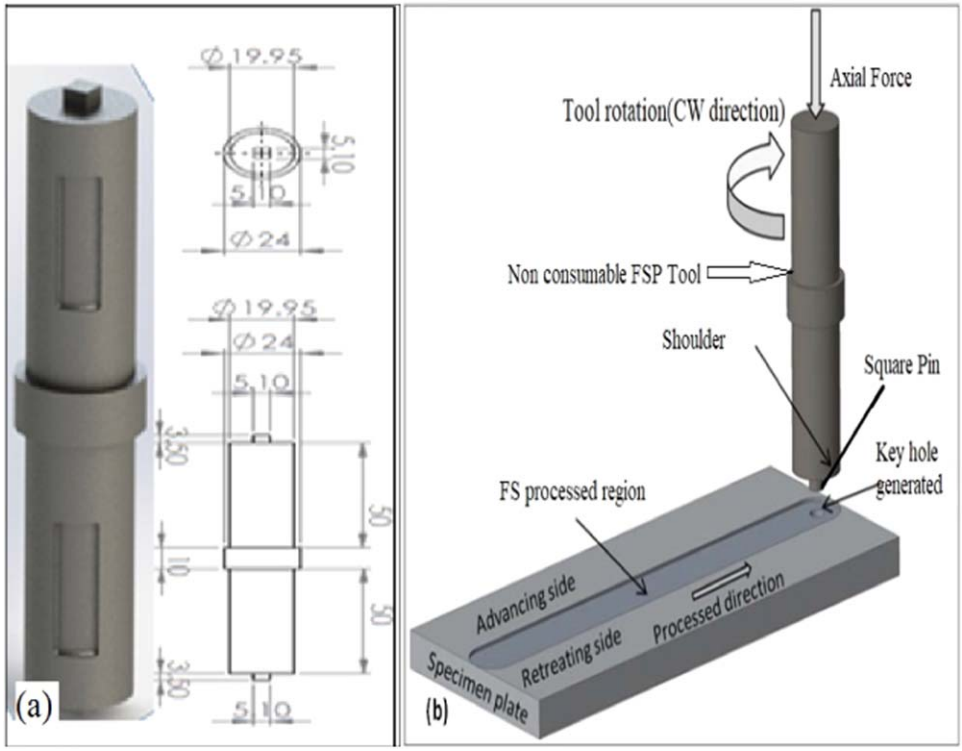


Figure 3. (a) Dimension of FSP square probe tool profile. (b) Schematic design of Friction stir processing.

CNC turning lathe machine tools was used to cover the groove to prevent the escape of B_4C nano-particles [42]. In figures 3(a), (b) shown, the dimension of square probe FSP tool and schematic design of the FSP, it initiates with a non-consumable rotating tool; inserted into the matrix material, which developed stringent plastic deformation of the matrix material due to conjoined action of tool stirring and friction generated heat over the matrix

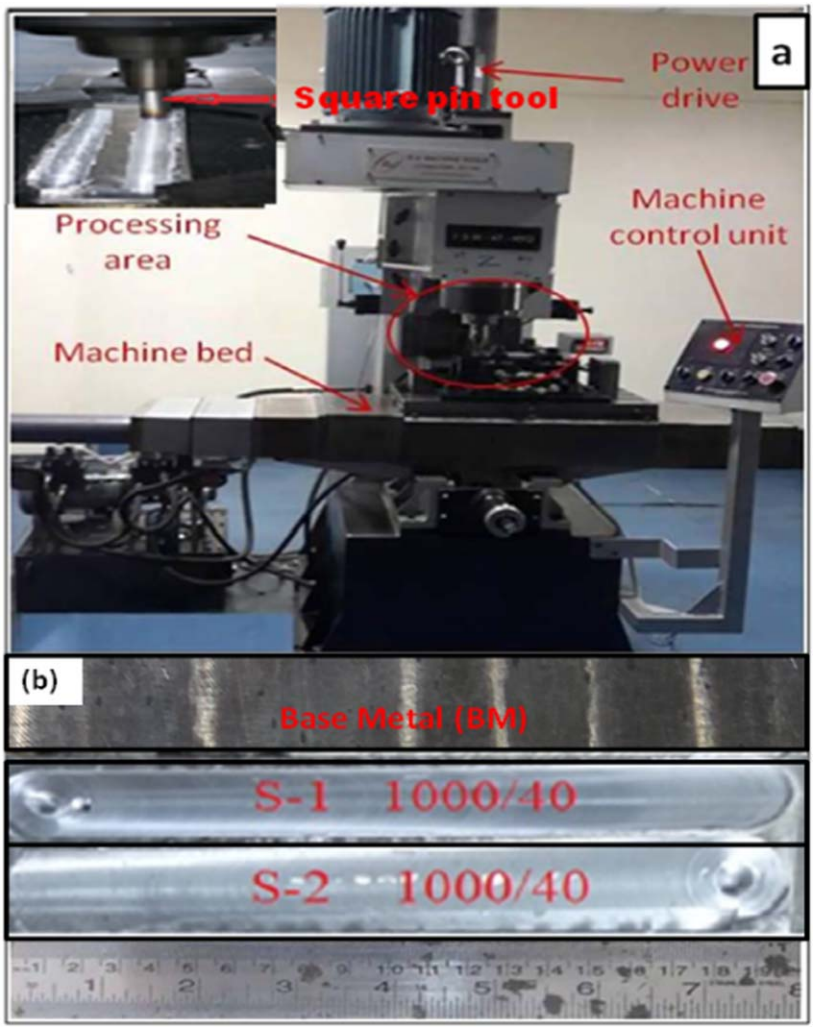


Figure 4. (a) FSP experimental setup (b) Sample of BM (Base Metal), S-1 (FSP on BM) and S-2 (BM + B_4C Nanoparticles).

material. Single-pass were perform using square tool pin of tool rotation and traverse speed of 1000 rpm and 40 mm min^{-1} respectively, tool tilt angle 2° were constant during experimentation. After the fabrication of AA7075-T6 and B_4C nano-ceramic surface composite, the sample was prepared from the FSPed region and then the sample was cut perpendicular to the processed direction using a wire EDM machine. The samples were prepared according to the dimension of microstructure, SEM, XRD. The samples were prepared for microstructural examination after polishing of different grade of emery paper and final polishing were done by wet polishing and etching.

Friction stir processing (FSP) is a new emerging, green and energy-efficient thermo-mechanical process for improvement in the material property by localized plastic deformation, which creates a microstructure with ultra-fine, equiaxed grains. In figures 4(a), (b), shown FSP experimental setup and processed sample, S-1 (FSP on BM) and S-2 (BM + B_4C Nanoparticles) and base metal(BM), the FSP process is a microstructural modification processing technique based on fundamental of Friction stir welding, a solid-state welding process. Friction stir processing uses the same process principles of FSW, but instead of joining the samples together, it modifies the local microstructure at the surface of a solid specimen to achieve specific and desired [38].

3. Results and discussion

3.1. Mechanical properties

Mechanical properties like tensile strength and microhardness value were notified from the experimental procedure. The average microhardness of the base metal was 102 Hv, the average hardness of sample (S-1) was 132 Hv, having a minimal concentration of B_4C nanoparticles, and the average hardness of sample (S-2) was 179 Hv.

The study of hardness characteristic concludes that the BM has the lowest hardness and the FSP sample varies with zero concentration and 0.5 moles of B_4C nanoparticles concentration. The tensile sample prepared as per ASTM standard and sample before and after the tensile test is shown in figures 5(a)–(c). The stress-strain diagram has shown in figure 6. The UTM result exhibit decrease in strength sample (S-1) and sample (S-2), but increase in ductility of S-1 (FSP on BM)

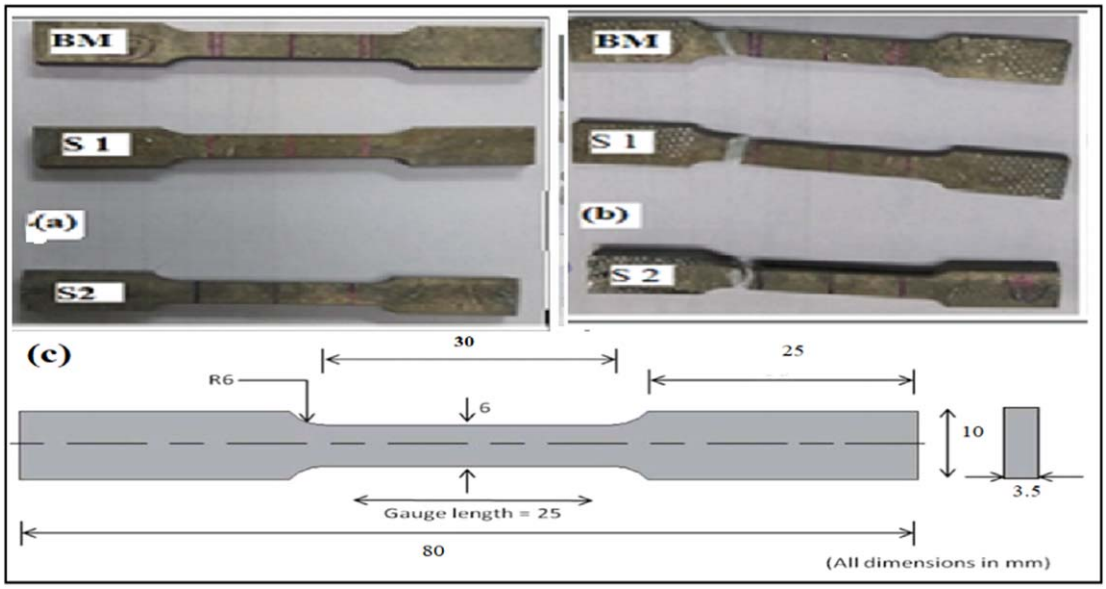


Figure 5. (a) Tensile test sample BM (Base Metal), S-1 (FSP on BM) and S-2 (BM + B_4C Nanoparticles) before fracture, (b) fracture after tensile test and (c) Prepared as per ASTM standard.

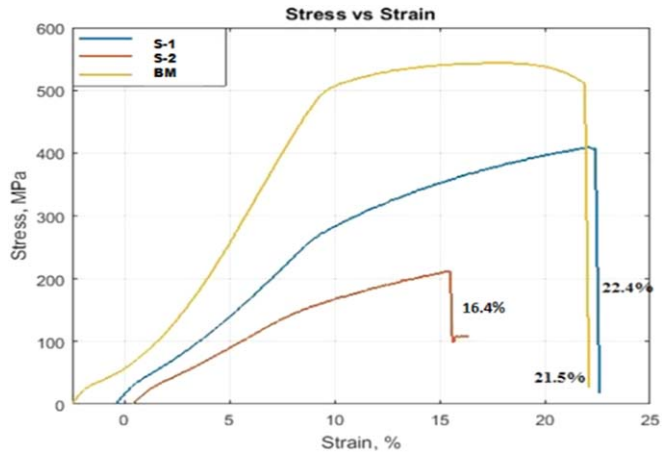


Figure 6. Stress-strain diagram shows the results of sample BM (Base Metal), S-1 (FSP on BM) and S-2 (BM + B_4C Nanoparticles) obtained from UTM test.

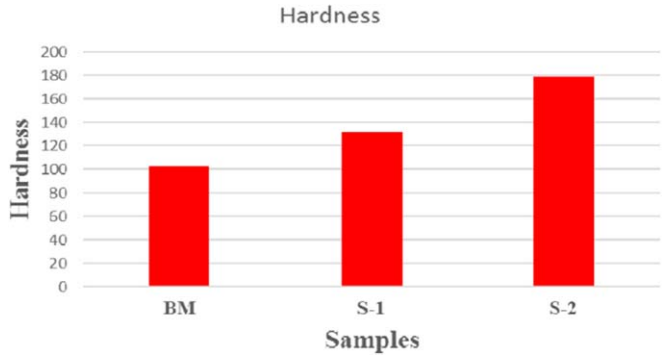


Figure 7. Effect of Microhardness at different sample.

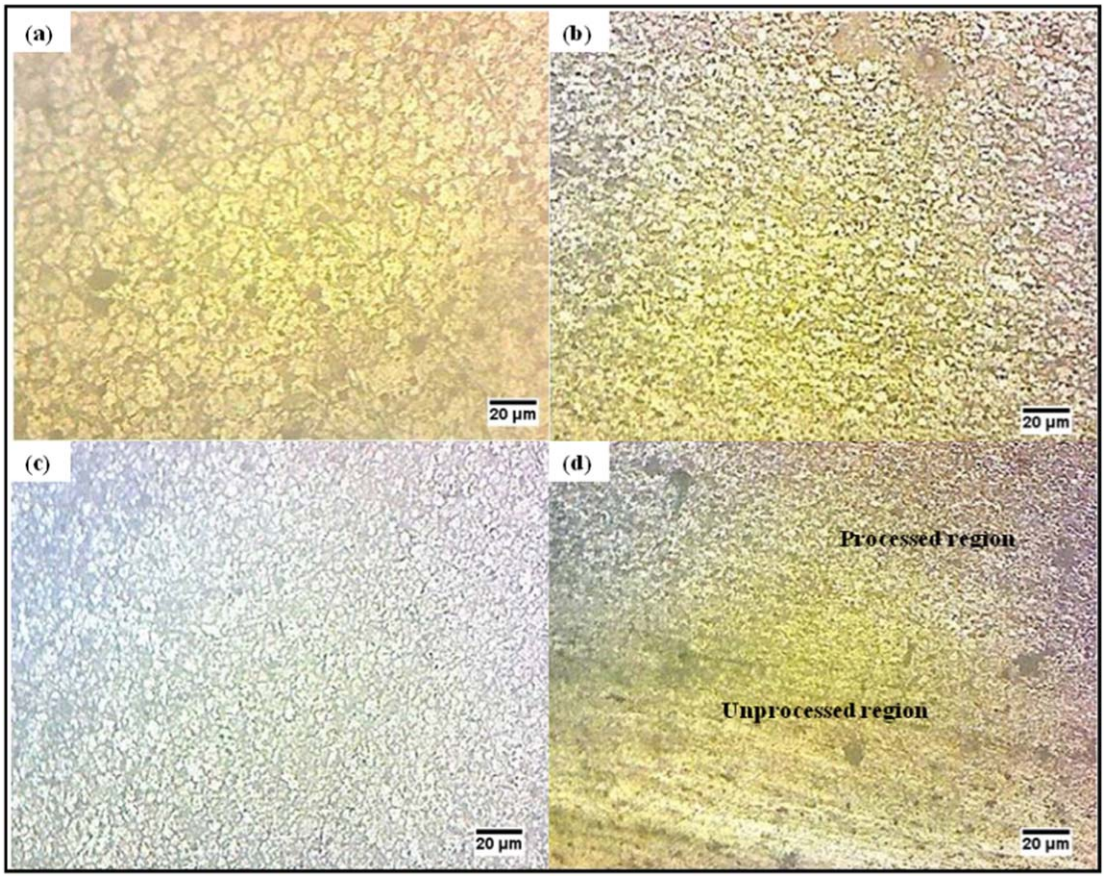


Figure 8. Microstructure images of different samples at 20 μm (a) Unprocessed surface (base metal), (b) Processed sample,S-1 (FSP on BM) (c) Processed sample, S-2 (BM + B_4C Nanoparticles) (d) Processed and unprocessed region.

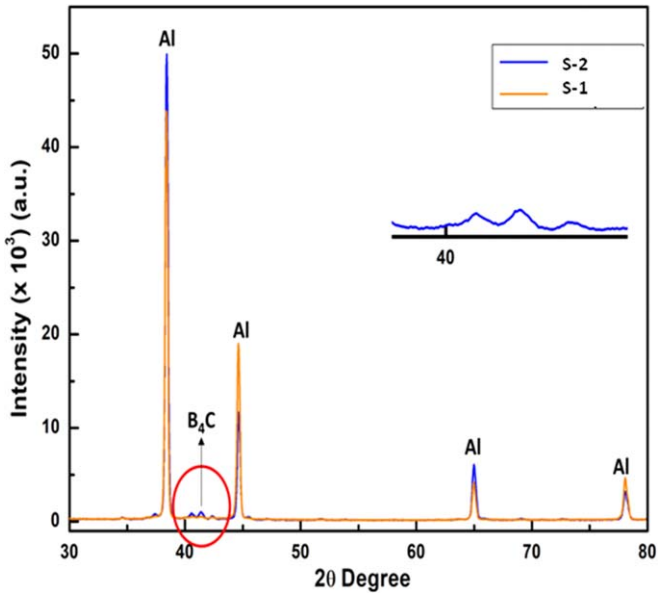


Figure 9. XRD pattern analyses of Sample S-1 (FSP on BM) and Sample S-2 (BM + B_4C Nanoparticles).

sample as compared to the base metal, in a sample S-2 (BM + B_4C Nanoparticles) with a 0.5 mole concentration of boron carbide nanoparticles, strength and ductility is lowest but significantly improve its hardness as compare in base metal (BM). The base metal (BM) exhibited elongation of 21.5% while the processed sample(S-1) showed the highest elongation of 22.4%.

The stress v/s strain graph plotted using Mat Lab is shown in figure 6. The graph compares the variation of tensile strength and percentage elongation for the three

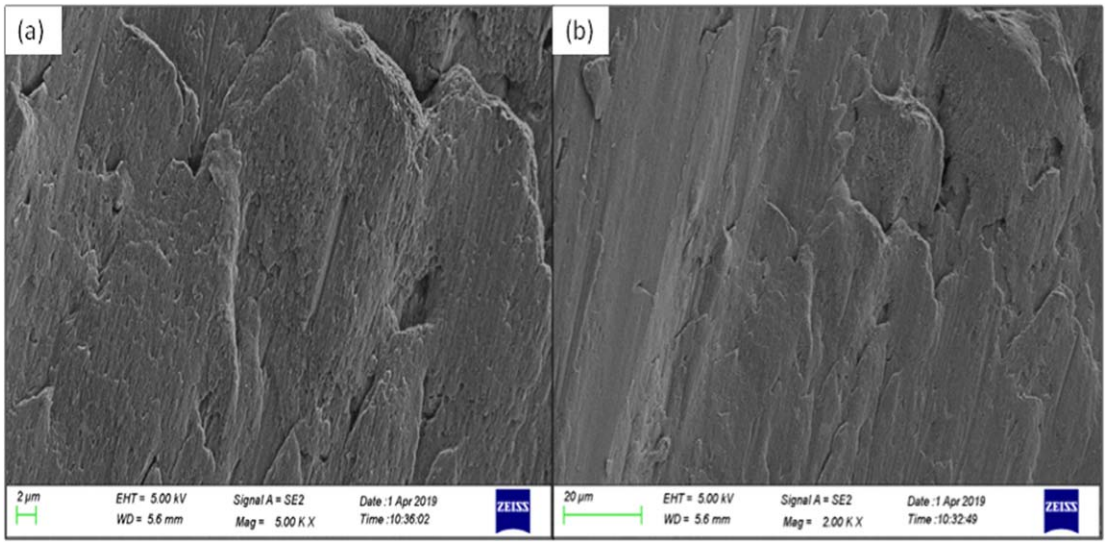


Figure 10. FESEM image of unprocessed surface (Base metal) (a) with 2 μm, (b) with 20 μm.

samples discussed. The tensile strength of sample S-2(BM + B_4C Nanoparticles) and S-1(FSP on BM) is decreasing as compared to the base metal (BM). This reduction in strength is a reason for grain growth that takes place during the Friction stir processing. Thus, a value of 284Mpa is observed for S-1 and an even smaller strength value for S-2 i.e. 152Mpa is observed. The change in ductility or percentage elongation behavior is characterized by the refinement of grain that takes place during FSP. Similar types of results have been reported by Faraji and Asadi [43] that the number of passes boosts the grain size refinement and thus due to the elimination of porosity a rise in ductility was observed. The percentage elongation for S-1 (FSP on BM) has therefore increased by 0.9%, i.e. arise from the base metal (BM) value. Yuvaraj *et al* [1, 32] have observed a similar type of report in their study of Al5083/B4C surface composite by FSP where the percentage elongation increases on performing FSP (without reinforcement). However, the presence of hard particles like B_4C nanoparticles in the present study contributes to reducing the ductility of the composite. Thus, the percentage elongation for S-2(BM + B_4C Nanoparticles) is very less i.e. 16.4% and the percentage elongation for S-1(FSP on BM) (22.4%) has increased by 4% rise from the base metal value.

3.2. Microhardness

The apparatus used for the testing was Fischerscope HM2000S microhardness testers, which measure the Vickers hardness (Hv). Vickers hardness test is an optical method, where the size of the indentation (the diagonals) is measured. The larger the size of the indentation left by intender on the surface of a workpiece or specimen, at a defined test force.

For experimental accuracy, the averages of four readings were taken for each sample. Microhardness test has been initiated at 0.300 N or 0.03 kg and the dwell time was 20 s. The average hardness was calculated using four readings on the same load. In figure 7, it has been presented that hardness variation is observed for base metal, without concentration S-1 (FSP on BM) and with concentration S-2 (BM + B_4C Nanoparticles) of reinforcement.

3.3. Microstructural analysis

For microstructure observations, Olympus GX41 inverted metallurgical optical microscope has been utilized and the microstructure images were captured at 20 μm. Width of the micrograph is measured as per the pixel to scale conversion according to 2 pixels per micron. Viewing the microstructure held to be imperative because of the momentous effects of the FSP technique; this was the resultant refinement of the grain structure. This would be translated into high-caliber mechanical properties imparted to the surface alteration of the material. As shown in figure 8(a) seen grain boundaries of unprocessed sample. In figure 8(b) processed sample, S-1 (FSP on BM) extreme plastic deformation and heating due to the friction during FSP between tool and matrix metal, lead to the origination of a recrystallized ultra-fine grained microstructure in the stirred zone of surface in comparison to the unprocessed sample. In figure 8(c) processed sample, S-2 (BM + B_4C Nanoparticles) has shown the recrystallized ultra-fine grain structure at same scale. The microstructure of processed and unprocessed region shown in figure 8(d) the well-defined nugget zone and thermos-mechanically affected zone (TMAZ) have been recognized.

3.4. XRD and EDS analysis

The FESEM micrograph of AA7075-T6/B4C composite using FSP was performed successfully and the FESEM images of the sample were taken from various

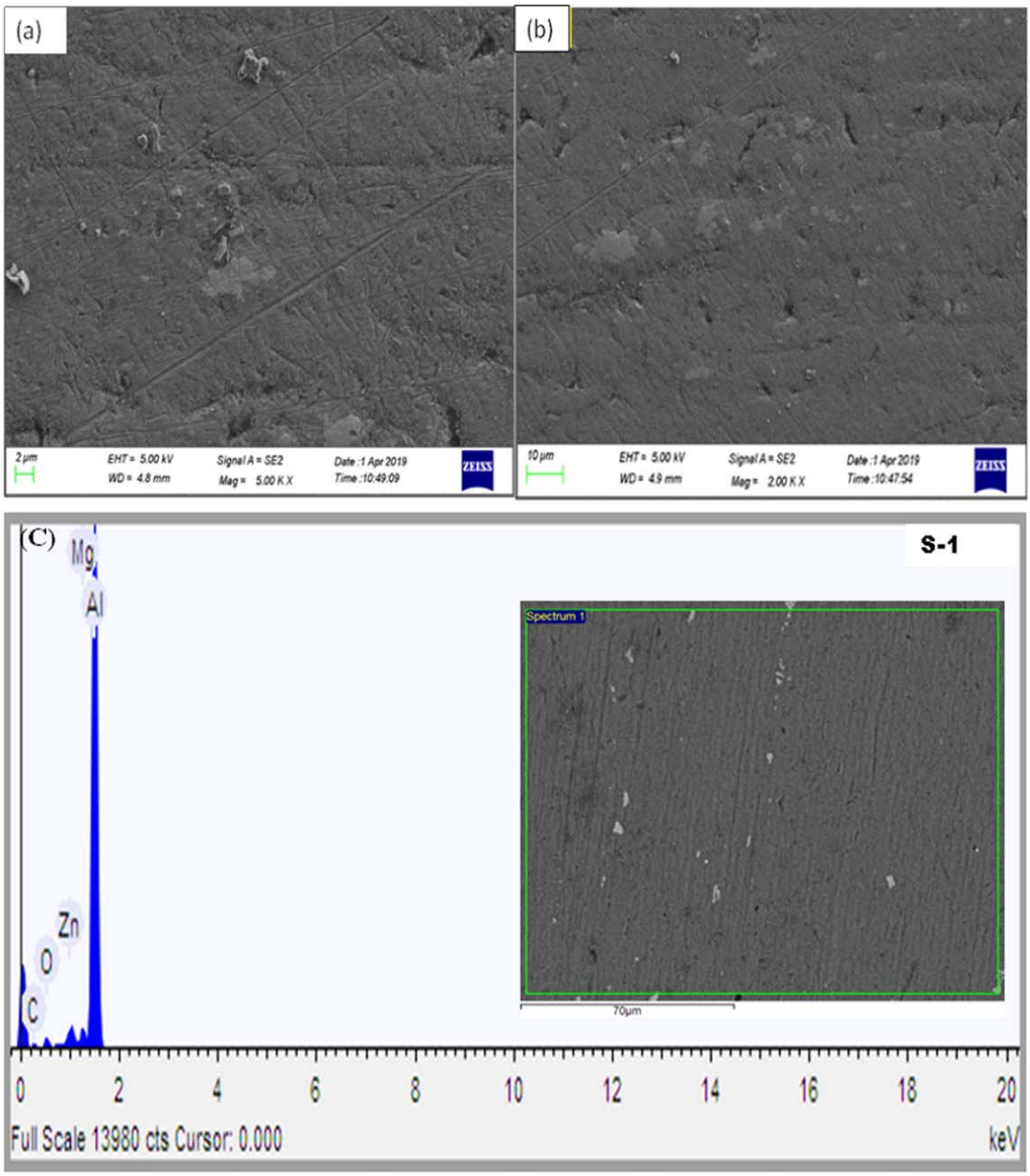


Figure 11. FESEM images of processed surface, S-1(FSP on BM) (a) with 2 μm, (b) with 10 μm and (c) EDS image.

regions. Figure 9 shows the XRD pattern and it has been confirmed the presence of boron carbide nanoparticle as reinforcement in the AA7075-T6 as a matrix material. Figures 10(a), (b) shows the FESEM micrograph of base metal and figures 11(a), (b) and 12(a), (b) show the effect of with or without concentration nanoparticles on the structure of FSP processed surface. In the figure 11(c) EDS is taken without a concentration of B_4C nanoparticles; its result shows that there are no B_4C particles present. In figure 12(c), the EDS result is taken for the processed sample with a 0.5-mole concentration of B_4C nanoparticles and it has clearly shown that the B_4C particle present in the sample(S-2). Moreover, as shown in figure 9 the highest Al peak is present and it was shifted towards 40° to 80°. A few weaker peaks were shown in figure 9, B_4C particle was observed due to 0.5-mole concentration of nanoparticles and no intermetallic elements were detected. Field Emission Scanning Electron Microscope (FESEM) model (ZeissSigma300) was used to examine the distribution of the nanoparticles. XRD (BRUKAR D8 ADVANCE) was used to identify the phase of reinforcement particles surface composite.

3.5. FESEM fractographs of tensile fracture

After the tensile test performed in UTM the next step is to cut the sample as per the standard dimension of the FESEM setup. Field emission scanning electron microscope (FESEM) model (ZeissSigma300) was used to examine the fracture point of BM alloys, S-1(FSP on BM) and S-2(BM + B_4C Nanoparticles). In figures 13(a) and (b) FESEM images with 1μm and

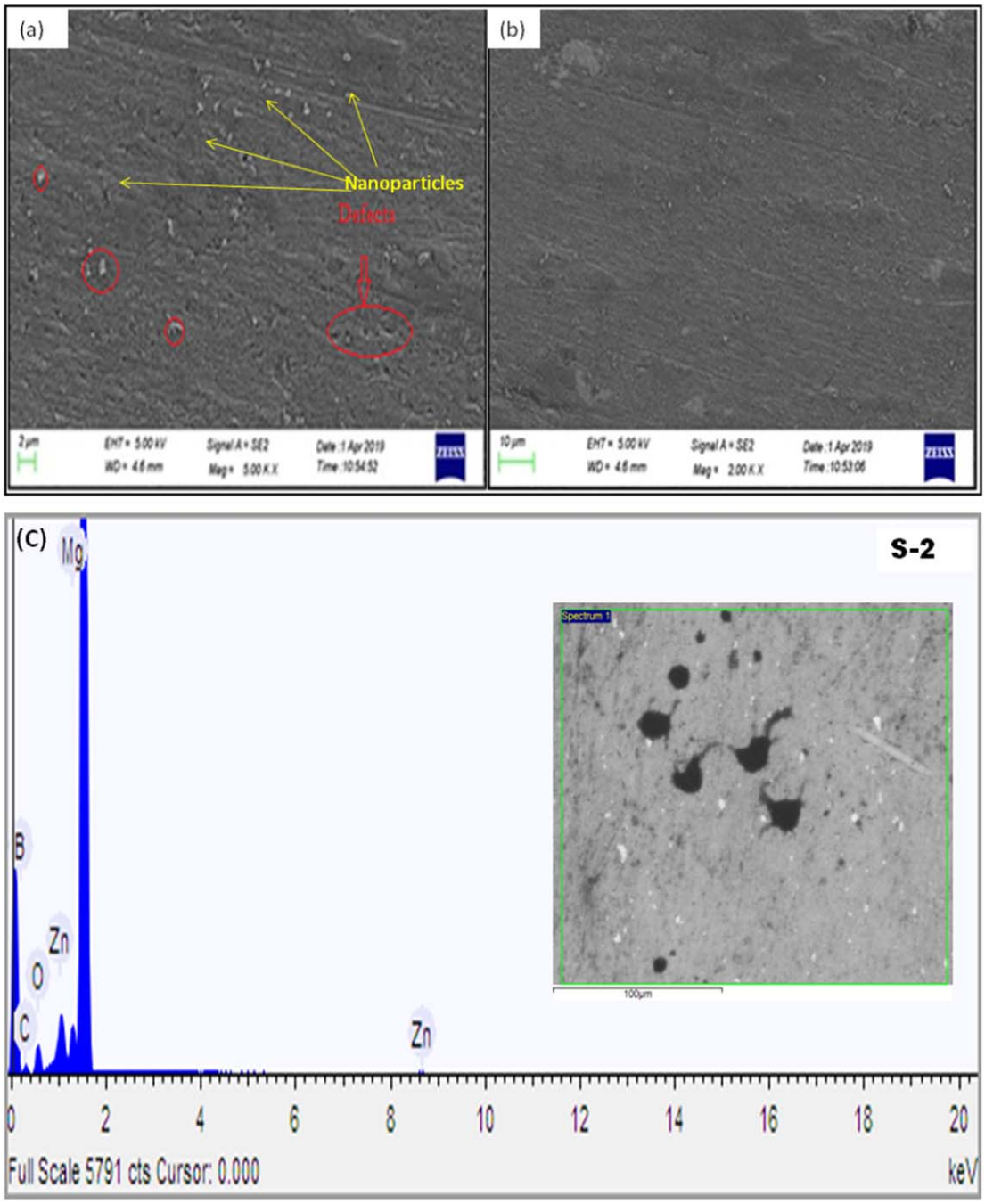


Figure 12. FESEM images of processed surface, S-2 (BM + B_4C Nanoparticles) (a) with 2 μm, (b) with 10 μm and (c) EDS image.

10 μm, are shown morphologies of the fracture surface of base metal (BM) has a small number of dimples which indicate the elongation of BM alloys. The traverse fracture of the sample, S-1(FSP on BM) shown in figures 14(a) and (b) FESEM images with 1μm and 10 μm, indicate the minuscule dimple and cleavage facets. The superior ductile behavior exhibited shown in figure 14(a), it justifies the appearance of the fractured surface of the sample, and S-1 (FSP on BM) has a large number of small dimples. The sample(S-1) undergoes significant elongation before fracture. The traverse fracture of the sample, S-2(BM + B_4C Nanoparticles) shown in figures 15(a) and (b) was characterized by the cleavage facets, large dimple and homogeneous distribution of boron carbide nanoparticle, which increases the hardness of the surface composite. After the fracture of the tensile test the area of fracture zone of the sample (S-2), which has a 0.5-mole concentration of boron carbide nanoparticles as reinforcement, crack formation occurred due to the brittle nature of the surface, which has significantly improved the hardness of the materials.

3.6. Coefficient of friction

Friction and wear are a result of complex and multiplex sets of microscopic interactions and thermo-mechanical embodiment between surfaces that are in

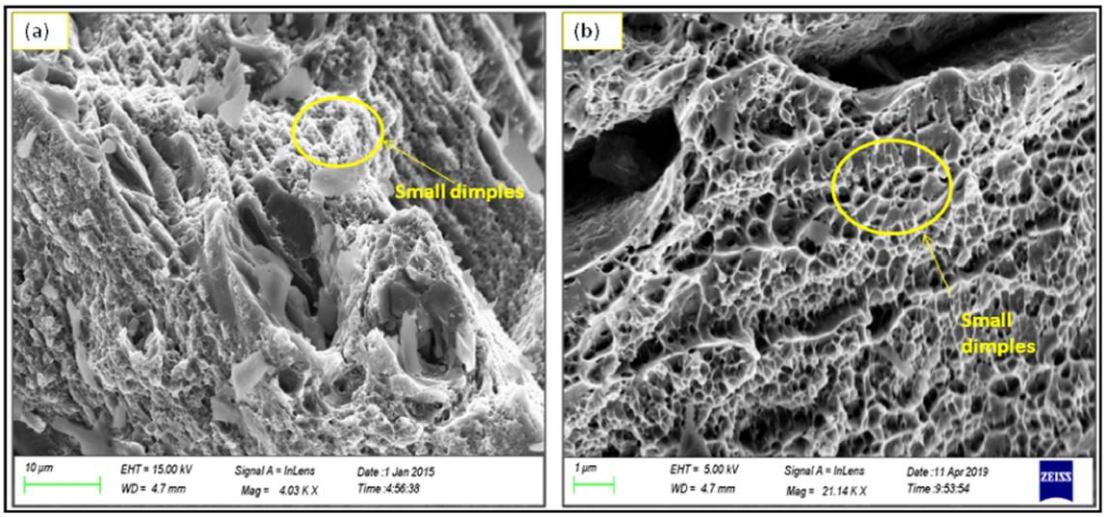


Figure 13. (a), (b) FESEM fractographs image of tensile fracture surface of unprocessed sample (Base metal).

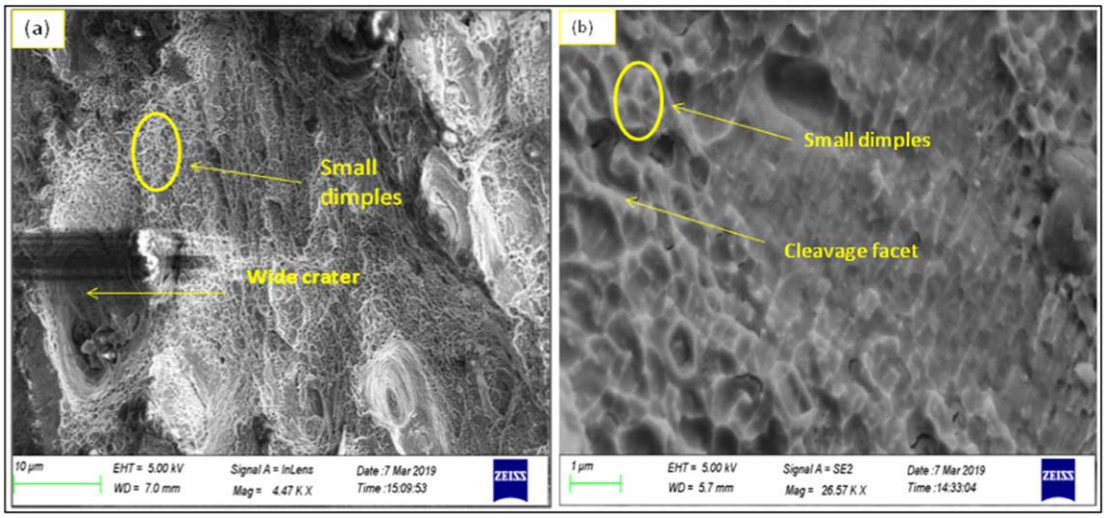


Figure 14. (a), (b) FESEM fractographs image of tensile fracture surface of processed sample, S-1 (FSP on BM).

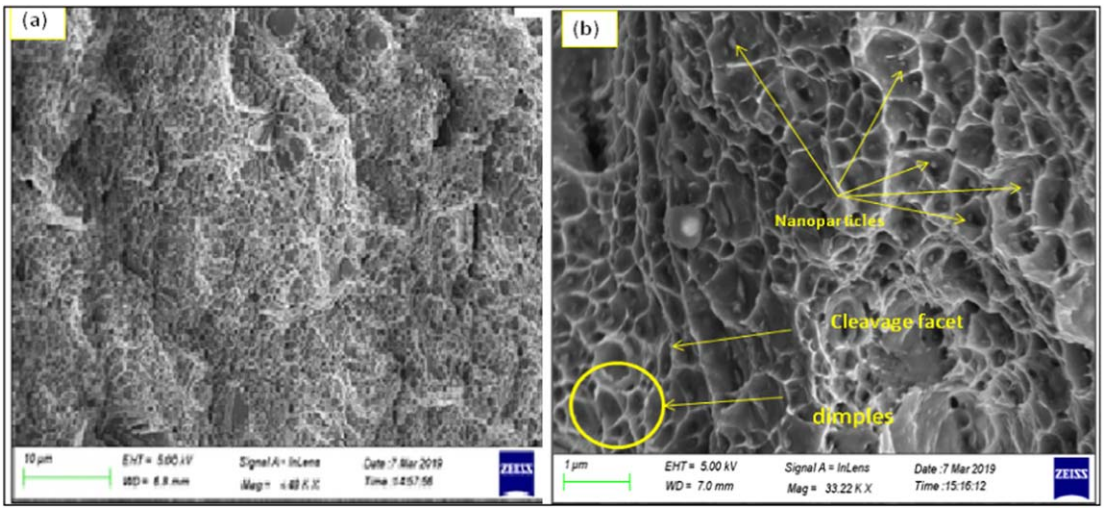


Figure 15. (a), (b) FESEM fractographs image of tensile fracture surface of processed sample, S-2 (BM + B_4C Nanoparticles).

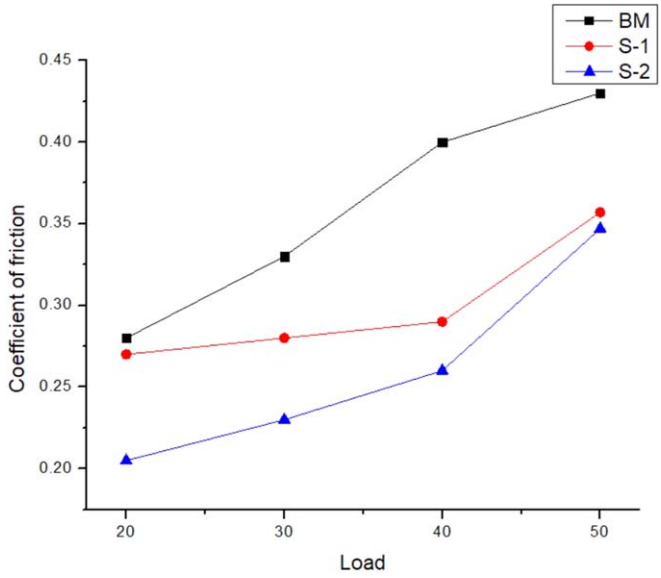


Figure 16. Effect of load applied on coefficient of friction at different sample.

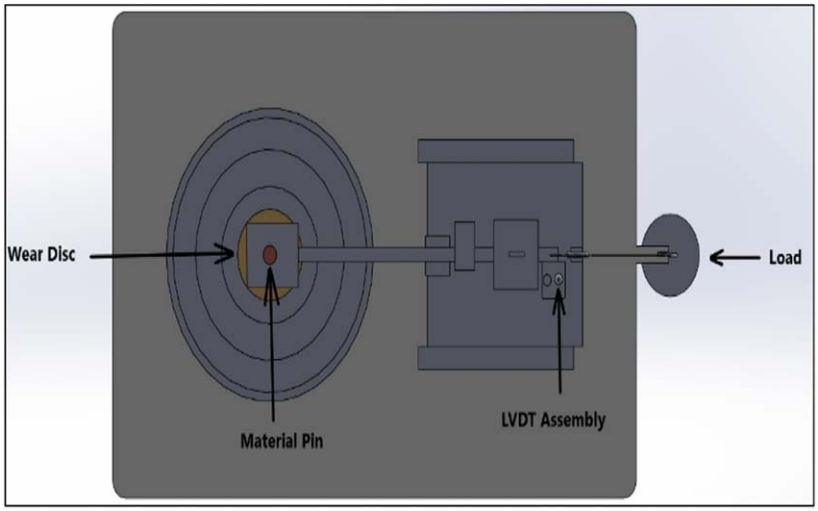


Figure 17. Schematic diagrams of Friction/Wear Apparatus –Disc Tribometer.

mechanical contact and drag against each other. These synergic actions are the result of the material characteristics, the geometrical and topographical properties of the surfaces, and the accompanying physical situations under which the surfaces are subjected to the sliding action against each other, e.g., loading, temperature, atmosphere, type of contact, etc.

Figure 16 shows the variation of the coefficient of friction of base metal (BM) aluminum alloy 7075-T6 and fabricated surface composite samples with or without concentration of nanoparticles under four normal sliding loads. At 20 N load there is large variation of friction is observed sample (S-1) and (S-2). The coefficient of friction samples (S-1) and (S-2) were lower than the base metal (BM) in all sliding loads. In base metal (BM) the coefficient of friction is linearly increase with increasing load. In sample(S-1) without concentration of B_4C nanoparticles, initially 20 N to 30 N the friction coefficient is approximately same and after that friction coefficient increases with increasing load. In sample (S-2) with 0.5 moles concentration of B_4C nanoparticles, coefficient of friction value with gradually increase with increasing load is observed due to the formation of homogeneous distribution of nanoparticles.

Aforementioned schematic pin on disc set-up in figure 17 illustrates the dry sliding wear tests performed which utilized the square probe over the disc tribometer as per ASTM G99-04 standard. Cylindrical wear test pin of 10 mm diameter and 6 mm height

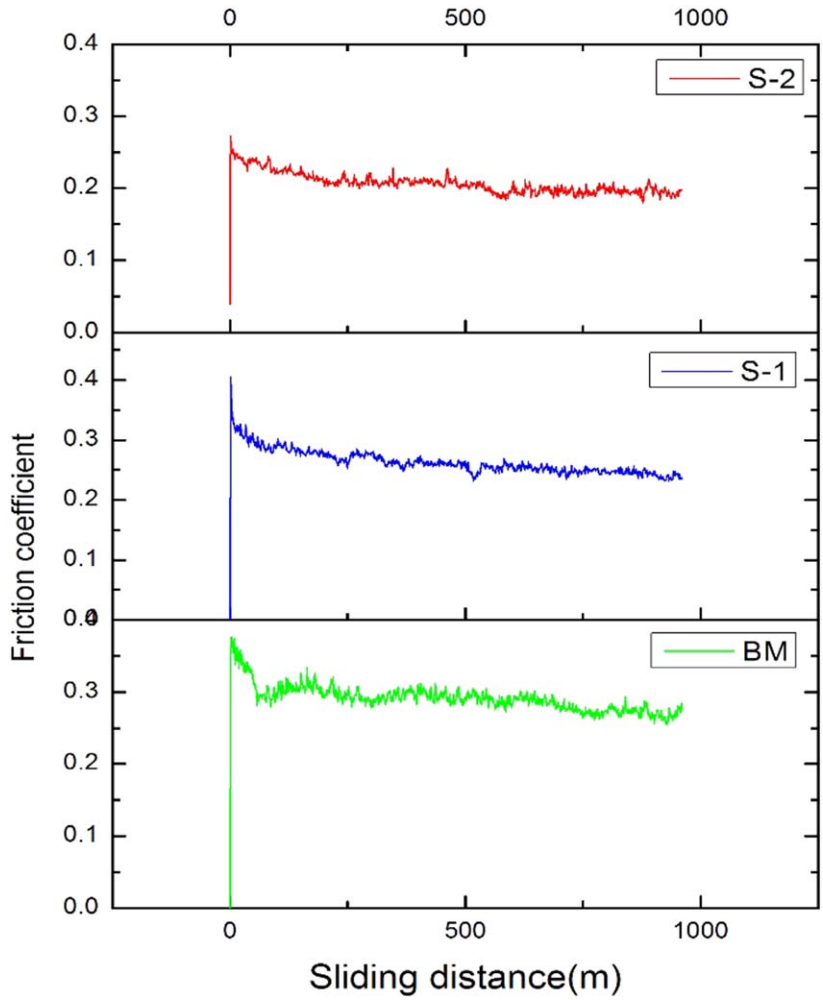


Figure 18. Variation of friction coefficient on sliding distance at different sample.

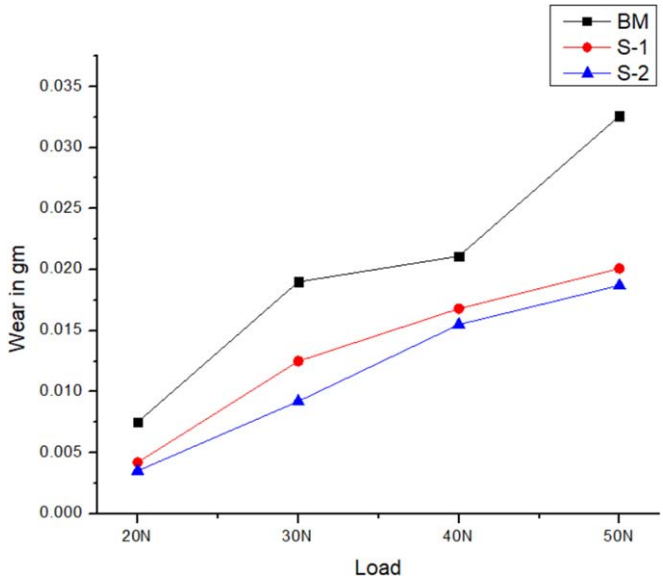


Figure 19. Effect of load on wear in gram at different sample.

were cut from the processed region by wire EDM. Wear test sample tested against EN-24 steel circular disc of 100 mm diameter hardened to 60–65 HRC. Wear test was carried out at a constant sliding speed of 1 m s^{-1} and normal load of 20,30,40 and 50 N with a constant sliding distance of 1000 m. Track diameter varies from 30 to 60 mm respectively was used in wear tests.

Before each wear test the pin and disc were cleaned with acetone before and after the test. All the pin samples were weighed before and after the wear tests performed by an electronic weighing machine. It is observed from the figure 18, that friction coefficient decreases with increases in sliding distance but on comparison of friction coefficient for BM, processed sample, S-1 (FSP on BM) and processed sample (S-2) with 0.5 mole concentration of B_4C nanoparticles. It is also, observed that there is large fluctuation in sample (S-2), due to the formation of hard particles in the nano surface composite. The friction coefficients of nano surface composite sample (S-2) are lower than those of BM alloys sample and sample, S-1(FSP on BM).

Friction and wear can be defined as the responses of a tribological system. Coefficient of friction and wear define the state of contact of bodies not the material properties of the bodies. Aforementioned experimental data analysis between S-1 and S-2 samples compared to BM is quite distinct and characteristic. The required hypothesis of performance of S-1 and S-2 compared to BM is keenly observed here in the wear versus load as shown in figure 19. The Friction and Wear data are in close correspondence to each other, highlighting that as load increases, friction and wear increase. Processed Sample, S-1 (FSP on BM) and Sample, S-2 (BM + B_4C Nanoparticles) processed with 0.5 mole concentration of B_4C nanoparticles, here both these samples have performed better than BM in having the minimized wear, hence maximizing their durability in the quantitative process of loading the metal matrix composite. At 20 N, the wear is quite approximately the same for S-1(FSP on BM) and S-2(BM + B_4C Nanoparticles) samples but as we progress, S-2 shows minimized wearing than S-1 sample by a safe margin showing the effectiveness of B_4C nanoparticle concentrations, whereas BM undergoes high amount of wearing which exemplifies its poor mechanical durability character. The margin of experimental wear observed here between S-1(FSP on BM) and S-2(BM +B_4C Nanoparticles) is appreciable.

4. Conclusion

The Self-Assemble Monolayer (SAM) technique of Al-B_4C nano-ceramic surface composite were successfully fabricated using Friction stir processing method. The mechanical properties of fabricated nano ceramic surface composites were evaluate the following outcomes.

1. The microstructure of the processed sample, (S-1) FSP without B_4C nanoparticles and sample sample, (S-2) with B_4C nanoparticles one pass exhibits fine grain structure, higher hardness (132 Hv) and (179 Hv), ultimate tensile strength (284Mpa) and (152Mpa) respectively in comparison to the base metal(BM) hardness (102 Hv), ultimate tensile strength (503Mpa).
2. The microhardness of the sample (S-2) with B_4C nanoparticles mixed with 0.5mole concentration of diethyl ether (DEE) is much higher hardness in comparison to the without concentration i.e. sample, (S-1) and base metal.
3. Significant grain refinement occurs in the FSP procedure to the intense plastic deformation recrystallization, which occurs due to frictional heating during FSP. This results in the generation of a recrystallized fine-grained microstructure within the stirred zone. Nanoparticles were seen in the processed samples (S-2) at high resolution. They were indicative of the presence of B_4C particles.
4. FESEM images, EDS composition and XRD pattern confirms the presence of B_4C nanoparticle as reinforcement in modified surface.
5. Frictional and Wear analysis highlights the mechanical durability and surface characteristics of BM and S-1, S-2 samples of metal matrix composites. Experimental data suggests that Sample (S-2) with B_4C nanoparticles has the least wearing and frictional coefficient due to the effective application and processing of B_4C nanoparticles compared to Sample, S-1 FSP without B_4C nanoparticles. BM has been reported to have least durability and highest frictional coefficient.

Acknowledgments

The authors acknowledge Laser Spectroscopy Laboratory (at the Delhi Technological University) for access to physical apparatus and fabricators and heartily send regards to Dr Mohan S Mehta, for assisting in processing of Self-Assembled Monolayer for the experimental analysis.

ORCID iDs

Ravi Butola https://orcid.org/0000-0002-6581-4131
Qasim Murtaza https://orcid.org/0000-0001-8784-2304
Ranganath M Singari https://orcid.org/0000-0003-4981-3857

References

[1] Yuvaraj N, Aravindan S and Vipin 2015 Fabrication of Al5083/B4C surface composite by friction stir processing and its tribological characterization *Journal of Material Research Technology*. **4** 398–410

[2] Miracle D B 2005 Metal matrix composites–from science to technological significance *Compos. Sci. Technol.* **65** 26–40

[3] Kalaiselvan K, Murugan N and Parameswaran S 2011 Production and characterization of AA6061–B4C stir cast composite *Mater. Des.* **32** 4004–9

[4] Abenojar J, Velasco F and Martinez M A 2007 Optimization of processing parameters for the Al + 10%B4C system obtained by mechanical alloying *J. Mater. Process. Technol.* **184** 441–6

[5] Khakbiz M and Akhlaghi F 2009 Synthesis and structural characterization of Al–B4C nano-composite powders by mechanical alloying *J. Alloys Compd.* **479** 334–41

[6] Coyal A, Yuvaraj N, Butola R and Tyagi L 2020 An experimental analysis of tensile, hardness and wear properties of aluminium metal matrix composite through stir casting process *SN Applied Sciences* **2** 892

[7] Butola R, Malhotra A, Yadav M, Singari R, Murtaza Q and Chandra P 2019 Experimental Studies on Mechanical Properties of Metal Matrix Composites Reinforced with Natural Fibres Ashes, SAE Technical paper. 2019-01-1123 (https://doi.org/10.4271/2019-01-1123)

[8] Butola R, Pratap C, Shukla A and Walia R S 2019 Effect on the mechanical properties of aluminum-based hybrid metal matrix composite using stir casting method *Mater. Sci. Forum* **969** 253–9

[9] Bradeswaran A and Elaya Perumal A 2013 Influence of B4C on the tribological and mechanical properties of Al 7075–B4C composites *Composites: Part B.* **54** 146–52

[10] Topan F, Kilicarslan A, Karaaslan A, Cigdem M and Kerti I 2010 Processing and microstructural characterization of AA 1070 and AA 6063 matrix B4Cp reinforced composites *Mater. Des.* **31** 87–91

[11] Kerti I and Toptan F 2008 Microstructural variations in cast B4C-reinforced aluminium matrix composites (AMCs) *Mater. Lett.* **62** 1215–8

[12] Ubaid F, Matli P R, Shakoor R A, Parande G, Manakari V, Mohamed A M A and Gupta M 2017 Using B4C nanoparticles to enhance thermal and mechanical response of aluminum *Materials* **10** 621

[13] Alihosseini H, Dehghani K and Kamali J 2017 Manufacturing seamless square tubes of B4C-reinforced aluminum composites by extrusion *Int. J. Adv. Manuf. Technol.* **90** 1921–30

[14] Liu B, Huang W M, Wang H W, Wang M L and Li X F 2013 Compressive behavior of high particle content B4C/Al composite at elevated temperature *Trans. Nonferrous Met. Soc. China* **23** 2826–32

[15] Moona M, Walia R S, Rastogi V and Sharma R 2019 Parameter optimization and characterization of environmental friendly aluminium hybrid metal matrix composites *Mater. Res. Express* **6** 1165d5

[16] Butola R, Ranganath M S and Murtaza Q 2019 Fabrication and optimization of AA7075 matrix surface composites using Taguchi technique via friction stir processing (FSP) *Eng. Res. Express* **1** 025015

[17] Chaudhary A, Dev A K, Goel A, Butola R and Ranganath M S 2019 The mechanical properties of different alloys in friction stir processing: a review *Materials Today: Proceedings* **5** 5553–62

[18] Yuvaraj N, Aravindan S and Vipin 2017 Comparison studies on mechanical and wear behavior of fabricated aluminum surface nano composites by fusion and solid state processing *Surface & Coatings Technology* **309** 309–19

[19] Chen X, Luais E, Darwish N, Ciampi S, Thordarson P and Gooding J J 2012 Studies on the effect of solvents on self-assembled monolayers formed from organophosphonic acids on indium tin oxide American Chemical Society *Langmuir* **28** 9487–95

[20] Hoqu E, De RJ, Bhushan B and T Mathieu H 2008 Self-Assembled Monolayers on Aluminum and Copper Oxide Surfaces: Surface and Interface Characteristics, Nano tribological Properties, and Chemical Stability *Applied Scanning Probe Methods IX. Nano Science and Technolgy* ed Tomitori M *et al* (Berlin Heidelberg: Springer) pp 235–81

[21] Fangstrom T, Kirrander A, Erikssona L A and Lunella S 1998 The reaction between aluminium and dimethyl ether Comparative study of density functional theory and EPR results *J. Chem. Soc. Faraday Trans.* **94** 777È782 777

[22] Wanga D, Nia Y, Huoa Q and Tallmanb D E 2005 Self-assembled monolayer and multilayer thin films on aluminum 2024-T3 substrates and their corrosion resistance study *Thin Solid Films* **471** 177–85

[23] Sharma V, Gupta Y, Kumar B M and Prakash U 2016 Friction stir processing strategies for uniform distribution of reinforcement in a surface composite *Mater. Manuf. Processes* **31** 1384–92

[24] Sun N and Apelian D 2015 Composite fabrication in cast Al A206 via friction stir processing *Int. J. Cast Met. Res.* **28** 72–80

[25] Khodabakhshi F, Simchi A, Kokabi A, Sadeghahmadi M and Gerlich A 2015 Reactive friction stir processing of AA 5052–TiO_2 nanocomposite: process–microstructure–mechanical characteristics *Mat. Sci. Technol.* **31** 426–35

[26] Hussain G, Hashemi R, Hashemi H and Al-Ghamdi K A 2016 An experimental study on multi-pass friction stir processing of Al/TiN composite: some microstructural, mechanical, and wear characteristics *Int. J. Adv. Manuf. Technol.* **84** 533–46

[27] Dehghani K and Mazinani M 2011 Forming nanocrystalline surface layers in copper using friction stir processing *Mat. Manuf. Processes.* **26** 922–5

[28] Barmouz M, Asadi P, Givi M B and Taherishargh M 2011 Investigation of mechanical properties of Cu/SiC composite fabricated by FSP: effect of SiC particles size and volume fraction *Mat. Sci. Eng.* A **528** 1740–9

[29] Joshi S, Singh R C and Chaudhary R 2019 Effect of rotational speed in Friction Stir Processing on the microstructural and mechanical characteristics of cast AS21A magnesium alloy *Mater. Res. Express* **6** 056554

[30] Jamshidijam M, Akbari-Fakhrabadi A, Masoudpanah S M, Hasani G H and Mangalaraja R V 2013 Wear behavior of multiwalled carbon nanotube/AZ31 composite obtained by friction stir processing *Tribol. Trans.* **56** 827–32

[31] Asadi P, Givi M B, Parvin N, Araei A, Taherishargh M and Tutunchilar S 2012 On the Role of cooling and tool rotational direction on microstructure and mechanical properties of friction stir processed AZ91 *Int. J. Adv. Manuf. Technol.* **63** 987–97

[32] Yuvaraj N, Aravindan S and Vipin 2017 Wear characteristics of Al5083 surface hybrid nano composites by friction stir processing *Trans. Indian Inst. Met.* **70** 1111–29

[33] Komarasamy M, Mishra R S, Baumann J A, Grant G and Hovanski Y 2013 Processing, microstructure and mechanical property correlation in Al-B4C surface composite produced via friction stir processing *Friction Stir Weld Process. VII* (Cham, Switzerland: Springer International Publishing) 39–46

[34] Liu Q, Ke L, Liu F, Huang C and Xing L 2013 Microstructure and mechanical property of multi-walled carbon nanotubes reinforced aluminum matrix composites fabricated by friction stir processing *Mater. Des.* **45** 343–8

[35] Ni D, Wang J, Zhou Z and Ma Z 2014 Fabrication and mechanical properties of bulk NiTip/Al composites prepared by friction stir processing *J. Alloys Compd.* **586** 368–74

[36] Du Z, Tan M-J, Guo J-F and Wei J 2016 Friction stir processing of Al–CNT Composites *Proceedings of the Institution of Mechanical Engineers, Part L. J. Mater. Des. Appl.* **230** 825–33

[37] Kurt A, Uygur I and Cete E 2011 Surface modification of aluminium by friction stir processing *J. Mater. Process. Technol.* **211** 313–7

[38] Mishra R S, Ma Z and Charit I 2003 Friction stir processing: a novel technique for fabrication of surface composite *Mater. Sci. Eng,* A **341** 307–10

[39] Mahmoud E R I, Takahashi M, Shibayanagi T and Ikeuchi K 2009 Effect of friction stir processing tool probe on fabrication

of SiC particle reinforced composite on aluminium surface *Sci. Technol. Weld. Joining* **14** 413–25

[40] Vijay S J and Murugan N 2010 Influence of tool pin profile on the metallurgical and mechanical properties of friction stir welded Al–10wt.% TiB2 metal matrix composite *Mater. Des.* **31** 3585–9

[41] Al-Ghamdi K A, Hussain G and Hashemi R 2015 Fabrication of metal-matrix AL7075T651/TiN nano composite employing friction stir process *Proc. IMechE Part B: J Engineering Manufacture.* **231** 1319–31

[42] Butola R, Murtaza Q and Singari R M 2019 Advances in computational methods in manufacturing *Lecture Notes on Multidisciplinary Industrial Engineering* ed R Narayanan *et al* (Singapore: Springer) pp 337–48

[43] Faraji G and Asadi P 2011 Characterization of AZ91/alumina nanocomposite produced by FSP *Mater. Sci. Eng.* A **528** 2431–40

2020 6th International Conference on Control, Automation and Robotics

Fractional Order Capacitors Based Filters Using Three OTAs

Gopal Singh
Electrical Engineering Department
Delhi Technological University
Delhi, India
e-mail: gopal_mt2k18@dtu.ac.in

Garima
Electrical Engineering Department
Delhi Technological University
Delhi, India
e-mail: garima.dtu@gmail.com

Pragati Kumar
Electrical Engineering Department
Delhi Technological University
Delhi, India
e-mail: pragatikumar@dtu.ac.in

***Abstract*—This paper is concerned with designing a fractional order filters in voltage mode (VM) employing three single output operational trans-conductance amplifiers (OTAs) and two fractional order capacitors. The presented structure can realize all second-order filters functions namely, fractional order low pass filter (FLPF), fractional order high pass filter (FHPF), fractional order band pass filter (FBPF), fractional order band reject filter (FBRF) and fractional order all pass filter (FAPF) in voltage mode (VM) using appropriate selection(s) of input signal(s). PSPICE simulation results using fractional order capacitor from α = 0.7 to α = 0.9 (designed using Valsa approximation) have been presented.**

Keywords-operational trans-conductance amplifier; fractional order capacitor; fractional order filters; multi input single output filters

I. INTRODUCTION

Ever since the generalization of the classical first order and second order filters came into fractional order domain [1]-[2], a multitude of fractional order filters with different properties have appeared in open literature. In a fractional order filter the designers have a more precise control over the pass-band/ stop-band characteristics because of the additional degree of freedom provided by the fractional order parameter 'α'. These filters have been realized using different types of active elements like operational amplifiers [3]-[4], current conveyors [5], current feedback amplifiers [6], operational trans-conductance amplifiers [7]-[19] and variety of other active elements. As the fractional order capacitors of arbitrary values and order are not available commercially, in most of the works presented on fractional order filters, the fractional order capacitor(s) have been simulated by semi-infinite RC networks in which the values of the RC components have been computed using some approximations [20]-[21]. Of the various active elements used for realization of fractional order filters, operational trans-conductance amplifiers (OTAs) are more versatile as the gain of this amplifier can be controlled externally by controlling its bias current. This feature is very useful in tuning the various parameters of the realized filters. A detailed review of the various fractional order filters realized with OTAs [7]-[19] has revealed the following:

(i) Single-input-single-output (SISO) type of fractional order filters utilizing six to eleven number of OTAs have been presented in [7]-[13].

(ii) Single-input-multiple-output (SIMO) type of fractional order filters have been realized in [14]-[15].

(iii) Multi-input-single-output (MISO) type of fractional order filter has been realized in [14]. This structure utilizes 5 OTAs and two fractional order capacitors while realizing different types of filters.

(iv) OTAs along with some other additional active elements [16]-[19] have been used to realize fractional order filters with different properties.

From the above discussion, it emerges that very little work has been done on the realization of multi-input-single-output type of fractional order filters using OTAs. These type of filter structures are very useful as different filter outputs may be obtained without changing the nature of the elements in different branches. Therefore, the main aim of this paper is to present a new multi-input-single-output type fractional order filter utilizing only three single-output OTAs and two fractional order capacitors.

II. CIRCUIT DESCRIPTION

The operational trans-conductance amplifier is a differential voltage controlled current source (DVCCS). Ideally in OTAs, the output current is a function of differential input voltage and is expressed as:

$$I_0 = g_m(V_1 - V_2) \tag{1}$$

The trans-conductance of this source can be controlled by an external bias current (I_{bias}) as given by equation (2).

$$g_m = \frac{I_{bias}}{2V_T} \tag{2}$$

978-1-7281-6139-6/20/$31.00 ©2020 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

The OTA is represented by the symbol shown below in Fig. 1.

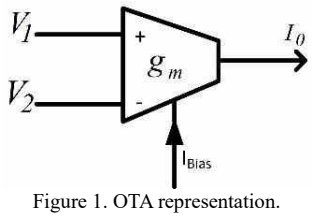


Figure 1. OTA representation.

The proposed filter configuration is shown below in Fig. 2. The configuration belongs to the class of multiple-input single-output (MISO) type of filter structures where, by appropriately choosing different inputs, various types of filter responses can be obtained. We have used two identical fractional order capacitors (C_α) whose driving point impedance is given by $1/Cs^\alpha$.

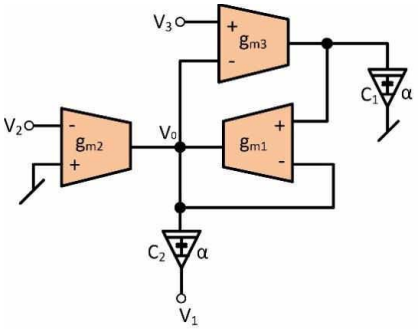


Figure 2. Proposed voltage mode filter configuration.

A routine analysis of the circuit given above results in the following output function:

$$V_0(s)=\frac{s^{2\alpha}(C_1C_2)V_1(s)-s^{\alpha}(C_2g_{m2})V_2(s)+(g_{m1}g_{m3})V_3(s)}{s^{2\alpha}C_1C_2+s^{\alpha}C_2g_{m1}+g_{m1}g_{m3}} \quad (3)$$

where g_{m1}, g_{m2}, g_{m3} are the trans-conductance's of the different OTAs. For any fractional-order filters with a transfer function T(s), the following important parameters should be determined [2]:

a) ω_m : defined as the frequency at which the magnitude response either has maxima or a minima and can be evaluated by solving the equation i.e., $\left(d|T(j\omega)|/d\omega\right)_{\omega=\omega_m}=0$.

b) ω_h : defined as the frequency at which the power drops to half the pass-band power, known as half power frequency (also known as cutoff frequency) and can be evaluated by solving the equation i.e., $|T(j\omega_h)|=\left(1/\sqrt{2}\right)T\left(j\omega_{passband}\right)$. The bandwidth of any filter can also be calculated using this half-power frequency.

A. Fractional Order Low-Pass Filter (FLPF)

If we select $V_{in}=V_3$ and $V_1=V_2=0$, the general transfer function given in equation (3) represents a fractional order low-pass filter (FLPF) transfer function described as:

$$\frac{V_0(s)}{V_{in}(s)}=\frac{g_{m1}g_{m3}}{s^{2\alpha}C_1C_2+s^{\alpha}C_2g_{m1}+g_{m1}g_{m3}} \quad (4)$$

Using the definition of cutoff frequency given above, value of cutoff frequency for FLPF for any order 'α' can be obtained from equation (4a):

$$Y^4+2\cos\left(\frac{\alpha\pi}{2}\right)Y^3+(1+2k\cos(\alpha\pi)Y^2+2k\cos\left(\frac{\alpha\pi}{2}\right)Y-k^2=0 \quad (4a)$$

where $Y=\frac{\omega_h^\alpha}{a}$, $k=\frac{b}{a}$, $a=\frac{g_{m1}}{c_1}$ and $b=\frac{g_{m3}}{c_2}$.

B. Fractional order High-Pass Filter (FHPF)

By selecting $V_{in}=V_1$ and $V_2=V_3=0$, the general transfer function represented by equation (3) can be converted into a fractional order high-pass filter (FHPF) transfer function described as:

$$\frac{V_0(s)}{V_{in}(s)}=\frac{s^{2\alpha}C_1C_2}{s^{2\alpha}C_1C_2+s^{\alpha}C_2g_{m1}+g_{m1}g_{m3}} \quad (5)$$

The value of cutoff frequency for FHPF for any order 'α' can be obtained from equation (5a):

$$\omega_{mFLPF}\cdot\omega_{mFHPF}=\omega_{hFLPF}\cdot\omega_{hFHPF}=(ab)^{\frac{1}{\alpha}} \quad (5a)$$

C. Fractional order Band-Pass Filter (FBPF)

If we select $V_{in}=V_2$ and $V_1=V_3=0$ then the general transfer function given in equation (3) gives the fractional order band-pass filter (FBPF) transfer function described as:

$$\frac{V_0(s)}{V_{in}(s)}=\frac{-s^{\alpha}C_2g_{m2}}{s^{2\alpha}C_1C_2+s^{\alpha}C_2g_{m1}+g_{m1}g_{m3}} \quad (6)$$

The value of maximum frequency for FBPF with order 'α' can be obtained from equation (6a):

$$\left(X^2-k\right)\left(X^2+\cos\left(\frac{\alpha\pi}{2}\right)X+k\right) \quad (6a)$$

where $X=\frac{\omega_m^\alpha}{a}$.

The value of half power frequency for FBPF for any order α can be obtained from equation (6b):

$$Y^4+2\cos\left(\frac{\alpha\pi}{2}\right)Y^3+\begin{pmatrix}-1+2k\cos(\alpha\pi)-8\cos^2\left(\frac{\alpha\pi}{2}\right)k\\-8\cos\left(\frac{\alpha\pi}{2}\right)k^{\frac{1}{2}}\end{pmatrix}Y^2+2k\cos\left(\frac{\alpha\pi}{2}\right)Y+k^2=0 \quad (6b)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

D. Fractional order Band-Reject Filter (FBRF)

When $V_{in} = V_1 = V_3$ and $V_2 = 0$, the general transfer function given in equation (3) represents a fractional order band-reject filter (FBRF) transfer function described as:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{s^{2\alpha}C_1C_2 + g_{m1}g_{m3}}{s^{2\alpha}C_1C_2 + s^{\alpha}C_2g_{m1} + g_{m1}g_{m3}} \tag{7}$$

The value of minimum frequency for FBRF for any order α can be obtained from equation (7a):

$$\left(X^2 - k\right)\left(\begin{array}{l} X^4\cos\left(\frac{\alpha\pi}{2}\right) + X^3 + \left[\begin{array}{l} 4k\cos\left(\frac{\alpha\pi}{2}\right) \\ -2k\cos(\alpha\pi)\cdot\cos\left(\frac{\alpha\pi}{2}\right) \end{array}\right]X^2 \\ +kX + k^2\cos\left(\frac{\alpha\pi}{2}\right) \end{array}\right) \tag{7a}$$

The value of half power frequency for FBRF for any order α can be obtained from equation (7b):

$$Y^4 - 2\cos\left(\frac{\alpha\pi}{2}\right)Y^3 - \left(1 - 2k\cos(\alpha\pi)\right)Y^2 - 2k\cos\left(\frac{\alpha\pi}{2}\right)Y + k^2 \tag{7b}$$

E. Fractional order All-Pass Filter (FAPF):

If we select $V_{in} = V_1 = V_2 = V_3$, the general transfer function given in equation (3) represents a fractional order all-pass filter (FAPF) transfer function described as:

$$\frac{V_0(s)}{V_{in}(s)} = \frac{s^{2\alpha}C_1C_2 - s^{\alpha}C_2g_{m2} + g_{m1}g_{m3}}{s^{2\alpha}C_1C_2 + s^{\alpha}C_2g_{m1} + g_{m1}g_{m3}} \tag{8}$$

The important frequency parameters of different fractional order filters, namely, maximum or minimum frequency (ω_m) and half-power frequency (ω_h) for the special case when α = 1 have been shown in Table I. For non-integer values of α, their values may be computed by solving the system of non-linear equations given in equation (4)-(7) for different filters by approximately mapping the values of g_{m1}, g_{m2} and g_{m3} into 'a', 'b' and k as defined therein.

III. Stability Analysis

Detailed analysis of stability for fractional order system has been presented in [22]. For a fractional order system, stability graph is plotted in W-plane as shown in Fig. 3. In our work, the stability depends on the coefficients of $s^{i\alpha}$ $(0 < \alpha < 1)$ as given in equation (9). By considering the coefficients of equation (9) positive, different cases for stability are shown in Table II and has been adopted from [2]. It may be also noted from Table II, that fractional order filters are stable if $\delta > \left|\frac{\alpha\pi}{2}\right|$ where $0 < \alpha < 1$. Stability plot of FLPF, FHPF, FBPF and FBRF is shown in Fig. 4 - 7 for α = 0.7 respectively, using MATLAB command forlocus.

$$D(s) = s^{2\alpha}C_1C_2 + s^{\alpha}C_2g_{m1} + g_{m1}g_{m3} \tag{9}$$

TABLE I. Parameters when α =1

Types of filters	ω_m	ω_h
FLPF	--	$\left(a\sqrt{\left(k \pm \frac{\sqrt{8k^2 - 4k + 1}}{2} - \frac{1}{2}\right)}\right)$
FHPF	--	$\frac{ab}{\left(a\sqrt{\left(k \pm \frac{\sqrt{8k^2 - 4k + 1}}{2} - \frac{1}{2}\right)}\right)}$
FBPF	$a\sqrt{k}$	$a\left(\frac{\sqrt{4k+1}}{2} \pm \frac{1}{2}\right)$
FBRF	$a\sqrt{k}$	$a\left(\frac{\sqrt{4k+1}}{2} \pm \frac{1}{2}\right)$

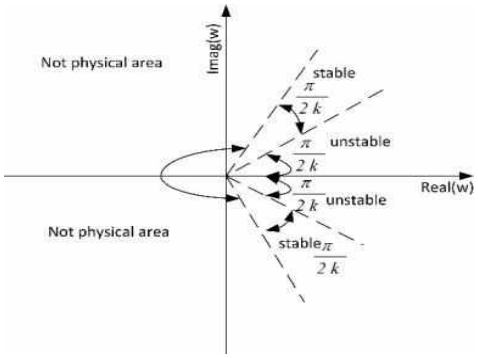


Figure 3. W-plane.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

TABLE II. RELATIONS, STABILITY CONDITION, ROOTS, POLE FREQUENCY ω_0 AND POLE QUALITY FACTOR Q FOR DIFFERENT CASES

Cases	Relations	Condition for stability and roots	ω_0, Q
1	$\frac{g_{m1}}{C_1} \geq 4\frac{g_{m3}}{C_2}$	$\alpha < 2$, $\eta_{1,2} = \frac{-\frac{g_{m1}}{C_1} \pm \sqrt{\left(\frac{g_{m1}}{C_1}\right)^2 - 4\left(\frac{g_{m1}g_{m3}}{C_1C_2}\right)}}{2} = g_{1,2}e^{j\pi}$	$\omega_{0_{1,2}} = g_{1,2}^{1/\alpha}$, $Q = \frac{-1}{2\cos(\pi/\alpha)}$
2	$\frac{g_{m1}}{C_1} < 4\frac{g_{m3}}{C_2}$	$\alpha < \frac{2\delta}{\pi}$, $\delta = \cos^{-1}\left(\frac{-\frac{g_{m1}}{C_1}}{\sqrt{\frac{g_{m1}g_{m3}}{C_1C_2}}}\right) > \frac{\pi}{2}$, $\eta_{1,2} = \sqrt{\frac{g_{m1}g_{m3}}{C_1C_2}}e^{\pm j\delta}$	$\omega_0 = \left(\sqrt{\frac{g_{m1}g_{m3}}{C_1C_2}}\right)^{1/\alpha}$, $Q = \frac{-1}{2\cos(\delta/\alpha)}$

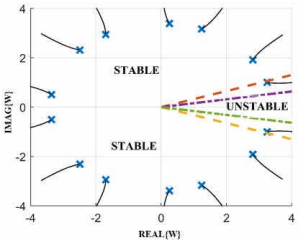


Figure 4. Stability plot of FLPF C=0.382μF and α=0.7.

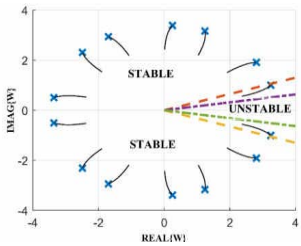


Figure 5. Stability plot of FHPF C=0.382μF and α=0.7.

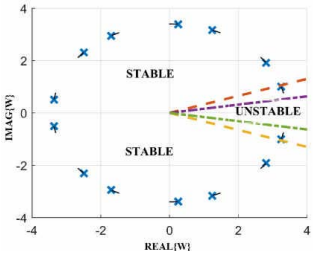


Figure 6. Stability plot of FBPF C=0.382μF and α=0.7.

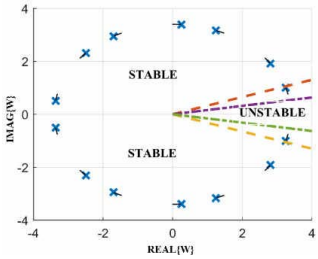


Figure 7. Stability plot of FBRF C=0.382μF and α=0.7.

IV. SIMULATION RESULTS

We have verified the workability of the proposed fractional order filters through PSPICE simulation using the macro model of OTA IC LM 13700. Equal valued fractional order capacitors ($C_1 = C_2$ =0.382uF/(rad/sec)$^{(1-\alpha)}$) are used in PSPICE simulations. These fractional order capacitors for different values of α (0.7, 0.8. 0.9) were designed using Valsa approximation method [21] of order 6, resulting in Foster type network shown in Fig. 8.

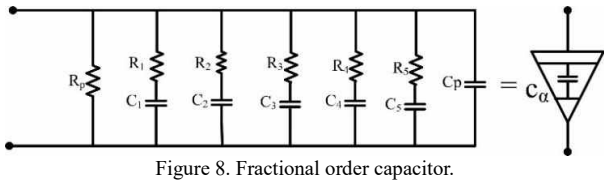


Figure 8. Fractional order capacitor.

All the filters presented in this paper were designed with values of $g_{m1} = g_{m2} = g_{m3}$ =5.48mA/V and R_{bias} =100kΩ.

The different frequency parameters for all the filters for different values of α = 0.7 to α = 1 have been listed in Table no. III and IV. From these tables it may be noted that the error between the theoretical values and the values obtained from PSPICE simulations is very small in case of all the filters (3.9%).

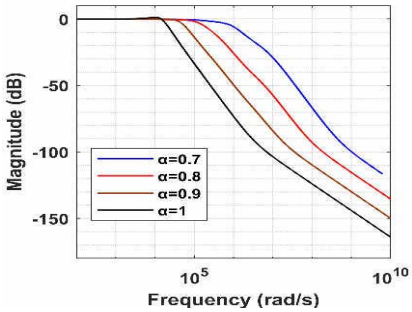


Figure 9. Magnitude response of FLPF for different α.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

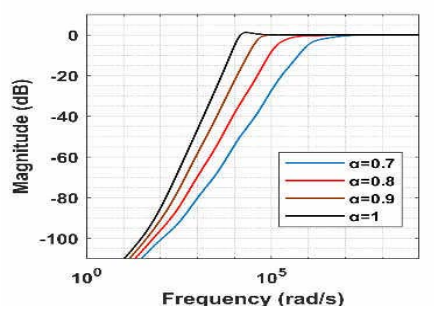


Figure 10. Magnitude response of FHPF for different α

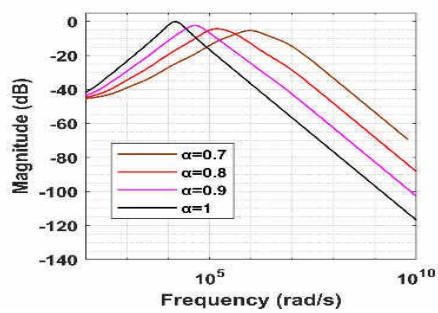


Figure 11. Magnitude response of FBPF for different α

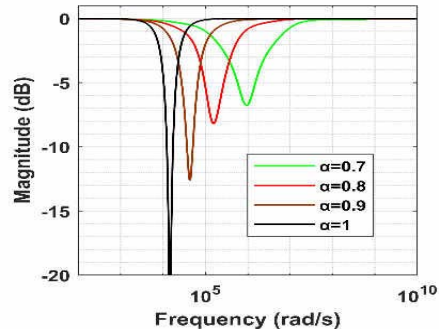


Figure 12. Magnitude response of FBRF for different α

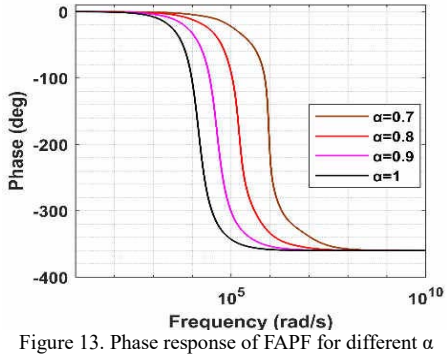


Figure 13. Phase response of FAPF for different α

TABLE III. CUTOFF FREQUENCY (ω_h) (IN KRAD/S)

α	FLPF		FHPF	
	PSPICE (ω_h)	Theoretical (ω_h)	PSPICE (ω_h)	Theoretical (ω_h)
0.7	486.63	487.81	1480.14	1500.46
0.8	129.67	130.13	202.35	202.87
0.9	46.12	46.87	38.63	38.33
1.0	14.9	14.34	14.1	14.34

TABLE IV. MAXIMUM/MINIMUM FREQUENCY (ω_m) (IN KRAD/S)

α	FBPF		FBRF	
	PSPICE (ω_m)	Theoretical (ω_m)	PSPICE (ω_m)	Theoretical (ω_m)
0.7	864.71	866.81	864.56	866.81
0.8	154.87	156.12	156.37	156.12
0.9	41.43	42.87	42.43	42.87
1.0	14.07	14.34	14.17	14.34

V. CONCLUSION

In this paper, two equal valued fractional order capacitors based filter using OTAs with electronic tunability of important frequency parameters has been presented and validated using PSPICE simulation results. When fractional order filter circuits gains momentum, more broad and difficult situations can be taken up by choosing different capacitors of different orders. Besides that, the advantages are:

(1) Considering the available filter design, they imposes the constraint on filter responses which can be easily removed in fractional order filters.

(2) Stability range and other important designing parameters are the function of α, which provides an extra degree of freedom to us.

REFERENCES

[1] Radwan, A. M. Soliman, and A. S. Elwakil, "First-order filters generalized to the fractional domain," Journal of Circuits, Systems, and Computers, 17, no. 01, pp. 55-66, 2008.

[2] A. G. Radwan, A. S. Elwakil, and A. M. Soliman, "On the generalization of second-order filters to the fractional-order domain," Journal of Circuits, Systems, and Computers, 18, no. 02, pp. 361-386, 2009.

[3] A. S. Ali, A. G. Radwan, and A. M. Soliman, "Fractional order Butterworth filter: active and passive realizations," IEEE journal on emerging and selected topics in circuits and systems, 3, no. 3, pp. 346-354, 2013.

[4] B. Maundy, A. S. Elwakil, and T. J. Freeborn, "On the practical realization of higher-order filters with fractional stepping," Signal Processing, 91, no. 3, pp. 484-491, 2011.

[5] F. Khateb, D. Kubánek, G. Tsirimokou, and C. Psychalinos, "Fractional-order filters based on low-voltage DDCCs," Microelectronics Journal, 50, pp. 50-59, 2016.

[6] T. Suksang, V. Pirajnanchai, C. Suppitaksakul, and W. Loedhammacakra, "Design and improve the performance of OTA low pass filter with Fractional-order step," In 2012 IEEE International Conference on Electron Devices and Solid State Circuit (EDSSC), 2012, pp. 1-4.

[7] G. Tsirimokou, C. Psychalinos, and A. S. Elwakil, "Fractional-order electronically controlled generalized filters," International Journal of Circuit Theory and Applications, 45, no. 5, pp. 595-612, 2017.

[8] M. R. Dar, N. A. Kant, F. A. Khanday, and C. Psychalinos, "Fractional-order filter design for ultra-low frequency applications," In 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016, pp. 1727-1730.

[9] R. Verma, N. Pandey, and R. Pandey, "Electronically tunable fractional order all pass filter," In IOP Conference Series: Materials

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

Science and Engineering, vol. 225, no. 1, p. 012229, IOP Publishing, 2017.

[10] D. V. Kamath, S. Navya, and N. Soubhagyaseetha, "Fractional order OTA-C current-mode all-pass filter," In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 383-387.

[11] M. V. Bhat, S. B. SS, and D. V. Kamath, "G m-C Current Mode Fractional All Pass Filter of order α (0< α< 1)," In 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 240-245.

[12] P. Bertsias, C. Psychalinos, A. S. Elwakil, and B. J. Maundy, "Simple Multi-Function Fractional-Order Filter Designs," In 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), 2019, pp. 1-4.

[13] I. E. Sacu, and M. Alci, "Low-power OTA-C Based Tuneable Fractional Order Filters," Electronic Components and Materials, 48, no. 3, pp. 135-144, 2018.

[14] R. Verma, N. Pandey, and R. Pandey, "Electronically tunable fractional order filter," Arabian Journal for Science and Engineering, 42, no. 8, pp. 3409-3422, 2017.

[15] E. Kaskouta, T. Kamilaris, R. Sotner, J. Jerabek, and C. Psychalinos, "Single-input multiple-output and multiple-input single-output fractional-order filter designs," In 2018 41st International Conference on Telecommunications and Signal Processing (TSP), 2018, pp. 1-4.

[16] R. Arya, G. Souliotis, S. Vlassis, and C. Psychalinos, "A 0.5 V tunable complex filter for Bluetooth and Zigbee using OTAs," Analog Integrated Circuits and Signal Processing, 79, no. 1, pp. 73-81, 2014.

[17] J. Jerabek, R. Sotner, D. Kubánek, J. Dvorak, L. Langhammer, N. Herencsar, and K. Vrba, "Fractional-order low-pass filter with electronically adjustable parameters," In 2016 39th International Conference on Telecommunications and Signal Processing (TSP), 2016, pp. 569-574.

[18] J. Dvorak, J. Jerabek, Z. Polesakova, D. Kubanek, and P. Blazek, "Multifunctional Electronically Reconfigurable and Tunable Fractional-Order Filter," Elektronika ir Elektrotechnika, 25, no. 1, 2019, pp. 26-30.

[19] J. Dvorak, Z. Polesakova, J. Jerabek, L. Langhammer, A. Kartci, and J. Koton, "Non-Integer-Order Low-Pass Filter with Electronically Controllable Parameters," In 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018, pp. 1-5.

[20] A. Oustaloup, F. Levron, B. Mathieu, and F. M. Nanot, "Frequency-band complex non-integer differentiator: characterization and synthesis," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 47, no. 1, pp. 25-39, 2000.

[21] J. Valsa, and J. Vlach, "RC models of a constant phase element," International Journal of Circuit Theory and Applications, 41, no. 1, pp. 59-67, 2013.

[22] A. G. Radwan, A. M. Soliman, A. S. Elwakil, and A. Sedeek, "On the stability of linear systems with fractional-order elements," Chaos, Solitons & Fractals, 40, no. 5, pp. 2317-2328, 2009.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 04:59:55 UTC from IEEE Xplore. Restrictions apply.

2020 6th International Conference on Control, Automation and Robotics

Fractional Order Oscillator Using OTAs

Vikramjeet Singh
Department of Electrical Engineering
Delhi Technological University
Delhi, India
e-mail: vikramjeet_mt2k18@dtu.ac.in

Pragati Kumar
Department of Electrical Engineering
Delhi Technological University
Delhi, India
e-mail: pragatikumar@dtu.ac.in

***Abstract*—A fractional order oscillator employing operational transconductance amplifiers (OTAs) and fractional capacitors has been presented in this paper. The Oustaloup approximation method is used to design the fractional order capacitors. Generalized fractional order theorems have been used to obtain the frequency of oscillation (FO) and condition of oscillation (CO). The stability of the proposed fractional order oscillator has been tested using pole-zero plots in the W-plane. PSPICE simulation and hardware result for the proposed oscillator have also been presented.**

Keywords-fractional calculus; fractional order capacitor; fractional order oscillator; Oustaloup approximation; operational transconductance amplifier

I. INTRODUCTION

Analog circuit realization of fractional order signal processing and signal generation circuits have started receiving renewed interest during the last decade [1]-[38]. Several fractional order filters and fractional order oscillators have been introduced in the open-literature wherein different types of amplifiers and fractional order capacitors have been used to realize these circuits. As fractional order capacitors are yet not available as standard circuit elements, these fractional order capacitors, have been realized using various approximation techniques which give a rational function approximation of the operator $s^{\alpha}(0<\alpha\leq 1)$ resulting in a Foster-like network emulating the behavior of a fractional order capacitor. Compared to fractional order filters, relatively less research work has been carried out on the realization of fractional order oscillators [3]-[4], [6], [13], [16-18], [21]-[33], [36]-[38]. In a fractional order oscillator, unlike a conventional oscillator, the FO, CO and the phase relationship between different voltages can be controlled with the help of the fractional order parameter 'α'. A control over the phase relationship is a very useful feature in applications like PSK modulation-demodulation, music synthesizers etc. Fractional order oscillators have been realized with various active building blocks like operational amplifiers [3]-[4], [6], [13], [16], [24]-[25], [32], [38], operational transresistance amplifiers [17], [32] current conveyors and their different variants [18], [21-23], [33], [29] current feedback amplifiers [28], voltage differencing inverting buffered amplifier [36], operational transconductance amplifiers [24] and current mirrors [30]. In most of these realizations the conventional capacitor(s) in an integer order realization of an existing oscillator has(ve) been replaced by fractional order capacitor(s). From the study of various fractional order oscillator circuits, it has been observed that very little work has been reported on the realization of fractional order oscillator circuits using operational transconductance amplifiers (OTAs). In such oscillator using control over the CO and FO of the fractional order oscillator can very easily be implemented. Therefore, the main aim of this paper is to present a new fractional order oscillator circuit using operational transconductance amplifier and fractional order capacitors in which the CO and FO for a specified value of the given fractional order parameter 'α' can be controlled by changing the bias currents of different operational transconductance amplifiers.

II. PROPOSED FRACTIONAL ORDER OSCILLATOR CIRCUIT

The proposed circuit uses five OTAs and two fractional order capacitors. An OTA is a differential voltage controlled current source whose circuit symbol is shown below in Fig.1 while its terminal relationships are given in equation (1).

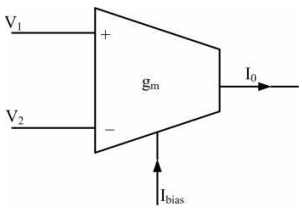


Figure 1. OTA circuit.

$$I_0 = g_m\left(V_1 - V_2\right),\ g_m = \frac{I_{Bias}}{2V_T} \tag{1}$$

A fractional order capacitor, on the other hand, whose circuit symbol is shown below in Fig. 2 is characterized by the driving point impedance.

$$Z(s) = \frac{1}{s^{\alpha}C_{\alpha}} \tag{2}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 05:00:21 UTC from IEEE Xplore. Restrictions apply.

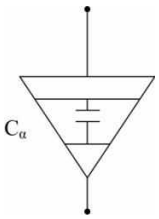


Figure 2. Fractional order capacitor.

The proposed fractional order oscillator is shown below in Fig. 3. The state equations of the oscillator circuit can be described by equation (3):

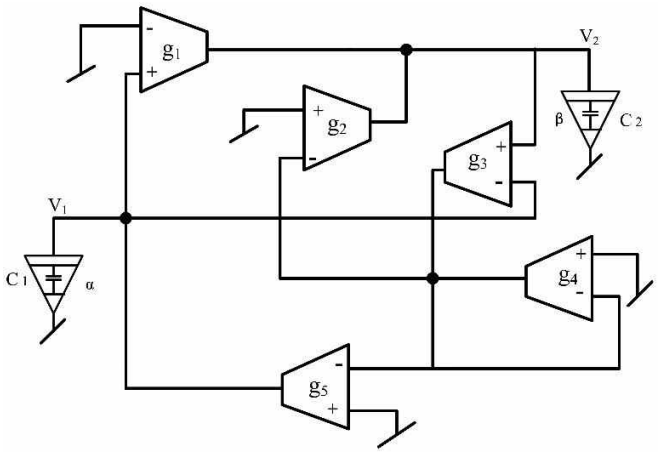


Figure 3.The proposed circuit of oscillator with two fractional order capacitors.

$$\begin{bmatrix} D^{\beta}V_2 \\ D^{\alpha}V_1 \end{bmatrix} = \begin{bmatrix} \dfrac{-g_2g_3}{g_4c_2} & \dfrac{g_1g_4+g_2g_3}{g_4c_2} \\ \dfrac{-g_5g_3}{g_4c_1} & \dfrac{g_5g_3}{g_4c_1} \end{bmatrix} \begin{bmatrix} V_2 \\ V_1 \end{bmatrix} \tag{3}$$

where D^{α} and D^{β} represent the fractional order derivative operator. By using the basic definition of the Laplace transform applicable to the fractional order derivative operator [4] the characteristic equation (CE) of the autonomous circuit represented by the fractional order state equations given above is as follows:

$$s^{\alpha+\beta} + s^{\alpha}\frac{g_2g_3}{g_4c_2} - s^{\beta}\frac{g_3g_5}{g_4c_1} + \frac{g_1g_5g_3}{g_4c_1c_2} = 0 \tag{4}$$

Putting $s = j\omega$ in equation (4) and separating real and imaginary part with help of Euler's relation we get the following two relationships.

$$\omega^{\alpha+\beta}\cos\left(\frac{(\alpha+\beta)\pi}{2}\right) + \omega^{\alpha}\frac{g_2g_3}{g_4c_2}\cos\left(\frac{\alpha\pi}{2}\right) - \omega^{\beta}\frac{g_3g_5}{g_4c_1}\cos\left(\frac{\beta\pi}{2}\right) + \frac{g_1g_5g_3}{g_4c_1c_2} = 0 \tag{5}$$

$$\omega^{\alpha+\beta}\sin\left(\frac{(\alpha+\beta)\pi}{2}\right) + \omega^{\alpha}\frac{g_2g_3}{g_4c_2}\sin\left(\frac{\alpha\pi}{2}\right) - \omega^{\beta}\frac{g_3g_5}{g_4c_1}\sin\left(\frac{\beta\pi}{2}\right) = 0 \tag{6}$$

The CO and FO of the fractional order oscillator must satisfy both (5) and (6) and are expressed as follows:

$$g_2 = \frac{-\omega^{\alpha+\beta}\sin\left(\frac{(\alpha+\beta)\pi}{2}\right) + \omega^{\beta}\frac{g_5g_3}{g_4c_1}\sin\left(\frac{\beta\pi}{2}\right)}{\omega^{\alpha}\frac{g_3}{g_4c_2}\sin\left(\frac{\alpha\pi}{2}\right)} \tag{7}$$

$$\omega^{\alpha+\beta}\sin\left(\frac{\beta\pi}{2}\right) + \omega^{\beta}\frac{g_3g_5}{g_4c_1}\sin\left(\frac{(\alpha-\beta)\pi}{2}\right) - \frac{g_1g_5g_3}{g_4c_1c_2}\sin\left(\frac{\alpha\pi}{2}\right) = 0 \tag{8}$$

From the above two equations (7) and (8) it may be observed that the transconductance g_2 may be selected in such a way that the circuit oscillates for a specified value of ω rad/sec for predefined values of α and β. We have shown below in Table I the CO, FO and the phase relationships between the two capacitor voltages for different values of α and β.

TABLE I. DIFFERENT CASES FOR GENERALIZED RELATIONS OF CO, FO AND PHASE

Oscillator parameter			
Cases	***CO(g_2)***	***FO(ω)***	***Phase(ϕ)***
1. α=β=1	$g_2=g_5$	$\omega = \sqrt{\frac{g_1g_3g_5}{g_4c_1c_2}}$	$\frac{\pi}{2}$
2. $\alpha = \beta \neq 1$	$g_2 = \frac{-\omega^{2\alpha}\sin(\alpha\pi) + \sin\left(\frac{\alpha\pi}{2}\right)\frac{g_5g_3\omega^{\alpha}}{g_4c_1}}{\sin\left(\frac{\alpha\pi}{2}\right)\frac{g_3\omega^{\alpha}}{g_4c_2}}$	$\omega = \left(\frac{g_1g_5g_3}{g_4c_1c_2}\right)^{\frac{1}{2\alpha}}$	$\frac{\alpha\pi}{2}$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 05:00:21 UTC from IEEE Xplore. Restrictions apply.

3. $\alpha = 1$ $0 < \beta < 1$	$g_2 = \dfrac{-\omega^{1+\beta}\cos\left(\frac{\beta\pi}{2}\right)+\sin\left(\frac{\beta\pi}{2}\right)\frac{g_5 g_3 \omega^{\beta}}{g_4 c_1}}{\frac{g_3\omega}{g_4 c_2}}$	$\omega^{1+\beta}\sin\left(\frac{\beta\pi}{2}\right)+\omega^{\beta}\frac{g_3 g_5}{g_4 c_1}\cos\left(\frac{\beta\pi}{2}\right)$ $-\frac{g_1 g_5 g_3}{g_4 c_1 c_2}=0$	$\frac{\pi}{2}$
4. $\beta = 1$ $0 < \alpha < 1$	$g_2 = \dfrac{-\omega^{\alpha+1}\cos\left(\frac{\alpha\pi}{2}\right)+\frac{g_5 g_3 \omega}{g_4 c_1}}{\sin\left(\frac{\alpha\pi}{2}\right)\frac{g_3\omega^{\alpha}}{g_4 c_2}}$	$\omega^{\alpha+1}+\omega\frac{g_3 g_5}{g_4 c_1}\cos\left(\frac{\alpha\pi}{2}\right)-\frac{g_1 g_5 g_3}{g_4 c_1 c_2}\sin\left(\frac{\alpha\pi}{2}\right)=0$	$\frac{\alpha\pi}{2}$

The electronic tunability of CO and FO for different values of 'α'by changing 'g_3' have been illustrated in Fig.4 -Fig.6.

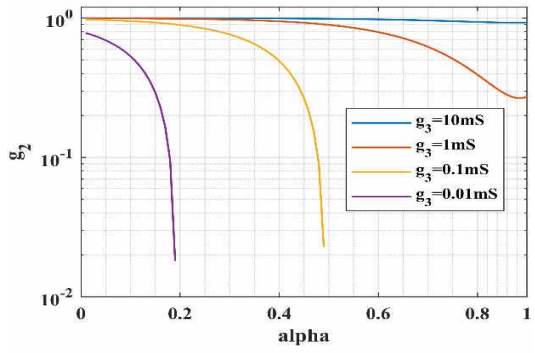


Figure 4. Tunability of CO with g_3 with α for case 2.

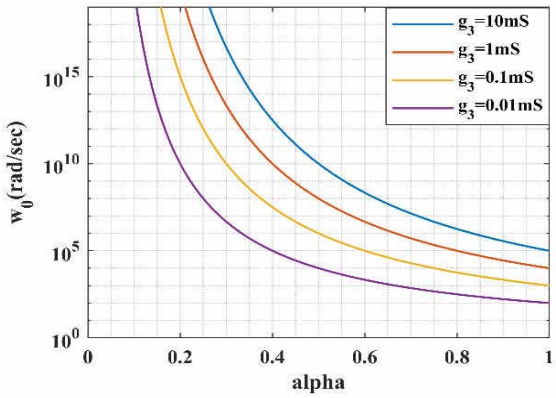


Figure 5. Tunability of FO with g_3 for case 2.

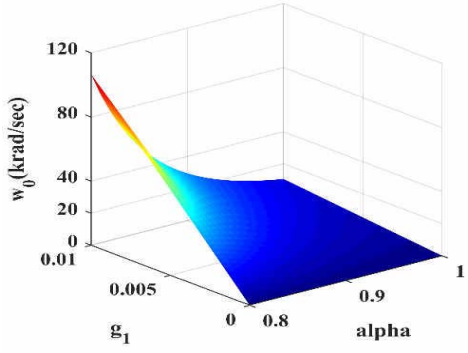


Figure 6. Tunability of FO with g_1 and α for case 2.

III. Stability Analysis

Stability of fractional order systems has been studied in detail [7], [20]. As the stability of any linear dynamic system depends on the location of the roots of the CE in the s-plane. It is necessary to know about the location of the poles in the s-plane. For analysis of the stability of fractional order systems the CE of the system is transformed into W-plane using the transformation $W = s^{\frac{1}{k}}$ where k is a positive integer related to the non-integer order 'α' by the relation α = n/k (n being another positive integer). The stable and unstable regions for system in W-PLANE are defined as $\pi/2k < |\phi_w| < \pi/k$ and $|\varphi_w| < \pi/2k$ respectively where φ_w represents the angle of the roots in the W-plane measured from the +ve real axis of the W-plane. Stability of fractional order oscillators have been studied in the W-plane for different values of 'α' by finding the location of the roots of the CE in W-plane. At least one pole must lie on the line $|\phi_w| = \pi/2k$ for system to be oscillatory and remaining poles should lie in the stable region as shown in Fig. 7. If even one pole lies in the region $|\phi_w| < \pi/2k$ the system become unstable.

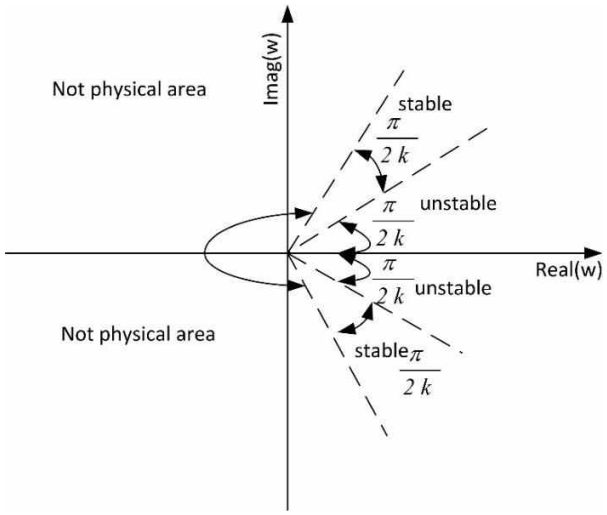
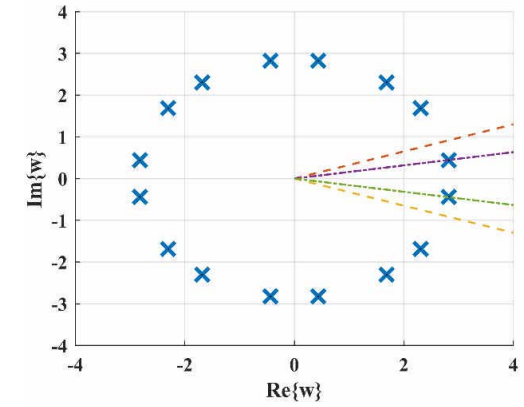


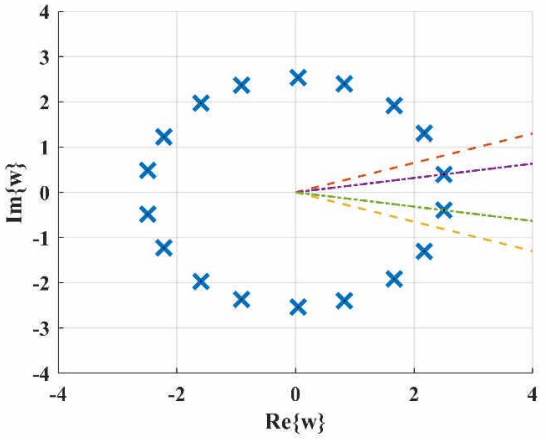
Figure 7. W-plane.

We have selected the value of k = 10 and plotted the location of the roots of the CE of the fractional order oscillator in the W-plane for different values of 'α'. And shown them in Fig. 8(a-c). It is noted from Fig. 8 that the fractional order oscillators presented in this paper are indeed stable!

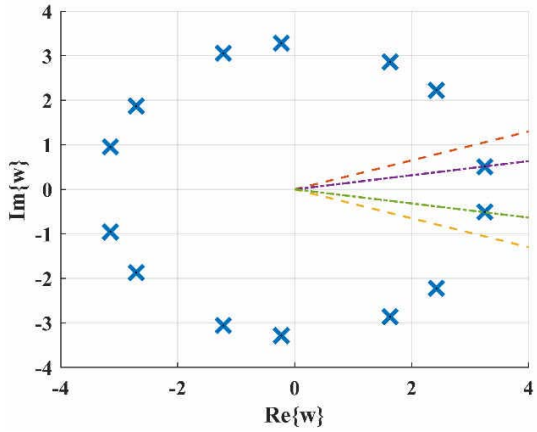
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 05:00:21 UTC from IEEE Xplore. Restrictions apply.



(a)



(b)



(c)

Figure 8. Poles in W-plane of oscillator for (a) α=β=0.8, (b) α=β=0.9 (c) α=β=0.7.

IV. Simulation Results

Workability of the fractional order oscillator presented in this paper has been verified by PSPICE simulations and experimental results. We have used CMOS implementation of an OTA [10] to test the design of the fractional order oscillator. Two identical valued fractional order capacitors of value $0.95\mu F(rad/sec)^{(1-\alpha)}$ designed using the Oustaloup approximation method were used in simulation studies (α = β= 0.9, 0.8 0.7). We have simulated the fractional order oscillator represented by case 2 from Table I (equal fractional-order capacitors) with (i) $\alpha=\beta$ =0.7 (ii) $\alpha=\beta$ =0.8 (iii) $\alpha=\beta$ =0.9 The designed FO and the corresponding values of transconductances meeting the CO and FO are given below in Table II.

TABLE II. Values of Transconductance at Different $\alpha=\beta$ and Frequency(ω_0)

C=0.95uF α=β	Values of transconductance (mS)					Theoretical value of FO(ω_0)
	g_1	g_2	g_3	g_4	g_5	
0.7	3.98	0.26	3.98	3.98	3.98	23.7KHz
0.8	4.13	1.33	4.13	4.13	4.13	5.61KHz
0.9	4.13	2.80	4.13	4.13	4.13	1.75KHz

Fig. 9-Fig. 11 show the output voltage waveforms for different values of the fractional order parameters.

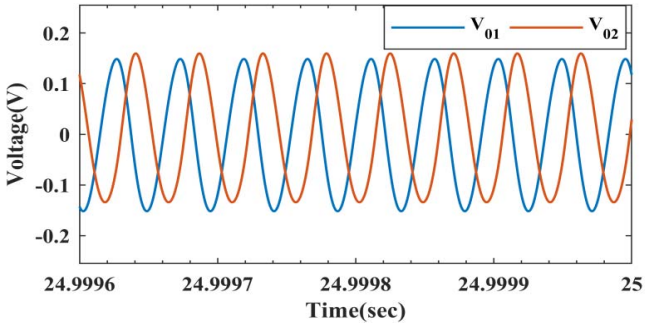


Figure 9. α=β=0.7 PSPICE Simulation of oscillator.

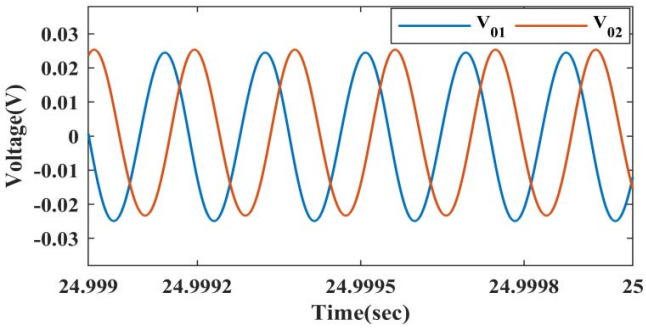


Figure 10. α=β =0.8 PSPICE Simulation of oscillator.

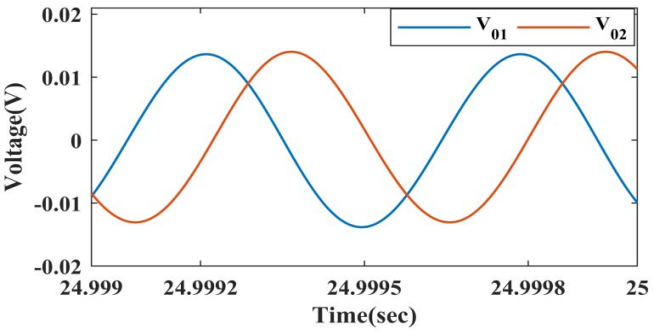


Figure 11. α=β =0.9 PSPICE Simulation of oscillator.

In Table III we have shown the values of FO measured in simulation along with the values of theoretical FO for different values of the fractional order parameters. From this table it is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 05:00:21 UTC from IEEE Xplore. Restrictions apply.

observed that the values obtained from simulation are very close to their theoretical counterparts.

TABLE III PSPICE AND THEORETICAL VALUES OF FO(ω_0) AT DIFFERENT α=β

C=0.95uFr(ad/sec) α=β	PSPICE value of FO(ω_0)	Theoretical value of FO(ω_0)
0.7	22KHz	23.7KHz
0.8	5.42KHz	5.61KHz
0.9	1.74KHz	1.75KHz

We have also realized a fractional order oscillator using LM13700 IC OTA biased with ± 15V. A fractional order oscillator with a nominal frequency of oscillation of 1.4MHz was designed by selecting two identical valued fractional order capacitors for α=β=0.8, with $C_1=C_2=10nF(rad/sec)^{(0.2)}$.

An eighth order approximation of the fractional order capacitor of value 10nF $(rad/sec)^{0.2}$ corresponding to 'α' = 0.8 using Oustaloup method [2] resulting into the circuit given below in Fig.12 was used to realize these capacitors. Values of different capacitors and resistors used in Fig.12 are given in Table IV.

TABLE IV. RC LADDER VALUES USING OUSTALOUP APPROXIMATION

R(Ω) / R_0=7Ω	10	56	330	1.7K	8.2K	39K	220K	3.3M
C(nF)	0.6	1	1.5	2	3	4	6.5	3.8

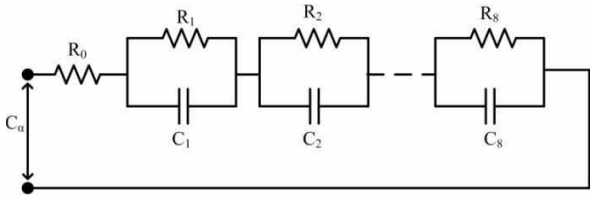


Figure 12. RC ladder circuit using Oustaloup approximation.

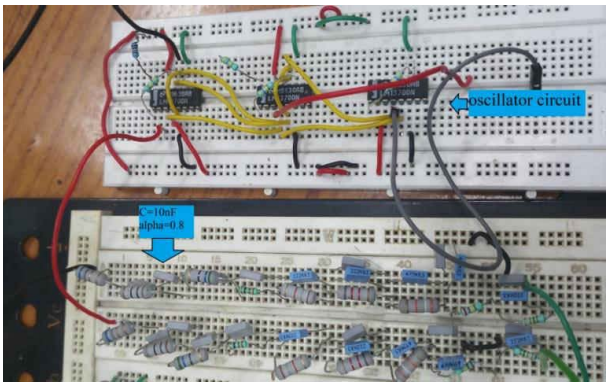


Figure 13. Circuit Implementation of fractional order oscillator.

The required values of different transconductances g_1=5.5mS, $g_3=g_4=g_5$=1.2mS, g_2=0.8mS were set by controlling the bias current of all the OTAs. The experimental set up of the oscillator circuit is shown below in Fig. 13. The output waveform of the oscillator is shown in Fig. 14. Fig. 15 shows the Lissajous pattern showing the phase relationship between the two output voltages while Fig. 16 shows the frequency spectrum of the voltage V_1. It may be observed from the experimental results indicated on the output waveform (obtained on Keysight DSOX3034T) that the values of experimentally obtained FO is equal to 1.456 MHz and the phase difference between the two output voltages is equal to 71.728º) that these results are in very close agreement with the corresponding designed value (1.4MHz and 72º).

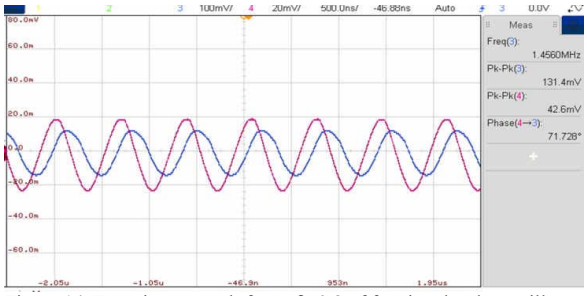


Figure 14. Experiment result for α=β=0.8 of fractional order oscillator.

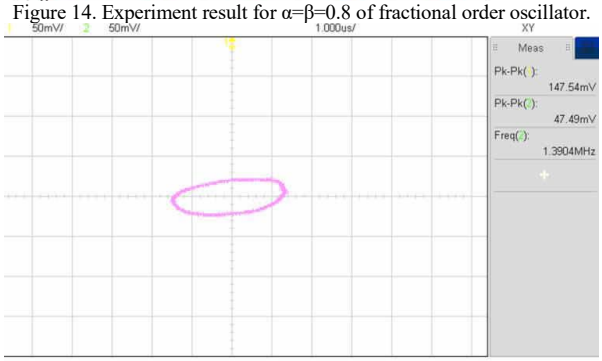


Figure 15. Lissajous pattern for α=β= 0.8 of fractional order oscillator.

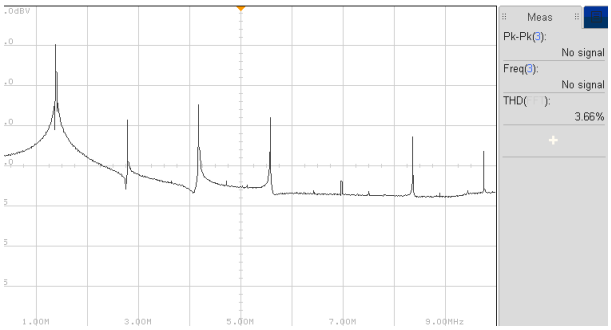


Figure 16. Frequency spectrum for α=β=0.8 of fractional order oscillator.

V. CONCLUSION

In this paper, an OTA based fractional order oscillator with electronic tunability of CO and FO has been presented. PSPICE simulation results and experimental results validating the operation of the presented circuit have been presented. The proposed oscillator circuit may be used in applications like music synthesizers, and PSK modulation and demodulation and has a better range of control of CO and FO because of presence of more number of transconductances.

REFERENCES

[1] A. Oustaloup, "Fractional order sinusoidal oscillators: optimization and their use in highly linear FM modulation," IEEE Transactions on Circuits and Systems 28.10, pp. 1007-1009,1981.

[2] A. Oustaloup, F. Levron, B. Mathieu and F. M. Nanot, "Frequency-band complex noninteger differentiator: characterization and synthesis," IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 47(1), pp.25-39,2000.

[3] W. Ahmad, R. El-Khazali, and A. S. Elwakil. "Fractional-order Wien-bridge oscillator," Electronics Letters 37.18, pp. 1110-1112,2001.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 05:00:21 UTC from IEEE Xplore. Restrictions apply.

[4] A. G. Radwan, A. M. Soliman, and A. S. Elwakil, "Design equations for fractional-order sinusoidal oscillators: four practical circuits examples," Int. J .Circ. Theory Appl., vol. 36, pp. 473-492, 2007.

[5] A. G. Radwan, A. M. Soliman, and A. S. Elwakil, "First-order filters generalized to the fractional domain," *Journal of Circuits, Systems, and Computers 17*, no. 01, pp.55-66, 2008.

[6] A. G. Radwan, A. S. Elwakil, & A. M. Soliman, "Fractional-order sinusoidal oscillators: design procedure and practical examples," IEEE Transactions on Circuits and Systems I: Regular Papers, 55(7), pp. 2051-2063, 2008.

[7] A. G Radwan, et al, "On the stability of linear systems with fractional-order elements," Chaos, Solitons & Fractals 40.5, pp. 2317-2328,2009.

[8] A. G. Radwan, A. S. Elwakil, and A. M. Soliman, "On the generalization of second-order filters to the fractional-order domain," *Journal of Circuits, Systems, and Computers* 18.02, pp. 361-386, 2009.

[9] T. J. Freeborn, B. Maundy, and A. S. Elwakil, "Field programmable analogue array implementation of fractional step filters," IET circuits, devices & systems 4.6, pp. 514-524, 2010.

[10] E. W. Greeneich, "*Analog integrated circuits*," Springer Science & Business Media, 2012.

[11] A. Soltan, A. G. Radwan, and A. M. Soliman, "Fractional order filter with two fractional elements of dependant orders," Microelectronics Journal 43.11, pp. 818-827, 2012.

[12] T. J. Freeborn, B. Maundy and A. Elwakil, "Fractional-step Tow-Thomas biquad filters," Nonlinear Theory and Its Applications, IEICE 3.3, pp.357-374, 2012.

[13] M. Brent, A. Elwakil, and S. Gift, "On the realization of multiphase oscillators using fractional-order all pass filters," Circuits, Systems, and Signal Processing 31.1, pp. 3-17,2012.

[14] M. C. Tripathy, K. Biswas, and S. Sen, "A design example of a fractional-order Kerwin–Huelsman–Newcomb biquad filter with two fractional capacitors of different order," Circuits, Systems, and Signal Processing 32.4, pp.1523-1536, 2013.

[15] A. Soltan, A. G. Radwan, and A. M. Soliman, "CCII based fractional filters of different orders," *Journal of advanced research 5.2*, pp.157-164, 2014.

[16] A. L Said, A. G. Radwan, A.H. Madian, and A. M. Soliman, "Fractional order two port network oscillator with equal order," In 2014 26th International Conference on Microelectronics (ICM), pp. 156-159, 2014.

[17] L. A. Said, A. G. Radwan A. H. Madian, and A. M. Soliman, "Fractional order Oscillators Based on Operational Transresistance Amplifiers," Int.J. Electron. Comm. (AEÜ), vol.69, pp. 988-1003, 2015.

[18] M. E. Fouda, A. Soltan, A. G. Radwan, A. M. Soliman, "Fractional-order multi-phase oscillators design and analysis suitable for higher-order PSK applications," Analog Integrated Circuits and Signal Processing, pp. 301-12, 2016.

[19] D. Kubanek, J. Koton, J. Jerabek, P. Ushakov and A. Shadrin, "Design and properties of fractional-order multifunction filter with DVCCs," 2016 39th International Conference on Telecommunications and Signal Processing (TSP), pp.620-624, 2016.

[20] M. S Semary, A. G. Radwan, and H. N. Hassan, "Fundamentals of fractional-order LTI circuits and systems: number of poles, stability, time and frequency responses," International Journal of Circuit Theory and Applications 44.12, pp. 2114-2133,2016.

[21] D. Kubánek, F. Khateb, G. Tsirimokou, and C. Psychalinos, "Practical design and evaluation of fractional-order oscillator using differential voltage current conveyors," Circuits, Systems, and Signal Processing 35, no. 6, pp. 2003-2016, 2016.

[22] L. A. Said, A. G. Radwan, A. H. Madian and A. M. Soliman, "Fractional order oscillator design based on two-port network," Circuits Syst. Signal Process. 35 pp. 3086–3112,2016.

[23] L. A. Said, A. G. Radwan, A. H. Madian, A. M. Soliman, "Fractional-order oscillator based on single CCII," in IEEE Int. Conf. Telecommun. Signal Process, pp. 603-606, 2016.

[24] M. R Dar, N. AliKant, F. A. Khanday, G. Tsirimokou, and C. Psychalinos,"Design of fractional-order multiphase sinusoidal oscillators," In 3rd International Conference on Electrical, Electronics, Engineering Trends, Communication, Optimization and Sciences (EEECOS 2016), pp. 1-5, 2016.

[25] L. A. Said, A. G. Radwan, A. H. Madian, and A. M. Soliman, "Three fractional-order-capacitors-based oscillators with controllable phase and frequency," Journal of Circuits, Systems and Computers 26, no. 10 (2017): 1750160, 2017.

[26] A. M. EL-Naggar, L. A. Said, A. G. Radwan, A. H. Madian, A. M. Soliman, "Fractional order four-phase oscillator based on double integrator topology," in IEEE Int. Conf. Modern Circ. and Syst. Technologies, pp. 1-4, 2017.

[27] A. Kartci, N. Herencsar, J. Koton & C. Psychalinos, "Compact MOSRC Voltage-Mode Fractional-Order Oscillator Design," in IEEE European Conf. Circ. Theor. Design, pp. 1-4, 2017.

[28] A. L. Said, A. G. Radwan, A. H. Madian, and A. M. Soliman, "Generalized family of fractional-order oscillators based on single CFOA and RC network," In 2017 6th International Conference on Modern Circuits and Systems Technologies (MOCAST), pp. 1-4, 2017.

[29] A. Pradhan, K. S. Subhadhra, N. Atique, R. K. Sharma, & S. S. Gupta, "MMCC-based current-mode fractional-order voltage-controlled oscillators," In 2018 2nd International Conference on Inventive Systems and Control (ICISC), pp. 763-768,2018.

[30] G. Tsirimokou, C. Psychalinos, A. Elwakil , & B. Maundy, "Fractional-Order Multiphase Sinusoidal Oscillator Design Using Current-Mirrors," In 2018 41st International Conference on Telecommunications and Signal Processing (TSP) ,pp. 1-4, 2018.

[31] A. Pradhan, and R. K. Sharma, "Generalised Fractional-Order Oscillators using OTA," In 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 402-406, 2018.

[32] O. Elwy, L. A. Said, A. H. Madian & A. G. Radwan, "Fractional-Order Relaxation Oscillators Based on Op-Amp and OTRA," In 2018 30th International Conference on Microelectronics (ICM), pp. 212-215 ,2018.

[33] S. K. Mishra, M. Gupta, & D. K. Upadhyay, "Design and implementation of DDCC-based fractional-order oscillator," International Journal of Electronics 106.4, pp.581-598, 2019.

[34] S. Mahata, R. Kar, & D. Mandal, "Optimal fractional-order highpass butterworth magnitude characteristics realization using current-mode filter," AEU-International Journal of Electronics and Communications, 102, pp.78-89, 2019.

[35] R. Verma, N. Pandey, & R. Pandey, " CFOA based low pass and high pass fractional step filter realizations," AEU-International Journal of Electronics and Communications, 99, pp. 161-176, 2019.

[36] A. Kartci, N. Herencsar, J. Dvorak, and K. Vrba, "VDIBA-Based Fractional-Order Oscillator Design," In 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), pp. 744-747, 2019.

[37] K. Bingi, R. Ibrahim, M. N. Karsiti, S. M. Hassan, I. Elamvazuthi, & A.M. Devan, "Design and Analysis of Fractional-order Oscillators Using Scilab," In 2019 IEEE Student Conference on Research and Development (SCOReD) ,pp. 311-316, 2019.

[38] G. M. Ahmed, L. A. Said, A. H. Madian, & A. G. Radwan, "Fractional-Order Oscillators Based on Double Op-Amp," In 2019 Fourth International Conference on Advances in Computational Tools for Engineering Applications (ACTEA), pp. 1-4, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on June 12,2020 at 05:00:21 UTC from IEEE Xplore. Restrictions apply.

Transportation Infrastructure Geotechnology
https://doi.org/10.1007/s40515-020-00113-y

TECHNICAL PAPER

Fuzzy-Based Model for Predicting Strength of Geogrid-Reinforced Subgrade Soil with Optimal Depth of Geogrid Reinforcement

Meenakshi Singh[1] · Ashutosh Trivedi[1] · Sanjay Kumar Shukla[2,3]

Accepted: 2 June 2020/Published online: 15 June 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

A fuzzy logic (FL)–based modeling approach is employed for geogrid-reinforced subgrade soil of unpaved roads. A review of the literature reveals that fuzzy logic has not been used for predicting the behavior of geogrid-reinforced subgrade. This paper presents FL-based two models with fuzzy Triangular and Gaussian membership functions for input and output variables. It consists of eight input parameters/factors, namely, reinforced/unreinforced section, depth of reinforcement, liquid limit, plastic limit, plasticity index, optimum moisture content, maximum dry unit weight, and soaked/unsoaked condition, and California bearing ratio (*CBR*) as an output parameter. The fuzzy rules are deduced from the experimental data. The laboratory *CBR* tests were performed on the subgrade soil reinforced with geogrid. The precision of models was examined by comparing the predicted *CBR* values with the experimental *CBR* values for Triangular and Gaussian membership functions. The sensitivity analysis reflects a set of dominant parameters. The results indicated a significant improvement in the *CBR* value of geogrid-reinforced subgrade soil due to the inclusion of geogrid. The range for optimal depth of geogrid reinforcement is found to be 36 to 60% of the thickness of the soil layer. The potentialities of FL were found to be satisfactory.

Keywords *CBR* value · Subgrade · Fuzzy logic · Geogrid reinforcement

✉ Meenakshi Singh
meenakshisingh_2k14@dtu.ac.in

Ashutosh Trivedi
atrivedi@dce.ac.in

Sanjay Kumar Shukla
sanjaykshukla1@gmail.com

Extended author information available on the last page of the article

1 Introduction

The use of geosynthetic reinforcement is one of the well-established techniques for subgrade soil improvement and road base reinforcement for over four decades. This technique has been developed extensively to improve the performance of both paved and unpaved roads. Geogrid is a major type of geosynthetic that is commonly used in unpaved roads for soil reinforcement to achieve the technical benefits and also for speeding the construction of roads over weak subgrade soils. This improvement is attributed to a set of mechanisms: improvement in load distribution through the base course, prevention of local shearing of the soil, reduction in shear stresses on the subgrade, and tensioned membrane effect (Giroud and Han 2004). In unpaved roads, the geogrid layers are mainly used as the reinforcement for granular layer and subgrade soil although it may play the role of a separator to some extent (Shukla 2016).

The California bearing ratio (*CBR*) of the subgrade soil is considered a key parameter in the design of flexible pavements of paved and unpaved roads. The subgrade plays a vital role in conveying the structural load to the pavement structure as it is subjected to the moving traffic load on roads. The load has to be transferred in such a way that the shear stresses and deformation developed in the subgrade soil are within the safe limits under adverse climatic and loading conditions. Under repeated/cyclic loading due to traffic, the deformation in the successive number of load cycles is controlled by a hardening parameter which is dependent on confining pressure and plasticity (Trivedi 2013). The effect of hardening parameter is such that it increases the elastic modulus in successive cycles and there is a diminishing settlement for the increasing number of cycles, which improves the strength of the soil. Hence, the subgrade soil may safely withstand the stress due to traffic loads. Analyzing and designing of unpaved roads with and without reinforcement layers is generally based on the bearing capacity of the subgrade (Perkins et al. 2012). The *CBR* test is one of the methods to determine the bearing capacity of the subgrade soil. When soft soil is used as a road subgrade, it undergoes excessive consolidation, settlement, and bearing capacity failure because of its low shear strength. Many experimental and numerical studies have been performed to investigate the reinforcing effect of geosynthetic on pavement structures (Negi and Singh 2019; Cuelho and Perkins 2017; Rashidian et al. 2016; Moghaddas-Nejad and Small 1996; Suku et al. 2017; Wu et al. 2015; Palmeira and Antunes 2010; Ibrahim et al. 2017). The results of such studies indicate that the incorporation of geogrid can improve the performance of paved/unpaved roads, increase the service life, and reduce the maintenance cost. Thus, it is necessary to understand the influence of geogrid on the engineering properties of subgrade soil and to identify the optimum location of geogrid reinforcement within a subgrade soil to get maximum benefit in terms of the increased strength of subgrade soil.

Soft computing techniques have been widely used in many areas of science and engineering applications for solving complex real-life problems. One of the soft computing techniques, namely, fuzzy logic, deals with the concept of partial truth theory to provide a methodology to model uncertainty and the human way of thinking, reasoning, and perception (Taghavifar and Mardani 2014). The neuro-fuzzy model has been applied in the application of geotechnical engineering and fuzzy logic in pavement engineering applications (Bagdatli 2018; Lee and Donnell 2007; Cabalar et al. 2012; Sandra and Sarkar 2015; Zehtabchi et al. 2018). Yet no major attempt has been

made to apply the fuzzy logic to model the effect of geogrid reinforcement on the strength behavior of subgrade soil. The literature review reveals that most research works focus on developing a prediction model to get the output. It has been investigated that the strength of subgrade soil is affected by the basic properties of the soil (Black 1962; Taskiran 2010; Venkatasubramanian and Dhinakaran 2011). Several researchers have shown the use of soft computing systems as an application to estimate the *CBR* value from the physical and compaction properties of the soil (Günaydin 2009; Gurtug and Sridharan 2002; Yildirim and Gunaydin 2011; Bhatt et al. 2014). Investigations were carried out on natal soils and the relationship between various parameters and *CBR* was predicted. Unfortunately, these models were found to be unsatisfactory (Stephens 1990). The prediction of *CBR* of subgrade soil by using genetic expression programming and artificial neural network was studied recently by Tenpe and Patel (Tenpe and Patel 2018). Still, no major effort has been made in the past to utilize the fuzzy logic for the estimation of *CBR* of reinforced subgrade soil with geogrid. *CBR* tests are routinely carried out to identify the *CBR* value of subgrade soil, which relates to the stiffness modulus and strength parameter of subgrade soil. The results of *CBR* tests are very important in geotechnical engineering, but many times engineers are facing difficulties in obtaining the desired *CBR* value for the design of pavements. Conducting this test either in the field or in the laboratory is laborious, tedious, and time consuming. Inaccurate results may be obtained due to the sample disturbance in the field and poor testing conditions in laboratory. Hence, it is advantageous to develop a fuzzy logic prediction model for *CBR* of subgrade soil reinforced with geogrid using the soil parameters and conditions. Moreover, consider the data used in subgrade soil strength modeling are associated with some error, which makes the fuzzy logic approach more suitable. The relationships between the strength of geogrid-reinforced subgrade soil and the various input parameters used in this study are not continuously increasing or decreasing. For example, depth of geogrid reinforcement is an important design parameter for the reinforced subgrade at which maximum strength can be obtained. For the geosynthetic reinforcement placed at depth below and above the optimal depth, strength of the subgrade soil may be lower than the maximum possible value. In such cases, the trend cannot be established easily and fuzzy logic can be applied to get the more realistic solution.

2 Objectives of the Present Study

The objective of the present study is to develop a model for predicting the *CBR* value of subgrade soil reinforced with the geogrid layers at various depths using a fuzzy logic approach. This fuzzy logic model evaluates the effect of the location of geogrid reinforcement and the properties of subgrade soil on the CBR value of geogrid-reinforced subgrade. Laboratory tests have also been carried out on reinforced and unreinforced subgrade soils. The model has the ability to predict the *CBR* value of the geogrid-reinforced subgrade without any laboratory tests. The database for developing the model was gathered from the literature and laboratory experimentations. The fuzzy logic approach has been used for the first time in this study for the estimation of the strength of geogrid-reinforced subgrade soil with optimal depth of geogrid reinforcement. The developed models also suggest the range of optimal depth of geogrid

reinforcement at which the maximum strength can be obtained. The relative importance of model input parameters for affecting the output (*CBR* value) was also evaluated.

3 Database Development and Laboratory Testing

3.1 Database

To establish a fuzzy logic model for predicting the *CBR* value of subgrade soil, the database of the number of *CBR* tests as performed on geogrid-reinforced subgrade soil is employed. The data used in this research were obtained from the literature (Duncan-Williams and Attoh-Okine 2008; Rajesh et al. 2016; Singh and Gill 2012; Nagrale et al. 2010) and also considering the results from the *CBR* tests that were conducted in this study. The success of the model in predicting the strength of geogrid-reinforced subgrade soil depends on the comprehensiveness of the training data. Availability of experimental data, selection of variables which affect the subgrade soil strength, properties of geosynthetic, and the ambiguity in testing methods are some factors which are responsible for setting the criteria for identification of data. To maintain the uniformity in database, similar compaction tests were adopted for establishing the model. The standard Proctor compaction tests were carried out as per IS: 2720 (Part VII-1980) and AASHTO T-99 for determining the maximum dry unit weight and optimum moisture content for the subgrade soil. The samples were compacted with optimum moisture content for determination of *CBR* value of geogrid-reinforced subgrade soil at maximum dry unit weight. For determining the soaked *CBR* value, the soil samples were soaked for 4 days before testing. The improvement in the *CBR* value of weak subgrade soil is expected to be greater than that of strong subgrade soil by the inclusion of geosynthetic reinforcement. Therefore, laboratory tests were conducted on weak subgrade soil in the present study. In fact, in real-life projects, engineers rarely recommend the use of geosynthetic reinforcement if the subgrade soil is strong. Geosynthetic properties are excluded from the database because of the inadequate information available in the selected studies. However, the depth of geogrid reinforcement, which controls the effect of reinforcement on pavement layers significantly, has been included. It has been assumed that the geogrid used in this study meets the basic survivability and strength requirements. Depth of reinforcement is an important parameter for the development of the model because the improvement in the strength of subgrade soil depends on the location of geosynthetic placement within the subgrade soil. Determining the optimal depth of geogrid reinforcement helps achieve the maximum gain in strength of subgrade soil. To better understand the effect of geogrid reinforcement on the subgrade strength, a wide range of variations in depth of geogrid reinforcement is considered for the development of models such as $H/2$, $H/3$, $H/4$, $H/5$, $2H/5$, $3H/5$, and $4H/5$, where H is the height of the specimen in the *CBR* mold. As shown in Table 1, several parameters, such as depth of geogrid reinforcement, unreinforced/reinforced section, unsoaked/soaked condition, and soil properties, are considered datasets that were used to propose a prediction model for strength of subgrade soil reinforced with geogrid. Database used for the construction of the model consists of some important geotechnical and geosynthetic parameters which are presented in Table 1.

Table 1 The database used for the development of models

Soil parameters	Depth of reinforcement (mm)	Section	Condition	*CBR* (%)
LL (%) = 21, *PL* (%) = 18, *PI* (%) = 3, *OMC* (%) = 12.8, and *MDU* (kN/m^3) = 18.93	–	Unreinforced	Soaked	4.6[a]
	–	Unreinforced	Unsoaked	6.9[a]
	62.5	Reinforced	Soaked	5.51[a]
	62.5	Reinforced	Unsoaked	7.97[a]
LL (%) = 45.5, *PL* (%) = 22.15, *PI* (%) = 23.35, *OMC* (%) = 16.1, and *MDU* (kN/m^3) = 17.9	–	Unreinforced	Soaked	1.82[b]
	–	Unreinforced	Unsoaked	3.37[b]
	62.5	Reinforced	Unsoaked	6.11[b]
	62.5	Reinforced	Unsoaked	5.38[b]
	62.5	Reinforced	Soaked	3.92[b]
	62.5	Reinforced	Soaked	3.46 [b]
LL (%) = 36.9, *PL* (%) = 18.4, *PI* (%) = 18.76, *OMC* (%) = 13.21, and *MDU* (kN/m^3) = 17.36	–	Unreinforced	Soaked	1.91[b]
	–	Unreinforced	Unsoaked	4.83[b]
	62.5	Reinforced	Unsoaked	8.76[b]
	62.5	Reinforced	Unsoaked	8.57[b]
	62.5	Reinforced	Soaked	5.01[b]
	62.5	Reinforced	Soaked	4.28[b]
LL (%) = 28, *PL* (%) = 15, *PI* (%) = 13, *OMC* (%) = 16, and *MDU* (kN/m^3) = 16.67	–	Unreinforced	Soaked	2.9[c]
	–	Unreinforced	Unsoaked	6.5[c]
	25	Reinforced	Unsoaked	16[c]
	50	Reinforced	Unsoaked	13.8[c]
	75	Reinforced	Unsoaked	10.9[c]
	100	Reinforced	Unsoaked	7.2[c]
	25	Reinforced	Soaked	9.4[c]
	50	Reinforced	Soaked	7.2[c]
	75	Reinforced	Soaked	5.8[c]
	100	Reinforced	Soaked	3.16[c]
LL (%) = 34, *PL* (%) = 22, *PI* (%) = 12, *OMC* (%) = 17, and *MDU* (kN/m^3) = 16.9	–	Unreinforced	Soaked	1.1[d]
	25	Reinforced	Soaked	1.52[d]
	50	Reinforced	Soaked	1.84[d]
	75	Reinforced	Soaked	2.24[d]
	100	Reinforced	Soaked	2.52[d]
LL (%) = 29, *PL* (%) = 20, *PI* (%) = 9, *OMC* (%) = 13.91, and *MDU* (kN/m^3) = 18.74	–	Unreinforced	Unsoaked	1.66[e]
	62.5	Reinforced	Unsoaked	2.21[e]
	62.5	Reinforced	Unsoaked	1.76[e]
	41.67	Reinforced	Unsoaked	1.76[e]
	41.67	Reinforced	Unsoaked	2.08[e]
	31.25	Reinforced	Unsoaked	2.04[e]
	31.25	Reinforced	Unsoaked	2.04[e]

[a] Williams and Okine (2008)

[b] Rajesh et al. (2016)

[c] Singh and Gill (2012)

[d] Nagrale et al. (2010)

[e] CBR test results of the present study

3.2 Laboratory Tests

The database consists of 7 *CBR* test results conducted on the soil which was excavated from the area closer to the Civil Engineering Department of Delhi Technological University, Delhi, India. The characterization of the excavated soil was done by the standard laboratory tests. The soil was classified as silty sand (SM) as per the Indian Soil Classification System (IS: 2720). The standard Proctor compaction test was carried out as per IS: 2720 (Part VII) to determine the maximum dry unit weight (*MDU*) and optimum moisture content (*OMC*) for the excavated soil. The *MDU* and *OMC* of soil were found to be 18.74 kN/m^3 and 13.91%, respectively. Geogrid used in this study is composed of polypropylene materials with square aperture of 12.5 mm × 12.5 mm. The ultimate tensile strength of geogrid in machine direction and cross-machine direction is 115 kN/m, while the mass per unit area of the geogrid is 405 g/m^2.

CBR tests were carried out to investigate the influence of geogrid reinforcement on penetration resistance of the soil sample at its maximum dry unit weight obtained from the compaction test. The *CBR* test was performed in accordance with IS: 2720 (Part 16)-1979 as shown in Fig. 1. Laboratory tests were conducted on both unreinforced and geogrid-reinforced soil samples to observe strength improvement. *CBR* mold was filled with subgrade soil mixed thoroughly with water corresponding to *OMC* and compacted to the maximum dry unit weight. Soil samples for unreinforced and geogrid-reinforced soils were prepared to investigate the effect of geogrid reinforcement on the strength of subgrade. For preparing geogrid-reinforced soil specimens, the mold was filled with the soil by placing a layer of geogrid reinforcement which was cut in a circular shape with a diameter slightly less than the mold diameter. Geogrid reinforcement layer was laid at different depths of 62.5 mm, 41.67 mm, and 31.25 mm from the top surface of the specimen. Both the unreinforced and geogrid-reinforced soil samples were compacted in three layers with the help of a 2.6-kg rammer falling from a height of 310 mm, and each layer was compacted with 56 blows. *CBR* value is defined as the ratio of load corresponding to 2.5-mm or 5-mm penetration into the soil with a 50-mm-diameter penetration plunger at a constant rate of 1.25 mm/min to that of the standard unit load. The load corresponding to the penetration ranging from 0 to 12.5 mm is measured in the experiment. *CBR* value was calculated by measuring the load corresponding to 2.5-mm and 5.0-mm penetration and maximum value was considered the *CBR* value, as recommended in the relevant standard.

4 Development of Fuzzy Logic Prediction Model

4.1 Fuzzy Logic Approach

The fuzzy logic approach deals with vagueness and imprecision. Fuzzy logic has the potential to generate solution when the available data are imprecise, ambiguous, incomplete, or not correct. Figure 2 illustrates the basic concept of the FL model with input variables and output. The fuzzy logic system consists of four sections (Zedeh and Kacprzyk 1992; Gopal 2009; Hossain et al. 2012):



Fig. 1 CBR test apparatus

fuzzification, rule base, fuzzy inference unit, and defuzzification. The fuzzification component represents the process of taking the crisp inputs and converts them into a linguistic variable with the help of membership functions that were stored in the fuzzy knowledge base. Fuzzy membership functions are available in many forms, but the preference of choosing the type of membership function for any practical application is provided in the form of simple linear functions, namely Triangular and Gaussian membership functions. The rule base defines a set of fuzzy if–then rules representing the relationship between input variables and output. Two types of fuzzy rule systems are available, namely Mamdani and Sugeno. Selection of an appropriate rule system is totally dependent on the users' problem. The fuzzy inference system consists of membership functions, operators, and a number of if–then rules that transform the available inputs to the output. Defuzzification is the process of converting the resulting output fuzzy variables from the fuzzy inference unit into a crisp value (numerical value). Structure identification and parameter prediction are the two phases of fuzzy modeling (Topçu and Saridemir 2008). The components of structure identification include appropriate selection of input parameters, selection of a particular type of fuzzy inference systems, assigning a type and number of membership function, defining of rules, and determining their number. Parameter prediction determines the value of output parameter corresponding to the input values of the proposed model. The methodology used to develop the fuzzy model is described in a flowchart as shown in Fig. 3.

4.2 Model Construction

Fuzzy logic approach has been used in the present study to establish a relationship between the *CBR* of geogrid-reinforced subgrade soil with the engineering properties of subgrade soil and depth of geogrid reinforcement. Mamdani-type fuzzy inference system was selected to develop the model using the results of experimental tests. Mamdani fuzzy inference system is the most commonly used inference engine in developing fuzzy models. A wide range and varieties of geosynthetics are available in the market, namely geotextile, geogrids, geomats, geocomposite, and geomembrane. Out of these, geogrid was selected for the development of a prediction model because geogrid has been widely used in unpaved roads with various applications. The fuzzy logic model was developed with eight input parameters and the *CBR* value of reinforced subgrade soil as the model single output parameter in MATLAB software. The input parameters are reinforced/unreinforced section (*R*/*UR*), depth of reinforcement (*DOR*), liquid limit (*LL*), plastic limit (*PL*), plasticity index (*PI*), optimum moisture content (*OMC*), maximum dry unit weight (*MDU*), and soaked/unsoaked condition (*S*/*US*). No data was available in the literature where the fuzzy model was developed to know the effect of geogrid reinforcement on the reinforced subgrade soil. However, it was available in the literature that the strength of subgrade depends on the properties of geogrid, location of reinforcement, and basic properties of soil such as liquid limit, moisture content, maximum dry density, plastic limit, *CBR*, and compaction. Hence, based on the literature review, the input parameters were selected, which significantly affects the reinforced subgrade soil to establish the fuzzy model for predicting the subgrade soil strength. The selection of realistic input parameters is essential for a successful prediction model. The statistical parameters for the input variables and output were calculated as summarized in Table 2.

4.3 Membership Functions

Membership functions of the model map an input value to its appropriate membership value which lies between 0 and 1. Selection of membership functions, as

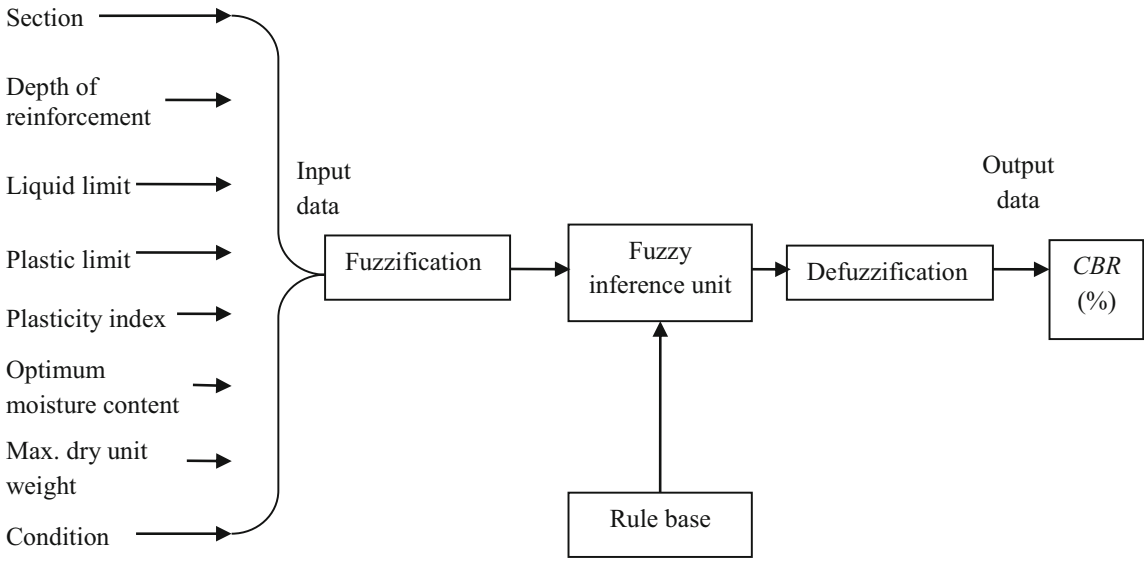


Fig. 2 Schematic representation of activity units in the fuzzy model with input variables and *CBR* as output

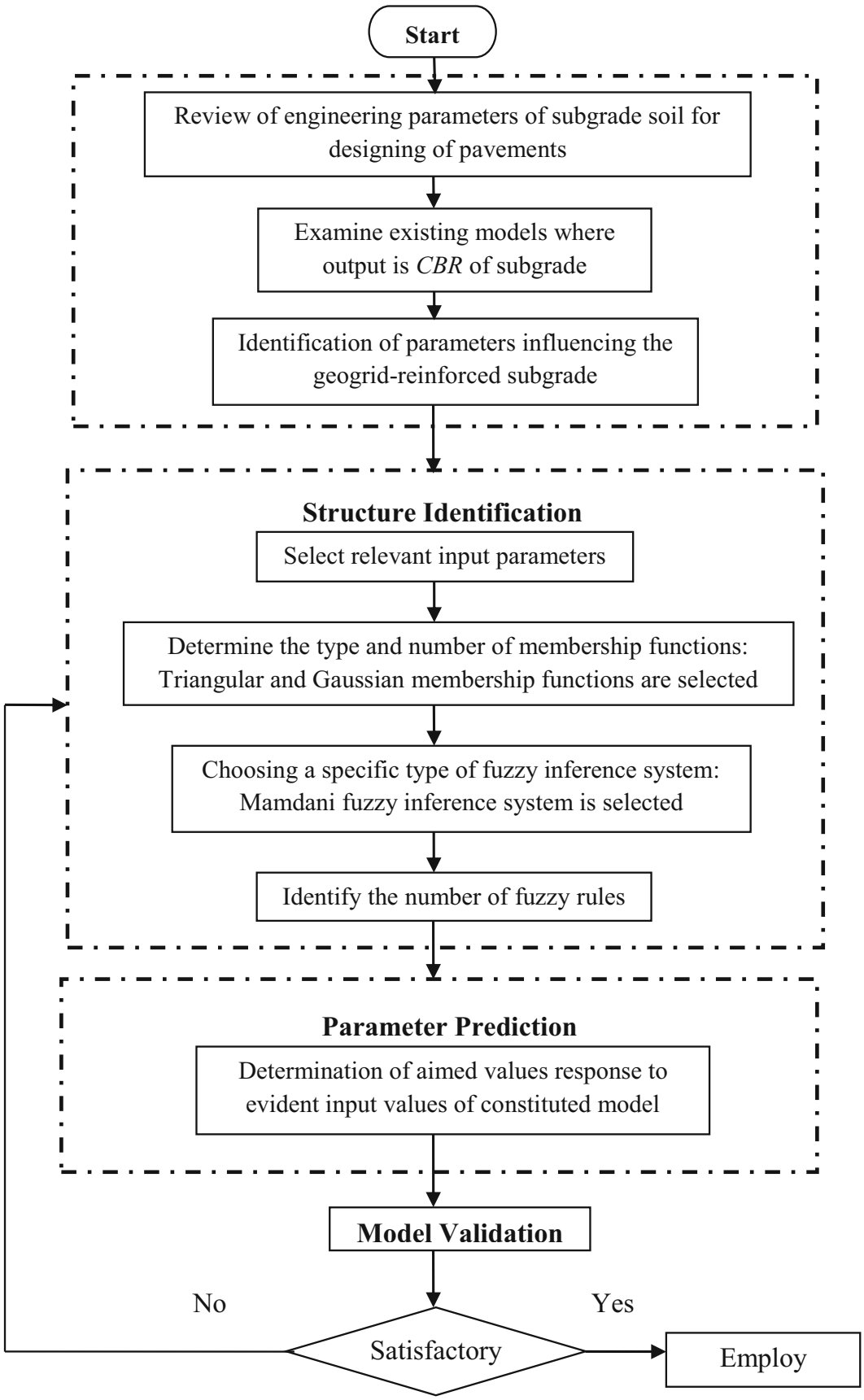


Fig. 3 Flowchart for the methodology adopted for the development of models

well as the number of membership function, is dependent on system knowledge, experience gained, and research conditions. Initially, fuzzification of all input

Table 2 Statistical parameters of variables used for the development of fuzzy model

Variables	Statistical parameters			
	Minimum	Maximum	Mean	Standard deviation
Section (reinforced/unreinforced)	0	1	0.5	0.707
Depth of reinforcement (mm)	20	110	65	63.639
Liquid limit (%)	20	50	35	21.213
Plastic limit (%)	10	25	17.5	10.606
Plasticity index (%)	1	25	13	16.970
Optimum moisture content (%)	11	20	15.5	6.363
Maximum dry unit weight (kN/m^3)	16	19	17.5	2.121
Condition (soaked/unsoaked)	0	1	0.5	0.707
CBR (%)	1	17	9	11.313

parameters and output parameter of the proposed model is occurred with the help of membership function After defining the range of values to all the input variables and output, assign appropriate membership functions to fuzzy variables. Triangular-shaped membership functions and Gaussian-shaped membership functions were used in the investigation for input and output variables as shown in Fig. 4. Membership value defines the degree an object belongs to a fuzzy set between 0 and 1, where 1 represents full membership and 0 expresses non-membership even if it allows having partial membership. Two fuzzy logic models were developed, one with Triangular membership function and the other with Gaussian membership function. The Triangular membership functions are formed with the help of a straight line and to understand the straight lines is simpler. The Triangular membership function is defined by a lower limit "a," an upper limit "b," and a value "m," where $a < m < b$ as shown in Eq. (1). The Gaussian membership functions are also popular because the patterns so formed are smooth and nonzero at all points. Gaussian membership function is defined by a central value "m" and a standard deviation $\sigma > 0$ as shown in Eq. (2).

$$f(x;a,m,b) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{m-a}, & a < x \le m \\ \dfrac{b-x}{b-m}, & m < x < b \\ 0, & x \ge b \end{cases} \tag{1}$$

$$f(x;\sigma,m) = e^{\frac{-(x-m)^2}{2\sigma^2}} \tag{2}$$

4.4 Rule Base

Fuzzy rule base consists of an entire group of if–then rules that include all possible fuzzy relations between input variables and output. Fuzzy sets and fuzzy operators are necessary to form rules to perform the fuzzy logic prediction model. If–then rules are used in developing the model to create the conditional statements that resemble human thinking rather than any mathematical rules. The general form of a fuzzy if–then rule statement is "if x is A then y is B" where A and B are linguistic variables, respectively. The if part of the rule statement "x is A" is defined as the antecedent and the then part of the rule statement "y is B" is defined as the consequent. The rules are responsible for

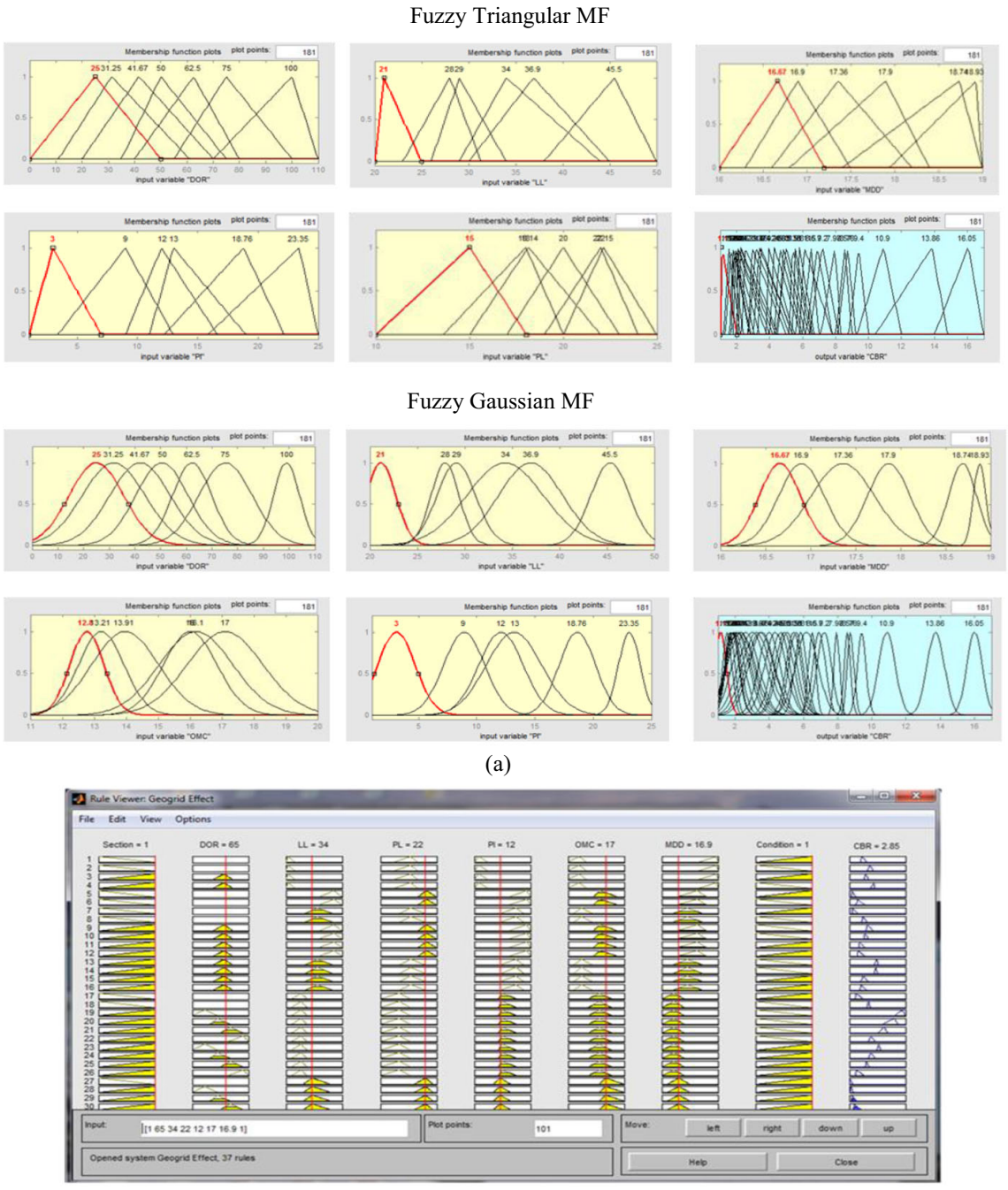


Fig. 4 (**a**) Few fuzzy Triangular and Gaussian membership functions for input variables (yellow background) and output variables (blue background). (**b**) Typical rule viewer window for the fuzzy logic model

determining the input and output membership functions; hence, 37 fuzzy rules were generated for implementation of the prediction model. The rule viewer of the geogrid-reinforced prediction model is shown in Fig. 4. Rule viewer can be shown after the assessment of the applied if–then rules in the fuzzy model and then the output can be computed. Generally, the more the number of rules, the greater would be the accuracy, but dealing with a larger rule base would be time consuming. With higher numbers of subsets, fuzzy logic system would work best when the set of rules is derived from the actual experimental results, which helps a lot in precise linking of various input parameters to the output. The fuzzy inference unit extracts the entire group of rules from the rule base and gains an understanding of converting inputs to the corresponding output. Center of gravity method was applied for converting linguistic output into a numerical value during the defuzzification process. Center of gravity method is mostly adopted for the defuzzification process.

5 Results and Discussion

5.1 Evaluation of Model Performance

The performance of both the models developed for the estimation of the strength of geogrid-reinforced subgrade soil was evaluated through the statistical criteria for performance measures, which include mean absolute error (*MAE*), root mean square error (*RMSE*), correlation coefficient (r), and coefficient of determination (R^2). These performance indicators are used for comparing the output results obtained from the established models with the actual outputs. These statistical performance measures are selected to evaluate the accuracy of the results obtained from the developed models and are summarized in Table 3. It can be observed that the satisfactory agreement between the models' predicted values and experimental data is found. If the value of correlation coefficient is more than 0.8, then it represents a strong correlation between the predicted and measured values of the database (Smith 1986). Based on the values of performance indicators, it can be concluded that the model with Triangular membership function performs better than the model with Gaussian membership function, in terms of a higher value of the coefficient of determination (R^2) and correlation coefficient (r) and lower value of errors. A prediction model with a higher coefficient of determination and lower error indicates good consent with laboratory test data. By the way, the magnitudes of errors are in the reasonable range for both the models and even very close to each other to use them for the prediction of geogrid-reinforced subgrade strength. The models were calibrated with input and output data present in the database of experimental results, and the same database has been used for the validation. Experimental values are compared with the predicted *CBR* values obtained from the models with Triangular and Gaussian membership functions shown in Fig. 5. It is concluded that fuzzy models are highly reliable to forecast the strength of geogrid-reinforced subgrade. The correlation coefficients between the predicted and experimental *CBR* values computed for geogrid-reinforced subgrade soil from both the models were found very close to unity as illustrated in Fig. 6. It can be observed that almost all the points were around the equality line. The results of the statistical indicators demonstrate that the fuzzy logic is a powerful tool for predicting the strength of

geogrid-reinforced subgrade soil and model's predicted values are near the experimental values.

5.2 Control Surfaces and Optimal Depth of Reinforcement

Figures 7 and 8 present the control surface diagrams for both the models with Triangular and Gaussian membership functions representing the interdependency of input parameters and output. It can be observed that each surface plot is showing the effect of two parameters at a time on the output. Surface diagrams are limited to only two input parameters because the visualization of figures will become more difficult if we increase the number of input variables by more than 2. Figure 7 shows the three-dimensional description of models with Triangular membership function which takes the combined effect of two input parameters on the strength of geogrid-reinforced subgrade soil. There is a particular range of possible defuzzified values for the depth of geogrid reinforcement in which the *CBR* increases gradually and beyond that range it gradually decreases. Therefore, it is important to place the reinforcement layer at the optimal depth of geogrid reinforcement to get the maximum strength of the geogrid-reinforced subgrade. For finding the optimal location of geogrid reinforcement, the predicted values of *CBR* are obtained from the proposed model for the given database by varying the depth of the reinforcement parameter in a range of 20 to 80 mm. Variation in the strength was observed at an interval of 10 mm from 20 to 80 mm. While calculating the optimal depth of geogrid reinforcement, only the parameter "depth of reinforcement" is varied and all other remaining parameters are kept constant as per the database. Optimal depth of geogrid reinforcement was found as 36 to 60% of the height of the subgrade soil specimen; hence, maximum strength can be gained if the geogrid is placed at a depth between 45 and 75 mm from the top surface of the soil specimen. As the optimum moisture content and liquid limit increase, initially the strength decreases then remains constant and then reaches its respective maximum value. The *CBR* reaches the peak when the section is reinforced and the least value is obtained when the section is unreinforced. Figure 8 shows the three-dimensional description of the model with Gaussian membership function which shows the strength of the reinforced soil is maximum at the maximum value of optimum moisture content then decreases suddenly as the optimum moisture content decreases. Similarly, the

Table 3 Statistical performance of the fuzzy models with varied membership functions

Developed models	Performance measures			
	Root mean square error (*RMSE*)	Mean absolute error (*MAE*)	Coefficient of determination (R^2)	Correlation coefficient (*r*)
Fuzzy model with Triangular membership function	0.4705	0.3897	0.9927	0.9963
Fuzzy model with Gaussian membership function	0.6207	0.4913	0.9773	0.9886

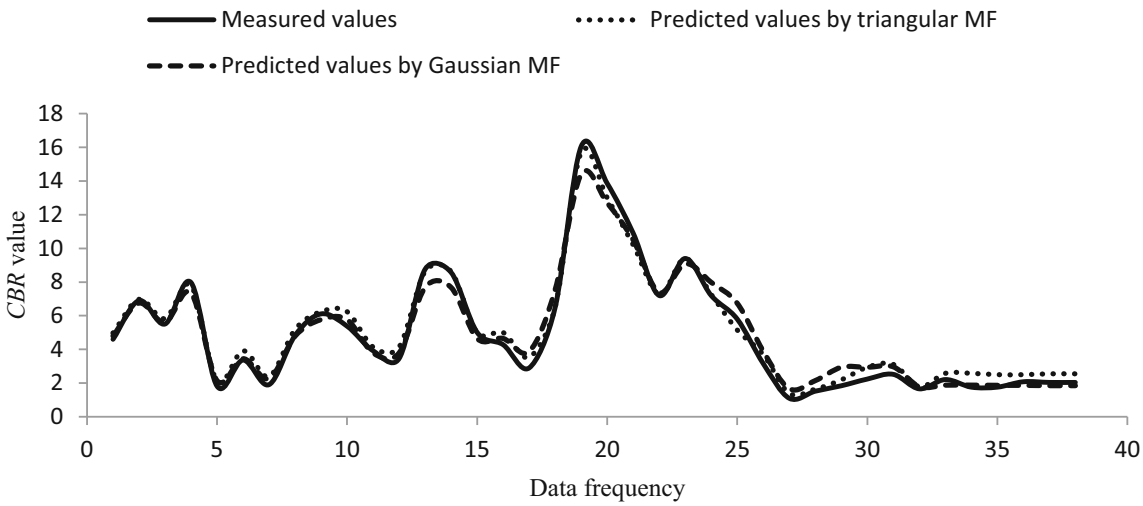


Fig. 5 Comparison between *CBR* value predicted by fuzzy model with Triangular membership function and Gaussian membership function with actual experimental values

maximum value of strength can be obtained if the reinforcement layer is placed within an optimal range, beyond which the effect is insignificant.

For the design of pavements, subgrade strength plays an important role which needs to be evaluated in terms of *CBR* value. The range considered for the *CBR* value of subgrade varies from 1 to 16% for the development of fuzzy logic model. Subgrade strength can be classified into various classes based on the *CBR* value as shown in Table 4. The strength of unreinforced subgrade which lies under class S1 has been improved to classes S2 and S3 due to the presence of geogrid reinforcement at depths $H/5$ to $4H/5$ and $H/2$, respectively. The strength of unreinforced subgrade which lies under class S2 has been improved to classes S3 and S4 due to the presence of geogrid reinforcement at depths $H/2$ and $H/5$ to $4H/5$, respectively. Similarly, the strength of unreinforced subgrade which lies in classes S3 and S4 has been improved to classes S4 and S5, respectively. It is clear from the results that inclusion of geogrid reinforcement to the subgrade improves its *CBR* value. The strength of geogrid-reinforced subgrade is improved due to good interlocking and frictional capability of geogrid against lateral movement. However, the dominating mechanism responsible for the enhancement in the *CBR* value of geogrid-reinforced subgrade is the maximum stress in the subgrade that is reduced because of the presence of a geogrid reinforcement layer within the subgrade (Cuelho and Perkins 2017; Singh et al. 2019). Geogrid reinforcement will help distribute the vertical stresses more uniformly. As per IRC: 37-2012, as the *CBR* value of subgrade soil increases, it will decrease the thickness of granular base course required for the pavement design. Ultimately, it will reduce the construction cost which helps the economy.

5.3 Sensitivity Analysis

Sensitivity analysis was conducted to investigate the effectiveness of the input parameters on the strength of geogrid-reinforced subgrade soil. The cosine amplitude method was implemented to evaluate the sensitivity of each input parameter on the *CBR* of geogrid-reinforced subgrade and to recognize the most sensitive

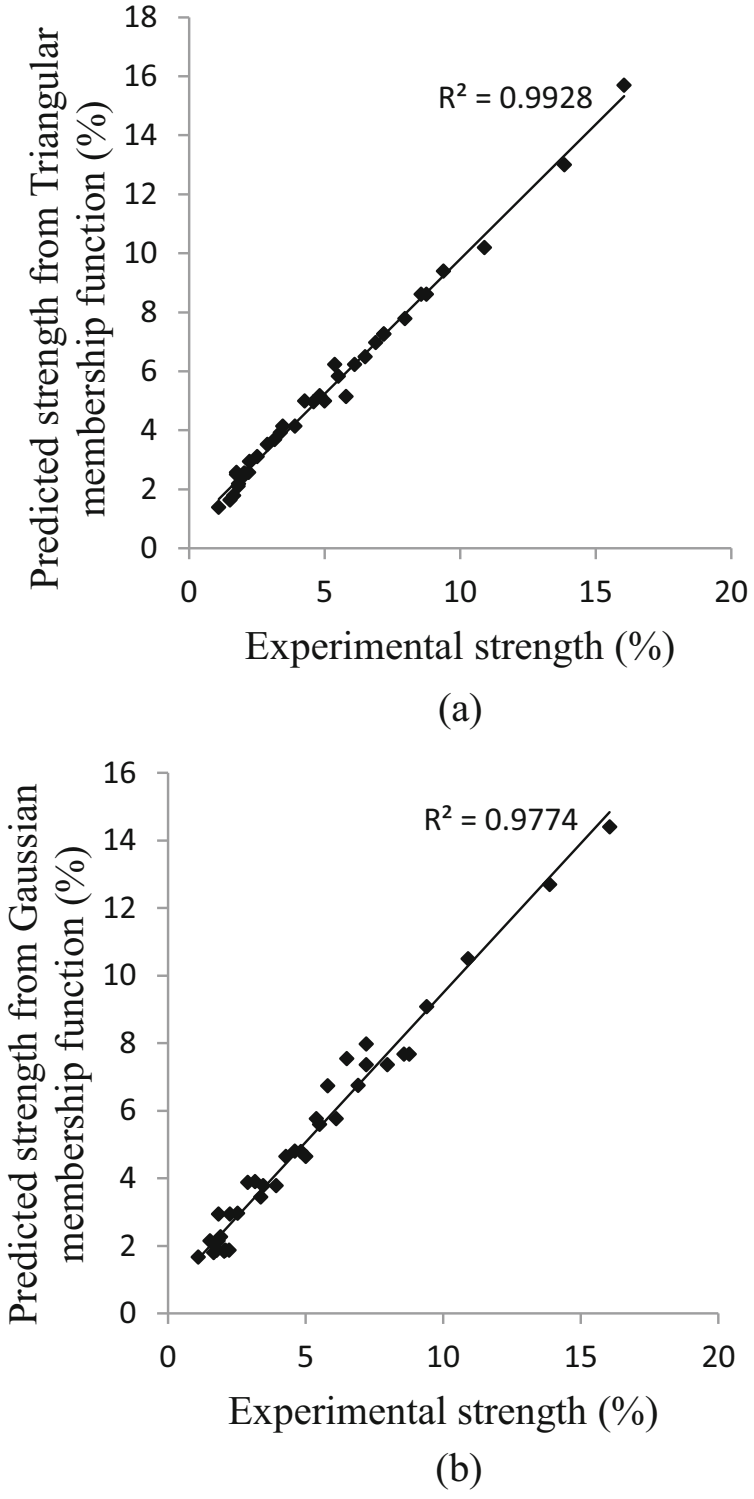


Fig. 6 Correlation between experimental and predicted values of *CBR* by (**a**) model with Triangular membership function, and (**b**) model with Gaussian membership function

factor affecting the performance. Assign a degree of sensitivity to each input parameter during the sensitivity analysis by establishing the strength of the relationship (r_{ij}) between the input parameters and *CBR* of geogrid-reinforced subgrade soil. The strength of the relationship between various input parameters and output can be estimated from Eq. (3).

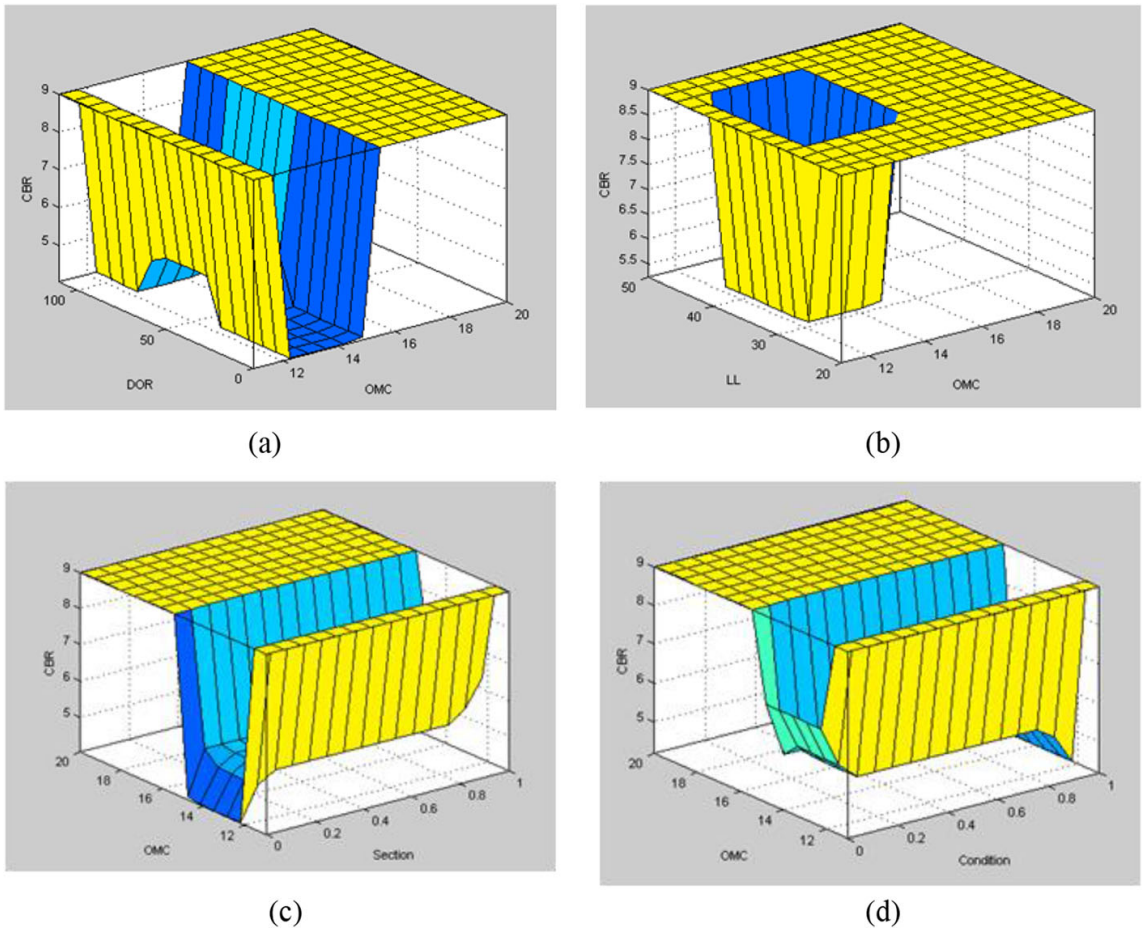


Fig. 7 Few control surfaces of the fuzzy model with Triangular membership function on *CBR* with combined effect of (**a**) *DOR* and *OMC*, (**b**) *LL* and *OMC*, (**c**) *OMC* and section, (**d**) *OMC* and condition

$$r_{ij} = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{m} x_{ik}^2 \sum_{k=1}^{m} x_{jk}^2}} \quad , \quad 0 < \text{rij} < 1 \tag{3}$$

where r_{ij} = strength of the correlation between input parameter and output, x_i = input parameter, x_j = output parameter, and m = number of input variables.

Figure 9 shows the effect of each input parameter used in the proposed fuzzy model on the output results. The value of r_{ij} for each input parameter is an indicator of the effect of that parameter on the output results. A higher value of r_{ij} represents a greater effect of that parameter on the output. According to the sensitivity analysis results, maximum dry density and optimum moisture content of subgrade soil have the most significant effect on the strength of reinforced subgrade, whereas the soaked/unsoaked condition has a marginal effect only. Liquid limit, depth of geogrid reinforcement, plasticity index, plastic limit, and reinforced/unreinforced section were also found to have a good contribution to the strength of reinforced subgrade. A very low difference is observed between the values of r_{ij} for seven input parameters except for one input, i.e., "soaked/unsoaked condition." Depth of reinforcement has a high influence on the reinforced subgrade strength because providing a reinforcement layer within the subgrade soil increases the stiffness of the soil which further increases the strength of the reinforced subgrade.

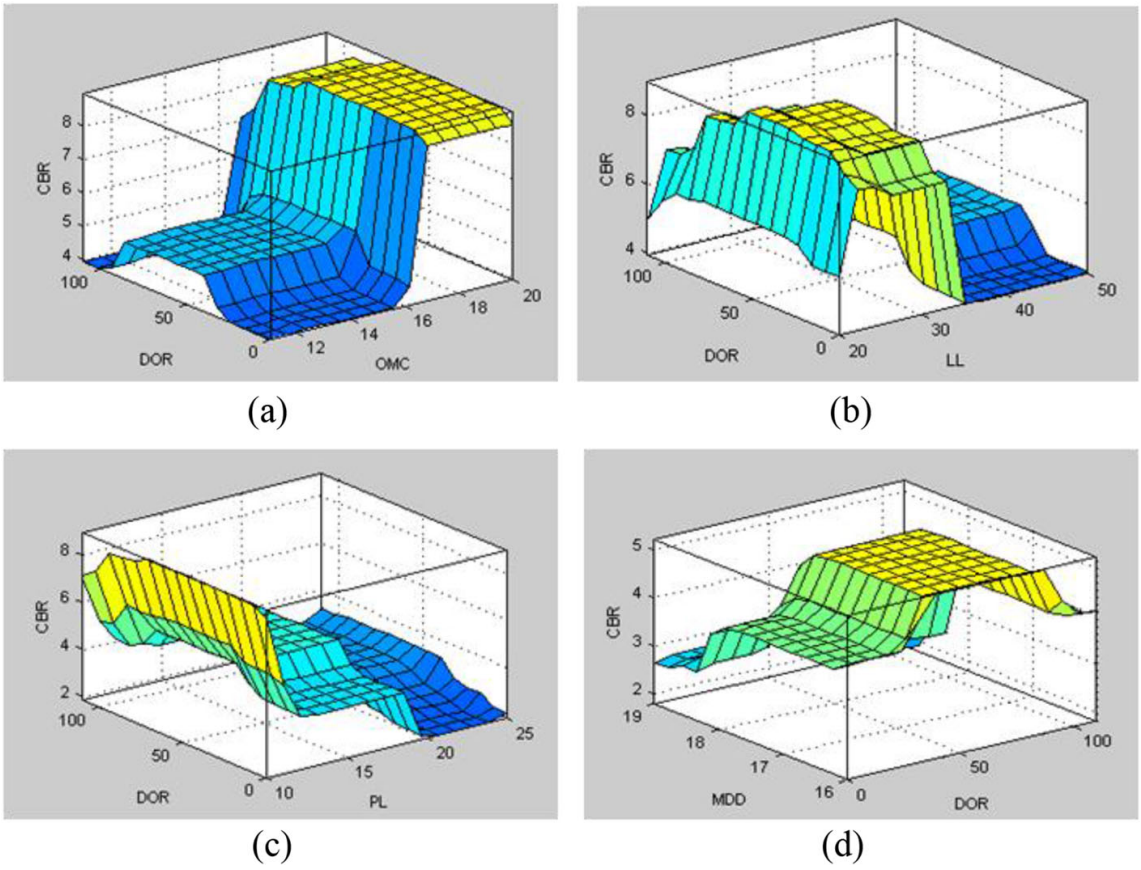


Fig. 8 Few control surfaces of the fuzzy model with Gaussian membership function on *CBR* with combined effect of (**a**) *DOR* and *OMC*, (**b**) *DOR* and *LL*, (**c**) *DOR* and *PL*, and (**d**) *MDD* and *DOR*

6 Conclusions

The present study is likely to be one of the very first research which aims at investigating the applicability of fuzzy logic model for estimating the geogrid-reinforced subgrade soil strength. To achieve this, two models were developed with Triangular and Gaussian membership functions with eight input parameters and one output parameter. Output results were obtained from the proposed model, and they were compared with the laboratory test results. The correlation coefficients of models were found to be 0.996 and 0.988, suggesting a good correlation between predicted *CBR* values and laboratory *CBR* values. Therefore, the strength of subgrade soil reinforced with geogrid can be estimated through fuzzy logic model depending on the subgrade

Table 4 Subgrade soil strength as classified into the classes

Quality of subgrade	Range of *CBR* value (%)	Class
Very low	< 2	S1
Low	3–4	S2
Average	5–6	S3
Good	7–9	S4
Very good	10–15	S5

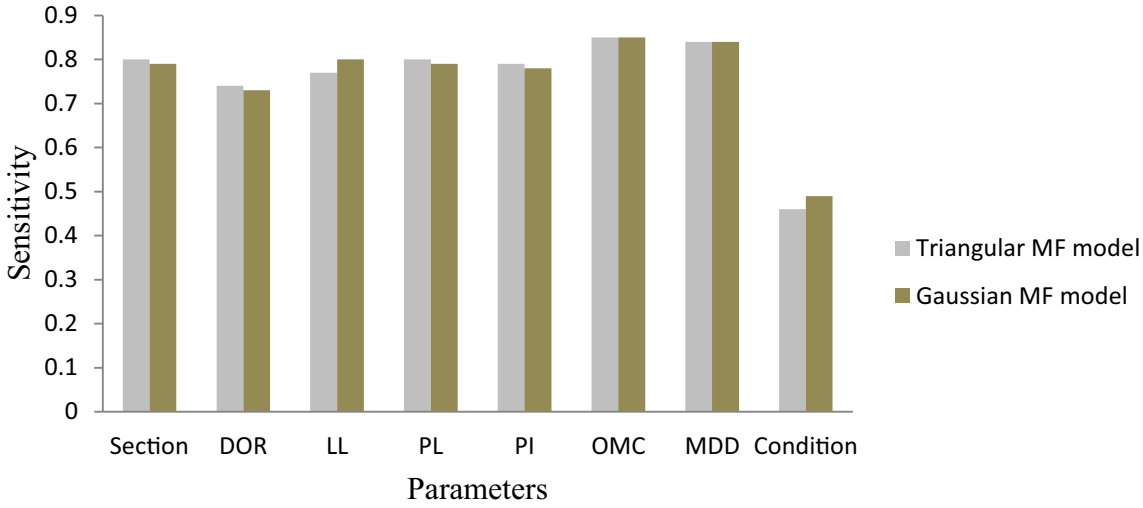


Fig. 9 Sensitivity analysis for various input parameters used in the model

soil properties and depth of geogrid reinforcement, thus saving cost and time. This study as presented also reveals that there is an improvement in the *CBR* value of subgrade soil by stabilizing it with geogrid. The potentialities of fuzzy logic were found to be satisfactory in predicting the behavior of geogrid-reinforced subgrade soil subjected to traffic loads. Based on the results and discussion presented earlier, the following conclusions can be drawn:

(a) Fuzzy logic modeling is a convenient approach for predicting the strength of geogrid-reinforced subgrade soil, thus saving time and resources in real-life projects.
(b) The correlation coefficient is found to be closer to unity for both the developed models having Triangular membership function and Gaussian membership function, thus validating the results.
(c) Sensitivity analysis evaluated the contribution of individual input parameters on the *CBR* value of geogrid-reinforced subgrade soil and it reveals that $r_{ij} > 0.7$ for all the input parameters except for "soaked/unsoaked condition." This demonstrates the good selection of input parameters for the development of models.
(d) As the availability of data on geogrid-reinforced subgrade soil increases in the literature, a more accurate prediction model can be proposed for a wider range.
(e) The range of the optimal depth of geogrid reinforcement is found as 36 to 60% of the thickness of the subgrade soil layer.

In future studies, more input variables can be included, such as specific property of geosynthetics, for developing the fuzzy logic model for predicting strength of geogrid-reinforced subgrade soil. The predictive capabilities of the proposed models with Triangular and Gaussian membership functions are limited to the range of the data used for their calibration. As more data become available, by including them for other types of soils and test conditions, the proposed models can be improved to make more accurate predictions for a wider range.

Data Availability All data used to generate the models during the present study appear in the submitted article.

References

Bagdatli, M.E.C.: Fuzzy logic-based life-cycle cost analysis of road pavements. J. Transp. Eng., Part B: Pavements. **144**(4), 04018050 (2018). https://doi.org/10.1061/JPEODX.0000081

Bhatt, S., Jain, P.K., Pradesh, M.: Prediction of California bearing ratio of soils using artificial neural network. American International Journal of Research in Science, Technology, Engineering & Mathematics. **8**(2), 156–161 (2014)

Black, W.P.M.: A method of estimating the CBR of cohesive soils from plasticity data. Geotechnique. **12**, 271–272 (1962)

Cabalar, A.F., Cevik, A., Gokceoglu, C.: Some applications of adaptive neuro-fuzzy inference system (ANFIS) in geotechnical engineering. Comput. Geotech. **40**, 14–33 (2012). https://doi.org/10.1016/j.compgeo.2011.09.008

Cuelho, E.V., Perkins, S.W.: Geosynthetic subgrade stabilization—field testing and design method calibration. Transp Geotech. **10**, 22–34 (2017). https://doi.org/10.1016/j.trgeo.2016.10.002

Duncan-Williams, E., Attoh-Okine, N.O.: Effect of geogrid in granular base strength—an experimental investigation. Constr. Build. Mater. **22**(11), 2180–2184 (2008). https://doi.org/10.1016/j.conbuildmat.2007.08.008

Giroud, J.P., Han, J.: Design method for geogrid-reinforced unpaved roads. I. Development of design method. J. Geotech. Geoenviron. **130**(8), 775–786 (2004). https://doi.org/10.1061/(ASCE)1090-0241(2004)130:8(775)

Gopal, M.: Digital control and state variable methods: conventional and intelligent control systems, 3rd edn. Tata McGraw-Hill Education Pvt. Ltd, Singapore (2009)

Günaydin, O.: Estimation of soil compaction parameters by using statistical analyses and artificial neural networks. Environ. Geol. **57**(1), 203–215 (2009)

Gurtug, Y., Sridharan, A.: Prediction of compaction characteristics of fine-grained soils. Géotechnique. **52**(10), 761–763 (2002)

Hossain, A., Rahman, A., Mohiuddin, A.K.M.: Fuzzy evaluation for an intelligent air-cushion tracked vehicle performance investigation. J. Terrramech. **49**(2), 73–80 (2012). https://doi.org/10.1016/j.jterra.2011.08.002

Ibrahim, E.M., El-Badawy, S.M., Ibrahim, M.H., Gabr, A., Azam, A.: Effect of geogrid reinforcement on flexible pavements. Innov. Infrastruct. Solut. **2**(1), 54 (2017). https://doi.org/10.1007/s41062-017-0102-7

Lee, D., Donnell, E.T.: Analysis of nighttime driver behavior and pavement marking effects using fuzzy inference system. J. Comput. Civ. Eng. **21**, 200–210 (2007). https://doi.org/10.1061/ASCE0887-3801200721:3200

Moghaddas-Nejad, F., Small, J.C.: Effect of geogrid reinforcement in model track tests on pavements. J. Transp. Eng. **122**(6), 468–474 (1996)

Nagrale, P.P., Sawant, P.H., Pusadkar, S.S.: Laboratory investigation of reinforced sub grade soils. Indian Geotechnical Conference, GEOtrendz. 637–640 (2010)

Negi, M.S., Singh, S.K.: Experimental and numerical studies on geotextile reinforced subgrade soil. Int. J. Geotech. Eng. 1–12 (2019). https://doi.org/10.1080/19386362.2019.1684654

Palmeira, E.M., Antunes, L.G.S.: Large scale tests on geosynthetic reinforced unpaved roads subjected to surface maintenance. Geotext. Geomembr. **28**(6), 547–558 (2010). https://doi.org/10.1016/j.geotexmem.2010.03.002

Perkins, S.W., Christopher, B.R., Lacina, B.A., Klompmaker, J.: Mechanistic-empirical modeling of geosynthetic-reinforced unpaved roads. Int J Geomech. **12**(4), 370–380 (2012). https://doi.org/10.1061/(ASCE)GM.1943-5622.0000184

Rajesh, U., Sajja, S., Chakravarthi, V.K.: Studies on engineering performance of geogrid reinforced soft subgrade. Transp. Res. Procedia. **17**, 164–173 (2016). https://doi.org/10.1016/j.trpro.2016.11.072

Rashidian, V., Naeini, S.A., Mirzakhanlari, M.: Laboratory testing and numerical modeling on bearing capacity of geotextile reinforced granular soils. Int. J. Geotech. Eng. **12**, 241–251 (2016). https://doi.org/10.1080/19386362.2016.1269042

Sandra, A.K., Sarkar, A.K.: Application of fuzzy logic and clustering techniques for pavement maintenance. Transp. Infrastruct. Geotech. **2**, 103–119 (2015). https://doi.org/10.1007/s40515-015-0021-z

Shukla, S.K.: An introduction to geosynthetic engineering. CRC Press, London (2016)

Singh, P., Gill, K.S.: CBR improvement of clayey soil with geo-grid reinforcement. International Journal of Emerging Technology and Advanced Engineering. **2**(6), 315–318 (2012)

Singh, M., Trivedi, A., Shukla, S.K.: Strength enhancement of the subgrade soil of unpaved road with geosynthetic reinforcement layers. Transp Geotech. **19**, 54–60 (2019). https://doi.org/10.1016/j.trgeo.2019.01.007

Smith, G.N.: Probability and statistics of civil engineering. Collins, London (1986)

Stephens, D.J.: Prediction of the California bearing ratio. J. S. Afr. Inst. Civ. Eng. **32**(12), 523–527 (1990)

Suku, L., Prabhu, S.S., Babu, G.L.S.: Effect of geogrid-reinforcement in granular bases under repeated loading. Geotext. Geomembranes. **45**(4), 377–389 (2017). https://doi.org/10.1016/j.geotexmem.2017.04.008

Taghavifar, H., Mardani, A.: Fuzzy logic system based prediction effort: a case study on the effects of tire parameters on contact area and contact pressure. Appl Soft Comput. **14**, 390–396 (2014). https://doi.org/10.1016/j.asoc.2013.10.005

Taskiran, T.: Prediction of California bearing ratio (CBR) of fine grained soils by AI methods. Adv. Eng. Softw. **41**(6), 886–892 (2010). https://doi.org/10.1016/j.advengsoft.2010.01.003

Tenpe, A.R., Patel, A.: Application of genetic expression programming and artificial neural network for prediction of CBR. Road Materials and Pavement Design. 1–18 (2018). https://doi.org/10.1080/14680629.2018.1544924

Topçu, I.B., Saridemir, M.: Prediction of mechanical properties of recycled aggregate concretes containing silica fume using artificial neural networks and fuzzy logic. Comput. Mater. Sci. **42**(1), 74–82 (2008). https://doi.org/10.1016/j.commatsci.2007.06.011

Trivedi, A.: Estimating in situ deformation of rock masses using a hardening parameter and RQD. Int J Geomech. **13**(4), 348–364 (2013)

Venkatasubramanian, C., Dhinakaran, G.: ANN model for predicting CBR from index properties of soils. Int. J. Civ. Struct. Eng. **2**(2), 605–611 (2011)

Wu, H., Huang, B., Shu, X., Zhao, S.: Evaluation of geogrid reinforcement effects on unbound granular pavement base courses using loaded wheel tester. Geotext. Geomembranes. **43**(5), 462–469 (2015). https://doi.org/10.1016/j.geotexmem.2015.04.014

Yildirim, B., Gunaydin, O.: Expert systems with applications estimation of California bearing ratio by using soft computing systems. Expert Syst. Appl. **38**(5), 6381–6391 (2011)

Zedeh, L., Kacprzyk, J.: Fuzzy logic for the management of uncertainty. John Wiley and Sons, Inc, New York (1992)

Zehtabchi, A., Hashemi, S.A.H., Asadi, S.: Predicting the strength of polymer-modified thin-layer asphalt with fuzzy logic. Constr Build Mater. **169**, 826–834 (2018). https://doi.org/10.1016/j.conbuildmat.2018.02.002

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

Affiliations

Meenakshi Singh[1] · **Ashutosh Trivedi**[1] · **Sanjay Kumar Shukla**[2,3]

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

[2] Discipline of Civil and Environmental Engineering, School of Engineering, Edith Cowan University, Joondalup, Australia

[3] School of Building and Civil Engineering, Fiji National University, Suva, Fiji

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Genetic CoCoA++: Genetic Algorithm based Congestion Control in CoAP

Rajesh Kumar Yadav
Department of Computer Science & Engineering
Delhi Technological University,
Delhi, India
rkyadav@dtu.ac.in

Navpreet Singh
Department of Computer Science & Engineering
Delhi Technological University,
Delhi, India
singhnav232@gmail.com

Prince Piyush
Department of Computer Science & Engineering
Delhi Technological University,
Delhi, India
princepiyush2398@gmail.com

***Abstract*— CoAP is the protocol used in various IoT device communications. Congestion in the network is a reason for packet loss during the transmission. Packet loss can be reduced via reducing congestion in the network. CoAP uses a simple binary exponential backoff factor which does not consider the dynamic nature of the network. A major enhancement for congestion control in the IoT network is achieved via CoCoA which somehow considers the network dynamics a little bit. New enhancements over this protocol are made overtime but those enhancements are based on the static observation of RTT. In this paper, a dynamic approach based on the observation of the network for a fixed interval of time, and based on the genetic algorithm is discussed. This approach also does not store previous states of the network for congestion control and RTO estimation. We worked on this approach and a significant decrease in RTO and PFR is observed via this approach.**

Keywords—CoAP; congestion control; IoT Network; COOJA; Contiki.

I. INTRODUCTION

Enlargement of the IoT technology space is happening at an exponential rate and hence the IoT devices, sensors, etc. The protocol which is being used widely for the IoT devices communication is the constrained application protocol (CoAP). It works on the top of the user datagram protocol (UDP). It is the application-layer protocol used for HTTP-like interactions that are IP-based. The technologies used for communication within a network of IoT devices have some limitations in the form of small data rates and comparatively worse bit error rates (BER). The network congestion problem plays a vital role while engineering a new communication paradigm. Network congestion occurs when the traffic load of the network reaches the capacity of the network. TCP takes care of the congestion by itself and a congestion control mechanism needs to develop for UDP. Traditional network traffic patterns are different from an IoT network traffic pattern. An individual IoT device does not create a large amount of data. It is the cluster of IoT devices that lead to congestion in the network. Reaction to the events can also become a reason for network congestion. The CoAP that was standardized by IETF provides a congestion control mechanism that operates only with two variables, namely binary exponential backoff factor (BEB) and retransmission timeout (RTO). The identification of congestion in the network is generally determined by the packet drops. The CoAP does not do dynamic calculations based on current network conditions. It just doubles the RTO. To overcome this disadvantage of CoAP, a version of CoAP with a better congestion control mechanism has been developed, one of them is CoCoA+ which has also been standardized by IETF CoRE WG TABLE III. CoCoA+ proposes a novel round trip time via considering the current network dynamics along with a variable backoff factor and aging mechanism. In large networks, CoCoA+ works better while in small networks they both work almost the same [2]. There are also some problems with CoCoA+ as it uses the measured RTT (Round Trip Time), which cannot be measured accurately for retransmitted messages, hence it may lead to incorrect output for RTO values in CoCoA+ [3].

Another problem with CoCoA+ is that very large changes in RTO estimates along with the default Stop-and-Wait behavior for the retransmissions in CoAP can violate the maximum delay requirements in the end to end IoT communications.

PROBLEM STATEMENT

CoAP nodes generally have an 8-bit microcontroller with very less amount of RAM and ROM 6 [4]. Congestion control methods that save previous states are not affordable for CoAP nodes. Hence, a method that is more effective and uses less memory is proposed in this paper.

The following structure has been followed in the study of this paper: Section (II) provides information about work done in this field. Section (III) presents the methodology of our proposed approach. Simulation setup for comparison is given in Section (IV) and results in Section (V).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:45 UTC from IEEE Xplore. Restrictions apply.

II. RELATED WORK

In this section, we will discuss the congestion control methods in the CoAP protocol. Firstly, we will discuss the general categorization of congestion control mechanisms. Then, we will move onto the CoAP protocol and traditional congestion control implementation in it. At the end of this section, we will take up the current advancements in this field and its drawbacks.

There are many different stages in IoT, for which different congestion control mechanisms are defined. The two major causes of congestion are increased request traffic and retransmissions of packets in the network.

Currently from the past few years, as we are more connected to each other than ever via the network, the data transmission is also elevated. This elevation affects the normal user-generated traffic by burdening it, due to which packet loss occurs more frequently. In these scenarios, we can use offloading methods to control congestion. Offloading is a technique to share loads of congested nodes with non-congested nodes so that the retransmissions are reduced and packets are reached on time. Here, the distribution of load from the congested node within the network takes place for the stable transmission of packets and scale down the delay in the network.

In IoT scenarios, where we have low power and lossy wireless networks, retransmission is a trivial routine. The major work in this field is done at the protocol level i.e. application layer and network layer. In this work, we modify and update the CoAP protocol to control the congestion at the level of the application layer. In protocol-based methods, we focus on the values of RTT and RTO to reduce the congestion and control the network traffic.

CoAP has a similar interaction as the client-server design of HTTP. The only difference between CoAP and HTTP is that, unlike HTTP, CoAP client-server interactions are non-synchronous and work on the top of the User Datagram Protocol (UDP) [4]. CoAP has its fundamental working between the application and the transport layer. The Hypothetical layer of CoAP consists of two reasonable sub-layers: request-response and messaging [2]. The message and response code are provided by the request-response layer to handle contrasting categories of messages. The non-synchronous nature of UDP is handled by the messaging layer. CoAP protocol has tags namely Confirmable(CON) and Non-Confirmable(NON) about messages which add reliability to this UDP based protocol [6]. CON-messages require ACK from the receiver, whereas, NON-messages don't require such confirmation.

The base CoAP controls congestion by restraining the number of parallel messages (parallel CON messages) and the rate of outgoing messages [2]. Firstly, applying a limit of one to active interactions per destination. Any request, CON or NON, waiting for an ACK or reply can be considered as active interaction. Before considering a failure, a CON message can be retransmitted up to four times if no ACK is received. The initial RTO value is randomly chosen from the range [2,3] s. And for subsequent retransmission, BEB (binary exponential backoff) is applied i.e. doubling the RTO value [2]. To make it more responsive and adaptive to the RTT information, a better congestion control mechanism is presented by August B. et al. i.e. CoCoA/CoCoA+ in [2] and [6].

CoCoA or CoCoA+ is a simple upgrade of traditional CoAP congestion control. CoCoA+ relieves the limits on messages (parallel and outgoing). CoCoA+ uses the RTT information which is available while calculating new RTO i.e adaptive RTO calculation. And instead of BEB, CoCoA+ uses VBF (Variable Backoff Factor). Also, if the RTO is not updated for a long time, it applies the RTO aging mechanism. Two separate estimators are provided by CoCoA+: (i) the strong RTO estimator i.e. estimated from the RTT information of packets that are successfully transmitted in the first transmission. (ii) the weak RTO estimator i.e. using RTT information of the packets that required retransmission (at most two) [2][6].

For calculation of RTO, according to RFC 6298 [7], two variables are maintained in CoAP source, SRTT for smoothed RTT and RTTVAR for RTT variation. Then, particular RTO (strong or weak) can be calculated as follows:

$$RTO_x = SRTT_x + K_x \times RTTVAR \quad (1)$$

Where x can be strong or weak, and as mentioned in RFC 6298 $K_s = 4$ and $K_w=1$. The lower K value for a weak RTO estimator minimizes the impact of high $RTTVAR_w$ in the case of large variations in RTT values of packets with retransmissions. At last, the overall RTO is calculated as follows:

$$RTO = \alpha_x \times RTO_x + (1 - \alpha_x) \times RTO \quad (2)$$

With $\alpha_s = 0.5$ and $\alpha_w = 0.25$. Low αw scales down the contribution of a weak RTO estimator.

Here, in CoCoA+ [6], the BEB is replaced by VBF for the calculation of RTO values of retransmitted packets. The VBF calculation is as follows:

$$VBF(RTO_{init}) = \begin{cases} 3 & RTO_{init} < 1s, \\ 2 & 1s \le RTO_{init} \le 3s, \\ 1.3 & RTO_{init} > 3s \end{cases} \quad (3)$$

VBF for large RTO is applied such that large idle time is avoided and for small RTO to avoid spurious retransmissions.

Also, the RTO aging mechanism is introduced in CoCoA+, which is used to update RTO if there is no new RTT information available for a long time i.e. RTO value becomes out-of-date. If an RTO estimation is small or large (below 1 s or above 3 s, respectively), and no new RTT measurement is made for 16 or 4 times the base RTO, respectively, the RTO value is modified to approach the default initial value [2][6]. Hence, RTO aging impacts the congestion control via reinitiating the RTO value to its default.

III. METHODOLOGY

The theory of natural evolution inspires genetic algorithms. Genetic algorithms depend on the fittest individuals in the current generation and the repetitive crossover and mutation [8][9]. In this paper, [10] CAIA delay-gradient (CDG) is used in the form of the Genetic algorithm and some modifications

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:45 UTC from IEEE Xplore. Restrictions apply.

for the UDP network are also made. CDG is used in TCP congestion control mechanisms, hence it needs modification for the UDP network.

RTT is a noisy signal for congestion control. So, a cleaner approach towards congestion detection involves RTT_{min} and RTT_{max} in a definite time interval. Based on RTT_{min} and RTT_{max} gradient of these two are maintained [10] where n is the n^{th} time interval x can be max or min:

$$g_{x,n} = RTT_{x,n} - RTT_{x,n-1} \quad (4)$$

After that moving average is calculated as follow:

$$\bar{g}_n = \sum_{i=n-a}^{n} \frac{g_i}{a} \quad (5)$$

where $g_i = g_{min,i}$ for calculating avg. for min or $g_i = g_{max,i}$ when calculating avg. for max.

In this paper, RTT min is tracked for 5 seconds and then in the process of creating fittest individuals in the Genetic Algorithm, repetitive crossovers are made every 5 seconds. After the RTT min observation, the crossover is done over RTO after measuring the RTTmin. Crossover is done concerning the previous and new RTTmin and then the probability of backoff (6) is calculated according to CDG.

Backoff probability varies with the difference in RTTmin with every iteration. Probability reaches 0 when there is no change in RTT_{min} and tries to reach 1 for higher values of change. This probability then decides the change in RTO after comparing it via a uniform random value Y.

$$g = RTT_{min}(curr) - RTT_{min}(curr - 1) \quad (6)$$

$$P[backoff] = 1 - \exp(-(g/G)) \quad (7)$$

RTT min is used as compared to RTT max because RTT max does not always give correct results while observing congestion. RTT max may also become too large in value because of other network conditions. IoT nodes do not have large memory spaces, this approach also reduces the memory usage as it does not store previous states of the network. TCP protocol maintains a window for packets the size of which is modified to control congestion. COAP is UDP based and hence cannot maintain a window. So, modifications in the form of RTO calculation are made by using PBF which is calculated as follows:

$$PBF = \begin{cases} 1.428, & g > 0 \text{ and } P[backoff] > Y, \\ 0.700, & \text{else}, \end{cases} \quad (8)$$

Where Y is a uniform random value.

The difference in RTTmin current and previous is also used as a condition to ensure that RTO is increased only when there is congestion. Hence this condition is typically used as a congestion notifier in the network. RTO is multiplied by PBF value of 1.42 when congestion in the network is detected, otherwise, it is reduced by a factor of 0.7.

Fig. 1 shows you the flowchart, where, the 5-second fixed interval is the observation duration for the RTTs. After the observations, minimum RTT is selected from all observed RTTs. After that difference between current and previous

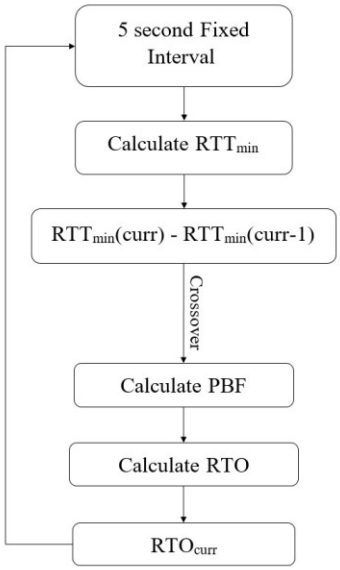


Fig. 1 Flowchart

RTT_{min} is calculated and this difference is used to calculate PBF after which RTO is calculated and this process continues.

IV. SIMULATION SETUP

This section will give a brief about how we evaluated our proposed method and how the simulation is set up.

Contiki OS [11], the open-source operating system for the Internet of Things, is perfectly suitable for any constrained application. It also provides all the tools and protocols we can use to evaluate our method. Here, we merged our modification in the base implementation of CoAP in Contiki OS by using the code of CoCoA/CoCoA+ openly provided by A. Betzler et al. [12]. Cooja [13], a tool provided by Contiki OS for sensor network simulation, is used for emulation of the wireless motes and simulation of the constrained application on a wireless sensor network. Cooja's main aspect is providing emulation for off-the-rack hardware of wireless sensor nodes that take all the hardware specifications and processing potential of actual nodes into consideration.

Here, we used the Z1 mote [14] type from all the mote types provided by Cooja. Specification of Z1 mote is provided in TABLE I. . Z1 motes are used for client and server nodes, as they provide more ROM for implementing and coding our method of congestion control in CoAP. As they also provide good RAM capacity, they are also used for the configuration of the border router. RPL border router [15], IPv6 Routing Protocol for Low-Power and Lossy Networks, is used as a border router in this setup to connect to the external network.

TABLE I. SPECIFICATIONS OF THE ZOLERTIA Z1 MOTE.

MCU	MSP430F2617
ROM	92 Kb
RAM	8 Kb
Radio Interface	CC2420

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:45 UTC from IEEE Xplore. Restrictions apply.

TABLE II. PARAMETERS USED IN SIMULATION

Parameters	Values
Operating System	Contiki-3.1
Network Simulator	Cooja
Radio Duty Cycling	NullRDC
Radio Medium	Unit Disk Graph Medium with Distance Loss (UDGM)
Mote Type	Zolertia Z1
Number of Nodes	Grid-36 and Dumbbell - 21
Interference range of nodes	20m
Transmission range of nodes	10m
Simulation Duration	30 minutes
Retransmission Limit	4

In this setup, the client node rhythmically sends the messages which are primarily targeted at the server node. Here, we performed the process with a random seed value and the results are plotted against the simulation clock.

The simulation parameters with respective values used for our simulation setup are given in the following table II. The parameter values are similar to those used in the evaluation of CoCoA+ by A. Betzler et. al. [6]. The radio medium parameter shows that the propagation loss model with distance loss is adopted in the simulation to make it close to the real scenarios. The sleep and wakeup cycle is handled by RDC (Radio Duty Cycling). The RDC method used in mote configuration operates these cycles. Here, NullRDC is used, where nodes never sleep and always listen for the messages i.e. no Radio Duty Cycling.

The setup can be evaluated in two types of scenarios (i) Static Topologies, where, the position of nodes is locked during the simulation and the nodes possess zero mobility, (ii) Mobile Topologies, where the nodes are mobile and have some velocity and move while the simulation is running. Here, we only worked on Static topologies. Mainly, we worked on two static topologies.

1) Grid Topology – These are popularly used in applications that require long-range and immense area coverage [6]. They are modeled in such a way that, there is always a node in the range of transmission of every node. To separate the network range from the transmission range of each node, the data packets carried over various nodes before delivering to the server. Here, the nodes with 4 or more neighbors are most likely to have congestion. The 6x6 grid we used consists of a border router (green) in the center, 2 sink nodes (orange) at 2 corners, and clients (pink) with a distance of 10m between each other. Shown in Fig 2.

2) Dumbbell Topology – This one is popular in evaluating the congestion algorithm in networks [6]. In this topology, the middle node, which connects the two halves of the dumbbell, acts as a link. This link needs to be shared generously to entertain all client transactions when the bandwidth requirements increase. The congestion control method used provides this sharing. Here, this topology contains 18 client-nodes (pink) and 2 server nodes (orange) and a border router(green) in the center, all together forming a dumbbell shape. The formation can be seen in Fig. 3

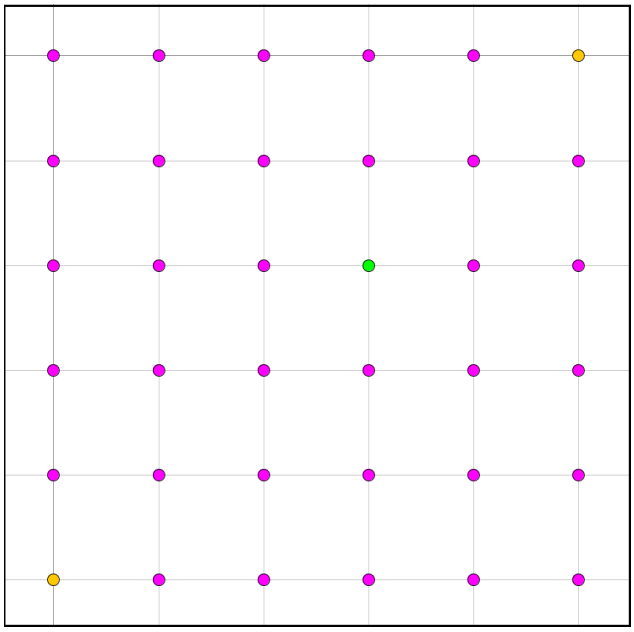


Fig. 2 Grid Network Topology

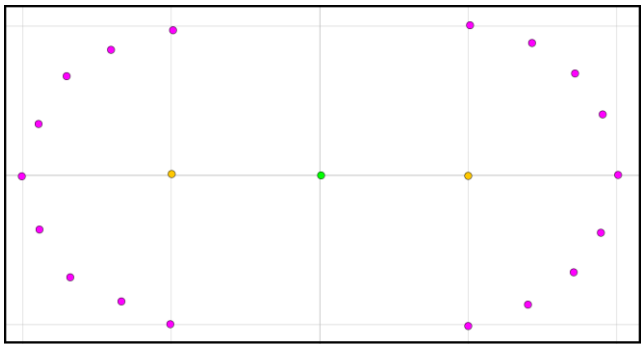


Fig. 3 Dumbell Network Topology

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:45 UTC from IEEE Xplore. Restrictions apply.

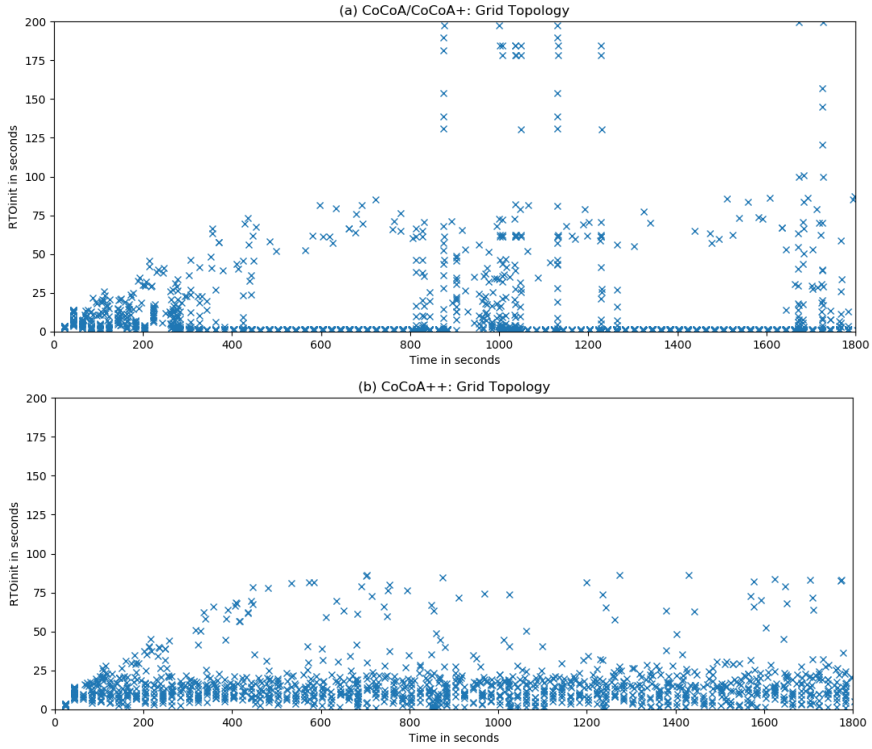


Fig. 4 Comparing (a) CoCoA/CoCoA+ & (b) Genetic CoCoA++ using Grid Topology

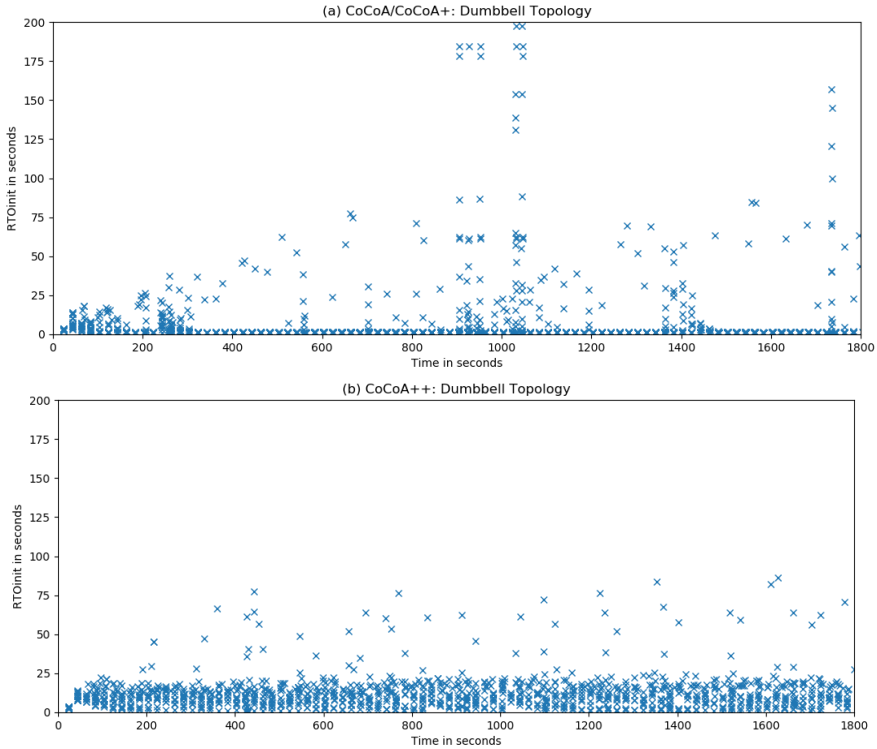


Fig. 5 Comparing (a) CoCoA/CoCoA+ & (b) Genetic CoCoA++ using Dumbbell Topology

V. RESULT

This section will be about the metrics and the comparisons of our approach with the CoCoA/CoCoA+. The network is simulated with the same parameters in both scenarios. We compared the shift in RTO value calculated by both techniques concerning the simulation time.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:45 UTC from IEEE Xplore. Restrictions apply.

TABLE III. RESULTS

Topologies	CoCoA /CoCoA+			Genetic CoCoA++		
	Max RTO(s)	Avg. RTO (s)	PFR	Max RTO (s)	Avg. RTO(s)	PFR
Dumbbell	505.39	28.38	0.0674	219.17	12.07	0.0333
Grid	507.41	46.97	0.1095	246.17	15.78	0.0945

Each simulation is run with a random seed value and the simulation clock is plotted over the x-axis along with the RTO values over the y-axis. The graphs in Fig. 4 & Fig. 5 show the RTO value fluctuation with the simulation clock in both the scenarios i.e. CoCoA/CoCoA+ & Genetic CoCoA++. Each dot in a graph shows the RTO value initiated at the time represented in the simulation clock axis. Also, we have observed the number of failed packets (timed out due to wrong RTO estimate), which then can be used to calculate PFR (Packet Failure Rate) i.e. the ratio of the total number of failed packets to the total number of packets generated in the complete simulation. TABLE III. contains the results obtained from the simulation. It contains the values of average and maximum RTO generated throughout the simulation clock within one experiment. Also, PFR is mentioned in this table. These results are shown comparing both the approaches in both topologies.

CONCLUSION AND FUTURE WORK

This work has introduced a congestion control scheme known as Genetic CoCoA++ that utilizes a CDG mechanism along with the genetic algorithm which figures out and tries to remove congestion in wireless sensor networks based on CoAP. We proved that congestion control mechanisms which are based on RTT_{max} may conclude congestion in the network from the noisy RTT sample. As delay gradient along with the genetic algorithm and probabilistic backoff factor are used here. After every iteration, better RTO is proposed to minimize congestion in the network.

The evaluation of the proposed scheme is done on static topologies. The results infer that Genetic CoCoA++ generates lower RTOs, therefore, it reduces the PFR also, hence improving the transmission rate.

In further researches, work can be done by altering the time interval for the observation period and the number of RTTs used in an iteration. Here, RTT_{min} and RTT_{max} have been used. RTT_{median} can be used. Feedback based Neural networks like recurrent neural network can also be trained over RTO or RTT. Also, we have not evaluated this method of congestion control in mobile scenarios. One can further compare this approach with mobile topologies like the Gauss–Markov mobility model, Manhattan grid mobility model, random waypoint mobility model (RWMM), etc.

REFERENCES

[1] C. Bormann, A. Betzler, G. Gomez, and I. Demrkol, "CoAP Simple Congestion Control/Advanced," Internet Draft (expires: April, 2017), October 2016. [Online]. Available: <https://www.ietf.org/id/Draft-ietf-core-cocoa-00.txt>.

[2] A. Betzler, C. Gomez, I. Demirkol and J. Paradells, "CoAP Congestion Control for the Internet of Things," IEEE Communications Magazine, vol. 54, no. 7, pp. 154–160, July 2016.

[3] J. J. Lee, K. T. Kim, and H. Y. Youn, "Enhancement of congestion control of Constrained Application Protocol/Congestion Control/Advanced for Internet of Things environment," International Journal of Distributed Sensor Networks, vol. 12, no. 11, November 2016.

[4] Kumar, R. Praveen, and S. Smys. "Analysis of Dynamic Topology Wireless Sensor Networks for the Internet of Things (IOT)." International Journal of Innovations in Engineering and Technology (IJIET), vol. 8, issue no. 4, August 2017 <http://dx.doi.org/10.21172/ijiet.84.006>

[5] E. Ancillotti and R. Bruno, "Comparison of CoAP and CoCoA+ congestion control mechanisms for different IoT application scenarios," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 1186-1192.

[6] Betzler, August & Gomez, Carles & Demirkol, Ilker & Paradells, Josep. (2015). CoCoA+: An advanced congestion control mechanism for CoAP. Ad Hoc Networks. 10.1016/j.adhoc.2015.04.007.

[7] RFC 6298, Computing TCP's Retransmission Timer <https://tools.ietf.org/html/rfc6298>.

[8] D. E. Goldberg, Genetic Algorithms in Search Optimization and Machine Learning, Addison-Wesley, 1989.

[9] A. E. Eiben and J. E. Smith, Introduction to Evolutionary Computing, Natural Computing Series, Springer, Berlin, Germany, 2003.

[10] D.A. Hayes, G. Armitage, Revisiting TCP congestion control using delay gradients,in: International Conference on Research in Networking, Springer, 2011, pp. 328-341.

[11] A. Dunkels, B. Grönvall, T. Voigt, Contiki – a lightweight and flexible operating system for tiny networked sensors, 2004.

[12] Implementation of CoCoA/CoCoA+ code by A. Betzler, <https://github.com/abetzler/er-cocoa>.

[13] F. Osterlind, A. Dunkels, J. Eriksson, N. Finne, T. Voigt, Cross-level sensor network simulation with COOJA, in: Proceedings 2006 31st IEEE Conference on Local Computer Networks, 2006, pp. 641–648. <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4116633>.

[14] Zolertia, Z1 Low-power Wireless Sensor Network Module, March 2010. <http://www.zolertia.com/products/z1>.

[15] T. Winter, P. Thubert, J. Hui, P. Kelsey, P. Levis, K. Pister, R. Struik, J. Vasseur, R. Alexander, RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks, Tech. Rep. 6550, RFC Editor, Fremont, CA, USA, March 2012. <http://www.rfc-editor.org/rfc/rfc6550.txt>.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:38:45 UTC from IEEE Xplore. Restrictions apply.

Soft Computing
https://doi.org/10.1007/s00500-020-05080-7

METHODOLOGIES AND APPLICATION

Global-best optimization of ANN trained by PSO using the non-extensive cross-entropy with Gaussian gain

Seba Susan[1] · Rohit Ranjan[1] · Udyant Taluja[1] · Shivang Rai[1] · Pranav Agarwal[1]

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
An efficient optimization of network weights has been the primary goal of the artificial neural network (ANN) research community since decades. The aim of every optimization problem is to minimize the network cost which is some form of error function between the desired and the actual network outputs, during the training phase. The conventional gradient-based optimization algorithms like backpropagation are likely to get trapped in local minima and are sensitive to choices of initial weights. The evolutionary algorithms have proved their usefulness in introducing randomness into the optimization procedure, since they work on a global search strategy and induce a globally minimum solution for the network weights. In this paper, we particularly focus on ANN trained by Particle Swarm Optimization (ANN-PSO), in which the local-best and global-best particle positions represent possible solutions to the set of network weights. The global-best position of the swarm, which corresponds to the minimum cost function over time, is determined in our work by minimizing a new non-extensive cross-entropy error cost function. The non-extensive cross-entropy is derived from the non-extensive entropy with Gaussian gain that has proven to give minimum values for regular textures containing periodic information represented by uneven probability distributions. The new cross-entropy is defined, and its utility for optimizing the network weights to a globally minimum solution is analyzed in this paper. Extensive experimentation on two different versions: the baseline ANN-PSO and one of its most recent variants IOPSO-BPA, on benchmark datasets from the UCI repository, with comparisons to the state of the art, validates the efficacy of our method.

Keywords ANN-PSO · Artificial neural network · Particle swarm optimization · IOPSO-BPA · Non-extensive cross-entropy with Gaussian gain

1 Introduction

The artificial neural networks (ANN) is one of the most well-researched tool for machine learning in data mining and pattern recognition. Its biggest advantage is that it 'learns from examples' without resorting to reasoning, which is suited for real-world problems that are exceedingly complex and cannot be mathematically defined (Zhang 2000). An example in (Susan and Singh 2011) demonstrates utility of ANN for a handwriting recognition problem, in which the best feature subset for the box partitioning method is learnt through a feedforward-backpropagation multi-layer perceptron (MLP) neural network, which cannot be otherwise argued with by logic. This is the reason that advanced forms of ANN in the form of deep neural networks are currently in trend for learning complex information from images (Liu et al. 2017). The optimization of neural networks implies learning the best set of synaptic weights during the training phase for accurate classification in the testing phase. This involves the iterative minimization of a predetermined cost function which is most often a measure of the error between the desired and computed class labels in the training phase. The conventional backpropagation algorithm introduced by Rumelhart et al. (1985) is an example of the gradient-based optimization methods. The gradient of the error cost function is incorporated into the weight update equations, and the whole process is mathematically analyzed and solved. The problem inherent with all derivative-based

Communicated by V. Loia.

✉ Seba Susan
seba_406@yahoo.in

[1] Department of Information Technology, Delhi Technological University, Bawana Road, Delhi 110042, India

Published online: 10 June 2020

methods is prevalent here too, that the solution (set of network weights) gets trapped in local minima and is sensitive to the initialization of weights. The solution also takes time to converge given the number of iterations involved that increase significantly with the number of extra hidden layers added. To achieve global minima, the randomization of neural networks has been investigated in Zhang and Suganthan (2016). Randomization is used to modify data distributions, and fine-tune the parameters and network configurations. However, in certain applications like deep neural networks, randomization leads to even lower performance than the conventional backpropagation (Chen et al. 2013). Evolutionary algorithms and swarm intelligence, which are specialized types of randomization algorithms, are derivative-free optimization tools that strive to achieve global minima. Evolutionary processes in bio-systems are the motivation for these algorithms loosely based on the principle of the 'survival of the fittest', that introduce randomization at certain points such as the mutation and reproduction steps in Genetic Algorithm (GA) (Goldberg and Holland 1988).

The evolutionary and swarm intelligence algorithms were thus naturally investigated for optimizing the ANN weights to globally optimum solutions, examples in literature being the application of GA in (Montana and Davis 1989), Particle Swarm Optimization (PSO) (Kennedy and Eberhart 1995) in (Mendes et al. 2002), and Ant Colony Optimization (ACO) (Dorigo et al. 2006) in (Socha and Blum 2007). Most of these works follow the same pattern for determining the synaptic weights, with each particle emulating a neural network and the local minima solutions of the particles being cumulatively analyzed to achieve the globally minimum solution (Chen et al. 2016). The local and global minima are iteratively updated till convergence is achieved. The applications of such hybrid networks incorporating evolutionary algorithms are spread over a wide spectrum of engineering fields such as energy (Zameer et al. 2017), signal processing and control (Raja et al. 2018), bioinformatics (Raja et al. 2017), and electromagnetic theory (Khan et al. 2015). Of all the bio-inspired algorithms, PSO that is based on stochastic processes is the least complex of all, with only the particle position and velocity being updated in each iteration and no complicated procedures of mutation and cross-over involved, hence the popularity of the Artificial Neural Networks trained by Particle Swarm Optimization (ANN-PSO) (Kiranyaz et al. 2009; McCaffrey 2013). In our work, we propose the computation of the global minima in ANN-PSO through non-extensive information-theoretic measures. We interpret this as being more meaningful, and validate our claim through rigorous experimentation. A new non-extensive cross-entropy function is defined for this purpose. Two different ANN training algorithms, ANN-PSO (Mendes et al. 2002) and IOPSO-BPA (Yaghini et al. 2013), are chosen for testing the new concept. The organization of the paper is as follows. The review of ANN-PSO and its variants is provided in Sect. 2. The new non-extensive cross-entropy is defined and analyzed in Sect. 3, and its application for the global-best optimization of ANN-PSO is described in Sect. 4. The results of the experimentation are described in Sect. 5 and the overall conclusions are drawn in Sect. 6.

2 ANN trained by particle swarm optimization (ANN-PSO): a review

We describe here in brief, the ANN-PSO algorithm (Mendes et al. 2002) and IOPSO-BPA (Yaghini et al. 2013) that are examples of ANN trained by Particle Swarm Optimization (PSO). PSO is a swarm intelligence algorithm introduced by Kennedy and Eberhart (1995), which mimics the movement of a flock of birds or a school of fish. The basic strategy in a flock of birds, while foraging for food, is to follow the bird that is nearest to the food. This is the principle followed in PSO, where each particle is searching for the optimum value of the cost function and has its own particle position and velocity at a given time. The particle remembers its personal-best optimum solution obtained so far (called lbest or local-best solution). Also, the best optimum solution obtained so far by any particle in the population is also iteratively recorded (called gbest or global-best solution). Both lbest and gbest play a crucial role in determining the next move made by each particle. The iterations continue until convergence is achieved for the gbest solution, or if the maximum number of allowed iterations is reached. The overall procedure is visually depicted in the form of a flowchart in Fig. 1. We next explain the process of network weight optimization of ANN by the PSO algorithm (ANN-PSO), as described by Mendes et al. (2002).

2.1 ANN-PSO

The training of artificial neural networks (ANN) using PSO for weight optimization was proposed by Mendes et al. (2002). The aim was to determine the optimal values of the network weights given the cost function to be minimized which is assumed as the fitness function of the PSO. The set of network weights denote the particle position in a particular iteration. Thus, the particle position, velocity, lbest, and gbest are all vectored quantities. The cost function is usually a measure of the error difference between the desired and actual neural network outputs, with the mean-square error and cross-entropy error being popular choices of the cost function. Each particle in the swarm

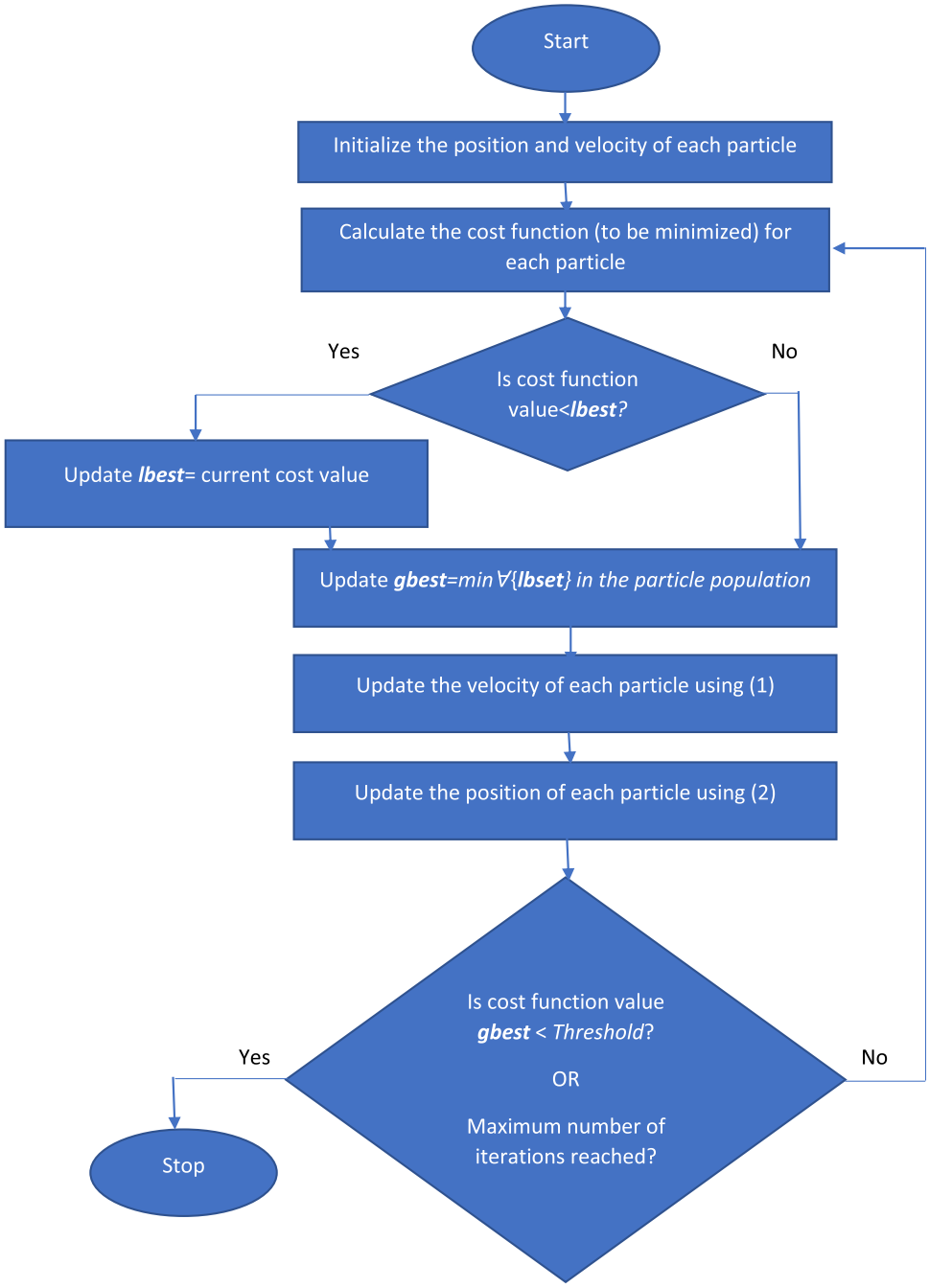


Fig. 1 Flowchart for Particle Swarm Optimization (PSO)

strives to minimize this fitness function, and moves its particle velocity and position in a direction to achieve this goal. There are two different goals of interest for each particle: its own personal-best position called lbest and the overall best position of the entire swarm population called the global-best or gbest. The velocity vector $\boldsymbol{v}$ and the position vector $\boldsymbol{x}$ of the particle are updated in each iteration as per (1) and (2), along with the iteratively improving values of lbest and gbest. The update equations for the vectored inputs is expressed as:

$$\boldsymbol{v}_{\text{new}} = w \times \boldsymbol{v}_{\text{old}} + c_1 \times r_1 \times (\text{lbest} - \boldsymbol{x}_{\text{old}}) + c_2 \times r_2 \times (\text{gbest} - \boldsymbol{x}_{\text{old}}) \quad (1)$$

$$\boldsymbol{x}_{\text{new}} = \boldsymbol{x}_{\text{old}} + \boldsymbol{v}_{\text{new}}. \quad (2)$$

In (1), w is a constant called the inertia weight, c_1 and c_2 are the cognitive and social coefficients respectively, and r_1 and r_2 are random numbers in the range [0,1).

The entire optimization process for the neural network is compiled in the form of a pseudo-code in Fig. 2. When the iterations terminate or the algorithm converges, the global-best position at that point indicates the optimum network

Fig. 2 Pseudo-code for ANN-PSO

```
For each particle [Neural Network]
  Initialize particle position [Neural network weight vector] and velocity
End

Do
  For each particle
    Calculate cost function [Cross-Entropy] value
    If value of the cost function < lbest
      set current cost function value as the new lbest
  End

  Choose the particle with the minimum cost function value among all particles as the gbest

  For each particle
    Calculate particle velocity according equation (1)
    Update particle position according equation (2)
  End

Until maximum iterations have been achieved or minimum error criteria is not attained
```

weight configuration. Some variants of ANN-PSO are discussed below with focus on a recent variant IOPSO-BPA that we implement using our approach for determining the global-best position.

2.2 Variants of ANN-PSO

A study of recent literature throws light on the rapid emergence of hybrid ANN-PSO as a quick and efficient learning tool for a variety of applications, such as predicting the shear strength of buildings (Chen et al. 2018), islanding detection in distribution systems (Raza et al. 2016), and electricity price forecasting (Anand and Suganthi 2017). Customized hybrid ANN-PSO models have also been introduced such as in (Gharghan et al. 2016) to improve the distance estimation efficiency between the bicycle and the coach in cycling events. A recent work by the authors (Susan et al. 2019), on ANN training stopping criterion, optimized ANN-PSO weights by monitoring the network weight entropy computed using the chain rule of entropy. The point of minimum weight entropy, as the iterations progressed, indicated the most optimal weight configuration. Several variants of PSO exist in literature, but only a few have been tried for training neural networks. Of all the variants of PSO introduced so far, opposition-based learning (Tizhoosh 2005) has consistently reported high performance. The particles demonstrating the worst performance are replaced by their opposite particles. The application of opposition-based learning to ANN training has been investigated in depth by Yaghini et al. (2011) and (2013). We focus on one recent algorithm on Opposition-based learning by Yaghini called IOPSO-BPA (Yaghini et al. 2013). The flowchart for IOPSO-BPA is shown in Fig. 3. The main difference in the procedure from ANN-PSO is in the opposite particle computation for each particle in the swarm that shows bad performance, and the selection of the fittest among the two. Also, if the position of a particle shows no improvement, then random perturbation is introduced with fine-tuning by backpropagation (BPA). In addition, the cognitive and social coefficients c_1, c_2 in (1) and (2) are considered not as constants, but are computed from time-varying equations (for a total of m iterations) as shown below:

$$c_1(t+1) = \frac{t}{m} \times (c_1(m) - c_1(t)) + c_1(t) \tag{3}$$

$$c_2(t+1) = \frac{t}{m} \times (c_2(m) - c_2(t)) + c_2(t). \tag{4}$$

The pseudo-code for IOPSO-BPA explains the process in more detail. For detailed mathematical analysis, the readers are referred to (Yaghini et al. 2013).

3 The proposed non-extensive cross-entropy for global-best optimization

The direction in which the particles move after each iteration, and how fast they move, directly impact the optimality of the solution obtained in ANN-PSO. Convergence to the globally minimum solution of the cost function is the aim of every optimization technique. The cost function or fitness function to be minimized for PSO is a measure of the difference between the target output t_i and the computed network output y_i, for $i = 1,2,\ldots,n$, where n is the number of output neurons. The output $\{y_i\}$ is softmaxed between 0 and 1, such that the sum of all output neurons is 1.

The cross-entropy is the most popular cost function for neural networks (Shore and Johnson 1981; McCaffrey

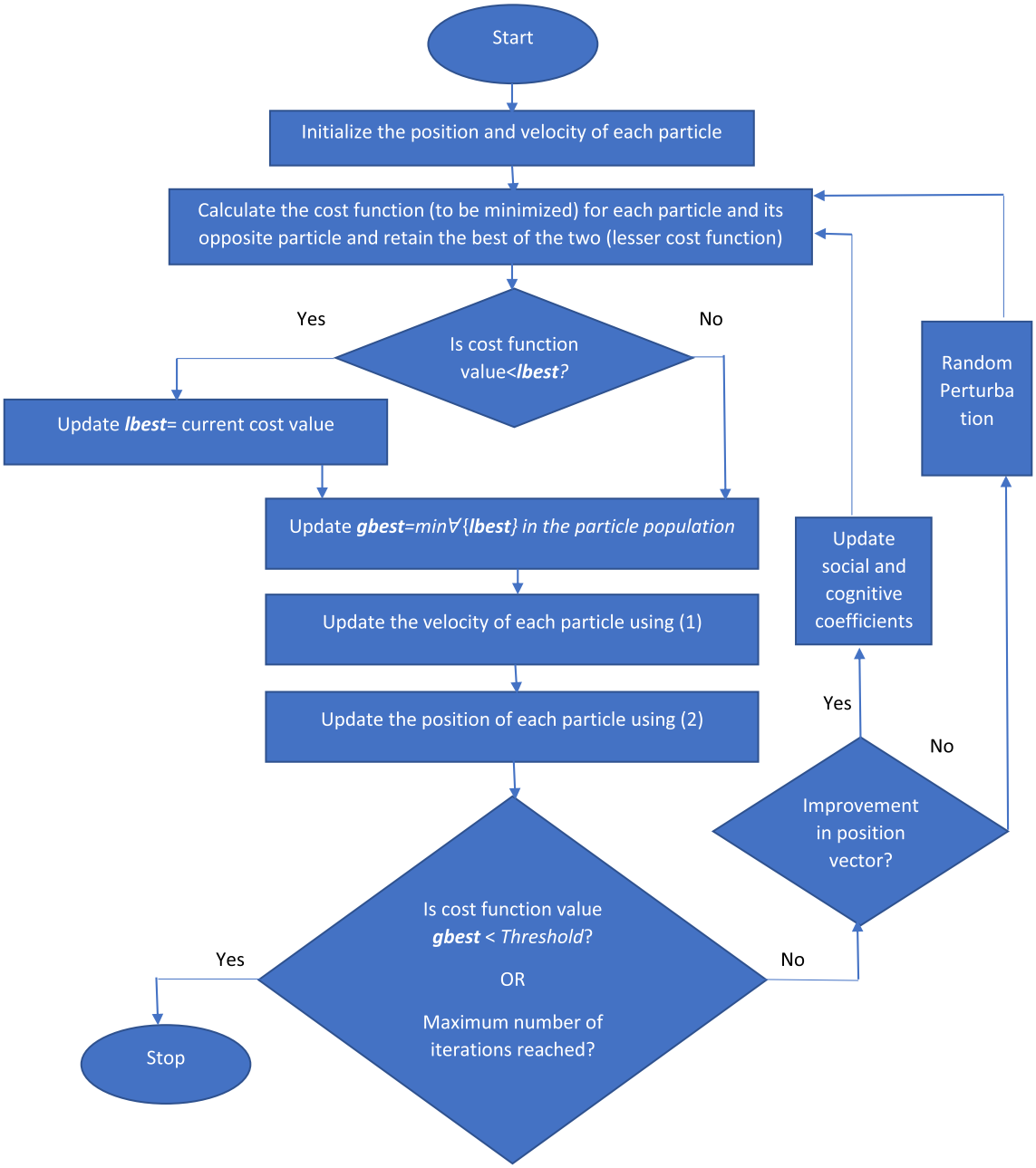


Fig. 3 Flowchart for the most recent variant of ANN-PSO-improved opposition-based Particle Swarm Optimization-Backpropagation Algorithm (IOPSO-BPA) where each particle is a neural network

2013). It was first referred to by Shannon as the relative entropy in (Shannon 1948) and interpreted as the Kullback–Leibler divergence between two probability distributions in (Kullback and Leibler 1951). It is defined by the formula:

$$H_{\text{cross}} = \sum_{i=1}^{n} t_i \log \frac{t_i}{y_i}. \tag{5}$$

Here, both $\{t_i\}$ and $\{y_i\}$ are complete probability distributions, such that $\sum_{i=1}^{n} t_i = 1$ and $\sum_{i=1}^{n} y_i = 1$.

Equation (5) can be rewritten as:

$$H_{\text{cross}} = \sum_{i=1}^{n} t_i \log t_i - \sum_{i=1}^{n} t_i \log y_i. \tag{6}$$

Since the target value $\{t_i\}$ assumes values of either 1 or 0, the first term in (6) is always zero. Thus, (6) reduces to:

$$H_{\text{cross}} = -\sum_{i=1}^{n} t_i \log y_i. \tag{7}$$

Equation (7) is the cross-entropy formula popularly used as the cost function in ANN as well as for ANN-PSO implementations. Cross-entropy minimization is advised as an apt solution for learning, as early as in (Shore and Johnson 1981), to predict an unknown distribution when a prior estimate and new information is available. In the previous work by the authors (Susan et al. 2019), the cross-entropy error for feedforward-backpropagation multi-layer perceptron network showed much higher performance as compared to the mean-square error (quadratic error) cost function. The cross-entropy is popular as a cost function, since it avoids the learning slowdown problem prevalent with the quadratic error cost function (Nielsen 2015). This involves very fast learning when the output neuron produces a huge error, and the learning slows down as the target output is reached. The process emulates the human tendency to learn quickly from mistakes, and compensates even a wrong choice of hyperparameters or initial weights, since the neuron quickly recovers from its wrong decision. The cross-entropy cost function in (5) and (7) is considered in (Dunne and Campbell 1997; Brescia et al. 2008) to be more apt for the multi-class neural networks than other forms of error functions.

Minimization of the entropy function in (7) is achieved when the individual terms in (7) approach either $0 \log 0$ or $1 \log 1$. Since $0 \log 0$ which indicates a 'match' condition, is not clearly distinguished, especially with respect to $1 \log 0$ which implies a 'mismatch between the target and computed outputs', hence a need arises for an alternative entropy expression especially when searching for the global-best position where we are likely to encounter situations of $0 \log 0$ or $1 \log 1$ (when actual output approaches target value). At this point, uneven probability distributions are encountered for $\{y_i\}$, and we need a suitable measure for representing the uneven probability distributions. One of the relevant measures introduced recently that interprets uneven distributions in periodic or regular data is the non-extensive entropy with Gaussian gain proposed by Susan and Hanmandlu (2013). The non-additive nature of this exponential entropy makes it a good identifier of correlations in sparse probability distributions associated with regular texture images, in which a particular pattern is periodically repeated. The cross-entropy cost function for global-best optimization in ANN-PSO is derived in our work from this parent entropy in (Susan and Hanmandlu 2013), as shown below.

Consider a complete probability distribution $\{p_i\}, i = 1, 2, ..., n$, such that the sum of all the probabilities is one, i.e., $\sum_i p_i = 1$. The non-extensive entropy with Gaussian gain (Susan and Hanmandlu 2013), for this probability distribution is defined as:

$$H(P) = \sum_{i=1}^{n} p_i \mathrm{e}^{-p_i^2}. \tag{8}$$

The entropy in Eq. (8) identifies regular patterns well as observed from the experiments on gray-scale and color textures (Susan and Hanmandlu 2013, 2018). Its bell-shaped Gaussian information gain function, $\mathrm{e}^{-p_i^2}$, highlights the low probabilities lying in the bell of the non-linear Gaussian curve. Therefore, it can represent well the correlations that occur within the sparse probability matrices of texture images. It has also been put into use to find the point of minimum entropy accurately in several applications such as the point of alarm in a moving crowd video (Susan and Hanmandlu 2015), monitoring a dynamically growing hidden layer in (Susan and Dwivedi 2014), to determine the optimum window size for localizing texture defects in (Susan and Sharma 2017), and for regularization of the fuzzy c-means clustering algorithm (Susan et al. 2015). In all these cases, the uneven probability distribution was well translated to a minimum entropy value that helped in correct identification of the pattern involved.

The new non-extensive cross-entropy with Gaussian gain is defined from (8) on the lines of (7) as:

$$H_{\mathrm{NEXTcross}} = \sum_{i=1}^{n} t_i \mathrm{e}^{-y_i^2}. \tag{9}$$

The subscript NEXT in (9) is derived from the word Non-EXTensive. The measure in (9) satisfies the two main properties of a cost function:

P1: The cost function in (9) is non-negative, i.e., $H_{\mathrm{NEXTcross}} = \sum_{i=1}^{n} t_i e^{-y_i^2} > 0, \forall y_i \in [0,1], \forall t_i \in \{0,1\}$.

P2: As $y_i \to 1$ when the target $t_i = 1$, the cost function achieves a minimum value of e^{-1}.

The above two properties are easy to prove. Next, we analyze the reason why the cross-entropy defined in (9) is a better guarantee for searching for the global minimum in ANN-PSO and other related swarm intelligence algorithms, as compared to the logarithmic version in (7). The use of the exponential Gaussian gain in (9) instead of the logarithmic gain in (7) provides a clear improvement from the confusing and indistinguishable cases of $\underset{t_i=1, y_i \to \{0,1\}}{t_i \log y_i} \to 1 \log 0 = 0(\text{mismatch}); 1 \log 1 = 0(\text{match})$ for the logarithmic Shannon cross-entropy in (7), to the clearly distinguishable cases of $\underset{t_i=1, y_i \to \{0,1\}}{t_i \mathrm{e}^{-y_i^2}} \to 1\mathrm{e}^{-0^2} = 1(\text{mismatch}); 1\mathrm{e}^{-1^2} = 0.3679(\text{match})$ in case of the proposed non-extensive cross-entropy in (9).

When $t_i = 1$, for any one particular output neuron i, all other t_i's being zero, the expression in (9) reduces to a simple Gaussian function as shown below:

$$H_{\text{NEXTcross}} = \mathrm{e}^{-y_i^2}. \tag{10}$$

The plot of this Gaussian function $\mathrm{e}^{-y_i^2}$ versus the logarithmic version $-\log y_i$ is shown in Fig. 4. The observations from the two learning curves as the output progresses to its target (of 1) are analyzed next. Initially, when the output y_i is close to 0, very far from its target output of 1, the logarithmic term rapidly falls as compared to the Gaussian term that is somewhat flat at values very close to 1. This implies that the logarithmic term is a faster learner when the results are bad. Thus, the ANN-PSO needs the logarithmic term as its lbest measure for lifting itself from the erroneous local conditions in the initial iterations. However, as the learning progresses and as the output y_i approaches 1, the Gaussian curve that follows a monotonically decreasing curve slows down as the curve approaches its target and is good for fine-tuning the results. From Fig. 4, it is observed that the Gaussian curve is concave and the logarithmic curve is convex. In literature, several decision models interpret searching for the global minima as a concave minimization problem (Horst 1984). Thus, the swarm particles in ANN-PSO that are searching for the global minimum are better guided by the Gaussian curve rather than the logarithmic curve.

4 Proposed methodology

We next outline the sequential steps followed for the implementation of the baseline ANN-PSO and the variant IOPSO-BPA incorporating the new cross-entropy cost function for locating the global-best position in both algorithms. In doing so, we follow the basic guidelines in the flowcharts/pseudocodes in Figs. 1, 2, 3, 4, and 5 and incorporate changes wherever necessary.

4.1 ANN-PSO algorithm incorporating proposed scheme

Step 1 Specify starting position and velocity of particles according to initial values of network parameters. Set the iteration counter to zero (iter=0).

Step 2 For each particle in the swarm, compute the best personal position over time lbest using the cross-entropy in (7) as the cost function. In addition, compute the overall best global position gbest using the new non-extensive cross-entropy in (9) as the cost function.

Step 3 If the cost function E_{train}(gbest(iter)) $<$ threshold (pre-defined value of Exit Error), go to Step 8; otherwise, go to Step 4.

Step 4 Calculate the new velocity of particles as per Eq. (1).

Step 5 Calculate the new position of particles as per Eq. (2).

Step 6 Increase the iteration counter (iterCounter++).

Step 7 If iteration counter is greater than maximum allowed number (iter $>$ m), go to Step 8; otherwise, go to Step 2.

Step 8 Stop the training.

4.2 IOPSO-BPA incorporating proposed scheme

Step 1 Specify starting position and velocity of particles according to initial value of parameters. Set iteration counter to zero (iter=0), and set the counter that calculates number of no improvements in gbest to zero (gbestNoImprovement = 0).

Step 2 To establish a better population, and calculate opposite position and velocity of particles. For a particle with position $x(t)$ and velocity $y(t)$ in the tth iteration, the opposite particle position (Yaghini et al. 2013) is computed as:

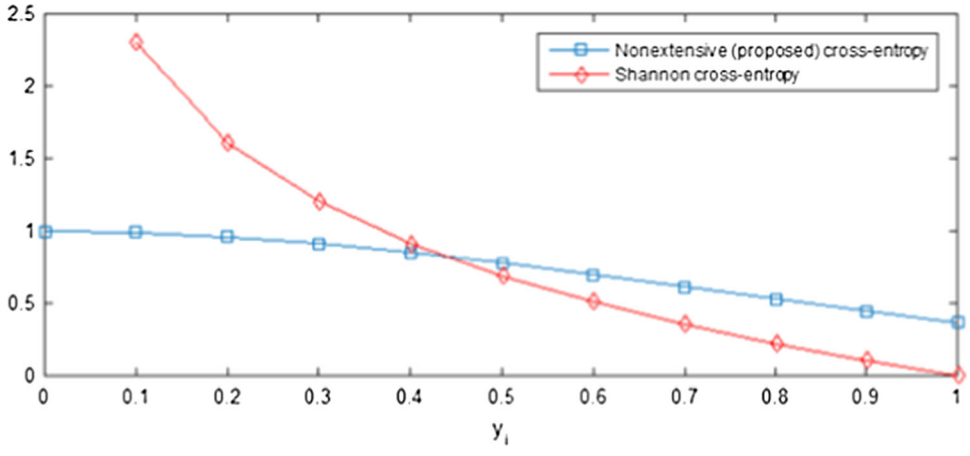


Fig. 4 Comparison of the information gains $-\log y_i$, and $\mathrm{e}^{-y_i^2}$ associated with the Shannon cross-entropy entropy and the proposed non-extensive cross-entropy, respectively, as y_i ranges from 0 to 1

Fig. 5 Pseudo-code for IOPSO-BPA

```
For each particle [Neural Network]
  Initialize particle position [Neural network weight vector] and velocity
END

Do
  For each particle
    Calculate cost function [Cross-Entropy] value
    Compute the opposite particle
    Calculate cost function [Cross-Entropy] value of the opposite particle
    Select either the particle or opposite particle for which the cost function is lesser
    New value of cost function=cost function of selected particle
    If value of the cost function < lbest
        set current cost function value as the new lbest
  End

  Choose the particle with the minimum cost function value among all particles as the gbest

  For each particle
    Calculate particle velocity according equation (1)
    Update particle position according equation (2)
    If no improvement in position vector, then introduce random perturbation followed by
      Backpropagation algorithm (BPA)
  End

Until maximum iterations have been achieved or minimum error criteria is not attained
```

$$\overline{x(t)} = x_{\max} + x_{\min} - x(t), \tag{11}$$

where $[x_{\max}, x_{\min}]$ is the initial particle position interval. Similarly, for a velocity interval of $[v_{\max}, v_{\min}]$, the opposite particle velocity (Yaghini et al. 2013) is computed as:

$$\overline{v(t)} = v_{\max} + v_{\min} - v(t). \tag{12}$$

Compute the cost function of the opposite particle using the cross-entropy in (7). If the cost function of opposite particle is lesser than that of the current particle, replace the current particle by the opposite particle and update the new particle position and velocity from (11) and (12), respectively.

Step 3 For each particle in the swarm, compute the best personal position over time lbest using the cross-entropy in (7) as the cost function. In addition, compute the overall best global position gbest using the new non-extensive cross-entropy in (9) as the cost function, and calculate the number of no improvements in gbest (gbestNoImprovement).

Step 4 If the number of no improvements in the best global position is greater than maximum allowed number (gbestNoImprovement >5), go to Step 5; otherwise, go to Step 6.

Step 5 Call random perturbation (Yaghini et al. 2013) for each particle in the swarm selected with a 0.5 probability. If the cost function of the global particle is now decreased, set the counter that calculates number of no improvement on gbest to zero (gbestNoImprovement = 0); otherwise, increment the counter by one (gbestNoImprovement + +).

Step 6 If the cost function E_{train}(gbest(iter)) < threshold (pre-defined value of Exit Error), go to Step 12; otherwise, go to Step 7.

Step 7 Update the cognitive and social coefficients: c_1-(iter+1), c_2(iter+1) as per Eqs. (3) and (4).

Step 8 Calculate the new velocity of particles as per Eq. (1).

Step 9 Calculate the new position of particles as per Eq. (2).

Step 10 Increase the iteration counter (iterCounter++).

Step 11 If iteration counter is greater than maximum allowed number (iter > m), go to Step 12; otherwise, go to Step 2.

Step 12 Stop the training.

For more details on IOPSO-BPA, the reader is referred to [16]. Random perturbation in Step 5 is an algorithm that resets the velocity of the particle to a new value (by adding a random number to each dimension of the current velocity vector) if the global-best position gbest shows no improvement over time.

5 Experimental results and discussions

The experiments are performed on a 2.4 GHz Intel core i7 PC with 8 GB RAM. The architecture of the neural network trained by PSO, for all our experiments, is fixed as: number of hidden layers = 1; number of hidden neurons = 6; maximum epoch/iteration $m = 700$; number of particles = 20; cognitive and social coefficients c_1 and $c_2 = \{1.49445, 1.49445\}$, inertia constant $w = 0.729$. For ANN trained by IOPSO-BPA, initialization of $c_1(0)$ and $c_2(0) = \{0.98, 0{,}02\}$. The set of above parameters is just an instantiation and these parameters can be varied as per choice to suit the machine learning application. For more instructions in this regard, the works of (Hajihassani et al. 2020; Kiranyaz et al. 2009; McCaffrey 2013) can be referenced. A trail-by-error approach is mostly followed for selecting the network hyperparameters though some works do address the issue by suggesting a systematic search (Nielsen 2015). Twelve datasets from the UCI repository (Frank and Asuncion 2010) are selected for our experiments, namely, iris, seeds, statlog heart, ionosphere, dermatology, beam balance, wine, glass identification, balance scale, breast cancer, Haberman's survival and teaching assistant. Alternate samples of each dataset are segregated for training and testing. The details of these datasets—the numbers of their samples, attributes, and class labels—are compiled in Table 1. Both the proposed ANN-PSO and IOPSO-BPA algorithms with the new non-extensive cost function measure embedded in them are implemented by C# software programs, as per the steps outlined in Sect. 4. The results are compared with the baseline methods ANN-PSO (Mendes et al. 2002) and IOPSO-BPA (Yaghini et al.

Table 1 Details of the 12 datasets from UCI repository (Frank and Asuncion 2010) used for the experimentation

Datasets	Number of instances	Number of features	Number of classes
Iris	150	4	3
Seeds	210	7	3
Statlog heart	270	13	2
Ionosphere	85	34	2
Dermatology	122	34	6
Beam balance	114	4	3
Wine	178	13	3
Glass identification	214	9	6
Balance scale	625	4	3
Breast cancer	138	10	2
Haberman's survival	152	3	2
Teaching assistant	152	5	3

Table 2 Validation percentage accuracy (%) of classification of all methods for the 12 datasets

Dataset	ANN-PSO (Mendes et al. 2002)	IOPSO-BPA (original cost function) (Yaghini et al. 2013)	IOPSO-BPA (with cross entropy)	Proposed ANN-PSO	Proposed IOPSO-BPA (with cross entropy)
Iris	94.67	16.00	89.33	**97.33**	**97.33**
Wine	66.67	72.22	**80.56**	66.67	61.11
Breast cancer	59.42	59.42	**71.01**	59.42	59.42
Ionosphere	69.00	**76.92**	52.46	74.35	**76.92**
Seeds	65.56	33.33	64.44	**86.67**	57.78
Balance scale	83.87	61.29	76.77	**89.68**	85.16
Statlog heart	55.56	54.07	72.59	68.89	**74.81**
Glass identification	36.54	50.00	42.31	**51.92**	50.00
Beam balance	**98.25**	64.90	94.74	**98.25**	89.47
Dermatology	29.5	65.57	52.46	**77.05**	62.30
Haberman's survival	53.95	51.32	**57.89**	51.32	51.32
Teaching assistant	36.84	31.58	36.84	42.11	**44.74**

(Best values for each dataset highlighted in bold)

Table 3 Cross-validation percentage accuracy (%) of classification of all methods for the 12 datasets

Dataset	ANN-PSO (Mendes et al. 2002)	IOPSO-BPA (original cost function) (Yaghini et al. 2013)	IOPSO-BPA (with cross entropy)	Proposed ANN-PSO	Proposed IOPSO-BPA (with cross entropy)
Iris	92.00	26.67	96.00	96.00	**96.33**
Wine	65.49	68.31	**69.01**	65.49	65.49
Breast cancer	59.42	59.42	59.42	59.42	**69.71**
Ionosphere	66.67	60.00	72.00	**73.33**	**73.33**
Seeds	69.17	49.17	80.83	**85.00**	62.50
Balance scale	77.19	55.65	71.86	**86.99**	86.14
Statlog heart	60.74	59.26	80.00	68.15	**81.11**
Glass identification	35.19	**51.92**	46.30	39.51	47.30
Beam balance	89.47	87.72	**96.49**	91.23	84.21
Dermatology	27.87	52.46	52.46	**65.57**	54.01
Haberman's survival	**64.47**	52.32	**64.47**	56.58	51.32
Teaching assistant	36.84	33.58	**50.00**	42.11	44.74

(Best values for each dataset highlighted in bold)

Table 4 Validation execution time (in s) for all methods for the 12 datasets

Dataset	ANN-PSO (Mendes et al. 2002)	IOPSO-BPA (original cost function) (Yaghini et al. 2013)	IOPSO-BPA (with cross entropy)	Proposed ANN-PSO	Proposed IOPSO-BPA (with cross entropy)
Iris	**10.7**	104.7	113.2	11.2	122.2
Wine	26.6	2209.0	2133.3	**24.7**	2132.3
Breast cancer	**11.8**	525.4	559.3	**11.8**	532.0
Ionosphere	**14.0**	1416.0	1780.1	15.3	1546.1
Seeds	12.3	245.9	306.5	**11.6**	266.2
Balance scale	**59.9**	2438.1	3005.2	68.2	3212.3
Statlog heart	**14.4**	442.3	431.1	14.6	425.3
Glass identification	**31.5**	2480.2	2755.4	33.6	3014.2
Beam balance	11.5	128.3	89.11	**9.0**	178.1
Dermatology	16.2	3141.0	4051.4	**14.5**	2186.3
Haberman's survival	**10.90**	199.1	233.1	**10.9**	222.3
Teaching assistant	**11.29**	393.2	495.7	11.4	471.9

(Best values for each dataset highlighted in bold)

2013), to mark the improvement obtained in each case. The accuracies obtained for all datasets are summarized in Tables 2 and 3 for validation and cross-validation, respectively. Cross-validation is achieved by simply swapping the training and testing subsets. The proposed global-best optimization using the new non-extensive cross-entropy significantly improves the results over that of the baseline algorithms, as observed from the higher classification accuracies in the last two columns of Tables 2 and 3. The proposed ANN-PSO scheme gives the highest accuracies for six datasets and the proposed IOPSO-BPA scheme gives the best results for four datasets in the validation experiments. In case of cross-validation, the proposed ANN-PSO scheme claims the best results for four datasets while the proposed IOPSO-BPA scheme gives the best results for four datasets. While ANN-PSO shows an all-round significant improvement in results with the new cost function definition, the IOPSO-BPA results are also found improved with an odd exception. Overall, cross-entropy error is noted to be an apt measure for the cost function as proved from the sharp increase in accuracies in

Table 5 Cross-validation execution time (in s) for all methods for the 12 datasets

Dataset	ANN-PSO (Mendes et al. 2002)	IOPSO-BPA (original cost function) (Yaghini et al. 2013)	IOPSO-BPA (with cross entropy)	Proposed ANN-PSO	Proposed IOPSO-BPA (with cross entropy)
Iris	**10.2**	101.5	112.6	11.2	122.2
Wine	26.6	1998.0	2231.0	**22.8**	2039.0
Breast cancer	**12.1**	524.4	593.9	12.2	560.1
Ionosphere	**14.0**	1416.0	1780.0	15.0	1546.0
Seeds	12.3	245.9	344.8	**10.7**	266.2
Balance scale	59.9	2248.0	2998.0	**59.8**	3104.0
Statlog heart	14.6	422.3	431.1	**12.0**	430.1
Glass identification	31.5	2389.0	2255.0	**28.6**	2967.0
Beam balance	11.5	128.0	89.0	**9.0**	178.0
Dermatology	16.0	3141.0	4051.0	**14.0**	2186.0
Haberman's survival	**10.1**	197.1	217.9	10.6	215.3
Teaching assistant	11.6	396.2	473.5	**11.5**	477.9

(Best values for each dataset highlighted in bold)

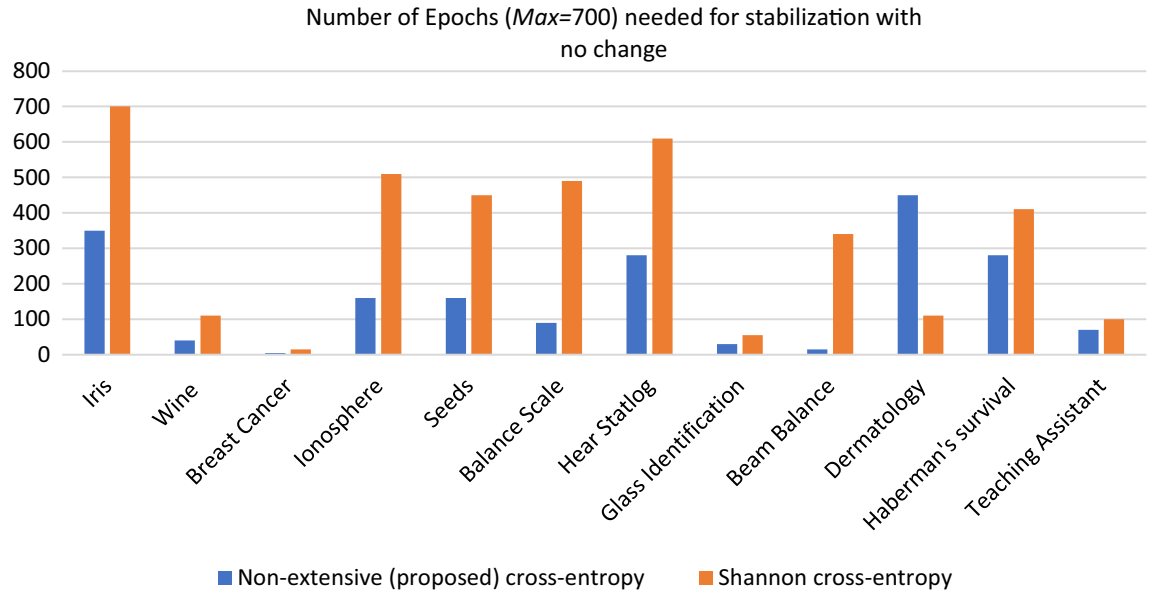


Fig. 6 Comparison chart of the average number of epochs required for ANN-PSO to stabilize when the global-best is determined by Shannon cross-entropy and the proposed non-extensive cross-entropy for all datasets

Tables 2 and 3 when the original cost function is substituted by the new cross-entropy error function.

The execution time is summarized in Tables 4 and 5 for all methods for 700 epochs. The new cross-entropy calculation takes up marginal computation time, and the computation time of one iteration of both the baseline algorithms and their improved versions is almost the same. The number of iterations, however, is expected to differ for some datasets since a different cost function is being used for the global-best calculation. It is interesting to note that the execution time is significantly reduced for several instances in Tables 4 and 5, by inserting the new cross-entropy. In fact, the proposed ANN-PSO takes minimal time for execution among all methods for six datasets in validation (Table 4) and eight of the datasets in cross-validation (Table 5). All the experiments are timed for 700 epochs, since the different cost functions are at different amplitudes and ranges and a common exit error threshold could not be determined. However, we constructed a comparison chart of the average number of epochs required for ANN-PSO to stabilize when the global-best is determined by Shannon cross-entropy and the proposed non-extensive cross-entropy in Fig. 6. The proposed measure stabilizes very quickly as observed from Fig. 6, and takes

but a few epochs to converge. It is our observation that an exit error of 0.4 is a good choice for our new measure, considering that $e^{-1} = 0.36$ is its lowest point.

6 Conclusion

A new non-extensive cross-entropy error function is defined in this paper, to find the global-best position in each iteration for the artificial neural networks trained by Particle Swarm Optimization (ANN-PSO). The use of non-extensive information-theoretic measures improves on the accuracy as proved by the experimental results. A new non-extensive cross-entropy is defined for determining the global-best position of the particle population. Experiments on benchmark datasets with comparison to the state of the art validate the usefulness of our method, since it converges quickly to the global minimum as verified from the high accuracies of the results obtained.

(In case of Funding) Funding: This study was not funded by any agency.

Compliance with ethical standards

Conflict of interest All five authors are associated with the Department of Information Technology, Delhi Technological University, Delhi.

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

References

Anand A, Suganthi L (2017) Forecasting of electricity demand by hybrid ANN-PSO models. Int J Energy Optim Eng 6(4):66–83

Brescia M, Cavuoti S, D'Angelo G, D'Abrusco R, Deniskina N, Garofalo M, Skordovski B (2008) The VO-Neural project: recent developments and some applications. arXiv preprint. arXiv:0806.1006

Chen D, Cao X, Wen F, Sun J (2013) Blessing of dimensionality: high-dimensional feature and its efficient compression for face verification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3025–3032

Chen Z, Zhan Z, Shi W, Chen W, Zhang J (2016) When neural network computation meets evolutionary computation: a survey. In: International symposium on neural networks, Springer, Cham, pp 603–612

Chen XL, Fu JP, Yao JL, Gan JF (2018) Prediction of shear strength for squat RC walls using a hybrid ANN–PSO model. Eng Comput 34(2):367–383

Dorigo M, Birattari M, Stutzle T (2006) Ant colony optimization. IEEE Comput Intell Mag 1(4):28–39

Dunne RA, Campbell NA (1997) On the pairing of the softmax activation and cross-entropy penalty functions and the derivation of the softmax activation function. In Proceedings of the 8th Australian conference on the neural networks, Melbourne, vol 181, p 185

Frank A, Asuncion A (2010) UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. https://archive.ics.uci.edu/ml. Accessed 24 May 2019

Gharghan SK, Nordin R, Ismail M, Ali JA (2016) Accurate wireless sensor localization technique based on hybrid PSO-ANN algorithm for indoor and outdoor track cycling. IEEE Sens J 16(2):529–541

Goldberg DE, Holland JH (1988) Genetic algorithms and machine learning. Mach Learn 3(2):95–99

Hajihassani M, Kalatehjari R, Marto A, Mohamad H, Khosrotash M (2020) 3D prediction of tunneling-induced ground movements based on a hybrid ANN and empirical methods. Eng Comput 36(1):251–269

Horst R (1984) On the global minimization of concave functions. Oper Res Spektrum 6(4):195–205

Kennedy J, Eberhart R (1995) Particle swarm optimization (PSO). In Proceedings of IEEE international conference on neural networks, Perth, Australia, pp 1942–1948

Khan JA, Raja MAZ, Rashidi MM, Syam MI, Wazwaz AM (2015) Nature-inspired computing approach for solving non-linear singular Emden-Fowler problem arising in electromagnetic theory. Connect Sci 27(4):377–396

Kiranyaz S, Ince T, Yildirim A, Gabbouj M (2009) Evolutionary artificial neural networks by multi-dimensional particle swarm optimization. Neural Netw 22(10):1448–1462

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86

Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE (2017) A survey of deep neural network architectures and their applications. Neurocomputing 234:11–26

McCaffrey JA (2013) Neural network training using particle swarm optimization. https://visualstudiomagazine.com/articles/2013/12/01/neural-network-training-using-particle-swarm-optimization.aspx. Accessed 24 May 2019

Mendes R, Cortez P, Rocha M, Neves J (2002) Particle swarms for feedforward neural network training. In: Proceedings of the 2002 international joint conference on neural networks. IJCNN'02 (Cat. No. 02CH37290), vol 2, pp 1895–1899

Montana DJ, Davis L (1989) Training feedforward neural networks using genetic algorithms. IJCAI 89:762–767

Nielsen MA (2015) Neural networks and deep learning, vol 25. Determination press, San Francisco

Raja MAZ, Shah FH, Alaidarous ES, Syam MI (2017) Design of bio-inspired heuristic technique integrated with interior-point algorithm to analyze the dynamics of heartbeat model. Appl Soft Comput 52:605–629

Raja MAZ, Shah AA, Mehmood A, Chaudhary NI, Aslam MS (2018) Bio-inspired computational heuristics for parameter estimation of nonlinear Hammerstein controlled autoregressive system. Neural Comput Appl 29(12):1455–1474

Raza S, Mokhlis H, Arof H, Naidu K, Laghari JA, Khairuddin ASM (2016) Minimum-features-based ANN-PSO approach for islanding detection in distribution system. IET Renew Power Gener 10(9):1255–1263

Rumelhart DE, Hinton GE, Williams RJ (1985) Learning internal representations by error propagation (No. ICS-8506). California University, San Diego

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Shore J, Johnson R (1981) Properties of cross-entropy minimization. IEEE Trans Inf Theory 27(4):472–482

Socha K, Blum C (2007) An ant colony optimization algorithm for continuous optimization: application to feed-forward neural network training. Neural Comput Appl 16(3):235–247

Susan S, Dwivedi M (2014) Dynamic growth of hidden-layer neurons using the non-extensive entropy. In: 2014 Fourth international conference on communication systems and network technologies IEEE, pp 491–495

Susan S, Hanmandlu M (2013) A non-extensive entropy feature and its application to texture classification. Neurocomputing 120:214–225

Susan S, Hanmandlu M (2015) Unsupervised detection of nonlinearity in motion using weighted average of non-extensive entropies. SIViP 9(3):511–525

Susan S, Hanmandlu M (2018) Color texture recognition by color information fusion using the non-extensive entropy. Multidimension Syst Signal Process 29(4):1269–1284

Susan S, Sharma M (2017) Automatic texture defect detection using Gaussian mixture entropy modeling. Neurocomputing 239:232–237

Susan S, Singh V (2011) On the discriminative power of different feature subsets for handwritten numeral recognition using the box-partitioning method. In: 2011 annual IEEE India conference IEEE, pp 1–5

Susan S, Sharawat P, Singh S, Meena R, Verma A, Kumar M (2015) Fuzzy C-means with non-extensive entropy regularization. In: 2015 IEEE international conference on signal processing, informatics, communication and energy systems (SPICES) IEEE, pp 1–5

Susan S, Ranjan R, Taluja U, Rai S, Agarwal P (2019) Neural net optimization by weight-entropy monitoring. Computational intelligence: theories, applications and future directions-volume II. Springer, Singapore, pp 201–213

Tizhoosh HR (2005) Opposition-based learning: a new scheme for machine intelligence. In: Computational intelligence for modelling, control and automation, 2005 and international conference on intelligent agents, web technologies and internet commerce, international conference on IEEE, vol 1, pp 695–701

Yaghini M, Khoshraftar MM, Fallahi M (2011) HIOPGA: a new hybrid metaheuristic algorithm to train feedforward neural networks for Prediction. In: Proceedings of the international conference on data mining, pp 18–21

Yaghini M, Khoshraftar MM, Fallahi M (2013) A hybrid algorithm for artificial neural network training. Eng Appl Artif Intell 26(1):293–301

Zameer A, Arshad J, Khan A, Raja MAZ (2017) Intelligent and robust prediction of short term wind power using genetic programming based ensemble of neural networks. Energy Convers Manag 134:361–372

Zhang GP (2000) Neural networks for classification: a survey. IEEE Trans Syst Man Cybern Part C Appl Rev 30(4):451–462

Zhang L, Suganthan PN (2016) A survey of randomized algorithms for training neural networks. Inf Sci 364:146–155

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

Life Sciences 256 (2020) 117912

Contents lists available at ScienceDirect

Life Sciences

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

Identification of novel class I and class IIb histone deacetylase inhibitor for Alzheimer's disease therapeutics

Rohan Gupta, Rashmi K. Ambasta, Pravir Kumar*

Molecular Neuroscience and Functional Genomics Laboratory, Department of Biotechnology, Delhi Technological University (Formerly DCE), Delhi 110042, India

ARTICLE INFO

Keywords:
Histone deacetylase
Machine learning
ChEMBL database
Alzheimer's disease
Virtual screening
IC_{50} value
HDAC inhibitor
Neurodegenerative disease

ABSTRACT

Histone deacetylase enzymes were prominent chromatin remodeling drug that targets in the pathophysiology of Alzheimer's disease associated with transcriptional dysregulation. *In vitro* and *in vivo* models of AD have demonstrated overexpression of HDAC activity. Non-specificity and non-selectivity of HDAC are the major problems of existing HDAC inhibitors. Hence, we aim to set up a methodology describing the rational development of isoform-selective HDAC inhibitor targeting class, I and class IIb. A convenient multistage virtual screening followed by machine learning and IC_{50} screenings were used to classify the 5064 compounds into inhibitors and non-inhibitors classes retrieved from the ChEMBL database. ADMET analysis identified the pharmacokinetics and pharmacodynamics properties of selected compounds. Molecular docking, along with mutational analysis of eleven compounds, characterized the inhibiting potency. Herein, for the first time, we reported ChEMBL1834473 (2-[[5-(4-chlorophenyl)-1,3,4-thiadiazol-2-yl]amino]-N-hydroxypyrimidine-5-carboxamide) as the isoform-selective HDAC inhibitor, which interact central Zn^{2+} atom. The negative energy and interacting residue of the ChEMBL1834473 with six HDAC isoform has also been tabulated and mapped. Moreover, our findings concluded histidine, glycine, phenylalanine, and aspartic acid as key residues in protein-ligand interaction and classify 2347 compounds as HDAC inhibitors. Later, a protein-protein interaction network of six HDAC with the key proteins involved in the progression of an AD and signaling pathway, which describes the relationship between ChEMBL1834473 and AD, has been demonstrated using PPI network where the chosen inhibitor will work. Altogether, we conclude that the compound ChEMBL1834473 may be capable of inhibiting all isoforms of class I and class IIb HDAC based on computational analysis for AD therapeutics.

1. Introduction

Alzheimer's disease (AD) is a fatal nervous system disorder that relates to memory deficit and destroys the thinking and learning ability [1]. The conceivable cause of AD is the loss of synaptic neurons due to the deposition of amyloid-β (Aβ) plaques and tau neurofibrillary tangles (NFT) [2]. Post-translational modification of histone and non-histone proteins has become the most significant area of research in the pathogenesis of AD [3,4]. Previous studies suggested the role of histone acetylation and deacetylation mechanisms playing an important role in memory impairment and recovery of learning and thinking ability [5]. Histone acetyltransferase (HAT) and histone deacetylase (HDAC's) were two key enzymes that promote chromatin relaxation and compaction respectively and thus regulate transcription and gene expression phenomenon. Any instabilities in the HAT and HDAC's activity leads to abnormal gene expression that may cause vital human disorders, for instances, cancer, neuropathy, diabetes, neoplasm, AD, and Parkinson's disease [6,7]. Till date, a total of 18 human HDACs have been identified, which were divided into two superfamilies and four classes based on structural, functional, and phylogenetic analysis [8]. Class I, Class II, Class IV were Zn^{2+} dependent while Class III, which only includes Sirtuins, were NAD^{+} dependent [9]. Class I HDAC includes HDAC1, HDAC2, HDAC3, and HDAC8 have sequence similarity to yeast Rpd3 protein whereas, Class II HDAC includes HDAC4, HDAC5, HDAC6, HDAC7, HDAC9, and HDAC10 have sequence similarity to yeast Hda1 protein [10]. Histone deacetylase inhibitors were the effective treatment against neurological disorders through two major neuroprotective mechanisms, (i) site-specific transcriptional activation of a diseased gene that ultimately leads to a controlled gene expression and (ii) regulating the homeostasis of histone acetylation [11]. Several studies show that the treatment with HDAC6 selective inhibitor resulted in the alleviated levels of Aβ degrading enzyme neprilysin (NEP), and administration of MS-275 in AβPP mouse model decreased the Aβ deposition which proves the possible role of HDAC inhibitors in the

* Corresponding author at: Molecular Neuroscience and Functional Genomics Laboratory, Shahbad Daulatpur, Bawana Road, Delhi 110042, India.
E-mail address: pravirkumar@dtu.ac.in (P. Kumar).

https://doi.org/10.1016/j.lfs.2020.117912
Received 1 May 2020; Received in revised form 15 May 2020; Accepted 1 June 2020
Available online 03 June 2020
0024-3205/ © 2020 Elsevier Inc. All rights reserved.

treatment of AD [12,13]. The HDAC2 inhibitor was found to promote the growth of dendritic spine density and further degrade the Aβ deposition [14,15], whereas inhibiting HDAC3 prevents the synaptic plasticity induced by amyloid oligomers [16]. In one study, it was found that inhibiting the activity of HDAC6 promotes the clearance of Aβ and tau [17]. HDAC inhibitors are classified into four groups based on their structure, catalytic activity, and a side chain. These include hydroxamic acids (Vorinostat, Belinostat) [18], benzamide (Entinostat, Mocetinostat, SK-7041) [19], carboxylates (Valproic acids) [20], and cyclic tetrapeptides (Trapoxin A, Romidepsin) [21]. Among all, short-chain carbolic acids were weak inhibitors, while cyclic tetrapeptides were structurally complex inhibitors and currently under clinical trials [22]. Though these inhibitors were different, they share three characteristic features termed as a Zinc-binding group (ZBG) [23] lipophilic linker and a hydrophobic cap [24]. ZBG interacts with central Zn^{2+} through coordination bond and formed hydrogen bonds with the residues near Zn^{2+}, while hydrophobic cap binds with the residues surrounding the active sites. However, *in silico* drug designing strategy, using machine learning (ML) techniques and molecular docking studies had always been providing the most promising results for the identification of the novel inhibitor [25,26]. Therefore, we aim to identify a unique novel broad range inhibitor for all six isoforms of HDAC class I and IIb histone deacetylase by virtual screening and machine learning for the therapeutics in neurodegenerative disorders based on ligand-drug designing technology.

2. Material and methods

2.1. Dataset preparation and data preprocessing

For evaluation of protein structure, high-resolution co-crystallized class I and class IIb HDAC protein structures were extracted from the RCSB PDB database (https://www.rcsb.org/) [27]. HDAC1 (5ICN), HDAC2 (5IWG), HDAC3 (4A69), HDAC8 (5FCW), belongs to class I HDAC while HDAC6 (5EEN) and HDAC10 (5TD7) constitute class IIb HDAC. These structures additionally went through the dock prep section of UCSF Chimera 1.10 (https://www.cgl.ucsf.edu/chimera/) [28] on account of including charges and missing residues. Active sites and conserved domain information of these structures were acquired from PDBSum (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html) [29]. Binding and catalytic mode information of class I and class II HDAC has been studied from different scientific sources. Protein PDB structures undergo energy minimization through ModRefiner: High-Resolution Protein structure Refinement (https://zhanglab.ccmb.med.umich.edu/ModRefiner/) from Zhang Lab available online freely [30].

2.2. Structure and sequence similarity descriptors of protein structures

To distinguish between the structural similarity and identity between different class protein sequences were obtained from NCBI protein database (https://www.ncbi.nlm.nih.gov/protein) and for multiple sequence alignments (MSA), freeware tools available online known as ClustalW tool (https://www.genome.jp/tools-bin/clustalw) [31] was used with a gap open penalty of 10 and gap extension penalty of 0.05. BLOSUM was used as a weight matrix in multiple sequence alignments. The sequence similarity annotation of HDAC isoforms was performed manually with the help of sequence alignment editor tool BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) available online [32].

2.3. Virtual screening of compounds based on IC_{50} value followed by machine learning classification

For dataset preparation, information related to known inhibitors of class I and class II HDAC were isolated from the DrugBank database (https://www.drugbank.ca/) [33]. Further, for preparation of training set and test set data, compounds with known inhibitory concentration known as IC_{50} value were isolated from Binding database (https://www.bindingdb.org/bind/index.jsp) [34] and ChEMBL database (https://www.ebi.ac.uk/chembl/) [35] developed by The European Molecular Biology Laboratory-The European Bioinformatics Institute (EMBL-EBI) respectively. A total of 3234 compounds was used as a training data set belonging to class I and Class II HDAC while 5064 chemical compounds were used as the test data set. In order to compute the 2D and 3D chemical, physical, and geometrical properties freely available software and tools were used such as DataWarrior (http://www.openmolecules.org/datawarrior/) [36] from open molecules, ACD ChemSketch (https://www.acdlabs.com/resources/freeware/chemsketch/) [37] used to draw the chemical structure and to calculate molecular properties, and ChemAxon Chemicalize (https://chemaxon.com/products/chemicalize) [38]. The training compounds were classified as inhibitors and non-inhibitors based on IC_{50} concentration. Compounds having the highest inhibitory effect that is $IC_{50} < 2500$ nM were classified as inhibitor/actives, while the compounds having weak inhibitory potential that is $IC_{50} > 2500$ nM were classified as non-inhibitor/decoys. A total of 37 attributes was calculated based on which test set data were classified as inhibitor or non-inhibitor. Machine learning models were applied to test set data in order to compute their binding effect with the help of freely available software Weka 3.8 (https://www.cs.waikato.ac.nz/ml/weka/) developed by The University of Waikato (New Zealand) [39]. Weka developed four classification model based on different algorithms, which are as follows; random forest model, deep learning model, logistic model, and k-star model. After classification, the data sorted out to reduce the number of compounds in order to obtain a novel inhibitor having high binding energy with both class I and class IIb HDAC.

2.4. Prediction of molecular properties, bioactivity score, and blood-brain-barrier (BBB) permeability

Pharmacokinetic and pharmacodynamics properties were used to study the pharmacological action of compounds administered to a particular cell. The compounds obtained from data sorting and filtering were subjected to analyze ADMET properties, bioactivity score, and BBB permeability using different online web servers such as admetSAR (http://lmmd.ecust.edu.cn/admetsar1/predict/) [40], Molinspiration (https://www.molinspiration.com/cgi-bin/properties) [41], and CBLigand version 0.90 online BBB predictor (https://www.cbligand.org/BBB/) [42]. This will be able to calculate the positive and negative pharmacological action of a compound with the help of graph-based signatures and identify properties such as intestinal absorption, water solubility, blood-brain barrier (BBB) permeability, central nervous system (CNS) permeability, total clearance, AMES toxicity, skin sensitization, and minnow toxicity.

2.5. Molecular docking studies and mutational effect

In order to check the binding efficiency of filtered compounds with class I and class II HDAC's isoforms, molecular docking studies were performed with the help of freely available web service SwissDock (http://swissdock.ch/) [43] developed by Swiss Institute of Bioinformatics with default parameters. Total binding energy and binding pose of ligand to specific HDAC were examined carefully. Before docking studies, ligand set preparation is done with the help of the OpenBabel tool (http://openbabel.org/wiki/Main_Page) [44], which converts the ligand format from mol to mol2, which is used by SwissDock server. Permanent alteration in the nucleotide sequence of genome termed as a mutation in protein structure causes the dysfunction or dysregulation of protein activity and structure. The catalytic sites or binding sites of different Histone Deacetylase possesses missense mutation as well as silent coding mutation. The data of amino acid mutation in HDAC were

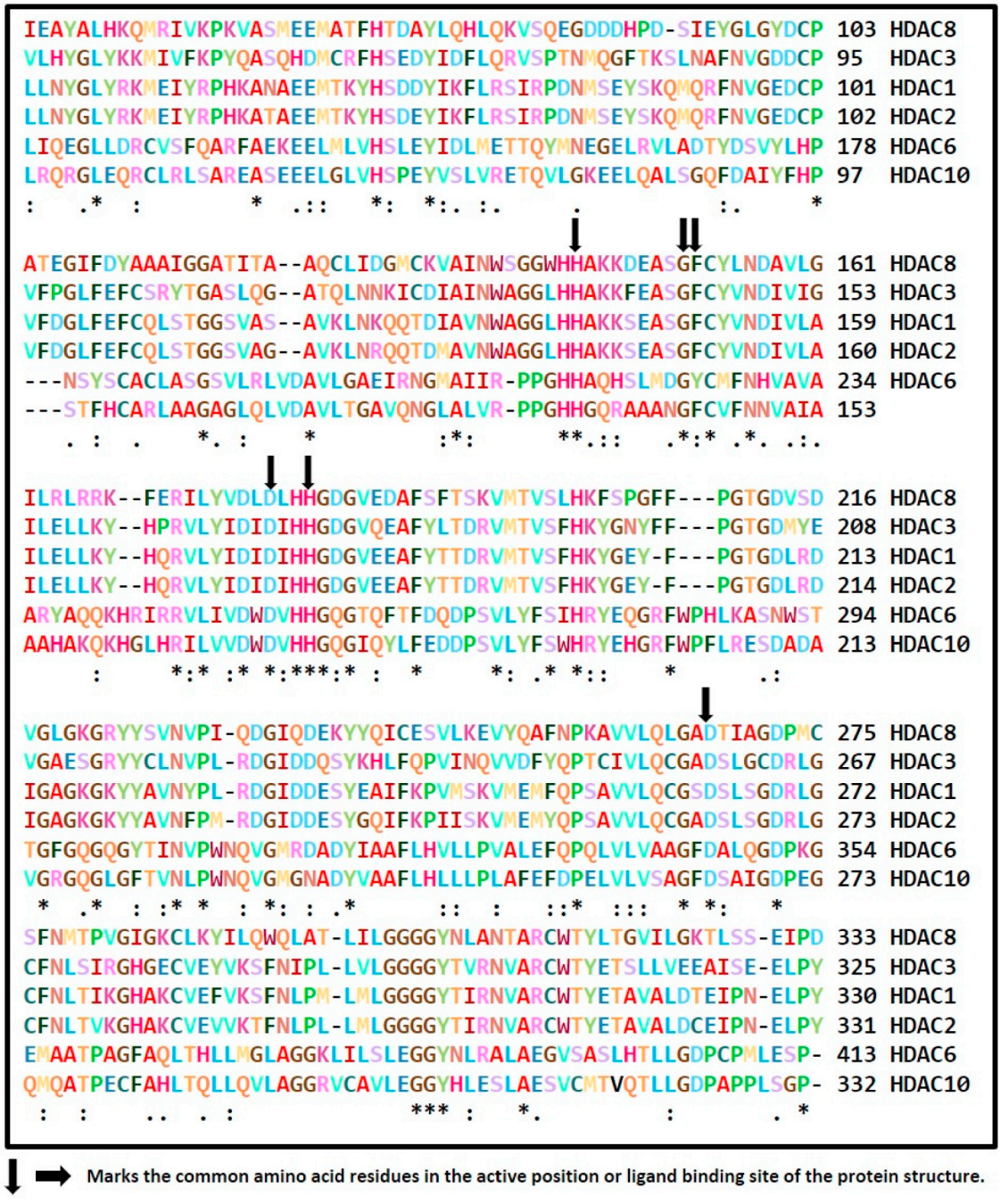


Fig. 1. The above image represents the multiple sequence alignment of six different isoforms of histone deacetylase belongs to class I and class IIb in which "*" represents the identical amino acid residues at a particular position in different isoforms whereas "▪" represents the similarity among different isoforms of histone deacetylase. The figure shows that amino acids, such as Aspartic acid, Histidine, Glycine, and Phenylalanine, were important amino acids in the process of ligand binding. Thus, a particular ligand should be designed in such a way that it will able to interact with these common amino acid residues.

retrieved from the COSMIC database (https://cancer.sanger.ac.uk/cosmic) [45]. UCSF Chimera was used to carrying out the amino acid substitution using the command "*swapaa (aa name): (position) (chain)*" or it can be done with the help of option under tools named as morph conformations. The modified structures were saved into Pdb file separately to carry out the molecular docking studies with SwissDock.

2.6. Protein-protein interaction network of class I and class II HDAC isoforms

The information related to interacting partners of class I and class II HDAC was isolated from the STRING database Version 11.0 (https://string-db.org/) [46]. In order to analyze the interaction pattern between known HDAC isoforms to these proteins, data were imported into freely available software Cytoscape Version 3.7.1 (https://cytoscape.

Table 1
Structural similarity descriptors of different isoforms of HDAC.

	Q9BY41.2	O15379.2	NP_004955.2	Q92769.2	Q8WUI4.2	Q9UKV0.2	P56524.3	Q9UQL6.2	Q96DB2.1	Q9UBN7.2	Q969S8.1
Q9BY41.2 [HDAC8]	100	40.98	40.54	40.7	21.18	22.92	22.42	21.76	23.1	21.99	19.76
O15379.2 [HDAC3]	40.98	100	59.1	59.1	20.77	23.4	22.87	23.85	22.76	21.35	17.88
NP_004955.2 [HDAC1]	40.54	59.1	100	85.06	22.25	24.92	23.95	23.45	24.68	23.06	20.2
Q92769.2 [HDAC2]	40.7	59.1	85.06	100	20.83	25.15	22.51	23.59	24.61	23.68	18.7
Q8WUI4.2 [HDAC7]	21.18	20.77	22.25	20.83	100	48.24	53.37	48.29	22.36	33.48	37.47
Q9UKV0.2 [HDAC9]	22.92	23.4	24.92	25.15	48.24	100	59.69	61.11	22.67	35.19	39.26
P56524.3 [HDAC4]	22.42	22.87	23.95	22.51	53.37	59.69	100	61.6	22.98	35.32	37.56
Q9UQL6.2 [HDAC5]	21.76	23.85	23.45	23.59	48.29	61.11	61.6	100	22.36	34.18	35.25
Q96DB2.1	23.1	22.76	24.68	24.61	22.36	22.67	22.98	22.36	100	26.17	23.99
Q9UBN7.2 [HDAC6]	21.99	21.35	23.06	23.68	33.48	35.19	35.32	34.18	26.17	100	41.58
Q969S8.1 [HDAC10]	19.76	17.88	20.2	18.7	37.47	39.26	37.56	35.25	23.99	41.58	100

Table 2
Characteristics features of machine learning classification models.

Characteristics	Random forest	Logistic model	K-star model	Deep learning method	Average
Accuracy	95.82	97.98	86.95	97.4	94.5375
ROC area	0.993	0.989	0.934	0.992	0.977
PRC area	0.993	0.985	0.938	0.992	0.977
Kappa statistic	0.9045	0.9543	0.7483	0.941	0.887025
TP rate	0.958	0.98	0.89	0.974	0.9505
FP rate	0.066	0.032	0.149	0.039	0.0715
Precision	0.958	0.98	0.889	0.974	0.95025
Mean absolute error	0.0951	0.0596	0.1169	0.0662	0.08445
Root mean squared error	0.1841	0.1476	0.3136	0.1574	0.200675
Relative absolute error	21.44%	13.44%	26.35%	14.93%	19.04%
Root relative squared error	39.09%	31.34%	66.60%	33.42%	42.61%

org/) [47] used for complex network analysis. It gives information related to the nature of the interaction between two or more proteins, whether they form a direct or indirect association and proximity between two proteins. Along with this signaling effect of ChEMBL1834473 was illustrated in Alzheimer's disease (AD) pathology involving inhibition of HDAC isoforms and their interacting partner.

2.7. Multiple target prediction for ChEMBL1834473

In order to identify the functional phenomenon of ChEMBL1834473 on multiple targets freely online webserver TargetNet (http://targetnet.scbdd.com/) used for predicting the binding probability of protein target to given molecule based on chemo genomics data from quantitative structure-activity relationship (QSAR) models [48]. The obtained data were minimized based on the probability score and the functional role of the target in AD pathology. The interaction patterns between multiple targets and AD pathological proteins such as amyloid precursor protein (APP), microtubule-associated protein tau (MAPT), Presenilin-1 (PSEN1), and Presenilin-2 (PSEN2) along with different class I and class II HDAC isoforms were also obtained from STRING database Version 11.0 and Cytoscape Version 3.7.1.

3. Results

3.1. Histidine, glycine, aspartic acid, and phenylalanine were key residue in efficient ligand binding

Binding site analysis of class I and class IIb showed that residues ranging from position 140 to position 250 considered as the active site for class I with amino acids such as Histidine, Glycine, and Aspartic acid. While residues were ranging from position 150 to position 300 were considered as the active site for class IIb. However, studies of catalytic sites explained that Histidine, Glycine, Aspartic acid, and Phenylalanine were important catalytic amino acids for HDAC class IIb. Zn^{2+} has centrally formed a coordination bond with one Histidine and two Aspartic acids, respectively, for both class I as well as class IIb HDAC. Moreover, HDAC forms a tunnel-shaped pocket, which allows the lipophilic part of the inhibitor to form a hydrogen bond with the selective receptor.

Final results of multiple sequence alignments concluded that HDAC1 and HDAC2 share 85.06% sequence similarity using the BLOSUM similarity matrix, which is highest as compared to other HDAC's in class I. The percentage of sequence similarity for HDAC1 reduces to 40.54% and 59.1% respectively with HDAC8 and HDAC3. Among class IIb the sequence similarity between HDAC6 and HDAC10 remains at 41.58%. The results explained the similarity and identity between different isoforms of HDAC. Also, a closer analysis of protein structures indicates that Histidine, Glycine, Phenylalanine, and Aspartic acid were the key residues that were common in their active sites. Thus, it may be concluded that for isoform-selective inhibitor design, ligand should selectively bind with Histidine, Glycine, Phenylalanine, and Aspartic acid as these were common amino acid residues that will align to each other in respective HDAC isoform as shown in Fig. 1. Hence, an inhibitor targeted against these specific amino acid sites, found in all HDAC isoforms will act as broad range inhibitors for all isoforms of HDAC (Table 1).

3.2. Virtual screening of ChEMBL compound library using a machine learning algorithm identifies 11 novel hits

In order to create machine learning models, the training set and test set should be carefully examined, and for this purpose, molecular descriptor studies of compounds were carried out. Training compounds were collected from the binding database based upon IC_{50} values. Total of 3234 compounds was used as training data in Weka 3.8, while 5064 compounds were collected as a test set data from various literature sources. The compounds were treated into DataWarrior to calculate total 37 physical and chemical descriptors in order to classify the compounds into inhibitors and non-inhibitors. Total 4 models named as a random forest model, deep learning model, KStar model, and logistic model were created to classify the test set data. Our generated model predicted that 2347 compounds were classified as inhibitors, while the

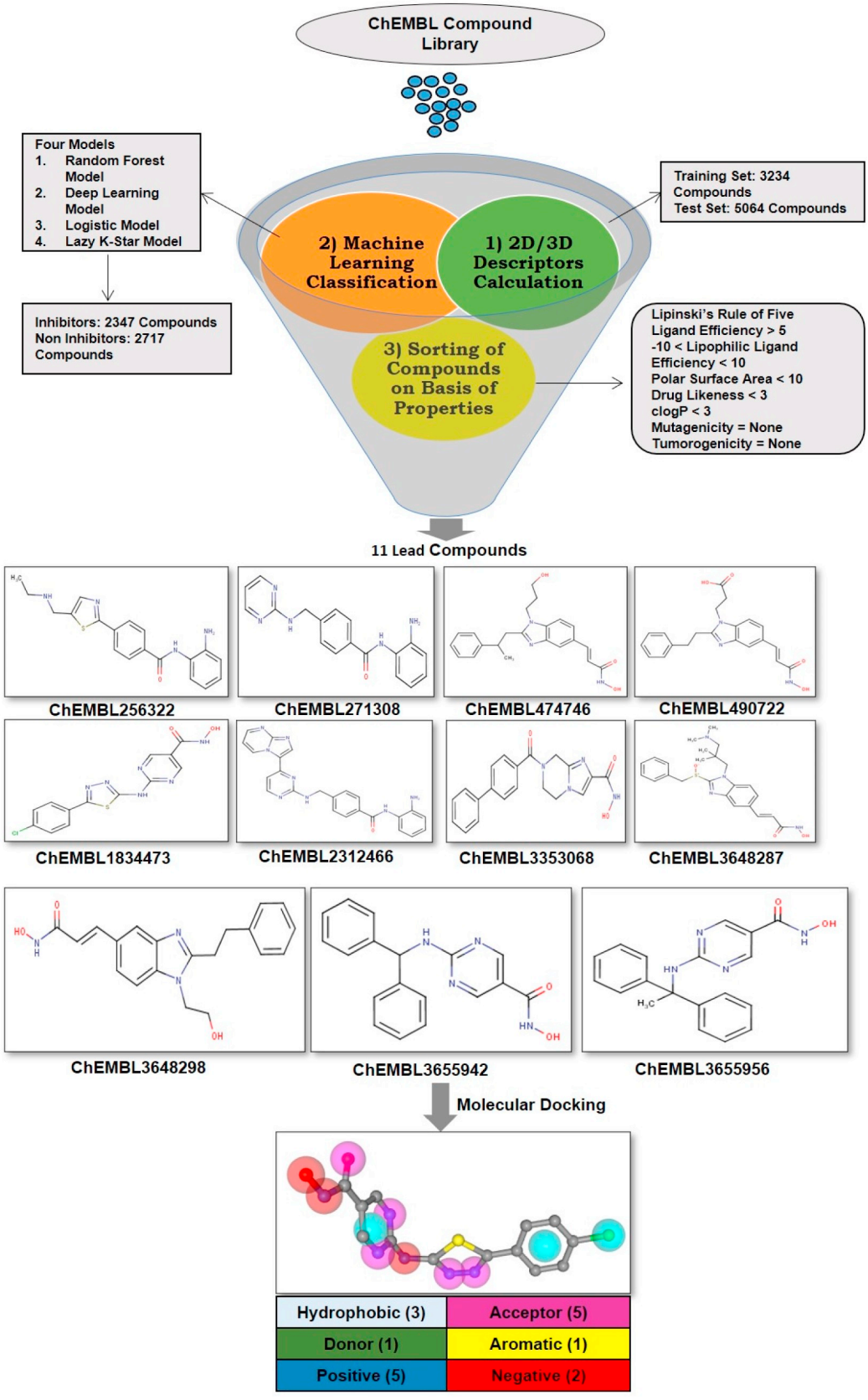


Fig. 2. Virtual Screening of ChEMBL compound libraries for HDAC Class I and Class IIb isoforms novel inhibitor by separated into two data sets that are training data set (3234 compounds) and test data set (5064 compounds). Molecular and chemical properties were evaluated and classify the test data set on the basis of training data into inhibitors (2347 compounds) and non-inhibitors (2717 compounds) using a machine learning algorithm. Afterward, inhibitors were screened on the basis of a prescribed range of drug properties for a compound to be an inhibitor. Eleven compounds were identified after the virtual screening, which was carried out for molecular docking analysis for further validation of compounds.

Table 3
37 calculated molecular descriptor of 11 novel compounds obtained from virtual screening.

	CHEMBL1834473	CHEMBL256322	CHEMBL271308	CHEMBL3353068	CHEMBL3648298	CHEMBL3655942	CHEMBL3655956	CHEMBL474746	CHEMBL490722
GPCR ligand	−0.34	−0.07	0.05	0.28	0.25	0.01	0.21	0.11	0.33
Ion channel modulator	−0.94	−0.28	−0.11	−0.07	−0.18	−0.37	−0.04	−0.24	−0.11
Kinase inhibitor	0.47	0.28	0.39	0.38	0.23	0.35	0.38	0.05	0.11
Nuclear receptor ligand	−0.61	−0.33	−0.29	−55	0.07	−0.27	0.09	0.05	0.17
Protease inhibitor	0.02	0	−0.14	0.52	0.34	0.32	0.58	0.29	0.39
Enzyme inhibitor	0.29	0.12	0.17	0.46	0.51	0.32	0.45	0.4	0.53
Dreiding energy	61.07	72.6	65.33	97.56	67.81	54.83	69.07	76.23	69.56
Van Der Waals volume	262.13	315.32	285	313.76	317.53	282.7	300.48	351.77	336.61
MMFF94 energy	119.11	138.89	93.76	258.54	177.1	137.04	153.58	167.19	136.59
Platt index	68	72	68	84	74	68	74	80	80
Randic index	11.15	12.17	11.72	13.15	12.71	11.78	12.13	13.62	13.56
Balban index	1.32	1.23	1.23	1.25	1.49	1.43	1.45	1.45	1.43
Harary index	77.59	86.99	81.32	100.73	93.15	84.69	91.83	104.69	103.43
Hyper Weiner index	6075	8151	7342	9505	8045	5362	5688	9668	9986
Szeged index	2081	2564	2436	3308	2592	2139	2282	3033	3082
Wiener index	1399	1771	1612	2073	1858	1404	1508	2213	2254
Wiener polarity	32	35	33	43	37	35	40	41	39
Total molweight	348.773	352.461	319.367	362.388	351.405	320.351	334.378	379.459	379.415
cLogP	2.241	2.6298	2.0012	2.1436	2.1121	2.2175	2.1407	2.994	2.0465
cLogS	−3.989	−3.817	−3.854	−2.593	−2.568	−3.814	−3.914	−3.206	−2.82
H-acceptors	8	5	6	7	6	6	6	6	7
H-donors	3	3	3	2	3	3	3	3	3
Total surface area	246.2	276.45	253.25	272.13	279.03	249.82	259.99	303.79	296.78
Relative PSA	0.46019	0.29864	0.28892	0.26469	0.24564	0.2842	0.27309	0.22562	0.27488
Polar surface area	141.16	108.28	92.93	87.46	87.38	87.14	87.14	87.38	104.45
Drug likeness	3.3625	3.041	1.225	4.6404	0.37885	0.34825	1.1164	0.25833	0.41437
LE from cLogP	0.51592	0.47083	0.49723	0.44047	0.45775	0.49468	0.47574	0.41763	0.42572
LLE from cLogP	6.4086	5.9503	6.6975	6.5253	6.5632	6.4366	6.5287	5.5297	6.6425
LELP from cLogP	4.3437	5.5854	4.0247	4.8666	4.6141	4.4827	4.4998	7.1691	4.8071
Mutagenic	None	None	None	None	None	None	None	None	None
Tumorigenic	None	None	None	None	None	None	None	None	None
Reproductive effective	None	None	None	None	None	None	None	None	None
Irritant	None	None	None	None	None	None	None	None	None
Shape index	0.69565	0.68	0.66667	0.62963	0.61538	0.54167	0.52	0.57143	0.57143
Metal-atoms	0	0	0	0	0	0	0	0	0
Rotatable bonds	4	6	5	3	7	5	5	8	8
Amides	0	1	1	1	0	0	0	0	0

Table 4
ADMET properties of 11 identified novel hits.

	Properties	1,834,473	2,312,466	256,322	271,308	3,353,068	3,648,287	3,648,298	3,655,942	3,655,956	474,746	490,722
Absorption	Water solubility (log mol/L)	−2.922	−2.893	−3.786	−3.2	−3.32	−3.053	−3.018	−3.681	−3.802	−2.995	1.078
	Caco2 permeability (log Papp in 10–6 cm/s)	0.804	2.012	0.921	1.026	1.204	0.962	0.811	0.779	0.86	1.065	1.122
	Intestinal absorption (% absorbed)	78.62	82.134	89.723	83.083	91.128	83.31	84.436	91.709	91.414	84.634	85.843
	Skin permeability (log Kp)	−2.742	−2.735	−2.759	−2.805	−2.735	−2.735	−2.735	−2.736	−2.739	−2.735	−3.534
Distribution	CNS permeability (log PS)	−3.242	−2.765	−2.249	−2.437	−2.489	−2.831	−2.794	−2.529	−2.48	−2.711	−3.016
Metabolism	CYP2D6 substrate	No	No	No	No	No	No	No	No	No	No	No
	CYP3A4 substrate	Yes	No	Yes	Yes	No	Yes	Yes	Yes	No	No	No
Excretion	Total clearance (log ml/min/kg)	0	0.98	1.055	0.249	0.467	0.264	1.139	0.396	0.398	0.901	0.633
	Renal OCT2 substrate	No	Yes	No	No	No	No	No	No	No	Yes	No
Toxicity	AMES toxicity	No	Yes	Yes	No	Yes	Yes	Yes	No	No	Yes	No
	Hepatotoxicity	Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No
	Skin sensitisation	No	No	No	No	No	No	No	No	No	No	No
	Minnow toxicity (log mM)	0.691	0.68	0.38	1.757	1.334	0.125	−0.264	2.054	2.472	1.051	3.097
BBB+/BBB−	Blood brain barrier (> 0.02)	0.028	0.049	0.06	0.082	0.033	0.051	0.067	0.106	0.079	0.064	0.027

remaining compounds were classified as non-inhibitors. To check the reliability and predictability of generating models, several statistical parameters such as root mean square error, relative squared error, root relative squared error, and precision was calculated carefully (Table 2). The compounds classified as inhibitors were further screened out based on Lipinski's rule of 5 and discussed molecular descriptors, as described in Fig. 2.

A total 11 compounds out of 2347 compounds were selected for a drug like analysis and docking studies. The compounds were selected based on several parameters such as ligand efficiency (LE) > 5, −10 < lipophilic ligand efficiency (LLE) < 10, mutagenicity should be none, clogP, drug-likeness < 3, and polar surface area < 10. Molecular descriptors of the novel compounds were calculated with ACD ChemSketch, Molinspiration, and ChemAxon's Chemicalize, which explains the topological descriptors of the novel compounds and the geometrical descriptors of novel lead compounds. Theoretically, for a compound to be high bioactive ligand clogP should be < 3, while LipE should be > 5. Hence after the virtual screening of the ChEMBL compound library with more than six thousand compounds, 11 novel hits were selected based on IC_{50} value by machine learning algorithm (Table 3). The eleven novel compounds were further analyzed by ADMET prediction for further compound selection for docking analysis and best inhibitor selection based on negative energy and binding site of novel inhibitor with and without a mutation.

3.3. ADMET and BBB prediction of eleven novel hits

ADMET analysis of novel compounds was carried out to check the pharmacokinetics and pharmacodynamics properties of lead compounds. In our study, we predict 14 ADMET properties of selected compounds out of them 4 each of absorption and toxicity properties and 2 each of distribution, metabolism, and excretion properties. For a compound, to be an effective drug it must fulfill these parameters which have their own range values such as Caco2 permeability > 0.80, intestinal absorption > 40%, skin permeability logKp < −2.5, BBB permeability > −0.3, CNS permeability > −2.0, and for Minnow toxicity logLC50 > −0.3. Hepatotoxicity predicts the action of compound on normal liver function. Furthermore, if the given compound is found to be AMES positive, then it will be considered as mutagenic. As we know that protein associated with AD will be found in the selective region of the brain, thus for a drug to be effective, it must pass the blood-brain barrier and central nervous system. Our study indicates that all 11 predicted compounds fulfill the eligibility criteria and show favorable results, which is given in the Table 4.

Therefore, we summarize that all eleven compounds fulfill the ADMET criteria for being a neurodegenerative HDAC inhibitor. After predicting favorable pharmacokinetics and pharmacodynamics properties, we carried out BBB permeability of 11 novels predicted hits, which is necessary for any compound to be said as a drug molecule. For BBB calculation support vector machine (SVM) approach with MACCS fingerprint was used within the CBLigand online tool with a cut off value of 0.02.

3.4. ChEMBL1834473 is a novel compound used as a drug against both class I and class IIb HDAC isoforms

Comparison between the full fitness of 11 novel compounds to 7 known inhibitors explained that the three compounds such as ChEMBL1834473, ChEMBL271308, and ChEMBL490722 have a higher binding affinity to both class I and class IIb HDAC due to which they have the strong potential to be used as isoform-selective HDAC inhibitors. Among the known inhibitors, entinostat exhibit the highest binding affinity to all forms of HDAC. The three novel compounds have the highest binding energy or a full fitness as compared to high-energy compounds among the known HDAC inhibitors. From the study of Binding pocket and binding site analysis, it is cleared that ChEMBL271308 can be used to inhibit HDAC class I and but it cannot be used for HDAC class IIb as the compound is unable to bind with central Zn^{2+} atom, which is characteristics features as discussed earlier.

Also, ChEMBL490722 can be used only with HDAC class IIb due to its specific binding with Zn^{2+} central atom present in the HDAC structure. In class I except HDAC8 other HDACs like HDAC1, HDAC2, HDAC3 bind with the novel compound to its specific position. Another compound named ChEMBL1834473 possesses accurate results that have the ability to binds with class I and class IIb specifically and selectively with the Zn^{2+} atom. Fig. 3 above shows the binding of ChEMBL1834473 with different isoforms of HDACs. Thus, from the studies, it can be said that ChEMBL1834473 can be used as an isoform-selective inhibitor of HDAC as a neuroprotective agent. Table 5 described the molecular docking energy of the novel compounds with six different isoforms of histone deacetylase.

Moreover, Table 6 shows the active binding amino acid residues of class I and class IIb HDAC, which takes part in interaction with ChEMBL1834473. Results in Fig. 4 part (B) shows that the novel drug-

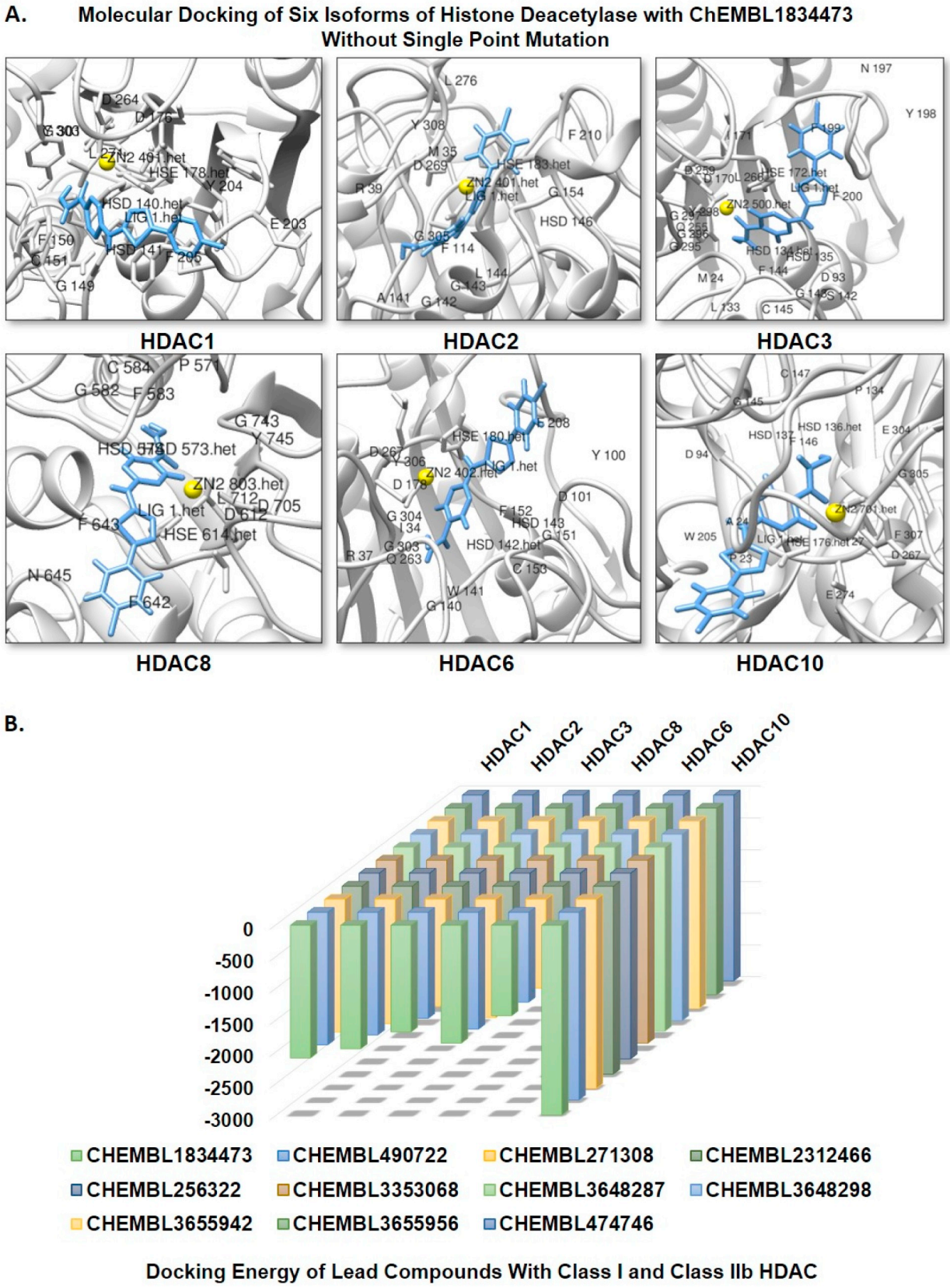


Fig. 3. (A) Represents the molecular docking of six different isoforms of Class I and Class IIb histone deacetylase with a novel lead compound ChEMBL1834473 without mutation on active sites of protein structure. (B) The bar graph shows the energy values of Class I and Class IIb HDAC with eleven different novel hits obtained from the virtual screening of compound libraries.

Table 5
Binding energy of 11 novel hits with class I and class IIb HDAC.

CHEMBLID	HDAC1	HDAC2	HDAC3	HDAC6	HDAC8	HDAC10
CHEMBL1834473	−2088.59	−1944.2	−1678.4	−1424.5	−1855.6	−2986.2
CHEMBL490722	−2087.795	−1930.08	−1669.18	−1414.11	−1839.5	−2950.6
CHEMBL271308	−2090.9	−1959.2	−1684.7	−1400.4	−1863.9	−2977.7
CHEMBL2312466	−2087.54	−1954.1	−1683.7	−1383.6	−1843.9	−2956.8
CHEMBL256322	−2037.16	−1860.7	−1637.2	−1363.5	−1792.6	−2922.1
CHEMBL3353068	−1990.64	−1836.9	−1579	−1321.7	−1770.7	−2881.7
CHEMBL3648287	−2028.31	−1882.6	−1611.6	−1368.8	−1790.3	−2895
CHEMBL3648298	−2055.88	−1903.1	−1623.84	−1385.1	−1802.4	−2934.1
CHEMBL3655942	−2075.97	−1920.6	−1668	−1412.2	−1839.3	−2947.7
CHEMBL3655956	−2060.5	−1913.3	−1649.5	−1401.4	−1844.7	−2938.5
CHEMBL474746	−2057.79	−1897.6	−1650.8	−1390	−1814.3	−2923.1

Table 6
Binding energy and residues of novel compound ChEMBL1834473 with six isoforms of HDAC.

HDAC class	HDAC type	Molecular docking energy (kcal/mol)	Interacting residues
Class I	HDAC1	−2088.59	HSD^{140}, PHE^{141}, GLY^{149}, PHE^{150}, CYS^{151}, ILE^{177}, ASP^{176}, HSE^{178}, GLU^{203}, TYR^{204}, PHE^{205}, LEU^{271}, GLY^{301}, TYR^{303}
	HDAC2	−1944.20	HSD^{145}, ILE^{180}, ASP^{181}, HSE^{183}, HSP^{184}, PHE^{268}, TYR^{271}, TYR^{284}, ARG^{304}
	HDAC3	−1678.40	ASP^{93}, HSD^{134}, HSD^{135}, GLY^{143}, CYS^{145}, ASP^{170}, HSE^{172}, TYR^{198}, PHE^{199}, PHE^{200}, ASP^{259}, LEU^{266}, GLY^{296}, TYR^{298}
	HDAC8	−1855.60	TYR^{100}, ASP^{101}, GLY^{140}, TRP^{141}, HSD^{142}, HSD^{143}, GLY^{151}, PHE^{152}, CYS^{153}, ASP^{178}, HSE^{180}, PHE^{208}, GLN^{263}, ASP^{267}, GLY^{303}, TYR^{306}
Class IIb	HDAC6	−1424.50	GLY^{609}, PHE^{610}, CYS^{611}, VAL^{616}, VAL^{619}, ALA^{620}, GLY^{643}, GLU^{651}, ASP^{652}, ALA^{743}, ASP^{747}
	HDAC10	−2986.20	PRO^{134}, HSD^{136}, HSD^{137}, GLY^{145}, PHE^{146}, CYS^{147}, HSE^{176}, TRP^{205}, ASP^{267}, GLU^{274}, GLU^{304}, GLY^{305}

like compound has the higher binding energy of full fitness as compared to the binding energy of other novel hits of HDAC class I and class IIb. Interpretation of docking results found out that in HDAC3, HDAC8, and HDAC10 mutation causes distortion in the structure due to which binding of several novel compounds that are ChEMBL1834473, binding does not take place in its active site, and thus it is unable to facilitate the interaction of compound to central Zn^{2+} atom.

From the molecular docking results, it can be predicted that the proposed compound does not have the tendency to bind with the central Zn^{2+} atom in its mutated state due to distortion in the structure. Altogether, the ChEMBL1834473 has been identified as the unique novel broad range inhibitor for all isoforms based on docking analysis, which is specific to the amino acid sequence present in all isoforms of HDAC based on mutation followed by docking investigation and negative energy analysis.

3.5. Interaction of ChEMBL1834473 with AD related proteins and its functional effect in AD progression

Six isoforms of HDAC, including Class I and Class IIb has an influential effect on AD-related proteins either directly or indirectly based on PPI analysis and interpretation. Fig. 5 part (A) shown above explains that modulation in one or more HDAC isoform has a direct functional impact on proteins such as Amyloid βeta Peptide Protein (AβPP), Metastasis Associated Protein 1 and 2 (MTA1 and 2), Tumor Protein 53 (TP53), E-3 Ubiquitin Protein Ligase mdm2 (MDM2), Microtubule Associated Protein Tau (MAPT), Glyceraldehyde 3-Phosphate Dehydrogenase (GAPDH), Beta-Secretase 1 Precursor (BACE1), and Histone-binding protein RBBP4 and RBBP7. However, proteins such as Alpha-Synuclein (SNCA), Brain-Derived Neurotrophic Factor (BDNF), Leucine-Rich Repeat Kinase 2 (LRRK2), ADAM Metallopeptidase Domain 10 (ADAM10), Clusterin (CLU), presenilin 1 and 2 (PSEN 1 and PSEN 2), Neurotrophic Receptor Tyrosine Kinase 2 (NTRK2) does not associate directly with HDAC isoforms, but they were considered as the first associates of HDAC isoforms through interaction with intermediate interactor.

Similarly, proteins such as Apolipoprotein E (ApoE), Bridging Integrator-1 (BIN1), Phosphoglycerate Kinase 1 and 2 (PGK1 and PGK2), ATP-Binding Cassette Transporter A7 (ABCA7), Phosphat idylinositol Binding Clathrin Assembly Protein (PICALM), and Complement Factor H (CFH) were considered as a second or third interactor of HDAC isoforms of Class I and Class IIb. They also do not show direct relation with histone deacetylase, but modulating the function of one or other HDAC will definitely alter the function of these proteins through a cascade of events.

Moreover, Part (B) of Fig. 5 explain the effect of ChEMBL1834473 administration on the expression of several proteins involved in neurodegenerative diseases, especially Alzheimer's disease, and ultimately decreases the pathology of neurodegenerative disease. For example, ChEMBL1834473 increases the expression of CLU and ABCA7, which decreases the Amyloid-β (Aβ) aggregation and promotes the clearance of Aβ in the brain, which inhibits the disease pathology. ChEMBL1834473 inhibits the expression of Glycogen Synthase Kinase 3 β (GSK3β), which inhibits the Tau aggregation and neurofibrillary tangle (NFT) formation, thus control the progression of the disease condition. Also, GSK3β increases PSEN activity, which inhibits the formation of Aβ oligomers. Therefore, the PPI network of HDAC isoforms clearly reveals the influence on proteins responsible for AD pathology, hence ChEMBL1834473 administration will definitely have a therapeutic effect on the neurodegenerative disorder and AD disease therapy based on computational analysis.

3.6. Interaction of novel compound ChEMBL1834473 with other targets minimizing AD pathology

Multiple targets of ChEMBL1834473 were identified using the inline freeware tool TargetNet, which constructs Multiple QSAR models using the Naïve Bayes algorithm based on chemogenomics information of the given drug molecule. The tool predicts 112 targets for ChEMBL1834473 with a probability range varying from 1 to 0.001. Higher the probability score, higher will be the chances of a target as an interacting partner of ChEMBL1834473. Screening of targets was done and targets having a probability < 0.90 were ruled out. Thus total of 30 targets was remaining out of which 13 targets have a maximum interaction probability score of 1 while the remaining 17 compounds have a probability

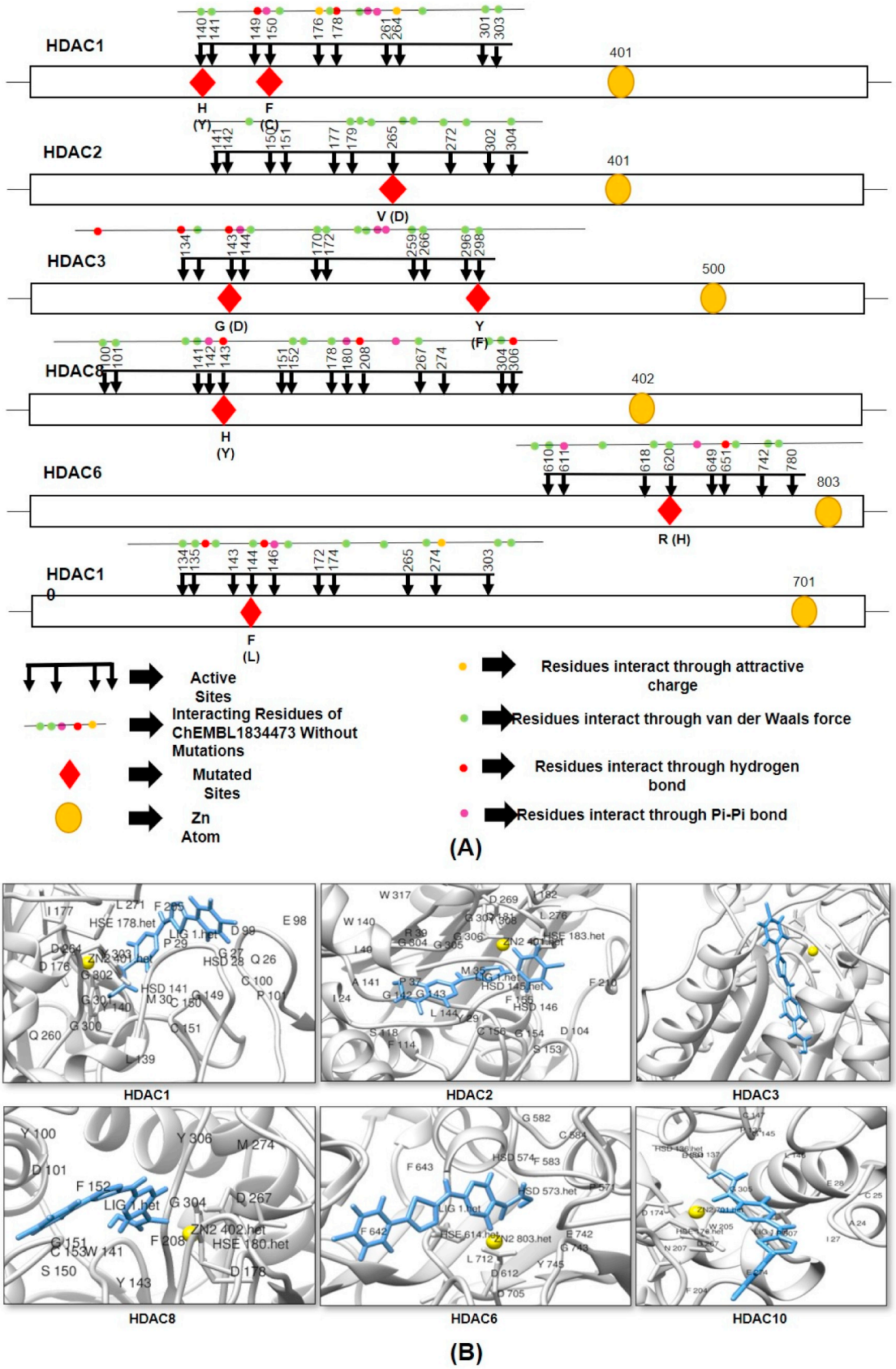


Fig. 4. (A) Bar diagram represents the six different isoforms of histone deacetylase showing active sites, mutated sites, binding residues of ChEMBL1834473, and central Zn atom along with the nature of bonding between novel lead compound and HDAC structure. (B) The figure shows the binding of a novel compound that is ChEMBL1834473 with six isoforms of HDAC after a single point mutation. Round shaped circle with yellow color shows the centrally bound zinc atom while the novel compound is shown in blue color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

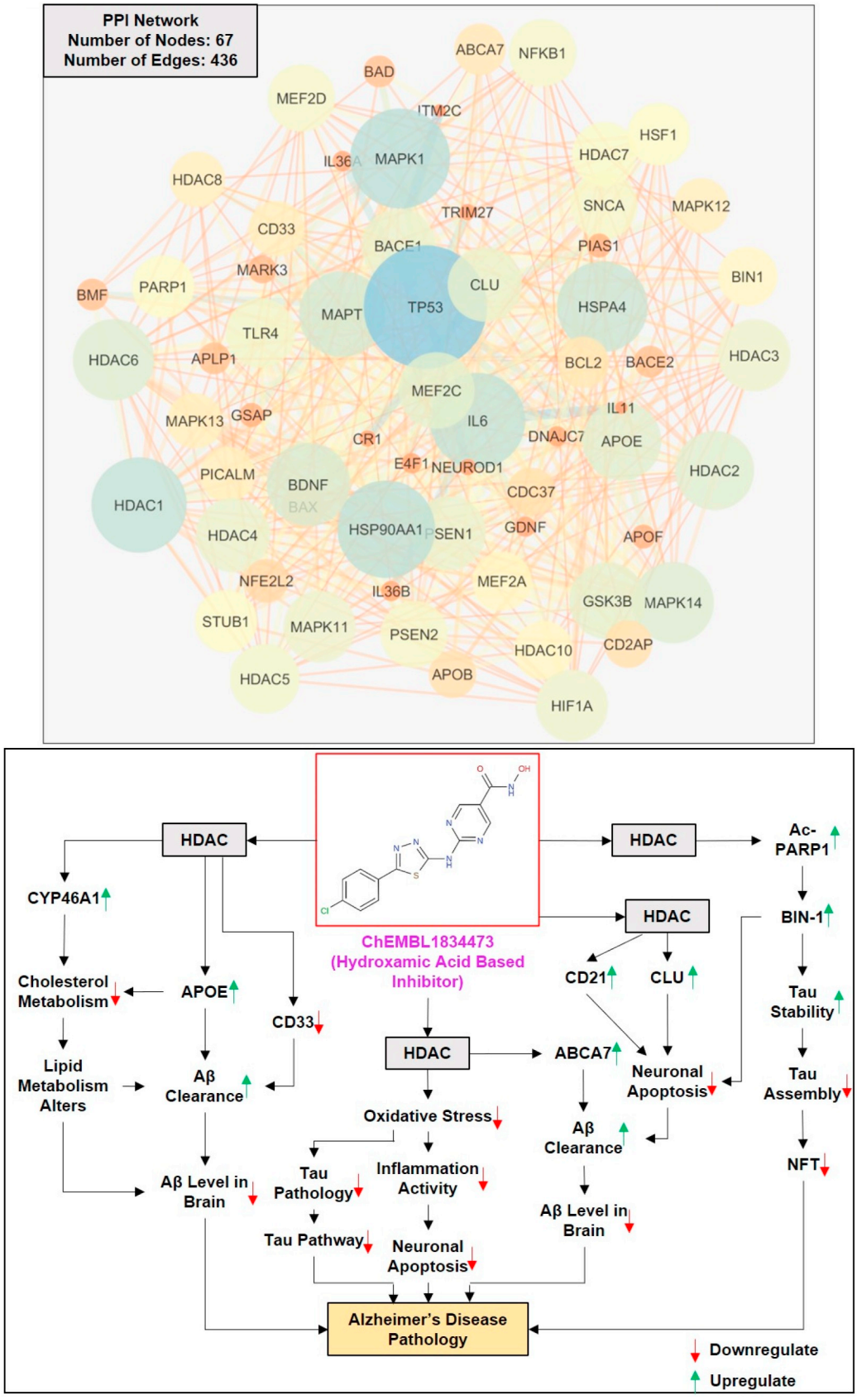
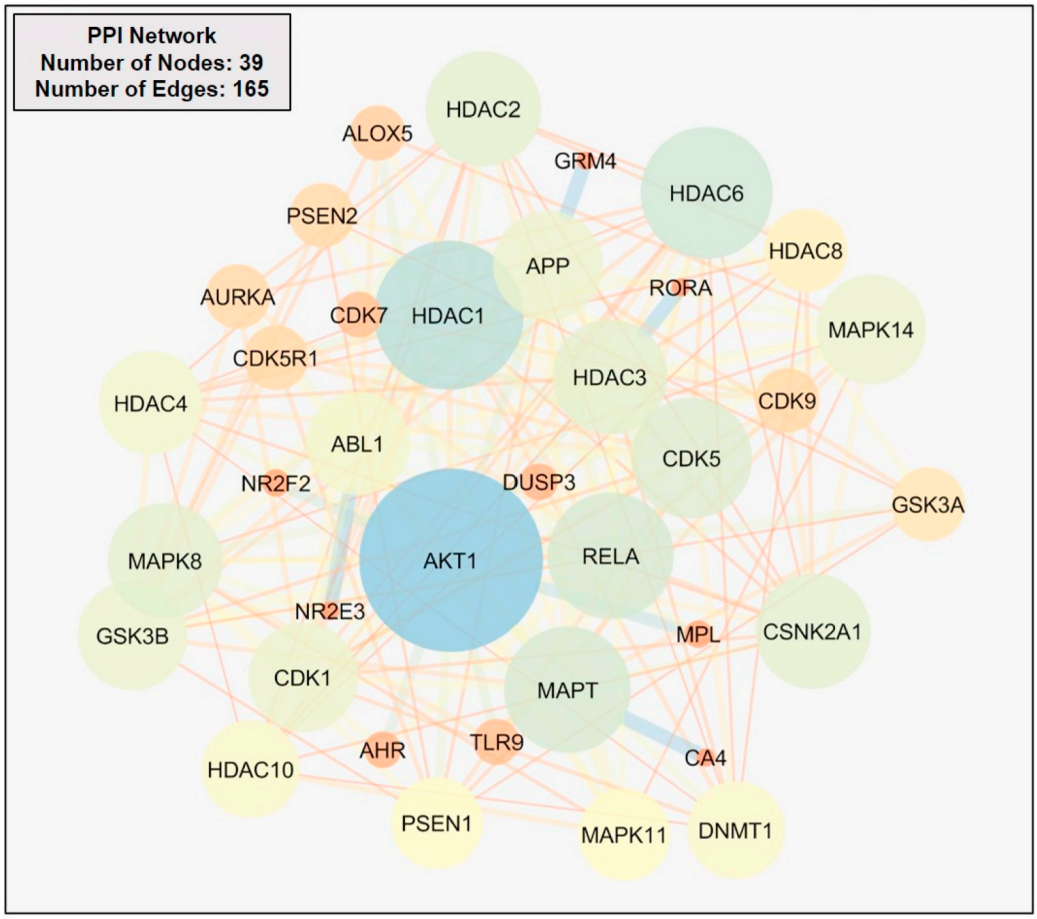


Fig. 5. (A) Protein-Protein interaction network of class I and class II HDAC's with different proteins implemented in Alzheimer's disease. The network consists of 67 nodes and 436 edges. The number of nodes signifies the total number of proteins involved in the network, while the number of edges represents the number of interactions between different proteins. In the given figure, the proteins were mapped with respect to size and color according to the eccentricity of a particular protein in the network. Similarly, the edges were mapped according to the distance between them or the association between two proteins. From the figure, it is easily understood that HDAC1, HDAC6, TP53, MAPK1, HSP90, IL6, HSPA4, MEF2C, BDNF, MAPT, and CLU were highly involved in the disease progression. While HDAC2, HDAC3, APOE, GSK3B, MAPK14, MEF2A, HSF1, and BIN1 were slightly less associated with the disease progression among the proteins involved in the network. (B) The figure represents the possible implication of novel inhibitor ChEMBL1834473 in the suppression of Alzheimer's disease pathology through interaction with class I and class II HDAC. HDAC involved in the deacetylation pathway of regulatory proteins such as PARP-1, BIN1, CLU. CD33, ABCA&, APOE, CYP46A1, and CD21. Variation in the expression level of these proteins initiates the signaling pathways such as cholesterol metabolism, oxidative stress, inflammation activity, which decreases the formation of neurofibrillary tangles and deposition of toxic amyloid-β protein. This mechanism decreases the neuronal apoptosis due to toxic effects mediated by hyperphosphorylated tau and Aβ aggregations.



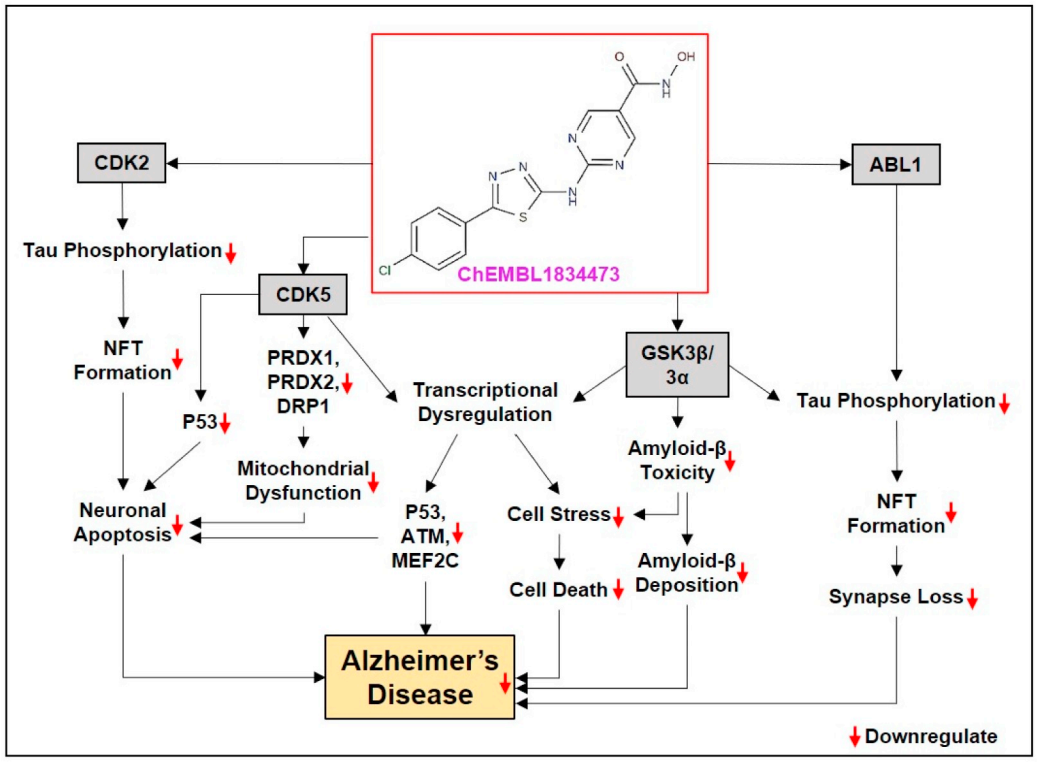


Fig. 6. (A) Protein-Protein interaction network of ChEMBL1834473 multiple targets in association with class I and class IIb HDAC together with the core proteins involved in the pathogenesis of AD such as APP, Tau, TREM2, PSEN1, and PSEN2. In the above network total number of proteins were 39 represented by number of nodes while there were 165 number of interactions which were represented by number of edges. The proteins in the network were colored according to the centrality of particular protein in the network. (B) Interaction patterns of ChEMBL 1834473 with its multiple targets in order to minimize the progression of AD. ChEMBL1834473 Inhibits CDK2 and CDK5 which causes the inhibition of tau phosphorylation and transcription of PRDX1, PRDX2, DRP1, and P53 respectively which causes the inhibition of neuronal apoptosis. Also, inhibition of CDK5, GSK3β, and GSK3α causes transcriptional regulation which in turn reduces cell stress and increases neuroprotection. Inhibition of Abl with ChEMBL 1834473 causes decrease in tau phosphorylation results in low NFT formation in order to increase neuroprotection.

score from 0.999 to 0.951. Multiple target prediction strategies validated our study as the results observed HDAC3 (P. Score = 1), HDAC1 (P. Score = 1), HDAC6 (P. Score = 1), and HDAC2 (P. Score = 0.999) as a potential target of ChEMBL1834473. Moreover, literature validation observed the potential of Cyclin-dependent kinase 2 (CDK2), Tyrosine-protein kinase ABL1 (ABL1), Cyclin-dependent-like kinase 5 (CDK5), Glycogen synthase kinase-3 alpha (GSKα), and Glycogen synthase kinase-3 β (GSK3β) in the pathogenesis of AD which can also be the potential target of ChEMBL1834473 as shown in Fig. 6(B). Moreover, the protein-protein interaction network was plotted between different targets and core AD proteins together with the HDAC class I and class IIb isoforms with the help of Cytoscape Version 3.7.1. A total 165 numbers of interactions were determined between 39 proteins, including multiple targets, class I and class IIb HDAC's, and core AD proteins. Altogether inhibition of these targets with ChEMBL1834473 prevents tau hyperphosphorylation and formation of Amyloid β aggregation, which reduces cell stress and, ultimately, neuronal apoptosis Fig. 6(A).

4. Conclusions

There are several important parameters are involved in the drug detection procedure, for instance, ligand-based drug interaction, machine learning approach, molecular docking based on the binding or full fitness energy of ligand to a receptor. Here we screen the data of chemical compounds from the ChEMBL compound library for the class I and class IIb HDAC by creating four machine learning models, for example, logistic model, k-star model, random forest model, and deep learning model. The reason behind opting these four machines models to classify the compounds into inhibitors and non-inhibitors based on their 2D/3D molecular and chemical properties. There ae 37 molecular descriptors belong to the physical, chemical, and biological properties, which were taken in order to classify the compounds. The average accuracy (94.53), precision (0.950), TP rate (0.950), FP rate (0.071), and area under the ROC curve of our generated models was quite high as compared to the previously created model for selecting a novel isoform-selective HDAC inhibitor. Consequently, 11 novel hits were identified through their 2D/3D structure and ADME properties. The protocol applied during the study was quite reliable and efficient as the whole study based on 37 different molecular descriptors of known inhibitors and novel compounds. Further, the data for testing the potential of inhibitors include > 6000 compounds. ADMET analysis of 11 novel hits explained the potential of compounds to be taken as a drug. For a chemical compound to be used as an efficient drug, it must have some absorption, distribution, metabolism, excretion, and toxicity values. Our results indicate that 11 novel hits shown a favorable range of values. Altogether, these compounds were subjected to molecular docking technique in order to check the binding efficiency of novel hits in the active position of six isoforms of HDAC. Through molecular docking studies, the novel compounds with binding energy or full fitness for both class I and class IIb HDACs were selected for comprehensive analysis. The predicted compounds had the characteristics feature that is a central Zn^{2+} atom in its binding region. Out of these 11 predicted compounds, ChEMBL1834473, ChEMBL490722, and ChEMBL271308 possess the greater binding energy for the class I and class IIb HDAC. However, due to insufficient binding of ChEMBL490722 and ChEMBL271308, these compounds were discarded and not used for further investigation. ChEMBL1834473 involved in hydrogen bonding with central Zn^{2+} atom and possess the basic pharmacophore sites and hence considered as a more potent isoform-selective HDAC inhibitor. Based on the analysis of the catalytic sites, it was found that ChEMBL1834473 forms a bonding with Histidine, Glycine, Phenylalanine, and Aspartic Acid. These bonding were carried out because of the conformational modification in the linker section of protein organization.

Moreover, to validate the inhibition effect of the compound when protein undergoes mutation, molecular docking of ChEMBL1834473 was carried out with the mutated structure of both class I and class IIb HDACs. Based on the energy values, it may be concluded that mutation in the HDAC isoforms decreases the protein stability and a single amino acid replacement or missense mutation took place in the active sites of HDACs. Docking energy analysis shows that the compound has quite comparable energy values as compared with the binding energy of 7 known inhibitors of HDAC. However, the proposed compound was able to promote the discrepancies in protein structure caused by the mutation, which prevents the binding of ChEMBL1834473 to six isoforms of HDAC. Therefore, upon the interpretation of these ligand-based drug designing and virtual screening and machine learning results, we identified ChEMBL1834473 as a uniquely broad range HDAC inhibitor which can be used to inhibit all isoforms of HDAC and can be qualify as a novel drug in the therapy of neurodegenerative disorders.

5. Future perspectives

In this study, we developed and identified a broad range inhibitor, which will be capable of inhibiting both class I and class II HDAC's isoform as a therapeutic agent for Alzheimer's disease. Here, we used a machine learning classification algorithm in order to categorize the sets of compounds into inhibitors and non-inhibitors and after multiple levels of screenings, we identified novel inhibitor known as ChEMBL1834473. This study is a novel approach to incorporate machine learning classification and virtual screening of compounds based on IC_{50} value and different types of molecular descriptors for the drug-discovery process. For the clinical purpose to treat neurodegenerative disorders, this study enables scientists to study the compound *in vitro* and/or *in vivo* studies. Although, we have validated our results from multiple *in silico* approaches such as multiple-target validation and literature validation, still there is a gap that has to be filled by cell culture and animal model studies. From *in silico* studies, we identified the common residues through which inhibitor can potentially bind to the target, but the molecular mechanism behind the inhibition of target expression should be validated only from *in vitro* studies. Moreover, the work will further emphasize the role of epigenetic alterations in the pathogenesis of neurodegenerative disorders such as Alzheimer's disease.

Authors' contributions

All authors had an equal role in the design, work, analysis, results interpretation, and manuscript writing.

Declaration of competing interest

The authors declare that there is no conflict of interests.

Acknowledgement

We would like to thanks senior management of Delhi Technological University for their constant support guidance and research fellowship to Rohan Gupta.

References

[1] K. Xu, X.-L.L. Dai, H.-C.C. Huang, Z.-F.F. Jiang, Targeting HDACs: a promising therapy for Alzheimer's disease, Oxidative Med. Cell. Longev. (2011), https://doi.org/10.1155/2011/143269.
[2] L. Zuo, B.T. Hemmelgarn, C.C. Chuang, T.M. Best, The role of oxidative stress-induced epigenetic alterations in amyloid-β production in Alzheimer's disease, Oxidative Med. Cell. Longev. (2015), https://doi.org/10.1155/2015/604658.
[3] X. Li, X. Bao, R. Wang, Neurogenesis-based epigenetic therapeutics for Alzheimer's disease (review), Mol. Med. Rep. (2016), https://doi.org/10.3892/mmr.2016.5390.
[4] R. Gupta, R.K. Ambasta, P. Kumar, Pharmacological intervention of histone deacetylase enzymes in the neurodegenerative disorders, Life Sci. (2020), https://doi.org/10.1016/j.lfs.2020.117278.

[5] R.T. Liu, L.B. Zou, Q.J. Lü, Liquiritigenin inhibits Aβ25-35-induced neurotoxicity and secretion of Aβ1-40 in rat hippocampal neurons, Acta Pharmacol. Sin. (2009), https://doi.org/10.1038/aps.2009.74.
[6] S. Sidoli, L. Cheng, O.N. Jensen, Proteomics in chromatin biology and epigenetics: elucidation of post-translational modifications of histone proteins by mass spectrometry, J. Proteome (2012), https://doi.org/10.1016/j.jprot.2011.12.029.
[7] S.L. Berger, The complex language of chromatin regulation during transcription, Nature (2007), https://doi.org/10.1038/nature05915.
[8] I.V. Gregoretti, Y.M. Lee, H.V. Goodson, Molecular evolution of the histone deacetylase family: functional implications of phylogenetic analysis, J. Mol. Biol. (2004), https://doi.org/10.1016/j.jmb.2004.02.006.
[9] A.J.M. de Ruijter, A.H. van Gennip, H.N. Caron, S. Kemp, A.B.P. van Kuilenburg, A.J.M.M. de Ruijter, A.H. van Gennip, H.N. Caron, S. Kemp, A.B.P.P. Van Kuilenburg, Histone deacetylases (HDACs): characterization of the classical HDAC family, Biochem. J. 370 (2003) 737–749, https://doi.org/10.1042/BJ20021321.
[10] E. Seto, M. Yoshida, Erasers of histone acetylation: the histone deacetylase enzymes, Cold Spring Harb. Perspect. Biol. (2014), https://doi.org/10.1101/cshperspect.a018713.
[11] G. Donmez, The neurobiology of sirtuins and their role in neurodegeneration, Trends Pharmacol. Sci. (2012), https://doi.org/10.1016/j.tips.2012.05.007.
[12] F. Ajamian, A. Salminen, M. Reeben, Selective regulation of class I and class II histone deacetylases expression by inhibitors of histone deacetylases in cultured mouse neural cells, Neurosci. Lett. (2004), https://doi.org/10.1016/j.neulet.2004.04.087.
[13] E. Hahnen, J. Hauke, C. Tränkle, I.Y. Eyüpoglu, B. Wirth, I. Blümcke, Histone deacetylase inhibitors: possible implications for neurodegenerative disorders, Expert Opin. Investig. Drugs (2008), https://doi.org/10.1517/13543784.17.2.169.
[14] C. Kerridge, N.D. Belyaev, N.N. Nalivaeva, A.J. Turner, The Aβ-clearance protein transthyretin, like neprilysin, is epigenetically regulated by the amyloid precursor protein intracellular domain, J. Neurochem. (2014), https://doi.org/10.1111/jnc.12680.
[15] Y.M. Sung, T. Lee, H. Yoon, A.M. DiBattista, J.M. Song, Y. Sohn, E.I. Moffat, R.S. Turner, M. Jung, J. Kim, H.S. Hoe, Mercaptoacetamide-based class II HDAC inhibitor lowers Aβ levels and improves learning and memory in a mouse model of Alzheimer's disease, Exp. Neurol. (2013), https://doi.org/10.1016/j.expneurol.2012.10.005.
[16] K. Krishna, T. Behnisch, S. Sajikumar, Inhibition of histone deacetylase 3 restores amyloid-β oligomer-induced plasticity deficit in hippocampal CA1 pyramidal neurons, J. Alzheimers Dis. (2016), https://doi.org/10.3233/JAD-150838.
[17] C. Cook, T.F. Gendron, K. Scheffel, Y. Carlomagno, J. Dunmore, M. DeTure, L. Petrucelli, Loss of HDAC6, a novel CHIP substrate, alleviates abnormal tau accumulation, Hum. Mol. Genet. (2012), https://doi.org/10.1093/hmg/dds125.
[18] K. Ververis, A. Hiong, T.C. Karagiannis, P.V. Licciardi, Histone deacetylase inhibitors (HDACIS): multitargeted anticancer agents, Biol. Targets Ther. (2013), https://doi.org/10.2147/BTT.S29965.
[19] M. Dokmanovic, C. Clarke, P.A. Marks, Histone deacetylase inhibitors: overview and perspectives, Mol. Cancer Res. (2007), https://doi.org/10.1158/1541-7786.MCR-07-0324.
[20] C. Monneret, Histone deacetylase inhibitors, Eur. J. Med. Chem. (2005), https://doi.org/10.1016/j.ejmech.2004.10.001.
[21] P.A. Marks, M. Dokmanovic, Histone deacetylase inhibitors: discovery and development as anticancer agents, Expert Opin. Investig. Drugs (2005), https://doi.org/10.1517/13543784.14.12.1497.
[22] P.M. Lombardi, K.E. Cole, D.P. Dowling, D.W. Christianson, Structure, mechanism, and inhibition of histone deacetylases and related metalloenzymes, Curr. Opin. Struct. Biol. (2011), https://doi.org/10.1016/j.sbi.2011.08.004.
[23] X. Hou, J. Du, R. Liu, Y. Zhou, M. Li, W. Xu, H. Fang, Enhancing the sensitivity of pharmacophore-based virtual screening by incorporating customized ZBG features: a case study using histone deacetylase 8, J. Chem. Inf. Model. (2015), https://doi.org/10.1021/ci500762z.
[24] O. Wiest, K. Chen, L. Xu, O. Wiest, Computational exploration of zinc binding groups for HDAC inhibition, J. Organomet. Chem. (2013), https://doi.org/10.1021/jo400406g.
[25] D.-F.F. Wang, P. Helquist, N.L. Wiech, O. Wiest, Toward selective histone deacetylase inhibitor design: homology modeling, docking studies, and molecular dynamics simulations of human class I histone deacetylases, J. Med. Chem. (2005), https://doi.org/10.1021/jm0505011.
[26] H. Tang, X.S. Wang, X.-P.P. Huang, B.L. Roth, K.V. Butler, A.P. Kozikowski, M. Jung, A. Tropsha, Novel inhibitors of human histone deacetylase (HDAC) identified by QSAR modeling of known inhibitors, virtual screening, and experimental validation, J. Chem. Inf. Model. (2009), https://doi.org/10.1021/ci800366f.
[27] S.K. Burley, H.M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J.M. Duarte, S. Dutta, Z. Feng, S. Ghosh, D.S. Goodsell, R.K. Green, V. Guranović, D. Guzenko, B.P. Hudson, T. Kalro, Y. Liang, R. Lowe, H. Namkoong, E. Peisach, I. Periskova, A. Prlić, C. Randle, A. Rose, P. Rose, R. Sala, M. Sekharan, C. Shao, L. Tan, Y.P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, M. Zhuravleva, C. Zardecki, RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy, Nucleic Acids Res. (2019), https://doi.org/10.1093/nar/gky1004.
[28] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, T.E. Ferrin, UCSF chimera—a visualization system for exploratory research and analysis, J. Comput. Chem. (2004), https://doi.org/10.1002/jcc.20084.
[29] R.A. Laskowski, J. Jabłońska, L. Pravda, R.S. Vařeková, J.M. Thornton, PDBsum: structural summaries of PDB entries, Protein Sci. (2018), https://doi.org/10.1002/pro.3289.
[30] D. Xu, Y. Zhang, Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization, Biophys. J. (2011), https://doi.org/10.1016/j.bpj.2011.10.024.
[31] J.D. Thomopson, D.G. Higgins, T.J. Gibson, Clustal W, Nucleic Acids Res. (1994), https://doi.org/10.1093/nar/22.22.4673.
[32] J.L. Phillips, S. Gnanakaran, BioEdit: an important software for molecular biology, Proteins Struct. Funct. Bioinforma. (2015), https://doi.org/10.1002/prot.24632.
[33] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. MacIejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, Di. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. (2018), https://doi.org/10.1093/nar/gkx1037.
[34] M.K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, J. Chong, BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology, Nucleic Acids Res. (2016), https://doi.org/10.1093/nar/gkv1072.
[35] A. Gaulton, A. Hersey, M.L. Nowotka, A. Patricia Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, E. Cibrian-Uhalte, M. Davies, N. Dedman, A. Karlsson, M.P. Magarinos, J.P. Overington, G. Papadatos, I. Smit, A.R. Leach, The ChEMBL database in 2017, Nucleic Acids Res. (2017), https://doi.org/10.1093/nar/gkw1074.
[36] T. Sander, J. Freyss, M. Von Korff, C. Rufener, DataWarrior: an open-source program for chemistry aware data visualization and analysis, J. Chem. Inf. Model. (2015), https://doi.org/10.1021/ci500588j.
[37] G.O. Spessard, ACD labs/LogP dB 3.5 and ChemSketch 3.5, J. Chem. Inf. Comput. Sci. (2002), https://doi.org/10.1021/ci980264t.
[38] ChemAxon.Ltd, chemicalize.org, J. Chem. Inf. Model. Softw. 52 (2) (2012) 613–615, https://doi.org/10.1021/ci300046g.
[39] Data Mining: Practical Machine Learning Tools and Techniques, (2016), https://doi.org/10.1016/c2009-0-19715-5.
[40] F. Cheng, W. Li, Y. Zhou, J. Shen, Z. Wu, G. Liu, P.W. Lee, Y. Tang, AdmetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties, J. Chem. Inf. Model. (2012), https://doi.org/10.1021/ci300367a.
[41] P. Ertl, Calculation of Molecular Properties and Bioactivity Score, © Molinspiration Cheminformatics, (2002).
[42] H. Liu, L. Wang, M. Lv, R. Pei, P. Li, Z. Pei, Y. Wang, W. Su, X.Q. Xie, AlzPlatform: an Alzheimer's disease domain-specific chemogenomics knowledgebase for polypharmacology and target identification research, J. Chem. Inf. Model. (2014), https://doi.org/10.1021/ci500004h.
[43] A. Grosdidier, V. Zoete, O. Michielin, SwissDock, a protein-small molecule docking web service based on EADock DSS, Nucleic Acids Res. (2011), https://doi.org/10.1093/nar/gkr366.
[44] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G.R. Hutchison, Open babel: an open chemical toolbox, J. Cheminform. (2011), https://doi.org/10.1186/1758-2946-3-33.
[45] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J.W. Teague, P.J. Campbell, M.R. Stratton, P.A. Futreal, COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer, Nucleic Acids Res. (2011), https://doi.org/10.1093/nar/gkq929.
[46] D. Szklarczyk, J.H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N.T. Doncheva, A. Roth, P. Bork, L.J. Jensen, C. Von Mering, The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible, Nucleic Acids Res. (2017), https://doi.org/10.1093/nar/gkw937.
[47] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. (2003), https://doi.org/10.1101/gr.1239303.
[48] Z.J. Yao, J. Dong, Y.J. Che, M.F. Zhu, M. Wen, N.N. Wang, S. Wang, A.P. Lu, D.S. Cao, TargetNet: a web service for predicting potential drug–target interaction profiling via multi-target SAR models, J. Comput. Aided Mol. Des. (2016), https://doi.org/10.1007/s10822-016-9915-2.

List of abbreviations

AD: Alzheimer's disease
Aβ: Amyloid β
NFT: neurofibrillary tangles
HAT: histone acetyltransferase
HDAC: histone deacetylase
NEP: neprilysin
ZBG: zinc binding group
Ml: machine learning
MSA: multiple sequence alignment
EMBL-EBI: The European Molecular Biology Laboratory-The European Bioinformatics Institute
BBB: blood brain permeability
CNS: central nervous system
QSAR: quantitative structure activity relationship
APP: amyloid precursor protein
MAPT: Microtubule Associated Protein Tau
PSEN1: Presenilin-1
PSEN2: Presenilin-2
LE: ligand efficiency
LLE: lipophilic ligand efficiency
SVM: support vector machine

MTA1: Metastasis Associated Protein 1
MTA2: Metastasis Associated Protein 2
TP53: Tumor Protein 53
MDM2: E-3 Ubiquitin Protein Ligase mdm2
GAPDH: Glyceraldehyde 3-Phosphate Dehydrogenase
BACE1: Beta-Secretase 1 Precursor
SNCA: alpha synuclein
BDNF: brain derived neurotrophic factor
LRRK2: Leucine-Rich Repeat Kinase 2
ADAM10: ADAM Metallopeptidase Domain 10
CLU: Clusterin
NTRK2: Neurotrophic Receptor Tyrosine Kinase 2
ApoE: Apolipoprotein E
BIN1: Bridging Integrator-1
PGK1: Phosphoglycerate Kinase 1
PGK2: Phosphoglycerate Kinase 2
ABCA7: ATP-Binding Cassette Transporter A7
PICALM: Phosphatidylinositol Binding Clathrin Assembly Protein
CFH: Complement Factor H
GSK3β: Glycogen Synthase Kinase 3 Beta
CDK2: Cyclin-Dependent Kinase 2
ABL1: Tyrosine-Protein Kinase ABL1
CDK5: Cyclin-Dependent-Like Kinase 5
GSK3α: Glycogen Synthase Kinase-3 Alpha

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Image Forgery Detection and Deep Learning Techniques: A Review

1st Ritu Agarwal
Department of Information technology
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road
Delhi 110042, India
ritu.jeea@gmail.com

2nd Deepak Khudaniya
Department of Information technology
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road
Delhi 110042, India
deepakkhudaniya_bt2k16@dtu.ac.in

3rd Abhinav Gupta
Department of Information technology
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road
Delhi 110042, India
gupta.abhinav2010@gmail.com

4th Khyati Grover
Department of Computer Engineering
Delhi Technological University
Shahbad Daulatpur, Main Bawana Road
Delhi 110042, India
khyati.grover@yahoo.com

***Abstract*—Due to the easily available software for tampering images, image manipulation has become quite common. Since the tampered images are non-distinguishable by the naked eye, they are being circulated on various platforms giving rise to rumors and misleading many. This has led researchers to work on various techniques for the detection of manipulated images with improved accuracy. Traditional works on image forgery detection are mostly based on extracting simple features that are specific for detecting some particular type of forgery. Recently, works on forgery detection based on neural networks have proved to be very efficient in detecting image forgery. Neural networks are capable of extracting complex hidden features of an image, thus giving better accuracy. Contrary to the traditional methods of forgery detection, a deep learning model automatically builds the required features, hence it has become the new area of research in image forgery. The paper initially discusses various types of image forgery techniques and later on compares different approaches involving neural networks to identify forged images.**

***Index Terms*—Copy move forgery, Image splicing, Convolutional neural network, Deep neural network**

I. INTRODUCTION

The emergence and widespread usage of social media platforms like Instagram, WhatsApp, Youtube, etc. have increased the number of images being uploaded and shared on these platforms. These images are used to spread information to a widespread audience and hence helps to form public opinion on a large scale. Due to this underlying property of forming mass opinions and easily available photo and video editing tools, many false images are circulated on these platforms daily. Not only on social media, but tampered images are also used in courtrooms, scientific journals, literature works, etc. Image forgery is a term that refers to manipulating or tampering the original image to hide some useful information or to showcase some false information. The purpose behind creating forged images could range from earning money, spreading rumors, or making false claims in one's favor. Due to the widespread use of images and serious consequences of forged images, many types of research have been done to detect manipulated images. Active approaches for forgery detection which include digital signatures and watermarking suffer from the drawback of inserting the watermark or the signature beforehand in the image which limits the scenarios where this technique would work. Various approaches have been proposed for passive based forgery detection. Popescu Farid [10] described a method for detecting duplicate regions present in an image by applying PCA to blocks of images and then sorting the blocks lexicographically. Amerini et al. [1] has used a SIFT based approach that solves the two-fold purpose of detecting the copy move forgery and to fetch the geometric transformation applied to build the forged image. Wei et al. [12] proposed a rotation angle estimation method that can detect splicing forgery. The algorithm is capable of blind detection of geometrical operations performed and to detect foreign areas present in the forged image. Ke et al. [7] used a technique to detect forged images by determining the consistency of shadow. It assumes that the tampered image has an inconsistent shadow due to splicing. The method has a shortcoming that it only works for forged images with shadow. Another method to identify forgery in images with reflections is developed by Farid [9] which detects geometrical inconsistencies based on 3-Dimensional geometry, reflection, and projection rules. Most of these methods involve the use of structural or geometrical properties of images to extract features. Image forgery detection methods that involve the use of deep learning techniques have become very popular nowadays. These approaches do not consider any statistical or geometrical calculations but tend to construct a model that automatically extracts useful features to perform the classification. The use of convolution neural networks has shown very good results over image manipulation detection. The paper presents an overview of basic concepts of image forgery and various research methodologies involving deep learning adopted so

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:31:53 UTC from IEEE Xplore. Restrictions apply.

far for image forgery detection. The paper is divided into the following sections. Section 2 discusses various techniques of image forgery detection. Section 3 presents an overview and general structure of the Convolutional Neural Network (CNN). Various research approaches for manipulation detection using CNN are compared and contrasted in section 4. The conclusion of the paper is presented in Section 5.

II. IMAGE FORGERY TECHNIQUES

To check an image for its originality, it is important to understand various existing techniques of image forgery. These techniques are broadly classified into active and passive techniques that are further classified.

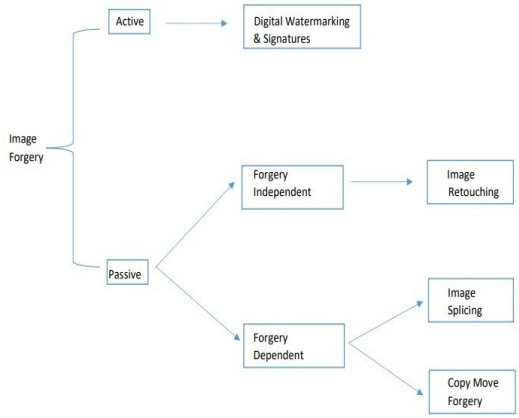


Fig. 1. Image forgery techniques

A. Active approach

In the active approach of forgery detection, the image is pre-processed and some cipher key is inserted in the image. The received image is authenticated by the use of that key [11]. Some examples of active approaches include digital watermarking and signatures.

B. Passive approach

In this type of forgery detection method, no pre-processing of the image is required. It is based on the fact that tampering the original image results in producing inconsistencies within its statistical properties or pixel intensities. These inconsistencies are detected to find whether the image has tampered or not. This technique is more popular than an active approach because of the non-requirement of any prior information of the image. This is further classified into two types-forgery independent and forgery dependent. Forgery dependent methods detect a specific type of forgeries whereas forgery independent methods detect forgery independent of their type.

1) Forgery dependent: These techniques are classified into copy-move and image splicing.

Image splicing

Splicing refers to cutting out a part of one image and putting it in the other, thus the tampered image is produced by a combination of two images. The produced image has two image sources. However, more than two image sources can also exist.



Fig. 2. Image Splicing

Copy move forgery

As the name itself indicates, this type of forgery involves copying one part of image and pasting it in some other location. The purpose is mainly to hide some important information in the image.

2) Forgery independent: Image retouching

This is the most widely used and most easy to do form of image tampering wherein the features of the image are changed in order to produce tampered image. This may involve rotation, resizing, changing the brightness, sharpness, color contrast etc of the image.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:31:53 UTC from IEEE Xplore. Restrictions apply.



Fig. 3. Copy move forgery



Fig. 4. Image Retouching

III. BASIC CNN ARCHITECTURE

CNN has recently become very popular due to their ability to learn the features of an image. They are widely being employed in image recognition and image processing tasks. Hence, many types of research for identifying image forgery are based on CNN.

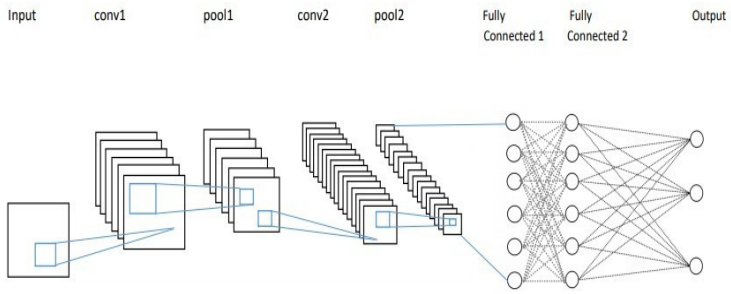


Fig. 5. Basic CNN Architecture

CNN has recently become very popular due to their ability to learn the features of an image. They are widely being employed in image recognition and image processing tasks. Hence, many types of research for identifying image forgery are based on CNN. Most CNN architectures are formed by a combination of one input and output layer, few convolution layers, and fully connected layers. Fig 5 shows a basic CNN architecture. The input image is visualized as a matrix of pixels. The matrix values are fed to the input layer which is composed of many units. Before feeding the matrix values to the input layer, these values are multiplied by the filter values and then summed up. The filter slides over the input image to produce the input matrix. While sliding, the corresponding cells in the input matrix and filter are multiplied and their sum is fed into the output matrix. After this, a pooling operation is done. The pooling can be max-pooling, average-pooling, or global-pooling. The process is shown in fig 6.

In this way, the feature map is generated which is fed into the input layer and the output is generated by applying the activation function to the input. Activation functions such as relu, sigmoid, tanh, etc are widely used. CNN uses a pooling layer which reduces the dimension of the data and serves the dual purpose of reducing the enormous amount of output to be fed into the next input layer and by losing some information, the problem of overfitting is solved. After passing through a series of hidden layers and pooling layers, the data goes into multiple fully connected layers where classification is performed. Finally, in the output layer, the error is calculated which is then propagated backward in the network to update the weights of the filter. Thus, a series of feedforward and backpropagation is done to minimize the error and train the network.

IV. COMPARATIVE STUDY OF FORGERY IDENTIFICATION TECHNIQUES USING NEURAL NETWORKS

The research method involves the use of CNN for the detection of spliced, retouching, and recompressed images [6]. The CNN

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:31:53 UTC from IEEE Xplore. Restrictions apply.

TABLE I

Paper	Targeted Forgery	Features	Layers	Dataset	Performance
Huang et al., 2018 [6]	Splicing, Retouching, and Recompressing	Softmax Classifier	5 Convolution Layers 2 Fully Connected Layers	CASIA v1.0	Accuracy 95%(Spliced) 93%(Retouched) 71%(Recompress)
Kim & Lee (2017) [8]	Median Filtering, Gaussian Blurring, AWGN, Image Resizing	High Pass Filter	2 Convolution Layers 2 Pooling Layers 2 Fully Connected Layers.	1000 Images of Boss Base 1.01	Accuracy:95%
Bayar & Stamm (2017) [2]	Median Filtering, Gaussian Blurring, AWGN, Resampling	Prediction Error Filters	1 newly proposed 2 convolution layers 2 max pooling 3 fully connected layers.	Collected from 12 different camera models.	Accuracy:99.10%
Zhang et al. (2018) [13]	Distinguishes forged and original edges of images	Works well on low-resolution images.	2 Convolution Layers 2 Fully Connected Layers	CASIA 2.0	Accuracy: 85.35%(JPEG) 82.93%(TIFF)
Hema Rajni (2019) [5]	Copy move and Splicing	Uses BDCT and ZM Polar	—	CASIA v1.0 CASIA v2.0	Accuracy(99.03%) TPR(98.91%) TNR(99.16%)
Zhou et al. (2017) [14]	Splicing	Effective in JPG Compression	7 Convolution Layers 2 Pooling Layers	CASIA v1.0 CASIA v2.0 Columbia Image Forgery Database	Accuracy: 97.62%(CASIA v1.0) 97.87%(CASIA v2.0) 96.38%(Columbia dataset)

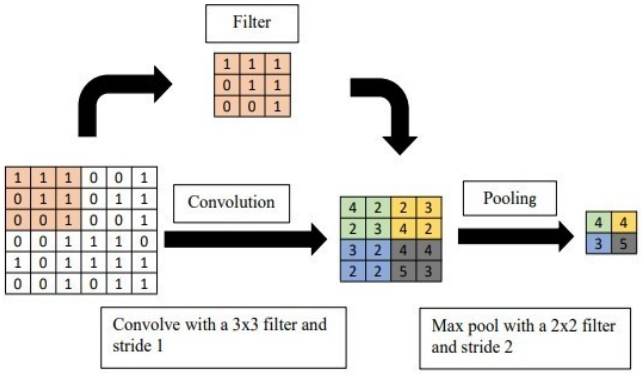


Fig. 6. Convolution operation and max-pooling

architecture is composed of 5 convolution layers, 2 fully connected layers along with a softmax classifier. For gradient descent, stochastic gradient descent with momentum (SGDM) optimizer was chosen. The method observes an accuracy of 95%, 93%, and 71% accuracy in detecting spliced, retouched, and recompressed images when trained and tested over CASIA v1.0 [3] dataset. The dataset was divided into training and testing set and accuracy was determined as how many percentages of testing images were classified correctly. However, some researchers [2, 4] suggest using only CNN for image forgery detection is not effective because CNN only learns semantics or the content of the image. It is unable to identify forged images that do not change the content of the image. For example, when forged images are produced by algorithms such as Median Filter, Gaussian Blur, Additive white Gaussian noise addition (AWGN), and Re-Sampling in which contents of the image remain the same, using CNN alone doesn't give the desired results. This effect was observed in [4] where CNN alone was not able to detect forged images produced by median filtering. Hence, some researches [8,2] use additional filters or additional layers along with CNN to extract the hidden features in an image. One approach using a High Pass filter [8]. A high pass filter, which is capable of extracting hidden features from the image is applied to the input image before feeding it into the input layer of CNN. Also, the Local response normalization (LRN) layer is applied after the pooling layer to standardize the brightness. Another approach involves using a new convolution layer along with the traditional convolution layer [2]. Manipulation of an image changes the local structural relationships of the pixels if not the content of the image. Hence, a new convolution layer is used to draw out those structural features which are independent of the image's content. Bayar & Stamm [2] used prediction error filters to specifically draw out manipulation detection features.

The above approaches work well on high-resolution images but are not as effective on low-resolution images because these approaches rely on sharpness of edges or resampling features [13]. Thus, a novel approach based on fast shallow CNN is proposed in [13] for detecting the boundaries of forged images in low-resolution images. The main difference has tampered and original low- resolution images are in their chroma and saturation. Since the research focuses only on the edges, the CNN model has composed of two convolutions and two fully connected layers along with skipping the pooling layer. Also, 20% of the filters use Laplacian kernels to extract resampling features and information from edges. To localize the tampered region, the basic approach is to use sliding window detection which extracts patches from the original image and feed them into the Shallow Convolutional Neural Network (SCNN), but this has a high time complexity, therefore fast SCNN approach as discussed in [13] feeds entire image in convolution layers to get feature vectors and then extract patches from the feature vectors to input into fully connected layers and constructing probability map. However, this kind

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:31:53 UTC from IEEE Xplore. Restrictions apply.

of approach works for only low-resolution images.

Another approach proposes the use of CNN combined with ZM-polar (Zernike moment) and block discrete cosine transform (BDCT) in forgery detection and classification of the type of forgery [5]. The image is converted to YCbCr form and to which de-correlation and BDCT are applied to generate feature vector set for the first CNN which classifies the image as original or forged. Another CNN model is trained to classify the image as copy-move or spliced. In the case of copy-move, features are extracted using a circular hough transform (CHT), and the CNN model is trained using the patch method. Zhou et al. [14] have used a block-based rich model convolution neural network (rCNN) method for spliced images. The entire image is cut into blocks and these blocks are fed to an rCNN model for processing. The rCNN model consists of 7 convolution layers to extract the desired features. The processed blocks are combined to form a feature vector of the original image. To solve the problem of different sizes of reconstructed features and loss of information during reconstruction, resizing and pooling operations are performed. The feature matrices are then reshaped to a one-dimensional feature vector which acts as input to the SVM classifier to classify images as forged or original. This method also proves its robustness for JPEG compressed images.

V. Results and Discussion

The major outcomings of our analysis of different research papers are:-

1. Deep Learning methods perform far better than traditional methods as they can extract hidden features, thus they give higher levels of accuracy.

2. These methods are flexible and can be modified a little to be used for different types of forgery detection.

A summary of review papers on forgery detection using deep learning techniques is presented in Table 1.

VI. Conclusion

The paper presented the classification of various types of image forgery techniques. Traditional approaches of forgery detection use handcrafted features for tampering detection. Instead, newer approaches based on deep neural networks (DNN) can outperform traditional methods due to their ability to extract complex features from the image. The paper discusses basic CNN architecture which is used by most deep learning approaches. Further, a comparative analysis of various deep learning methods, their effectiveness, and their limitations are discussed. Some methods classify the image as original or tampered whereas some methods can detect the tampered region also. Hence, we conclude that while DNN approaches have their shortcomings and need further improvement, but still they can perform better than the traditional approaches. However, most of the deep learning approaches detect an only specific types of forgeries. Hence, in the future, there is a need to build more generalized algorithms that can detect various types of forgeries simultaneously.

References

[1] Amerini, I., Ballan, L., Caldelli, R., Del Bimbo, A., Serra, G. (2011). A SIFT-based forensic method for copy-move attack detection and transformation recovery. IEEE Transactions on Information Forensics and Security, 6(3 PART 2), 1099–1110. https://doi.org/10.1109/TIFS.2011.2129512

[2] Bayar, B., & Stamm, M. C. (2016). A deep learning approach to universal image manipulation detection using a new convolutional layer. IH and MMSec 2016 - Proceedings of the 2016 ACM Information Hiding and Multimedia Security Workshop, 5–10. https://doi.org/10.1145/2909827.2930786

[3] Box, P. O. (2013). CASIA IMAGE TAMPERING DETECTION EVALUATION DATABASE Jing Dong , Wei Wang and Tieniu Tan Institute of Automation , Chinese Academy of Sciences. 422–426.

[4] Chen, J., Kang, X., Liu, Y., & Wang, Z. J. (2015). Median Filtering Forensics Based on Convolutional Neural Networks. IEEE Signal Processing Letters, 22(11), 1849–1853. https://doi.org/10.1109/LSP.2015.2438008

[5] Hema Rajini, N. (2019). Image forgery identification using convolution neural network. International Journal of Recent Technology and Engineering, 8(1 Special Issue 4), 311–320.

[6] Huang, N., He, J., & Zhu, N. (2018). IEEE International Conference On Big Data Science And Engineering A Novel Method for Detecting Image Forgery Based on Convolutional Neural Network. 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), 1702–1705. https://doi.org/10.1109/TrustCom/BigDataSE.2018.00255

[7] Ke, Y., Qin, F., Min, W., & Zhang, G. (2014). Exposing image forgery by detecting consistency of shadow. The Scientific World Journal, 2014(d), 1–9. https://doi.org/10.1155/2014/364501

[8] Kim, D. H., & Lee, H. Y. (2017). Image manipulation detection using convolutional neural network. International Journal of Applied Engineering Research, 12(21), 11640–11646.

[9] O'Brien, J. F., & Farid, H. (2012). Exposing photo manipulation with inconsistent reflections. ACM Transactions on Graphics, 31(1), 1–11. https://doi.org/10.1145/2077341.2077345

[10] Popescu, A. C., & Farid, H. (2004). Exposing Digital Forgeries by Detecting Duplicated Image Regions. Technical Report, TR2004-515, Department of Computer Science, Dartmouth College, Hanover, New Hampshire, 2000, 1–11. http://os2.zemris.fer.hr/ostalo/2010marceta/Diplomski_files/102.pdf

[11] Prayla Shyry, S., Meka, S., & Mogant I, M. (2019). Digital image forgery detection. International Journal of Recent Technology and Engineering, 8(2 Special Issue 3), 658–661. https://doi.org/10.35940/ijrte.B1121.0782S319

[12] Wei, W., Wang, S., Zhang, X., & Tang, Z. (2010). Estimation of image rotation angle using interpolation-related spectral signatures with application to blind detection of image forgery. IEEE Transactions on Information Forensics and Security, 5(3), 507–517. https://doi.org/10.1109/TIFS.2010.2051254

[13] Zhang, Z., Zhang, Y., Zhou, Z., & Luo, J. (2018). Boundary-based Image Forgery Detection by Fast Shallow CNN. Proceedings - International Conference on Pattern Recognition, 2018-Augus(February), 2658–2663. https://doi.org/10.1109/ICPR.2018.8545074

[14] Zhou, J., B, J. N., & Rao, Y. (2017). Block-Based Convolutional Neural Network. 65–76. https://doi.org/10.1007/978-3-319-64185-0

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:31:53 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Instant Access Memory Design based on an FPGA

Saharsh Aggarwal
Electronics and Communication Engineering
Delhi Technological University,
New Delhi, India
saharsh.aggarwal@gmail.com

Deva Nand
Electronics and Communication Engineering
Delhi Technological University,
New Delhi, India
devkamboj07@gmail.com

Megha Rayer
Electronics and Communication Engineering
Delhi Technological University,
New Delhi, India
megharayer@gmail.com

Saket Kr. Jha
Electronics and Communication Engineering
Delhi Technological University,
New Delhi, India
jhasaket139@gmail.com

***Abstract*— Since the days when they were only a small bunch of logic gates that could design any Boolean Function, Field Programmable Gated Array (FPGA) have come a long way and evolved into many complex chips. The FPGAs have now become a very essential part of the semiconductor and Very Large Scale Integration industry. This paper presents a memory design, based on FPGA which can locate the address required by CPU almost instantly. This memory design works efficiently in Memory intensive jobs also. In this paper, the same hardware architecture has been designed and synthesized with different memory sizes and then they are compared by analyzing the number of lookup tables (LUTs) they use, their on-chip power and the delays associated with them.**

Keywords—FPGA, Decoder, LUT, VLSI, on-chip power, delay, memory.

I. INTRODUCTION

An FPGA is an integrated circuit which can be configured and programmed after manufacturing. It can be programmed according to the requirements of the designer or the customer. FPGAs are made of numerous Programmable Logic Blocks that are wired together to make up the functionality desired by the designer. For instance, FPGAs can have about 10 million transistors on a chip which measure only about 1.8 cm x 1.5 cm x 0.5 microns [1]. FPGAs are reprogrammable which makes them cost-efficient and this coupled with their high integration density makes them suitable for use in the VLSI industry [2][3]. In this paper, the configuration of FPGA is specified using Verilog-HDL on Xilinx- Vivado Design Suite.

Memory provides a means for storing binary data. It consists of storage locations in which numeric or alphanumeric information may be stored. It can also store programs which are sets of instructions that a computer executes to achieve the desired result.

Memory is one of the largest fields in the semiconductor industry. Three types of memory contribute to a major part of the global memory revenue. They are Static Random Access Memory, Dynamic Random Access Memory, and NAND flash memory. They target applications of on-chip cache, main memory, and massive data storage respectively [4][5].

All kinds of memory consist of memory locations to store data and each such location is associated with an address which is unique to that particular location. Every memory access takes place in the following order:

- The CPU puts the desired address on the address bus
- The CPU generates the *Rd* (or *Wr*) signal depending on whether it wants to read from (or write into) the memory.
- In case of a write operation, the CPU puts the data it wants to write on the data bus.
- The Memory detects the *Wr* signal and interprets it as a Write operation. Consequently, it stores the provided data in the address also provided by the CPU
- In case of a read operation, the Memory detects the *Rd* signal and interprets it as a Read operation. Consequently, it searches for the address provided by the CPU and sends back the corresponding data.
- The CPU then reads the data from the data bus.

If the memory has n address bits, then it will have 2^n different memory locations. For example, a 32K byte memory needs to have an address bus of width 15 bits and each memory location needs to have a capacity to store 8 bits. As the memory size increases, the number of LUTs employed in the FPGA and eventually the on-chip power also increases.

In this paper, memory is implemented as an array of block cells. Each cell has a unique address and is located using Row and Column Decoders. In section II of this paper, the modules used in the design are shown and discussed while section III provides details of the methodology used for memory design. Section IV and V presents synthesis reports and simulation results respectively followed by conclusion in section VI.

II. MODULES USED IN DESIGN

A. Constructing Memory from Scratch

The procedure to make up Memory from scratch used in this paper is shown in Fig. 4. Three modules namely decoder, memory_cell and tri_state_buffer are made and all of them are instantiated in the top module called memory.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:59 UTC from IEEE Xplore. Restrictions apply.

All the modules have their respective functions and together they constitute memory which provides instant access.

B. Decoder

The decoder converts binary data to their decimal equivalent. Each possible combination of input is assigned a unique line on the output. So if there are N input lines the decoder will have 2^N output lines. Decoders can vary in size as 2x4 Decoder, 3x8 Decoder, 4x16 Decoder and so on [6][7][8].

In our memory design, two instantiations of *Decoder* are used: One as a Row Decoder and the other as Column Decoder. The memory is designed as a 2-Dimensional Array of *Memory Cells*. For a memory of N^2 size, each Row Decoder output line is connected to N different *Memory Cells* in the same row and similarly, each Column Decoder output line is connected to N different *Memory Cells* in the same column. This way each *Memory Cell* is connected to a unique pair of Row Decoder output line and Column Decoder output line.

The MSB half of the address bits are fed to the Row Decoder and the LSB half of the address bits are fed to the Column Decoder.

Fig. 1. Shows the diagram of a *Decoder*. All the N input lines are uniquely mapped to 2^N output lines.

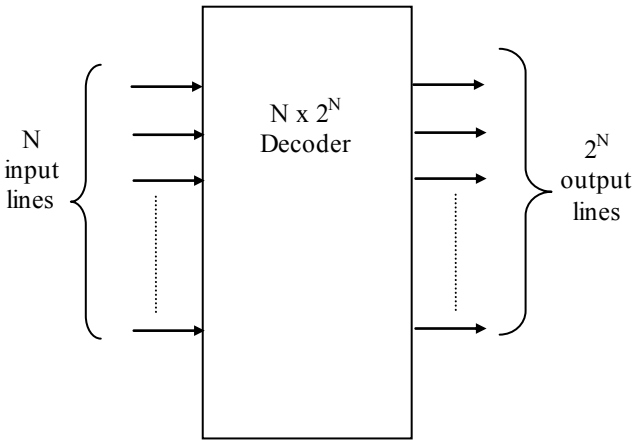


Fig. 1. Decoder

C. Memory Cell

Memory Cell is the basic, most fundamental building block of our memory design. It stores 8-bit data and has certain input signals, *row* and *column*, which when enabled, activates the Cell. Then depending on the *Rd or Wr* signal, Read or Write operation takes place. The signals *row* and *column* are connected to a unique pair of Row Decoder output line and Column Decoder output line. The *Memory Cell* also has an *Enable* output which is used to activate the *Tri-State Buffer* connected to the output of this particular cell. The module *Memory Cell* also has an input *data_in* which is used during a Write operation to receive data to be stored and an output *data_out* which is used to show the stored data during a Read operation. The *Memory Cell* stores data in a register called *data* which can store 8 bits. This way, each *Memory Cell* can store 8 bits i.e. 1 byte. So the number of *Memory Cells* in the memory decides the size of the memory. For example, 4096 *Memory Cells* make up a memory of size 4K bytes.

Fig. 2 shows the block diagram of *Memory Cell*.

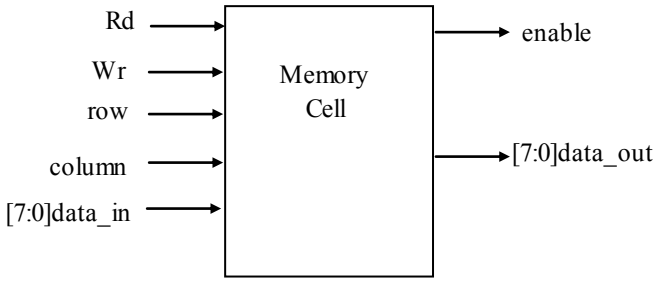


Fig. 2. Memory Cell

D. Tri-State Buffer

Tri-State Buffers act as normal buffers when their enable pin is high i.e. they give output equal to input and when their enable pin is low, they let their output pin float at high-impedance. *Tri-state Buffers* can be used to drive a port through multiple ports [9][10][11].

The outputs of each *Memory Cell* is assigned to the final output port through individual *Tri-State Buffers*. The *enable* output pin of each *Memory Cell* is used as the *enable* pin of their *Tri-State Buffer*.

Fig. 3 shows the *Tri-State Buffer*.

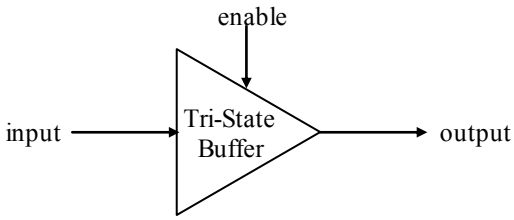


Fig. 3. Tri-State Buffer

III. METHODOLOGY

In this paper, a 2-Dimensional Memory Design using Row and Column Decoders is presented. This design does not make use of a clock which is why it executes the operation almost instantly. Fig. 4 shows the flowchart of the procedure to make such a 2-Dimensional Memory Design. The detailed procedure is explained below.

A. Create Respective Modules

The Modules of *Memory Cell, Decoder* and *Tri-State Buffer* are made in Verilog.

B. Instantiate Memory Cell

- *Memory Cell* is instantiated as many times as is required using a for-loop inside a generate statement.
- Verilog Generate statements are used to write synthesizable RTL. They can be used to create multiple instantiations or conditional instantiations of a module. Here, a generate statement is used to create multiple instantiations of *Memory Cell* and *Tri-State Buffer*.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:59 UTC from IEEE Xplore. Restrictions apply.

C. *Instantiate Tri-State Buffer*

- Tri-State Buffers are used to drive the final output of the design. The outputs of all the individual *Memory Cells* drive the output of the design through *Tri-State Buffers.*
- A multi-input OR gate could also have been used in place of *Tri-State Buffers.* This OR gate will take all the outputs of individual *Memory* Cells and OR them, but that would increase the hardware requirements of the design. An OR gate would also produce erroneous outputs if the outputs of the inactive *Memory Cells* float.

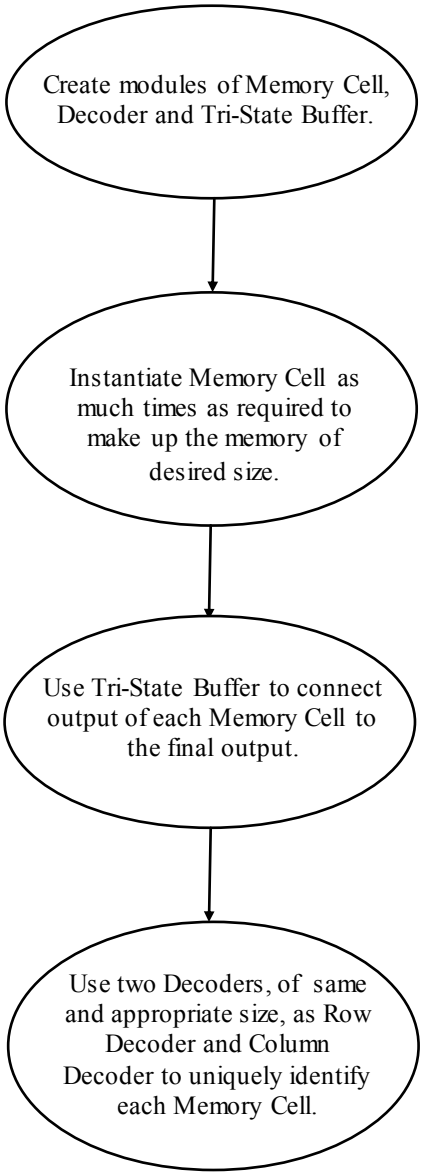


Fig. 4. Flowchart of Procedure

D. *Use Row Decoder and Column Decoder*

- *Decoder* is instantiated twice in the top module as Row Decoder and Column Decoder. The Row Decoder takes the MSB half of the address bits and selects the row of the desired *Memory Cell.* The Column Decoder takes the LSB half of the address bits and selects the column of the desired *Memory Cell.*
- Each *Memory Cell* is associated with a unique row and column pair and hence, together Row Decoder and Column Decoder activates the *Memory Cell* with the desired address.

E. *Block Diagram of 2-Dimensional Memory Design*

Memory Cells are organised in a square array with equal number of rows and columns. A memory of size N^2 is organised by having N rows and N columns. Each row is connected to one of the outputs of Row Decoder and each column is connected to one of the outputs of Column Decoder [12].

Fig. 5 shows the block diagram of 16 Byte 2-Dimensional memory. According to requirements, memory space can be increased by increasing the number of *Memory Cells* and the size of the *Decoder.*

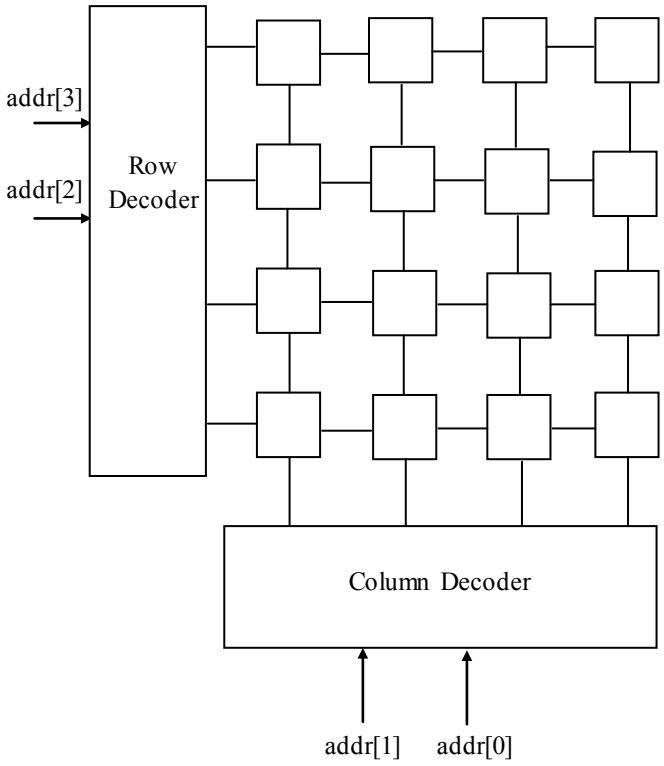


Fig. 5. Block Diagram of 16 Byte (4x4) 2-Dimensional Memory Design

IV. SYNTHESIS REPORTS

In this paper, different sizes of memory are synthesised and then their delays, on-chip power and number of LUTs used are compared. The device used is xc7a200tffg1156-3 (active) which belongs to the family: Artix-7, package: ffg1156, speed grade: -3.

Table 1 gives a clear comparison between different memory sizes.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:59 UTC from IEEE Xplore. Restrictions apply.

TABLE I. SYNTHESIS RESULTS

Memory Size (in bytes)	Number of LUTs used	On-Chip Power (in Watt)	Maximum Delay (in ns)
16	192	1.304	6.712
64	778	3.554	6.712
256	2840	15.561	9.216
1024	11160	60.566	9.216
4096	42543	215.184	9.216

Fig. 6. Synthesis Results

A. *Number of LUTs used*

LUT stands for Look-Up Table. A Look-Up Table can be imagined as a customised truth table which is loaded when the FPGA is powered on with values that are relevant to the functionality required.

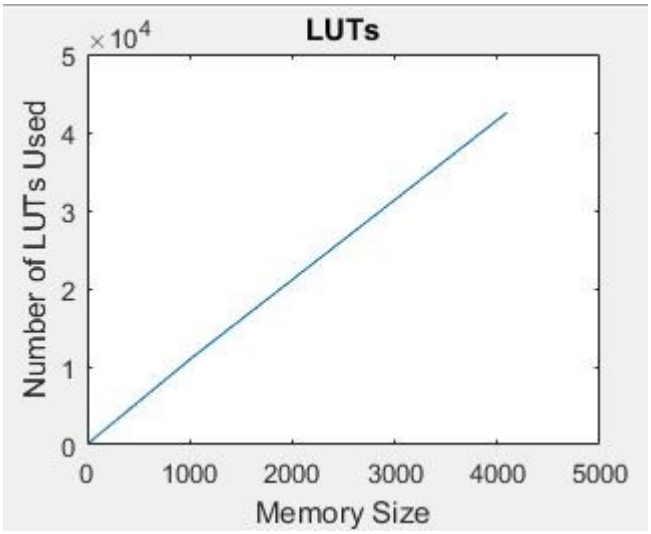


Fig. 7. Number of LUTs used vs Memory Size

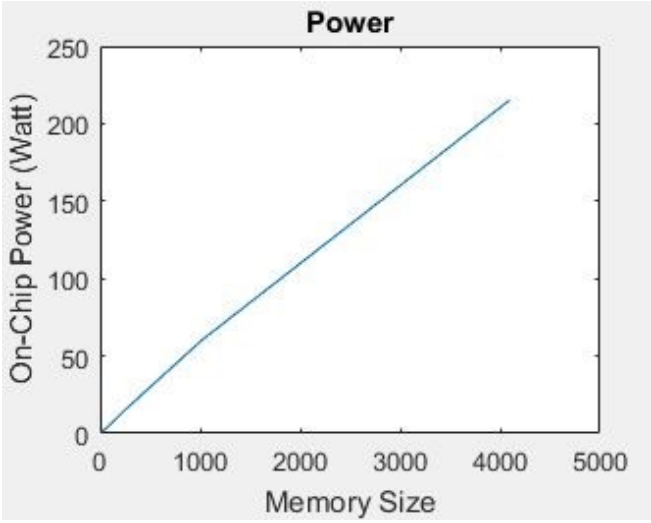


Fig. 8. On-Chip Power vs Memory Size

Fig. 7 shows an almost linear relationship between the number of LUTs used and Memory Size.

B. *Total On-Chip Power*

On-Chip Power shows the power consumed within the device. Fig. 8 shows an almost linear relationship between On-Chip Power and Memory Size.

C. *Maximum Delay*

This shows the Delay involved in the longest path in the design. Fig. 9 shows that Maximum Delay first increases then becomes constant as the Memory Size increases.

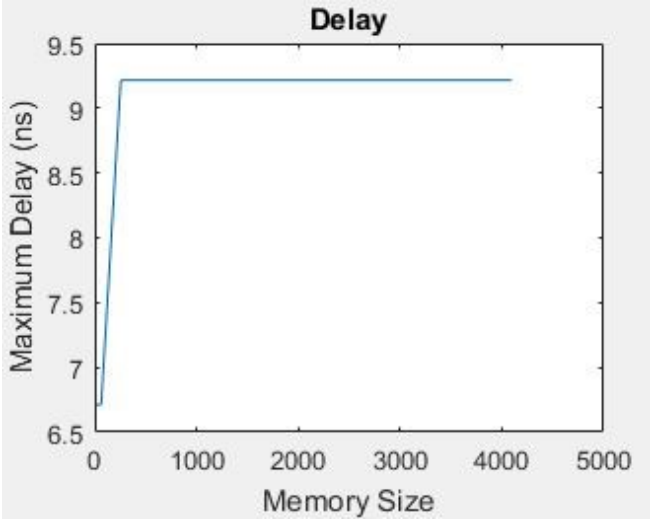


Fig. 9. Maximum Delay vs Memory Size

V. SIMULATION RESULTS

This section shows the simulation results for Read and Write operations. It also shows that both the Read and Write operations were performed instantly.

Fig. 10 shows the Write Operation. At 10ns, the *Wr* signal went high and the *addr* signal was assigned the address of the *Memory Cell* that was to be written. At the same instant, the *data* register of this particular *Memory Cell* was updated with the value that was sent to the Memory via *data_in* signal and the Write Operation was successfully executed.

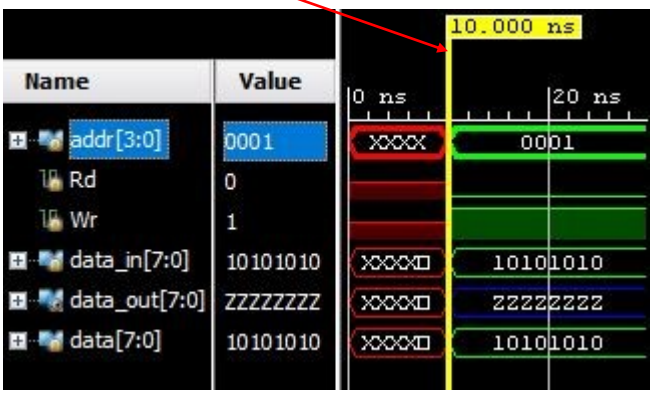


Fig. 10. Write operation

Fig. 11 shows the Read Operation. At 50ns, the Rd signal went high and the *addr* signal was already assigned the address of the *Memory Cell* that was to be read. At the same instant, the *data_out* of the Memory was assigned the value stored in the register of this particular *Memory Cell* and the Read Operation was successfully executed.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:59 UTC from IEEE Xplore. Restrictions apply.

Fig. 12 shows the schematic of a 4-byte memory.

VI. CONCLUSION

Memory design based on FPGA which can locate the address required by CPU almost instantly is presented in this paper. This proposal of memory design works efficiently in memory-intensive jobs also. In this proposal, a hardware architecture has been designed and the same is synthesized with different memory sizes which are then compared by analysing the number of lookup tables (LUTs) used, on-chip power and the delays associated. This comparison is then reported for more clarity.

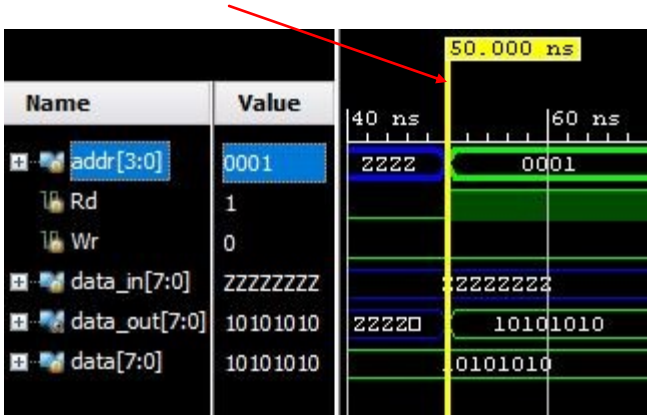


Fig. 11. Read Operation

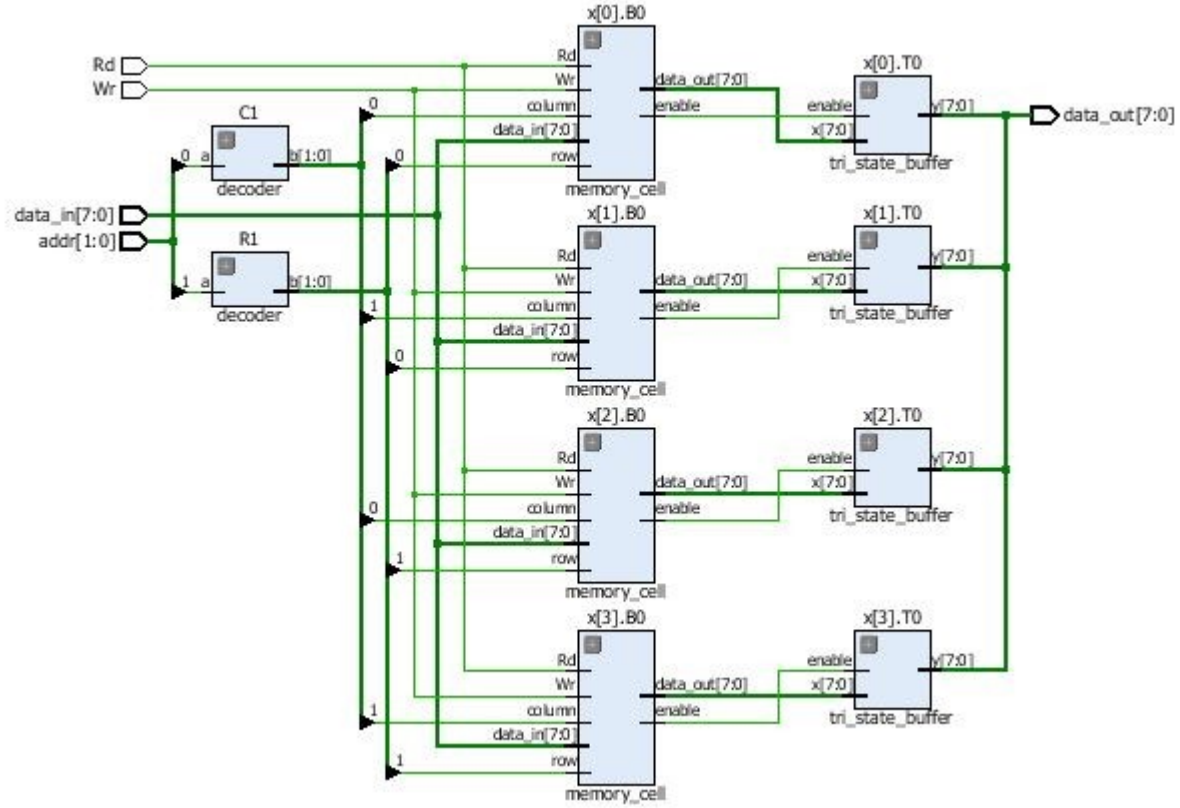


Fig. 12. Schematic of proposed memory design

REFERENCES

[1] S. Brown and J. Rose, "FPGA architectural research- Survey" in IEEE Design & Test of Computers, vol. 13, issue 4, 1996, p.42-57

[2] E. Monmasson and M. N. Cirstea, "FPGA design methodology for Industrial Control Systems" in IEEE transactionson Industrial applications, vol. 54, issue 4, August 2007, p.1824-1842.

[3] Atitallah, R. B., Viswanathan, V., Belanger, N., & Dekeyser, J.-L. (2018). FPGA-Centric Design Process for Avionic Simulation and Test. IEEE Transactions on Aerospace and Electronic Systems, 54(3), 1047–1065.

[4] Y. Chen , H. Helen Li, I. Bayram and E. Eken, "Recent Technology advances of Emerging memories" in IEEE Design & Test of Computers, vol.34, issue 3, June 2017, p.8-22.

[5] Woo Young Choi, "Three-Dimensional Stackable Electromechanical Nonvolatile memory cell(H cell) for four bit operation" in IEEE Electron Device Letters, vol. 31, issue 1, Jan 2010, p29-31.

[6] S. H. Teen, J. H. Lim and L. L. Lim, "IC Layout Design of Decoder Using Electric VLSI Design System" in International Journal of Electronics and Electrical Engineering vol. 3, No. 1, February, 2015, p.54-60

[7] Ming Zhao ,Xiaolin Zhang ,Ling Zhao and Chen Lee, "Design of a High Throughput QC-LDPC Decoder with TDMP Scheduling" in IEEE Transactions on Circuits and Systems II: Express Briefs vol. 62 , issue 1 , Jan. 2015 p.56-60.

[8] Malviya, Deepti & Majumdar, Shubhankar & Mishra, Mayank & Bansod, Prashant. (2012), "Comparative Study on Implementation of Various Decoder Architecture" at ICICES Chennai, February 2012, DOI:13140/2.1.4511.5524.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:59 UTC from IEEE Xplore. Restrictions apply.

[9] Neeta Pandey, Bharat Choudhary, "Improved tri-state buffer in MOS current mode logic and its application", Analog Integrated Circuits and Signal Processing, vol. 84 ,2015 , pp. 333.

[10] Badel, S., & Leblebici, Y. (2007), "Tri-state buffer/bus driver circuits in MOS current-mode logic" 2007 Ph.D Research in Microelectronics and Electronics Conference
doi:10.1109/rme.2007.4401856.

[11] Kirti Gupta, Neeta Pandey, Maneesha Gupta, "Low-power tri-state buffer in MOS current mode logic", Analog Integrated Circuits and Signal Processing, vol. 75 ,2013, pp. 157.

[12] T. Alexoudi, S. Papaioannou, T. G. T. Kanellos, A. Miliou and N. Pleros, "Optical RAM Row Access with WDM-Enabled All-Passive row/column Decoders",in IEEE Photonics Technology Letters, vol. 26, No. 7, April 1 2014, p.671-674.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:59 UTC from IEEE Xplore. Restrictions apply.

IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)
e-ISSN: 2278-1684,p-ISSN: 2320-334X, Volume 17, Issue 3 Ser. II (May - June 2020), PP 01-13
www.iosrjournals.org

Linking TQM Critical Success Factors to strategic goal: Impact on Organizational Performance

Rakesh Kumar[1], R. S. Mishra[2]

***Abstract*:** *This paper aims to examines and establish the relationship between critical success factors (CSFs) of total quality management (TQM), strategic goal and organizational performance. Using key enablers and drivers collectively said CSFs as measures, the suggested model appraises three dimensions of performance: financial, non-financial and operational. A total of 500 questionnaires have been distributed to a FMCGs company in India and effective sample of 165 usable completed surveys (33.0 percent usable response rate). Statistical analysis and correlation were used to predict and estimate the relationship. The model was assessed using Analysis of Moment Structure (AMOS) based Structural Equation Modeling (SEM). The findings showed that TQM practices have a significant impact on financial and operational performance.*

Keywords: *Critical Success Factors, CSFs, Performance, Total Quality Management, TQM, FMCGs*

Date of Submission: 16-05-2020 Date of Acceptance: 31-05-2020

I. Introduction

In examining influences on performance measurement (PM) practices, greater attention must be paid to the existing systems of manufacturing & management that operate within organizations viz inventory management system, maintenance management, quality management system, financial management, supply chain management etc. In various sectors such as manufacturing and service in order to deal with changes in the marketplace and focus on product and service quality total quality management (TQM) practices have been widely acknowledged as disciplined management process. A set of guidelines TQM provides which help to improve the organization performance, G. Muruganantham et al. (2018). The today's dynamic business environment making it necessary to the organization to maintain and develop effective TQM, which allows developing effective ways to improve customer satisfaction with the product quality, providing increased efficiency and competitiveness. For measurement to be effective, it has to be linked to strategy, focusing on processes and be geared towards positive action and improvement. Zairi's (1994). Performance Measurement is to be a top-down approach, where the top level management believes in the process and implements it, on the entire organization. Incorporating the best practices of continuous performance management, first set objective and key results aligned to business strategic goal, second, check on regular basis on performance, third reach out for feedback from stakeholders, then finally evaluate performance & update objective and key results. Performance should always be measured to help to figure out if meeting strategic goals. Zairi's (1994) developed analogy of comparing measurement to the umbilical cord that links a mother to its baby is apt in this context. Mothers, in a similar manner to organizations, have to look after themselves in such a way that whatever they do and whatever they eat, is not going to harm their baby (in this case, the business). The umbilical cord, or (performance measurement), is the mechanism by which the baby grows and the relationship with the mother remains a close one. Michel J.Lebas (1995) expressed the general agreement of people on performance that what performance really means and stated that it can mean anything from efficiency, to robustness or resistance or return on investment, or plenty of other definitions never fully specified. The objectives of any performing firm or system achieves contain (1) *Goal to* be reached, as well as, (2) factors/ elements of *time* at which the target or milestones to that aim are reached and (3) rules about a preference ordering about the *ways to get there.* These three elements indicate that objectives and the definition of performance rest on the definition (choice) of a causal model linking inputs and outcomes through selected causal relationships. Performance is something each firm, each stakeholder, each organizational actor defines. Performance is never objective; it is only a way of defining where one wants to go.

The purpose of this research is to examine and establish relationship between TQM strategic goal with its drivers and enablers, which is vial of the success of TQM. It also looks for a contingent relationship between organizational performance (the outcome variable) and the match between TQM goal and the critical factors variables described above as drivers and enablers. The framework for the research is illustrated in Figure 1. The critical success factors of TQM are to align the measures and metrics to the strategic goal of the organization. The function / department level measures also need to precisely identified which provide adequate insight in to the progress towards achieving / not achieving of the goals & objectives. The next section of the

paper briefly reviews the relevant literature and develops the research hypotheses. The research method applied is described in section three. Section four presents our results. The final section concludes the paper.

Therefore, organizations must align their critical success factors with their strategic goals and to do so, they have to act upon these individuals' factors, in order to improve performance (Figure 1). From a performance perspective of implemented TQM, therefore, once the goal is aligned with the CSFs have been identified, organizations will seek to modify their CSFs and steer it towards the desired performance level with the target competency. For this purpose, performance assessors may have to adapt their previous levels of both declarative and, above all, CSFs, with support from all the level of employee (Figure 1(a) and 1(b)).

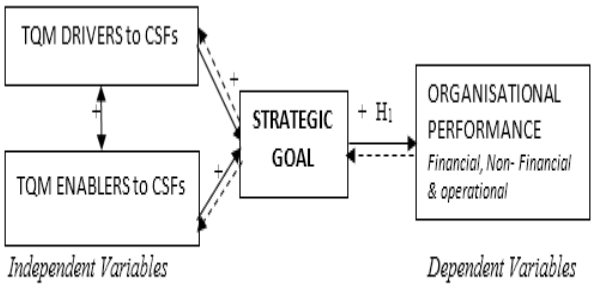


Figure 1 (a)

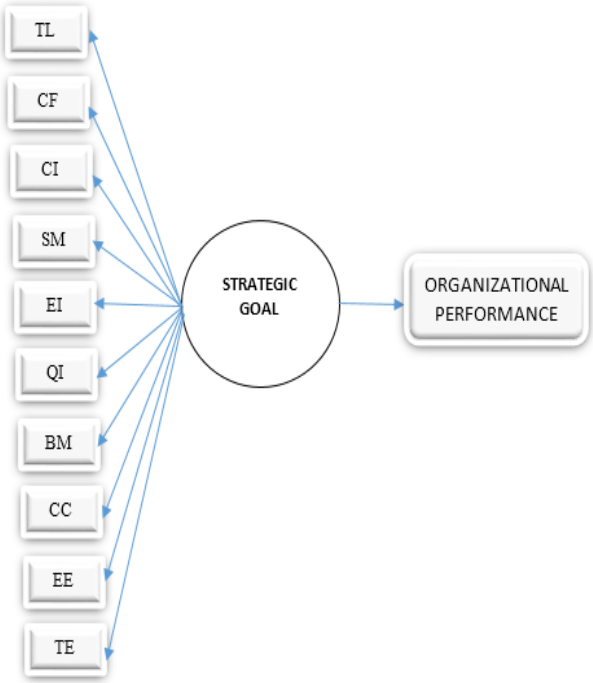


Figure 1 (a) & (b) Framework for the study

II. Literature Review And Hypothesis Development

As a prelude to the study of the relationship between TQM CSFs and organizational performance, it is first necessary to briefly present the literature on each of these concepts to identify the state of the art in the research that specifically examines this relationship.

TQM and Organizational Performance

The trace of the concept of TQM dates back to the early 20th Century when Walter Shewart, in the early 1920s, first introduced the concept of statistical process control (SPC) to monitor quality in mass production manufacturing (Shewart, 1931). Later it was followed by various quality management gurus and practitioners who all advocated with their own views and various approaches to TQM. Crosby (1979), the four absolutes, Deming (1986), fourteen points, Feigenbaum (1993), total quality control, Ishikawa (1985), quality control circles, Juran J.M. (1981), quality triology and Taguchi (1986), loss function, have prescribed different

techniques and organizational requirements for effective implementation of TQM. In a study by Adam (1994) examined multiple quality and productivity approaches are correlated to eight quality, three operating, and three financial performance measures for 187 US business firms. Results indicate a strong relationship between a quality improvement approach and performance quality. The relationship between a quality improvement approach and operating or financial performance is weaker, but significant. Productivity improvement approaches also help predict quality, operating, and financial performance. Depending upon the selected measure(s) of quality, operating and financial performance, an appropriate approach to quality improvement can also be selected. Madu et al. (1996) used an empirical study to test if there are any significant associations between quality dimensions (such as customer satisfaction, employee satisfaction and employee service quality) and organizational performance. The study aimed to validate or refute some of the claims that are largely made by quality experts and practitioners regarding the importance of total quality management to organizational success. Samson and Terziovski (1999) attempted to find the relationships between the various TQM practices, individually and collectively, and company performance of around 1200 Australian and New Zealand manufacturing organizations. The results showed that the intensity of TQM practice does contribute significantly to the performance. In another investigation, Terziovski et al. (1999) tested the relationship between TQM practice and organizational performance with and without the covariates, company size, industry type and ISO 9000 certification status. The authors concluded that there were significant differences in the relationship between TQM and organizational performance across industry type and size, especially on the effect of defect rates, warranty costs and innovation of new products.

TQM Critical Success Factors (CSFs)

The concept of Critical Success Factors (CSFs) (also known as Key Results Areas, or KRAs) was first developed by D. Ronald Daniel, in his article "Management Information Crisis" (Harvard Business Review, September-October 1961).The Critical Success Factors (CSFs) is defined as "the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization" by Rockart (1979) also concluded that CSFs are "areas of activity that should receive constant and careful attention from management". According to manufacturing management point of view CSFs can be derived through a document review and analysis of the goals and objectives of key management personnel, as well as interviews with those individuals about their specific domain and the barriers they encounter in achieving their goals and objectives. So, critical success factors may be also defined as the handful of key areas where an organization must perform well on a consistent basis to achieve its mission.(Brotherton and Shaw, 1996) advocated about the CSFs that they are the actions and processes that can be controlled by the management to achieve the organization's goals.(Rockart, 1979) considered it as a crucial to the success of a program, and if the objectives associated with these factors are not achieved, the application program will also lead to failure. Boynton and Zmud (1984), suggested that the CSFs are "those few things that must go well to ensure success". The critical success factors are essential parts that must be addressed by management or the manager to ensure that 'things must go right' for a project or activity to achieve management objectives and business growth. Critical success factors are the few key areas of activity in which favorable results are absolutely necessary for a particular manager to reach his goals. Critical success factors are the basic structural variables which will most affect the success or failure of the pursuit of strategic goals. CSFs also provide an important engine for performance goals. Munro and wheeler (1980) suggested that managers and organizations should use CSFs to achieve higher performance. (Zairi and Leonard 1994) defined TQM drivers as the critical areas of managerial planning and action that must be fostered in order to achieve effective quality management within a business unit, ensuring its successful implementation. Here short definitions are given to distinguish between them as:

Drivers - drivers are the key inputs and activities that drive the operational and financial results of an organization. Drivers vary significantly by industry, but they can all be determined using the same type of root cause analysis. Drivers impact all financial aspects of a business: revenues, expenses, and capital costs.

Enablers – Enablers are external conditions or organizational strengths that facilitate an organization's ability to accomplish its goals or objectives. Enabler of TQM system to Increase operational efficiency, empower employees to be more effective with a measurable impact on their satisfaction, generate increased revenue and profit, fix problems, increase base or loyalty by delivering improved customer-client relationships, reinvent and reinvigorate brands, products or services, improve the effectiveness of leadership or management and embed ethical or sustainable practices within organization.

A significant number of literatures on TQM critical success factors (CSFs) is available and have been adopted by organizations in their business. Few of them are summarized as

Table 1 TQM Critical Success Factors

Author	Identified Critical Success Factors	Number of factors
Carmona F. J. (2016)	Top management commitment and leadership, TQM philosophy adoption,	10

	Customer involvement, Supplier involvement, Open and flexible structure, Employee education and training, Empowerment, Benchmarking, Process improvements, Zero-defects mentality	
Fotopoulos and Psomas (2009)	leadership; strategic quality planning; employee management and involvement; supplier management; customer focus; process management; continuous improvement; information and analysis; and knowledge and education	09
Kanagi Kanapathy (2008)	Top management support ,Quality information availability, Quality information usage, Employee training, Employee involvement, Product/process design , Supplier quality, Customer orientation	08
Lenka and Suar (2008)	transformational leadership; customer orientation; human resource management; organizational culture; continuous improvement; and quality measurement	06
Vouzas and Psychogios (2007)	Total employee involvement, Continuous improvement, Continuous training, Teamwork, Empowerment, Top-management commitment and support, Democratic management style, Customer/citizen satisfaction, Culture change.	09
Yusuf et al. (2007)	leadership and commitment; customer focus; continuous improvement; get things right first time; just-in-time; competitive benchmarking; cost of quality; employee involvement; teamwork; training; communication; and recognition and reward	12
Al-Marri, Ahmed, and Zairi (2007)	top-management support; customer focus; strategy; benchmarking; employee involvement; recognition and reward; problem analysis; quality technologies; service design; services capes; service culture; social responsibility; human resource management; continuous improvement; quality department; and quality systems.	16
Mahapatra and Khan (2006)	Leadership and top-management commitment; customer focus and satisfaction; policy and strategy planning; human resource management; process management and control; product=service design and control; continuous improvement; supplier management; training; employees satisfaction; employees participation; employee appraisal, reward and recognition; quality culture; quality assurance; quality system; impact on society; teamwork; flexibility; zero defect; and benchmarking.	20
Saravanan and Rao (2004)	Top-management commitment and leadership; benchmarking; customer focus and satisfaction; service marketing; social responsibility; human resource management; employee satisfaction; service culture; service scape; continuous improvement; technical system; and information and analysis	12
Hasan and Kerr (2003),	Top management commitment , Employee involvement, Training, Supplier Quality, Quality cost, Service Design, Quality Techniques, Benchmarking, Customer satisfaction.	09
Sila and Ebrahimpour (2002)	25 CSFs is listed after analyzing 286 of 347 articles , across 76 studies	25
Motwani (2001)	top-management commitment, quality measurement and benchmarking, process management, product design, employee training and empowerment, supplier quality management, and customer involvement and satisfaction	07
Tsang and Antony (2001)	customer focus; continuous improvement; teamwork and involvement; top-management commitment and recognition; training and development; quality systems and policies; supervisory leadership; communication within the company; supplier partnership; measurement and feedback; and cultural change.	11
Kanji and Wallace (2000)	top-management commitment; customer focus and satisfaction; qualityinformationand performance measurement; human resource management; employee involvement; teamwork; process management; quality assurance; zero defects; and communication	10
Sohal and Terziovski (2000),	Leadership support, Staff involvement, Availability of money, Available information, Customer involvement, Supplier involvement, Government assistance	07
Tamimi (1998),	Top management commitment, Supervisory leadership, Education, Cross functional communications to improve quality, Supplier management, Quality training, Product/service innovation, Providing assurance to employees	08
Thiagarajan and Zairi (1997)	Leadership, Employee involvement, Middle management role, Training and education, Rewards and recognition, Teamwork, The role of employee unions, Policy and strategy	08
Black and Porter (1996),	Strategic quality management, corporate quality culture, Quality improvement measurement system, communication of improvement information, Operational quality planning, External interface management, Supplier partnerships, People and customer management, Customer satisfaction orientation	09
Zairi and Youssef (1995)	People and customer management, Supplier partnerships, Communication of improvement information, Customer satisfaction orientation, External interface management, Strategic quality management, Teamwork structures for process improvement, Operational quality planning, Improvement measurement systems, Corporate quality culture	10
Badri et al. (1995)	Role of divisional top management and quality policy, Role of quality department, Training, Product/service design, Supplier quality management, Process management/operating procedures, Quality data and reporting, Employee relations	08

Flynn, Schoeder, & Sakakibara, (1994)	Top Management Support, Quality information availability, Employee Involvement, Process/Product Design, Supplier Quality, Customer Orientation, Process Management	07

Relationship between CSFs and Strategic Goal

Hinterhuber (2004) defines strategy as a way of using the resources and capabilities of an organization, this definition was in the context of business management research and practice. But in general, the use of the term 'strategic' implies that there is an overall goal, or some specific vision concerning the nature of success. All factors and their actions contribute to a common vision, an overall goal. Setting strategic goals facilitates the identification of appropriate business strategies and focuses management attention and available resources on their accomplishment, enabling subsequent goal realization. However, little is known empirically about the extent to which strategic goals enable desired performance to be achieved and factors that may link this relationship. It is often true that the strategic goals of an organization are not achieved due to the complex interactions within the organizational hierarchy.

When managers set goals, they also implicitly consider what they need to do to be successful at achieving the goals. Thus, it is likely that managers consciously consider their CSFs during goal setting and consequently create the bond between goals and CSFs that is needed to contribute to accomplishing the organization's mission. The strong relationship between strategic goals and CSFs results from the fact that managers are the origin of both goals and CSFs. In this way, the influence of CSFs on goal achievement is made explicit, even if the actual CSFs are not. Organizations that have been successful at achieving their goals have also likely achieved their CSFs, albeit in a less observable way. Thus, goals sometimes resemble CSFs because they embody the importance of a key performance area. Usually a goal is immediately discernible from a CSF because of its specificity. A CSF for the organization may be more general and is likely to be related to more than one goal. A successful PM system plays a predominant role in assisting executives track corporate performance to determine the extent to which strategic goals have been reached (Koufteros et al., 2014). Until recently, limited work has been carried out on mechanisms required to link critical success factors with strategic goals in properly managing their performance. Hence, more research needs to be conducted to provide further insight into what integration of organizational systems and practices can help companies in attaining this strategic goal (Cleary, 2015).

Based on the literature review, a conceptual framework is developed to investigate the relationship between selected TQM critical success factors and organizational performance. The framework is depicted in Fig-1. Based on the research objective the research hypothesis is developed as;

Alternate hypothesis(H_1): There are significant positive relationships between TQM CSFs (drivers and enablers as H_{11}, H_{12}, H_{13}, H_{14}, H_{15}, H_{16}, H_{17}, H_{18}, H_{19} and H_{20}) and strategic goal.

III. Research Approach

This research adopted a cross departmental survey of existing FMCGs firm in India. Surveys allow the collection of large amount of data from a sizable population in a highly economical way. The study was used to provide analysis of how various TQM drivers and enablers, which is collectively called critical success factors contributed to the performance improvement in FMCGs firms. The characteristics of the organization differs its tool's critical success factors.

Measurement Instrument. It was decided on Top-management commitment and leadership, Customer focus, Continuous improvement and innovation, Supplier management, Employee involvement, Benchmarking, Culture and communication, Employee encouragement and motivation and Training and education to consider as the critical success factors of TQM practices based on the literature review. Multiple performance factors are also included, namely, operational performance, financial performance and non-financial performance, to cover all aspects of organizational performance. The items of the questionnaire are also prepared with the help of available literature mentioned in table 1. The items included a five point Likert-type scale anchored from (1) strongly disagree to (5) strongly agree, which indicates respondents' opinion with each item, respectively.

With the objective to establish link the TQM drivers and enablers, which emerge from the strategic goal, this study conducted as the algorithm suggested above with strict guidelines pertaining to research. Ensuring research rigor, we conducted this study adhering to strict guidelines pertaining to research.

Population, Sample and Data collection process

A sample of all the employee who were engaged for production and quality department was selected making a total of 165 respondents. The method used was systematic random sampling method because in this method each and every member of the population had an equal chance of being selected as a sample. They were asked to complete the questionnaires. The selected sample was deemed adequate for general conclusions about the entire population. The sample was also adequate for the statistical tools which were used in the data analysis.

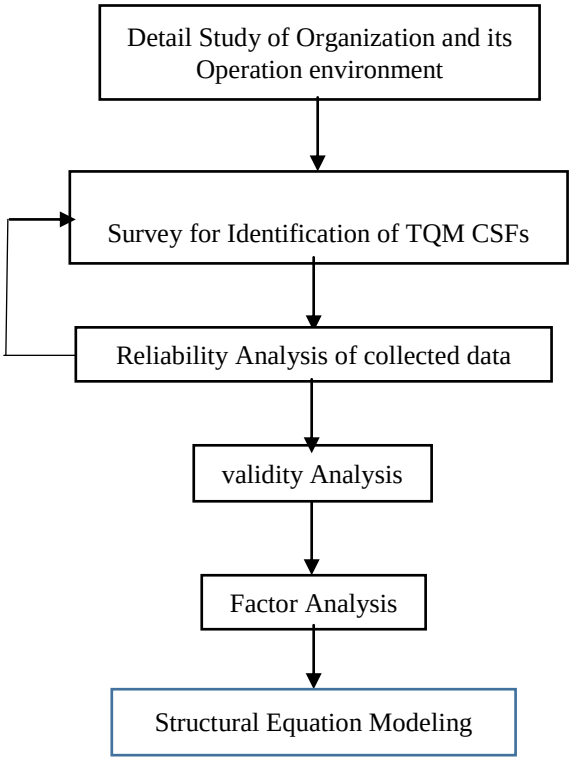


Figure 2. Proposed algorithmic steps for study of TQM CSFs

The questionnaire was considered most appropriate because it allowed for collection of data from many respondents within a short time and provided a high degree of data standardization and adoption of generalized information amongst any populations. The questionnaire consisted of both closed and open-ended questions since this led to control over the data collected. Section A of the questionnaire sought to obtain firm data. Section B sought to obtain data on the TQM drivers and enablers and Section C acquired data on the firm's performance. The respondents filled in the questionnaire as the research assistants awaited. This helped reduce the instances of non-response. Where necessary, the questionnaires were left and picked up after a week and this ensured a high proportion of usable responses.

Linking applicability

This section describes the linking of CSFs with the strategic goal, as well as how it was applied. The authors performed a pre-testing process in steps wise. First, for the purpose of assessing the face validity of the questionnaire, the authors conducted reliability analysis of the collected data. Table 9 provides a summary of this integration. Overall, this exposé provides a valuable application and template for other firms.

Identify TQM CSFs

Identifying Critical Success Factors enable to track and measure the progress toward achieving strategic goals - and, ultimately, to fulfilling organization's mission. CSFs also provide a common point of reference so that everyone knows exactly what's most important, ensuring that tasks and projects are aligned across teams and departments. The identification of critical success factors of TQM, this useful approach at this stage include brainstorming or the nominal group technique, keeping in view the goal of the organization. All feasible drivers and enablers should be identified that could potentially influence the desired outcomes. The overriding objectives of the performance assessment in case were to reduce costs along the total quality management, with however at the same time improving (or at least not deteriorating) performance levels and effectiveness. These objectives were kept in mind when considering potential TQM critical success factors for performance. After a review of potential critical success factors of TQM, we deemed the ten critical success factors as identified earlier by various researchers. These ten critical success factors, as well as their definitions and corresponding abbreviations used in the ensuing discussion, are summarized in Table 2.

Table 2. Identified Critical Success Factors (CSFs) of TQM

	CSFs Label	CSFs Code	CSFs Definition
TQM *DRIVERS*	1)Top-management commitment and leadership	TL	Top management consist of individuals with good leadership skills, who can deliver results, who are committed, who can contribute to growth of the company, who shares the vision of the company .
	2)Customer focus	CF	This meant thinking not in terms of what company sold first, but whom company sold to first. Company asked questions such as what problems do customers face, how do company address them, and what solutions will help.
	3) Continuous improvement and innovation	CI	Continuous improvement, also known as Kaizen. An organization must always continue to innovate its processes. Improvement is a continuous process and there is never an end to it.
	4)Supplier management	SM	The dyadic relationship between a buyer and its supplier, it falls short of addressing the importance of extended networks beyond the immediate dyadic relationship (i.e., how a supplier's relationship with its suppliers or other buying firms affects its performance).
	5)Employee involvement	EI	Employees, also known as internal customers form the organization. Involvement of employees thus can literally make or break mission. Make sure that the employees are motivated and are taking ownership of the objective.
	6)Quality Information /Information Quality	QI	the desirable characteristics of the system inputs; that is, management reports and Web pages. For example: relevance, understandability, accuracy, conciseness, completeness, currency, timeliness, and usability.
	7)Benchmarking	BM	Benchmarking helps organizations to identify the gaps in its performance when compared with another organization
TQM *ENABLERS*	8) Culture and communication	CC	Organizational communication—including managerial, interpersonal and other forms of communication—carries policies, strategies, instructions and information across the organization and tends to play an important role in the relationship between culture and organizational performance. Effective communication at work can be nurtured through an organizational culture that takes into account internal beliefs of employees.
	9) Employee encouragement and motivation	EE	Employee encouragement and motivation and their ability collectively participate into organizational performance and in their tasks assigned by the manger are to purpose get maximum productivity.
	10) Training and education	TE	Training and education as strategies to improve organizational effectiveness and competitiveness

Instrumental at this stage, pertaining to the reliability analysis,was that as much detail as possible for each critical success factors was collected. The opportunity to gather this information at this stage is utilized, since there is usually an underlying rationale for work persons to suggest a particular factor as being important. Specifically, for each identified critical success factor.

Reliability analysis of collected data

Variables derived from test instruments are declared to be reliable only when they provide stable and reliable responses over a repeated administration of the test. Cronbach's alpha is an index

of reliability associated with the variation accounted for by the true score of the "underlying construct." Construct is the hypothetical variable that is being measured (Hatcher, 1994). In this regard, the Cronbach's α coefficient was used to test the reliability of all the constructs and their specific dimensions. α scores for all the main variables exceeded the recommended cut-off point of 0.70 (Nunnally et al., 1967).

The higher the score, the more reliable the generated scale is. Nunnally (1978) has indicated 0.7 to be an acceptable reliability coefficient. The Cronbach's alpha values of the variables are summarized in Table 1 and Table 2. The alpha values ranges from shows good internal consistency of the data collected. The data collected in survey are evaluated for reliability using Cronbach's alpha-coefficient method. The test was conducted in SPSS Statistics v.23 software. Reliability tests were performed separately for the items of each critical success factors (10 independent variables) and dependent variables (financial, non-financial and operational performance).

Validity Analysis

Validity explains how well the collected data covers the actual area of investigation (Ghauri and Gronhaug, 2005). Validity basically means "measure what is intended to be measured" (Field, 2005). The main types of validity namely are; face validity, content validity, construct validity and criterion validity. The approach to establish content validity involves literature reviews, as for this paper and then follow-ups with the evaluation by expert. This study extracted ten prompt critical success factors of TQM for the study listed in table 2. Kerlinger, F.N (1986) argues that content validity is representative of the content. Thus, content

Table 3. Reliability Analysis of CSFs (Input variables)

Input Variables Label Code	CRITICAL SUCCESS FACTORS									
	1	2	3	4	5	6	7	8	9	10
	TL	CF	CI	SM	EI	QI	BM	CC	EE	TE
No of items	4	4	4	3	4	3	4	4	4	4
Cronbach's alpha	0.718	0.720	0.735	0.716	0.728	0.765	0.752	0.723	0.720	0.735

Table 4. Reliability Analysis of Output variables

Output Variables Label	ORGANISATIONAL PERFORMANCE		
	Financial Performance	Non-financial Performance	Operational Performance
Code	FP	NFP	OP
No of items	06	05	06
Cronbach's alpha	0.771	0.724	0.712

validity of an instrument depends on the adequacy of a specified domain of content that is sampled (Nunnally, J.C.). Bush (1985) pointed out that content validity refers to the degree that the instrument covers the content that it is supposed to measure. It also refers to the adequacy of the sampling of the content that should be measured (Polit D.F. et al. 1991). Cronbach (1971) noted, when the items of a test are judged to adequately represent well-defined domains of content, it is permissible to view responses to these items as generalizable samples of the responses examinees would exhibit if they were tested on all of the items constituting these domains. An evaluation of the theoretical and practical soundness of using each of the mentioned notions to define content validity suggested that these notions are best regarded as definitions of concepts other than content validity.

Factor Analysis

Factor analysis is a variable reduction procedure that allows one to explore the interrelationships among variables in a dataset (Dunteman, 1989; Stevens, 1996). Significant clusters of variables, or factors, are identified through optimally weighted linear combinations of observed variables that maximize the amount of explained variance. First the factor analysis was carried out to reduce the TQM drivers, enablers and strategic goal scales to a smaller number of underlying factors. For this principal component analysis was used to identify factors with eigenvalues of at least 1, Hair JF el al. (1992), and Varimax rotation was used to obtain interpretable factor loadings. Bagozzi R.P. et al. (1988) advocated in the interests of convergent and discriminant validity, only items that had a factor loading of at least 0.50 and did not have a loading in excess of 0.40 on a second factor were retained.

Reliability Analysis

Applying SPSS, the principal components analysis (PCA) was conducted to measure the underlying dimension of ten constructs. The constructs validity was measured using Bartlett's test of Sphericity and Kaiser–Mayer–Olkin (KMO) measure of the sampling adequacy of individual variables. KMO overall should be 0.6 or over to perform factor analysis (Ozdamar, 2002). According to the results of Bartlett's test of Sphericity and KMO revealed that both are significant and suitable for the factor analysis (Table 2). The cumulative variance explained is 66.59% which exceeds the acceptable limit of 60% (Ozdamar, 2002). The value of Bartlett's test of Sphericity indicate sufficient correlation between the variables, it shows 4333.93 and significant ($p > 0.000$). The factor loading of all items of each scale exceeds 0.5 (Hair, Black, Babin, Anderson, & Tatham, 1998). This data analysis demonstrates that measurements possessed an acceptable convergent validity. The composite reliability of the measurements must reach 0.6 or above (Fornell & Larcker, 1981). The results indicated that all the latent variables reached (0.817–0.88) the standard or above.

Table 5

KMO Bartlett's test.	
Kaiser–Mayer–Olkin measures of sampling adequacy	0.785
Bartlett's Test of Sphericity	

Approx. chi-square	4305.939
Degree of Freedom	435
Sig	0.00

Table 6

Goodness of fit indices	Measurement model	Structural model	Recommended value
χ 2/degree of freedom	1.156	1.782	< 3.0
CFI (comparative fit index)	0.985	0.937	>0.90
TLI (Tusker–Lewis fit index)	0.983	0.943	>0.90
IFI (incremental fit index)	0.985	0.974	>0.90
RMSEA (root mean square error)	0.078	0.072	<.08
GFI (goodness fit index)	0.930	0.943	>0.90
SRMR (root mean square residual)	0.46	0.39	<.50

Structural Equation Modeling (SEM):

The SEM model was employed to examine the relationship between constructs developed by study. Hence SEM analysis was performed by AMOS 24 version and analyses simultaneously goodness-of-fit indices. The results supported with Goodness fit indices. For the whole model statistical results shows that Chi-square/df = 1.156, CFI = 0.985, TLI = 0.983, IFI = 0.985, GFI = 0.930, RMSEA = 0.20, SRMR = 0.46 (Table 4). Hu and Bentler (1999) mentioned that RMSEA, TLI, and CFI are necessary to value for the model fit. According to the study, we hypothesized ten paths including seven TQM drivers and three enablers. Using the SEM investigated that impact of drivers, enablers and firm performance. Results exhibits all the paths are significant ($p < 0.05$). A SEM model divulges the critical success factors of TQM is directly and positively affects the strategic goal which further affect operational, financial and non-financial performances.

Table 7. Mean, Standard deviations and correlations

Variables	Mean	SD	1	2	3	4	5	6	7	8	9
1.Top-management commitment and leadership	6.09	1.65	-								
2. Customer focus	0.86	0.34	0.19*	-							
3.Continuous improvement and innovation	4.93	1.23	0.15*	0.22**	-						
4.Supplier management	5.62	1.17	0.17*	0.09	0.58	-					
5.Employee involvement	5.87	0.72	0.19*	-0.01	0.05	0.19*	-				
6.Quality Information /Information Quality	4.67	0.81	0.24**	0.04	0.02	0.14	0.24**	–			
7.Benchmarking	5.35	0.93	0.07	0.12	0.39**	0.26**	0.40**	0.16*	–		
8. Culture and communication	5.43	1.08	0.22**	0.23**	0.25**	0.21**	0.25**	0.36**	33**	–	
9. Employee encouragement and motivation	3.53	1.04	0.06	0.21**	0.22**	0.29**	0.16*	0.17*	0.19*	0.26**	–
10. Training and education	2.62	1.54	0.49**	0.28**	0.14	0.21**	0.17*	0.17*	0.17*	0.16*	0.23**

N=165 (Questionnaire distributed to 500 persons of the organization out of which 165 provided their response)
*P<0.05; **P<0.01.

The descriptive statistics for the study variables are shown in Table 8. In order to test the all ten hypotheses, mediated regression models were computed (e.g. James and Brett, 1984; Barron and Kenny, 1986). A variable is said to function as a mediator when variations in the independent variables significantly account for variations in the proposed mediator, variations in the mediator significantly account for variations in the dependent variable, and, when controlling for the mediator, a previously significant relationship between the independent variable(s) and the dependent variable decreases or becomes insignificant. One can test these conditions by regressing the mediator on the independent variable, regressing the dependent variable on the independent variable, and regressing the dependent on both the independent variable and the mediator (Barron and Kenny, 1986).

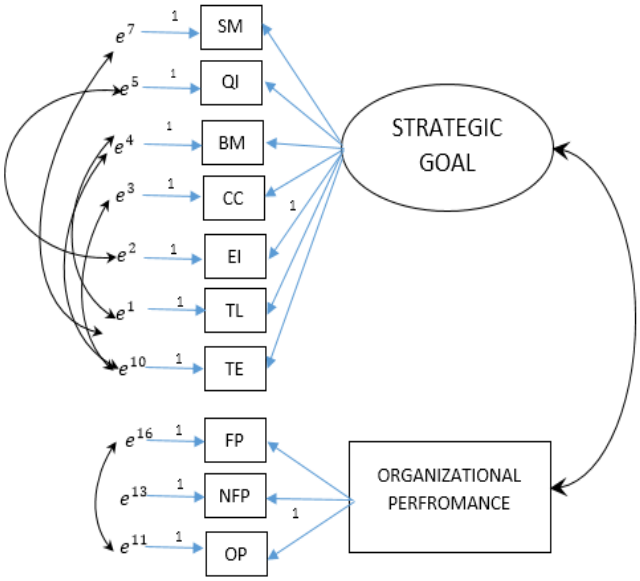


Figure 3. Measurement model for effect of TQM Strategic Goal on Organizational Performance

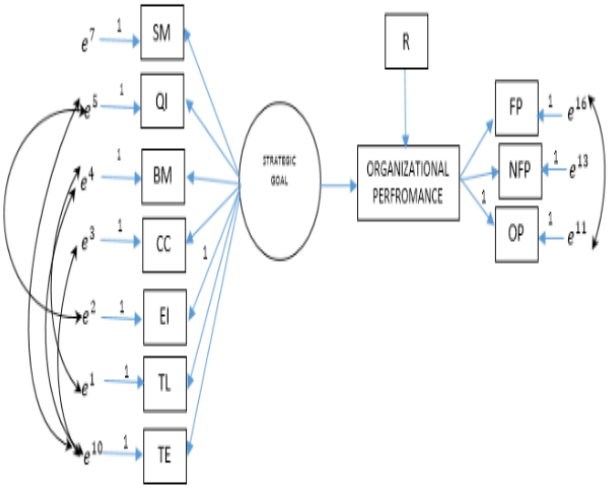


Figure 4. Structural model for effect of Critical Success Factors on Organizational Performance

Table 8. Construct Structural Model

Links in the model	Hypotheses	Standardized parameter estimate	T value	P value
SG -> OP	H_1	0.958	4.599	****
BM -> SG	H_{17}	0.218	2.222	0.026
QI -> SG	H_{16}	0.313	2.972	0.003
SM -> SG	H_{14}	0.447	3.820	****
CC -> SG	H_{18}	0.437	3.154	****
EE -> SG	H_{19}	0.589	4.399	****
TL -> SG	H_{11}	0.478	3.967	****
TE -> SG	H_{20}	0.526	3.776	****

Table 8 shows that the hypothesis H_1 is significant at $p < 0.05$. This indicates strategic goal has significant and direct effect on organizational performance. Also according to hypotheses H11 to H17, the TQM critical success factors selected has significant positive effect on Strategic Goal of TQM.

IV. Conclusion And Future Scope

Quality is considered by many researchers and industrial practitioner as one of the effective strategic weapon for improving productivity and enhancing reliability in the organization. A framework based on the attribute theory was used in an Indian FMCGs organization. The various empirical research has shown how the organizational performance is affected by various quality dimensions.

Seven TQM drivers, three enablers and three organizational performance measures were conceptualized for this study. From data analysis an empirically valid and reliable measurement instrument was developed. The purpose of this study was to link the critical success factors of TQM to the financial, non-financial and operational performance of the organization. Reliability and validity analysis was carried out for the constructs. All the factors were found to have acceptable reliability coefficients. Factor analysis was carried to test the construct validity of each of the factors.

The performance of an organization is the sum of all desired, but also undesired results, which were brought about by all efforts that were made possible within the organization, with the resources that were employed for this purpose.

Organizations should strive, as much as possible, to achieve only the results that add value. To be able to achieve only results that add value, with strategic goal we mean the specification of a certain value. the organization can follow the following steps:

First, the organization has to make sure that an individual employee knows exactly in which way he can influence the performances and how these performances add value, for this Training and education is required. Subsequently, the department or team (Top level management) in which the employee works also needs to know exactly what the team performance or department performance is and how the team or department can influence this performance, for this Quality Information is required. Here, primary activities that directly influence the performance level directly are most important. Secondary activities that have a less direct influence on the performance level (supplier management, benchmarking) are also important, but do not have an immediate impact on the results; Finally, it is necessary to define the strategic goal and critical success factors in a clear way.

Another interesting finding of this research is relationship between financial performances; non-financial performance and operational performance. The results revealed that these relationships are positive and significant.

As a future scope, this study may be extended by considering the drivers and enablers as external and internal one. As the performance of the organization vastly depends upon the critical success factors, so the externals factors must be given weightage as well as internal factors. The organizational performance may also be categorized as innovation and export performance, because FMGSs products having very short life cycle with heavy global demand.

This study, of course, is not without its limitations. First, the respondents are from same enterprise and some unbalanced distribution occurs. In future studies, we need to test our proposed model with larger and more balanced datasets. Second, the drivers and enablers vary according to the characteristics of the organization, namely, quality awareness and its impact, setting standards and requiring improvement, and collaboration, is not considered in this study. In future studies, we need to identify two clusters of TQM adopters among FMCGs industries, or non-FMCGs industries, and test our proposed model in a more robust manner. Third, we need to examine the interaction effect between the following two factors in our future studies: TQM Drivers and TQM enablers.

REFERENCES

[1]. Adam, E.E., Jr, "Alternative quality improvement practices
[2]. and organization performance", *Journal of Operations Management,* Vol.12, 1994, pp. 27-44.
[3]. Al-Marri, K., Ahmed, A. M. M. B., & Zairi, M., "Excellence in service: An empirical study of the UAE banking sector", International Journal of Quality and Reliability Management, Vol.24, 2007, 164–176.
[4]. Boynton, A. C., & Zmud, R. W., "An assessment of critical success factors", Sloan Management Review, 1984, Vol.25(4), pp.17-27.
[5]. Bush CT. Nursing research. Virginia: Reston Publishing Company; 1985.
[6]. Black, S. A. & Porter, L. J.," Identification of the critical
[7]. factors of TQM," Decision Sciences, Vol. 27(1), 1996, pp. 1–21.
[8]. Brotherton, B., & Shaw, J., "Towards an identification and classification of critical success factors in UK hotels plc.," International Journal of Hospitality Management, Vol.15(2), 1996, pp.113- 135.
[9]. Bagozzi RP, Yi Y., "On the evaluation of structural equation models," Academy of Marketing Science, Vol. 16(1), 1988, pp. 74–94.
[10]. Barron, R.M. and Kenny, D.A., "The moderator mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations", Journal of Personality and Social Psychology, Vol. 51, 1986, pp. 1173–1182.
[11]. Cronbach, L. J. Test validation. In R. L. Thomdike (Ed.), Educational measurement (2nd ed.). Washington DC: American Council on Education, 1971.
[12]. Crosby, B.P., "*Quality is Free",* New York, McGraw-Hill, 1979.
[13]. Christos B. Fotopoulos, Evangelos L. Psomas,"The impact of "Soft" and "Hard" TQM elements on quality management results", International Journal of Quality and Reliability Management, Vol. 26(2), 2009,pp.150-163.

[14]. Cleary, P., "An empirical investigation of the impact of management accounting on structural capital and business performance", Journal of Intellectual Capital, Vol. 16 No. 3, 2015, pp. 566-586.
[15]. Carmona-Márquez,F.,Leal-Millan, Vázquez-Sánchez,A., Leal-Rodríguez, A. and Eldridge, S. (2016), "TQM and business success: Do all the TQM drivers have the same relevance? An empirical study in Spanish firms", *International Journal of Quality & Reliability Management*, Vol. 33 No. 3, pp. 361-379
[16]. Deming, W.E., "*Out of the Crisis*", Cambridge, MA, MIT Centre for advanced Engineering, 1986.
[17]. Dunteman, G. H., "Principal components analysis", Beverly Hills: Sage Publications, 1989.
[18]. Feigenbaum, A.V., "*Total Quality Control*", New York, McGraw-Hill, 3rd ed.,1993.
[19]. Field, A. P., "Discovering Statistics Using SPSS", Sage Publications Inc., 2005.
[20]. Ghauri, P. & Gronhaug, K., "Research Methods in Business Studies", Harlow, FT/Prentice Hall, 2005.
[21]. G. Muruganantham, S. Vinodh, C. S. Arun and K. Ramesh, "Application of interpretive structural modelling for analyzing barriers to total quality management practices implementation in the automotive sector", Total Quality Management, Vol. 29, No. 5, 2018, pp.524–545.
[22]. Hair JF, Anderson RE, Tatham RL, Black WC., "Multivariate data analysis with readings", New York, NY: Macmillan; 1992.
[23]. Hatcher, L., "A step-by-step approach to using the SAS(R) system for factor analysis and structural equation modeling", Cary, NC: SAS Institute, 1994.
[24]. Hinterhuber, H.H., "Strategische Unternehmensführung: I. Strategisches Denken" (strategic management: I. strategic thinking (original in German)). 7, Berlin, New York: Walter de Gruyter, 2004.
[25]. Ishikawa, K., "*What is Total Quality Control? The Japanese Way* (Englewood Cliþ s, NJ, Prentice Hall), 1985.
[26]. Juran, J.M., "Product quality: a prescription for the west (Part I)", *Management Review*, Vol. 70, 1981, pp. 8-14.
[27]. James, L.R. and Brett, J.M., "Mediators, moderators, and tests for mediation", Journal of Applied Psychology, Vol. 69,1984, pp. 307–321.
[28]. Jaideep Motwani,"Critical factors and performance measures of TQM", The TQM Magazine, Vol. 13(4), 2001, pp. 292-300.
[29]. Joanna Hing Yee Tsang, Jiju Antony, "Total quality management in UK service organisations: Some key findings from a survey", Journal of Service Theory and Practice, Vol. 11(2), 2001, pp.132-141.
[30]. Kerlinger FN., "Foundations of behavioral research", New York: CBS Publishing, 3rd ed., 1986.
[31]. Kanji, G.K., Wallace, W., "Business Excellence Through customer Satisfaction", Total Quality Management, Vol. 11 (7), 2000, pp.979-998.
[32]. Koufteros, X., Verghese, A.J. and Lucianetti, L., "The effect of performance measurement systems on firm performance: a cross-sectional and a longitudinal study", Journal of Operations Management, Vol. 32 No. 6, 2014, pp. 313-336.
[33]. Malcolm C. Munro and Basil R. Wheeler., "Planning, Critical Success Factors, and Management's Information Requirements", JSTOR, Vol. 4, No. 4, 1980, pp. 27-38.
[34]. Michel J. Lebas, "Performance measurement and performance management", Int. J. Production Economics. Vol. 41, 1995, pp.23 35.
[35]. Mohamed Zairi Mohamed A. Youssef, "Benchmarking critical factors for TQM", Benchmarking for Quality Management & Technology, Vol. 2 Iss 1, 1995, pp. 5 – 20.
[36]. Masood A. Badri Donald Davis Donna Davis, "A study of measuring the critical factors of quality management", International Journal of Quality & Reliability Management, Vol. 12 Iss 2, 1995, pp. 36 – 53.
[37]. Madu, C.N., Kuei, C.H. & Jacob, R.A., "An empirical assessment of quality dimensions on organizational performance", *International Journal of Production Research*, Vol. 34, 1996, pp. 1943-1962.
[38]. M. Hasan R.M. Kerr, "The relationship between total quality management practices and organisational performance in service organisations", The TQM Magazine, Vol. 15 Iss 4, 2003, pp. 286 – 291.
[39]. Nunnally, J.C., Bernstein, I.H. and Berge, J.M.T.,"Psychometric Theory", McGraw-Hill, New York, NY. Vol. 2, 1967.
[40]. Nunnally, J.C., "Psychometric theory", New York: McGraw-Hill Book Company, ed.2nd, 1978.
[41]. Polit DF, Hungler BP. Nursing research principles and methods. 4thed. Philadelphia: ID Lippincott Company; 1991.
[42]. Rockart, J. F., "Chief executives define their own data needs", Harvard Business Review. Vol. 57, 1979, pp.81-93.
[43]. Shewart, W. A., "*Economic Control of Quality of Manufactured Product",*New York, Van Nostrand, 1931.
[44]. Stevens, J., "Applied multivariate statistics for the social sciences", Mahwah, NJ: Lawrence Erlbaum Associates, 1996.
[45]. Samson, D. & Terziovski, M., "The relationship between total quality management practices and operational performance, *Journal of Operations Management*, Vol. 17, 1999, pp. 393- 409.
[46]. Sohal, A. S. & Terziovski, M., "TQM in Australian manufacturing: Factors critical to success", International Journal of Quality and Reliability Management, Vol.17, 2000, pp.158–167.
[47]. Saravanan, R. & Rao, K. S. P., "An instrument for measuring total quality management
[48]. implementation in service-based business units in India", Proceedings of the International Conference on Manufacturing and Management, Tamilnadu, India, 2004.
[49]. S.S.Mahapatra& M.S.Khan, "A framework for analysing quality in education settings", European Journal of Engineering Education, Vol. 32, Iss.2, 2007, pp.205-217.
[50]. Taguchi, G., "*Introduction to Quality Engineering, Designing Quality into Products and Processes",* (New York, Unipub), 1986.
[51]. Thiagarajan, T. & Zairi, M., "A review of TQM in practice: Understanding the fundamentals through examples of best practice applications—Part 1", The TQM Magazine, Vol. 9, 1997, pp.270–286.
[52]. Tamimi, N., "A second-order factor analysis of critical TQM factors", International Journal of Quality Science, Vol. 3(1), 1998, pp.71–79.
[53]. Terziovski, M. & Samson, D., "The link between total quality management practice and organizational performance", *Internal Journal of Quality and Reliability Management*, Vol. 16, 1999, pp. 226- 237.
[54]. U Lenka, D Suar , "A holistic model of total quality management in services", The Icfaian Journal of Management Research, Volume 7,Issue 3, 2008, pp.57-72.
[55]. Yusuf, Y., Gunasekaran, A., & Dan, G., "Implementation of TQM in China and organizational performance: An empirical investigation", Total Quality Management, Vol.18, 2007, pp.509–530.
[56]. Zairi, M., "Measuring performance for business results", London: Chapman Hall, 1994.
[57]. Zairi, M. and Leonard, P., "Practical Benchmarking: The Complete Guide", Chapman & Hall, London, 1994.

About Authors

Mr. Rakesh Kumar received Bachelor's in Industrial Engineering and Master's in Mechanical Engineering with specialization in Computer Integrated Manufacturing from India. He is currently a research scholar at Delhi Technological University in Department of Mechanical Engineering under the supervision of Dr. R.S. Mishra. His research interest includes Total Quality Management, Reliability Engineering, Operations and Supply chain, Fuzzy Logic, AI and Decision Science.

R.S. Mishra (Professor) received the Ph.D. degree in Mechanical engineering from IIT Delhi, India and currently at Delhi Technological University serving since 1997. He having 150 research papers in peer reviewed in International journals and 75 research papers in proceedings of International and National conferences. Prior to joining DTU, he served 09 years at different engineering colleges of India, viz, REC Hamirpur, PAU Ludhiana (Punjab) and HAU, Hisar (Haryana.).
His research area includes Green Technology, Refrigeration & Air Conditioning, Metal Cutting Technology, Total Quality Management, Solar Energy Technology, Bio Fuels, power Plant engineering.

Rakesh Kumar, et. al. "Linking TQM Critical Success Factors to strategic goal: Impact on Organizational Performance." *IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE)*, 17(3), 2020, pp. 01-13.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Localization of User in an Indoor Environment Using Machine Learning Classification Models

Manish Kumar
Computer Science And Engineering,
Delhi Technological University,
Delhi, India
manishkumar59497@gmail.com

Manish Rawat
Computer Science And Engineering,
Delhi Technological University,
Delhi, India
manishrawat2674@gmail.com

Prashant Giridhar Shambharkar
Computer Science And Engineering,
Delhi Technological University,
Delhi, India
prashant.shambharkar@dtu.ac.in

***Abstract*—There is a wide range of applications like detecting the position of people in intelligent home systems, locating persons in specific areas, finding the number of people using an internet access point, etc. for our present investigation. Here, based on Wi-Fi signal values we are discovering users in an indoor location. In this paper, we are using the Wi-Fi signal strength obtained on the mobile devices of various people through different routers present in an office like the place and using this data to build an effective model to predict the user's location based on that. The signal strength values collected from routers are associated with the person's position and that association can be considered as a grouping problem. We have used five classification models, "logistic regression, k-nearest neighbor, support vector machine, kernel support vector machine, and naïve bayes" to obtain a model that results in better prediction of the classifier.**

Keywords—data preprocessing; dimensionality reduction; kernel support vector machine; naïve bayes

I. INTRODUCTION

Wide possibilities currently present in mobile computing have been enabled due to the advancements in location-based services. Many methods are being used to give users with services based on their locations. Plans like these have given a great increase in e-business profits, position-based recommendation systems, embedded smart systems, and many other areas. Many other technologies e.g. GPS, Wi-Fi, and Bluetooth can be used to give such facilities [1]. This method uses the global positioning system (GPS) techniques for locating people within certain ranges, and due to weak satellite signals, it cannot be applied in indoor locations. Bluetooth can be used to fill this purpose, but it usually works for small ranges. Thus, using Wi-Fi entry locations for user localization can be a good method [2]. These techniques guess the way people behave by observing their regular activities. Such systems that detect locations of people could also be used to provide rehabilitation to people in panic situations and disasters [3]. The methods introduced SVM method and presented improved classifying results contrasted to other previous classification methods [4]. The method aimed to classify the location of outdoor and indoor places using the position data of mobile devices and proposed learning methods for classification. The method proposed a procedure based on various machine learning techniques for indoor localization and showed its easy adaptation to the environment [5]. This method proposed the particle swarm optimization (PSO) strategy which views particles moving in the search space or population of candidate solutions and is improved to their local best or global best based on their velocity and position values [6]. The method introduced an adaptive approach of PSO to check the fuzzy system used in different views of the PSO by using multiple benchmark functions [7]. The method proposed a back-propagation approach based on the bee colony to train artificial neural networks [8]. The method used the PSO and cuckoo swarm method to learn an MLP and then showed the by using optimization methods we can enhance the performance of the neural network marginally [9]. The need for our issue is in detecting users at correct positions using better learning techniques. There is a lack of the proper utilization of optimization techniques that teach the model correctly in most of the techniques proposed.

For predicting the location of people rightly, a clear and compatible model needs to be taught and then used in some tracing or observing instruments. We take the strength of Wi-Fi signal collected by different routers from a specific area and teach our classification models so that it would, later on, estimate the user's location for some not known data values set of the strength of signals. We are considering a dataset from UCI ML repository. The data is collected at an office in New Delhi, INDIA. There are 7 Wi-Fi routers devices in this office and signal values that are calculated out of these devices are used to classify the location of a person in the cafeteria, meeting room, main hall, or the refreshment area. W1 is the signal value collected from the device 1, W2 is the signal value collected from the device 2 and same for the remaining devices. The class marks belong to the cafeteria, meeting room, main hall, and the refreshment area, are marked 1, 2, 3, and 4 accordingly. In this example scenario, collected wireless signal values shown on Android mobile devices. In some places, the values of signals were collected by sampling the strength of the wireless signal in a fixed interval of time (1 s period is taken). It was then again simulated for different places and so that one thousand and six hundred examinations done for this scenario

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:49 UTC from IEEE Xplore. Restrictions apply.

for these seven routers. The model evolved here, can, therefore, be used again as per the situation of the closed area and the count of wireless router devices in the actual scenario. The data collected is used to create a sample classification dataset with signal values from seven routers used as the input parameters. This dataset now will be used to predict the user's location into four types of classes. After getting a solid dataset maintained, we now teach our different classification models using a method that increases the probability of defining the correct class label perfectly.

Our objective is to take this user localization based on the Wi-Fi strengths of user devices as a classification problem. The data set consists of signal strength of seven Wi-Fi routers WS1, WS2,.., WS7. We will use logistic regression, k-nearest neighbor, support vector machine, kernel support vector machine, and naïve bayes as our five classification techniques to train our model to predict the location of a new user using our trained model with maximum precision. We will also analyze the results of all five classification techniques used.

II. Literature Review

The method discussed here uses meta-heuristic approaches that ensured plenty of solutions for optimization procedures. Particle Swarm Optimization (PSO) method had been proposed in this approach that finds a population of possible solutions or particles that are present in our search space which is upgraded to its local best or global best based on position and velocity conditions [10]. Major roadblocks that the standard PSO algorithm often encounters is the capabilities to investigate new search spaces. To examine the fuzzy system used in diverse use cases of the PSO, this method instigated an adaptive PSO strategy that uses more than one benchmark functions [11]. A randomized PSO which shows the requirement to generate new search spaces by creating some instability parameter in the velocity part of the PSO had been suggested in this approach [12]. Even using these algorithms does not guarantee to achieve the global optimum. So some hybrid PSOGSA method is also used previously [13]. It helps GSA to get out of the local optimum value which results in enhancement of the accuracy of neural networks, this issue has been vanquished. By examining these factors, they presented the FPSOGSA that intensifies the possibility of a higher convergence rate by vanquishing the possibilities of confining itself in the local minima. Similar to PSOGSA, it moderately decreases the error rate instead of accomplishing saturation at higher iterations. It acquires superior convergence, which intensifies the capability for the optimization of the neural networks due to which the error generated in the fuzzy neural network is reduced. Many similar techniques are being leveraged to train a variety of classifiers. Artificial neural network (ANN) is yet another model which is a non-parametric and non-linear. It is very simple to use and perceive. Using the backpropagation (BP) algorithm in ANN, various forecasting, and classification problems are solved. They apply optimization methods to accomplish higher correctness rates and faster convergence. Hence, the hybrid PSOGSA algorithm has been used as an optimization method and on the other side, the ANN approach is used for training. Uncertainties cannot be handle using this strategy.

III. Implementation

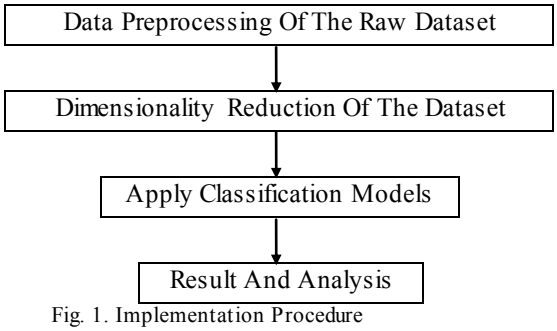


Fig. 1. Implementation Procedure

Our implementation is a four-step process. The first step of data preprocessing is performed to convert the raw form of data into a usable form. Since our dataset is multidimensional, we have applied a dimensionality reduction technique to convert it into the 2D dataset. After this step, we have applied the five classification models and further performed the analysis of the results obtained.

IV. Data Preprocessing Of The Dataset

To build a machine learning model some pre-processing of present data is an important step as without doing data pre-processing our machine learning model will not work properly. The pre-processing of data necessitates the modification of raw data into a comprehensible format. Real-world data is ordinarily incomplete, erratic, and may contain many flaws. All these concerns can be rectified with the help of pre-processing the data.

A. Data Cleaning

Raw data may contain missing or inessential values. This may lead to inconsistencies in a dataset. All these issues are handled in the data cleaning step. All the missing values are filled by following proper criteria and all the inessential values are removed.

B. Creating Test And Training Sets

In this step, we bisect a data set into a training set and testing set in which a major section of the data is used for training(70%-80%) purpose and a smaller section of the data is used for testing(20%-30%) purpose.

C. Feature Scaling

When the independent variables are measured at different scales then this may result in one variable dominating other variables. Thus feature scaling is a crucial step where all the data values are transformed to a certain range. This ensures all the variables having an equal contribution to training the dataset. There are two widely used methods of feature scaling:

Normalization

$$X_{Norm} = (X - Min(X))/(Max(X) - Min(X)) \quad (1)$$

Standardization

$$X_{stand} = (X - Mean(X))/(Standard\ Deviation(X)) \quad (2)$$

Here, Min(X) and Max(X) represent the minimum and

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:49 UTC from IEEE Xplore. Restrictions apply.

maximum value present in the dataset X. Mean(X) gives the mean value and Standard Deviation(X) gives the standard deviation of the dataset X.

V. DIMENSIONALITY REDUCTION TECHNIQUE

A. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a popularly used dimensionality reduction technique. In this approach out 'n' independent variables of the dataset, we select 'p' new variable where p<=n, that helps in differentiating classes of the dependent variable to reduce overfitting and cost of computation [14]. According to mathematics, we need to find a new feature space to project the data to increase distance among a different set of classes. Well, the first step is to find a method to define the separation capacity of each new feature space possible. So one method could be the distance between the projected means of each class. But this is not as good because it does not take the spread of data into account.

LDA and PCA (Principal Component Analysis) are closely related to each other. However, the major difference is LDA attempts to maximize the difference between the classes whereas PCA doesn't use a class label in its implementation. The fact that the dependent variable or class label is considered makes LDA a supervised model and on the other hand, PCA doesn't consider class labels which makes it a supervised model.

VI. CLASSIFICATION MODELS

A. Logistic Regression

The logistic regression model is a supervised machine learning model. It is used to determine the binary output based on the set of given independent variables. Logistic regression output comprises of only two classes that are either True or False**,** while linear regression predicts continuous value as output. It fits an "S" shaped curve into the data. Based on the probability of occurrence of an event, it fits data to the logistic function. Logistic regression has many practical applications in the fields of medical and social sciences. data may contain missing or inessential values. This may lead to inconsistencies in a dataset. All these issues are handled in the data cleaning step. All the missing values are filled by following proper criteria and all the inessential values are removed.

B. K-Nearest Neighbor

K-NN is a simple to use and understand machine learning classification model. In this, our approach is to determine the K training examples whose features are relatively similar for the given test sample. The way the value of K is selected is very critical as if the value of K is very small then it may lead to overfitting problem and if the values are very large then the classifier may misclassify the test sample because the number of nearest neighbors may include data points that lie far away from the neighborhood of the test sample. The distance between two data points is computed using the Euclidean measure. Hence, the test samples are classified based on the class labels of these K closest training samples.

It does not make any explicit assumptions about the relation between the independent and the dependent variable hence it is called non-parametric classification technique. It is also called a lazy-learning technique as the classifier makes predictions based on local information without building explicit models.

C. Support Vector Machine

One another supervised machine learning model is the support vector machine(SVM). The major use case of SVM is in classification and regression problems [15]. In this approach, every single data point is represented in the n-dimensional graph, where n is the number of attributes of the dataset with the values of each attribute being the value of respective coordinates. Further, our goal is to determine the maximum margin hyper-plane which separates the two classes in the best possible way. SVM can also be applied in the case of categorical data by initiating dummy variables for each categorical feature value present in the dataset. SVM does not have issues related to dimensionality problems. It also works well in case of high dimensional data. It has many practical applications such as hand-written digits recognition, classification of images, categorization of texts, etc.

D. Kernel Support Vector Machine

Kernel strategies are a category of the algorithm used in sample analysis in machine learning. One of the members of this category is Support Vector Machine (SVM). Locating and studying standard styles of family members (for example classifications, clusters, principle components, ratings, correlations) in datasets are some major tasks of sample analysis. For many of the algorithms that are capable of solving this kind of task, the records present in raw illustration need to be directly changed into characteristic vector illustrations using a feature map defined by users. In evaluation, kernel techniques need a person-particular kernel simply, i.e., a characteristic feature which tells the similarity over pairs of facts points in raw datasets. Kernel features have a major contribution in kernel strategies because it allows them to work in a high dimensional, clear function area without calculating the coordinates of the records in that space, but alternatively through clearly calculating the inner merchandise among the pictures of all pairs of statistics inside the function area [16]. This operation is less expensive than the specific calculations of the coordinates. This method is known as the "kernel trick". Kernel features were added to collect data, graphs, text, pictures, in addition to vectors.

E. Naïve Bayes

A simple method for building classifiers is Naïve Bayes. It is a model that puts important labels to difficult times, which is present in the form of vectors of feature values, wherein the labels of each class are taken from some fixed set [17]. No single algorithm is present for building such classifiers, however a class of algorithms mainly based on a not unusual

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:49 UTC from IEEE Xplore. Restrictions apply.

precept: there is an expectation that the cost of a specific characteristic is not dependent on the fee of every other characteristic by all naïve bayes classifiers, given the impressiveness variable. As an example, if the fruit is red, round, and about 10 cm in diameter it can be considered to be an apple. Each of the above characteristics helps in deciding if the given fruit is apple independently. It does not consider any correlations between the color, shape, and diameter features.

For some varieties of possibility models, naive bayes classifiers can be used for putting the data sets in classes correctly. In many sensible programs, the approach of most chance is used to estimate parameters for naive bayes models. Regardless of its basic design and extremely simple assumptions made by naive bayes classifiers, it has worked very well in many complex actual-international situations. A study conducted in 2004 for evaluating the problems of the bayesian classification, showed that there is some good amount of theoretical backing for the amazing efficacy of naïve bayes classifier. Nonetheless, a comprehensive assessment with other classification algorithms in 2006 founded that some tactics can outperform bayes classification, which include faster timber or random forests. A bonus of naive bayes is that it most effective calls for a few numbers of education data to find the parameters essential for class.

VII. Result And Analysis

A. *Logistic Regression*

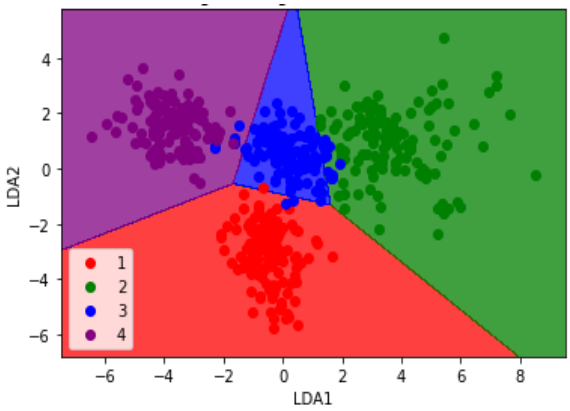


Fig. 2. Logistic Regression (*Test Set*)

In the above graph, LDA1 and LDA2 represent two-dimensionally reduced components. Here the model applied is Logistic Regression. The colors used in the above figures red, blue, green, and violet are used to represent four classes that are 1,2,3 and 4 of the output.

Table I. Prediction Table for Logistic Regression

Class	Number Of Correct Predictions	Number Of Incorrect Predictions
1	131	1
2	121	1
3	110	18
4	118	0

Total Number Of Correct Predictions = 480

Test Set Size = 500

Accuracy of Model = (480/500)*100 = 96%

B. *K-Nearest Neighbors*

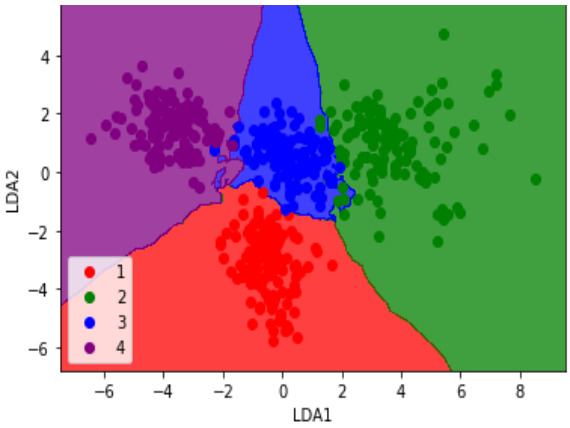


Fig. 3. K-Nearest Neighbor (*Test Set*)

Here we have applied the K-nearest neighbor classification model. The colors used in the above figures red, blue, green, and violet are used to represent four classes that are 1,2,3 and 4 of the output.

Table II. Prediction Table for K-Nearest Neighbor

Class	Number Of Correct Predictions	Number Of Incorrect Predictions
1	131	1
2	112	10
3	124	4
4	117	1

Total Number Of Correct Predictions = 484

Test Set Size = 500

Accuracy of Model = (484/500)*100 = 96.8%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:49 UTC from IEEE Xplore. Restrictions apply.

C. *Support Vector Machine*

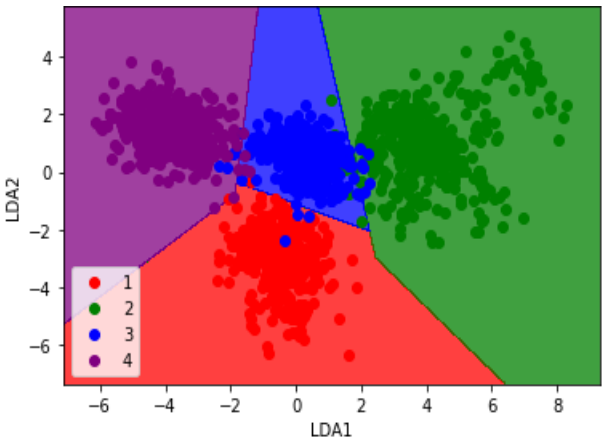


Fig. 4. Support Vector Machine (*Test Set*)

Here we have applied the support vector machine classification model. The colors used in the above figures red, blue, green, and violet are used to represent four classes that are 1,2,3 and 4 of the output.

Table III. Prediction Table for Support Vector Machine

Class	Number Of Correct Predictions	Number Of Incorrect Predictions
1	131	1
2	114	8
3	124	4
4	116	2

Total Number Of Correct Predictions = 485

Test Set Size = 500

Accuracy of Model = (485/500)*100 = 97%

D. *Kernel Support Vector Machine*

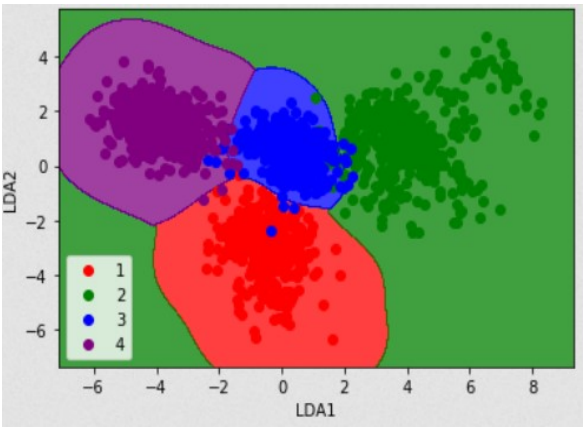


Fig. 5. Kernel Support Vector Machine (*Test Set*)

Here we have applied the kernel support vector machine classification model. The colors used in the above figures red, blue, green and violet are used to represent four classes that are 1,2,3 and 4 of the output.

Table IV. Prediction Table for Kernel Support Vector Machine

Class	Number Of Correct Predictions	Number Of Incorrect Predictions
1	132	0
2	115	7
3	124	4
4	116	2

Total Number Of Correct Predictions = 487

Test Set Size = 500

Accuracy of Model = (487/500)*100 = 97.4%

E. *Naïve Bayes*

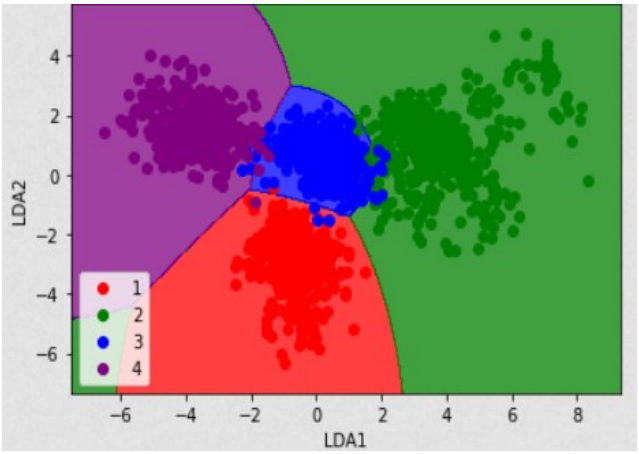


Fig. 6. Naïve Bayes (*Test Set*)

Here we have applied naïve bayes machine classification model. The colors used in the above figures red, blue, green, and violet are used to represent four classes that are 1,2,3 and 4 of the output.

Table V. Prediction Table for Naïve Bayes

Class	Number Of Correct Predictions	Number Of Incorrect Predictions
1	131	1
2	120	2
3	122	6
4	115	3

Total number of correct predictions = 488

Test Set Size = 500

Accuracy of Model = (488/500)*100 = 97.6%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:49 UTC from IEEE Xplore. Restrictions apply.

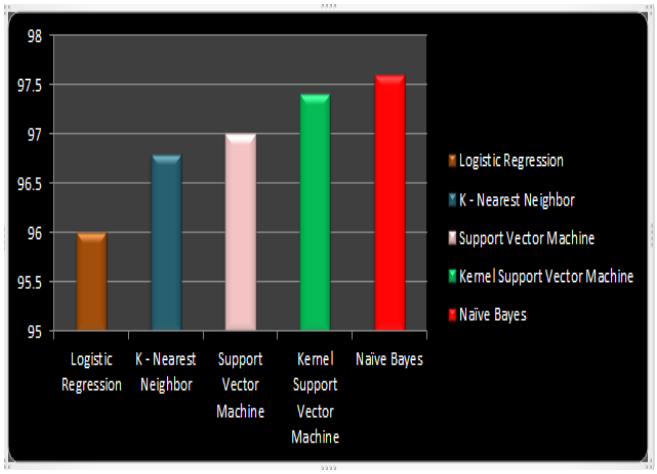


Fig. 7. Analysis Of Models

In *Fig. 6*, graph demonstrates the estimate of accuracies of the five classification methods that we have applied. We can notice that there is the least error when we have trained our dataset using naïve bayes model. We obtained accuracies of 96%, 96.8%, 97%, 97.4%, and 97.6% by applying logistic regression, k-nearest neighbor, support vector machine, kernel support vector machine, and naïve bayes models respectively. So the Naïve Bayes model came out to be the best model with maximum prediction accuracy of 97.6% among all other models used.

VIII. CONCLUSION

Our work deals with the task of detecting the user's location based on the Wi-Fi signal strengths of the routers present in the same locality. For this, we have taken this task of identification of the user's location as a classification problem. To do this classification of user based on signal strengths we have used five famous machine learning classification models which are support vector machine, logistic regression, KNN, kernel support vector machine, and naïve bayes. We train all these models on the available data set and the test results show that Support Vector Machine and Naïve Bayes have a maximum prediction accuracy of 97.6% among the five used models. This work has various real-world implementations. This can be used for objectives like detecting the position of people in intelligent home systems, tracking people in specific areas, finding the number of users on an internet access point, etc. These position perception systems are also used in conditions like disasters where people require needful rehabilitation. This can also be used in giving location-based services to people, location-based recommender systems.

Since the dataset used in our work is not very big and complex, an extension for the near future is the use of more powerful and optimized classification models that can handle big and complex datasets.

REFERENCES

[1] Bulusu, N., Heidemann, J., Estrin, D.: GPS-less low-cost outdoor localization for very small devices. IEEE Personal Commun. 7(5), 28–34, 2000.

[2] Salazar, A.M., Warden, D.L., Schwab, K., Spector, J., Braverman, S., Walter, J., Ellenbogen, R.G.: Cognitive rehabilitation for traumatic brain injury a randomized trial. JAMA 283(23), 3075–3081, 2000.

[3] Nguyen, N.T., Bui, H.H., Venkatsh, S., West, G.: Recognizing and monitoring high-level behaviors in complex spatial environments. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, II-620. IEEE, 2003.

[4] Cho, S.B.: Exploiting machine learning Techniques for location recognition and prediction with smartphone Logs. Neurocomputing, 2015.

[5] Zou, H., Lu, X., Jiang, H., Xie, L.: A fast and precise indoor localization algorithm based on an online sequential extreme learning machine. Sensors 15(1), 1804–1824, 2015.

[6] Eberhart, R.C., Kennedy, J.: A new optimizer using particle swarm theory. In: Proceedings of the Sixth International Symposium on Micro Machine and Human Science, vol. 1, pp. 39– 43, 1995.

[7] Shi, Y., Eberhart, R.C.: Fuzzy adaptive particle swarm optimization. In: Proceedings of the 2001 Congress on Evolutionary Computation, vol. 1, pp. 101–106. IEEE, 2001.

[8] Nandy, S., Sarkar, P.P., Das, A.: Training a feed-forward neural network with artificial bee colony based backpropagation method. arXiv preprint arXiv:1209.2548, 2012.

[9] Kawam, A.A., Mansour, N.: Metaheuristic optimization algorithms for training artificial neural networks. Int. J. Comput. Inf. Technolgy 1, 156–161, 2012.

[10] Jayant G. Rohra , Boominathan Perumal , Swathi Jamjala Narayanan , Priya Thakur , and Rajen B. Bhatt. User Localization in an Indoor Environment Using Fuzzy Hybrid of Particle Swarm Optimization & Gravitational Search Algorithm with Neural Networks. *Springer Nature Singapore, 2017,* DOI 10.1007/978-981-10-3322-3_27, 2017.

[11] Pei, L., Liu, J., Guinness, R., Chen, Y., Kuusniemi, H., Chen, R.: Using LS-SVM based motion recognition for smartphone indoor wireless positioning. Sensors 12(5), 6155–6175, 2012.

[12] Liu, H., Abraham, A., Zhang, W.: A fuzzy adaptive turbulent particle swarm optimisation. Int. J. Innovative Comput. Appl. 1(1), 39–47, 2007.

[13] Mirjalili, S., Hashim, S.Z.M., Sardroudi, H.M.: Training feedforward neural networks using hybrid particle swarm optimization and gravitational search algorithm. Appl. Math. Comput. 218(22), 11125–11137, 2012.

[14] Raji Ramachandran, Gopika Ravichandran & Aswathi Raveendran, Evaluation of Dimensionality Reduction Techniques for Big data, 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), 2020.

[15] Raj, J. S., & Ananthi, J. V., Recurrent Neural Networks And Nonlinear Prediction In Support Vector Machines, Journal of Soft Computing Paradigm (JSCP), 1(01), 33-40, 2019.

[16] Gong Cheng & Xiaojun Tong, Fuzzy Clustering Multiple Kernel Support Vector Machine, 2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), 2018.

[17] Akarshita Tripathi, Saumya Yadav & Rajiv Rajan, Naïve Bayes Classification Model for the Student Performance Prediction, 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:49 UTC from IEEE Xplore. Restrictions apply.

Materials Research Express

PAPER • OPEN ACCESS

Mechanical and tribological properties of AA7075-T6 metal matrix composite reinforced with ceramic particles and aloevera ash via Friction stir processing

To cite this article: Lakshay Tyagi *et al* 2020 *Mater. Res. Express* **7** 066526

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.

This content was downloaded from IP address 14.139.251.100 on 19/06/2020 at 07:06

Materials Research Express

PAPER

Mechanical and tribological properties of AA7075-T6 metal matrix composite reinforced with ceramic particles and aloevera ash via Friction stir processing

Lakshay Tyagi[1], Ravi Butola[1] and Avinash Kumar Jha[2]

[1] Mechanical, Production & Automobile Engineering Department, Delhi Technological University, New Delhi, India
[2] Civil Engineering Department, Delhi Technological University, New Delhi, India

E-mail: ravibutola33855@gmail.com and avinash.jha.6696@gmail.com

Keywords: aluminium, wear, microhardness, friction stir processing, SiC, Aloevera ash

OPEN ACCESS

RECEIVED
13 February 2020

REVISED
11 May 2020

ACCEPTED FOR PUBLICATION
12 June 2020

PUBLISHED
19 June 2020

Original content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

Abstract

Mechanical and wear properties of AA7075-T6 reinforced with SiC and Aloevera ash, fabricated using Friction stir processing (FSP) are investigated in this study. Due to less density, easy availability, and cost-effectiveness, aloevera ash is considered as one of the reinforcements. FSP is done using a square tool pin profile, at different tool rotational speeds. To study wear behaviour, pin on disc test is carried out on High Temperature Rotary Tribometer at 20N, 30N and 40N applied load. Wear increases on increasing the applied load and at 20N load Al+SiC/Aloevera ash composite, processed at 600 tool rpm gave the best results due to the formation of oxide tribolayer. At 30N and 40N applied load Al +SiC composite, processed at 900 tool rpm showed the least wear because of proper scattering of ceramic particles due to high tool rotational speed. Coefficient of friction increases on increasing the applied load and all fabricated composite samples showed a lesser coefficient of friction than the base metal. Microhardness, ductility and Ultimate tensile strength increases on the addition of reinforcement and had a direct relation with tool rpm. Wear morphology was analysed using Scanning Electron Microscope (SEM). Energy Dispersive Spectroscopy (EDS) analysis after wear shows the presence of C, Fe, O, Mg, Zn, Si, Al elements and confirms the formation of an oxide layer which is responsible for decreasing wear loss.

1. Introduction

Aluminium is one of the most abundant metal found on the Earth. It finds wide application in the automobile sector, defence sector, nuclear industry for neutron absorption etc [1]. Aluminium parts are widely used in machines and aluminium composites are widely used in industries [2]. Wear of aluminium is one the serious concern, while it cannot be stopped but can be minimised. In order to minimise the wear Aluminium Metal Matrix Composites (AMMC) are being used instead of Aluminium metal.

In recent years, many researchers have been conducted to fabricate aluminium composite with desirable mechanical and tribological properties. Zawawi *et al* [3] discussed the wear rate and coefficient of friction effect of Al_2O_3-SiO_2 composite nano lubricant. They found that pure lubricant had a higher coefficient of friction and wear rate than composite nano lubricant. Naveed *et al* [4] used Graphite as reinforcement keeping SiC content as constant and Aluminium as the base metal. They concluded that wear properties of Aluminium Matrix Composite were enhanced and properties further improved on heat treatment. Raghavendra *et al* [5] coated Ni-Al_2O_3 nano particles by electro-deposition process on aluminium, Ni-Al_2O_3 electrodeposited composite showed better wear resistance than Ni coating. Razzaq *et al* [6] on observing AA6063 alloy reinforced with different wt.% of fly ash, found that wear rate decreased on increasing fly ash content and it increased on increasing the applied load. Nagaral *et al* [7] used SiC and Graphite as reinforcement and Al6061 as the base metal. They stated that Al6061%-6%SiC-6%graphite showed less volumetric wear than Al6061 lattice. They also concluded that on increasing applied load, sliding distance and disc speed volumetric wear increases. Gireesh

© 2020 The Author(s). Published by IOP Publishing Ltd

et al [8] reinforced Al6061 Metal Matrix Composite with aloevera and found that wear resistance, density and ultimate tensile strength showed no significant difference whereas ultimate yield strength and hardness increased as compared to pure Aluminium based Metal Matrix Composite reinforced with aloevera. Mittal *et al* [9] used SiC, Al_2O_3 and red mud as reinforcement, AA7075 as the base metal and found that wear rate decreased with increasing wt.% of Al_2O_3 and SiC with red mud it decreases till 7.5% wt. and on further adding wear rate increases. Baradeswaran *et al* [10] found that wear resistance increased on increasing B_4C content which was used as reinforcement and AA7075 as base alloy.

Raman *et al* [11] fabricated AMMC using SiC as reinforcement, processed through stir casting and observed that wear rate decreased and wear resistance increased on increasing SiC content. Sun *et al* [12] on evaluating the microstructure of AA206 during FSP, they observed that grain size and pore morphology of processed material was decreased. Ning Sun *et al* [13] investigated FSP effect on mechanical and structural properties of age hardenable AA206, tensile and ductile properties were observed to enhance and solidification shrinkage porosity almost eliminated. Singla *et al* [14] used SiC as reinforcement and Al as base metal. They observed that volume wear increased linearly with sliding distance for both composite and base metal. Wear Rate increased with normal load, but the increment was less for composite than that for pure aluminium. Alizadeh *et al* [15] used B4C nanoparticles as reinforcement and Al as the base metal. They observed that mechanical properties i.e. yield strength, hardness and tensile strength improved and wear rate was observed to be less than that coarse grain Al matrix. Gireesh *et al* [16] reinforced aluminium with fly ash as well as with aloevera separately and concluded that aluminium reinforced with aloevera provided better wear resistance than aluminium reinforced with fly ash. Butola *et al* [17] found that tool profile in FSP is the least influential factor, whereas tool rotational speed of FSP tool influence microhardness the most.

Miranda *et al* [18] formed aluminium based FGM composite using FSP. They concluded that because of FSP there were decreased number of cracks and voids, and surface hardness of composite increased. Vijayairl *et al* [19] reinforced LM25AA with 5% SiC by stir casting and then FSP was carried out on this composite with different shoulder diameter. They concluded that for better microhardness and tensile strength shoulder diameter to pin diameter (D/d) ratio should be 3 and there was very fine distribution of SiC particles in the matrix after FSP. Dinaharam [20] reinforced AA6082 with ceramic particles i.e. SiC, Al2O3, B4C, TiC and WC via friction stir processing. He found that hardness and tensile strength increased in all reinforced samples. Best hardness and tensile strength results were observed for AA6082/TiC AMC. Khodabakshi *et al* [21] fabricated Al-matrix nanocomposite ultra-fine grain by FSP. Al-Mg alloy matrix 3.5 vol% SiC and multiple FSP passes were done. Tensile properties improved and hardness increased for ultra-fine grained Al-matrix. Due to multiple passes, SiC was well distributed throughout the matrix. Salehi *et al* [22] by FSP fabricated AA6061/SiC nanocomposite and determined the factors that influenced ultimate tensile strength. Optimisation of parameters was done by Taguchi. The factor that influenced the most was rotational speed, followed by transverse speed, then pin profile and the factor that influenced the least is tool penetration depth. 1600 rpm rotational speed, 40 mm min^{-1} transverse speed, 0.3 mm tool penetration depth and threaded type tool profile were the optimum parameters obtained for maximum tensile strength.

Rathee *et al* [23] fabricated AA6061/SiC composite via FSP and to obtain maximum microhardness optimised the design of experiment. Maximum microhardness was obtained at nugget zone and optimum parameters obtained were 50 mm min^{-1} transverse speed, 2.50 tilt angle and 1400 rpm of tool rotational speed. Deepak *et al* [24] by FSP fabricated Al5083 with nano SiC as reinforcement and observed that the hardness of fabricated composite was more than of base metal, but wear resistance was less due to higher frictional force and coefficient of friction. Salehi *et al* [25] by FSP fabricated Al6061composite having SiC nanoparticles, with SiC wt% varying from 0%–18%. Microhardness of the composite was more than that of base metal. They also proposed that microhardness is more when inter-particles spacing is less. Kurt *et al* [26] by FSP fabricated commercially pure aluminium with SiC particles at different tool rotational and transverse rate. The decrease in grain size and increase in hardness was observed. 1000 RPM and 20 mm min^{-1} were the optimum parameters for the particle's distribution throughout the matrix. Paidar *et al* [27] reinforced AA5182 with WC using FSW at different tool traverse speed. They found that the composite had enhanced microhardness, wear-resistance and grain refinement at decreased traverse speed. Paidar *et al* [28] fabricated Al/B4C composite via FSP, by varying number of passes. Grain size and wear loss decreased, whereas UTS value increased with the increasing number of passes. Devaraju *et al* [29] by FSP, with SiC and Gr as reinforcement formed aluminium composite. To optimise vol% of reinforcement and tool rotational speed Taguchi was used in order to enhance the mechanical and tribological properties. Wear and hardness increased at respective optimum conditions but tensile properties were lower than base metal at its optimum conditions. Izadi *et al* [30] by FSP formed Al-SiC composite with vol% varying from 4–16 vol% and found that there were pores and lack of consolidation when processed 16 vol% SiC and all processed samples showed increased hardness and it increased with increase in SiC content. Adilokht *et al* [31] reinforced AA350 with SiC and MoS2 using FSP at constant tool rotational and travel speed. Hardness and wear resistance of processed composite was more than base metal due to the presence

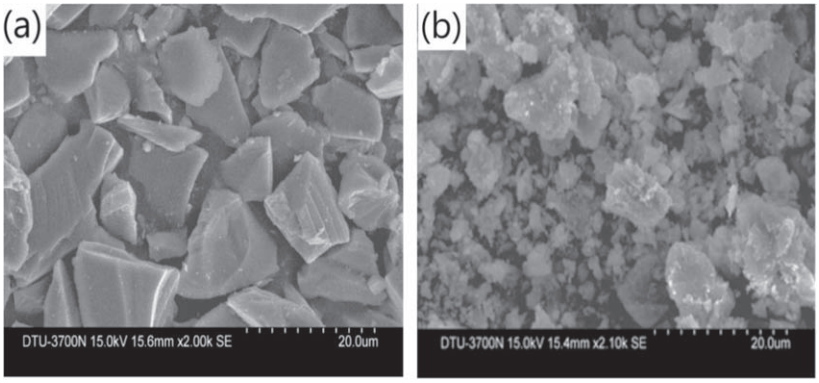


Figure 1. SEM images of (a) SiC particles at 20 μm (b) Aloevera ash particles at 20 μm.

Table 1. Composition of AA7075-T6.

Element	Si	Mg	Fe	Zn	Al
Weight % composition	1.514	1.664	1.744	4.156	Balance

of reinforcement. Butola *et al* [32] reinforced AA6061 with variety of agro-waste materials i.e. jute, bagasse and banana, in order to improve its mechanical properties by stir casting process. Mechanical properties of the material were enhanced with the introduction of these materials and the best results obtained were for composite reinforced with banana ash.

In the present study, mechanical and tribological properties of aluminium reinforced with ceramic and natural reinforcement (Aloevera ash) is studied. Effect of FSP tool rotational speed on the distribution of reinforcement particle is studied. One of the most important factors for any industry is the cost-effectiveness of its product without compromising with its properties. In manufacturing industry constant attempts are being made in order to produce a low-cost AMMC to achieve this several attempts have been made by researchers like taking bagasse ash, red mud, fly ash, rice husk etc as reinforcement [16]. Aloevera is less dense, eco-friendly, easily available material and it also costs less these properties of aloevera attracted us to use aloevera ash as reinforcement as it can be a good alternative for various reinforcements.

2. Experimental procedure

2.1. Matrix material

Aluminium alloy 7075-T6 is used as a matrix material in this study. On aluminium, process like FSP can easily be performed [33]. The plate of length 200 mm, width 80 mm and height 6 mm is used as the base plate. AA-7075-T6 has properties like high fatigue strength, creep resistance, high formability [34]. The chemical composition of aluminium 7075-T6 is mentioned in table 1.

2.2. Reinforcement

Powdered SiC and Aloevera ash are used as reinforcement. SiC due to its semiconductor characteristics possesses interesting electrical properties. Fine powder of SiC having a particle size of 40μm is used; SEM image of the SiC powder is shown in figure 1(a). Aloevera is a naturally occurring medicinal plant. It has high water content and is rich in nutrients, minerals and vitamins . It is easy to grow plant, which can easily be found throughout the world. The aloevera is first dried in sunlight and then it is burned. The powder left after burning of aloevera is aloevera ash, which is composed of preserved minerals like Ca, Cl, Cr, Cu, Fe, Mg, K, P, Na, Zn and Mn, most of which act as antioxidants [35, 36]. After burning the resulting white ash is collected and ball milled for 30 h in order to produce a fine powder of size 40 μm. The SEM image of the resulting aloevera ash used is shown in figure 1(b).

2.3. Fabrication of composites

All samples are fabricated using Friction stir processing at different rpm i.e. 600, 900. FSP is the method of fabricating composite by plastic deformation of the matrix surface by rotating a tool at very high rpm, the set up

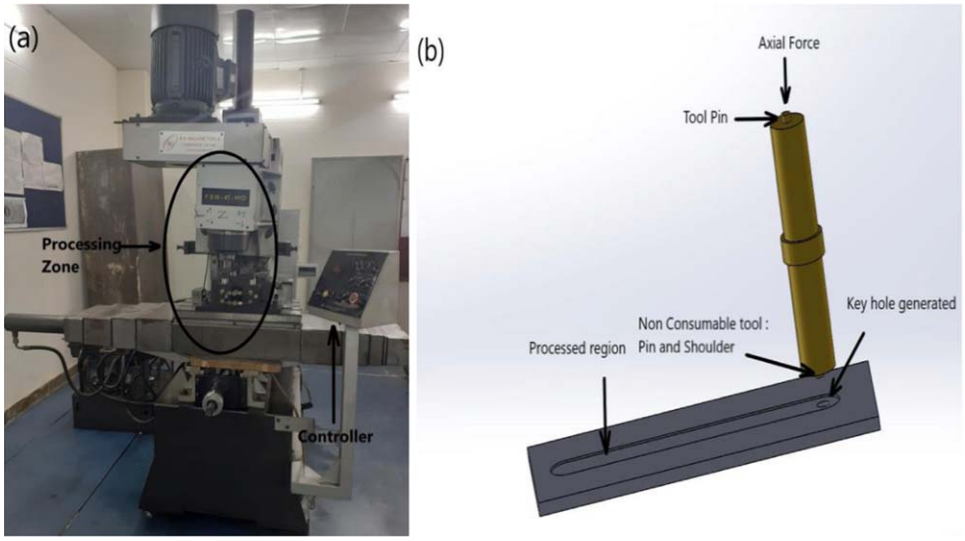


Figure 2. (a) FSP set up (b) FSP tool and processed plate design.

Table 2. Groove design.

Parameters	Values
Dimension of plate	200 * 80 * 6 (mm^3)
Groove length	160 mm
Groove width	2 mm
Groove depth	3.5 mm

Table 3. Process parameters.

Process parameters	Values
Tool material	H13 tool steel
Tool pin profile	Square
Tool shoulder profile	Flat
Tool pin length	3.5 mm
Tool shoulder diameter	19.95 mm
Tool plunge depth	0.20 mm
Tilt angle	2^0 constant
Reinforcement	SiC, Aloevera ash

can be seen in figure 2. Different properties of friction stir processed alloy are due to the different speed and angle of the FSP tool [37]. The square tip tool made of H13 tool steel of total length of 117 mm and tip length of 3.5 mm is used for fabricating the above samples, as square tip minimises the defects in the matrix, uniformly distributes the reinforcement and increases hardness [21, 38]. Summary of groove design is shown in table 2. H13 tool steel is also used in friction stir welding and manufacturing welding joints [39]. The other process parameters are mentioned in table 3. Even sized grooves of depth 3.5 mm and width 2 mm are machined on aluminium 7075-T6 plates. Various calculations related to grooves i.e. volume of fraction, area of groove and projected tool pin area are done in equations (1)–(3). In these grooves, reinforcement is filled in powdered form and by using FSP these are mixed.

$$\text{Area of groove} = \text{Depth of groove} * \text{Width of groove} \ldots \quad (1)$$

$$\text{Projection area of tool pin} = \text{diagonal of square base} * \text{tool pin length} \ldots \quad (2)$$

$$\text{Fraction Volume} = \frac{\text{Area of groove}}{\text{Projection area of tool pin}} \ldots \quad (3)$$

Sample 1 contains 2 wt.% SiC reinforcement with FSP done at 600 rpm. In, Sample 2 FSP is done at 600 rpm and SiC + Aloevera ash (2 wt.%) are used as reinforcements. In sample 3 and sample 4 FSP is done at 900 rpm,

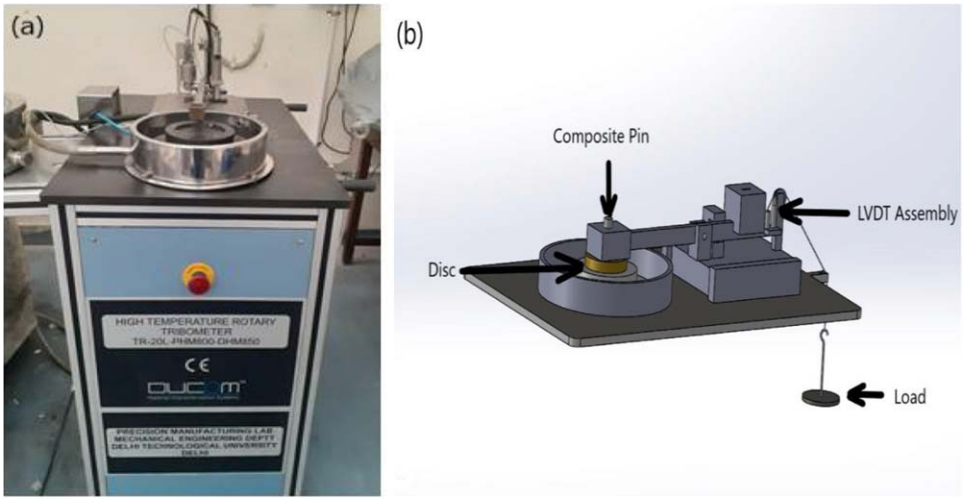


Figure 3. (a) High temperature rotary tribometer Setup (b) Schematic design of tribometer.

Table 4. Composition of tested samples.

Sample	Composition	Tool rpm
Sample 1	AA7075-T6 + 2wt% SiC	600
Sample 2	AA7075-T6 + 1wt% SiC + 1wt% Aloe-vera ash	600
Sample 3	AA7075-T6 + 2wt % SiC	900
Sample 4	AA7075-T6 + 1wt% SiC + 1wt% Aloe-vera ash	900

Table 5. Test conditions for pin on disc experiment.

S. No.	Test Conditions	Values
1.	Load	20 N, 30 N, 40 N
2.	Pin diameter	8 mm
3.	Pin height	6 mm
4.	Track diameter	30 mm, 40 mm, 50 mm
5.	Sliding distance	950 m
6.	Rpm	637, 478, 382
7.	Time	16 min
8.	Disc roughness	0.2 μm
9.	Disc hardness	57 HRC

used SiC (2 wt.%) and SiC + Aloevera ash (2 wt.%) as reinforcements respectively. Table 4 shows the chemical composition of all tested samples.

2.4. Pin on disc set up

Wear is studied using high temperature rotary tribometer manufactured by DUCOM. Disc of 100 mm diameter made of EN24 metal is used and this disc is cleaned using acetone before the experiment. Samples used to perform testing are taken in the shape of cylindrical pins with 8 mm diameter and 6 mm height. Test conditions and set up used for the pin on disc experiment are shown in table 5 and figure 3 respectively. Wear and coefficient of friction are calculated from the data and analysed in this study.

2.5. Tensile testing

Samples of length 80 mm and thickness 3.5 mm with flat base are prepared according to ASTM standard shape for the tensile testing, shown in figure 4. Samples are gripped from both sides and a constantly increasing load is applied until the breaking point is reached. Changes in the sample are recorded throughout the testing and then

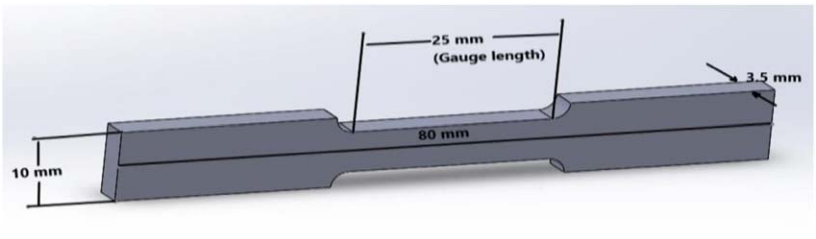


Figure 4. Design of tensile sample.

analysed. All samples i.e. base metal, sample 1, sample 2, sample 3, sample 4 has been tested and the parameters recorded were ultimate tensile strength and stress-strain curve.

2.6. Microhardness

Microhardness is tested using Fischer scope HM2000 S microhardness tester according to ASTM: E8/E8M-011 standards. By analysing the indentation depth formed by indenter due to a constant force of 200.5 KN, applied on it for 20 s, microhardness value at a point is obtained. Microhardness is examined for 4 different points on a sample and an average of them is considered as the true microhardness of the sample. Microhardness is analysed for base metal, sample 1, sample 2, sample 3 and sample 4.

2.7. Wear morphology

The worn-out samples are examined using Scanning Electron Microscope (SEM) focused electron beam is used to produce the image in order to examine the wear surface of the samples. SEM images are produced in order to analyse the wear effect and Energy Dispersive x-ray Spectroscopy (EDS or EDX) is done to find the compound present after wear. EDS basic principle is that every element has a unique atomic structure, it works on interaction between the x-rays and samples.

3. Results and discussion

3.1. Wear

From figure 5 it can be concluded that with increasing load wear increases for all samples, the increase is different for samples depending upon the reinforcement and tool rpm of FSP. Similar pattern was observed when SiC and Al_2O_3 were reinforced with aluminium matrix and an increase in wear volume was observed with an increase in load [40].

With the addition of reinforcement wear decreases, as load is distributed on uniformly scattered hard ceramic particles. Base metal shows the maximum wear loss for all tested load. Of all tested samples, minimum wear at 20 N load is observed for Al+SiC/Aloevera ash processed at 600 tool rpm, due to less ageing and formation of the oxide layer which acts as a solid lubricant and reduces wear [41]. For 30 N and 40 N load, optimum results are given by Al+SiC processed at 900 tool rpm, as hard ceramic SiC particles are properly dispersed throughout the matrix and prevent direct contact of aluminium with the disc. On comparing wear loss of Al+SiC composite i.e. sample 1 and sample 3, it can be observed that wear is less for sample processed at 900 rpm than sample processed at 600 rpm for all tested loads. This is due to proper dispersion of fine hard SiC particles in the matrix at higher tool rpm, with the increased FSP tool rpm number of cavities and defects in the metal matrix are minimum [42]. For Al+SiC/Aloevera ash composites i.e. sample 2 and sample 4, the samples processed at 600 rpm showed less wear than sample processed at 900 rpm. This might be attributed to the softening of the matrix because of heat generated in the nugget zone, due to over ageing it causes matrix softening [43].

3.2. Coefficient of friction with sliding distance

Variation of coefficient of friction as a function sliding distance of AMMCs is displayed in figure 6, the total sliding distance was 950 m. There is a large rise in the initial coefficient of friction value of the base metal. This can be due to the large force required to overcome the adhesion between the contact surfaces [40]. After 100 m there is large fluctuation and friction coefficient value goes down and fluctuates within the limits. Ceramic SiC particles dispersed throughout the matrix assist the load and reduce the contact between the aluminium surface and disc surface. There is a sudden rise in the friction coefficient value of Al+SiC composite i.e. sample 1 and 3. The reinforcement particles are very well scattered throughout the matrix and when the disc surface comes in

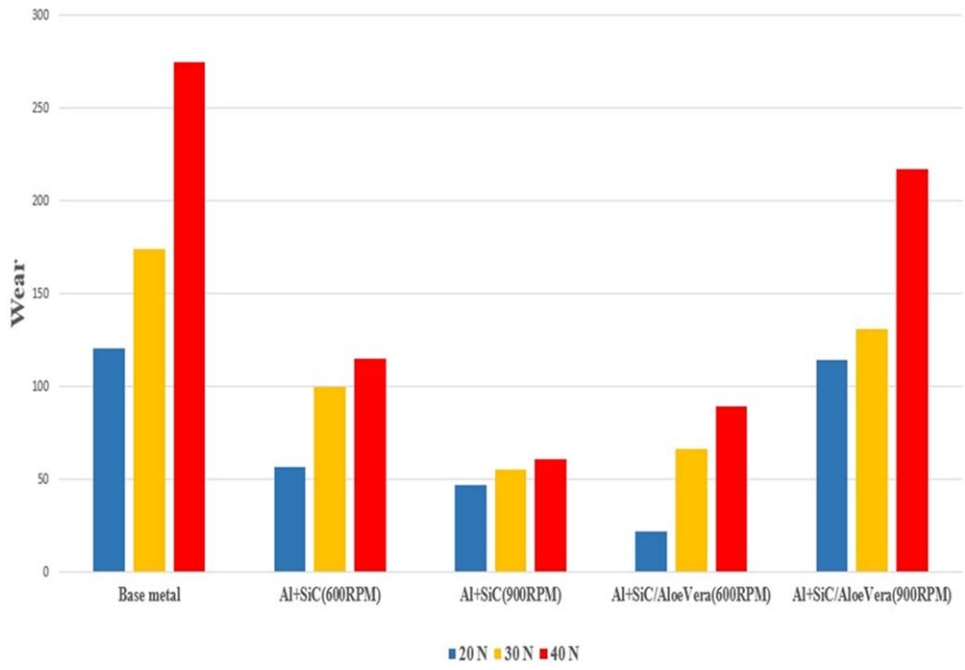


Figure 5. Wear rate graph of all worn out samples.

contact with these well dispersed hard SiC particles the coefficient of friction increases and this results in less wear loss and more wear resistance [44]. Sudden fluctuation can be seen in coefficient of friction of Al+SiC/Aloevera ash composite for both 600 and 900 tool rpm. This fluctuation is due to ploughing off of oxide layer which results in the generation of surface pits. Presence of these pits causes mismatching of the surface resulting in fluctuating value of coefficient of friction. SiC prevent the drastic metal removal and ploughing off of layers hence preventing pits formation, so less fluctuation in the value of friction coefficient is recorded [44].

3.3. Effect of load

Variation of coefficient of friction with load can be seen in figure 7(a), it is more pronounced at lower load. Base metal shows the highest coefficient of friction at all tested loads. Coefficient of friction increases with increasing load. For composite processed at higher tool rpm there is a consistent increase in the coefficient of friction with increasing load. Samples containing aloevera ash as one of their reinforcement i.e. sample 2 and 4, showed comparatively lower coefficient of friction due to the formation of the oxide layer, it acts as a lubricating hence reduces the coefficient of friction. Effect of load on wear rate is shown in figure 7(b). It can be concluded that with increasing load wear rate increases and base metal shows the highest wear rate at all tested load. Due to well distribution of SiC particles throughout the matrix sample 3 has least wear rate at 30 N and 40 N and sample 2 showed minimum wear rate at 20 N due to formation of a lubricating oxide layer. Sample 4 showed the highest wear rate of all reinforced sample justifying the matrix softening due to ageing by FSP.

3.4. Tensile properties

Selection of engineering material is an important task and it all depends on the properties of the material. The tensile test is important in determining the properties of a material. Tensile strength is tested for the base metal as well as for processed composites and, on the basis of the readings obtained the stress-strain curve of the respected samples is presented in figure 8.

From figure 8 it can be observed that ductility increased for all fabricated composite. The composites fabricated at higher tool rpm i.e. sample 3 and sample 4, are more ductile than composite fabricated at low rpm i.e. sample 1 and sample 2. It can be summarised that ductility and UTS are enhanced with addition of reinforcements (SiC and aloevera ash), and further enhancement was observed on processing the material at higher tool rpm. Similar type of results were reported by Zhou *et al* [45] when reinforced Al7085 with TiC via FSP. Arab *et al* [46] in their study showed that ductility improved when AA1100 was FSPed without any reinforcement, and further enhanced when AA1100 was reinforced with various novel fibres through FSP, reflecting that reinforcement plays a vital role in enhancing the properties along with the process.

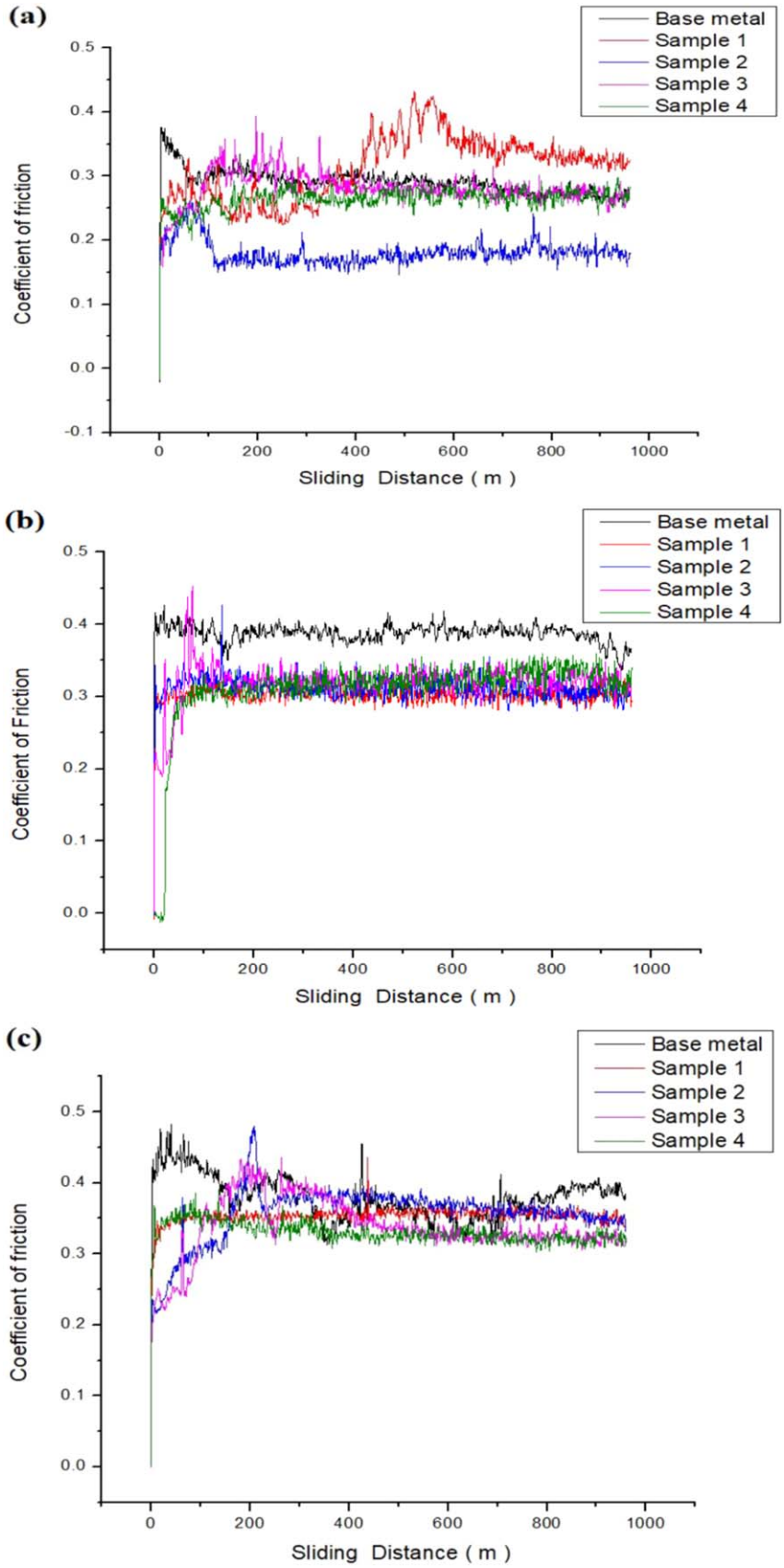


Figure 6. Coefficient of friction of samples at (a) 20 N (b) 30 N (c) 40 N.

3.5. Microhardness

Hardness refers to the property of the material to resist the deformation due to the force applied. Microhardness of all samples is shown in figure 9. Microhardness for base metal is 102 Hv and for sample 1 is 121.3 Hv which is

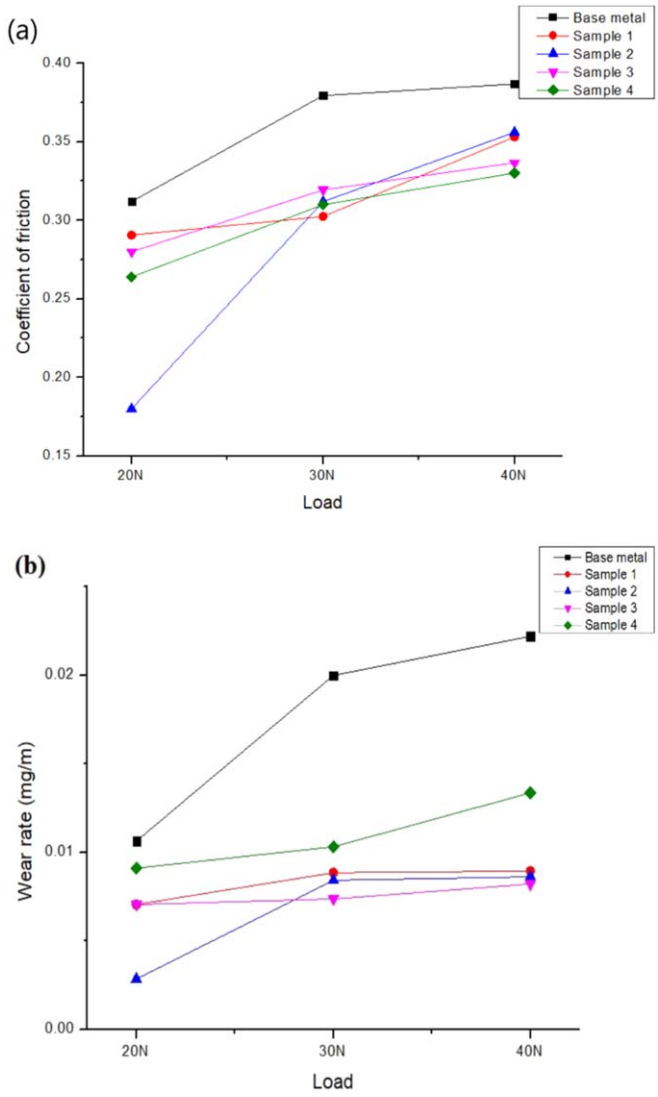


Figure 7. Effect of load at (a) coefficient of friction (b) wear rate.

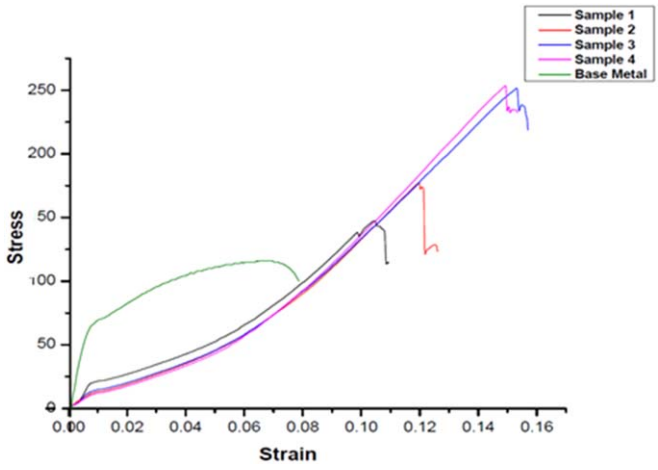


Figure 8. Stress-Strain curve of processed sample and base metal.

19% more than base metal, as well dispersed hard reinforcement particles support the matrix base and thus increase the microhardness [29]. Sample 2 has 18% more microhardness than base metal due to the presence of well-dispersed reinforcement particle. For sample 3 microhardness value is 152 H and for sample 4 it is 137.82

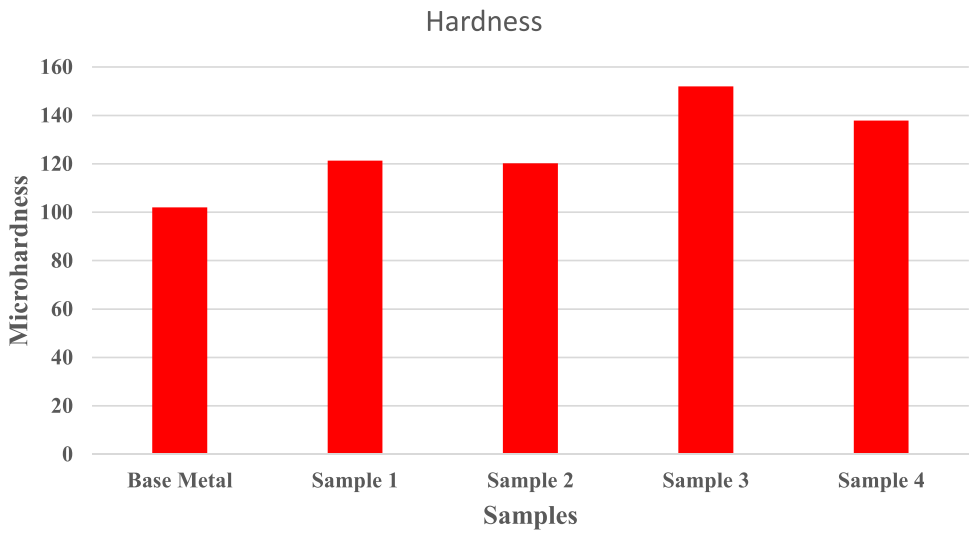


Figure 9. Microhardness of all samples.

which is more than the recorded value for base metal, sample 1 and sample 2. It can be observed that on adding reinforcement microhardness of material increases and it further increases on increasing the tool rpm on which it is processed, as material FSPed at higher tool rpm have finer grains than those FSPed at lower rpm [21] and microhardness is enhanced due to uniform distribution of these fine grains throughout the matrix [26].

3.6. Wear morphology

SEM analysis is done in order to discern the effect of reinforcements, tool rotational speed and applied load on wear behaviour of the fabricated AMMCs. Figures 10(a)–(h) illustrates the worn surface of all fabricated composites at different applied load i.e. 20 N and 40 N. Wear tracks can be seen in the SEM images (figures 10(b)–(e), (h)) and the white particles in the images shows the formation of carbide in the samples (figures 10(b), (d), (e), (g)) during wear of disc materials. Traces of adhesive wear can be seen in the images majorly at higher applied load. Figure 11 shows the EDS analysis of the fabricated composite, it is done in order to find the constituent of the samples after wear. Presence of C, O, Mg, Al, Si, Fe, and Zn can be seen in the EDS of the wear debris. Presence of oxygen confirms the formation of oxide tribolayer. This tribolayer is formed by the reaction of oxygen and aluminium or iron. This reaction might have occurred during the wear test due to the large heat generation between the surfaces. Fe in EDS signifies the formation of Fe layer which also acts as a solid lubricant, hence reducing wear and coefficient of friction. Such high content of Fe in the samples is due to the transfer of elements between wear pin and disc in the early phase of wear testing. Similar types of results were reflected by Alam *et al* and Coyal *et al* [47, 48] in their study.

4. Conclusions

- Wear resistance of all fabricated composite was more than that of base metal.
- At 20N applied load Al+SiC/Aloevera ash composite processed at 600 tool rpm showed the least wear loss, due to oxide tribolayer formation. At higher applied load Al+SiC, processed at 900 rpm had the best wear resistance, as at higher tool rpm there is proper dispersion of reinforcement.
- Wear rate and coefficient of friction both increases with increasing the applied load.
- Of all tested samples base metal had the maximum value for both coefficient of friction and wear rate for all applied load.
- Ductility and ultimate tensile strength increased with the addition of reinforcement.
- processed at higher tool rpm are more ductile and have higher ultimate tensile strength than those processed at lower tool rpm.
- Microhardness increased with the introduction of reinforcement. Microhardness is more for samples processed at higher tool rpm due to the finer grains obtained because of high tool rotation.
- SEM analysis of wear surface shows the proper distribution of reinforcement and traces of adhesive wear.

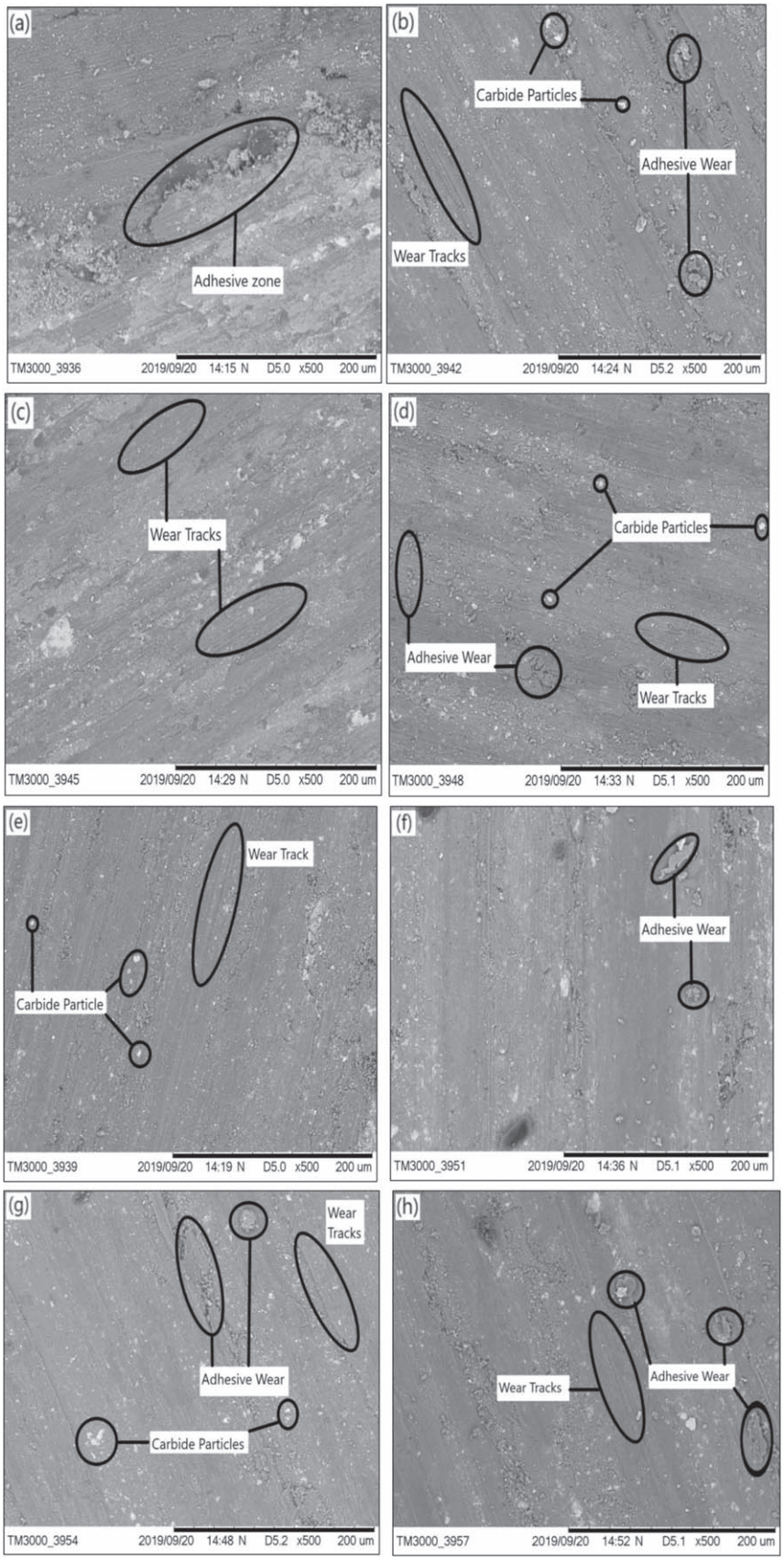


Figure 10. (a) SEM of sample 1(20 N load) (b) SEM of sample 1(40 N load) (c) SEM of sample 2(20 N load) (d) SEM of sample 2(40 N load) (e) SEM of sample 3(20 N load) (f) SEM of sample 3(40 N load) (g) SEM of sample 4(20 N load) (h) SEM of sample 4(40 N load).

- EDS of the sample after wear indicates the presence of C, O, Fe, Zn, Si and Al, confirming the formation of oxide tribolayer and validation the reduced wear loss and coefficient of friction.

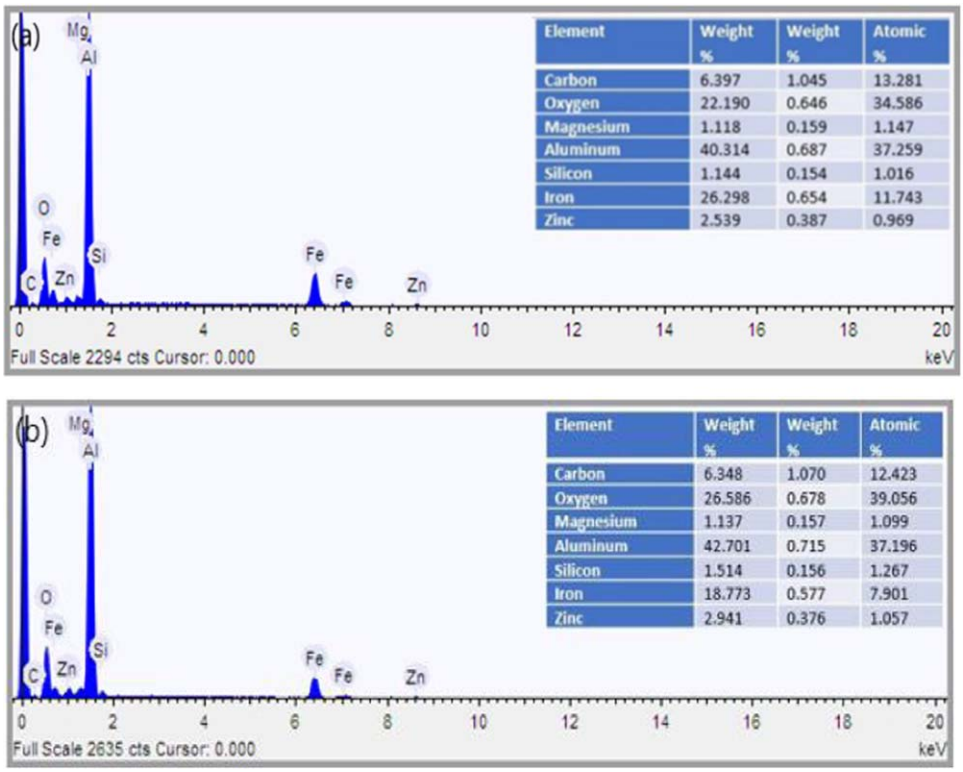


Figure 11. EDS analysis of worn out surface (a) sample 1 (b) sample 2.

ORCID iDs

Lakshay Tyagi https://orcid.org/0000-0001-5739-9807
Ravi Butola https://orcid.org/0000-0002-6581-4131

References

[1] Abenojar J, Velasco F and Martínez M A 2007 Optimization of processing parameters for the Al+10% B4C system obtained by mechanical alloying *J. Mater. Process. Technol.* **184** 441–6
[2] Butola R, Singari R M, Bandhu A and Walia R S 2017 Characteristics and properties of different reinforcements in hybrid aluminium composites: a review *IJAPIE-SI-MM* **511** 71–80
[3] Zawawi N N M, Azmi W H, Redhwan A A M and Sharif M Z 2017 Coefficient of friction and wear rate effects of different composite nanolubricant concentrations on aluminium 2024 plate *IOP Conf. Series: Materials Science and Engineering* vol 257
[4] Naveed M and Khan A R A 2016 Dry sliding wear of heat treated hybrid metal matrix composites *IOP Conf. Series: Materials Science and Engineering* vol 149
[5] Raghavendra C R, Basavarajappa S and Sogalad I 2018 Study on influence of Surface roughness of Ni-Al2O3 nano composite coating and evaluation of wear characteristics *IOP Conf. Series: Materials Science and Engineering* vol 310
[6] Razzaq A M, Majid D L, Manan N H, Ishak M R and Basheer U M 2018 Effect of Fly ash content and applied load on wear behaviour of AA6063 aluminium alloy *IOP Conf. Series: Materials Science and Engineering* vol 429
[7] Nagaral M, Auradi V, Parashivamurthy K I, Shivananda B K and Kori S A 2018 Dry sliding wear behaviour of aluminium 6061-SiC-graphite particulates reinforced hybrid composites *IOP Conf. Series: Materials Science and Engineering* vol 310
[8] Hima Gireesh C and Ramji K 2019 Experimental Investigation on Mechanical Properties of an Al6061 Hybrid Metal Matrix Composite *International Journal of Engineering and Advanced Technology (IJEAT)* **2** 1519–22
[9] Mittal P 2018 Wear behaviour of aluminium 7075 based composites reinforced with SiC, red mud and Al2O3 *International Journal for Research in Applied Science and Engineering Technology* **6** 2802–16
[10] Baradeswaran A and Elaya Perumal A 2013 Influence of B4C on the tribological and mechanical properties of Al 7075– B4C composites *Composites Part B: Engineering* **54** 146–52
[11] Rahman M and Rashed H 2014 Characterization of silicon carbide reinforced aluminum matrix composites *Procedia Engineering* **90** 103–9
[12] Sun N and Apelian D 2018 Friction stir processing of aluminum alloy A206: Part I—Microstructure evolution *Int. J. Metalcast.* **13** 234–43
[13] Sun N, Jones W and Apelian D 2018 Friction stir processing of aluminum alloy A206: Part II—tensile and fatigue properties *Int. J. Metalcast.* **13** 244–54
[14] Singla M, Singh L and Chawla V 2009 Study of wear properties of Al-SiC composites *J. Miner. Mater. Charact. Eng.* **08** 813–21
[15] Alizadeh A and Taheri-Nassaj E 2012 Mechanical properties and wear behavior of Al–2wt.% Cu alloy composites reinforced by B4C nanoparticles and fabricated by mechanical milling and hot extrusion *Mater. Charact.* **67** 119–28

[16] Hima Gireesh C, Durga Prasad K, Ramji K and Vinay P 2018 Mechanical characterization of aluminium metal matrix composite reinforced with aloe vera powder *Materials Today: Proceedings* **5** 3289–97
[17] Butola R, Singari R M and Murtaza Q 2019 Fabrication and optimization of AA7075 matrix surface composites using Taguchi technique via Friction stir processing (FSP) *Engineering Research Express* **1** 025015
[18] Miranda R, Santos T, Gandra J, Lopes N and Silva R 2013 Reinforcement strategies for producing functionally graded materials by friction stir processing in aluminium alloys *J. Mater. Process. Technol.* **213** 1609–15
[19] Vijayavel P, Balasubramanian V and Sundaram S 2014 Effect of shoulder diameter to pin diameter (D/d) ratio on tensile strength and ductility of friction stir processed LM25AA-5% SiCp metal matrix composites *Materials & Design* **57** 1–9
[20] Dinaharan I 2016 Influence of ceramic particulate type on microstructure and tensile strength of aluminum matrix composites produced using friction stir processing *Journal of Asian Ceramic Societies* **4** 209–18
[21] Khojastehnezhad V M, Pourasl H H and Barenji R V 2017 Effect of tool pin profile on the microstructure and mechanical properties of friction stir processed Al6061/Al2O3—TiB2 surface hybrid composite layer *Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications* **233** 900–12
[22] Salehi M, Saadatmand M and Aghazadeh Mohandesi J 2012 Optimization of process parameters for producing AA6061/SiC nanocomposites by friction stir processing *Transactions of Nonferrous Metals Society of China* **22** 1055–63
[23] Rathee S, Maheshwari S, Noor Siddiquee A, Srivastava M and Kumar Sharma S 2016 Process parameters optimization for enhanced microhardness of AA 6061/ SiC surface composites fabricated via Friction Stir Processing (FSP) *Materials Today: Proceedings* **3** 4151–6
[24] Deepak D, Sidhu R S and Gupta V K 2013 Preparation of 5083 Al-SiC surface composite by friction stir processing and its mechanical characterization *International Journal of Mechanical Engineering* **3** 1–11
[25] Salehi M, Farnoush H and Mohandesi J 2014 Fabrication and characterization of functionally graded Al–SiC nanocomposite by using a novel multistep friction stir processing *Mater. Des.* **63** 419–26
[26] Kurt A, Uygur I and Cete E 2011 Surface modification of aluminium by friction stir processing *J. Mater. Process. Technol.* **211** 313–7
[27] Paidar M, Asgari A, Ojo O and Saberi A 2018 Mechanical properties and wear behavior of AA5182/WC nanocomposite fabricated by friction stir welding at different tool traverse speeds *J. Mater. Eng. Perform.* **27** 1714–24
[28] Paidar M, Ojo O, Ezatpour H and Heidarzadeh A 2019 Influence of multi-pass FSP on the microstructure, mechanical properties and tribological characterization of Al/B4C composite fabricated by accumulative roll bonding (ARB) *Surf. Coat. Technol.* **361** 159–69
[29] Devaraju A, Kumar A and Kotiveerachari B 2013 Influence of rotational speed and reinforcements on wear and mechanical properties of aluminum hybrid composites via friction stir processing *Mater. Des.* **45** 576–85
[30] Izadi H, Nolting A, Munro C, Bishop D, Plucknett K and Gerlich A 2013 Friction stir processing of Al/SiC composites fabricated by powder metallurgy *J. Mater. Process. Technol.* **213** 1900–7
[31] Alidokht S, Abdollah-zadeh A, Soleymani S and Assadi H 2011 Microstructure and tribological performance of an aluminium alloy based hybrid composite produced by friction stir processing *Mater. Des.* **32** 2727–33
[32] Butola R, Pratap C, Shukla A and Walia R 2019 Effect on the mechanical properties of aluminum-based hybrid metal matrix composite using stir casting method *Material Science Forum* **969** 253–9
[33] Butola R, Malhotra A, Yadav M, Singari R, Murtaza Q and Chandra P 2019 Experimental studies on mechanical properties of metal matrix composites reinforced with natural fibres ashes *SAE Technical Paper Series* - 11
[34] Ikumapayi O M, Akinlabi E T and Majumdar J D 2019 Influence of carbonaceous agrowastes nanoparticles on physical and Mechanical properties of friction stir processed AA7075-T651 metal matrix composites *Surf. Topogr.: Metrol. Prop.* **7** 035013
[35] Haque M, Islam M, Jalil M and Shafique M 2014 Proximate analysis of aloe vara leaves *IOSR Journal of Applied Chemistry* **7** 36–40
[36] 2016 Some Drugs And Herbal Products (Volume 108) (France: IARC Monographs)
[37] Chaudhary A, Kumar Dev A, Goel A, Butola R and Ranganath M 2018 The mechanical properties of different alloys in friction stir processing: a review *Materials Today: Proceedings* **5** 5553–62
[38] Moradi Faradonbeh A, Shamanian M, Edris H, Paidar M and Bozkurt Y 2018 Friction stir welding of Al-B4C composite fabricated by accumulative roll bonding: evaluation of microstructure and mechanical behavior *J. Mater. Eng. Perform.* **27** 835–46
[39] Butola R, Murtaza Q and Singari R M 2019 Advances in computational methods in manufacturing *Lecture Notes on Multidisciplinary Industrial Engineering*. (Singapore: Springer) ed R Narayanan *et al* pp 337–48
[40] Mahmoud E, Takahashi M, Shibayanagi T and Ikeuchi K 2010 Wear characteristics of surface-hybrid-MMCs layer fabricated on aluminum plate by Friction stir processing *Wear* **268** 1111–21
[41] Yuvaraj N, Aravindan S and Vipin 2016 Wear characteristics of Al5083 surface hybrid nano-composites by friction stir processing *Trans. Indian Inst. Met.* **70** 1111–29
[42] Raaft M, Mahmoud T, Zakaria H and Khalifa T 2011 Microstructural, mechanical and wear behavior of A390/graphite and A390/Al2O3 surface composites fabricated using FSP *Materials Science and Engineering:* A **528** 5741–6
[43] Devaraju A, Kumar A, Kumaraswamy A and Kotiveerachari B 2013 Influence of reinforcements (SiC and Al2O3) and rotational speed on wear and mechanical properties of aluminum alloy 6061-T6 based surface hybrid composites produced via Friction stir processing *Mater. Des.* **51** 331–41
[44] Dolatkhah A, Golbabaei P, Besharati Givi M and Molaiekiya F 2012 Investigating effects of process parameters on microstructural and mechanical properties of Al5052/SiC metal matrix composite fabricated via Friction stir processing *Mater. Des.* **37** 458–64
[45] Zhou D, Wang J, Lu Y and Bai Z 2020 Microstructure evolution and mechanical properties of friction stir processed TiC/7085Al nanocomposites *Mater. Res. Express* **7** 036515
[46] Arab S, Karimi S, Jahromi S, Javadpour S and Zebarjad S 2015 Fabrication of novel fiber reinforced aluminum composites by friction stir processing *Materials Science and Engineering:* A **632** 50–7
[47] Alam M, Arif S and Ansari A 2018 Wear behaviour and morphology of stir cast aluminium/SiC nanocomposites *Mater. Res. Express* **5** 15
[48] Coyal A, Yuvaraj N, Butola R and Tyagi L 2020 An experimental analysis of tensile, hardness and wear properties of aluminium metal matrix composite through stir casting process *SN Applied Sciences* **2** 892

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/335029636

Modeling and Development of an Autonomous Underwater Vehicle ARYA for Object Recognition

Research · July 2019

DOI: 10.35940/ijrte.B2843.078219

CITATIONS
0

READS
38

6 authors, including:



Vipul Garg
Delhi Technological University
3 PUBLICATIONS **1** CITATION
SEE PROFILE



Deepanshu Bansal
Delhi Technological University
1 PUBLICATION **0** CITATIONS
SEE PROFILE



Aakash Kumar
Delhi Technological University
3 PUBLICATIONS **2** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Modeling and Development of an Autonomous Underwater Vehicle ARYA for Object Recognition View project

Design and Development of a seven DoF Semi-Autonomous Robotic Manipulator View project

All content following this page was uploaded by Vipul Garg on 23 May 2020.

The user has requested enhancement of the downloaded file.

Modeling and Development of an Autonomous Underwater Vehicle ARYA for Object Recognition

Aditya Natu, Anuj Badhwar, Vipul Garg, Upasana Biswas, Deepanshu Bansal, Aakash Kumar

Abstract: ***Arya is an autonomous underwater vehicle (AUV) modeled and developed by team DTU-AUV comprising of undergraduate students from multidisciplinary backgrounds of Delhi Technological University (DTU), India, to participate in an IEEE backed Singapore AUV Challenge (SAUVC). This paper entails the rationale and methodology employed to design and integrate various systems onboard. Significant improvisations have been made in the structural design of the vehicle to enhance its hydrodynamic stability and maneuverability to perform discrete tasks in comparison to the previous vehicles developed by the team. The focus is laid on the embedded and power system to enhance reliability, modularity, and power distribution. The software stack is designed to run in decentralized multi-threaded agent architecture, with the threads handling pressure sensor, cameras, control system, IMU, mission planner each performing input and output operations in continuous loops. PID control algorithms achieve the desired dynamic control. The vision system is devised to monitor the marine environment and detect underwater contoured objects.***

Index Terms: AUV, Design, Power System, PID, ROS, Vision.

I. INTRODUCTION

The uninterrupted industrialization in developing countries has not only pushed the development to the shores of the vast water bodies but also increased the dependency on the same for generating hydropower to satisfy the surge in energy demands of the industries and humanity. These majorly include the bridged structures for transport, dams to spawn the hydroelectric energy and the pipelines for supplying oil and gas. The reliability of these structures over a significant period calls for regular monitoring and subsequent maintenance to prevent catastrophic events. The hefty costs involved in such crewed missions demand cost-effective solutions for long term usage. Autonomous systems assist in setting up such large-scale monitoring programs, which not only reduces human interference in the job but also expands the reach envelope for successfully carrying out the mission. In recent decades, various research and experimental activities have been carried out to develop such autonomous systems, which have shown the potential to execute tasks with ease and augmented accessibility.

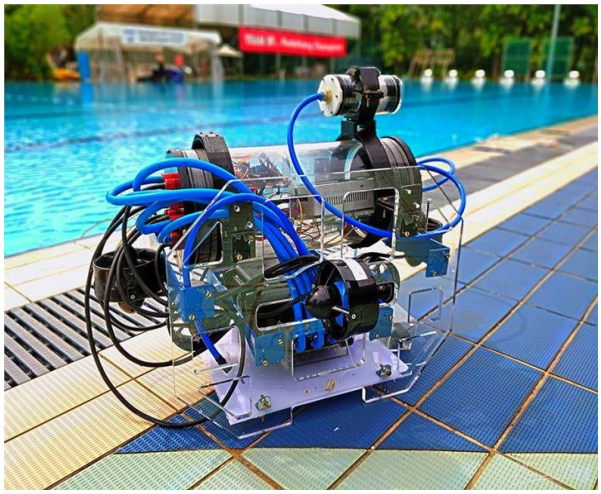


Figure 1: ARYA AUV

A pipeline detection autonomous underwater vehicle (PDAUV), having an advanced vision and acoustical system, was developed recently to inspect the unsatisfactory conditions of damns in China [1]. The severity of the risk associated with the cracks propagating aggressively over the dam walls can be predicted by identifying, evaluating, and categorizing the various cracks based on their length and depth [2]. A hybrid AUV/ROV named TriMARES is designed and developed for periodic monitoring of large dam reservoir [3]. Photomosaic of the walls of the dam has been generated by blending the different images gathered during the inspection [4]. Stereo vision based 3D structural map has been produced for underwater structures for review [5].

This project has been taken up with a similar objective of designing and developing a compact AUV with efficient vision algorithms for detection of contoured structures in the marine environment. The vehicle has been modeled based on foundational design concepts of underwater vehicles while seeking motivation to improvise on the previous vehicle developed by the team, ZYRA [6]. An elegant electronic system is placed onboard, enhancing the reliability and accessibility of the system to debug and repair. Vision algorithms have been devised on a ROS framework for better communication and integration.

Revised Manuscript Received on July 21, 2019.

Aditya Natu, Department of Mechanical Engineering, Delhi Technological University, Delhi, India.
Anuj Badhwar, Department of Applied Mathematics, Delhi Technological University, Delhi, India.
Vipul Garg, Department of Mechanical Engineering, Delhi Technological University, Delhi, India.
Upasana Biswas, Department of Electrical Engineering, Delhi Technological University, Delhi, India.
Deepanshu Bansal, Department of Electrical Engineering, Delhi Technological University, Delhi, India.
Aakash Kumar, Department of Mechanical Engineering, Delhi Technological University, Delhi, India.

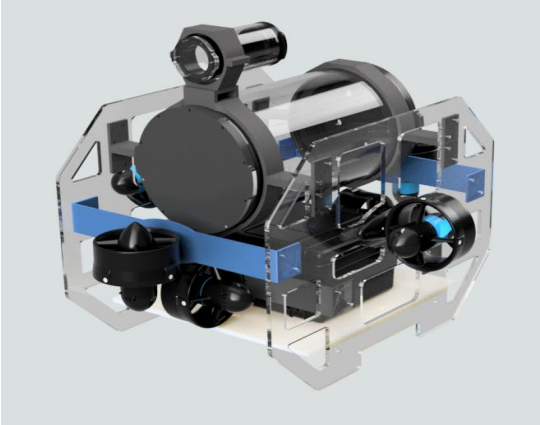


Figure 2: CAD Model of ARYA

This paper discusses the in-depth description of each system designed and integrated into the vehicle. The first section represents the structural design of the vehicle along with validated finite element analysis to assess the structural performance. The successive section describes the electronic stack developed to power and operate the entire modular system. The last section talks about the software framework devised to implement distinctive vision algorithms for precisely carrying out the detection related missions.

II. MECHANICAL DESIGN

The mechanical system of ARYA majorly comprises of the structural framework and the watertight hulls to house the electronic system on-board. The rationale behind the design methodology is to develop a stable vehicle with positive buoyancy, which assists in modeling and operating the control system of the AUV. This is achieved by keeping the desired aspect ratio for hydrodynamic stability and optimally placing the components to maintain the desired relative positions of the center of gravity and buoyancy. The vehicle adopts an open-frame structure to reduce the drag force experienced by it. The design of each subsystem is discussed below in detail.

A. Structural Frame

The AUV employs an acrylic framework to support all the subsystems rigidly. The design of the frame complies to develop a lightweight structure without compromising of the structural strength. The mentioned material is chosen based on its mechanical properties such as strength-to-weight ratio and corrosion resistant abilities. The modeled design has been critically analyzed for its structural performance at greater depths, with emphasis laid on the induced stresses and the factor of safety.

The frame is topologically optimized for trimming the material concerning locations of insignificant stress concentration. The laser cut process efficiently manufactured it and the assembly is done with the mounts of other subsystems using L-shaped brackets and Allen bolts, made of stainless steel, along with nylon lock nuts to enhance the rigidity of the structure and prevent undue vibrational effects of the system due to loosening of nuts. These bolts have lever arch mechanisms facilitates to apply increased tightening torque [7].

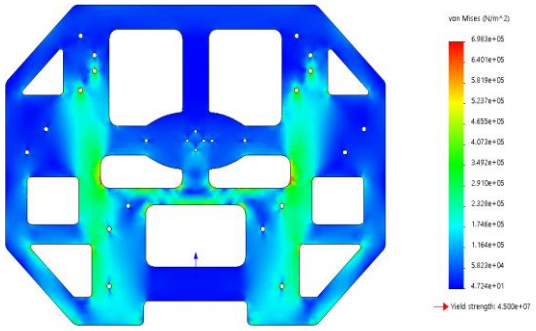


Figure 3: Stress Analysis of Frame

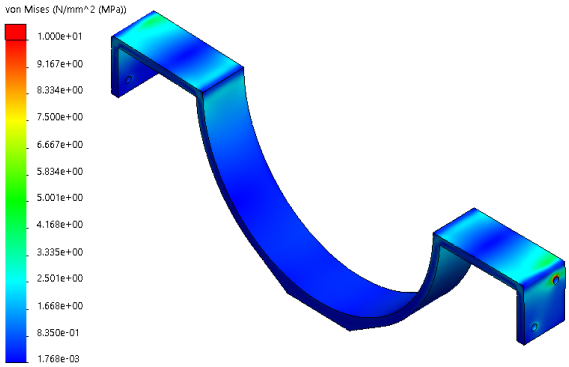


Figure 4: Stress Analysis of Enclosure Clamp

B. Pressure Hulls

The need to enclose the delicate cum costly electronic constituents, namely the sensors and the processing unit calls for developing watertight enclosures that can sustain immense pressure magnitudes. The AUV utilizes cylindrical canisters of 3 and 6-inch diameters, for housing camera and electronic stack respectively, to satisfy the system requirements, specifically due to the uniform pressure distribution over its surface and effective volume utilization [8]. The hulls are made up of hollow acrylic tube, at the end of which aluminum flanges are inserted. End Caps, having intricate cutouts to mount the connectors, are effectively clamped on to the flanges with the help of retaining bolts. O-rings are placed at both facial and radial grooves machined in the flange. These provide an effective seal under varying differential pressures. The batteries are placed in the pelican case set below the electronic hull. The canisters and case cumulatively account for the source of buoyancy, which assists in retrieving the vehicle in case of any sudden failure.

C. Propulsion System

Six T100 Blue Robotics thrusters are used to propel the AUV. The thrusters are placed parallel to the translational axes, namely the surge, heave, and sway. They are oriented as pairs to provide either a couple or combined translational force to maneuver the AUV. The vehicle has five degrees of freedom ruling out the roll motion, which is not required because of the stability the AUV achieves due to the symmetrical and balanced placement of components about the surge axis. This orientation allows for proper

maneuverability and controllability of the AUV through asymmetrical thrusting and varying the amount of thrust.

The propulsive efficiency of the vehicle is proportional to the power required to propel, which itself depends on the resistance it experiences during its motion [9]. The open frame enhances this performance characteristic of the vehicle by reducing the frontal area thus reducing the magnitude of pressure drag and also offers smooth flow trajectories of fluid particles through the body which prevents any unwanted deviations on the motion. The validation of the ideology mentioned above is done by performing a CFD analysis of the assembled system of ARYA. The resultant velocity and pressure contours are represented below in Fig.5 & Fig.6.

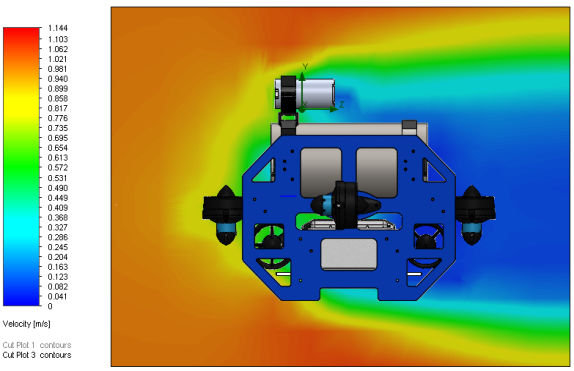


Figure 5: Velocity Contour for Surge Motion

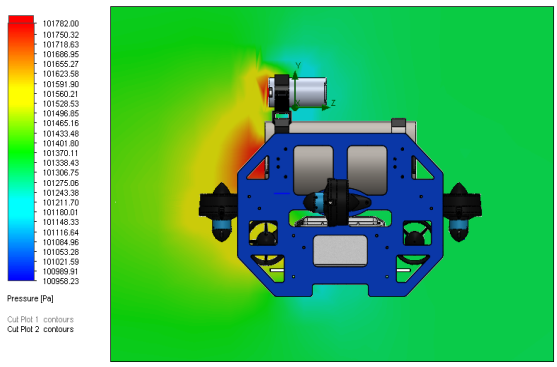


Figure 6: Pressure Contour for Surge Motion

III. EMBEDDED SYSTEMS.

A. Power System

Li-Po batteries have emerged as a promising rechargeable power source for the applications which required high energy density. It is deduced from [10] that the Lithium-ion batteries have relatively very high specific power as compared to other alternatives. They have real advantages which can be utilized by the AUV industry. Low self-discharge increases the life of the battery. Considering all the positive aspects of Lithium-ion batteries, 3 Lithium polymer batteries with 8000mAh current rating and 11.1V operating voltage have been used in ARYA to power the overall system of the vehicle. The power is distributed in a way such that the crucial components will not shut down unless the mission is completed and therefore two batteries are allocated for the actuators, and one battery is used to power the single board computer, microcontroller, and- sensors. It reduces the load from the power system and ensures the proper functioning of each subsystem.

B. Power Distribution System

It becomes necessary to provide an optimum operating voltage to each component of the integrated system to maximize efficiency. It can be achieved only through an efficient power distribution system which takes a tiny part for its functioning. To reduce the complexity of the electronic structure, highly power optimized step-down and step-up converters have been used. To power the NUC, an adjustable 150W DC-DC boost converter is employed to provide a 19V consistently. Universal Battery Eliminator Circuit (UBEC) has been interfaced with Arduino mega, the microcontroller of the ARYA, as it has in-built filter capacitors. The operating voltage for the actuator of the dropper mechanism is achieved through LM2596 step-down DC-DC voltage converter.

C. Battery Protection Circuit

The cutoff circuit of ARYA is designed keeping in mind the safety of batteries and electronic circuitries. The cutoff circuit consisted of diodes, MOSFETs, comparators, and relays. The diodes are used to prevent back current going to the batteries. Two comparators are used which keeps the battery voltage between a minimum and maximum threshold. Considering 3S Li-Po batteries, one of the comparator is set to a minimum threshold of 3.7V and the other to a maximum threshold of 12V. A combination of MOSFET and relay acted as the switch and made sure that the battery disconnects whenever it malfunctions. The MOSFET and relay are selected, keeping in mind the current required by electronic circuitries (around 30A) and Power distribution board.

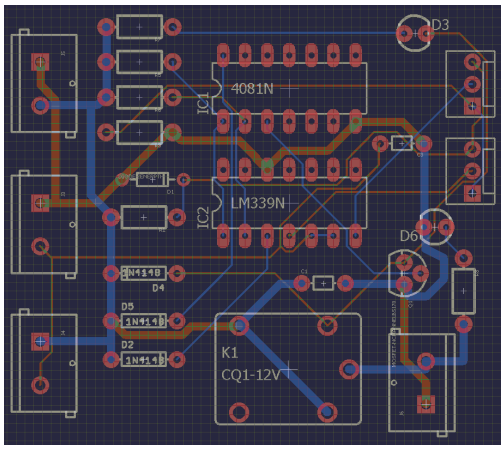


Figure 7: Layout of Battery Protection Board

After the battery is disconnected, switching circuit is used to switch to spare battery. The switching circuit consisted of simple electronic relays which acted as switch.

D. Kill Switch

An external switch is an essential component of every system which obstructs or kills the power supply. It prevents the failure of the whole in case of emergencies. In ARYA, a manually operated T7 series toggle switch from OTTO Is employed between the battery and the processing components. It can withstand 16A current which is adequately high for them. When it is toggled, it

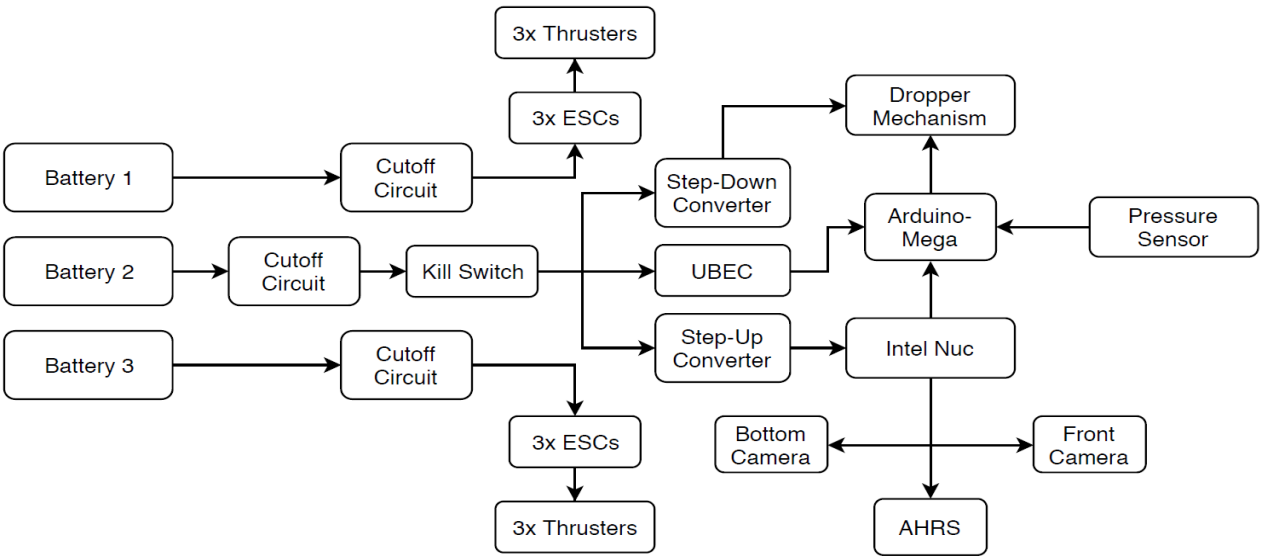


Figure 8: Electronic Architecture of ARYA

kills the power supply of Arduino, Intel NUC, and the servo actuated dropper mechanism such that no signals can be commanded to the thrusters and eventually they will stop and the ARYA will rise because of its positive buoyancy.

E. Control Units

- **Onboard Single Central Processing Unit:** A light & sturdy single-board computer (Intel NUC corei7) development platform is used as the computing center of the AUV.
- **Microcontroller:** The vehicle uses Arduino Mega 2560 microcontroller based on Atmega 2560. The Arduino receives signals from the SBC via UART mechanism and controls servo such as thrusters.
- **Thruster and motor control:** Brushless thrusters T100 of BLUE ROBOTICS used for providing movement is controlled using Brushed ESC's (electronic speed control) at different speed levels. The thrusters speed is set by the microcontroller using PWM. Power is supplied to the thrusters directly from batteries. The ESC's used are reliable, fast, and compact with proper error indicators.

F. Sensors

- **Attitude Heading and Reference System:** The VN-100 is a high-performance10 Axis Inertial Measurement Unit (IMU) and Attitude Heading Reference System (AHRS). It provides highly accurate and stable data. It has onboard hard & soft iron calibration capabilities, which compensates the distortion created from the magnetic field of electric motors and batteries.
- **Pressure Sensor:** *Bar30* pressure sensor from Blue Robotics has been incorporated for a precise pressure and temperature measurements, which together give us a resolution of 0.2 mbar.
- **Cameras:** Two Logitech Quick Cam 5000 USB cameras are used for vision sensing. These have meager power requirements and are user-configurable. One covers the front view and other covers bottom view. The cameras are internally mounted, thereby not interfering with the streamline motion of the body. The vehicle can switch its monitoring between the two according to the tasks.



Figure 9: AHRS [11]

IV. SOFTWARE SYSTEM

A. Stack Structure

The stack is divided into the three separate packages which handle different tasks according to their functionalities. It is developed using the Kinetic version of the Robot Operating System, which is chosen because of the excellent support available. ROS enabled the design to be modular through discrete and parallel-running nodes with communication implemented through custom messages.

- **Software Module -** It acquires live input feed from the camera for guiding the vehicle as per the task received from the task planner. After multiple preprocessing steps applied to improve the quality of the image, image-processing algorithms are implemented to extract features and acquire the guide data.
- **Navigation Module-** It receives the guide data published by the software module and accordingly calculates the appropriate actuator command values.
- **Control Module -** It has the task of maintaining stable yaw and pitch control of the vehicle through a PID algorithm and feeds the PWM values to the microcontroller.

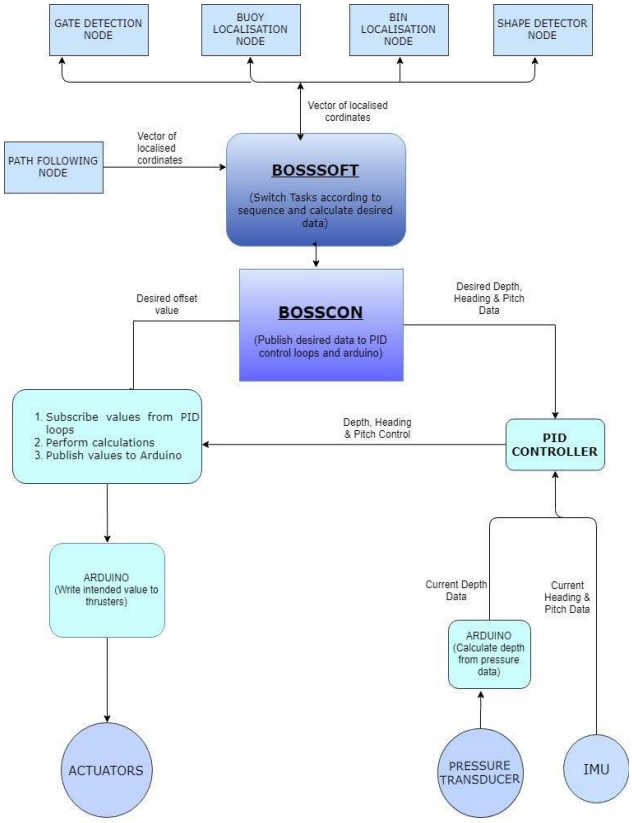


Figure 10: Software Framework of ARYA

B. Image Processing

Due to light absorption, reflection, unsuitable lighting conditions and scattering by water, challenges are faced in developing vision software on images captured in an underwater environment such as distortion in color, contrast is contrast limited adaptive histogram equalization known as CLAHE [12] which is an alternative to adaptive histogram equalization [13] implemented by limiting the enhancement of contrast and in doing so, limits the unwanted amplification of noise.

Color-based segmentation of the task objects is executed by using Otsu's method of adaptive thresholding [14] on the hue-channel of the image in HSV (hue, saturation, value) color space [15-18]. Additional extraction of our desired contour is done by checking the morphological features of all the contours.

Localization and guidance to the vehicle towards the desired point are achieved through the use of a visual serving algorithm known as VisP [19] which produces accurate relative three-dimensional pose estimation of our vehicle using the acquired two-dimensional feature data feedback.

C. PID Control Algorithm

For achieving precise yaw and pitch control, we implement a feedback-based control loop algorithm called proportional-integral-derivative (PID) controller [20-21] on the depth values, pitch angles and yaw angles of the AUV operated on as process variables. The pitch of the vehicle is maintained at a constant stable angle for minimum deviation in vehicle movement while the yaw angle and depth is varied by the navigation module to steering the vehicle towards the desired coordinates. The algorithm consists of three parameters to be tuned- Proportional controller (K_p), Derivative controller (K_d) and Integral controller (K_i) which respectively work on the proportional, derivative and integral value of the error between the desired set point and current state point of the process variable and accordingly vary the actuator commands. The parameters are tuned though Ziegler-Nicholls's tuning method by observing the oscillations of the vehicle in an underwater environment.

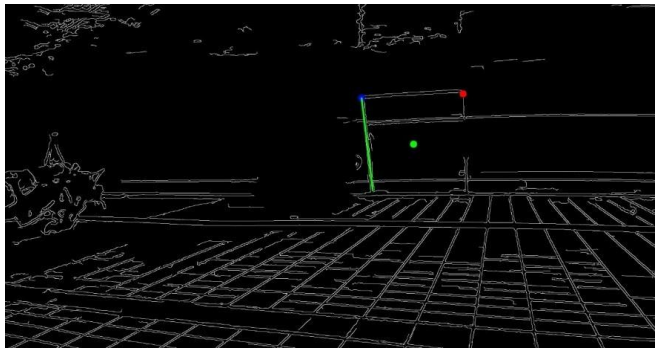


Figure 11: Gate Detection

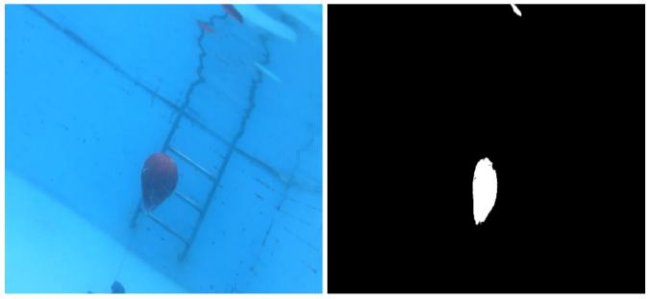


Figure 12: Segmentation of buoy using adaptive thresholding

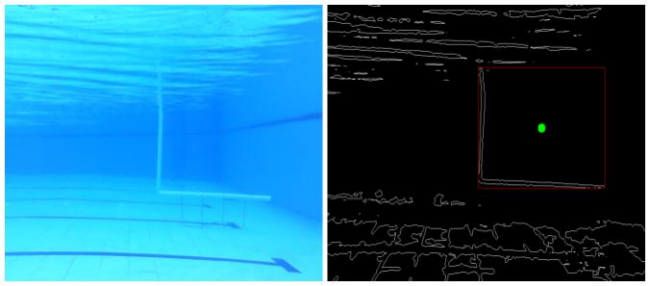


Figure 13: Localization of gate centre using contour validation

V. CONCLUSION

The design of ARYA is a result of rigorous iterations with the commitment to develop a sturdy yet lightweight-cum-compact vehicle. The vehicle meets the industrial standards in terms of safety and eco-friendliness to operate in shallow water, with a depth range of 5-25m. The assembled framework of the vehicle is utterly stable underwater to hover or maneuver to perform varied functions. A compact electronic stack ensures optimal utilization of space to place all the electronic components in it while avoiding clustering of wire harnesses. The Li-Po batteries efficiently power the stack through the watertight connectors mounted on the endcaps of the pressure hulls. The effective formulations of the software

and vision algorithms precisely detect contoured objects and maneuver the vehicle with ease. The PID controllers implemented prevents undesirable deviations in the motion of the AUV. The development of ARYA provides a favorable platform for the team to advance further and improvise the system by integrating distinctive advanced sensors for broadening the application horizons of the AUV in monitoring and detecting the marine habitat and structures.

ACKNOWLEDGMENT

This research and development project is financially supported and supervised by Delhi Technological University. We also thank Pacific Sports Complex, Delhi for letting us availing the facility of swimming pool for testing the vehicle.

REFERENCES

1. Jiang, Chun-meng, Lei Wan, and Yu-shan Sun. "Design of motion control system of pipeline detection AUV." *Journal of Central South University* 24, no. 3 (2017): 637-646.
2. Neto, Edson Cavalcanti, Rejane C. Sá, Guilherme C. Holanda, Francisco AX da Mota, Antonio T. Varela, André LC Araújo, Ítalo J. Loiola, Rogério Oliveira, R. Auzuir, and Victor Hugo C. de Albuquerque. "Autonomous underwater vehicle to inspect hydroelectric dams." *International Journal of Computer Applications* 101, no. 11 (2014).
3. Cruz, Nuno A., Aníbal C. Matos, Rui M. Almeida, Bruno M. Ferreira, and Nuno Abreu. "TriMARES-a hybrid AUV/ROV for dam inspection." In *OCEANS'11 MTS/IEEE KONA*, pp. 1-7. IEEE, 2011.
4. Neto, Edson Cavalcanti, Rejane C. Sá, Guilherme C. Holanda, Francisco AX da Mota, Antonio T. Varela, André LC Araújo, Ítalo J. Loiola, Rogério Oliveira, R. Auzuir, and Victor Hugo C. de Albuquerque. "Autonomous underwater vehicle to inspect hydroelectric dams." *International Journal of Computer Applications* 101, no. 11 (2014).
5. Negahdaripour, Shahriar, and Pezhman Firoozfam. "Multi-camera inspection of underwater structures." U.S. Patent 7,496,226, issued February 24, 2009.
6. Gupta, Aayush, Vatsal Rustagi, Raj Kumar Saini, Akshay Jain, Prateek Murgai, Aditya Rastogi, Pronnoy Goswami et al. "Delhi Technological University: Design and Development of the Littoral AUV Zyra 2.0."
7. Inoxmare. "Hexagon Socket Screws – What Benefits?" inoxmare.com https://blog.inoxmare.com/eng/2016/03/03/hexagon-socket-screws-what-benefits/ (accessed June 1, 2019).
8. Moore, Steven W., Harry Bohm, Vickie Jensen, and Nola Johnston. *Underwater robotics: science, design & fabrication*. Monterey, CA: Marine Advanced Technology Education (MATE) Center, 2010.
9. Diesel, M. A. N. "Turbo, Basic principles of ship propulsion. December 2011. Copenhagen, Denmark."
10. Morris, Melissa, and Sabri Tosunoglu. "Comparison of rechargeable battery technologies." In *ASME Early Career Technical Conference, Atlanta, Georgia*, pp. 2-12. 2012.
11. Vectornav. "VN-100 IMU/AHRS." vectornav.com. https://www.vectornav.com/products/vn-100 (accessed Jan. 1, 2018).
12. Yadav, Garima, Saurabh Maheshwari, and Anjali Agarwal. "Contrast limited adaptive histogram equalization based enhancement for real time video system." In *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 2392-2397. IEEE, 2014.
13. Pizer, Stephen M., E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. "Adaptive histogram equalization and its variations." *Computer vision, graphics, and image processing* 39, no. 3 (1987): 355-368.
14. Priya, M. S., and Dr GM Kadhar Nawaz. "Multilevel Image Thresholding using OTSU's Algorithm in Image Segmentation." *International Journal of Scientific & Engineering Research* 8, no. 5 (2017).
15. Liu, Zhong, Weihai Chen, Yuhua Zou, and Cun Hu. "Regions of interest extraction based on HSV color space." In IEEE 10th International Conference on Industrial Informatics, pp. 481-485. IEEE, 2012.
16. Wikipedia. "Color space" https://www.wikipedia.org. https://en.wikipedia.org/wiki/Color_space100 (accessed Jan. 10, 2018).
17. Wikipedia. "HSL and HSV" https://www.wikipedia.org. https://en.wikipedia.org/wiki/HSL_and_HSV (accessed Jan. 10, 2018).
18. Sural, Shamik & Vadivel, A & Majumdar, Arun. (2005). Histogram Generation from the HSV Color Space.
19. Marchand, Eric. "Visp: A software environment for eye-in-hand visual servoing." In Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C), vol. 4, pp. 3224-3229. IEEE, 1999.
20. Wikipedia. "PID controller" https://www.wikipedia.org. https://en.wikipedia.org/wiki/PID_controller (accessed Jan. 8, 2018).
21. ROS.org. "pid" http://wiki.ros.org. http://wiki.ros.org/pid (accessed Jan. 10, 2018).

AUTHORS PROFILE



Mr. Aditya Natu is a Delhi Technological University graduate having pursued Production and Industrial Engineering, and embarking his career in Planys Technologies as a Mechanical Design Engineer. He has been actively involved in various technical projects namely on Autonomous Underwater Vehicle and Racing Go-Kart and has lead the respective teams to various national and international competitions. He has successfully published research papers based on the work pursued by him on the Structural Design of Go-Karts and recently on Regenerative Braking in Commercial Vehicles. He is currently pursuing research on Design and Controls of Underwater Vehicles, and have successfully completed a research internship at CSIR National Institute of Oceanography, India.



Mr. Anuj Badhwar is a Delhi Technological University graduate, and currently working in Amazon as a Software Development Engineer. He is an avid computer vision and deep learning researcher for the past 2 years, and has published a paper in NTIRE, CVPRW 2019 in the field of Deep Learning.



Mr. Vipul Garg is a Mechanical Engineering student at Delhi Technological University (formerly DCE). He is currently working as team leader and embedded department head in Team DTU-AUV. For the last two years, He has been pursuing active research in the field of mechatronics.



Ms Upasana Biswas is an Electrical Engineering student at Delhi Technological University (formerly DCE). She has been pursuing active research in the field of computer vision and has worked in the software department of Team DTU-AUV.



Mr. Deepanshu Bansal is currently pursuing Electrical Engineering from Delhi Technological University. He has worked in the Embedded and control Department of Autonomous Underwater Vehicles for two years. And is currently working on navigation of Autonomous Grounded Vehicles using state machine model.



Mr. Aakash Kumar is a Mechanical Engineering student at Delhi Technological University (formerly DCE). He has been in a research field from last two years and published technical and review papers in the field of Autonomous Underwater Vehicles and in the field of Hydrogen fuel cell technology (published in SAE WCX 2018).

View publication stats

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Multi-Agent based Stable Clustering and Collision Detection In VANET

Nikhil Pratap Singh Rathore
Department of Information Technology
Delhi Technological University
New Delhi, India
nikstobu@gmail.com

Abhinav Yadav
Department of Information Technology
Delhi Technological University
New Delhi, India
abhinav.yadav.1998@gmail.com

Anamika Chauhan
Associate Professor
Department of Information Technology
Delhi Technological University
New Delhi, India
anamika@dce.ac.in

***Abstract*— One of the most important components of the mobile network is the Vehicular ad-hoc network (VANET). VANET facilitates the communication between vehicles and agents present in the roadside units. We introduce clustering as an important feature in dealing with the frequently shifting VANET topology through local coordination. The strategy uses mobile and static both agents to create communication between the vehicles and the Roadside units using a multi-agent interaction technique. The cluster size is chosen by the RSU manager and the cluster head is decided based on the adjacent node's connectivity and the vehicular speed. Once the cluster is created, we vary the cluster size on the speed factor and therefore enhance the cluster lifetime and reduce the cluster overhead. We also introduce the concept of finding traffic congestion and the shortest path for data transmission between nodes. The entire technique is simulated in NS2 using C++ and TCL languages by considering parameters like throughput, security, power consumption, cluster lifetime, communication range, etc.**

Keywords— ***VANET; RSU; TA; CH; DCH; etc.***

Introduction

VANET or vehicular ad-hoc network vehicles is a subsidiary class of Mobile Ad-hoc networking. It is the technology that enables communication taking place between any set of vehicles and RSU's. VANET is supposed to be a revolution in the field of intelligence incorporation to the transportation systems which has broadened the definition of a vehicle much more than just a moving object [1]. The main highlight of its introduction is the possibility of increased safety of the driver by sensory detection of unforeseen dangers. Other applications can be categorized into entertainment, networking, weather forecasting, and management of incoming traffic [2]. The inclusion of Roadside Unit(s) and the internet comes as a firm support system. The architecture of VANET is designed to encourage vehicle to vehicle propagation alongside dissemination among RSUs and IEEE 802.11p standardized hybrid communication. The network is complicated because of the need for frequent changes in network structures (topology) as it is dependent on the speed of the object of interest. The collection, exchange, and averaging of data are subject to the formation of Clusters. The proposed study uses this essential clustering method for the formation of excellent communication in VANET. The connections and signal dissemination between a vehicle and RSU present in VANET are established with the help of external software agents. The work is focused on the enhancement of the head lifetime of a cluster and its stability using the speed of vehicle and cluster's radius as the basis. This paper elaborates on multiple factors such as relative speed and selection of CH which are further dependent. Along with an overview of VANET, issues related to security including spoofs and data modification is detailed. The application of Wireless Sensor Network or WSN can be found in various areas of medicine, defense, data transmission, etc. particularly because of its low maintenance and installation cost. The objective of the proposed method is to facilitate communication among nodes after cluster formation with minimized data loss and higher throughput. Also, the formed clusters can deal with the fault and traffic collision.

I. LITERATURE SURVEY

In the previous works done on VANET, it is possible to create stable clusters and also enable node to node which is a vehicle to vehicle communication [1]. The clusters are enhanced in the work done so far; their lifetime has increased even in the ever-changing network topology. When a vehicle leaves a cluster, it gets disconnected from the cluster network and it gets really difficult to manage such complications. The dynamic clustering technique is a solution to this which has been devised in the paper. Clusters have been created based on the relative speeds of the vehicles. They devised methods for applying to simple and a lane and also fail to provide safety and data packet losses when nodes communicate.

There has been sufficient work on effective routing algorithms for a node to node communication in VANET. The vehicle Agency algorithm in the paper [2] explains how the clusters are created and RSU's decide the CH and this CH with the RSU's are responsible for vehicular communication. But this algorithm fails when there is a traffic collision and does not tell much about traffic congestion. Congestion in the network,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:10 UTC from IEEE Xplore. Restrictions apply.

node depletion, and limited processing due to limited battery resources leads to a sharp increase in packet loss along with degradation of channel capacity which in turn end up affecting the energy efficiency of sensor nodes. While there are several congestion control mechanisms present [3] [4], the proposed method tackles the detection and notification of congestion while keeping rate adjustment mechanisms in mind. The solution considers the utility of all nodes and their periodically incoming data packets apart from inter-arrival to service time ratio as the matrix of congestion. In the exponential research arena of Wireless Sensor Networks, Fault Tolerance is considered to be a vital parameter as it does not work in an open-loop fashion. [5] and majorly drawn on WFCC [4], the results of stimulation are achieved. The paper [6] explains the need for facilitating reliable communication in VANET. The existing routing protocols face many challenges as explained in the paper [7].

In our work, we devise an algorithm that considers the mobility of the BS and the nodes. This includes mobile as well as static software agents. The algorithm used creates stable clusters ready for communication-based on factors like speed and weight factor. It finds the shortest path for message delivery and also the traffic congestion in the communication range. Whenever a traffic collision is detected the proposed method looks for an alternative path for data transmission and reroute path for vehicles in the lane [8] [9]. We increase the throughput, security and enable fault tolerance in vehicular communication

II. SYSTEM ANALYSIS

The used routing protocols in the existing methods do not consider the mobility of BS and the sensor nodes and thus cannot be directly implemented on a mobile WSN. None of the reported algorithms can explain all the goals simultaneously. The goals are:

1. Maintaining the node mobility and connectivity via alternate paths.
2. Overcoming the less-reliable wireless connections and diminishing the message overhead.
3. With the BS and the node mobility ensuring the reliability in an energy-efficient manner.

While making the routing decisions we consider the mobility of the BS node and the sensor nodes. The concept of Deputy collector head (DCH) also known as cluster management nodes and the cluster head (CH) is used to increase the lifetime of the cluster.

A. Mathematical Model

The mathematical model describes the entire cluster's lifetime. Clusters consist of vehicles moving in the same direction with varying speeds. The trusted agent (TA) receives the authentication request from the on-board units and keeps information such as car ID, speed of the car, etc.

Cluster radius $\propto$ Vehicle density*mean speed of on-board units

Therefore, the radius and the size of the cluster keep on changing dynamically.

Weight factor $\propto$ neighbor count $\div$ near to the above-average speed

The weight factor is a value assigned to the data points to distinguish them based on importance. In CH selection the Weighting factors are the vehicular speed and the Neighbor count. The weight factor is forwarded to the TA which is the RSU in this case which selects the CH.

After cluster formation cluster maintenance comes into play. A new vehicle enters the communication range (Rc), the vehicle is added to the cluster member (CM) list, and if leaves it is deleted from the CM list. If there are two CH in an Rc cluster management agent (CMA) merges both clusters and if CH fails to work, CMA selects new CH.

B. Algorithm

1. Define a transport agent for sending (TCP)
2. Attach the transport agent to sender node using
3. Define another transport agent for receiving (TCP)
4. Attach the transport agent to the receiver node
5. Connect the sending and receiving transport agents
6. Define Application instance
7. Attach transport agent to application agent
8. Set the flow color
9. Define a transport agent for sending using (TCP)
10. Attaching transport agent to the sender node
11. Defining a transport agent for receiving (TCP)
12. Attaching transport agent to the receiver node
13. Connecting sending and receiving transport agents
14. Defining Application instance
15. Attaching transport agent to application agent
16. Setting flow color
17. Setup a UDP connection and attach agent nodes
18. Setup a CBR over UDP connection with packet size=1000 and rate=1Mb
19. Schedule events for the CBR and FTP agents (start CBR)
20. Process the encryption and decryption of a file using RSA
21. initialize the parameters for the RSA algorithm
    e=3; d=11787, n=17947
22. accept the user's choice. E to encrypt or D to decrypt.
23. if the choice is E
    accept the name of the file to encrypt from the user and the absolute path of the file.
    open the file and read the contents (read mode)
    open another (write mode)
    read the input files
    split the file contents into constituent characters
    for (i=1 to e) process character-wise and encrypt

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:10 UTC from IEEE Xplore. Restrictions apply.

for each character convert each character to ASCII

24. if the choice is D

    run program to decrypt a file using RSA
    initialize the parameters for the RSA algorithm
    e=3; d=11787; n=17947
    accept file name to decrypt and the absolute file to be decrypted
    open the file and read the contents (read mode)
    open another file and append new file (write mode)
    read the input files
    split file contents into constituent characters and process character-wise.

25. Finish

III. SIMULATION AND RESULTS

The entire simulation process is done in NS2. The vehicles are moving in multi-direction and the number of nodes is given by N within a communication range 'Rc'. The Rc is given for an RSU, and the vehicle speed V varies between 40kmph to 120 kmph. The simulation can detect traffic congestion and collision. After an accident occurs, the algorithm can reroute the vehicles.

A. Simulation Inputs

A few input values are listed in the table below:

Quantities	**Proposed method Values**	**Existing method Values**
Range of communication (Rc)	Up to 1200 meters	Up to 1000 meters
No. of clusters (Cn)	7	2
Size of network	1000 msq	1000msq
Vehicle speed(V)	40-120kmph	40-120kmph
Number of nodes(N)	27	9
Network Bandwidth	100mbps	100mbps
Simulation time	300 sec	300 sec
Queue Limit (Ql)	50	30
Link-type	Simplex	simplex
Tail drop	10ms	unspecified

B. Execution Parameters

There are the following performance parameters in the simulation process:

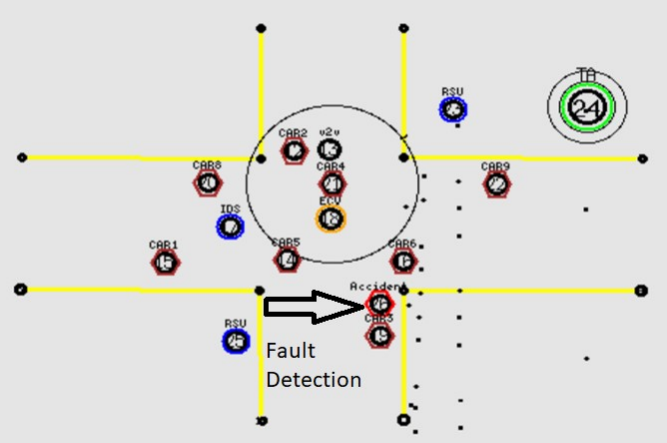


Fig1: Traffic collision Detection

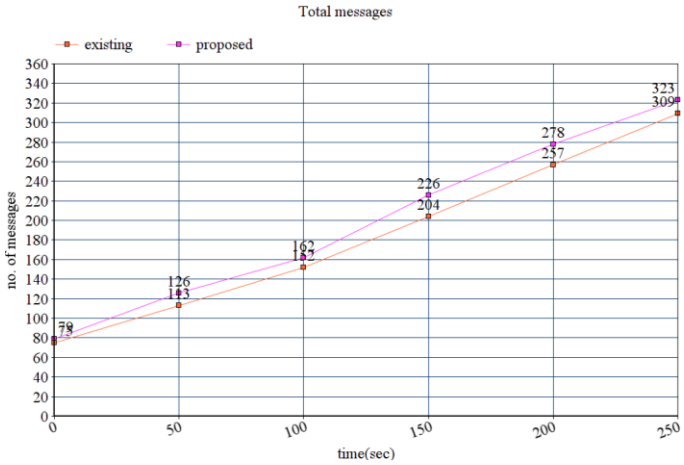


Fig2: total messages sent

1. Cluster formation time: cluster formation time taken by the RSU. With time the number of nodes increases and therefore the cluster formation time increases. The vehicle speed also increases the cluster formation time.
   Cluster formation time $\propto$ V
   Cluster formation time $\propto$ N
2. Cluster lifetime: The time till there is no change in the cluster head of the cluster. Mobility of vehicle plays a major role in deciding the cluster lifetime, decreases with increasing speed.
   Cluster lifetime $\propto 1/V$
3. Cluster head selection time: CH selection time taken by the road side unit. The CH selection time increases with the increasing N and/or V.
   CH selection time $\propto$ V
   CH selection time $\propto$ N
4. Packet data loss: The number of packets lost while

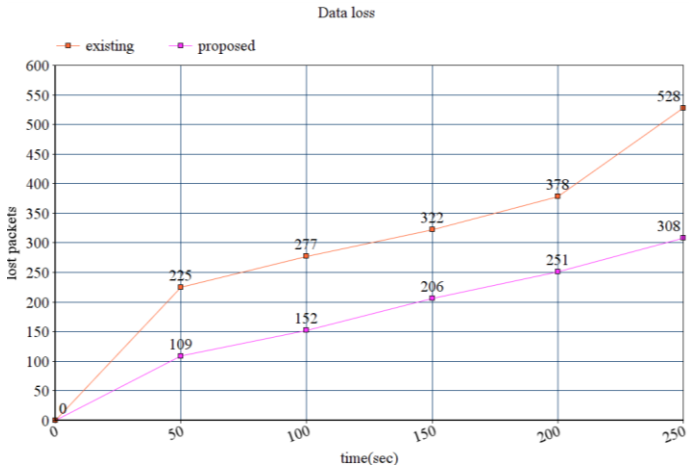


Fig3: Data loss over a period of time

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:10 UTC from IEEE Xplore. Restrictions apply.

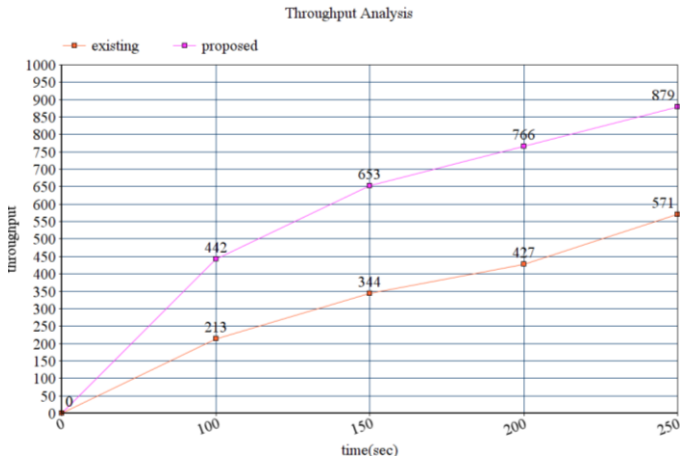


Fig4: Throughput analysis

transferring packets from source node to destination nodes. If the number of vehicles increase in the cluster the packet data loss increases but the loss is small in the proposed method as compared to the existing method.

5. Throughput: The rate of processing of the data packets is higher in the proposed method than the existing one at any point in time.

IV. CONCLUSION AND FUTURE WORK

We presented a traffic collision detection technique in addition to the multi-agent-based stable clustering in VANET. The technique uses the on-board and Roadside units. The cluster size is varied based on the mean speed of all the vehicles in the communication range. The cluster head is appropriately chosen based on the weight factor. The node with the highest weight factor is the CH. The transfer of control messages during the formation of the cluster between vehicles and RSU's is secured and encrypted which in turn increases the network routing overhead. The clusters formed are long-living clusters. The algorithm can deal with traffic collisions and also enables the rerouting of vehicles. The entire simulation is done using the NS2 network simulator and the proposed method works perfectly with throughput, data loss, cluster lifetime, cluster creation time. Power consumption is high in the proposed method. As the power factor is not that important factor presently but it can be minimized in future works.

V. REFERENCES

[1] Li, G. Li, J. Li and B. Yu, "Lower bound of weighted fairness guaranteed congestion control protocol for wsns," *In 2012 Proceedings IEEE INFOCOM,* pp. 3048-3050. IEEE, 2012.

[2] Li, David Chunhu, Li-Der Chou, Li-Ming Tseng and Yi, "A bipolar traffic density awareness routing protocol for vehicular ad hoc networks," *Mobile Information Systems 2015,* 2015.

[3] Y. Jennifer, B. Mukherjee and D. Ghosal, "Wireless sensor netwrok survey," *Computer Networks 52,* vol. 12, pp. 2292-2330, 2008.

[4] Hubballi, Pampapati, Ravik and A. V. Sutagundar, "Agent based dynamic clustering for hybrid VANET (ADCHV)," *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT),* no. IEEE, pp. 382-386, 2016.

[5] Borasia, Shivangi and V. Rajsinghania, "A review of congestion control mechanism for wireless sensor networks," *Technology Systems and management,* no. Springer, BErlin Heidelberg, pp. 201-206, 2011.

[6] A. Bavalatti and A. V. Sutagundar, "Multi-Agent based stable clustering in VANET," *International Conference On Smart Technologies for Smart Nation,* pp. 1033-1038. IEEE, 2017.

[7] Awang, Azlan, . K. Husain, Nidal Kamel and Son, "Routing in vehicular ad-hoc networks: A survey on single-and cross-layer design techniques, and perspectives," *IEEE Access 5,* pp. 9497-9517, 2017.

[8] Q. K. Naseer, A. H. Abdullah and J. Lloret, "Road Perception based geographical routing protocol for vehicular ad hoc networks," *International Journal of Distributed Sensor Networks 12,* vol. 2, p. 2617480, 2016.

[9] B. J. and D. Manivannan, "Unicast routing protocols for vehicular ad hoc networks : A critical comparison and classification," *Pervasive and Mobile computing 5,* vol. 1, pp. 1-18, 2009.

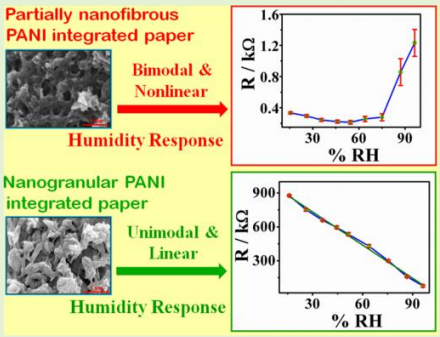
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:34:10 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

Nanoengineered Conductive Polyaniline Enabled Sensor for Sensitive Humidity Detection

Samar S. Sandhu[¥1], Saurabh Kumar[¥1, 3], Shine Augustine[1], Udiptya Saha[1], Kamal Arora[2], Sayan Bayan[4], Samit K. Ray[4, 5], Nitin K. Puri[2] and Bansi D. Malhotra[*1]

***Abstract*— We report results of the studies relating to the fabrication of a paper-based humidity sensing element comprising a nanostructured polyaniline (PANI) coating on a filter paper substrate. The electrical conductivity of the PANI integrated conductive paper increases from 1.9 × 10^{-6} Scm^{-1} to 1.1 × 10^{-1} Scm^{-1} on modification of the polymerization protocol. This increase in conductivity is ascribed to the change in morphology of the PANI coating from nanogranular (N_g) to partially nanofibrous (pN_f). We investigate the effect of observed morphologies of PANI-paper composites on their resistive-type humidity sensing performances. The partially nanofibrous (pN_f) PANI-paper exhibits a bimodal humidity response due to polymer-swelling effect at higher humidity (≥ 55% RH). However, the nanogranular (N_g) PANI-paper yields a unimodal, linear humidity response within the humidity range, 16-96.2% RH, with a sensitivity of 9.79 kΩ/% RH (R^2 ≈ 0.996). Hence, the N_g PANI-paper is a promising alternative to the conventional humidity sensors fabricated on ITO glass/bare glass/PET substrates.**



***Index Terms*— Conductive paper, humidity sensor, *in situ* polymerization, nanostructured PANI.**

I. INTRODUCTION

Paper has been used for many decades as a supporting platform for the fabrication of analytical devices. Its flexible, foldable and lightweight nature makes it an ideal substrate for designing portable devices with electronic, optical, optoelectronic and electrochemical applications [1-3]. Besides this, paper has a superhydrophilic, mesh-like, hierarchical structure comprising randomly interconnected cellulosic fibers [1]. The intrinsically high porosity and wicking capability facilitates uniform adsorption of hydrophilic precursors, thereby making it a suitable substrate for *in situ* growth of nanomaterials like metal oxides, conductive polymers etc. [4]

Conductive polymers have been vastly employed for flexible electronics, biosensing and environmental monitoring applications [5-6]. Their light weight, low production cost and good electrical conductivity make them an ideal choice to form conductive networks over flexible substrates such as paper. Additionally, the electronic, optical, physicochemical and electrochemical properties of conductive polymers can be tuned by varying their oxidation levels and doping levels [7-9]. Compared to conductive polymers like polypyrrole, polycarbazoles, polythiophenes and phthalocyanines, PANI has attracted special interest owing to its unique redox conductance tunability along with a relatively higher attainable electrical conductivity and primary doping level [10-18].

[1]Humidity sensing is an integral part of environmental monitoring with a crucial role in the maintenance of drug storage areas, hospitals, food storage and packaging, electronic device fabrication and air conditioning systems etc. [12] The commercially available humidity sensors including those based on tin oxide, infrared, catalytic and dew point sensing mechanisms have high fabrication and maintenance costs, significant power consumption and require high temperature for synthesis and/or operation. Hence, considerable efforts have been directed towards the development of humidity sensitive organic materials

Manuscript resubmitted for peer review on 18th April 2020.

[1]Nanobioelectronics Laboratory, Department of Biotechnology, Delhi Technological University, Delhi 110042, India
[2]Nanomaterials Research Laboratory, Department of Applied Physics, Delhi Technological University, Delhi 110042, India
[3]Centre for Nano Science and Engineering (CeNSE), Indian Institute of Science, Bangalore 560012, India
[4]Department of Physics, Indian Institute of Technology, Kharagpur 721301, India
[5]S. N. Bose National Centre for Basic Sciences, Kolkata 700106, India
Samar S. Sandhu- samarsandhu31@gmail.com
Saurabh Kumar- sau2203@gmail.com
Shine Augustine- shine2089@gmail.com
Udiptya Saha- udiptya999@gmail.com
Kamal Arora- kamal.arora8@hotmail.com
Sayan Bayan- sayan.bayan@gmail.com
Samit K. Ray- samit@bose.res.in
Nitin K. Puri- nitinkumarpuri@dtu.ac.in
*Corresponding Author: Bansi D. Malhotra-bansi.malhotra@gmail.com
Scopus ORCID ID: 0000-0002-0957-2684

XXXX-XXXX © XXXX IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

(phthalocyanine along with doped conductive polymers like polypyrrole, polythiophene and especially PANI), organic composites, organic blends, and organic-inorganic composite materials [11, 13-18]. The main advantage of incorporating organic materials in humidity sensors is that they facilitate room temperature measurements at low voltages, whereas the conventional inorganic materials-based sensors normally operate at higher voltages and/or elevated temperatures. This makes the conventional humidity sensors relatively uneconomical and power inefficient [12]. Gomez *et al.* first reported a thermally inkjet-printed PANI modified conductive paper for impedimetric humidity sensing at room temperature [19]. Further, Morais *et al.* fabricated an inkjet-printed, PANI/PEDOT:PSS modified conductive paper for electrochemical impedimetric humidity sensing at room temperature [20]. However, both these humidity sensors exhibited poor humidity sensitivity like other PANI-based humidity sensors including freestanding PANI-chitin blends [21], PANI microaggregates [13] and PANI-MWCNT composite films [22] fabricated on flexible PET substrates. Recently, Manjunatha *et al.* reported the use of PANI-TaS_2 composite pellets as room-temperature resistive-type humidity sensors. However, the composite exhibited brittleness, undesirably high resistance (~few hundred MΩ) and non-linear response across the dynamic sensing range [23]. To overcome the limitations of PANI-based humidity sensors, we have directed the morphology and appropriately selected the primary dopant of the PANI coating to optimize the PANI-paper composite-based resistive-type humidity sensor [17]. Notably, the morphology of a humidity sensing element has been found to significantly affect its performance [13, 16]. PANI microaggregates offer large electrical resistance (~MΩ) and poor humidity sensitivity, which make them undesirable for resistive-type humidity sensing [13, 24]. Further, 1D nanostructured networks of PANI experience a significant polymer-swelling effect, which causes an abrupt increase in their electrical resistance and thus, cannot be used for high humidity measurement (*i.e.* ≥ 50-55% RH) [13, 25-26]. However, 3D PANI nanostructures may perhaps bear greater resistance to polymer swelling, owing to their relatively aggregated and disordered morphology. Hence, a bulk nanostructured PANI coating may provide a balance between humidity sensitivity, electrical conductivity and morphological stability to the humidity sensing element. However, a significant disadvantage of conducting polymer-based humidity sensors is that they exhibit non-specific response to various electron donating and electron withdrawing analytes [11]. This could perhaps be restricted by fabricating p-n heterojunction nanohybrids comprising p-type PANI and possibly n-type rGO [27-28].

By varying the *in situ* polymerization parameters, we synthesized two distinct morphologies of PANI coatings on filter paper namely, partially nanofibrous (pN_f) and nanogranular (N_g) PANI. We then investigated the effect of morphology of the two PANI-paper composites on their resistive-type humidity sensing performances. We found that filter paper played an important role towards the enhanced humidity sensitivity and durability of the PANI-paper composite due to increased water vapor adsorption capacity, enhanced mechanical flexibility and environmental stability [29-31]. The novelty of this work lies towards the design of a one-step fabrication process to synthesize a swelling-resistant N_g morphology of PANI on a paper substrate that could be employed as a chemiresistive humidity sensing element.

II. MATERIALS AND METHODS

A. Materials

Whatman-1 filter paper (thickness: 180 µm, pore size: 11 µm) was used as the substrate for *in situ* growth of PANI. For synthesizing PANI: aniline ($C_6H_5NH_2$) (Central Drug House, India) was purified by distillation prior to its use, ammonium persulfate (APS) and formic acid (HCOOH) were purchased from Fisher Scientific, India and were used as received. Deionized (DI) water (resistivity = 18.2 MΩ-cm) was used throughout the experiments. For creating different % RH conditions, salts such as lithium chloride (LiCl), potassium acetate (CH_3COOK), magnesium chloride ($MgCl_2{\cdot}6H_2O$), potassium carbonate (K_2CO_3), magnesium nitrate ($Mg(NO_3)_2{\cdot}6H_2O$), cobalt chloride ($CoCl_2{\cdot}6H_2O$), sodium chloride (NaCl), potassium chloride (KCl) and sodium sulfate (Na_2SO_4) (Fisher Scientific, India) were purchased and used as received. All the chemicals used were of analytical grade.

B. In situ syntheses of the PANI-paper composites

Fig. 1 represents the various steps employed for *in situ* syntheses of the two PANI-paper composites (using protocols A and B, respectively) followed by the fabrication of a resistive-type humidity sensing setup.

Protocol A presents the preparation of a partially nanofibrous (pN_f) PANI-paper composite using rapid mixing, static polymerization of aniline on paper at 0 ^oC. Whereas, Protocol B describes the preparation of a nanogranular (N_g) PANI-paper composite using dropwise mixing, periodically agitated polymerization of aniline on paper at 5 ^oC. The optical images of the synthesized PANI-paper composites are shown in Fig. 2a. The detailed descriptions of protocols A and B are presented in the Supporting Information.

C. Construction of the humidity sensing platform

A laminated wooden piece of size (6 cm×2.4 cm) was used as a support on which two pieces of adhesive copper tape (one-side conducting) of size (2.5 cm×1.4 cm) each were stuck, leaving an insulating gap of 0.3 cm between them. A strip of either PANI integrated conductive paper (humidity sensing element) of size (2.4 cm×1.2 cm) was then placed over this gap and was covered with another two pieces of copper tape of size (2.4 cm×0.8 cm) each on either side of the gap. This left an area (1.2 cm×0.3 cm) of the conductive paper surface exposed for humidity sensing. Lastly, two 80 cm long Ni-Cr alloy-based wires (each with a resistance = 0.18 Ω) were soldered using a Sn-Ag alloy-based solder wire (resistivity ≈ 12.3 µΩ-cm) to the respective bottom pieces of copper tape. This resulted in negligible electrical resistance between the sensing element and the wire leads.

D. Fabrication of the humidity sensing setup

We assembled a simple resistive-type humidity sensing setup. Ten controlled-humidity enclosures were constructed using sealed containers, each filled with a different saturated

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

salt solution to attain different % RHs. A commercial thermohygrometer was used to monitor the % RH and temperature inside each sealed chamber. A potentiostat (Autolab) was used to monitor the resistance of the humidity sensing element. The detailed descriptions of the humidity sensing setup and protocol are presented in Section III of the Supporting Information.

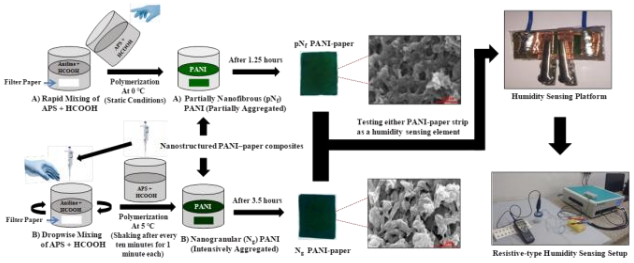


Fig. 1. *In situ* syntheses of the PANI-paper composites followed by the fabrication of a resistive-type humidity sensing setup.

III. RESULTS AND DISCUSSIONS

A. Optical and morphological studies

Fig. 2(a-b) shows the optical images of the synthesized PANI-paper composites. It was observed that PANI coatings were uniformly deposited on/within the paper substrates and the obtained composites exhibited the mechanically flexible nature. The comparative morphologies of these two PANI-paper composites were analyzed using their FESEM images [Fig. 2, (c-d) verses. (e-f)], that were synthesized using the polymerization protocols A and B, respectively. The PANI-paper composite obtained through protocol A exhibited a residual, mesh-like nanofibrous network forming the inner layer of the PANI coating, with outer layers of progressively aggregated nanofibers, termed as pN$_f$ PANI-paper. The obtained morphology could be largely ascribed to the heterogeneous nucleation mechanism governing the *in situ* polymerization of PANI on the cellulosic matrix [29, 32]. At the beginning of each of the *in situ* syntheses, the anilinium cations were physisorbed onto the superfluous acid-activated hydroxyl groups of the cellulosic matrix. Due to hierarchical microstructure of the paper, 3-D adsorption of monomers perhaps occurred across it. Subsequently, these hydrogen-bonded hydroxyl groups formed the heterogeneous nucleation sites on paper. After initiation of the heterogeneous polymerization, the nanofibrous PANI network formed the primary layer of the PANI coating on paper. However, this PANI network underwent partial aggregation due to hydrophilic nature of the cellulosic matrix, which initially promoted the agglomeration of hydrophobic primary oligomeric nucleates into clusters/islands [32]. Next, the high initial concentration of aniline monomers and APS (~ 0.421 M each) led to additional secondary nucleation of the oligomeric nucleates (formed in the bulk solution-medium), over the heterogeneously polymerized PANI network on paper. Subsequently, secondary growth of PANI chains resulted in the rough and multilayered profile of this PANI coating [32-34]. Thus, we obtained a partially nanofibrous PANI morphology for the PANI-paper composite synthesized by rapid mixing, static polymerization (protocol A). Contrastingly, the PANI-paper composite obtained through dropwise mixing, periodically agitated polymerization (protocol B) exhibited a uniform distribution of nearly discrete PANI clusters (of sizes ranging across hundreds of nanometres), with a few nanostructured interconnections between them, termed as N$_g$ PANI-paper. These clusters appeared to be composed of PANI nanogranules, which fused to form these bulk nanostructures [35]. Hence, intensive secondary PANI growth led to the evolution of nearly discrete, 3D nanogranular PANI clusters along with a few nanostructured PANI interconnections spread across the macroporous cellulosic network. The detailed morphological evolutions of the pN$_f$ and N$_g$ PANI coatings on paper have been discussed in the Supporting Information.

The mean bulk electrical conductivities of pN$_f$ PANI-paper strips and N$_g$ PANI-paper strips [no. of strips (n) = 3, at ~ 50% RH, 25 ^{o}C] were found to be 1.04×10^{-1} Scm^{-1} and 1.88×10^{-6} Scm^{-1}, respectively. Further, even after 40 days their electrical conductivities remained nearly the same (0.93×10^{-1} Scm^{-1} and 1.74×10^{-6} Scm^{-1}, respectively) (see the Supporting Information for conductivity calculations). Next, we investigated the crystallographic and compositional differences between Whatman-1 filter paper, pN$_f$ PANI-paper and N$_g$ PANI-paper by obtaining their X-ray diffraction (XRD) patterns and Fourier transform infrared (FT-IR) spectra. The changes in crystallographic characteristics (such as interplanar spacings, crystallite size and interchain spacing) after *in situ* growth of each PANI coating (pN$_f$ and N$_g$) on paper were determined using the XRD patterns and are discussed in Section VI of the Supporting Information. Additionally, the growth of each PANI coating on paper led to the evolution of characteristic reflectance peaks (in the FT-IR spectra) corresponding to the quinonoid (Q) and benzenoid (B) ring stretching within PANI chains (at 1581 cm^{-1} and 1488 cm^{-1} for pN$_f$ PANI-paper verses 1584 cm^{-1} and 1499 cm^{-1} for N$_g$ PANI-paper), respectively. Further, two modified and enhanced reflectance peaks (obtained at 1313 cm^{-1} and 1248 cm^{-1} for pN$_f$ PANI-paper verses 1310 cm^{-1} and 1248 cm^{-1} for N$_g$ PANI-paper) were attributed to the C–N and C–N$^{+\bullet}$ bond stretching within the PANI chains, respectively [36-37]. These two pairs of reflectance peaks were used to calculate the oxidation level and doping level in each PANI coating (pN$_f$ and N$_g$), respectively (Section VII of the Supporting Information for the detailed analyses).

B. X-ray photoelectron spectroscopic (XPS) studies

The chemical compositions of both PANI coatings (pN$_f$ and N$_g$) were investigated using XPS. The presence of C and N originated from the PANI backbone; O was largely due to the presence of formate along with a few hydrogen sulfate and sulfate (from residual H_2SO_4: a decomposition by-product of APS) counter anions on primary-doped PANI; lastly, S solely represented these residual counter anions [Fig. 3(a-b)] [38-40]. Fig. 3c shows the N (1s) core level spectrum of the pN$_f$ PANI coating, deconvoluted into four characteristic peaks: imine nitrogen (C=N moiety, N1 at 398.41 eV), amine nitrogen (C–NH moiety, N2 at 399.03 eV), polaronic nitrogen

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

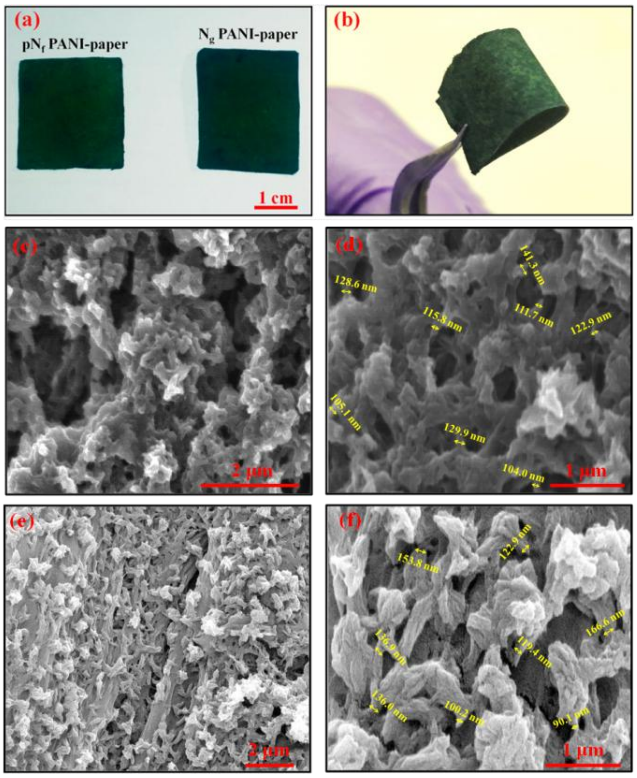


Fig. 2. (a) Optical image of the HCOOH-doped PANI-paper composites, (b) Optical image displaying the mechanically flexible nature of the PANI-paper composites, (c-d) FESEM images of the pN_f PANI-paper composite, (e-f) FESEM images of the N_g PANI-paper composite.

($C–N^{+\bullet}$ moiety, N3 at 400.38 eV) and bipolaronic nitrogen ($C=N^{+}$ moiety, N4 at 401.24 eV), respectively. Similarly, Fig. 3d shows the N (1s) core level spectrum of the N_g PANI coating, deconvoluted into five characteristic peaks: imine nitrogen (C=N moiety, N1 at 398.60 eV), amine nitrogen (C–NH moiety, N2 at 399.29 eV), polaronic nitrogen ($C–N^{+\bullet}$ moiety, N3 at 400.32 eV), bipolaronic nitrogen ($C=N^{+}$ moiety, N4 at 401.37 eV) and an additional peak corresponding to the π-π^{*} shake-up satellite (N5 at 403.27 eV), arising due to a low doping level in the N_g PANI coating [40-41]. The oxidation and doping levels of the PANI coating are crucial factors that can influence the hydrophilicity, crystallographic attributes and electrical conductivity of the PANI-paper composite [42-44]. A variation in these levels would in turn influence the humidity sensing performance of the PANI coating [10, 12-13, 21]. The oxidation levels of both PANI coatings were calculated using the following equation: -

PANI oxidation level = (N1+N3+N4) / (N1+N2+N3+N4) (1)

(Where N1, N2, N3, N4 represent the intensities of the N (1s) deconvoluted peaks as stated above) [40]. For the pN_f PANI coating, the value obtained was **0.526**, which is close to the desired value of 0.5 (*i.e.* 50% oxidation level) of the emeraldine state of PANI [40]. However, the value obtained for the N_g PANI coating was **0.706**, which was much above the required value of 0.5 (*i.e.* 50% oxidation level) [40]. A higher oxidation level (versus the pN_f PANI coating) signified the conversion of a fraction of neutral amine moieties (N2) to neutral imine moieties (N1) within the N_g PANI coating [39-41]. This increase in the oxidation level could be ascribed to the relatively longer polymerization period and higher polymerization temperature of the N_g PANI coating (3.5 hrs. at 5 ^{o}C vs. 1.25 hrs. at 0 ^{o}C), within the polymerization medium comprising APS in superfluous formic acid solution (~ 18.467 M) [32, 36]. Importantly, formic acid has been previously reported to produce a higher oxidation level of PANI in comparison with other inorganic acids like HCl, H_2SO_4 etc.[36] Additionally, the increase in oxidation level from the partially N_f PANI coating to the N_g PANI coating is in agreement with their FT-IR spectroscopic investigations (refer to the Supporting Information).

The doping levels were calculated according to the equation reported by Li *et al.*[45] and Mao *et al.*[43]:-

PANI areal doping level = (Area N3+Area N4) / Area (N1+N2+N3+N4) (2)

The obtained value for the pN_f PANI coating was **0.430** (*i.e.* 43% polaronic/bipolaronic nitrogen sites), which is the highest reported doping level for any PANI-paper composite. This superior primary doping ability of formic acid is crucial towards fabricating a highly sensitive and reliable (*i.e.* high repeatability and low hysteresis during dynamic response) PANI-based humidity sensor [21], by affording a high concentration of the emeraldine salt defect states (polarons and bipolarons) [39, 43]. Further, the obtained areal doping level for the N_g PANI coating was **0.273** (*i.e.* 27.3% polaronic/bipolaronic nitrogen sites). Similarly, we calculated the doping levels based on the intensities of the deconvoluted N (1s) peaks (Figs. 3c and 3d). For the pN_f PANI coating, we obtained a value of **0.323**, suggesting that about one in every three nitrogen sites were polaronic/bipolaronic defect states [39, 41, 45]. Further, the value obtained for the N_g PANI coating was **0.133**, which implied that more than one in every eight nitrogen sites were polaronic/bipolaronic defect states. This drastic reduction in the doping level from 43% (pN_f PANI coating) to 27.3% (N_g PANI coating) could be correlated with the larger extent of secondary nucleation involved in the growth of the N_g PANI coating, which resulted in intensive aggregation of its PANI chains. Importantly, PANI chain aggregation not only de-doped the charge carrier-defect states (N3 and N4), but also reduced the accessibility of the neutral imine-nitrogen sites (N1), thereby hindering the approach of the primary dopant acid moieties towards these imine-nitrogen sites. This resulted in significant increase in the fraction of neutral imine-nitrogen sites (N1), which is evident in Fig. 3d via the highest intensity peak (N1) obtained for the N_g PANI coating [29, 35-36]. The higher oxidation level and lower doping level of the N_g PANI coating (versus the pN_f PANI coating) is in agreement with previously reported results for the H_2SO_4-doped PANI film [40] and HCl-doped PANI films [39]. Further, the trends observed for the oxidation and doping levels of both PANI coatings were found to be in agreement with the values obtained using their FT-IR spectroscopic analyses (please see the Supporting Information).

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

We deconvoluted the characteristic C (1s) peaks corresponding to both PANI coatings to confirm the observed findings (Fig. S3, Supporting Information). Finally, we analyzed the degrees of chemical impurities and structural defects present in each PANI coating, by determining their *C/N* ratios and chain-coupling parameters, respectively (please see the Supporting Information).

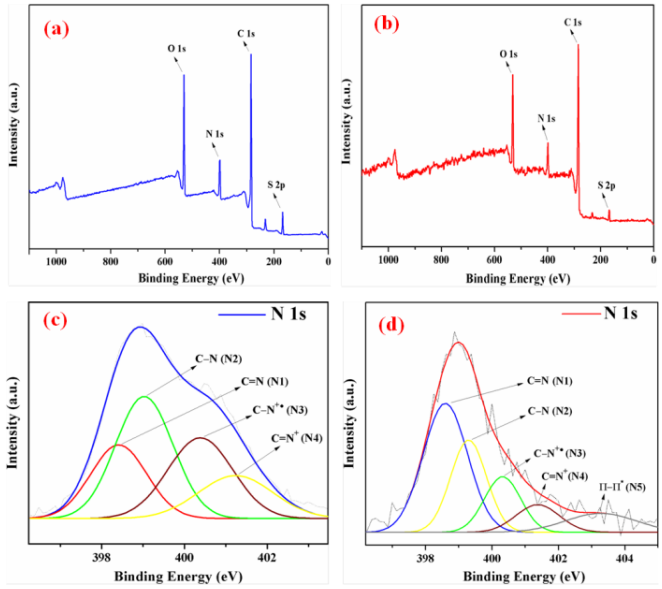


Fig. 3. (a) XPS wide scan survey spectrum of the pN_f PANI coating, (b) XPS wide scan survey spectrum of the N_g PANI coating, (c) XPS core level N (1s) deconvoluted spectrum of the pN_f PANI coating, (d) XPS core level N (1s) deconvoluted spectrum of the N_g PANI coating.

C. Humidity sensing studies

The Resistance verses % RH response characteristics were successively obtained (at a set potential of +1V) for the pN_f and N_g PANI morphologies of the PANI-paper composites and their respective potentials as effective humidity sensors were determined. As seen in Fig. 4a, the pN_f PANI-paper did not yield a unimodal humidity response beyond 54.5% RH. The obtained bimodal response can be attributed to the predominant polymer-swelling effect within the pN_f PANI network at higher humidity (*i.e.* ≥ 55% RH) [13].

Polymer swelling is known to adversely affect the crystallinity and electrical conductivity of a hygroscopic polymer network, which further gets magnified as the ambient humidity exceeds the threshold humidity (for polymer chain hydration) [13]. This triggers chain distortion in the hydrated conductive polymer network, thereby restricting the mobility of charge carriers. The pN_f PANI coating bore an appreciable ordering of polymer chains due to the residual 1-D nature of a compact mesh of partially aggregated nanofibers. This made it increasingly susceptible to the polymer-swelling effect, when chain hydration took place at higher humidity [13]. Moreover, the resultant chain distortion led to the increase in inter-chain distances within this PANI coating. This proved to be an irreversible process as the resistance of the pN_f PANI-paper continued to increase, even when the humidity was reverted to below 55% RH [26]. Hence, it can be concluded that the polymer-swelling effect got intensified, once the threshold humidity for water absorption was surpassed in the pN_f PANI coating containing 1-D nanostructural content [25]. Further, it appeared that the large intrinsic packing disorder in the intensively aggregated morphology of the N_g PANI coating increased its resistance to the polymer-swelling effect [13]. Hence, the N_g PANI-paper showed no appreciable side-effects of polymer swelling and gave a unimodal, linear humidity response within the investigated range of 16-96.2% RH. As seen in Fig. 4b, the resistance of the N_g PANI-paper exhibited more than a tenfold decrease ($\Delta R = R_{16} - R_{96.2}$) from an initial resistance of about 877.19 kΩ (R_{16}) at 16% RH, to a final resistance of about 76.92 kΩ ($R_{96.2}$) at 96.2% RH. The humidity response obtained was attributed to the following phenomena: "PEACE" (Proton Exchange Assisted Conduction of Electrons) and secondary doping mechanism of adsorbed water vapor molecules; Grotthuss and primary doping mechanism of condensed layers of water molecules; secondary doping mechanism of absorbed water molecules; cellulose-mediated inter-domain tunnelling of PANI charge carriers along with the superhydrophilicity and porosity of the cellulosic matrix [12-13, 21, 24, 31, 46-49]. The detailed humidity sensing mechanisms are presented in Section VIII of the Supporting Information.

The resistance versus % RH response characteristic of the N_g PANI-paper was linearly fitted as Resistance = -9.79(% RH) + 1026.21 kΩ ($R^2 \approx 0.996$ and no. of paper strips tested, N = 3). The obtained sensitivity of 9.79 kΩ/%RH was a significant improvement over other PANI-based resistive-type humidity sensing elements, like PANI-grafted cellulose film with a sensitivity of 0.26 kΩ/%RH [50] and PANI–Chitin blend with a sensitivity of 7.12 [$\Delta(\log_{10}R)$/%RH] [21]. Moreover, the observed sensing response [$(\Delta R/R_{16})$%] of 90.29% was higher than that obtained by using PANI-metal oxide composite pellets/films like PANI–Cr_2O_3 pellets (87%) [51], PANI–TiO_2 film (84.21%) [52], PANI–CeO_2 pellets (60%) [53], PANI–Nb_2O_5 film (80%) [54] and PANI–Pr_2O_3 pellets (85%) [55], over a wider linear sensing range of 16-96.2% RH. The obtained linear sensing range (16-96.2% RH) exhibited a significant increase as compared to other PANI-based impedimetric humidity sensing elements like PANI–CMC film (50-80% RH) [56] and PANI–Orange dye film (30-70% RH) [57]. Lastly, the theoretical humidity detection limit for the N_g PANI-paper was calculated from the resistance vs. % RH response characteristic (Fig. 4b) to be 4.32% RH [58].

To analyze the potential of the N_g PANI-paper as a chemiresistive humidity sensing element, a dynamic response study was performed within approximately 16-65.5% RH. As shown in Fig. 4c, the response time ($t_{95\%}$) and recovery time ($t_{105\%}$) were found to be 1300s and 2809s, respectively. On comparison of the successive dynamic response cycles, we found good repeatability (R.S.D. in the final response/recovery conductance values was 8.34%/4.71% over n = 3 cycles) and low hysteresis (maximum hysteresis = 6.8% RH at 65.5% RH) within the investigated humidity range [31, 59]. The response and recovery times of our paper-based humidity sensor were much improved than those of porous silicon particles on paper [60] and the response time was similar to that of silver nanoparticles'-based ink on paper [61]. The longer response/recovery times obtained as compared to

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

other paper-based humidity sensors was a result of the large thickness of the N_g PANI-paper (> 180µm), which resulted in slow adsorption/desorption of water vapor on/from it [21]. Further, the significant absorption of water molecules into the hygroscopic, partially amorphous PANI-paper network (at higher humidity) led to increased retardation of the desorption process, thereby resulting in a slower recovery [21]. The large thickness and hygroscopic nature of the composite also aggravated the degree of hysteresis, when it was exposed to higher humidity than about 65.5% RH (maximum hysteresis = 11.8% RH at 75.5% RH) (Fig. 4d) [21, 59-60]. These humidity cycling results were in agreement with those obtained for other PANI-containing humidity sensors including PANI–Chitin blend [21] and PANI–Orange dye composite [57]. Hence, by using a lower thickness (current thickness = 180µm) and enhanced crystalline form of cellulose paper, we could reduce the response/recovery times, while minimizing hysteresis over the entire humidity sensing range [21].

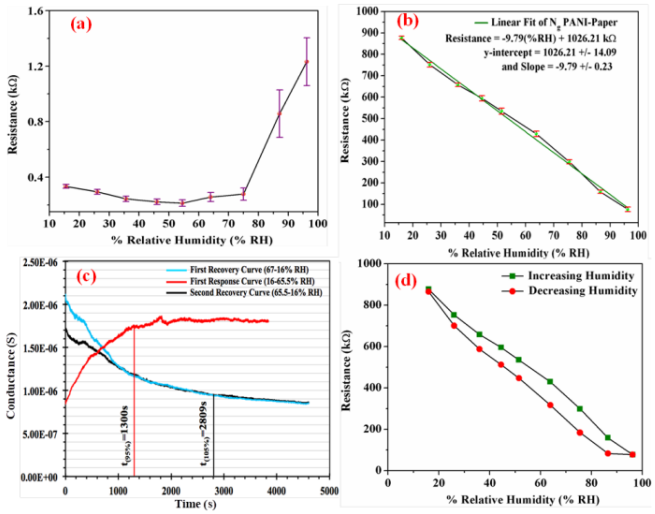


Fig. 4. (a) Resistance versus % RH response characteristic of the pN_f PANI-paper composite (Error bars represent S.D. for N = 3), (b) Resistance versus % RH response characteristic of the N_g PANI-paper composite (Error bars represent S.D. for N = 3), (c) Dynamic response study of the N_g PANI-paper composite and (d) Hysteresis characteristic of the N_g PANI-paper composite for humidity cycled between 16-96.2% RH.

Finally, a selectivity study for the N_g PANI-paper was performed at 64.5% RH (maintained with saturated $CoCl_2.6H_2O$ solution) by individually introducing 133.34 ppm each of SO_2, CO_2, CH_4, H_2 and NH_3 into the 150 ml humidity-controlled PVC chamber [27]. The sensing element exhibited reasonable selectivity to water vapor ($[(\Delta R/R_{16})\%] = 51.07\%$), with deviation less than 9% for all other gases except NH_3 ($[(\Delta R/R_{16})\%] = 4.45\%$). This behaviour of the N_g PANI coating is substantiated by previous works on PANI-based ammonia gas sensors including PANI-DBSA nanorods [62] and PANI-rGO nanocomposite films [63]. The results (Fig. S4) and explanations have been included in Section VIII of the Supporting Information.

Table S3 (Supporting Information) compares results of the studies conducted on the N_g PANI-paper with those reported for other cellulose paper-based humidity sensors. To the best of our knowledge, the sensitivity obtained (within the linear sensing range) by the N_g PANI integrated conductive paper is the highest amongst all the low-voltage (*i.e.* magnitude of operational voltage ≤ 5V), resistive, impedimetric and conductometric paper-based humidity sensors reported till date.

IV. CONCLUSION

We have synthesized two distinct morphologies of PANI coatings on a paper substrate by varying the parameters governing the *in situ* chemical oxidative polymerization mechanism of aniline. The nature of the humidity response obtained for each PANI-paper composite has been found to be governed by the intimate relationship between the polymer-swelling effect and the morphology of the respective PANI coating [13]. Accordingly, the swelling-resistant N_g morphology of PANI on paper has been utilized as a chemiresistive humidity sensing element, which can be operated at room temperature with a 1V DC supply. This N_g PANI integrated conductive paper has resulted in a unimodal humidity response, with more than a tenfold decrease in its resistance across a wide linear sensing range of 16-96.2% RH ($R^2 \approx 0.996$). The measured sensitivity of 9.79 kΩ/%RH, with a sensing response of 90.29% has been found to be higher than those obtained using several other PANI-based organic and inorganic hybrids [21, 50-51, 55]. Moreover, an improved combination of linearity, sensitivity and linear sensing range has revealed that PANI-paper composites hold promise as superior paper-based humidity sensors. The enhanced sensing performance of the N_g PANI-paper composite has been ascribed to the novel combination of a swelling-resistant morphology of the N_g PANI coating (comprising bulk nanostructures) with the superhydrophilic filter paper substrate. Lastly, attempts should also be made to fabricate a more sensitive humidity sensor based on N_g PANI, by increasing its doping level while maintaining its morphology. This can perhaps be realized by introducing other organic acid dopants (e.g. PTSA, SSA etc.) along with formic acid [43].

ASSOCIATED CONTENT

Supporting Information

Additional details regarding the apparatus used, *in situ* syntheses of the PANI-paper composites, fabrication of the humidity sensing setup and protocol for the humidity sensing studies, morphological studies, electrical conductivity studies, X-ray diffraction analyses, Fourier transform infrared spectroscopic investigations, X-ray photoelectron spectroscopic studies, humidity sensing mechanisms and selectivity study; along with a comparative table (Table S3) of low-voltage, cellulose paper-based humidity sensors have been presented.

Author contributions

[¥] Samar S. Sandhu and Dr. Saurabh Kumar contributed equally to this work.

ACKNOWLEDGMENT

The authors wish to thank Prof. Yogesh Singh, Vice Chancellor, Delhi Technological University (DTU), Delhi,

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

India for providing the necessary facilities. Saurabh Kumar acknowledges the Department of Science and Technology, New Delhi, India for the DST INSPIRE Faculty Award (DST/INSPIRE/04/2017/002750). B.D.M. thanks Science & Engineering Research Board (Govt. of India) for the award of a Distinguished Fellowship (SB/S9/YSCP/SERB-DF/2018(1)).

REFERENCES

[1] D. Tobjörk, and R. Österbacka, "Paper electronics," *Advanced Materials,* vol. 23, no. 17, pp. 1935-1961, 2011.

[2] S. Kumar, P. Rai, J. G. Sharma *et al.*, "PEDOT: PSS/PVA-Nanofibers-Decorated Conducting Paper for Cancer Diagnostics," *Advanced Materials Technologies,* vol. 1, no. 4, pp. 1600056, 2016.

[3] R. B. Channon, M. P. Nguyen, A. G. Scorzelli *et al.*, "Rapid flow in multilayer microfluidic paper-based analytical devices," *Lab on a Chip,* vol. 18, no. 5, pp. 793-802, 2018.

[4] Y. Zhang, L. Zhang, K. Cui *et al.*, "Flexible Electronics Based on Micro/Nanostructured Paper," *Advanced Materials*, pp. 1801588, 2018.

[5] I. Shown, A. Ganguly, L. C. Chen *et al.*, "Conducting polymer-based flexible supercapacitor," *Energy Science & Engineering,* vol. 3, no. 1, pp. 2-26, 2015.

[6] J. G. Ibanez, M. E. Rincón, S. Gutierrez-Granados *et al.*, "Conducting Polymers in the Fields of Energy, Environmental Remediation, and Chemical–Chiral Sensors," *Chemical reviews,* vol. 118, no. 9, pp. 4731-4816, 2018.

[7] C. Dhand, M. Das, M. Datta *et al.*, "Recent advances in polyaniline based biosensors," *Biosensors and Bioelectronics,* vol. 26, no. 6, pp. 2811-2821, 2011.

[8] E. Kang, K. Neoh, and K. Tan, "Polyaniline: a polymer with many interesting intrinsic redox states," *Progress in Polymer Science,* vol. 23, no. 2, pp. 277-324, 1998.

[9] I. Fratoddi, I. Venditti, C. Cametti *et al.*, "Chemiresistive polyaniline-based gas sensors: A mini review," *Sensors and Actuators B: Chemical,* vol. 220, pp. 534-548, 2015.

[10] T.-F. Wu, and J.-D. Hong, "Humidity sensing properties of transferable polyaniline thin films formed at the air–water interface," *RSC Advances,* vol. 6, no. 99, pp. 96935-96941, 2016.

[11] M. Das, S. Ghosh, and S. Roy, "Non-covalent functionalization of CNTs with polycarbazole: a chemiresistive humidity sensor with tunable chemo-electric attributes at room temperature," *New Journal of Chemistry,* vol. 42, no. 9, pp. 6918-6931, 2018.

[12] S. Jain, S. Chakane, A. Samui *et al.*, "Humidity sensing with weak acid-doped polyaniline and its composites," *Sensors and Actuators B: Chemical,* vol. 96, no. 1-2, pp. 124-129, 2003.

[13] F.-W. Zeng, X.-X. Liu, D. Diamond *et al.*, "Humidity sensors based on polyaniline nanofibres," *Sensors and Actuators B: Chemical,* vol. 143, no. 2, pp. 530-534, 2010.

[14] N. Fatima, F. Aziz, Z. Ahmad *et al.*, "Compositional engineering of the pi-conjugated small molecular VOPcPhO: Alq 3 complex to boost humidity sensing," *RSC Advances,* vol. 7, no. 32, pp. 19780-19786, 2017.

[15] J. Miao, Y. Chen, Y. Li *et al.*, "Proton Conducting Polyoxometalate/polypyrrole Films and their Humidity Sensing Performance," *ACS Applied Nano Materials*, 2018.

[16] E. Raza, M. Asif, F. Aziz *et al.*, "Influence of thermal annealing on a capacitive humidity sensor based on newly synthesized macroporous PBObzT2," *Sensors and Actuators B: Chemical,* vol. 235, pp. 146-153, 2016.

[17] Y.-N. Guo, Z.-Y. Gao, X.-X. Wang *et al.*, "A highly stretchable humidity sensor based on spandex covered yarns and nanostructured polyaniline," *RSC Advances,* vol. 8, no. 2, pp. 1078-1082, 2018.

[18] D. Zhang, D. Wang, X. Zong *et al.*, "High-performance QCM humidity sensor based on graphene oxide/tin oxide/polyaniline ternary nanocomposite prepared by in-situ oxidative polymerization method," *Sensors and Actuators B: Chemical,* vol. 262, pp. 531-541, 2018.

[19] T. Gomes, C. Constantino, E. Lopes *et al.*, "Thermal inkjet printing of polyaniline on paper," *Thin Solid Films,* vol. 520, no. 24, pp. 7200-7204, 2012.

[20] R. M. Morais, M. dos Santos Klem, G. L. Nogueira *et al.*, "Low cost humidity sensor based on PANI/PEDOT: PSS printed on paper," *IEEE Sensors Journal,* vol. 18, no. 7, pp. 2647-2651, 2018.

[21] A. Ramaprasad, and V. Rao, "Chitin–polyaniline blend as humidity sensor," *Sensors and Actuators B: Chemical,* vol. 148, no. 1, pp. 117-125, 2010.

[22] S. Kundu, R. Majumder, R. Ghosh *et al.*, "Relative humidity sensing properties of doped polyaniline-encased multiwall carbon nanotubes: wearable and flexible human respiration monitoring application," *Journal of Materials Science,* vol. 55, no. 9, pp. 3884-3901, 2020.

[23] S. Manjunatha, T. Machappa, Y. Ravikiran *et al.*, "Polyaniline based stable humidity sensor operable at room temperature," *Physica B: Condensed Matter,* vol. 561, pp. 170-178, 2019.

[24] S. T. McGovern, G. M. Spinks, and G. G. Wallace, "Micro-humidity sensors based on a processable polyaniline blend," *Sensors and Actuators B: Chemical,* vol. 107, no. 2, pp. 657-665, 2005.

[25] A. Balkan, E. Armagan, and G. O. Ince, "Synthesis of coaxial nanotubes of polyaniline and poly (hydroxyethyl methacrylate) by oxidative/initiated chemical vapor deposition," *Beilstein journal of nanotechnology,* vol. 8, pp. 872, 2017.

[26] M. Joulazadeh, A. H. Navarchian, and M. Niroomand, "A comparative study on humidity sensing performances of polyaniline and polypyrrole nanostructures," *Advances in Polymer Technology,* vol. 33, no. S1, 2014.

[27] S. Y. Park, J. E. Lee, Y. H. Kim *et al.*, "Room temperature humidity sensors based on rGO/MoS2 hybrid composites synthesized by hydrothermal method," *Sensors and Actuators B: Chemical,* vol. 258, pp. 775-782, 2018.

[28] N. D. K. Tu, J. Choi, C. R. Park *et al.*, "Remarkable conversion between n-and p-type reduced graphene oxide on varying the thermal annealing temperature," *Chemistry of Materials,* vol. 27, no. 21, pp. 7362-7369, 2015.

[29] Z.-l. Mo, Z.-l. Zhao, H. Chen *et al.*, "Heterogeneous preparation of cellulose–polyaniline conductive composites with cellulose activated by acids and its electrical properties," *Carbohydrate Polymers,* vol. 75, no. 4, pp. 660-664, 2009.

[30] W. Hu, S. Chen, Z. Yang *et al.*, "Flexible electrically conductive nanocomposite membrane based on bacterial cellulose and polyaniline," *The Journal of physical chemistry B,* vol. 115, no. 26, pp. 8453-8457, 2011.

[31] J.-W. Han, B. Kim, J. Li *et al.*, "Carbon nanotube based humidity sensor on cellulose paper," *The Journal of Physical Chemistry C,* vol. 116, no. 41, pp. 22094-22097, 2012.

[32] I. Sapurina, and J. Stejskal, "The mechanism of the oxidative polymerization of aniline and the formation of supramolecular polyaniline structures," *Polymer International,* vol. 57, no. 12, pp. 1295-1325, 2008.

[33] H. D. Tran, Y. Wang, J. M. D'Arcy *et al.*, "Toward an understanding of the formation of conducting polymer nanofibers," *ACS nano,* vol. 2, no. 9, pp. 1841-1848, 2008.

[34] D. Li, and R. B. Kaner, "Shape and aggregation control of nanoparticles: not shaken, not stirred," *Journal of the American Chemical Society,* vol. 128, no. 3, pp. 968-975, 2006.

[35] N. D. Tissera, R. N. Wijesena, S. Rathnayake *et al.*, "Heterogeneous in situ polymerization of polyaniline (PANI) nanofibers on cotton textiles: Improved electrical conductivity, electrical switching, and tuning properties," *Carbohydrate Polymers,* vol. 186, pp. 35-44, 2018.

[36] E. Gomes, and M. Oliveira, "Chemical polymerization of aniline in hydrochloric acid (HCl) and formic acid (HCOOH) media. Differences between the two synthesized polyanilines," *Am. J. Polym. Sci,* vol. 2, no. 2, pp. 5-13, 2012.

[37] A. Dan, and P. Sengupta, "Synthesis and characterization of polyaniline prepared in formic acid medium," *Journal of applied polymer science,* vol. 91, no. 2, pp. 991-999, 2004.

[38] M. Trchová, and J. Stejskal, "Polyaniline: The infrared spectroscopy of conducting polymer nanotubes (IUPAC Technical Report)," *Pure and Applied Chemistry,* vol. 83, no. 10, pp. 1803-1817, 2011.

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2020.3001599, IEEE Sensors Journal

[39] S. Golczak, A. Kanciurzewska, M. Fahlman *et al.*, "Comparative XPS surface study of polyaniline thin films," *Solid State Ionics,* vol. 179, no. 39, pp. 2234-2239, 2008.

[40] S. N. Kumar, F. Gaillard, G. Bouyssoux *et al.*, "High-resolution XPS studies of electrochemically synthesized conducting polyaniline films," *Synthetic Metals,* vol. 36, no. 1, pp. 111-127, 1990.

[41] S. Kumar, K. K. Jagadeesan, A. G. Joshi *et al.*, "Immuno-CoPS (conducting paper strips) for futuristic cost-effective cancer diagnostics," *RSC Advances,* vol. 3, no. 29, pp. 11846-11853, 2013.

[42] K. Deb, A. Bera, and B. Saha, "Tuning of electrical and optical properties of polyaniline incorporated functional paper for flexible circuits through oxidative chemical polymerization," *RSC Advances,* vol. 6, no. 97, pp. 94795-94802, 2016.

[43] H. Mao, X. Wu, X. Qian *et al.*, "Conductivity and flame retardancy of polyaniline-deposited functional cellulosic paper doped with organic sulfonic acids," *Cellulose,* vol. 21, no. 1, pp. 697-704, 2014.

[44] A. A. Qaiser, M. M. Hyland, and D. A. Patterson, "Surface and charge transport characterization of polyaniline− cellulose acetate composite membranes," *The Journal of physical chemistry B,* vol. 115, no. 7, pp. 1652-1661, 2011.

[45] J. Li, X. Qian, L. Wang *et al.*, "XPS characterization and percolation behavior of polyaniline-coated conductive paper," *BioResources,* vol. 5, no. 2, pp. 712-726, 2010.

[46] J. Travers, and M. Nechtschein, "Water effects in polyaniline: a new conduction process," *Synthetic Metals,* vol. 21, no. 1-3, pp. 135-141, 1987.

[47] A. MacDiarmid, and A. J. Epstein, "The concept of secondary doping as applied to polyaniline," *Synthetic Metals,* vol. 65, no. 2-3, pp. 103-116, 1994.

[48] B. Lubentsov, O. Timofeeva, S. Saratovskikh *et al.*, "The study of conducting polymer interaction with gaseous substances IV. The water content influence on polyaniline crystal structure and conductivity," *Synthetic Metals,* vol. 47, no. 2, pp. 187-192, 1992.

[49] Z. Chen, and C. Lu, "Humidity sensors: a review of materials and mechanisms," *Sensor letters,* vol. 3, no. 4, pp. 274-295, 2005.

[50] S. Shukla, "Synthesis of polyaniline grafted cellulose suitable for humidity sensing," *Indian Journal of Engineering & Materials Sciences,* vol. 19, pp. 417-420, 2012.

[51] K. Sajjan, A. S. Roy, A. Parveen *et al.*, "Analysis of DC and AC properties of a humidity sensor based on polyaniline–chromium oxide composites," *Journal of Materials Science: Materials in Electronics,* vol. 25, no. 3, pp. 1237-1243, 2014.

[52] S. Kotresh, Y. Ravikiran, H. R. Prakash *et al.*, "PolyanilineTitanium dioxide composite as humidity sensor at room temperature," *Nanosystems: Physics, Chemistry, Mathematics,* vol. 7, no. 4, pp. 732, 2016.

[53] B. Geeta, K. Bikshalu, V. Rajendar *et al.*, "Nanostructured conducting polyaniline (NSPANI)/CeO2 nanocomposites for humidity sensors application," *Journal of Materials Science: Materials in Electronics,* vol. 29, no. 1, pp. 374-381, 2018.

[54] S. Kotresh, Y. Ravikiran, S. Vijayakumari *et al.*, "Polyaniline niobium pentoxide composite as humidity sensor at room temperature," *Adv. Mater. Lett,* vol. 6, pp. 641-645, 2015.

[55] S. Nagaraju, A. S. Roy, J. Kumar *et al.*, "Humidity sensing properties of surface modified polyaniline metal oxide composites," *Journal of Engineering,* vol. 2014, 2014.

[56] S. Kotresh, Y. Ravikiran, H. R. Prakash *et al.*, "Humidity sensing performance of spin coated polyaniline–carboxymethyl cellulose composite at room temperature," *Cellulose,* vol. 23, no. 5, pp. 3177-3186, 2016.

[57] M. T. S. Chani, K. S. Karimov, F. Khalid *et al.*, "Orange dye—polyaniline composite based impedance humidity sensors," *Chinese Physics B,* vol. 22, no. 1, pp. 010701, 2013.

[58] Y. H. Kim, S. J. Kim, Y.-J. Kim *et al.*, "Self-activated transparent all-graphene gas sensor with endurance to humidity and mechanical bending," *ACS nano,* vol. 9, no. 10, pp. 10453-10460, 2015.

[59] P. Zhu, Y. Liu, Z. Fang *et al.*, "Flexible and highly sensitive humidity sensor based on cellulose nanofibers and carbon nanotube composite film," *Langmuir,* vol. 35, no. 14, pp. 4834-4842, 2019.

[60] T. Jalkanen, A. Määttänen, E. Mäkilä *et al.*, "Fabrication of porous silicon based humidity sensing elements on paper," *Journal of Sensors,* vol. 2015, 2015.

[61] A. Quddious, S. Yang, M. M. Khan *et al.*, "Disposable, Paper-Based, Inkjet-Printed Humidity and H2S Gas Sensor for Passive Sensing Applications," *Sensors,* vol. 16, no. 12, pp. 2073, 2016.

[62] A. Yadav, J. Kumar, M. Shahabuddin *et al.*, "Improved Ammonia Sensing by Solution Processed Dodecyl Benzene Sulfonic Acid Doped Polyaniline Nanorod Networks," *IEEE Access,* vol. 7, pp. 139571-139579, 2019.

[63] C.-T. Lee, and Y.-S. Wang, "High-performance room temperature NH3 gas sensors based on polyaniline-reduced graphene oxide nanocomposite sensitive membrane," *Journal of Alloys and Compounds,* vol. 789, pp. 693-696, 2019.



SAMAR S. SANDHU received his B.Tech. degree in engineering physics from the Delhi Technological University, Delhi, India, in 2018. He is currently pursuing a master's degree in nanoengineering at the University of California San Diego, La Jolla, CA, USA. From 2016 to 2019, he was an undergraduate research student at the NanoBioelectronics Laboratory and the Nanomaterials Research Laboratory, at the Delhi Technological University, Delhi, India. His research interests include the development of portable/wearable environmental monitoring and health monitoring devices, top-down and bottom-up nanofabrication techniques for obtaining functional nanostructures along with their diverse electroanalytical applications.



BANSI D. MALHOTRA received his PhD from the University of Delhi, Delhi in 1980. He has published more than 300 papers in refereed international journals, has filed 12 patents, and has recently co-authored text books on "Nanomaterials for Biosensors: Fundamentals and Applications" and "Biosensors: Fundamentals and Applications". He is a recipient of the National Research Development Corporation Award 2005 for invention on "Blood Glucose Biochemical Analyzer" and is a Fellow of the Indian National Science Academy, the National Academy of Sciences, India and an Academician of Asia Pacific Academy of Materials. His current research activities include Nanobiomaterials, Biosensors, Bio-fuel cells, Ordered Molecular Assemblies, Conducting Polymers, Langmuir-Blodgett Films, Self-assembled Monolayers, Nano-Biotechnology, Biomedical Engineering, and Biomolecular Electronics. He is currently a SERB Distinguished Fellow (Govt. of India) at the Department of Biotechnology, Delhi Technological University, Delhi, India.

1558-1748 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 06,2020 at 05:14:00 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Natural Language Processing based Visual Question Answering Efficient: an EfficientDet Approach

Rahul Gupta
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
Rahulgupta100689@gmail.com

Parikshit Hooda
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
Phooda804@gmail.com

Sanjeev
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
Sanjeevdhangar89@gmail.com

Nikhil Kumar Chikkara
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
nikhilkumarchikkara@gmail.com

***Abstract*— This research paper is intended to propose a technique to solve the task of image-based question answering using EfficientDet and Bidirectional LSTM (BiLSTM). Visual Question Answering (VQA) requires a combination of Image recognition and Natural Language Processing (NLP) techniques. The proposed technique uses EfficientDet for image processing and BiLSTM for question processing. The model acts efficiently because of efficient image processing. The model takes an image and a question as input processes them independently, then fuses the processed versions and predicts the answer to the question. It outputs the result in an open domain based on input questions on the image. This study intends to help the visually impaired person. The model achieves similar performance as its contemporary models on the benchmark VQA dataset.**

Keywords*— *Natural Language Processing (NLP), Visual Question Answering (VQA), Long Short-term Memory (LSTM)

I. INTRODUCTION (*HEADING 1*)

Visual Question Answering (VQA) is the task to answer natural-language queries (or a question) on the given image. Imagine a computer system that looks at an image and answers natural language questions related to it. Questions can be open-ended or structured [1]. A question can be arbitrary and may involve many Computer Vision sub-problems. Taking the Fig, 1 as an example, we see the following sub-problems:

- Attribute Classification: "What is the cat's color?", and "Is the man wearing spectacles?"
- Object recognition: "What is the hand-holding?"
- Scene classification: "Where is the man sitting?"
- Counting: "How many zebras are there?"
- Object detection: "Are there any zebras in the image?"

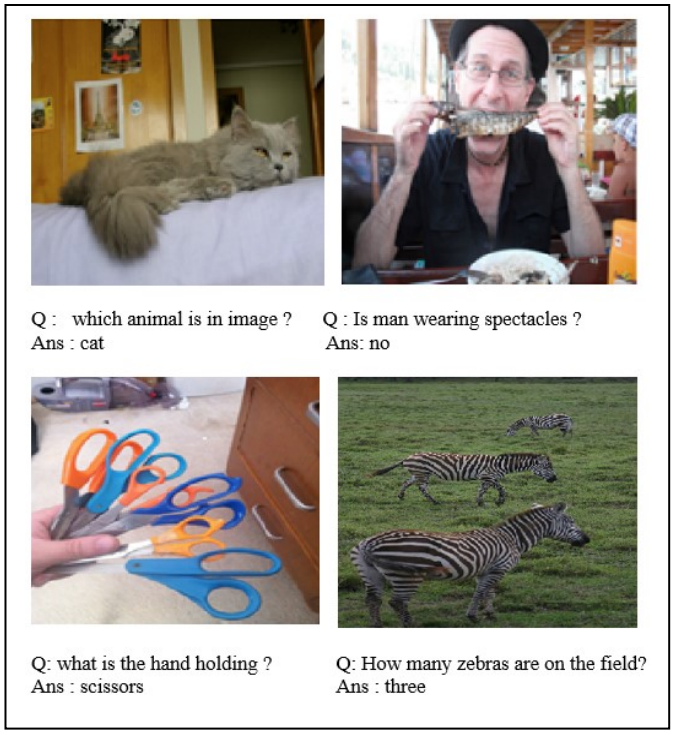


Fig. 1. Example of image-based question answering

To answer the above questions, the model must have a good understanding of the question as well as image features. The the answer can be structured and open-ended as well, corresponding to the form of the question. VQA is a multidisciplinary Artificial Intelligence (AI) research problem. As the dataset used for VQA is large and can expand considering the advent of big data, there is a need for an efficient and scalable VQA model. Recently, EfficientDet [2]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:38 UTC from IEEE Xplore. Restrictions apply.

achieved a state of the art performance on object detection task with 10 times

better efficiency. EfficientDet is based on a new family of Convolutional Neural Networks(CNN) called the EfficientNets [3]. Bidirectional LSTM [4] networks have been used recently and have gained prominence due to their ability to retain sequential information. This work proposes to leverage the EfficientDet's performance and BiLSTM's memory to Visual based question answering. It's a multidisciplinary artificial intelligence research problem. Generally, VQA is defined as a system that takes as input an image and a natural language question about the image and generates an accurate natural language answer as the output.

This paper is divided into the following sections: Section 2 details the related work has already done, Section 3 describes the Methodology of our technique, and Section 4 details the results of our work.

II. RELATED WORK

VQA has been exploited to a deeper level in recent years [1], [5], [6]. There are various models present for VQA. Most of the models are based on the attention mechanism [7]–[9]. The first study on visual-based question answering was done in [1]. Ideas about combining image and linguistic features were discussed based on late fusion. Later, attention models as top-down attention model [5], RCNN [5], visual attention using capsnet [10] models were presented to solve the VQA problem. It showed some promising results. (the red part correction here)
It helps in providing the optimal number of regions for attention positioning. The issue with attentional positioning is that sometimes, the regions specified by the attention model are poorly aligned and it becomes hard to detect objects and adhere to the visual features associated with the object in the image.

Many datasets [11] were used during the evolution of the VQA task too. However, now the VQA dataset based on Microsoft Common Objects in Context [12] dataset has gained prominence because of the sheer volume of images and image classes. There are two versions of the VQA dataset, v1.0, and v2.0. The training is done on datasets from MSCOCO-2014 while testing is done on MSCOCO-2015 for the VQA v2.0 dataset [11], [12]. It contains 82,783 training and 40,504 validation images. There is a minimum of 3 questions to each image for the VQA dataset. Each question was answered independently by 10 human judges and their responses were recorded to form the dataset.

III. METHODOLOGY

The overall architecture of the model which is depicted in Fig 4. consists of 4 major components:

- A preprocessing layer that converts the dataset into a suitable form by processing questions and images.
- Bidirectional LSTM layer for extraction of temporal relationships between words of the question, which acts on word embeddings generated using GloVe [13].
- An EfficientDet based image features' extraction component to efficiently process images.
- The final layer to combine the question and image features and compile the model.

A. Preprocessing Layer

The dataset is first preprocessed by collecting question id, image path, question, and answer into an object form in a different file. Initially, the question text is concatenated with the answer text. The formed string is turned to lower case and unnecessary words such as stop words are removed. The processed string is then tokenized and a dictionary of words is created. Each token is then mapped to an index and stored.

Then, every answer is stored as an index-word pair. A list of top 1000 candidate answers is generated and stored for later use. The candidate answers list contains top 1000 answers that appear in the dataset, in order.

B. Bidirectional LSTM Layer

An LSTM network is used to extract contextual feature representation of the questions in the dataset. A unit of an LSTM structure is a memory cell which preserves the state of a sequence. Internally, an LSTM unit consists of three gates and a memory cell [4].

A bidirectional LSTM, as shown in Fig. 2 maintains contextual information from the past and the future. The preprocessed question sentence is represented by S,

$$S = [w_o, w_1, w_2, \dots w_T] \quad (1)$$

where T is the total length of the sentence and w_i represents the word token at index i. The sentence S is passed through GloVe embedding to obtain the word embeddings X' of sentence S. Representing the word's representation at index i by x'_i, the question representation is the collection of word representation for all words in the question, ordered by index. The collection is then denoted by X'

$$X' = (x'_0, x'_1, x'_2, \dots, x'_T) \quad (2)$$

The collection of word representations is then fed into the forward and backward layer of BiLSTM, where a single word embedding goes into one LSTM memory cell. Representing the forward layer and backward layer of BiLSTM by $\vec{H}$ and $\overleftarrow{H}$ respectively, we have

$$\vec{H} = \left[\overrightarrow{h_0}, \overrightarrow{h_1}, \overrightarrow{h_2}, \dots, \overrightarrow{h_T}\right] \quad (3)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:38 UTC from IEEE Xplore. Restrictions apply.

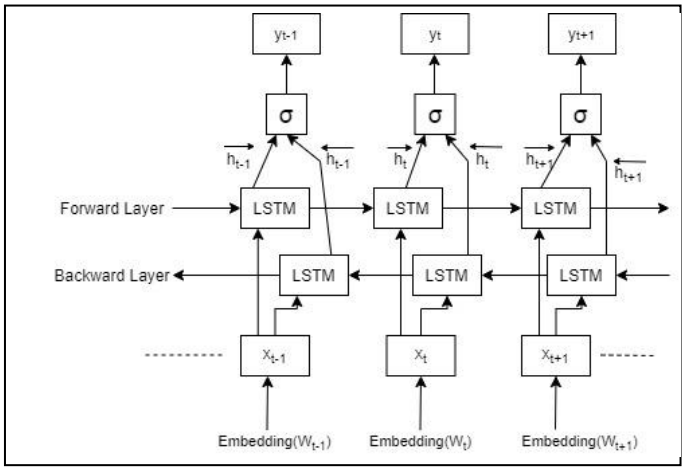


Fig. 2. The architecture of Bidirectional LSTM

$$\overleftrightarrow{H} = [\overrightarrow{h_T}, \overleftarrow{h_{T-1}}, \overleftrightarrow{h_{T-2}}, \dots, \overleftarrow{h_0}] \quad (4)$$

where, $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ denote the forward and backward LSTM feature for word at index i, respectively. Here, T is the length of the pre-processed question.
The output from feeding the word representation at index i into forward and a backward layer of the BiLSTM network is then

passed through tanh activation function after concatenating both layers' output for the word, which is denoted by y_i. So, y_i is the output from the BiLSTM network for word token at index i.

$$h_i = \overrightarrow{h_i} \oplus \overleftarrow{h_i} \quad (5)$$

$$y_i = \tanh(h_i + b) \quad (6)$$

Here, b is biased and is a learnable parameter, learned on the training data.

The collection of y_i that represents the question are concatenated to yield the final question representation.

$$y = [y_0, y_1, y_2, \dots y_T] \quad (7)$$

This output from this model is fed into the final layer for training the VQA model.

C. EfficientDet Layer

EfficientDets are a lightweight CNN architecture that achieves state-of-the-art accuracy with an order of magnitude better efficiency. The baseline architecture is scaled using the compound scaling method to obtain more information from the image. It uses a compound scaling method that uses a compound coefficient to uniformly scale the model. EfficientDet is used for image processing. It has two major components: (1) EfficientNet based backbone, (2) Bidirectional Feature Pyramid Network (BiFPN) [4] for image feature extraction. EfficientNet [3] backbone is a lightweight CNN architecture that achieves state-of-the-art accuracy on Imagenet [14] with an order of magnitude better efficiency. EfficientNet backbone is essentially a series of convolution (MBConv) blocks with activation at the end of each block. The baseline EfficientNet architecture is scaled using the compound scaling method [3] to allow capturing of more information from the image, resulting in performance increment. To scale the EfficientNet to capture more features, the scaling factors realized were:

$$\alpha=1.2,\ \beta=1.1,\ \gamma=1.15$$

Here, α is the depth factor, β is the width factor, γ is the resolution factor.

As they are optimal, we decide to use them in our model. The EfficientNet used is pretrained on Retinanet for this paper. It is used with BiFPN as a feature extractor. The features of an input image are extracted from EfficientNet using BiFPN. These features are then repeatedly fused by using Normalized Fast Fusion [2] technique to further improve efficiency.

$$O = \Sigma_i \frac{w_i}{\varepsilon + \Sigma_j \omega_j} \cdot I_i \quad (8)$$

The output, O is between 0 and 1. The fused features are fed to the final layer.

D. Final Layer

The final model is shown in Fig. 2. The question and image representations are extracted from the Bidirectional LSTM layer and the EfficientDet component respectively. The output from both layers is concatenated to yield a single model. The resulting model uses sigmoid activation and Adam optimization with loss as 'categorical_crossentropy'. The final component is a Softmax activation unit which gives a probability distribution on the candidate answer list generated in the preprocessing layer. The model is then trained.

IV. EXPERIMENTS

A. Datasets

We evaluate our model on the VQA v2.0 dataset [12] provided by visual. The dataset splits are downloaded from the official website. The VQA v2.0 dataset contains images from the MSCOCO dataset. The question-answer pairs are words or phrases. The dataset has 443,757 training question-answer pairs, 214,354 validation question-answer pairs, 447,793 testing pairs. The dataset also contains 82,783 training images, 40,504 validation images, and 81,434 testing images.

B. Evaluation Metrics

During the construction of the VQA dataset, the accuracy metric used was

$$Acc(ans) = min\{\frac{\#humans\ that\ said\ ans}{3}, 1\} \quad (9)$$

Where '#humans that said ans' is the number of humans that gave 'ans' as the answer. This metric reflects inter-human variability in phrasing the answers. For our model, the answer is first processed to normalize it, and then the percentage of right answers generated by the model is calculated. The answer generated by the model is matched with the answers in the dataset to label it as right or wrong.

C. Result

The experimental results on VQA v2.0 real-image dataset are presented in Table 1. The experimental results show that our model gives similar performance as the contemporary VQA

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:38 UTC from IEEE Xplore. Restrictions apply.

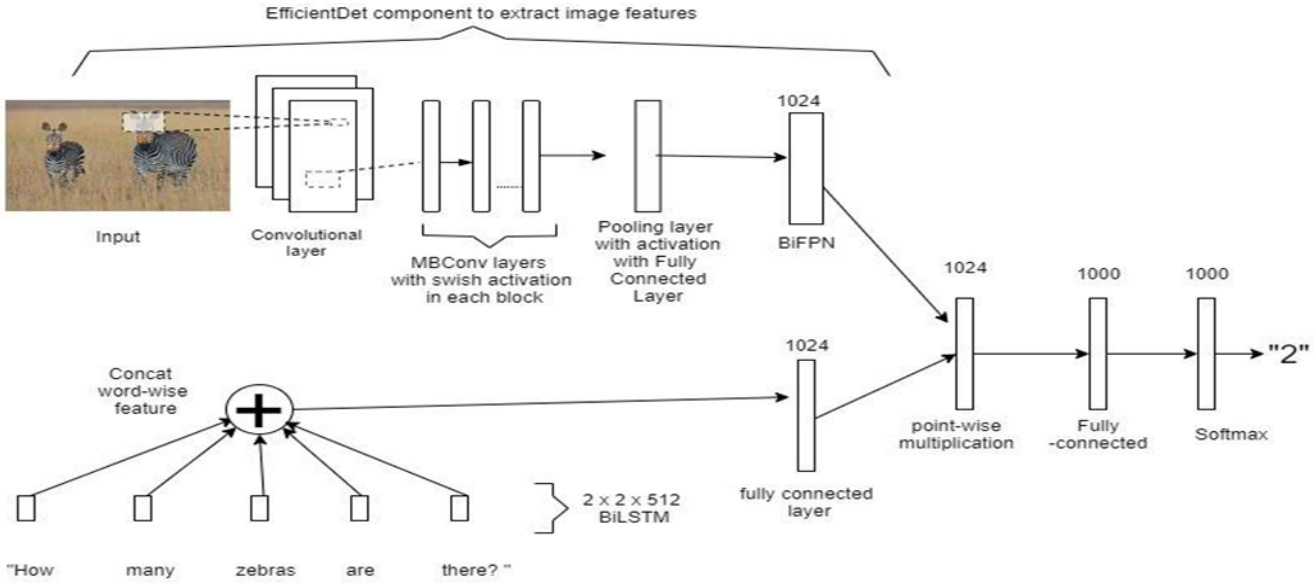


Fig 3 Overall architecture of the model

models in the joint embedding domain exclusively [1]. Our model outperforms the baseline Deeper LSTM + Norm I by 81% and 7.05% on Val and test split, respectivelyThe model was trained for 40 epochs with a batch size of 512 on Google Colab Pro and Keras. Significant improvements in evaluation metrics are observed on the VQA v2.0 dataset.

V. CONCLUSION

In this paper, we presented an improved model that uses EfficientNet and BiLSTM to extract more accurate image features and question features respectively. Experimental results show that our model achieves better performance than models that use VGG19 and ResNet152(used by deeper LSTM + norm I) as the image modeling layer for the VQA task on VQA v2.0 dataset because of EfficientDet's superior accuracy. Our model is effective in contextual feature extraction of questions because of the BiLSTM layer that can process past and future information which works well on sequential modeling tasks. Our models performance is less than the current state-of-the-art models like Oscar [15] (73.82%, because of the usage of detected object tags as anchor points to easy alignment's learning), LXMERT [16] (72.54%, because of learned intra-modality and cross-modality relationships) and other attention-based models.

TABLE I. TABLE FOR ACCURACY COMPARISON FOR DIFFERENT MODELS

a. Val split according to visualqa split
b. Test split according to visualqa split
c. Split according to original paper [17]

Model name	Split	Accuracy
Deeper LSTM + Norm 1[3]	Val[a]	51.62
	Test[b]	54.22
LSTM + VGGNet [7]	Val[a]	54..48
	Test[b]	-
MCB [8]	Val[a]	59.14
	Test[b]	-
SAN(2, LSTM) [15]	Val[c]	57.3
	Test[c]	-
Our Model	Val[a]	59.43
	Test[b]	61.27

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:38 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] A. Agrawal *et al.*, "VQA: Visual Question Answering," *arXiv:1505.00468 [cs]*, Oct. 2016.

[2] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and Efficient Object Detection," Nov. 2019.

[3] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *arXiv:1905.11946 [cs, stat]*, Nov. 2019.

[4] I. Chowdhury, K. Nguyen, C. Fookes, and S. Sridharan, "A cascaded long short-term memory (LSTM) driven generic visual question answering (VQA)," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 1842–1846, doi: 10.1109/ICIP.2017.8296600.

[5] P. Anderson *et al.*, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *arXiv:1707.07998 [cs]*, Mar. 2018.

[6] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge," *arXiv:1708.02711 [cs]*, Aug. 2017.

[7] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee, "Visual Question Answering using Deep Learning: A Survey and Performance Analysis," *arXiv:1909.01860 [cs]*, Aug. 2019.

[8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding," *arXiv:1606.01847 [cs]*, Sep. 2016.

[9] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep Modular Co-Attention Networks for Visual Question Answering," *arXiv:1906.10770 [cs]*, Jun. 2019.

[10] "(21) Dynamic Capsule Attention for Visual Question Answering." [Online]. Available: https://www.researchgate.net/publication/335229436_Dynamic_Capsule_Attention_for_Visual_Question_Answering. [Accessed: 24-Apr-2020].

[11] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual Question Answering: A Survey of Methods and Datasets," *arXiv:1607.05910 [cs]*, Jul. 2016.

[12] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Computer Vision and Image Understanding*, vol. 163, pp. 3–20, Oct. 2017, doi: 10.1016/j.cviu.2017.06.005.

[13] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global Vectors for Word Representation," in *EMNLP*, 2014, doi: 10.3115/v1/D14-1162.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

15] X. Li *et al.*, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," *arXiv:2004.06165 [cs]*, Apr. 2020

16] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," *arXiv:1908.07490 [cs]*, Dec. 2019.

[17] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Question Answering," *arXiv:1511.02274 [cs]*, Jan. 2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:28:38 UTC from IEEE Xplore. Restrictions apply.

#NotAWhore! A Computational Linguistic Perspective of Rape Culture and Victimization on Social Media

Ashima Suvarna*
Delhi Technological University
Department of Computer Engineering
asuvarna31@gmail.com

Grusha Bhalla*
Delhi Technological University
Department of Computer Engineering
grushabhalla@gmail.com

Abstract

The recent surge in online forums and movements supporting sexual assault survivors has led to the emergence of a 'virtual bubble' where survivors can recount their stories. However, this also makes the survivors vulnerable to bullying, trolling and victim blaming. Specifically, victim blaming has been shown to have acute psychological effects on the survivors and further discourage formal reporting of such crimes. Therefore, it is important to devise computationally relevant methods to identify and prevent victim blaming to protect the victims. In our work, we discuss the drastic effects of victim blaming through a short case study and then propose a single step transfer-learning based classification method to identify victim blaming language on Twitter. Finally, we compare the performance of our proposed model against various deep learning and machine learning models on a manually annotated domain-specific dataset.

1 Introduction

Global statistics indicate that 35% of women worldwide have experienced sexual violence at some point in their lives[1]. Popular hashtags like '#metoo', '#sexualharassment' on Twitter have encouraged victims to share their stories of sexual assault and formally report them. However, the backlash faced by the victims has been staggering. Victims of sexual assault are often held culpable for the assault, and are attacked on social forums by extremists. With the rise of such crimes, it is important to devise a computational framework that can identify and prevent online victimization of sexual assault survivors who choose to report the crime. Ambiguous interpretations of rape culture and victim blaming makes manually sorting and identifying such information an arduous task. Hence, in our work we have attempted to identify an objectively grounded definition of victim blaming for further research in this domain.

Victim blaming occurs when the victim of a crime or any wrongful act is held entirely or partially at fault for the harm that befell them (Coates et al., 2006). Additionally, "slut shaming" is a popular form of victim blaming which refers to attacking a person's character on the basis of sexual activity, real or perceived (Ringrose and Renold, 2012). Victims of sexual assault are initially hesitant to make a sexual assault complaint and often encounter victim blaming and slut shaming attitudes when they finally do (Ahrens, 2006). This blame can appear in the form of toxic social responses from medical professionals, the media, the judiciary or a growing majority of online activists on social media platforms (Campbell et al., 2009).

Social platforms like Twitter and Facebook provide victims with a 'virtual bubble' to recollect the assault stories and seek emotional help. The victim blaming faced by these victims, however, discourages them from disclosing their personal stories and further seeking medical help (Verdun-Jones and Rossiter, 2010). Therefore, in this work, we propose a method to identify such language on Twitter and protect the victims who choose to disclose their plight. We propose a Twitter-specific classification model which can exclusively identify victim blaming tweets. The key contributions of this work are:

- Our work is the first attempt in devising a computational framework for identifying victim blaming language.
- We provide a manually annotated dataset that contains 5,070 tweets for further research in this domain.
- We propose a single step transfer learning based classification method that identifies victim blaming language and labels it. It obtained

*Authors contributed equally

[1] http://worldpopulationreview.com/countries/rape-statistics-by-country

superior results to many deep learning and machine learning based approaches.

2 Related Work

Prior research has shown sexual assault is a crime that women are most afraid of (Koss, 1993). Oftentimes, victims of sexual assault are subjected to humiliation, blaming because of which, reasonable doubt is created about their credibility (Ullman, 2000). Popular theories such as the "just world" (Lerner, 1980) theory and the "invulnerability" (Andrew et al., 2003) theory explain the psychological motivation behind victim blaming. The "just world" theory states that people get what they deserve and deserve what they get, thereby shifting the blame of the crime to the victim while the "invulnerability" theory states that people blame the victim to project their own sense of invulnerability.

Victim Blaming, therefore, stems from an individual's personal sense of insecurity and acts as a silencing function for most victims who are discouraged to disclose their personal stories or seek any help online (Ahrens, 2006). Since victims fail to obtain the required medical assistance, they become highly susceptible to emotional difficulties that manifest as depression in the short term and acute psychological difficulties in the long run (Verdun-Jones and Rossiter, 2010). In addition to this, chronosystem factors like past instances of victimization and sexual revictimization affect the mental health outcomes of the survivor often leading to suicidal behaviour, substance use, depression etc. (Campbell et al., 2009).

With the advent of social media, a new medium has presented itself for victim blaming to occur. Social platforms like Twitter, Facebook and Reddit provide a space to publicly post comments and present an insight into community opinions for researchers and social scientists. Due to its increasing popularity, Twitter is being used for research in opinion mining (Andleeb et al., 2019), keyword extraction (Biswas, 2019), hate speech detection (Badjatiya et al., 2017) etc. It has been widely used for research on sexual violence (Wekerle et al., 2018) as well as suicidal ideation using linear and ensemble classifiers (Sawhney et al., 2018). Research has also been focused on hate speech detection for Twitter using deep learning techniques, classifying tweets as sexist, racist or neither (Badjatiya et al., 2017). Balakrishnan et al. (2020) have used Naive Bayes, Random Forest and J48 for detection of cyberbullying. Due to the similarities between victim blaming and cyberbullying, we have used Naive Bayes as one of our baseline models. Schrading (2015) analyzes discussions on domestic abuse across social media, using LSTM, Naive Bayes, Logistic Regression, and SVM which has been used as a baseline against our proposed method because of its good performance.

2.1 Motivation: Weibo Victim Scandal

Liu Jingyao, a 21 years old student at the University of Minnesota accused Liu Qiangdong the founder of Chinas largest company JD.com, of raping her. She did not report her case immediately as she was afraid that she would be blamed. After the case became popular in China, people commented things like The woman looks disgusting, She is a slut etc. on Weibo, the Chinese equivalent of Twitter. She suffered from post traumatic stress disorder and insomnia because of this.[2]

The Weibo Case study is a classic example of the drastic effects of victim blaming on the victim and it's prevalence in our society. Reporting of instances of sexual violence has shown to pre-empt blame in the talk of women reporting blame which further shows that victim blaming itself is marked by specific topics and framing of sentences that shifts the blame onto the victim. Parameters like location of incident, state of victim etc. can be used for identifying such instances (Stubbs-Richardson et al., 2018). In lieu of these specific markers of victim blaming language which can further lead to biased reporting in offline media as well, we feel it is imperative to study about it in detail. Previous works in hate speech detection classified tweets as racist or sexist only (Badjatiya et al., 2017). These works generalize all instances of sexism under one classification. However, recent research has shown important sub-classifications of sexism that may be important for online media research. (Parikh et al., 2019) classifies tweets into 14 sub categories of sexism and we identify that not all 14 categories may have as drastic effects as victim blaming and slut shaming. Victim Blaming on online media directly leads to psychological disturbances for the victim and biased responses from authorities seeking legal action for such crimes (Gruenewald et al., 2004). This does not undermine the severity of other categories but rather establishes why victim blaming

[2]https://www.nytimes.com/2019/12/13/business/liu-jingyao-interview-richard-liu.html

Clothing, makeup of victim (short skirt, skimpy clothes, v-neck shirts)
Victim's physiological state at time of incident (drunk, sad, depressed)
Victim's former/current job as a prostitute
Victim's sexual history or promiscuity
Victim's upbringing as explanation for behaviour (raised by lesbians, rich)
Locations that suggest victim culpability (bar, road, club)
Use of loaded terms to describe rape (alleges, accuses, she says)
self-reporting (Victim chooses to report crime on online media first)

Table 1: Coding Instrument to Identify Victim Blaming Language

should be studied separately and not as a specific instance of hate speech due to sensitivity and specific topics related to this issue. Models specific to each sub-category may not seem feasible and scalable but viewing at the issue of victim blaming from a psychological perspective we feel research in this domain is essential to devise computational frameworks to identify and prevent victim blaming on social media extensible to offline media reporting.

3 Dataset Construction

Creating our victim-blaming dataset entails a two step process: collection of data and data annotation. Due to the lack of prior work, we create a custom dataset by crawling English tweets from Twitter using the Twitter API[3] that mentioned major hashtags related to sexual harassment. A total of 4,242 tweets were scrapped from November 6, 2019 to November 19, 2019 which contained 'metoo. Also, 413 tweets were scrapped containing 'sexualharassment. We further observed that victim blaming tweets contained derogatory terms like 'whore' therefore, we used common words related to sexual harassment (rape, slut, whore) for increasing the number of positive samples. A total of 732 tweets were added after scrapping tweets from November 16, 2019 to November 29, 2019. Tweets are unstructured and noisy in nature due to the use of informal language prevalent on social media (Zappavigna, 2012). For pre-processing them, stop words are removed along with non alphanumeric text as they provide little information. URL's, emojis like ':)' and the symbol '#' are removed using regex. Usernames are also removed for anonymity. The tweets are then tokenized to run the various models. Tokenization and removal of stop words is done using NLTK. The average number of characters in a tweet is 33 [4]. After pre-processing, we remove tweets which have less than 33 characters. The final dataset contains 5,070 tweets out of which 1562 were classified as positive samples. Under the guidance of a social scientist, a coding scheme to identify victim blaming language is devised, taking into account gender related discourse and campaigns as well as psychological analysis of rape culture and victimization from social media. Table 1 summarizes the key identifiers of victim-blaming language.

We follow a two-phase annotation process for the classification of each tweet in the final dataset. In Phase 1, two annotators, a psychology student and a social science student identify tweets which contain features listed in Table 1 and subsequently classify them by marking victim blaming tweets as '1' and other tweets as '0'. The inter annotator agreement, measured by the average of the Cohens Kappa (Cohen, 1960) is 0.712. In case of a disagreement between the annotators, a third annotator who is a social science student classifies the ambiguous tweet. Phase 2 involves checking that all the tweets have been classified and correct symbol that is, either '1' or '0' have been used for it. Table 2 shows some samples from the dataset that were labeled '1' by the annotators.

4 The Victim Blaming Language Model

NLP problems are solved using shallow machine learning models or deep learning models along with manually crafted features which fail to capture the linguistic information properly due to dimensionality problems (Khatua et al., 2018). Twitter follows a unique semantic and linguistic format (Hu et al., 2013) that makes it imperative to devise frameworks that incorporate the specific linguistic

[3] https://developer.twitter.com/

[4] https://techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/

Tweets	Label
She went with them and got drunk. She slept with them and when they called her a whore accused them of rape	1
She is a dumb British whore who filed fake claim	1
Harvey Weinstein is a very evil man, but some of the celebrity women apart of the #Metoo movement that were "sexually assaulted" knew exactly what they were doing.	1
Another perverted Democratic! Let's see if Hollywood will say something or if #MeToo is only for Republicans.	0
They definitely raped her. Why else would they call her a whore?	0
Women don't admit they lost their power to men via rape for the fun of it. If a woman says she was raped, believe her.	0

Table 2: Examples from Annotated Dataset(Paraphrased to preserve anonymity)

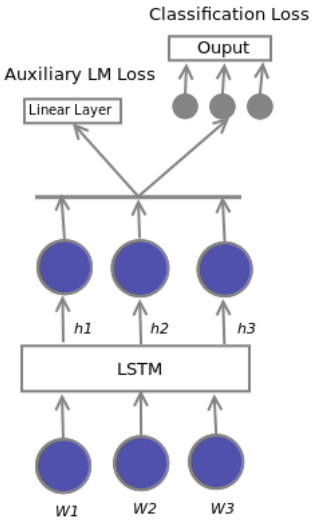


Figure 1: High Level Overview of Victim Blaming Classification Method. Blue indicates transferred layers and grey indicates randomly intialized layers

styles used on Twitter. Recent advancements in using transfer learning for tweet stance classification shows that enriching models with Twitter linguistics can improve performance (Zarrella and Marsh, 2016) Additionally, victim blaming language exhibits specific topics and syntax as shown in the coding instrument in Table 1. However, popular text classification models have failed to incorporate the subtle nuances of victim blaming language on social media specifically Twitter. Therefore, we propose a transfer learning based model that addresses this issue.

We propose a simple yet effective classification method based on single step transfer learning (Chronopoulou et al., 2019). State-of-the-art transfer learning methods employ language models (LM) trained on generic corpora with additional fine tuning of LMs for task specification. In our method, we combine the task-specification and language modelling with the help of an auxiliary loss function that is adjusted during training for task and linguistic adaptation. This prevents catastrophic forgetting and allows our model to be trained on a social media specific large corpora for e.g.: Twitter or Reddit standard datasets and then be adapted to target tasks on domain specific smaller corpora. (Chronopoulou et al., 2019) theorizes that the prevention of catastrophic forgetting for machine translation tasks is because the language model objective acts as the regularizer that limits the loss of generalizable features and evidence for the same is presented in their work. We intuitively inferred that a similar approach would be effective in capturing the subtle topics of victim blaming on Twitter due to the additional language modelling step that guides the training across the text classification task.

LM Pretraining: We train a word-level LM which consists of an embedding LSTM layer, 2 hidden LSTM layers and a linear layer.

Transfer learning and Auxiliary Loss: We transfer the weights of the pre-trained model and add an additional LSTM layer.

We introduce an auxiliary LM loss during training to incorporate the contribution of the pre-trained language model in the classification method. The joint loss is the sum of classification loss, L_{CLF} and auxiliary LM loss, L_{LM}.

$$L = L_{CLF} + L_{LM}$$

We consider equal contribution of both the loss values to effectively capture language modelling information and classification information specific

Parameter	Value
Activation function	ReLU
Dropout	0.4
Batch size	64
Epochs	20
Optimizer	Adam
Learning rate	0.0005

Table 3: Parameters for CNN Architectures

to the nature of our dataset. A High level overview of our method is shown in Figure 1.

5 Experiments

5.1 Baselines

Traditional Machine Learning (ML) Approaches
We have used two machine learning models, Support Vector Machine (**SVM**) and Naive Bayes (**NB**).

SVM: For feature extraction, TF-IDF has been used on word unigrams that is fed to the SVM

NB: Similar to SVM, TF-IDF has been used for feature extraction for classification

LSTM-Based Architectures

LSTM: The word embeddings for all the words in a post are fed to a vanilla LSTM

TextbiRNN: This is an improvement on a vanilla RNN. The word embeddings for all the words in a post are fed to a bi-directional LSTM.

CNN-Based Architectures

TextCNN: Convolutional filters we applied to the word vectors of a post followed by max-pooling layers as described by (Kim, 2014)

CharCNN: A sequence of encoded characters are fed into a CNN as described by (Zhang et al., 2015)

fasttext: fasttext classifier is used for text classification which takes into account n-grams of words to incorporate local word order (Grave et al., 2016)

5.2 Implementation of Victim-Blaming Classification Method

To pretrain the language model we create a dataset[5] of 1 million English tweets scraped from Twitter, including approximately 1M unique tokens. We use 50K most frequent tokens as our vocabulary. We then use our Victim-Blaming dataset for classification. To pre-process the tweets we use regex to remove usernames, urls and emojis. In addition to this, we use NLTK for stop word removal and tokenization of the tweets. For neural models, we use an LM with embedding size of 300, 2 hidden layers, dropout of 0.3 and batch size of 64. We add an LSTM of size 100 with a softmax classification layer on top of the transferred LM. In pretraining, and pretrained layers (of transferred model), Adam was used with a learning rate of 0.0001. For the newly added LSTM and classification layers, Adam with learning rate of 0.0005 was used. For developing our models, we use Pytorch and Sci-kit learn.

[5]https://www.kaggle.com/paoloripamonti/twitter-sentiment-analysis

5.3 Results

Table 4 describes the performance of the baseline models in comparison to our proposed approach across the accuracy metric. The models were trained over 60% of the dataset while 20% was held out for test and 20% was used as dev split to optimize the parameters across all the models tested. The proposed approach outperforms all baselines including RNNs, CNNs, LSTMs and traditional ML approaches SVM and NB. Fasttext model is able to generate domain specific embeddings due to the nature of embedding construction that benefits the unpredictable and unstructured Twitter semantics. CharCNNs usually have a high perplexity due to the character-by-character prediction, however, they presented similar results to the fasttext model which are better than the other baseline models. Our method shows better results when compared with the baseline models. Since we do not have a lot of data, the baseline models fail to identify linguistic features of twitter language which are significantly different from normal conversational language. Our method takes care of this using the auxiliary loss function. The language model trained on generic corpora is successfully able to capture these features and can therefore perform better when retrained for classification. In comparison to our baselines, our model architecture is simpler and computationally inexpensive.

5.4 Error Analysis

It has been observed that sometimes incidents of victim blaming are either self reported or reported by a third person. Some tweets may cite previous instances of victim blaming to speak against victim blaming. Since, these tweets consist of the marked topics and linguistic framing encoded in the coding instrument in Table 1, the proposed model classifies

Approach	Accuracy
NB	0.60
SVM	0.73
LSTM	0.74
TextbiRNN	0.74
TextCNN	0.75
CharCNN	0.77
fasttext	0.78
Proposed Method	**0.82**

Table 4: Performance Comparisons on Victim Blaming Dataset

such tweets as positive examples. This is a typical form of error encountered in even hate speech detection tasks (MacAvaney et al., 2019) where keywords marked for positive examples leads to classification errors. It should further be discussed whether such examples should be classifies as positive or negative during the annotation process and requires extensive social and psychological research. Some systemic errors we explored during our experiments:

- *She may tweet against you in #MeToo if you are not careful* : This tweet was annotated as 1, that is, victim blaming tweet since it contains implicit victim blaming. The proposed model wrongly classifies it as 0 as it lacks specific topics and keywords that the model has learned during training. This error arises due to the model failing to effectively capture sarcasm.

- *All the desi feminists....using woman card for personal gains and abusing #Metoo.*: This tweet is not classified as victim blaming by the model, however, it is annotated as victim blaming due to implicit gaslighting of victims who choose to report it. It was mis-classified since it is not directly threatening or accusing a victim. The researchers also feel this might be an oversight between the annotators while data annotation as this may be a case of general sexism and not victim blaming. To address this type of error we plan to extend our work into a multi-label categorization task which considers sub-categories of victimization, that is, secondary, primary and gender based victimization in rape cases.

- *Even if I agree, most of what it would take for that to be a valid viewpoint, you still mean that "you canŕ rape a whore." Justice should be principle based not tribe based.*: This tweet is not victim blaming but citing an instance of victim blaming directly which leads the model to classify it as victim. This is an instance which is very common in hate speech tasks as well where citing or using such phrase and words to talk against the hate or victim blaming language leads to false positives during classification.

- *I'll be a good boy and take it silently if you rape my cunt*: This tweet contains vulgar language that is identified by the model as victim blaming erroneously. This tweet is annotated as 0, not victim blaming but the specific words like 'cunt' or 'take it' has clearly confused the model as it is failing to capture long sequences here and decipher the meaning of the tweet wholly.

6 Conclusion and Future Work

In this work, we established the need to devise a computationally effective method to identify victim blaming language on Twitter. To achieve this, we proposed a single step transfer learning based classification method that effectively captures the unique linguistic structures of twitter data and victim blaming language. On a manually annotated dataset, our proposed approach could achieve significant improvement over existing methods that rely on custom textual features and popular deep learning based methods. The prevalence of rape culture and the subsequent victim blaming on unsolicited social media forums like Twitter has not been studied from a computational linguistic perspective before. Our work, therefore presents an extensive study of popular text classification methods on a niche' dataset with victim blaming semantics and further presents the significance of using a simple transfer learning approach to capture Twitter semantics on a limited dataset. We anticipate that this study encourages further research on how victims of sexual assault are portrayed on social media. Our future agenda includes further bifurcating and exploring the specific types of victim blaming and the efficacy of the proposed approach on such a multi label classification task. We plan to explore the different weighting factors for the language modelling loss and classification loss described in section 4 to determine if weighting factors can help customize the auxiliary loss for different tasks.

References

Courtney E. Ahrens. 2006. Being silenced: The impact of negative social reactions on the disclosure of rape. In *American Journal of Community Psychology*, pages 23(2), 31–34.

Aslam Andleeb, Qamar Usman, Ayesha Khan Reda, Pakizah Saqib, Aleema Aleena Ahmad, and Aiman Qadeer. 2019. Opinion mining using live twitter data. In *IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*.

Bernice Andrew, Chris R. Brewin, and Suzanna Rose. 2003. Gender, social support, and ptsd in victims of violent crime. In *Journal of Traumatic Stress*, pages (4), 16, 421–427.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, page 759760.

Vimala Balakrishnan, Shahzaib Khan, and Hamid R. Arabnia. 2020. Improving cyberbullying detection using twitter users psychological features and machine learning. In *Computers Security*, page Volume 90.

Saroj Kumar Biswas. 2019. Keyword extraction from tweets using weighted graph. In *Advances in Intelligent Systems and Computing*.

Rebecca Campbell, Emily Dworkin, and Giannina Cabral. 2009. An ecological model of the impact of sexual assault on women's mental health. In *Trauma Violence Abuse*, pages 10(3), 225–46.

Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An embarrassingly simple approach for transfer learning from pretrained language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2089–2095, Minneapolis, Minnesota. Association for Computational Linguistics.

Linda Coates, Cathy Richardson, and Allan Wade. 2006. Reshaping responses to victims of violent crime. In *Cowichan Bay, B.C., Canada*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Edouard Grave, Tomas Mikolov, Armand Joulin, and Piotr Bojanowski. 2016. Bag of tricks for efficient text classification. In *EACL*.

Tara L Gruenewald, Margaret E Kemeny, Najib Aziz, and John L Fahey. 2004. Acute threat to the social self: Shame, social self-esteem, and cortisol activity. *Psychosomatic medicine*, 66(6):915–924.

Y. Hu, K. Talamadupula, and S. Kambhampati. 2013. Dude, srsly?: The surprisingly formal nature of twitter's language. *Proceedings of the 7th International Conference on Weblogs and Social Media, ICWSM 2013*, pages 244–253.

Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of silence breakers: Exploring sexual violence on twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 397–400.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

M. P Koss. 1993. Rape: Scope, impact, interventions, and public policy responses. In *American Psychologist*, pages 48(10), 10621069. American Psychological Association.

Melvin J. Lerner. 1980. The belief in a just world: A fundamental delusion.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8).

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. *arXiv preprint arXiv:1910.04602*.

Jessica Ringrose and Emma Renold. 2012. Slut-shaming, girl power and sexualisation: Thinking through the politics of the international slutwalks with teen girls. In *Gender and Education*.

Ramit Sawhney, Prachi Manchanda, Raj Singh, and Swati Aggarwal. 2018. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, Melbourne, Australia. Association for Computational Linguistics.

Nicolas J. Schrading. 2015. Analyzing domestic abuse using natural language processing on social media data.

Megan Stubbs-Richardson, Nicole E Rader, and Arthur G Cosby. 2018. Tweeting rape culture: Examining portrayals of victim blaming in discussions of sexual assault cases on twitter. *Feminism & Psychology*, 28(1):90–108.

S.E Ullman. 2000. Psychometric characteristics of the social reactions questionnaire. In *Psychology of Women Quarterly*.

Simon N. Verdun-Jones and Katherine Rossiter. 2010. The psychological impact of victimization: Mental health outcomes and psychological, legal, and restorative interventions.

Christine Wekerle, Negar Vakili, Sherry Stewart, and Tara Black. 2018. The utility of twitter as a toolfor increasing reach of research on sexual violence. In *Child abuse neglect*, page 85.

Michele Zappavigna. 2012. *Discourse of Twitter and social media: How we use language to create affiliation on the web*, volume 6. A&C Black.

Guido Zarrella and Amy Marsh. 2016. Mitre at semeval-2016 task 6: Transfer learning for stance detection. *arXiv preprint arXiv:1606.03784*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc.

The Visual Computer
https://doi.org/10.1007/s00371-020-01873-x

ORIGINAL ARTICLE

Novel approach for multimodal feature fusion to generate cancelable biometric

Keshav Gupta[1] · **Gurjit Singh Walia**[2] · **Kapil Sharma**[3]

© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract
Biometric systems provide various benefits over traditional pin-based authentication systems. However, the issue of data privacy and theft is of great concern. To resolve these issues, a novel cancelable multimodal biometric system is proposed that combines multiple traits by means of a projection-based approach. The proposed approach generates a cancelable biometric feature that is used to obtain revocable and noninvertible templates. Cancelable features are generated by projecting the feature points onto a random plane obtained using a user-specific key. The point of projection is then transformed into cylindrical coordinates and a combined cancelable feature is obtained. Extensive experiments are performed over 3 chimeric multimodal databases and results reveal high performance. The average DI and EER achieved by the proposed method are 16.63 and 0.004, respectively. Also, the proposed method is successfully analyzed for privacy concerns, namely revocability, non-invertibility, and unlinkability. Moreover, the proposed system demonstrated tolerance against various security attacks like brute force attacks, attacks via record multiplicity, and substitution attacks.

Keywords Cancelable template · Feature fusion · Multimodal biometric

1 Introduction

Biometric authentication systems are gaining high popularity and public acceptance due to a wide variety of uses. It includes identification, authentication, forensics, and access control. With the increase in usage, the number of threats and attacks like spoofing attacks, template thefts, etc. are also increasing. Thus, securing the biometric system itself is a challenge as biometric data for an individual is irreplaceable. If it is compromised, then it cannot be changed or replaced making its protection a very crucial task.

✉ Kapil Sharma
kapil@ieee.org

Keshav Gupta
keshavgupta101@gmail.com

Gurjit Singh Walia
gurjit.walia@gmail.com

[1] Department of Computer Science, Delhi Technological University, New Delhi, India

[2] SAG, Defence Research and Development Organization, New Delhi, India

[3] Department of Information Technology, Delhi Technological University, New Delhi, India

To overcome problems such as spoof attacks, poor input data quality, non-universality, etc. multimodal biometric systems came into existence [1]. Here, information from complementary and fairly independent modalities was fused to reach a final decision [2]. This fusion process may take place at the feature level, score level, decision level, sensor level, and rank level. In feature level fusion, extracted features from various biometric modalities are fused to generate a more discriminative feature. On the other hand, matching score values from multiple classifiers is combined in score level fusion [3]. At decision level fusion, results from different classifiers are combined together using methods like majority voting [2, 4]. Fusion at feature level generally provides better results as compared to other fusion techniques since the information available about biometric data is very high at this level [4, 5]. But to achieve fusion at feature level is strenuous as in many cases the features are not compatible. Generally, multimodal systems are very helpful in dealing with above-stated issues but no protection is available for biometric data itself in case of template hacking or database stealing. The security of biometric data should be of top priority as biometric information cannot be replaced easily. If biometric templates are stolen, the user's identity

Published online: 13 June 2020

is compromised for multiple applications and subject to cross-application attacks as well. Overall, it not only threatens the security, but it may also incur a significant financial or social loss. To resolve these issues, cancelable biometrics is widely used.

Cancelable biometric systems use a pseudo-biometric template instead of the original template for matching and verification purposes. These pseudo-templates are generated from original templates using various transformation mechanisms. For instance, in non-invertible geometric transformations, a feature domain transformation is applied using transformation techniques such as Cartesian, polar and functional transformation [6]. Random projection is also used as a non-invertible transformation [7, 8] wherein an extracted feature $x \in F_n$ is projected to a random subspace $Y \in F_n \times N$ with $n < N$ and all element of y are independently realized from a random variable as $z = Yx$ where z is random projection vector. A random convolution method was proposed in [9] to produce cancelable templates wherein a user-specific kernel is used to encrypt the biometric data. Bio-convolving is also a convolution-based approach for generating cancelable templates [10]. In this, biometric template was segmented into various sequences and a transformed sequence is generated using linear convolution. An extension of random projection is bio-hashing [11] wherein a bio-hash template was generated using a user-specific random number. Random permutation of features is also been used by many researchers for generating cancelable templates [12]. There are also various salting methods where a random pattern or noise is mixed with the original template. The techniques used for the cancelable biometric template must exhibit properties, viz. revocability, security, diversity, and accuracy. Recently, researchers are also working at hybrid methods where more than one technique is used to generate cancelable templates.

In the proposed work, a multimodal biometric system is introduced using two complementary biometric traits, viz. fingerprint and iris where complementary information is fused using feature-level fusion. The proposed system is highly secure as well as tolerant of template hacking by means of cancelable templates. Cancelable templates are generated using a novel transformation technique where the feature vectors from two modalities are used to generate abscissa and ordinate of a feature point. These feature points are orthogonally projected on a random plane generated using a user-specific key. The projected point is then transformed into a cylindrical coordinate system to receive θ, ρ, z values. In the next step, θ values for every feature point are combined together to generate a transformed template which is irreversible in nature. The proposed method is faster, cheaper and provides high performance. Also, the cancelable templates generated are non-invertible, revocable and diverse in nature. The proposed method is evaluated on multimodal chimeric datasets obtained from benchmarked images. A brief summary of the proposed work is discussed as follows:

- A novel multimodal biometric system using two complementary features, namely fingerprint and iris is proposed. A cancelable biometric template feature is generated by combining the information from these traits.
- Cancelable biometric templates are generated from the feature vectors of the modalities using a user-specific random key set. Here, the feature points from the corresponding feature vectors used as abscissa and ordinate are projected on a random plane that is generated using a user-specific random key set.
- The point of projection on a random plane is transformed into a cylindrical coordinate system to generate r, θ, z values. A different plane is generated for each feature point to generate unique points of projection. From every point, θ values are combined together to generate cancelable template.
- The proposed method is evaluated over three multimodal chimeric datasets created using benchmarked images. The results reveals high performance, high privacy, high security and low error rate compared to state-of-the-art techniques.

The rest of the manuscript is arranged in the following manner: In Sect. 2, various state-of-the-art methods for generating cancelable templates and feature-fusion techniques are discussed and compared. In Sect. 3, details of the proposed biometric system are discussed at length. Section 4 contains the experimental validation, database design and performance analysis of the proposed method. Finally, Sect. 5 contains the conclusion and future scope of the work done.

2 Related work

Cancelable biometrics is used to protect templates by performing the matching and storage in a different domain [13]. Generally, in the case of template stealing, the stolen template is revoked and an entirely different template is created by altering the key. Thus, it fulfills the necessary requirements for template protections, namely revocability, non-invertibility, diversity and accuracy [13, 14]. Transformation performed on biometric data can be mainly classified as non-invertible transforms and biometric salting. Biometric salting can be additionally classified as projection-based transformations, noise-based transformation, and convolution-based transformation.

Transformations based on random projection projects biometric data to a random sub-space using various transformation techniques. The most popular technique in this category is bio-hashing [11] which projects the features into

the orthonormal sub-space. It provides good discrimination capability and high performance. However, this method suffers from the problem of irreversibility if both template and transformation matrix are compromised. In [15], authors proposed an extension of the bio-hash technique to address the issue of stolen-token scenarios. In this, the authors used a novel multi-state discretization technique instead of a simple threshold scheme. Pillai et al. proposed a new variant of random projection, namely Sectored random projection [7] technique where the biometric feature is segmented into various sectors, then on each sector, the random projection is applied and concatenation was performed to generate the cancelable template. This method not only caters to the issue of useful iris area reduction, but it is also robust to common iris outliers. To improve non-invertibility, Teoh et al. [16] introduced a novel technique, random multi-space quantization (RMQ), wherein the biometric feature vector was mapped with a sequence of random sub-spaces using a pseudo-random sequence. In the second step, quantization was performed based on a threshold value. Teoh et al. also proposed a multi-space random projections technique [17] which is a two-factor cancelable formulation and feature vector obtained from biometric modalities is projected to multiple random subspaces based on a user-specific pseudo-random number. Wang et al. proposed a random projection with vector translation method [18] for cancelable biometrics wherein biometric data was projected using a Gaussian random variable. Also, Paul et al. [19] proposed a random cross-folding method using random projection and selection to create cancelable templates. In order to make it a user-dependent dynamic process, Yang et al. proposed a nonlinear projection process in which the projection vector was dynamically decided by feature vector itself [20]. In general, most of the projection techniques discussed are vulnerable to attacks such as inverse operations if both templates and transformation matrix are stolen [21]. Accordingly, researchers have proposed other methods like random convolution transformations for generating cancelable templates.

In random convolution (RC) based transformations, cancelable templates are generated by convolving the biometric feature using a random kernel. Savvides et al. [9] used minimum average correlation energy (MACE) filters as a convolution method to generate cancelable templates. In [22], authors used the concept of locality sensitive hashing to generate secure, cancelable iris features. Maiorana et al. proposed a bio-convolving technique where a set of non-invertible transformations were performed on sequence-based biometric representation [10]. But if kernel used for transformation is known, it can be vulnerable to inverse attacks. To overcome this issue, an algorithm using curtailed circular convolution was proposed [23]. The algorithm convolved input binary features in a circular manner using random binary strings imparting non-invertibility. In [24], the authors used the quality of fingerprint features to determine the presence of live fingerprint. Similarly, Ali et al. [25] used a user key set to modify minutiae information from fingerprint to produce cancelable templates. Further, Trivedi et al. [26] used a binary user key on fingerprint modality to generate cancelable templates. On the other hand, Wu et al. [27] presented ECG as a biometric and generated revocable templates using signal subspace collapsing. The cancelable biometrics is also widely used with multimodal biometric systems to remove various limitations imposed by unimodal systems.

Rathgeb and Busch used adaptive bloom filters to transform iris feature for both eyes of a single subject and fused them at feature level [28]. This method provided improved performance but vulnerable to template linking. A random permutation principal component analysis (RP-PCA) method was introduced by Kumar et al. [29] to generate cancelable biometric using face, iris, and ear modality. The accuracy of the system was unaffected, and the robustness of the system was improved. Also, Dwivedi and Dey [30] created a hybrid scheme for a cancelable multi-biometric system combining mean-closure weighting (MCW) with Dempster-Shafer (DS) theory. This scheme showed robustness against score variability and considerable performance improvement over uni-modal counterparts. Similarly, Walia et al. [31] proposed a cancelable biometric system by performing cross-diffusion of graphs. Later, PCR-6 was used to fuse belief masses from individual classifiers. Experimental results demonstrated better performance than many existing techniques. Kaur and Khanna [32] proposed a novel random distance technique for transformation in a multi-biometric scenario using face, palmprint, and finger-vein as input modalities. Similarly, walia et al. [33] used key images to generate cancelable features and dimension reduction. A multifold random projection was introduced by Paul and Gavrilova [34] for a multimodal biometric system with improved recognition performance. Chin et al. [35] proposed a template protection scheme wherein original fingerprint and palmprint templates were arranged in random rectangles using user-specific keys. Later, statistical features were extracted and fused at the feature level to generate cancelable templates. Similarly, Gomez-Barrero et al. [36] used bloom filters on face-finger vein and face-iris to generate protected templates and a weighted feature level fusion was performed to generate the multimodal cancelable template. With increasing attacks on biometric systems, detection of fake biometric is equally important as the protection of biometric data.

Recently, deep learning methods are widely used for feature representation. For instance, Essam et al. [37] identified various regions in face image using multiple CNNs. Later, bio-convolving encryption was used to generate cancelable templates. Also, Das et al. [38] proposed a finger-vein

identification method based on deep learning. Similarly, in [39], a novel architecture ScoreNet was introduced for unconstrained ear recognition wherein deep learning and hand-crafted methods were used together. Further, Liu et al. [40] presented a detailed discussion about various deep learning methods for feature representation.

Based on the above discussion, it is evident that securing biometric systems from various attacks should be of top priority. Accordingly, a multimodal biometric system with cancelable templates and liveness detection is proposed in this manuscript which complements the advantages of various techniques and improves the overall performance of the biometric system. A detailed explanation of the proposed methodology is provided in the next section.

3 Proposed multimodal cancelable biometric system

Figure 1 represents the architecture of the proposed system. In this, a cancelable template is generated for Iris(i) and Fingerprint(p) features using the proposed technique with user-specific keys.

The proposed cancelable biometric system takes two biometric modalities, viz. iris and fingerprint and feature vectors are extracted. For extracting features from iris modality, image pre-processing combined with local binary pattern (LBP) [41] is performed [42]. LBP not only provides low computation but also immune to changes in image grey levels. For fingerprint modality, the input query image is preprocessed by performing binarization and thinning operations. Next, feature extraction is performed to generate minutiae-based features from the input image.

Feature points at *i*th position of the iris feature vector is considered as abscissa and fingerprint feature as ordinate. Combining them together, a point (χ_i, ψ_i) is defined in a cartesian coordinate system. Each point thus obtained is projected onto a plane corresponding to each feature point obtained using a user-specific key. The points of projection thus obtained are transformed into a cylindrical space to obtain corresponding feature points. The azimuth values are considered for generating cancelable templates so as to achieve non-invertibility. A similarity score is calculated by matching the generated feature with the stored templates in the database. Finally, the match score is compared with an optimal threshold value to reach a final decision. The proposed system is non-invertible in nature as only the azimuth values are considered for generating the cancelable templates. The in-depth details of the proposed system are presented in the next subsection.

3.1 Multimodal feature extraction

Biometric modalities, viz. iris and fingerprint are processed using feature extraction techniques to determine individual feature vectors. For fingerprint features are extracted using a minutiae-based technique which is widely used by researchers [43, 44] as it provides low complexity with high performance. For this, first of all, binarization and thinning operations are performed as a pre-processing step on input query

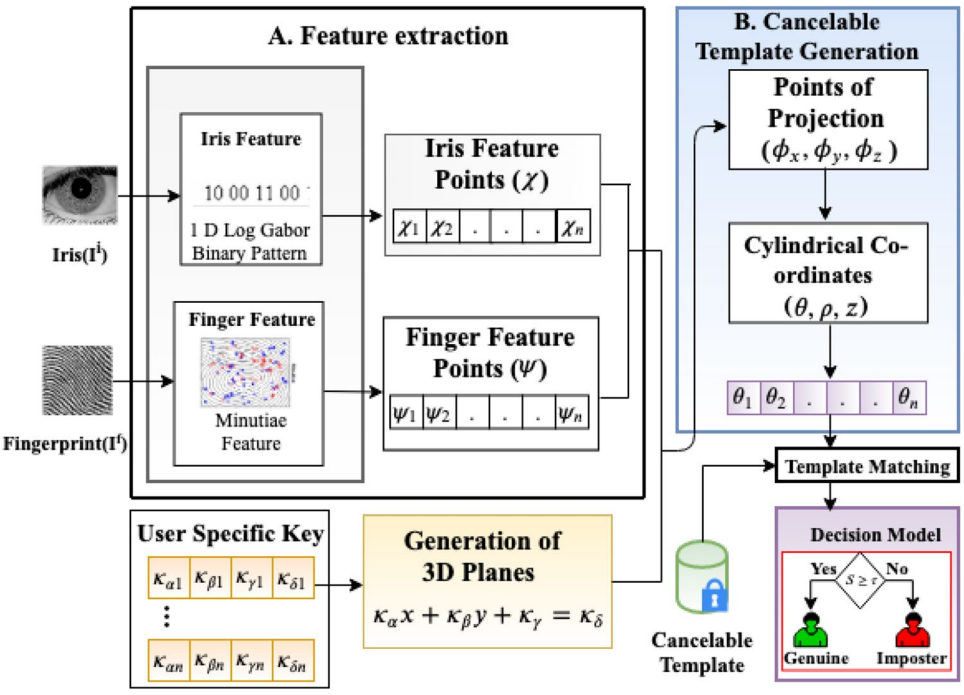


Fig. 1 Overview of the proposed cancelable template generation scheme. Extracted features from the input query image are projected onto a plane obtained using a user-specific key. The points of projection are transformed into cylindrical coordinates to generate cancelable templates. These are compared with stored templates to evaluate similarity score to reach a final decision

fingerprint image I^f. Binarization operation helps in increasing the contrast between ridges and valleys as shown in Eq. 1

$$B(x,y) = \left\{ \begin{array}{ll} 1, & \text{if } I(x,y) \geq h \\ 0, & \text{otherwise} \end{array} \right\} \tag{1}$$

where $I(x,y)$ shows intensity value at pixel position (m,n) and h represents the value of threshold. Also, a thinning operation is used to reduce ridges to the unit-pixel thickness and is carried out using in-built morphological functions in the MATLAB platform on binary images. Rutovitz crossing number (CN) is computed by locating the minutiae over the thinned image using a sliding window of size 3×3 in an anti-clockwise manner [44]. The CN defines the minutia type and is calculated using Eq. 2

$$\text{CN} = \frac{1}{2} \sum_{k=1}^{8} |p_k - p_{k-1}| \tag{2}$$

where p_k is the pixel values of immediate neighbors for pixel k. The value of CN is used to classify ridge pixel as isolated, continuing, ending, crossing point and bifurcation. Further, minutia is represented as a vector $m = [x, y, \text{CN}, \theta]$ having (x,y) as the pixel coordinates and θ as angle of orientation. For input fingerprint query image I^f, an extracted feature vector, η^f is created by combining n minutiae using Eq. 3

$$\eta^f = [m_1, m_2, \ldots m_n] \tag{3}$$

For iris feature extraction, input iris image I_i is pre-processed involving localization and normalization processes. In the first step, an integro-differential operator is used for localizing iris and pupillary boundaries. In the second step, Daugman's rubber sheet model [45] is used to normalize the localized iris into a fixed-sized rectangular block. Further, the processed image is quantified using the histogram of LBP. From LBP histogram values $l_1, l_2 \ldots l_n$, feature vector for iris η^i is generated using Eq. 4

$$\eta^i = (l_1, l_2 \ldots \ldots l_n) \tag{4}$$

This creates a unique pattern, generating iris feature η^i. The extracted feature vectors η^i and η^f are fused together using the proposed method to generate a cancelable feature which is discussed in the next subsection.

3.2 Proposed multimodal feature fusion

The fusion process is very important in a multimodal biometric system for making a decision. Here, we have proposed a feature level fusion method with template protection. The proposed method generates a cancelable template which is revocable, non-invertible and robust to various types of attacks such that true biometric feature will not be revealed to the attacker. Every user is provided with a complex unique key (κ^k), where $k \in [1, N]$ used to create random 3-D planes as shown in Eq. 5.

$$\kappa^k = \begin{bmatrix} \kappa^k_{\alpha 1} & \kappa^k_{\beta 1} & \kappa^k_{\gamma 1} & \kappa^k_{\delta 1} \\ \kappa^k_{\alpha 2} & \kappa^k_{\beta 2} & \kappa^k_{\gamma 2} & \kappa^k_{\delta 2} \\ . & . & . & . \\ \kappa^k_{\alpha n} & \kappa^k_{\beta n} & \kappa^k_{\gamma n} & \kappa^k_{\delta n} \end{bmatrix} \tag{5}$$

Here, k represents the kth user. The length of the key is equivalent to the length of feature vectors and consists of n rows containing 4 co-efficient values namely $\kappa^k_{\alpha i}$, $\kappa^k_{\beta i}$, $\kappa^k_{\gamma i}$, and $\kappa^k_{\delta i}$ where $i \in [1, n]$ and having values randomly distributed in the range $[-1000, 1000]$. The user key is used to generate a random plane defined using Eq. 6 for each feature point as shown above in Fig. 1

$$\kappa_{\alpha i} x + \kappa_{\beta i} y + \kappa_{\gamma i} z = \kappa_{\delta i} \tag{6}$$

where $i \in [1, n]$. The random planes generated using user-specific keys are used in combining multiple features using the proposed approach. In this, feature vectors obtained during the feature extraction process are fused together by means of the proposed projection-based approach. Each feature point of the iris feature vector is considered as abscissa and each feature point of the fingerprint feature vector is considered as ordinate in a cartesian coordinate system. In case of different feature size, padding can be used so that every feature vector contain an equal number of feature points. Moreover, the abscissa and ordinates at the corresponding positions are combined to describe a point(Q) (χ, ψ) such that any user can be defined as in Eq. 7.

$$\upsilon = \left[(\chi_1, \psi_1)\ (\chi_2, \psi_2)\ (\chi_3, \psi_3)\ .\ .\ (\chi_n, \psi_n) \right] \tag{7}$$

For simplicity, the above equation can also be represented as shown in Eq. 8

$$\upsilon = \left[Q_1\ Q_2\ Q_3\ .\ .\ Q_n \right] \tag{8}$$

Also, as discussed above, each user is provided with a unique user-specific key (κ^k) of dimension $4 \times n$. The key is used to create a different random plane corresponding to each point Q. Further, an orthogonal projection is performed from each of these points(Q) on the corresponding plane and point of projection(P) is obtained as shown in Fig. 2

For k^Pth user, each orthogonal projection from point $Q^k_i(\chi, \psi)$ on the random planes obtained using key κ^k generates a point $P^k_i(\phi_x, \phi_y, \phi_z)$ as in Eq. 9.

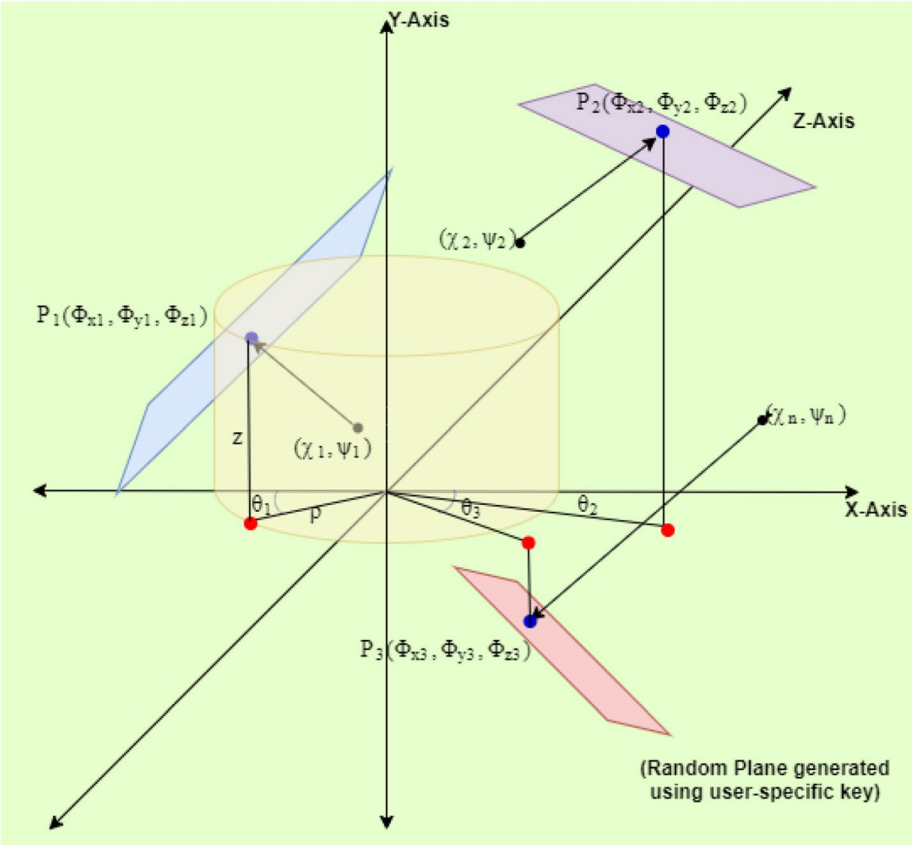


Fig. 2 Projection of feature points on random planes

$$\begin{bmatrix} P_1^k & P_2^k & . & . & P_n^k \end{bmatrix} = \begin{bmatrix} (Q_1^k & Q_2^k & . & . & Q_n^k \end{bmatrix} \begin{bmatrix} \kappa_{\alpha 1}^k & \kappa_{\beta 1}^k & \kappa_{\gamma 1}^k & \kappa_{\delta 1}^k \\ \kappa_{\alpha 2}^k & \kappa_{\beta 2}^k & \kappa_{\gamma 2}^k & \kappa_{\delta 2}^k \\ . & . & . & . \\ \kappa_{\alpha n}^k & \kappa_{\beta n}^k & \kappa_{\gamma n}^k & \kappa_{\delta n}^k \end{bmatrix} \tag{9}$$

In the next step, each point of projection (P_i^k) is transformed into cylindrical co-ordinates using a function f defined as in Eq. 10

$$P_i^k(\theta, \rho, z) = f(P_i^k(\phi_x, \phi_y, \phi_z)) \tag{10}$$

where ρ and θ are determined using Eqs. 11 and 12, respectively.

$$\rho = \sqrt{{\phi_x}^2 + {\phi_y}^2} \tag{11}$$

$$\theta = \tan^{-1} \frac{\phi_y}{\phi_x} \tag{12}$$

In the cylindrical coordinate system, point P is represented as θ, ρ, z. In the last step, the azimuth(θ) values corresponding to each point of projection(P_i^k) for kth user, are concatenated together to generate a fused vector ζ^k that is cancelable and non-invertible in nature as shown in Eq. 13.

$$\zeta^k = (\theta_1, \theta_2, \theta_3, \ldots \theta_n) \tag{13}$$

The purpose of conversion to a 3D cylindrical coordinate system is to provide a higher dropout ratio as compared to the 2D coordinate system. Since only azimuth values are used to generate the cancelable template, a dropout ratio of 66.6% is achieved. On the other hand, a 2D point of projection would have led to a dropout ratio of 50% only. Further, in case of template theft, if azimuth (θ) values are compromised, it will lead to ambiguous values of ϕ_x and ϕ_y as evident from equation 12 and original feature points will not be exposed. Thus, the generated template is highly non-invertible and robust against theft. Also, the feature vectors from both Iris and Fingerprint modality of dimension $1 \times n$ are combined and converted into a single vector of dimension $1 \times n$. The cancelable feature obtained is compared with

the stored templates to generate a final match score (S). The final decision is performed based on an optimal threshold value (τ), if the match score (S) is greater than τ, then it is considered as Genuine else imposter. The next section provides details of the experimental validation of the proposed method.

4 Experimental validation

The experimental validation of the proposed multimodal biometric system is performed over three multimodal chimeric databases. During privacy analysis, the proposed system is analyzed for unlinkability, non-invertibility, and revocability. Also, the system is analyzed for various attacks like record multiplicity, substitution, and brute force attacks to establish the robustness of the system. On the other hand, various performance metrics like equal error rate (EER), decidability index (DI), and recognition index (RI) are determined to estimate the system's performance. Also, the proposed system's performance is compared with other state-of-the-art methods.

4.1 Database and experimental design

The multimodal datasets are obtained using images from various benchmark datasets for experimental validation. Here, the Chimeric datasets are created by uniquely combining benchmark datasets, namely MCYT bimodal database [46], IITD PolyU iris database [47], Casia iris database (Casia-IrisV1, http://biometrics.idealtest.org/), FVC2006 DB1-A fingerprint database [48], and MMU2 iris database [49]. Sample images from the mentioned benchmarked datasets are shown in Fig. 3.

MCYT fingerprint database contains images captured using two different sensors, namely CMOS-based device and an optical capture device. For every finger, 12 samples are taken with both sensors from 330 subjects. The resolution of image samples is 300×300 and 256×400 for sensors 1 and 2, respectively. FVC2006 DB1-A is obtained using an electric Field sensor having an image resolution of 250 dpi. It contains 12 fingerprint samples from 140 subjects. All acquired samples are 256 gray-levels fingerprint images in BMP format. The IITD PolyU iris database contains 5 eye images of both eyes from 224 subjects. The images are captured using a digital CMOS camera of size 320×240 pixels. Casia IrisV1 database consists of a total of 756 iris images obtained from 108 subjects. All sample images have a resolution of 320×280 and are in BMP format. MMU2 Iris database contains 995 iris images from 100 volunteers having resolution 320×238 in BMP format.

The experimental validation is performed on three chimeric datasets namely D1, D2, and D3. D1 contains fingerprint samples of N different subjects from the MCYT database (sensor 1) and IITD iris PolyU database. Thus, a virtual multimodal dataset is created for N subjects by combining the above-mentioned datasets. Similarly, D2 is created by combining N distinct subjects from the MCYT database (sensor 2) and the Casia Iris-V1 database. Also, the D3 dataset is obtained by combining N distinct subjects from the MMU2 iris database and the FVC2006 DB1-A database. Further, all the subjects in D1, D2, and D3 databases are completely different. In addition, fivefold cross-validation is carried out to obtain balanced results. The proposed system

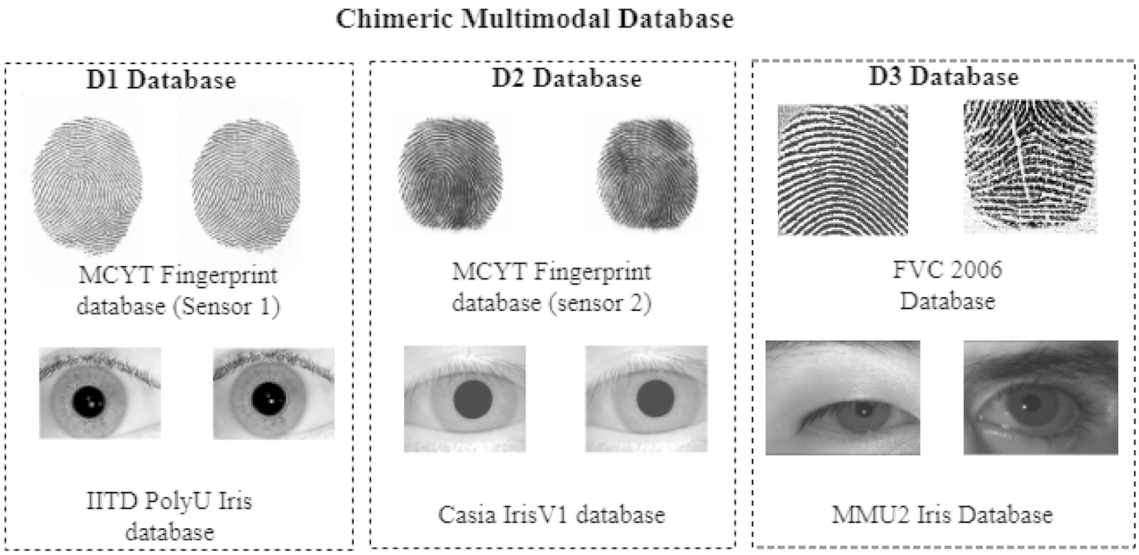


Fig. 3 Sample multimodal database images from the benchmark datasets

is implemented over a hardware configuration of the Intel i3 processor and 4GB RAM using the MATLAB 2018a platform. The performance analysis of the proposed system is discussed in the next section.

4.2 Performance analysis

The proposed system's performance is quantitatively analyzed by means of various performance metrics, namely decidability index (DI), equal error rate (EER), and recognition index (RI). Also, the results thus obtained are compared with other state-of-the-art techniques. During this process, the same techniques for generic feature extraction and score calculation are used for evaluating other methods.

4.2.1 Performance metrics

The difference between the imposter and genuine scores is quantified using decidability index (DI) and is calculated using Eq. 14

$$\mathrm{DI} = \frac{|\mu_g - \mu_i|}{\sqrt{(\sigma_g^2 - \sigma_i^2)/2}} \tag{14}$$

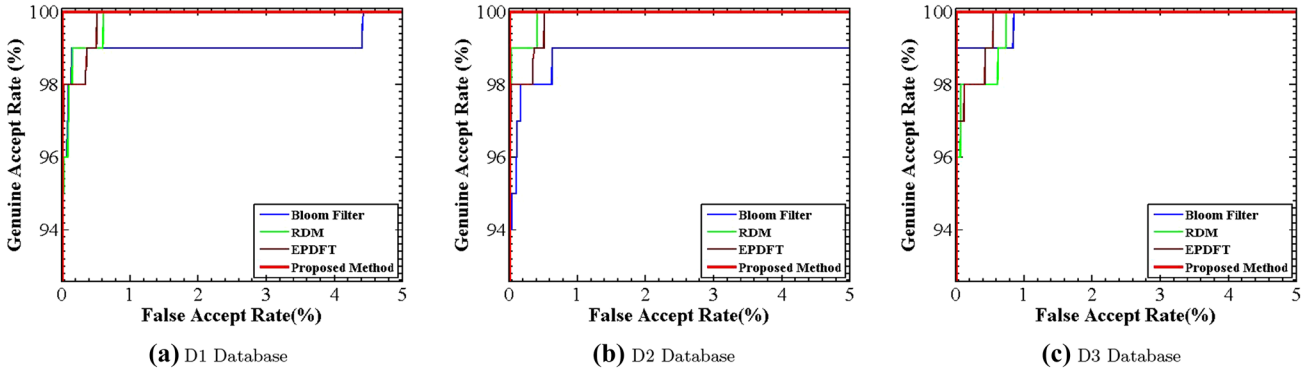
Here, μ_g, μ_i denotes the mean values, and σ_g and σ_i denotes the standard deviation values of score distributions for genuine and imposter, respectively. A high decidability index value suggests a higher ability of the classifier to separate genuine from imposters. Values of decidability index are shown in Table 1 for individual and proposed classifier over chimeric datasets D1, D2, and D3. Equal error rate (EER) is calculated by plotting ROC curves wherein the false acceptance rate (FAR) is plotted against the genuine acceptance rate (GAR). It provides the measure of the accuracy of the proposed biometric system. Recognition index (RI) provides the recognition rate at rank-1 and used for performance evaluation. Cumulative matching characteristics (CMC) curves show the relationship between rank and the recognition rate.

4.2.2 Accuracy analysis

The accuracy of the proposed system is analyzed and compared with state-of-the-art techniques using performance metrics, namely EER, DI, and RI. The performance metric

Table 1 Comparison of EER, DI, and RI values for D1, D2 and D3 databases

Method	DB								
	D1 Database			D2 Database			D3 Database		
	EER	DI	RI	EER	DI	RI	EER	DI	RI
RDM [32]	0.60 ± 0.23	7.28 ± 0.35	97 ± 0.25	0.40 ± 0.19	7.68 ± 0.37	98 ± 0.75	0.71 ± 0.21	7.11 ± 0.19	99 ± 0.50
Bloom filter [36]	0.97 ± 0.20	5.77 ± 0.42	96 ± 0.50	1.14 ± 0.24	4.98 ± 0.28	97 ± 0.45	0.92 ± 0.17	5.33 ± 0.31	98 ± 0.58
EPDFT [50]	0.49 ± 0.32	7.96 ± 0.42	98 ± 1.11	0.78 ± 0.24	6.21 ± 0.46	97 ± 0.90	0.61 ± 0.25	5.79 ± 0.18	97 ± 0.80
Proposed method	0.005 ± 0.004	22.14 ± 0.34	99 ± 0.94	0.003 ± 0.005	18.45 ± 0.43	99 ± 0.95	0.004 ± 0.10	9.71 ± 0.22	99 ± 0.90



(a) D1 Database **(b)** D2 Database **(c)** D3 Database

Fig. 4 Performance comparision of evaluated methods: ROC curves for D1, D2 and D3 database. **a** ROC curves for various cancelable biometric techniques over Database D1. **b** ROC curves for various cancelable biometric techniques over Database D2. **c** ROC curves for various cancelable biometric techniques over Database D3

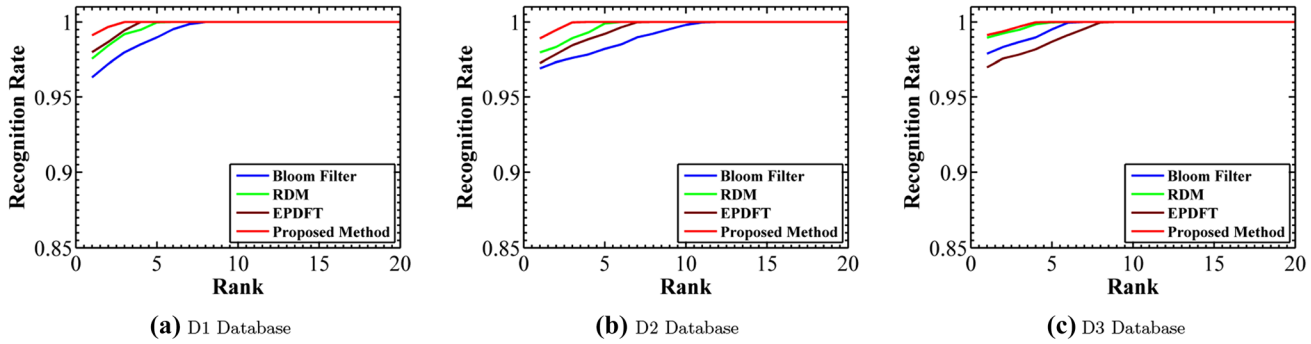


Fig. 5 CMC curves for D1, D2 and D3 databases. **a** Comparison of CMC curves of various cancelable biometric techniques over Database D1. **b** Comparison of CMC curves of various cancelable biometric techniques over Database D2. **c** Comparison of CMC curves of various cancelable biometric techniques over Database D3

values for various methods viz. Random distance method (RDM) [32], bloom filters [36] and enhanced partial discrete Fourier transform (EP-DFT) [50] are compared with the proposed method in Table 1.

The EER value for the proposed score level fusion method is 0.005 for the D1 database, 0.003 for the D2 database and 0.004 for the D3 database which is lowest in comparison with other state-of-the-art methods. The efficiency of the proposed system is also depicted by a high decidability value of 22.14 for the D1 database, 18.45 for the D2 database and 9.71 for the D3 database and supported by ROC and CMC curves in Figs. 4 and 5, respectively. The variation among EER values is due to the use of different datasets and fivefold cross-validation.

Accuracy analysis reveals that the limitations of individual classifiers were effectively addressed by the proposed fusion method providing higher accuracy and more reliable results. The proposed system also improves the privacy of every user making it robust against various issues like template thefts and safeguarding the identity of every user which is also discussed in the next subsection.

4.3 Privacy analysis

The performance analysis confirms the high accuracy of the proposed system by means of various performance metrics. The proposed biometric system also ensures user privacy by exhibiting the properties like non-invertibility, revocability, and unlinkability.

4.3.1 Non-invertibility

In order to fulfill the criteria of non-invertibility, it must be infeasible to create the original biometric traits even if the key and the transformed template both are compromised. In a situation when both user-specific key and transformed templates are stolen, non-invertibility in the proposed method is achieved by choosing only the azimuth(θ) values during the generation of cancelable templates. Since 67% of the information is discarded, it is not possible to trace back to the original points of projection. Even if points of projection are estimated, it is impossible to find the actual source of projection as there can be infinite points over that line connecting the feature point, Q and projected point on the plane,

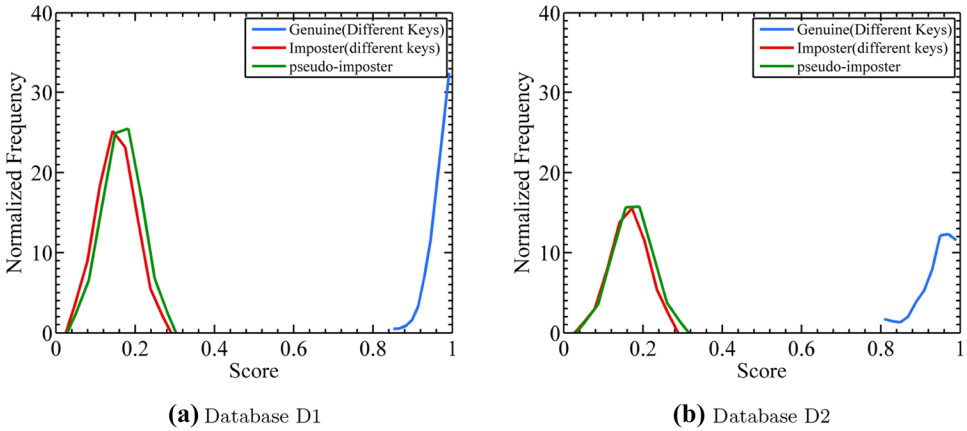


Fig. 6 Comparision of genuine, imposter and pseudo-imposter distribution for **a** D1 Database and **b** D2 Database

P. Thus, the proposed method exhibits non-invertibility of biometric templates.

4.3.2 Revocability

For a cancelable biometric system, if the stored templates are stolen, then they are discarded and new templates are generated using a new set of keys. Revocability states that templates generated from the same features should not be correlated. The revocability test is performed to measure the difference between the newly generated template and the old template. In order to check the revocability of the proposed system, 100 different keys were used to obtain 100 transformed templates. Every new user-specific key generates a different random plane and hence different templates. The distribution of imposter vs pseudo-imposter is shown in Fig. 6 which shows that the pseudo-imposter distribution is very similar to imposter distribution. It shows that there is no correlation between the old and new transformed templates. This suggests that transformed templates are treated as different individuals but they were created from the biometric features of the same subject. Thus, the revocability analysis shows that the stolen template can be easily replaced by a new template by using a different set of keys in the proposed system.

4.3.3 Unlinkability

Unlinkability states that multiple biometric templates of a single subject must be unlinkable provided different keys are used. This secures the identity of the subject when it is enrolled in multiple applications. In order to analyze unlinkability, the procedure described in [51] is adopted. For this, pseudo-genuine scores are introduced which refers to the match scores between the different templates of the same user by using different user-specific keys. Also, the pseudo-imposter scores are also calculated between different templates generated using a different user-specific key. In this scenario, if we plot the pseudo-imposter and pseudo-genuine distribution, the overlapping nature of both the distribution suggests that the templates generated from the same or different users are indistinct in nature which is also evident from Fig. 7. On the other hand, if pseudo-genuine and pseudo-imposter distributions are separated, it will be easier to identify the templates from the same user. The struggle in differentiating the templates leads to the unlinkability of the system.

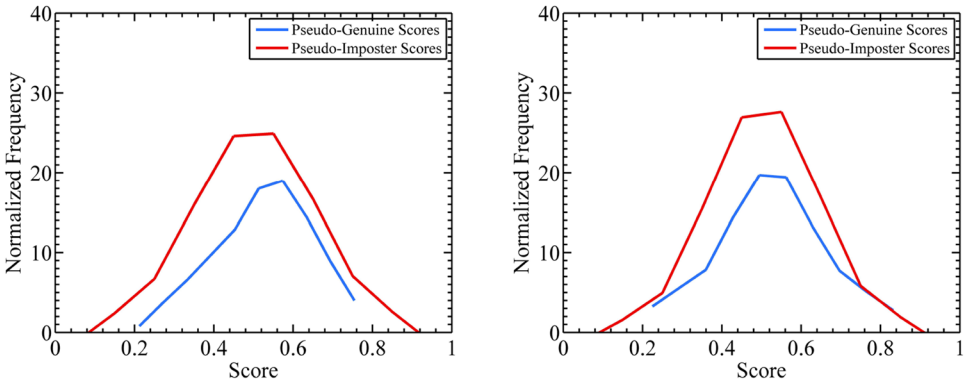
Privacy analysis clearly indicates that the proposed system preserves the privacy of each user by means of revocability, unlinkability, and non-invertibility. The complementary traits are fused together generating a cancelable template making system highly secure and robust against various attacks which are discussed in the next subsection.

4.4 Security analysis

The security of the biometric system is of utmost priority. No adversary, in any case, may be able to break through the system as it will arise many security problems. The proposed biometric system is also analyzed against many such attacks.

4.4.1 Brute force attack

Brute Force attack assumes that the adversary possesses no previous knowledge about the transformed or the original biometric feature. Each possible combination is used to generate and match a transformed template. In the worst-case scenario, let us assume the intruder knows the length of the template and the maximum value and the minimum value of the transformed template. In the proposed approach, the maximum and minimum values of template elements are -3.14165834 and 3.14165834 in DB1 with a precision of each element set to 8 decimal digits which amount to a total



(a) Database D1 **(b)** Database D2

Fig. 7 Comparision of pseudo-genuine and pseudo-imposter distribution for **a** D1 Database and **b** D2 Database

of 2^{29} possibilities for each element in the template. Hence, to generate a template it would require an exceptionally high brute force effort which makes it computationally infeasible.

4.4.2 Attacks via record multiplicity

During attacks via record multiplicity (ARM), the adversary possesses multiple transformed templates of the same user and tries to establish a connection so as to develop an image of the original biometric trait. For example, let us take two transformed templates T1 and T2 created using a different key for the same user. Also, during privacy analysis, unlinkability between these templates was proven experimentally. Further, the *i*th value of T1 depends upon the point of projection P_1 and it cannot be connected to *i*th value of T2 as it depends upon its points of projection $P_2.P_1$ and P_2 will always be different as different keys are chosen. Therefore, the attacker cannot mount ARM attack even after having different copies of transformed templates from the same user.

4.4.3 Blended substitution attacks

In the blended substitution, the attacker combines its data with user data in a single template. The blended template allows both users and attackers to authenticate against the same ID simultaneously. In the proposed approach, blended substitution is not feasible since, with only half of the user or attacker's attribute, the matching score would be rejected as an imposter for both user and attacker.

In sum, the proposed system is an accurate, highly robust and reliable solution that can be employed in critical applications. The generated cancelable templates are highly non-invertible, unlinkable and revocable and hence address privacy and security concerns.

5 Conclusion and future directions

In this paper, a multimodal biometric system is proposed that ensures not only high performance but also robustness to security and privacy concerns. The feature size is reduced to half thereby requiring low computational and space complexity. The security analysis of the proposed fusion mechanism shows that the approach is able to defend various attacks, thereby, assuring user-privacy and data-security. Moreover, the generated templates are highly revocable, thus can be regenerated in case the data gets compromised. Hence, the proposed approach can be deployed for biometric authentication in security-critical applications. Experimental analysis shows the effectiveness of the proposed method in comparison to state-of-the-art solutions. On average, the proposed fusion achieves DI of 16.63 and EER of 0.004.

In future, we will look forward to incorporate image quality as a reliability factor into the proposed cancelable biometric system. It will help in enhancing the adaptivity of the system so that it may address various other challenges like poor quality input images, improper sensor interaction, etc. The reliability factor can be learned over a time period by estimating image quality for specific users and can make the system highly reliable and robust. The work can also be extended to integrate different security levels in the proposed framework.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Gupta, K., Walia, G.S., Sharma, K.: Quality based adaptive score fusion approach for multimodal biometric system. Appl. Intell. **50**, 2824–2836 (2019)
2. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recogn. Lett. **24**(13), 2115–2125 (2003)
3. Gupta, K., Walia, G.S., Sharma, K.: Multimodal Biometric System using Grasshopper Optimization. In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 387–391. IEEE (2019)
4. Ross, A., Jain, A.K.: Multimodal biometrics: an overview. In: 12th European Signal Processing Conference, 1221–1224. IEEE (2004)
5. Haghighat, M., Abdel-Mottaleb, M., Alhalabi, W.: Discriminant correlation analysis: real-time feature level fusion for multimodal biometric recognition. IEEE Trans. Inf. Forensics Secur. **11**(9), 1984–1996 (2016)
6. Ratha, N.K., Chikkerur, S., Connell, J.H., Bolle, R.M.: Generating cancelable fingerprint templates. IEEE Trans. Pattern Anal. Mach. Intell. **29**(4), 561–572 (2007)
7. Pillai, J.K., Patel, V.M., Chellappa, R., Ratha, N.K.: Sectored random projections for cancelable iris biometrics. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1838–1841 (2010)
8. Pillai, J.K., Patel, V.M., Chellappa, R., Ratha, N.K.: Secure and robust iris recognition using random projections and sparse representations. IEEE Trans. Pattern Anal. Mach. Intell. **33**(9), 1877–1893 (2011)
9. Savvides, M., Kumar, B.V., Khosla, P.K.: Cancelable biometric filters for face recognition. In: Proceedings of the 17th International Conference on Pattern Recognition. ICPR 2004, vol. 3, pp. 922–925. IEEE (2004)
10. Maiorana, E., Campisi, P., Fierrez, J., Ortega-Garcia, J., Neri, A.: Cancelable templates for sequence-based biometrics with application to on-line signature recognition. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. **40**(3), 525–538 (2010)
11. Jin, A.T.B., Ling, D.N.C., Goh, A.: Biohashing: two factor authentication featuring fingerprint data and tokenised random number. Pattern Recogn. **37**(11), 2245–2255 (2004)
12. Zuo, J., Ratha, N.K., Connell, J.H.: Cancelable iris biometric. In: 2008 19th International Conference on Pattern Recognition, pp. 1–4. IEEE (2008)

13. Nandakumar, K., Jain, A.K.: Biometric template protection: bridging the performance gap between theory and practice. IEEE Signal Process. Mag. **32**(5), 88–100 (2015)
14. Patel, V.M., Ratha, N.K., Chellappa, R.: Cancelable biometrics: a review. IEEE Signal Process. Mag. **32**(5), 54–65 (2015)
15. Teoh, A.B.J., Yip, W.K., Toh, K.-A.: Cancellable biometrics and user-dependent multi-state discretization in BioHash. Pattern Anal. Appl. **13**(3), 301–307 (2010)
16. Teoh, A.B., Goh, A., Ngo, D.C.: Random multispace quantization as an analytic mechanism for biohashing of biometric and random identity inputs. IEEE Trans. Pattern Anal. Mach. Intell. **28**(12), 1892–1901 (2006)
17. Teoh, A.B.J., Yuang, C.T.: Cancelable biometrics realization with multispace random projections. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **37**(5), 1096–1106 (2007)
18. Wang, Y., Plataniotis, K.N.: An analysis of random projection for changeable and privacy-preserving biometric verification. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **40**(5), 1280–1293 (2010)
19. Paul, P.P., Gavrilova, M., Klimenko, S.: Situation awareness of cancelable biometric system. Vis. Comput. **30**(9), 1059–1067 (2014)
20. Yang, B., Hartung, D., Simoens, K., Busch, C.: Dynamic random projection for biometric template protection. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS), pp. 1–7. IEEE (2010)
21. Lacharme, P., Cherrier, E., Rosenberger, C.: Preimage attack on biohashing. In: 2013 International Conference on Security and Cryptography (SECRYPT), pp. 1–8. IEEE (2013)
22. Sadhya, D., Raman, B.: Generation of cancelable Iris templates via randomized bit sampling. IEEE Trans. Inf. Forensics Secur. **14**(11), 2972–2986 (2019)
23. Wang, S., Hu, J.: Design of alignment-free cancelable fingerprint templates via curtailed circular convolution. Pattern Recogn. **47**(3), 1321–1329 (2014)
24. Sharma, R.P., Dey, S.: Fingerprint liveness detection using local quality features. Vis. Comput. **35**(10), 1393–1410 (2019)
25. Ali, S.S., Ganapathi, I.I., Prakash, S., Consul, P., Mahyo, S.: Securing biometric user template using modified minutiae attributes. Pattern Recogn. Lett. **129**, 263–270 (2020)
26. Trivedi, A.K., Thounaojam, D.M., Pal, S.: Non-Invertible cancellable fingerprint template for fingerprint biometric. Comput. Secur. (2020). https://doi.org/10.1016/j.cose.2019.101690
27. Wu, S.-C., Chen, P.-T., Swindlehurst, A.L., Hung, P.-L.: Cancelable biometric recognition with ECGs: subspace-based approaches. IEEE Trans. Inf. Forensics Secur. **14**(5), 1323–1336 (2019)
28. Rathgeb, C., Busch, C.: Cancelable multi-biometrics: mixing iriscodes based on adaptive bloom filters. Comput. Secur. **42**, 1–12 (2014)
29. Kumar, N., Singh, S., Kumar, A.: Random permutation principal component analysis for cancelable biometric recognition. Appl. Intell. **48**(9), 2824–2836 (2018)
30. Dwivedi, R., Dey, S.: A novel hybrid score level and decision level fusion scheme for cancelable multi-biometric verification. Appl. Intell. **49**(3), 1016–1035 (2019)
31. Walia, G.S., Rishi, S., Asthana, R., Kumar, A., Gupta, A.: Secure multimodal biometric system based on diffused graphs and optimal score fusion. IET Biom. **8**(4), 231–242 (2019)
32. Kaur, H., Khanna, P.: Random distance method for generating unimodal and multimodal cancelable biometric features. IEEE Trans. Inf. Forensics Secur. **14**(3), 709–719 (2018)
33. Walia, G.S., Jain, G., Bansal, N., Singh, K.: Adaptive weighted graph approach to generate multimodal cancelable biometric templates. IEEE Tran. Inf. Forensics Secur. (2019). https://doi.org/10.1109/TIFS.2019.2954779
34. Paul, P.P., Gavrilova, M.L.: A novel cross folding algorithm for multimodal cancelable biometrics. Int. J. Softw. Sci. Comput. Intell. **4**(3), 20–37 (2012)
35. Chin, Y.J., Ong, T.S., Teoh, A.B.J., Goh, K.: Integrated biometrics template protection technique based on fingerprint and palmprint feature-level fusion. Inf. Fusion **18**, 161–174 (2014)
36. Gomez-Barrero, M., Rathgeb, C., Li, G., Ramachandra, R., Galbally, J., Busch, C.: Multi-biometric template protection based on bloom filters. Inf. Fusion **42**, 37–50 (2018)
37. Abdellatef, E., Ismail, N.A., Elrahman, S.E.S.A., Ismail, K.N., Rihan, M., El-Samie, F.E.A.: Cancelable multi-biometric recognition system based on deep learning. Vis. Comput. **2**, 1–13 (2019)
38. Das, R., Piciucco, E., Maiorana, E., Campisi, P.: Convolutional neural network for finger-vein-based biometric identification. IEEE Trans. Inf. Forensics Secur. **14**(2), 360–373 (2018)
39. Kacar, U., Kirci, M.: ScoreNet: deep cascade score level fusion for unconstrained ear recognition. IET Biom. **8**(2), 109–120 (2019)
40. Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., Pietikäinen, M.: Deep learning for generic object detection: a survey. Int. J. Comput. Vis. **128**, 1–58 (2019)
41. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
42. Daugman, J.G.: High confidence visual recognition of persons by a test of statistical independence. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1148–1161 (1993)
43. Farina, A., Kovacs-Vajna, Z.M., Leone, A.: Fingerprint minutiae extraction from skeletonized binary images. Pattern Recogn. **32**(5), 877–889 (1999)
44. Sudiro, S.A., Paindavoine, M., Kusuma, T.M.: Simple fingerprint minutiae extraction algorithm using crossing number on valley structure. In: IEEE Workshop on Automatic Identification Advanced Technologies, pp. 41–44 (2007)
45. Daugman, J.: How iris recognition works. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 21–30 (2004)
46. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, I., Igarza, J.-J., Vivaracho, C., Escudero, D., Moro, Q.-I.: MCYT baseline corpus: a bimodal biometric database. IEE Proc. Vis. Image Signal Process. **150**(6), 395–401 (2003)
47. Kumar, A., Passi, A.: Comparison and combination of iris matchers for reliable personal authentication. Pattern Recogn. **43**(3), 1016–1026 (2010)
48. Cappelli, R., Ferrara, M., Franco, A., Maltoni, D.: Fingerprint verification competition 2006. Biom. Technol. Today **15**, 7–9 (2007)
49. MMU2 Iris Image Databases. http://pesona.mmu.edu.my/ccteo/. Accessed June 2019 (2008)
50. Yang, W., Wang, S., Hu, J., Zheng, G., Valli, C.: A fingerprint and finger-vein based cancelable multi-biometric system. Pattern Recogn. **78**, 242–251 (2018)
51. Jin, Z., Hwang, J.Y., Lai, Y.-L., Kim, S., Teoh, A.B.J.: Ranking-based locality sensitive hashing-enabled cancelable biometrics: index-of-max hashing. IEEE Trans. Inf. Forensics Secur. **13**(2), 393–407 (2017)

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



Keshav Gupta is currently pursuing his Ph.D. degree in the Computer Science and Engineering Department, Delhi Technological University, New Delhi, India. He completed his M. Tech. in software engineering from Delhi Technological University, New Delhi. His current research interests include biometric systems, pattern recognition, image processing and machine learning. He has published many research papers in reputed international journals and conferences.



Dr. Gurjit Singh Walia received his Ph.D. degree in the field of computer vision from Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, and the M.E. degree in electronics from Punjab Engineering College, Chandigarh. He is working as a Senior Scientist with Defence Research and Development Organization, New Delhi. His current research interests include machine learning, pattern recognition, and information security. He has published over 30 research papers in international journals and conferences. He is Fellow of IETE, Fellow of Institute of Engineering, Senior Member of IEEE, Member of CSI and CRSI. He is reviewer of various IEEE transactions, Elsevier and Springer journals.



Dr. Kapil Sharma is working as a Professor and the Head of Department of Information Technology, Delhi Technological University, Delhi. He has completed Ph.D. Degree in Computer Science from M.D. University, India. He has obtained his B.E. and M.Tech. degrees in Computer Science from M.D. University and IASE, India, respectively. His research interests include system design, pattern recognition, computer vision and soft computing. He has published more than 100 research papers in refereed journals.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09013-2

Novel approach with nature-inspired and ensemble techniques for optimal text classification

Anshu Khurana [1] · Om Prakash Verma [2]

Received: 31 May 2019 / Revised: 1 March 2020 / Accepted: 1 May 2020
Published online: 13 June 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Text classification reduces the time complexity and space complexity by dividing the complete task into the different classes. The main problem with text classification is a vast number of features extracted from the textual data. Pre-processed dataset have many features, some of which are not desirable and act only like noise. In this paper, a novel approach for optimal text classification based on nature-inspired algorithm and ensemble classifier is proposed. In the proposed model, feature selection was performed with Biogeography Based Optimization (BBO) algorithm along with ensemble classifiers (Bagging). The use of ensemble classifiers for classification delivers better performance for optimal text classification as compared to an individual classifier, and hence, improving the accuracy. Ensemble classifiers combines the weakness of individual classifiers. The individual classifiers are unable to improve the classification results when compared to ensemble classifier. The selected features, after feature selection using BBO algorithm, are classified into various classes using six machine learning classifier. The experimental results are computed on ten text classification datasets taken from UCI repository and one real-time dataset of an airlines. The four different measures namely; Accuracy, Precision, Recall and F- measure are used to validate performance of our model with ten-fold cross-validation. For feature selection process, a comparison is performed among state-of-the-art algorithms available in the literature. Results shows that BBO for feature selection outperforms the other similar nature-based optimization techniques. Our proposed approach of BBO with ensemble classifier is also compared with techniques proposed by other researchers and we analyzed the results quantitatively and qualitatively.

Keywords Text classification · Feature selection · Nature-based optimization · Machine learning classifier · Ensemble classifier · Biogeography based optimization

✉ Anshu Khurana
anshuchugh@gmail.com

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

[2] Department of Electronics and Communication, Delhi Technological University, New Delhi, India

1 Introduction

As there is a rise in research areas related to data mining, the advancement in Information Communication and Technology (ICT), gives opportunities to all users to access the information quickly and at a faster rate. Due to increase in demand for information, there is a simultaneously increase in the number of text documents, as the availability and accessibility of the digital information are saved and organized in the forge of text [26]. The application area of text classification is widely falling in the era of text analysis, and according to our survey, many machine learning supervised algorithms are applied for text classification such as Naive Bayes [10] [7], Decision Tree [18], K- Nearest Neighbor [7], Random Forest (RF) [18], Support Vector Machine [5, 7, 9], and ensemble classifiers [39]. There are many related issues arises for text classification majorly due to high-dimensional feature set [59]. Large feature space decreases the efficiency of the algorithm [18], hence reduces the efficacy of text classification. The presence of noise and un-useful features for dataset can also decrease the efficiency to classify text into various classes. Feature selection mainly selects the desired and important features which provides necessary related information [14], hence increases the efficiency of text classification. Feature selection techniques are broadly classified into the filter approach and wrapper approach [31]. To extract only desirable features, the filter approach and the wrapper approach have many feature selection techniques. Some of the filter techniques used are Information Gain (IG), Poisson Distribution, improved Gini Index (GINI), Odds Ratio (OR), Chi-Square, Binomial Hypothesis testing, and many other techniques [64]. The proposed nature-based optimization, used for feature selection, is categorized into the wrapper approach. Many evolutionary optimization algorithms such as: Genetic Algorithm (GA) [61], Firefly Algorithm [63], Particle Swarm Optimization (PSO) [68], Ant Colony Optimization (ACO) [37], and Biogeography- Based Optimization [54] has been applied in literature for feature selection. These nature-based approaches are used for high dimensional datasets by researchers and practitioners. Wrapper methods are known for high precision in classification. Hybrid approaches are also used for the feature selection [19]. The hybrid approach produces the best performance among all the approaches, as it combines the wrapper approach and filter approach. In [54], Biogeography based optimization solved the problem of best selection of sensor for aircraft engine. The BBO outperformed when compared to other nature-based optimization techniques such as GA, PSO and Differential Evolution. In many classification problems other than text classification, BBO is used for optimization and has proved best among others [3, 41, 42, 55]. Hence, in our proposed method, we have chosen the BBO algorithm for feature selection.

To the best of our knowledge, the BBO algorithm is not applied for the text classification. In this work, we proposed a hybrid BBO for feature selection with ensemble classifier. The extracted features are stored in an array and passed for training to six machine learning algorithms namely: Naïve Bayes (NB), K-Nearest Neighbor (k-NN), and Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT) and ensemble classifier.

The rest of this article is organized in the following order. A brief related work on feature selection techniques for text classification and ensemble methods is presented in section 2. The brief theoretical background of all used algorithms is presented in section 3. The proposed methodology and performance measures is presented in section 4. A performance comparison with other techniques and complete illustration of the results along with the statistical significance of our proposed technique is presented in section 5. Finally, the conclusion of the proposed technique is present in section 6.

2 State-of-the-art literature and main contribution of the paper

2.1 Related work on feature selection

The text classification optimization is performed over the years with many algorithms [25]. Optimization of classification can be performed by reducing undesirable features or feature selection. Feature selection can be performed implicitly as well as explicitly in classification techniques. Nag *et al*. [38] have performed feature selection implicitly in genetic programming by applying changes in mutation operator. Feature selection has three approaches, namely; filter approach [2, 21], wrapper approach [28, 32, 43] and embedded approach [60, 65]. The filter based approach mainly depends on the selection criteria of the dataset and not on the learning algorithm [57].

The categories of wrapper approach such as PSO [15, 27], GA, ACO, Differential Evolution [10, 69] mainly inclined towards the design of algorithms that can be used for the optimization of a large datasets. Hence, the wrapper approach produces more accurate results than a filter approach.

Zhang *et al*. [68] proposed the feature selection technique which results in cost reduction as well. They have used probability-based encoding and Pareto domination relationship. Based on the performance measure, their proposed approach yields the best results in comparison to five multi-objective algorithms. But their proposed model failed to improve the measures for low-dimensional datasets. The sparse solution produced by pareto front was producing the worst hyper volume values. In [19], the study gives a new idea of hybrid search methodology, the advantages of cooperation of both genetic approach and filter feature selection method. They have proposed a model with rank-based feature extraction by assigning the ranks in descending order according to their importance. The authors have used Naive Bayes and associative classification for training and testing of Arabic text dataset. Their research shows that the proposed enhanced genetic algorithm provides better results as compared to classical genetic algorithm. But their results of Precision, Recall and F-measure shows that the proposed model didn't work for all text datasets and all categorization techniques. However, their model works well on high-dimensional datasets and not co-related classes.

Researchers in [50], proposed a new model of feature selection with a Chaotic Crow Search Algorithm. Their proposed model works with best chaotic map. They have compared their mean fitness value with other metaheuristic algorithms such as: Chicken Swarm Optimization [22], Gray Wolf Optimizer [13], Sine Cosine algorithm [23], Particle Swarm Optimization [30], Whale Optimization algorithm [48], Artificial Bee Colony Optimization [52], Moth Flame Optimization [67] and Flower Pollination algorithm [49]. Their approach performed better than all the algorithms used for comparison. Through algorithm different chaotic maps were adopted. Selection of optimal chaotic map was difficult and time consuming process, as every dataset didn't improved result with same chaotic map.

In [10], researchers used Genetic Algorithms, Differential Evolution [56] and Simulated Annealing [46] for optimization and Fine-Tuning Naive Bayesian algorithm [11, 12] for training on 53 datasets. They have also applied their model on 18 text datasets. Their multi-parent differential approach gives better accuracy with fine-tuning Naive Bayes algorithm when compared with Bernoulli NB [29, 66], Multinomial NB [24, 33, 45]. They have validated their classification model using accuracy measures only, which is not sufficient measure for generalization. The computational time was quite high for the tuning of NB classifier. In the study [35], researchers have compared local feature selection techniques with

global feature selection technique using transformation method [53]. They have performed their experiment on different applications of flat and hierarchical databases, and final computed results showed that local transform outperforms the global transform. Jiang *et al.* [26], presented a hybrid model of softmax regression and deep belief network. They performed feature extraction using a deep belief network, and the selected features are classified using softmax regression. They have also performed parameter tuning using Broyden-Fletcher-Goldfarb-Shanno algorithm. The computed results of classification accuracy showed that fine-tuning of the hybrid algorithm performs better than the classical approach of Support Vector Machine and K-Nearest Neighbor.

In [69] zorarpaci *et al.* performed dimensionality reduction by using a hybrid combination of Artificial Bee Colony algorithm with Differential Evolution. Their hybrid method gives improved accuracy and run time performance. In [7], the study introduced the idea of using ACO for optimum feature subset. They have used a ranking method to select relevant features. The authors have used NB, k-NN and SVM as classifiers on 15 datasets for the classification of text. Through the experiment results, authors have proved that their proposed method found better classification performance. But, the proposed model failed with the k-NN classifier for some of the corpus. The above models proposed in [7, 26, 35, 69] worked only with low dimensional datasets, as performance measures fail to produce good results with high-dimensional datasets.

2.2 Related literature on ensemble methods

Ensemble methods are used to improve the accuracy as it combines the output of various weak learning classifiers. In [5], the technique uses the ensemble classifiers for sentiment classification. They have used NB, Bagging and SVM with vote algorithm as base classifier for ensemble classifier. Optimization of SVM parameters is performed to improve classification accuracy, and from experimental results, it was proved that multiple classifiers show good results.

In [39], study shows how the keyword extraction method is combined with an ensemble. All the base classifiers are compared with five ensemble methods using performance measure: Precision, Recall, F-measure and Area Under Curve (AUC). The final conclusion was made with the result shown by declaring two ensemble classifiers Bagging and Random Subspace with RF as base classifier, produces the best results. However, only statistical feature extraction method can be used with their model and results are based on accuracy measures. In [44], author proposed a model using five classifiers namely: SVM, Rule-Based classifiers, Induction based classifier, General Inquirer based classifier and Statistics based classifier. The results shows that the use of an ensemble classifier improves the performance of text classification. Both models proposed in [5, 44] failed to improved performance on high dimensional datasets. In the [64], the different subset of features namely; word relation based feature subset and parts of speech are extracted. The different base classifiers SVM, Maximum entropy and NB were used with ensemble learners. The results have proved the effectiveness of ensemble classifiers. Although, their results are proved only for accuracy measures.

In [6], the authors have used the feature representation scheme for sentiment classification. Empirical analysis shows that the use of set ensemble classifiers improves the classification performance. In another study [62], the five base learners are applied on three ensemble classifiers namely; NB Maximum Entropy, k-NN and SVM. In [16], the researchers proposed a Bayes model which is an average of different ensemble methods for sentiment classification.

However, the research presented in [6, 16, 62], ensemble classifier didn't work on large datasets and diverse features. The models presented in [16, 62] also worked only on balanced datasets and failed to improve classification performance for multi-class datasets. Hence, we proposed a model that works on high dimensional datasets and analyzed results with four performance measures. The applicability on practical problems is proved by testing the proposed model on real-time dataset of an airlines.

2.3 Main contribution of paper

Following are the main contribution of this research for optimization of text classification:

- A novel algorithm is proposed for feature selection based on BBO. This algorithm is a nature-based algorithm that selects optimal subset of features. The algorithm produces outstanding results when combined with other nature-based algorithms used for feature selection. The BBO has proven best among other optimization techniques. To get an optimal performance, 50 runs of every model were performed.
- Three nature-inspired algorithms implemented for feature selection are; GA, PSO and BBO. The classification is performed using six machine learning classifiers. The model is evaluated on ten text datasets and one real-time dataset. We have chosen four performance measures namely: Accuracy, Precision, F-measure and Recall for the validation and analysis of classification performance of the proposed model.
- To the best of our knowledge, the nature-inspired algorithm with an ensemble classifier is accustomed for first time for the optimization of text classification. The BBO is used for feature selection. The ensemble classifier are known to produce better accuracy than individual classifier. The validation of a model is computed on ten open source datasets and one real-time dataset for text classification.
- Implementation and validation of the proposed model is carried out on all ten text classification datasets with 10-fold cross-validation. To compute the optimal performance of text classification, 50 runs were performed. The 10-fold cross-validation method is used to minimizing the biasing in results.
- Implementation of our proposed model was carried out on a real-time dataset of airlines using 10-fold cross-validation. Testing our model on a real dataset validates its applicability on practical problems. The performance of the proposed model in case of real dataset shows its applicability scope and significance.
- Performance of the proposed model is compared with other nature-inspired algorithms. Comparison is performed using four performance metrics and our model produced better results across all metrics. Comparison with the techniques proposed by other researchers shows the uniqueness and enhanced performance given by the proposed model. We have used the Friedman test for statistical analysis to validate the chosen technique qualitatively.

3 Theoretical foundations

Our approach of feature selection with BBO algorithm in combination with the ensemble classifier is compared with similar state-of-the-art-algorithms. We have compared the feature selection technique by implementing GA, PSO and BBO. The selected classifiers for training and testing are NB, k-NN, SVM, RF, DT and ensemble (Bagging). To validate the technique

we have chosen performance measures namely: Precision, Recall, Accuracy and F-measure. All these performance values are calculated with every model using ten-fold cross-validation. In this paper, we have created 35 models of every dataset. Experimental results are validated using a statistical test, i.e. Friedman test. All the algorithms implemented in the paper are briefly described in the following subsection and the detailed working of the proposed technique is described in the next section.

3.1 Feature selection methods

There are various existing feature selection methods [36] which extract only relevant and important features from the data, as discussed earlier. These techniques helps to improve the accuracy of text classification. In this paper, we have used the wrapper approach. Wrapper approach mainly consists of nature-inspired algorithms that can be used for feature selection. The comparison has been performed on all the selected datasets of text. The computation has been done on a large number of features by performing feature selection methods.

3.1.1 Genetic algorithm

Genetic Algorithm [8, 17] is used as feature selection technique, where all features are considered as the input. By using feature selection approach, the main goal is to improvise the general performance using optimization. By following the procedure of genetic algorithm for feature selection, the main work is to produce the best feature subset based on the fitness function. GA starts with random initialization of population. At every generation, new individuals are selected depending on the value of the fitness function. In this study, we have taken rank-based fitness selection. After fitness assignment, the selection operator chooses the individual that recombines with the other individuals to form next generation. Roulette wheel selection is performed in which individuals are selected randomly with the help of the spinning movement of a wheel. Finally, the two parents were randomly taken for the crossover operation, which is responsible for reproducing the new fittest subsets known as children of parents. But, the working of mutation operator depends on a subset (single) which randomly adjusts some preferred values of features for the survival of the fittest. The flowchart of GA used as feature selection approach is shown in Fig.1 and algorithm for GA as feature election is explained in Algorithm 1

3.1.2 Particle swarm optimization

The other nature-inspired algorithm is Particle Swarm Optimization [40]. It chooses random particles and the velocity associated with it [27]. Position at generation t is denoted as $X_i(t) = [x_{i1}(t), x_{i2}(t)\ldots\ldots\ldots x_{id}(t)]$, where $X_{id}(t)$ is the position of i^{th} particle in correspondence to the d^{th}dimension and velocity as $V_i(t) = [\, v_{i1}(t), v_{i2}(t). \ldots v_{id}(t)]$, where V_{id} is the velocity of i^{th} particle with respect to the d^{th} dimension. Then at generation $t+1$ the updated position and updated velocity is described by the following equations:

$$V_{id}(t+1) = wV_{id}(t) + c_1 rand1(p_{bestid}(t) - X_{id}(t)) + c_2 rand2(g_{bestid}(t) - X_{id}(t)) \quad (1)$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (2)$$

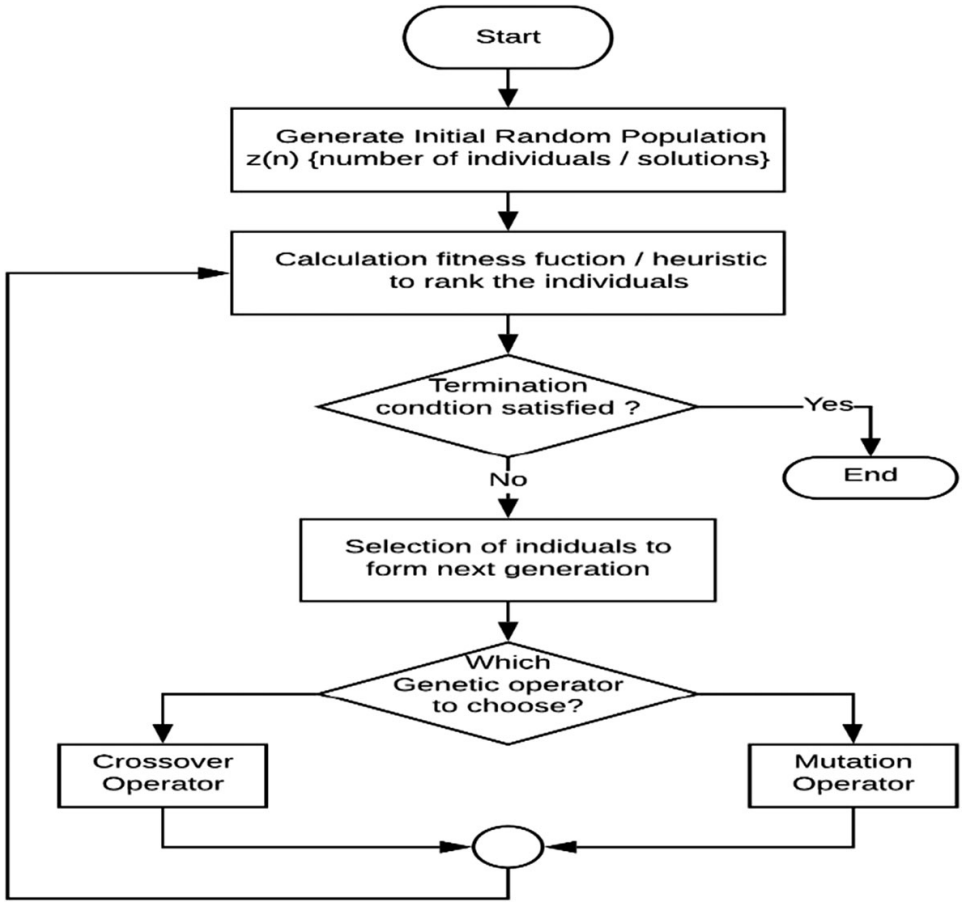


Fig. 1 Flow chart of genetic algorithm for feature selection

where $p_{best\ id}(t)$, which represents the best position of individual i in d dimensional space until generation t [68] and $g_{best\ id}(t)$, which represents the best position of the group in d dimension until generation t. The non- negative velocity coefficients, c_1 is the cognitive parameter and $c_2$is the social parameter, w is an inertia weight, $rand1$ and $rand2$are two random values ranging between 0 and 1. The flowchart of PSO as feature selection is described in Fig.2.

3.2 Biogeography based optimization

Biogeography based optimization is a nature-inspired evolutionary algorithm which mainly finds a solution to the problem of global optimization. The convergence of the algorithm depends on immigration (of species) and emigration (of species) that reside on different islands according to the various factors considered by the species for a more friendly life. It gives the candidate a solution which can be referred as "habitat" and known as Habitat Suitability Index (HSI). The various factors like temperature, rainfall etc. on which migration on the island by species depends is Suitability Index Variables (SIV). HSI is k-dimension vector from which initial population is randomly generated. The improved suitability increases the count of species that emigrates raising the HSI and decreasing the immigration count. The habitats with high HSI refers as good solutions and they are densely populated. The habitat with low HSI refers as poor solutions and they are sparsely populated. There is some information sharing from high HSI to low HSI to control the emigration rate.

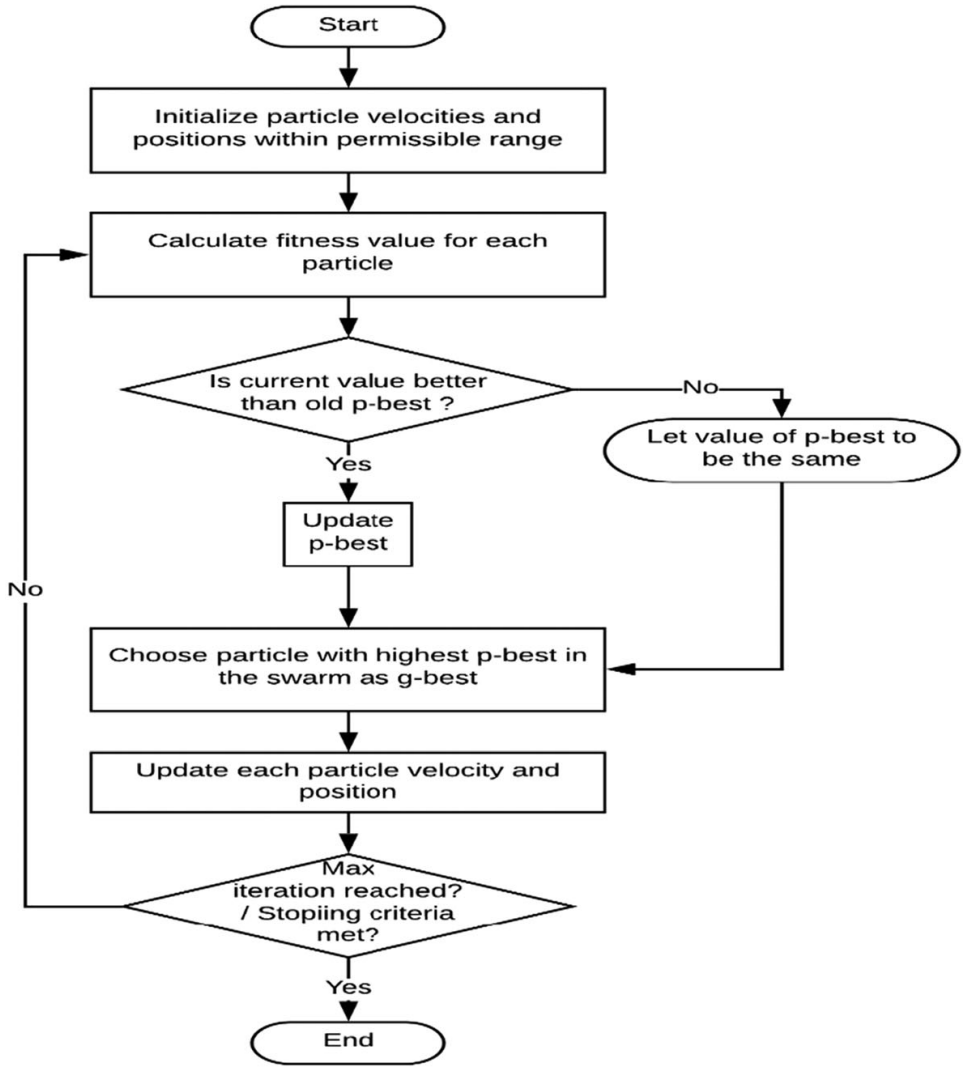


Fig. 2 Flow chart of PSO algorithm for feature selection

In BBO, mutation and migration [20, 54] of species is calculated using SIV vector, H(A). By considering the values of SIV, an optimal solution is obtained. Information

1. To determine the Habitat (H_l) with the probability $\propto$ (λ)
2. **if** H_l *is chosen* **then**

if *rand.random(0.0,1.0)*$< \lambda$ **then**

for $k \leftarrow 1$ **to** N **do**

Determine the Habitat (H_k) with the probability $\propto$ (μ)

if (*) is chosen* **then**

if *rand.random(0.0,1.0)* $< \mu$ **then**

(5)

end

end

end

end

end

Fig. 3 Pseudo-code for BBO migration

```
for y=1 to m do
    The parameters λ and μ are used to calculate probability P_i

     Select Habitat H_i(y)(SIV) with the probability ∝P_i
    if          == selected then
        if rand.random(0.0,1.0)<μ then
            Replace              with randomly produced SIV
        end
        end
end
```

Fig. 4 Pseudo-code for BBO mutation

sharing among the island is possible with migration. Islands with a high value of HSI have low immigration rate due to less number of resources and high population. The immigration rate λ, which signifies the living of species on other island and emigration rate μ, signifies the leaving of species from the native island are consider as objective functions. When there exist zero species on any island then only maximum immigration, I is possible. When habitat supports the maximum number of species then rate of

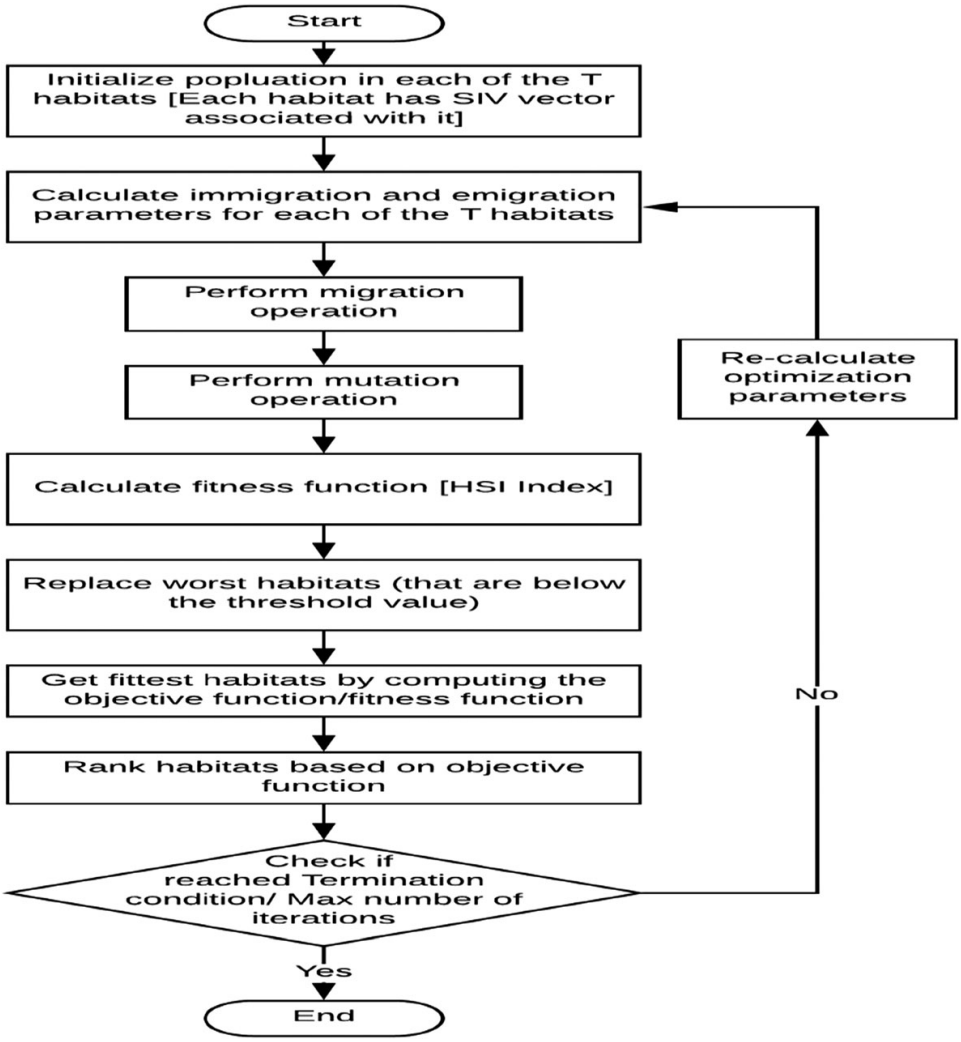


Fig. 5 Flowchart of BBO as feature selection

immigration starts falling to zero, due to lack of resources and high population, species start moving to other islands, and also emigration reaches its maximum possible value E. Alternatively, the rate of emigration is zero when there is no existence of any single species on the island. Variation in immigration is due to factors like climate change, resource scarcity, etc.

In BBO, λ_k is the probability of replacing a given independent candidate chosen from K-th candidate solution. After performing random selection with roulette wheel and independent variable candidate (which is chosen) has to be replaced, the probability of emigrating candidate solution (selected) is proportional to the emigration probability, which is performed using roulette wheel selection is given as:-

$$Prob(x_k)selected for emigration = \mu_k / \sum_{j=1}^{A} \mu_j \tag{3}$$

where, $tk = 1 \ldots n$ are the number of species,

and $kj = 1, 2 \ldots A$, A is represented as the number of candidate solutions in the population.

In BBO, migration is calculated using probability among species with consideration of suitability factors. Due to some low suitability factors on an island, the species tend to migrate or move to any other or new island. The migration process modifies the habitat of island. Migration uses probability to change a habitat H_k. Then based on immigration rate λ, the probability $H_k$is modified proportionally, and changes in λ is done with respect to (probability)$H_t \propto \mu$, emigration rate. Migration of species is mapped into habitat vector H(A), immigration rate and emigration rate is represented as:-

$$\lambda_p = I(1-p/k)\mu_p = E(p/k) \tag{4}$$

where, I is considered as the maximum possible immigration rate, E is considered as the maximum possible emigration rate, p is the number of species of p-th individual and k is the maximum number of species. Pseudocode for the migration is described in Fig.3 as

Mutation is a probability operation that is responsible for variations in habitat.

SIV's. The mutation rate is calculated by:-

$$m(s) = m_{max}(1-P_i)/P_{max} \tag{6}$$

where m is mutation operator, m_{max} is user-defined parameter, P_i is the probability of *ith* species and P_{max} is probability of the maximum number of species. Mutation is likely to increase the melange among the population. Pseudocode for mutation is described in Fig. 4.

3.2.1 The BBO as feature selection

Flowchart of the BBO as feature selection is shown in Fig. 5.

Parameters considered for BBO algorithm are described as follows:-

Parameters	Represented as
Number of habitats, habitat []	T
Feature Set	SIV[] (Each habitat have SIV[] associated with it)
Maximum number of iterations	K

Parameters	Represented as
Immigration	λ []
Emigration	μ[]
Maximum immigration	$\lambda = 500$
Maximum emigration	$\mu = 600$
Array to store crucial/redundant SIV factor	gb []

Working of BBO algorithm for feature selection explained in Fig.6.

3.3 Classification algorithms

To evaluate the classification performance, we have chosen supervised machine learning classifiers. The six best performing machine learning classifiers used for text classification in state-of-the-art are considered. The six classifiers we have chosen are Naive Bayes [4], K-Nearest neighbor [40], Support Vector Machine [51] [5], Random Forest [1], Decision Tree [18] and ensemble classifier [5].

1. Initialize BBO parameters
2. Generate initial population
2.1 Random population in each habitat (habitat [ ]): H_1, H_2,...,H_T
2.2 Each habitat index is associated with SIV values(feature)
2.3 Initial immigration rate of each habitat =λ [ ]
2.4 Initial emigration rate of each habitat =μ[ ]
2.5 Each index of Habitat is associated to a feasible answer for the given problem.
3. Calculate the fitness of each habitat (solution) of the population.
3.1 It is based on the classification rate of the evolved subset of features.
3.2 Calculate HSI value of each habitat of the population: hab_index [ ]
3.3 Rank habitats based on their HSI value.
Set iteration variable K: depicts the maximum number of iterations and k is the variable which is gradually increased.
while $k < T$ **do**
Calculate emigration rate (μ_j) and immigration rate (λ_i) for each habitat(solution) of the population. Here, j = 1,2,...T and i = 1,2,...T
H_j is selected where selection criterion is based on emigration rate (μ_j) for the emigration
H_i is selected where selection criterion is based on immigration rate (λ_i) for the immigration
To perform migration operation
To perform mutation operation on H_i
Produce new population by replacing previous (old) H_i from previous population with new H_i
Recalculate the values for habitats, compute their corresponding HSI values Increment k(the iteration variable)
end
4. Get fittest habitats based on the threshold value.
5. Ranking is performed second time with respect to population
// The SIV value that causes the change in hab_index [], is a crucial feature (and the redundant features are removed)
for $k < N$ **do**
if *SIV[k]==crucial value* **then**
gb[k]=1
else gb[k]=0
end end
6. Based on gb[]=1 the crucial feature are kept and the redundant features are removed
7. Dataset with reduced features is stored in SIV[] (Each index of hab_index [] corresponds to the list of features stored in SIV[])

Fig. 6 BBO algorithm for feature selection

3.4 Performance measure

To evaluate the performance of our hybrid classification model, we have considered four measures to verify the model behavior. The four performance measures chosen are Precision, Recall, Accuracy and F-measure. In many studies [10, 14, 31], only Accuracy is the main performance measure to evaluate the model but choosing Precision, Recall and F-measure for multi-class classification verifies the model Accuracy. We have computed all performance measures with the help of confusion matrix consist of following instances:

- If an instance is predicted as positive and also lies in the positive class in its correct actual state it is known as True Positive (TP).
- If an instance is predicted as negative but lies in the positive class in its correct actual state it is known as False Negative (FN).
- If an instance is predicted as positive but belongs in the negative class in its correct actual state it is known as False Positive (FP).
- If an instance is predicted as negative and also belongs in the negative class in its correct actual state it is known as True Negative (TN).
- The performance measures are briefly described in the following subsections.

3.4.1 Accuracy

It is calculated by taking the ratio of instances that are correctly predicted and the entire number of instances (both correctly predicted and incorrectly predicted) [10, 14, 18]. Accuracy is calculated as:-

$$Accuracy\ (A) : [\mathrm{TN}] + [\mathrm{TP}]/[\mathrm{TN}] + [\mathrm{TP}] + [\mathrm{FN}] + [\mathrm{FP}] \tag{7}$$

3.4.2 Precision

Precision is calculated by taking the ratio of true positive instances and the entire instances which are predicted as positive [7].

$$Precision\ (P) : [TP]/[TP] + [FP] \tag{8}$$

3.4.3 Recall

It is calculated by taking the ratio of true positive instances and the entire instances which are positive in their correct actual state [32].

$$\mathrm{Recall(R)} : [\mathrm{TP}]/[\mathrm{FN}] + [\mathrm{TP}] \tag{9}$$

3.4.4 F-measure

F-measure is given by the harmonic mean of Precision and Recall [18, 59].

$$F\text{–}measure\ (F\ ) : 2[P][R]/[R] + [\mathrm{P}] \tag{10}$$

3.5 Dataset description

In this paper, we have considered ten text classification dataset from UCI repository [10] and one real-time dataset of airlines from MOA repository [34]. Description of all datasets is given in Table 1.

3.6 Statistical test

The use of a statistical test evaluates the techniques and validates the model performance. We have chosen a Friedman test for the evaluation of performance measure and checked whether there is any significant difference when compared to other techniques. The Friedman test [58], is a non-parametric test and it is preferred over the parametric test as for the parametric test we have to consider some data assumptions. We have computed results with degree of freedom as 3, significant value (α=0.05) and the obtained p-value equals to 0.001. Higher rank indicate the significance of that particular technique. With the help of rank, we can find out the significant difference between the different techniques. The following hypothesis is considered:

Null Hypothesis There exists no significant difference between the algorithms GA, PSO, BBO and without optimization technique.

Alternate hypothesis There is a significant difference between the algorithms; GA, PSO, BBO and without optimization techniques.

4 Experimental method

In this paper, the ten text classification datasets are collected from the UCI repository and one real-time dataset from MOA. For a given dataset with features s_1, s_2, s_3.. s_n, we

Table 1 Description of dataset

Dataset	Instances	Classes	Features	Missing values
tr11	414	9	6430	0
tr12	313	8	5805	0
tr21	336	6	7903	0
tr23	204	6	5833	0
tr31	927	7	10,129	0
tr41	878	10	7455	0
tr45	690	10	8262	0
oh0	1003	10	3183	0
oh10	1050	10	3239	0
oh15	914	10	3101	0

used a feature subset selection technique. For feature selection, we have used GA, PSO and BBO and compared the performance measures. To completely validate the results we have used 10-fold cross validation. In this research, firstly we calculated all performance measures without using any feature selection technique on all datasets. Then, we performed feature selection on all datasets using GA, PSO and BBO to check the improved performance. By comparing the performance values of our proposed algorithm using BBO for the feature selection with GA and PSO, we concluded that BBO produces significantly better results. We selected the essential features by applying feature selection techniques individually and then pass the new dataset with selected features to classifiers. In total, 20 models were applied to all datasets to find the best model for the optimal solution.

4.1 Proposed hybrid model BBO-bagging for feature selection and classification

Ensemble techniques give promising research models in many areas especially in machine learning and pattern recognition [39]. The major goal of ensemble classifier is to combine the decision of predictions of several weak classifiers (base learner). Combination of classifiers enhances the robustness and Accuracy of a model than an individual classifier. There are mainly two categories of ensemble classifiers namely: averaging methods and boosting methods. In averaging methods, several predictive models are made independently and final results are concluded after averaging the results. Bagging and forest of randomized trees are examples of averaging methods. In boosting method, base learners are implemented sequentially with a goal to combine several weak models to produce strong ensemble classifier. We have selected Bagging as an ensemble classifier for our model as Bagging produces better results with strong and complex models compared to boosting technique. Bagging is also known as a bootstrap aggregating technique [5, 39] which combines the individual classifiers trained on different training sets to build a highly improved predictive model. When a weak learning classifier is under the training phase, the size of every new sample is equal to the size of original training set. Simple randomized sampling with replacement is deployed for producing new training sets. The aggregation of considered individual classifiers is performed to produce predictive results with the technique of majority voting.

In the proposed model, feature selection is performed using BBO as explained earlier in Fig. 6. and final subset of features are stored in array SIV. Each index of hab_index is associated to the list of features stored in SIV. Initialization of parameters is performed for every base learner classifier. Initially, D is taken as an empty set of classifiers and C is the number of classifiers to be train. Bootstrap samples S_1, $S_2 \ldots S_c$, which we have to train, are selected from Habitat []. A classification model, D_k, is built utilizing bootstrap samples by adding the current classifier to D. Now, aggregation of results of individual classifiers is performed with majority voting. Class with the highest number of votes is assigned to the test instance. Proposed hybrid approach of BBO and Bagging has improved the performance values further. Ensemble classifier was also combined and compared with GA and PSO. Separate 15 new models were designed for ensemble technique. Flow chart of the proposed methodology is shown in Fig.7. Algorithm for the proposed model is elaborated in Fig.8.

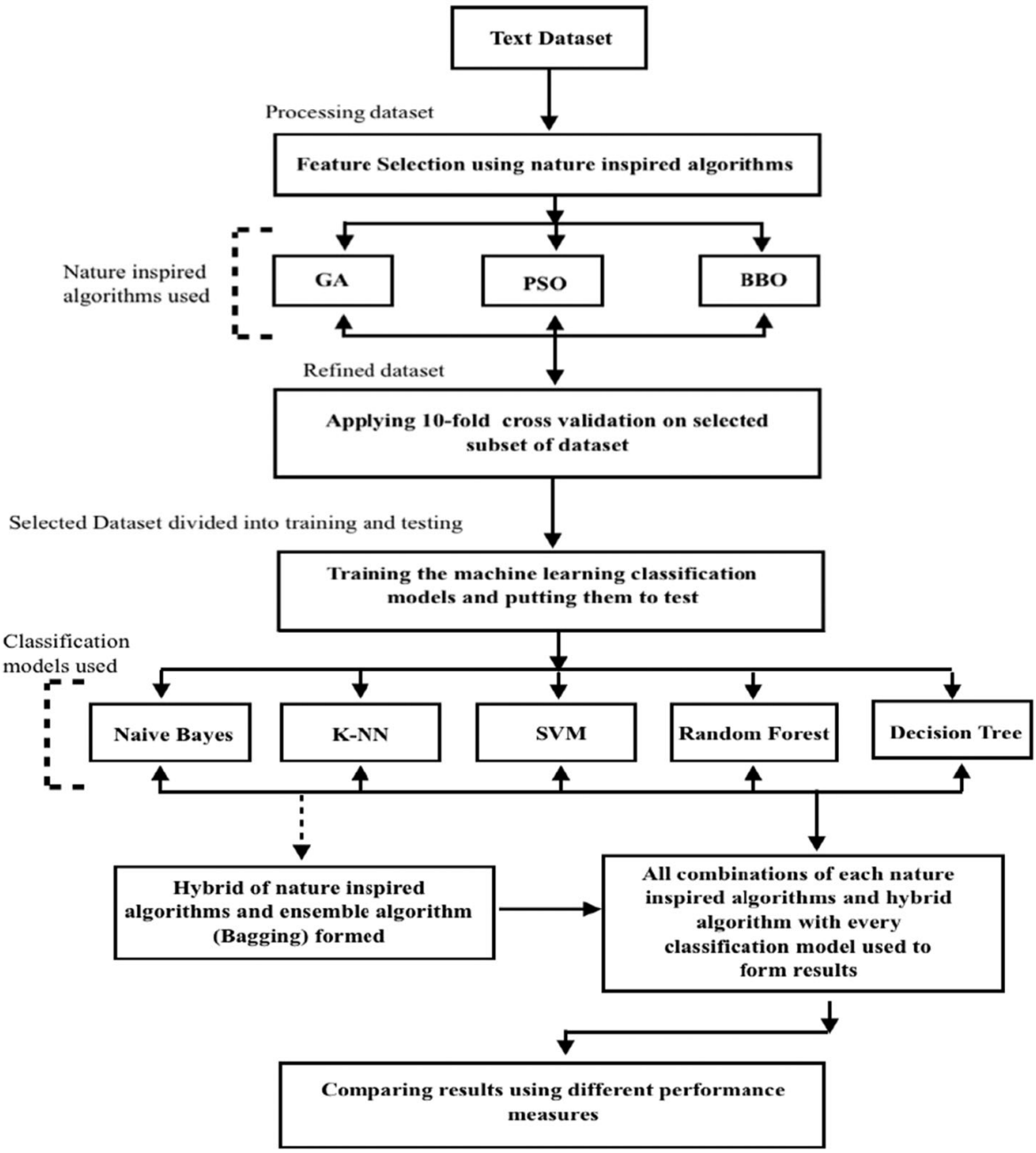


Fig. 7 Flow process of the methodology followed

5 Result

The implementation of the proposed model is performed on Windows 10 64-bit OS, Intel Core i5-5200U CPU @2.20GHz processor. Python3 is selected as a stimulation tool for analysis because of predefined libraries and functions for data analysis, hence providing user-friendly environment. The standard libraries of python namely: numpy (used for objects creation and arrays), pandas (for memory management), Scikit-learn or sklearn which have supervised and unsupervised machine learning algorithms, matplotlib for generating graphs were included. In our method GA, PSO and BBO are used for feature selection. For the proposed model, parameter [number of generations] is set to 100, since no improvement in results was shown after 100 number of iterations.

Input: hab_index[] with reduced feature set: stored in SIV[] from Figure 6
Output: Precision, Recall, Accuracy and F-measure
Result: Predicted model correctness

1. Initialize the parameters foe ensemble classifier [We choose a base learner classifier] (Bagging: ensemble algorithm)
2. D // initially an empty pool of classifiers
3. Let C be the number of classifiers we have to train
4. **for** $j=1,2,..C$ **do**
   From instance space (given by hab_index[]), a bootstrap sample is taken,
   The sample is denoted by BS_j
   // Training phase
   The selected instance space (consist of reduced subset of features) is passed on to the Bagging phase. From the training set BS_j build a classifier, denoted by D_j .
   This classifier is added to the current pool of the classifiers (ensemble), D= D U D_j .

end

5. **for** $j=1,2,..C$ **do**
   Return D.
   //Classification phase
   Run $D_1 D_C$ on the input y (y is the test instance that needs to be labeled).
   The label for y instance is decided depending on which class gets the maximum number of votes.

end

6. Calculate the performance measure based on the correctness of the class predicted by the model

Fig. 8 Hybrid BBO-ensemble Algorithm

The primary objective of our model is to improve the performance of text classification. The NB, k-NN, SVM, RF, DT and Bagging classifiers are used for classification purpose after feature selection. Figure 9, 10, 11 and 12, shows the comparison of experimental values of all performance measures trained and computed on six ML techniques on four datasets with each feature selection technique and where no feature selection techniques are used. Due to the constraint of space, we are unable to show the results of every dataset. There could be observed significant improvement after using feature selection approach when compared with values obtained without feature selection approach. Our approach of using BBO has proven to produce the best results in all dataset with the rise in values of performance measure from 15-20%

On comparing the performance measure values in tr11, tr21, tr31, oh10, oh15, there is a high improvement in all measures when feature selection is done using BBO. Although, there is improvement in results when feature selection is performed with GA and PSO, but significant improvement was shown when feature selection is performed with BBO. In tr12, BBO produces higher performance measures. BBO shows best results with all classifiers except SVM for the precision measure. In tr23, BBO is best with all classifiers except with RF for Accuracy measure. In tr41, tr45 and oh10 precision is not improved when feature selection is performed using GA and compared to without feature selection approach, but feature selection with BBO shows great improvement in results for all performance measures.

In oh0, improvement is shown when values are compared with feature selection and without feature selection. Feature selection with BBO produces better results with all

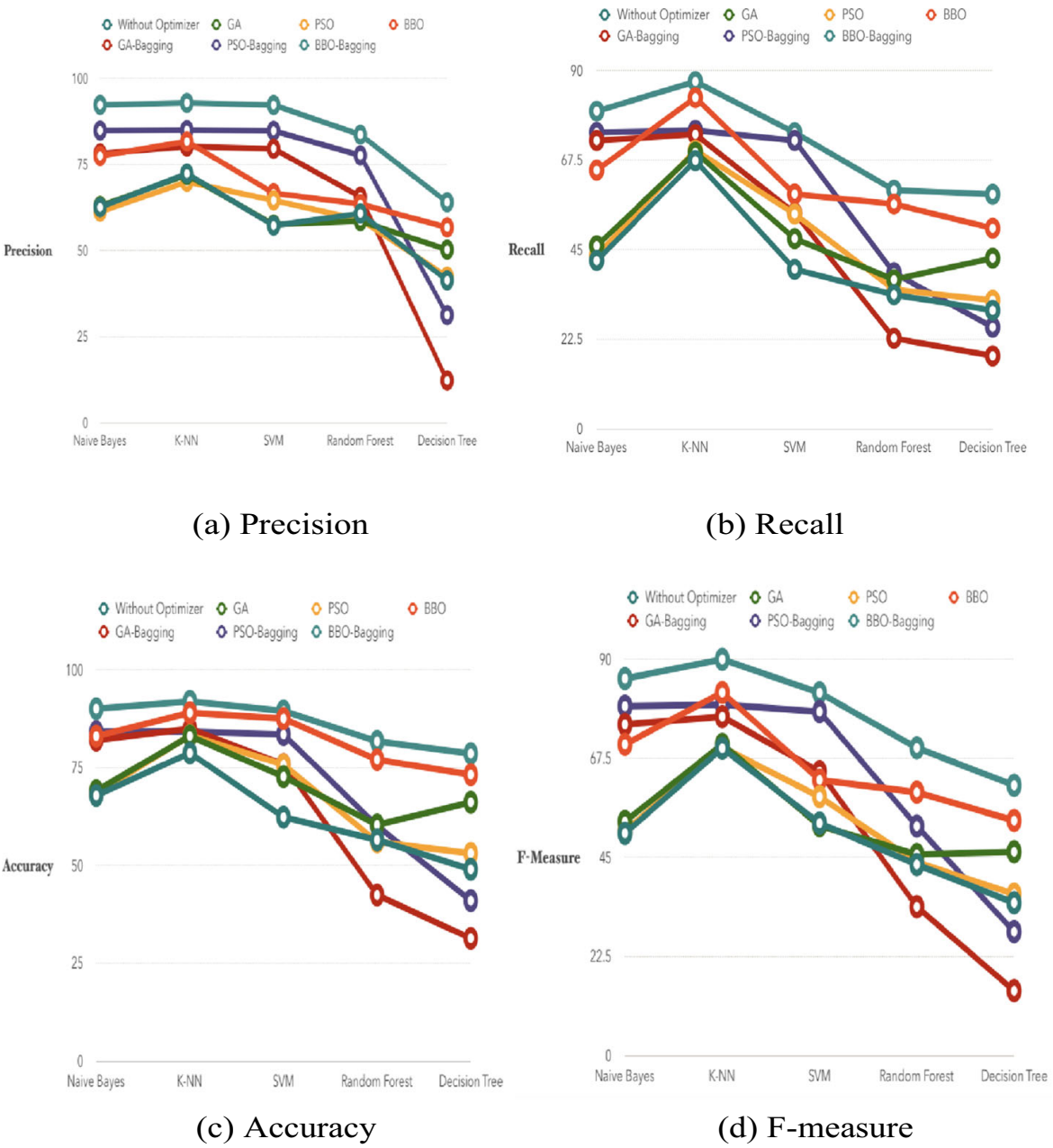


Fig. 9 Performance measure of tr11. (**a**) Precision. (**b**) Recall. (**c**) Accuracy. (**d**) F-measure

performance measures improvement from 5% to 20% except the precision value of SVM. Our model is a hybrid of ensemble technique-Bagging with BBO. The hybrid model is used to test the impact on classification performance, and measure values proved its optimal nature. All performance measures are computed and compared with each other when feature selection is done using GA, PSO and BBO. BBO in all performance measures shown best results with each dataset.

After performing experimental analysis, the proposed model, where feature selection was performed using BBO, produced best results with ensemble classifiers with the base learner as k-NN except for oh10 and oh15. There is a rise in Accuracy from 5% to 20%, a rise in Precision from 5% to 20%, a rise in Recall from 5% to 25% and a rise in F-measure from 5% to 30%. In dataset Tr31, only precision values (Bagging with k-NN) are not good when compared to values of Bagging with Naive Bayes. In oh10, Bagging with SVM gives better results than any other classifier. For the high value of Accuracy and low value of F-measure,

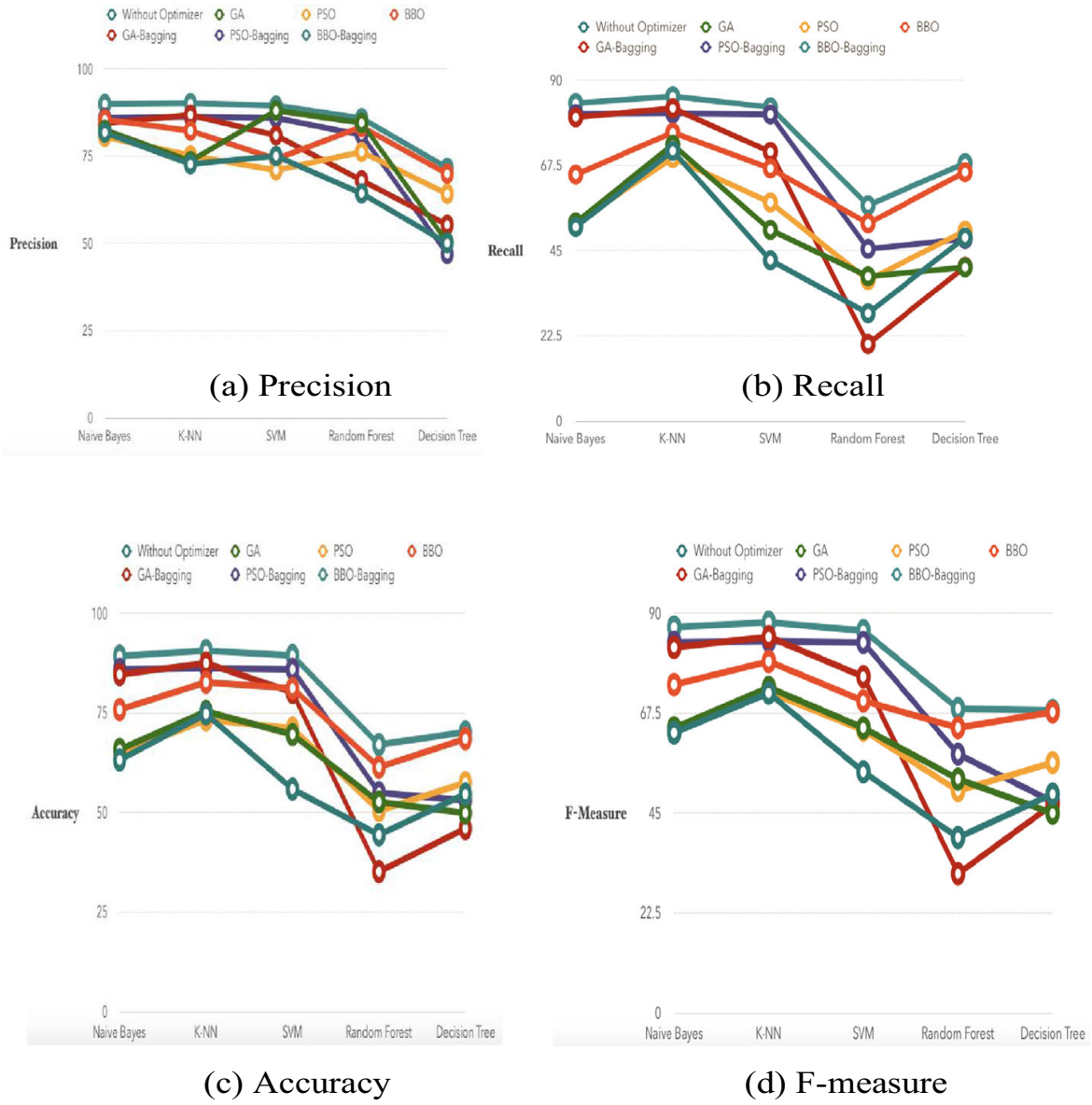


Fig. 10 Performance measure of tr12. (**a**) Precision. (**b**) Recall. (**c**) Accuracy. (**d**) F-measure

there is a class imbalance problem that will be considered for future work. To further validate the experimental results, we applied the Friedman test in Minitab statistical analysis software. The ranking achieved by our technique and other optimization technique is summarized in Table 3. The results are compared at a 95% confidence level. The p-values ($p<0.001$) indicates that the given technique is significantly better than the other technique available in the literature.

5.1 Comparison of different feature selection techniques

The average values of performance metric is given in Table 2 for all the datasets. Average values for all the individual classifiers were computed with feature selection techniques. After computing the empirical results, the statistical analysis shows the order of rank for all the three feature selection algorithms based on average performance values is BBO > PSO > GA. All datasets were validated using 10-fold cross validation method. To verify the results of the techniques, statistical Friedman test is employed.

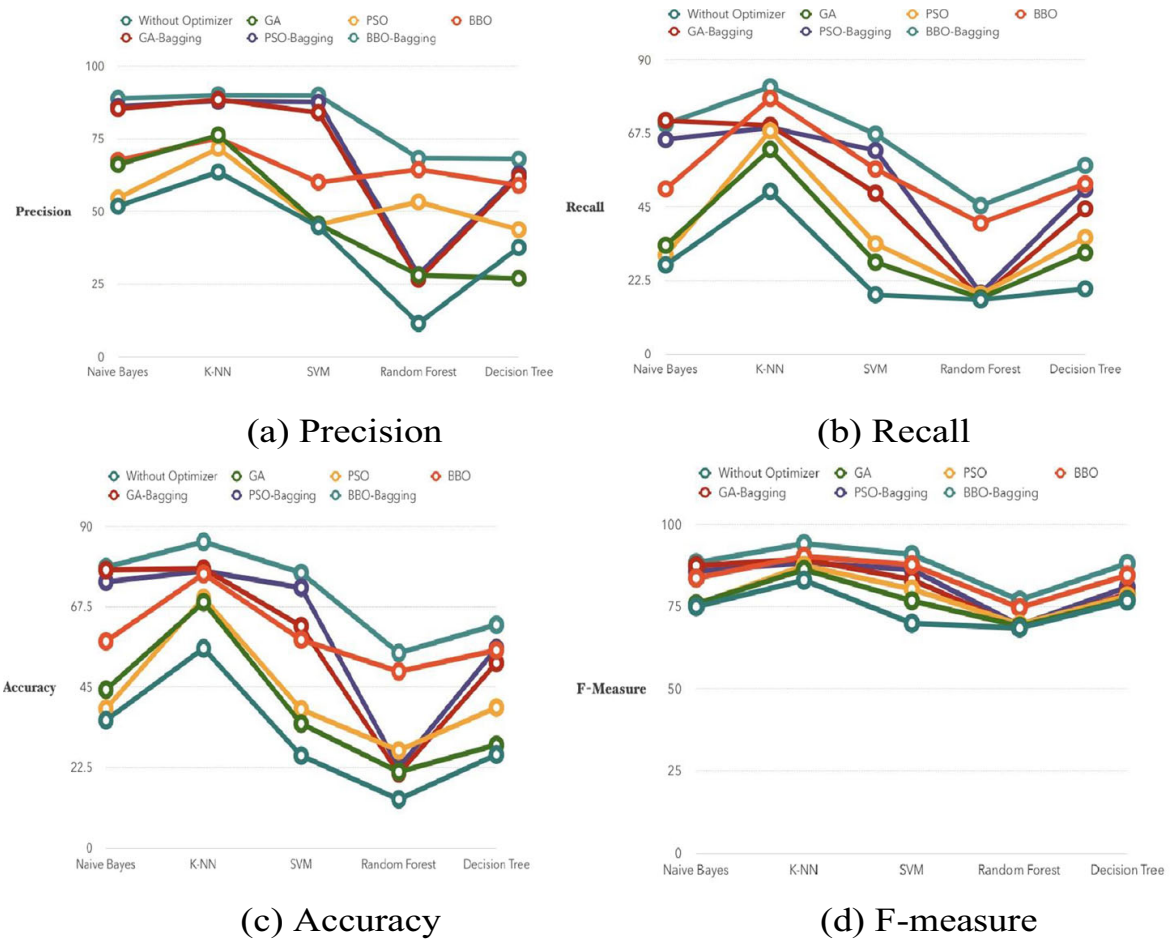


Fig. 11 Performance measure of tr21. (**a**) Precision. (**b**) Recall. (**c**) Accuracy. (**d**) F-measure

According to the hypothesis taken in section 3.5, null hypothesis is rejected, the results are given in Table 3. From the results, it can be predicted that there is a significant difference in techniques used for feature selection i.e. GA vs. PSO vs. BBO (with α=0.01). Feature selection with BBO produce the best results for all the classification techniques. The high convergence rate of BBO and ability to retain previous solutions is one of the reasons behind its superior performance. Due to lack of space we have omitted some data.

5.2 Comparison of different hybrid nature-inspired algorithms and ensemble classifier

Table 4 shows the average performance measure values of all the datasets for Bagging technique with all feature selection techniques. The conclusion can be drawn that the hybrid model of BBO-ensemble is better model than GA-ensemble and PSO-ensemble.

To verify the technique, Friedman test is performed and results obtained are given in Table 5. From experimental results and Friedman test results, it can be concluded that BBO-Bagging>PSO-Bagging>GA-Bagging. After comparing the values in Table 2 and Table 4, it can be seen that the proposed model stands out with a rise of approximately 20% in all the performance metrics. This is possible because of combination of the best converging algorithm, BBO with ensemble classifier (combines advantages of individual classifier), results in an overall superior model with greater advantages and fewer weaknesses.

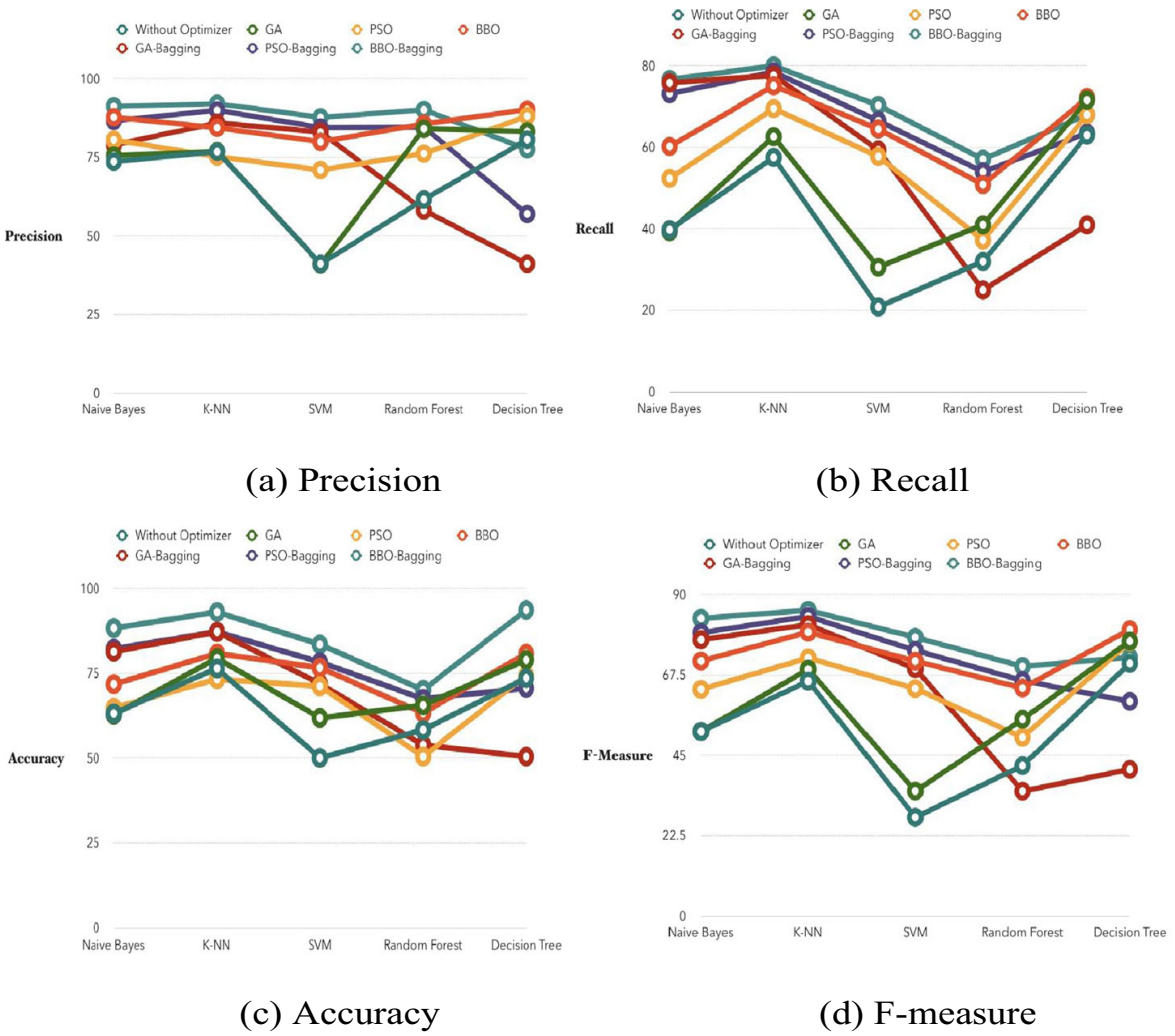


Fig. 12 Performance measure of tr23. (**a**) Precision. (**b**) Recall. (**c**) Accuracy. (**d**) F-measure

Figures 9, 10, 11 and 12, shows the classification performance in terms of Precision, Recall, Accuracy and F-measure of some datasets. Figure 13 shows the results of a real dataset of airlines. From the results it can be concluded that our proposed technique is best for optimal classification. To represent the value obtained by each technique with respective classifier we have used different colors. Colors chosen are dark sea-green (without optimizer), green (GA), yellow (PSO), orange (BBO), red (GA-Bagging), blue (PSO-Bagging) and turquoise (BBO-Bagging)

Fig.9 represents all four performance measures for the dataset tr11. From the Fig. 9, it can be clearly concluded that our model outperformed with a rise in all measures ranging from 15% to 30%. Best values for Precision, Recall, Accuracy and F-measure are 92.89, 87.26, 92 and 89.99 respectively. Our model with base classifier k-NN performed best while worst

Table 2 Average Performance values in percentage for feature selection techniques using individual classifier

Optimizer	Average Precision	Average Recall	Average Accuracy	Average F-measure
GA	61.6	44.71	70	51.81
PSO	62.25	51.97	69.69	56.65
BBO	71.21	61.19	80.05	65.82

Table 3 Friedman Test results for feature selection techniques

Technique	Precision(Ranks)	Recall(Ranks)	Accuracy(Ranks)	F-measure(Ranks)
without optimizer	1.59	1.18	1.26	1.18
GA	1.99	2.2	2.38	2.14
PSO	2.52	2.62	2.42	2.68
BBO	3.9	4	3.94	4

results were obtained with decision tree (as base classifier for Bagging). Base classifiers as NB and SVM for Bagging also obtained near values to k-NN.

Fig. 10 shows the results for dataset tr12. Our model outperformed in all the measures with rise of 15% to 35%. Best results are obtained with k-NN as a base classifier with values 90.21, 85.8, 90.67, and 87.95 for Precision, Recall, Accuracy and F-measure respectively. SVM and NB also performed well near to k-NN for all performance measures. The worst base classifier for this dataset is random forest despite the high precision value of 85.93, as it fails to perform for all other measures.

Fig. 11 depicts the results for dataset tr21. The performance of all measures rises with our model from 10% to 40% for Precision with best value of 90.07. Rise in Recall from 15% to 40% with best value of 81.77. There is a significant rise of Accuracy and F-measure values from 10% to 20% with best values of 94.21 and 85.72 respectively. All best values are obtained using k-NN as base classifier. SVM didn't perform well for all measures for this dataset.

Figure 12 shows the results of all performance measures for dataset tr23. Our model outperformed with base classifier k-NN when compared with other techniques. The best value obtained for each performance measure is 92.11, 80, 93.09 and 85.63 for Precision, Recall, Accuracy and F-measure respectively. NB as a base classifier for Bagging in our model obtained near values to best values. There is a rise of 10% to 40% for each performance measure.

The other datasets also performed best with our model. To reduce the length of the paper, we have not presented the result for other dataset. Dataset tr31, tr41and tr45 performed best with our model (base classifier as k-NN or NB. Oh10 and oh15 showed best results with our model having SVM as base classifier. Oh0 obtained best values when BBO-Bagging combined with base classifiers k-NN or SVM. Hence, after analysis, we recommend using SVM as base classifier to Bagging for low dimension datasets while k-NN or NB works better for high-dimensional dataset as a base classifier to Bagging for our proposed model.

5.3 Comparison with other proposed models

In [10], authors have used GA, Simulated Annealing, Differential Evolution and multi-parent mutation and crossover for optimization and tuning of Naive Bayes on 11 text classification

Table 4 Average Performance values in percentage for feature selection techniques using ensemble classifier

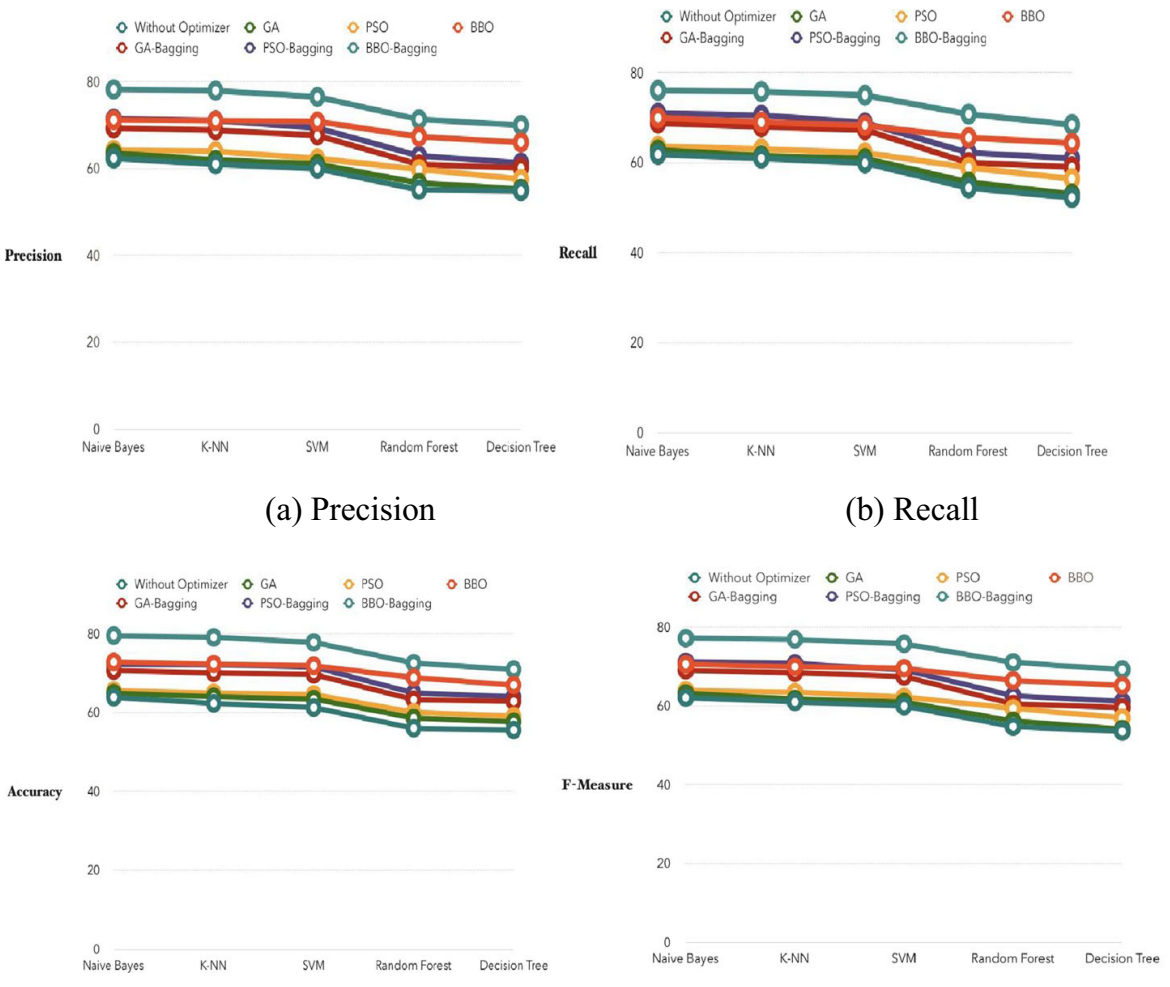
Optimizer	Average Precision	Average Recall	Average Accuracy	Average F-measure
GA-Bagging	69.23	52.61	70.55	59.79
PSO-Bagging	73.55	59.33	75.35	65.68
BBO-Bagging	83.87	70.67	85.16	76.71

Table 5 Friedman Test results for feature selection techniques

Technique	Precision(Ranks)	Recall(Ranks)	Accuracy(Ranks)	F-measure(Ranks)
GA-Bagging	1.1	1.14	1.16	1.08
PSO-Bagging	1.94	1.88	1.86	1.92
BBO-Bagging	2.96	2.98	2.98	3

datasets. Their experimental results shows that their approach of multi-parent mutation and crossover with Naive Bayes is the best approach among other optimization algorithms used for tuning. Our proposed model of optimal text classification has shown better results in all performance measures when compared to other proposed model. The comparison of results are presented in Table 6.

In [26], the study has proposed a novel approach of a hybrid model on deep belief network and softmax regression. They introduced an approach of deep belief network for dimensionality reduction and softmax regression for classification. They have tested their model on two datasets i.e. Reuters- 21,578 and 20-Newsgroups. We have also implemented our model on these two datasets to compare the classification Accuracy and results proved that our model performed better as shown in Table 7. The study [47], proposes four term-weighting methods and calculate the missing terms of weight. The best performance of classification was shown



(a) Precision (b) Recall

(c) Accuracy (d) F-measure

Fig. 13 Performance measure of airlines. (**a**) Precision. (**b**) Recall. (**c**) Accuracy. (**d**) F-measure

Table 6 Comparison of performance measures with other method

Dataset	Precision	Precision	Recall	Recall	Accuracy	Accuracy	F-measure	F-measure
	our approach Diab	our approach Diab [10]	our approach Diab	our approach Diab [10]	our approach Diab	our approach Diab [10]	our approach Diab	our approach Diab [10]
tr11	92.89	85.5	87.26	82.3	92	87.93	89.99	83
tr21	90.07	86.2	81.77	77.3	94.21	84.19	85.72	81
tr31	89.4	83.2	83.23	79	96.89	92.77	83.88	76
tr41	94.33	89.1	83.2	80.3	92.42	91.46	88.42	83.5
tr45	94.33	89	83.13	80.1	92.42	88.70	88.42	85

by SVM classifier. They have used Reuters-21,578 and 20-Newsgroups. We compared the Accuracy of their approach with our approach and our proposed hybrid approach outperforms their approach as shown in Table 7.

In [50], researchers propose a novel approach of the chaotic crow search algorithm to overcome the problem of low convergence rate and trapping in local optima. They have used 20 datasets and we have compared results of our approach on four clinical datasets. Our approach produced much better results than their proposed approach as shown in Table 8.

5.4 Result of proposed approach on a real dataset of airlines

We have perform the testing of our proposed model on a real dataset of airlines [34]. The task of correctly predicting the labels and reducing the feature dimensionality leads to optimal performance for the text classification. Airlines dataset has 539,384 records. The objective was to predict whether the flight is delayed or not, in accordance with other attributes. The result of our proposed model on this dataset is shown in Fig. 13.

The comparison is done with the results obtained from state-of-the-art algorithms of feature selection techniques (with individual classifiers) with our proposed model. It can be observed that the proposed model produces the best classification performance among all other classification models. After comparing the results, we can conclude that our model gives best results for optimal text classification. The best values which we obtained are with base classifier NB. We obtained the values as 79.96, 76.1, 79.55 and 77.16 for Precision, Recall, Accuracy and F-measure respectively. k-NN also showed almost near values to NB. Hence, our model can be applied for large datasets as well.

Table 7 Comparison of classification Accuracy

Dataset	our approach	jiang [26]	Sabbah [47]
20-Newsgroups	92.57	85.57	55
Reuters-21,578	92.11	86.80	88

Table 8 Comparison of classification Accuracy

Dataset	our approach	Sayed [50]
Parkinson's Disease detection dataset(PDD)	94.25	90.78
Single proton emission computed tomography(Spect)	86.2	81.5
Stalog(Heart)	82.6	78.84
Indian Liver patient dataset	89.71	71.68

5.5 Result summary

- Our proposed model performed feature selection with BBO and produces the better classification results when used with machine learning classifiers. All performance values rise from of 5% to 20%.
- In proposed model, we use an ensemble classifier with a feature selection approach using BBO. The model produces better improvement in classification results as compared to individual classifier.
- When experimental values were compared, we found the best results were produced using Bagging with base classifier as k-NN or NB for high dimensional sets and SVM for low dimensional datasets.
- The ensemble classifier-Bagging is paired with other five machine learning classifiers. This results in the enhancement of the classifiers performance in terms of Accuracy, Precision, Recall and F-measure.
- The feature selection method and different classifiers, individual and ensemble, have a different impact on classification results.
- The application on real dataset of airlines signifies that the proposed model can be applied to more real-world practical problems.
- Decision tree and Random forest are not recommended to use with our model as base classifier because of their poor performance.
- All results are validated using 10-fold cross-validation.

6 Conclusion

This research has explored the novel approach of nature-based optimization as feature selection with an ensemble classifier for optimal text classification. The proposed model of BBO and ensemble classifier is compared with other individual models with state-of-the-art algorithms available in the literature on all datasets. Based on the obtained results, the analysis shows that the performance of BBO as feature selection technique is better than GA and PSO. Performance of text classification is measured in terms of Accuracy, Precision, Recall and F-measure. This model is also verified on a real-time dataset of airlines. Our model can be used in practical problems for optimization in text classification. The model is a highly competitive optimal text classification method as compared to other algorithms.

The future work will include more ensemble classifiers with other upcoming enhanced optimization algorithms. Default parameters are considered for all the algorithms.

Parameter tuning can be performed for better results. Imbalanced data is also responsible in some of the datasets for irregular change in Accuracy and F-measure. Hence classification performance can be improved by balancing the dataset.

References

1. Akinyelu, AA, Adewumi, AO (2014): Classification of phishing email using random forest ma- chine learning technique. Journal of Applied Mathematics 2014
2. Balasubramanian, V, Ho, SS, Vovk, V (2014): Conformal prediction for reliable machine learn- ing: theory, adaptations and applications. Newnes
3. Bhattacharya A, Chattopadhyay PK (2010) Solving complex economic load dispatch problems using biogeography-based optimization. Expert Syst Appl 37(5):3605–3615
4. Boynukalin, Z (2012): Emotion analysis of turkish texts by using machine learning methods. Middle East Technical University
5. Catal C, Guldan S (2017) Product review management software based on multiple classifiers. IET Softw 11(3):89–92
6. Da Silva NF, Hruschka ER, Hruschka ER Jr (2014) Tweet sentiment analysis with classifier ensembles. Decis Support Syst 66:170–179
7. Dadaneh BZ, Markid HY, Zakerolhosseini A (2016) Unsupervised probabilistic feature se- lection using ant colony optimization. Expert Syst Appl 53:27–42
8. Das AK, Sengupta S, Bhattacharyya S (2018) A group incremental feature selection for classification using rough set theory based genetic algorithm. Appl Soft Comput 65:400–411
9. Demidova, L, Nikulchev, E, Sokolova, Y (2016): The svm classifier based on the modified particle swarm optimization. arXiv preprint arXiv:1603.08296
10. Diab DM, El Hindi KM (2017) Using differential evolution for fine tuning näıve bayesian classifiers and its application for text classification. Appl Soft Comput 54:183–199
11. El Hindi K (2014) Fine tuning the näıve bayesian learning algorithm. AI Commun 27(2):133–141
12. El Hindi K (2014) A noise tolerant fine tuning algorithm for the näıve bayesian learning al- gorithm. Journal of King Saud University-Computer and Information Sciences 26(2):237–246
13. Emary, E., Zawbaa, H.M. (2016), Hassanien, A.E.: Binary grey wolf optimization approaches for feature selection. Neurocomputing 172, 371–381
14. Eroglu DY, Kilic K (2017) A novel hybrid genetic local search algorithm for feature selection and weighting with an application in strategic decision making in innovation management. Inf Sci 405:18–32
15. Esseghir MA, Goncalves G, Slimani Y (2010) Adaptive particle swarm optimizer for feature selection. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, pp 226–233
16. Fersini E, Messina E, Pozzi FA (2014) Sentiment analysis: Bayesian ensemble learning. De- cision support systems 68:26–38
17. Fister Jr, I., Yang, X.S., Fister, I., Brest, J, Fister, D (2013): A brief review of nature-inspired algorithms for optimization. arXiv preprint arXiv:1307.4186
18. Fong S, Wong R, Vasilakos A (2016) Accelerated pso swarm search feature selection for data stream mining big data. IEEE Trans Serv Comput 1:1–1
19. Ghareb AS, Bakar AA, Hamdan AR (2016) Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Syst Appl 49:31–47
20. Gong W, Cai Z, Ling CX (2010) De/bbo: a hybrid differential evolution with biogeography- based optimization for global numerical optimization. Soft Comput 15(4):645–665
21. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3(Mar): 1157–1182
22. Hafez AI, Zawbaa HM, Emary E, Mahmoud HA, Hassanien AE (2015) An innovative approach for feature selection based on chicken swarm optimization. In: 2015 7th Inter- national Conference of Soft Computing and Pattern Recognition (SoCPaR). IEEE, pp 19–24
23. Hafez AI, Zawbaa HM, Emary E, Hassanien AE (2016) Sine cosine optimization algorithm for feature selection. In: 2016 International Symposium on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, pp 1–5
24. Jiang L, Cai Z, Wang D, Zhang H (2012) Improving tree augmented naive bayes for class probability estimation. Knowl-Based Syst 26:239–245

25. Jiang L, Li C, Wang S, Zhang L (2016) Deep feature weighting for naive bayes and its application to text classification. Eng Appl Artif Intell 52:26–39
26. Jiang M, Liang Y, Feng X, Fan X, Pei Z, Xue Y, Guan R (2018) Text classification based on deep belief network and softmax regression. Neural Comput & Applic 29(1):61–70
27. Kennedy J (2011) Particle swarm optimization. In: Encyclopedia of machine learning. Springer, pp 760–766
28. Li X, Rao Y, Xie H, Lau RYK, Yin J, Wang FL (2017) Bootstrapping social emo- tion classification with semantically rich hybrid neural networks. IEEE Trans Affect Comput 8(4):428–442
29. Liang, HS, Xu, JM, Cheng, YP (2007): An improving text categorization method of na ve bayes. Journal of Hebei University (Natural Science Edition) (3), 24
30. Lin SW, Ying KC, Chen SC, Lee ZJ (2008) Particle swarm optimization for parameter determination and feature selection of support vector machines. Expert Syst Appl 35(4):1817–1824
31. Lin KC, Zhang KY, Huang YH, Hung JC, Yen N (2016) Feature selection based on an improved cat swarm optimization algorithm for big data classification. J Supercomput 72(8):3210–3221
32. Liu Y, Bi JW, Fan ZP (2017) Multi-class sentiment classification: the experimental com- parisons of feature selection and machine learning algorithms. Expert Systems with Ap- plications 80:323–339
33. McCallum A, Nigam K (1998) A comparison of event models for naive bayes text classification. In: et al. (ed) AAAI-98 workshop on learning for text categorization, vol. 752. Citeseer, pp 41–48
34. Mehta, S. (2017), et al.: Concept drift in streaming data classification: algorithms, platforms and issues. Procedia computer science 122, 804–811
35. Melo A, Paulheim H (2019) Local and global feature selection for multilabel classification with binary relevance. Artif Intell Rev 51(1):33–60
36. Mirończuk MM, Protasiewicz J (2018) A recent overview of the state-of-the-art elements of text classification. Expert Syst Appl 106:36–54
37. Moradi P, Gholampour M (2016) A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy. Appl Soft Comput 43:117–130
38. Nag K, Pal NR (2016) A multiobjective genetic programming-based ensemble for simultane- ous feature selection and classification. IEEE transactions on cybernetics 46(2):499–510
39. Onan A, Korukŏglu S, Bulut H (2016) Ensemble of keyword extraction methods and classi- fiers in text classification. Expert Syst Appl 57:232–247
40. Pai PF, Chen CT, Hung YM, Hung WZ, Chang YC (2014) A group decision clas- sifier with particle swarm optimization and decision tree for analyzing achievements in mathematics and science. Neural Comput & Applic 25(7–8):2011–2023
41. Panchal, V, Singh, P, Kaur, N, Kundra, H (2009): Biogeography based satellite image classifi- cation. arXiv preprint arXiv:0912.1009
42. Panchal V, Kundra H, Kaur A (2010) An integrated approach to biogeography based opti- mization with case based reasoning for retrieving groundwater possibility. Int J Comput Appl 1(8):975–8887
43. Pinheiro RH, Cavalcanti GD, Correa RF, Ren TI (2012) A global-ranking local feature selection method for text categorization. Expert Syst Appl 39(17):12851–12857
44. Prabowo R, Thelwall M (2009) Sentiment analysis: a combined approach. Journal of Infor-metrics 3(2): 143–157
45. Rennie JD, Shih L, Teevan J, Karger DR (2003) Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of the 20th international conference on machine learning (ICML-03), pp 616–623
46. Russell SJ, Norvig P (1995) A modern, agent-oriented approach to introductory artificial intelligence. SIGART Bulletin 6(2):24–26
47. Sabbah T, Selamat A, Selamat MH, Al-Anzi FS, Viedma EH, Krejcar O, Fujita H (2017) Modified frequency-based term weighting schemes for text classification. Appl Soft Comput 58:193–206
48. Sayed GI, Darwish A, Hassanien AE, Pan JS (2016) Breast cancer diagnosis approach based on meta-heuristic optimization algorithm inspired by the bubble-net hunting strat- egy of whales. In: International Conference on Genetic and Evolutionary Computing. Springer, pp 306–313
49. Sayed SAF, Nabil E, Badr A (2016) A binary clonal flower pollination algorithm for feature selection. Pattern Recogn Lett 77:21–27
50. Sayed GI, Hassanien AE, Azar AT (2019) Feature selection via a novel chaotic crow search algorithm. Neural Comput & Applic 31(1):171–188
51. Sboev A, Litvinova T, Gudovskikh D, Rybka R, Moloshnikov I (2016) Machine learning models of text categorization by author gender using topic-independent features. Procedia Computer Science 101:135–142
52. Schiezaro M, Pedrini H (2013) Data feature selection based on artificial bee colony algorithm. EURASIP Journal on Image and Video Processing 2013(1):47

53. Silla CN, Freitas AA (2011) A survey of hierarchical classification across different application domains. Data Min Knowl Disc 22(1–2):31–72
54. Simon D (2008) Biogeography-based optimization. IEEE Trans Evol Comput 12(6):702–713
55. Simon D (2011) A probabilistic analysis of a simplified biogeography-based optimization algo- rithm. Evol Comput 19(2):167–188
56. Storn R, Price K (1997) Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359
57. Su CT, Lin HC (2011) Applying electromagnetism-like mechanism for feature selection. In- formation Sciences 181(5):972–986
58. Tang, X., Chen, L. (2018): Artificial bee colony optimization-based weighted extreme learning machine for imbalanced data learning. Cluster Computing pp. 1–16
59. Tang B, He H, Baggenstoss PM, Kay S (2016) A bayesian classification approach using class-specific features for text categorization. IEEE Trans Knowl Data Eng 28(6):1602–1606
60. Uysal AK (2016) An improved global feature selection scheme for text classification. Expert Syst Appl 43: 82–92
61. Uysal AK, Gunal S (2014) Text classification using genetic algorithm oriented latent semantic features. Expert Syst Appl 41(13):5938–5947
62. Wang G, Sun J, Ma J, Xu K, Gu J (2014) Sentiment classification: the contribution of ensemble learning. Decis Support Syst 57:77–93
63. Wang H, Zhou X, Sun H, Yu X, Zhao J, Zhang H, Cui L (2017) Firefly algorithm with adaptive control parameters. Soft Comput 21(17):5091–5102
64. Xia R, Zong C, Li S (2011) Ensemble of feature sets and classification algorithms for senti- ment classification. Inf Sci 181(6):1138–1152
65. Yang B, Zhang Y, Li X (2011) Classifying text streams by keywords using classifier ensemble. Data Knowl Eng 70(9):775–793
66. Ying, L (2007): Analysis on text classification using naive bayes. Computer Knowledge and Technology (Academic Exchange) **11**
67. Zawbaa HM, Emary E, Parv B, Sharawi M (2016) Feature selection approach based on moth-flame optimization algorithm. In: 2016 IEEE Congress on Evolutionary Computation (CEC). IEEE, pp 4612–4617
68. Zhang Y, Gong DW, Cheng J (2017) Multi-objective particle swarm optimization approach for cost-based feature selection in classification. IEEE/ACM Transactions on Computa- tional Biology and Bioinformatics (TCBB) 14(1):64–75
69. Zorarpacı E, Ö Zel SA (2016) a hybrid approach of differential evolution and artificial bee colony for feature selection. Expert Syst Appl 62:91–103

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



Anshu Khurana She is pursuing her doctoral degree in the Department of Computer Science and Engineering from Delhi Technological University. She is working as TRF in the Department of Computer Science and Engineering. She completed her M.Tech in Computer Science and Engineering from M.D.University in 2010. She served as contractual faculty in the Department of Information Technology from Aug, 2010- May, 2013. Her area of interest lies in the area of Nature Based Optimization, Text Classification, Machine Learning, Evolutionary Computing and Soft computing.



Om Prakash Verma He is **p**resently working on diverted capacity from DTU as Principal of G B Pant Engineering College Govt. Of NCT,Delhi (Okhla) since April 2017. He worked as Professor & Head of Department of Computer Science and Engineering from August 2014–April 2017 and of Department of Information and Technology from Nov 2007- April 2017. He held prestigious position of Dean of Continuing Education from Feb 2013- Dec 2016. He has completed his Ph.d from Delhi University and M.Tech from IIT, Delhi. His areas of interest are Image Processing, Soft Computing, Machine learning, Evolutionary Computing. His H-index is 19 and i10-index is 31 as reperted by google scholar. He has more than 100 published papers in reputed journals and international conferences.

arXiv:2006.02111v1 [math.CV] 3 Jun 2020

ON A SPECIAL TYPE OF NON-MA-MINDA FUNCTION

S. SIVAPRASAD KUMAR AND SHAGUN BANGA

Abstract. This paper deals with a special type of non-Ma-Minda function, introduced here with many fascinating facts and interesting applications. It is much akin in all aspects but differs by a condition from its Ma-Minda counterpart. Further, we consider the function: $1-\log(1+z)$, a non-Ma-Minda of the type introduced here, to define a subclass of starlike functions in a similar fashion as we do with Ma-Minda function and is studied for establishing inclusion and radius results. Apart from that, we also deal with the majorization and Bloch function norm problems for the same class. In addition, we obtain the bounds of fourth coefficient: a_4 and second Hankel determinant: $H_2(2)$ for the functions belonging to a newly defined class using convolution, which generalizes many earlier known results and its association with the non-Ma-Minda function is also pointed out.

1. Introduction

Let $\mathcal{A}$ be the set of all normalized analytic functions f of the form $f(z)=z+a_2z^2+a_3z^3+\cdots$ defined on the unit disk $\mathbb{D}=\{z:|z|<1\}$. The subset $\mathcal{S}$ of $\mathcal{A}$ denotes the class of univalent normalized analytic functions. We say, f is subordinate to g, denoted by $f\prec g$, if there exists a Schwarz function ω with $\omega(0)=0$ and $|\omega(z)|<1$ such that $f(z)=g(\omega(z))$, where f and g are analytic functions. Moreover, if g is univalent in $\mathbb{D}$, $f(z)\prec g(z)$ if and only if $f(0)=g(0)$ and $f(\mathbb{D})\subseteq g(\mathbb{D})$. Ma-Minda [9] introduced an analytic univalent function ϕ, called Ma-Minda function, such that $\phi(\mathbb{D})$ has positive real part, symmetric with respect to the real axis, starlike with respect to $\phi(0)=1$ and $\phi'(0)>0$. Let the Taylor series expansion of such $\phi(z)$ be of the form:

$$\phi(z)=1+B_1z+B_2z^2+B_3z^3+\cdots\ (B_1>0). \tag{1.1}$$

Further, Ma and Minda introduced the following classes involving ϕ:

$$\mathcal{S}^*(\phi)=\left\{f\in\mathcal{S}:\frac{zf'(z)}{f(z)}\prec\phi(z)\right\}\ \text{and}\ \mathcal{C}(\phi)=\left\{f\in\mathcal{S}:1+\frac{zf''(z)}{f'(z)}\prec\phi(z)\right\}, \tag{1.2}$$

and obtained distortion and growth theorems for the classes defined in (1.2), in addition to coefficient problems for these classes. Recently, these classes were studied widely for a suitable choice of ϕ. For instance, Aouf et al. [2], studied the class $\mathcal{S}^*(q_c)$ when $\phi(z)=q_c=\sqrt{1+cz}$ $(0<c\leq 1)$, Robertson [16] introduced the class of starlike functions of order alpha $(0\leq\alpha\leq 1)$, denoted by $\mathcal{S}^*(\alpha)$ by opting $\phi(z)$ to be $(1+(1-2\alpha)z)/(1-z)$ and when $\phi(z)=((1+z)/(1-z))^\eta$, $\mathcal{S}^*(\phi)$ reduces to the class of strongly starlike functions of order η, which can be represented in terms of argument as:

$$\mathcal{SS}^*(\eta):=\{f\in\mathcal{A}:|\arg zf'(z)/f(z)|<\eta\pi/2,\ (0<\eta\leq 1)\},$$

introduced in [4]. Consider the class $k-\mathcal{SP}(\alpha,\beta)$ which is introduced in [17], for $k=1$, it reduces to

$$1-\mathcal{SP}(\alpha,\beta):=\{f\in\mathcal{A}:zf'(z)/f(z)\prec\phi(z)\},$$

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

Key words and phrases. Ma-Minda function, Carathéodory coefficients, typically real function, Hankel determinant, radius problems.

where

$$\phi(z) = \alpha + \frac{2(\alpha-\beta)}{\pi^2}\left(\log\frac{1+\sqrt{\nu(z)}}{1-\sqrt{\nu(z)}}\right)^2 \text{ with } \nu(z) = \frac{z+\rho}{1+\rho z},$$

$$\rho = \left(\frac{e^A-1}{e^A+1}\right)^2 \text{ and } A = \sqrt{\frac{1-\alpha}{2(\alpha-\beta)}}\pi.$$

Ever since Ma-Minda introduced the analytic univalent function ϕ with certain conditions, a good many sub-classes of starlike functions emerged out involving it. These classes were further studied, particularly in establishing the inclusion, distortion, and growth theorems, as well as the radius, subordination, and many more related results as seen above. Also recently in [5, 10], the authors studied extensively the class $\mathcal{S}^*(\phi)$ for their chosen ϕ in establishing the radius, inclusion and differential subordination results. Many authors have determined the coefficient bounds for the classes associated with $\phi(z)$ (see [7, 11, 14, 15, 19]). Also, in [3], authors have dealt with the differential subordination results for various classes associated with Ma-Minda function.

Indeed, the governing conditions of Ma-Minda function, when altered or relaxed, would lead to a non-Ma-Minda function, which will be interesting when have some meaningful use. This very idea led us to examine the conditions for alteration and found that $\phi'(0) > 0$, imposed on Ma-Minda functions has a role confined to distortion and growth theorems only. However, it has no influence in establishing the coefficient, radius, inclusion, subordination, and other results for the classes $\mathcal{C}(\phi)$ and $\mathcal{S}^*(\phi)$. This very fact, which is under gloom until now, has been brought to daylight in this paper by considering a non-Ma-Minda function Φ, obtained just by altering only one condition, namely $\Phi'(0) < 0$, which is merely a rotation. We now give the name for the function Φ, a non-Ma-Minda of type **A**, where **A** here signifies the condition $\Phi'(0) < 0$.

Thus a non-Ma-Minda of type **A** function differs from its Ma-Minda counterpart by mere a rotation and is therefore non-typically real, but still, image domain invariant and rest all properties are intact. We now premise the above notion in the following definition:

Definition 1.1. An analytic univalent function Φ defined on the unit disk $\mathbb{D}$ is said to be a *non-Ma-Minda of type **A*** if it has positive real part in $\mathbb{D}$, $\Phi(\mathbb{D})$ is symmetric with respect to the real axis, starlike with respect to $\Phi(0) = 1$ and $\Phi'(0) < 0$. Further, it has a power series expansion of the form:

$$\Phi(z) = 1 + \sum_{n=1}^{\infty} C_n z^n = 1 + C_1 z + C_2 z^2 + \cdots \ (C_1 < 0). \tag{1.3}$$

The class of all such functions of non-Ma-Minda of type **A**, we shall denote it by $\not{\mathcal{M}}_{\mathbf{A}}$, provided $\mathcal{M}$ stands for the class of Ma-Minda functions.

With the advent of $\not{\mathcal{M}}_{\mathbf{A}}$, it is envisaged that a new avenue opens up for further explorations parallel to $\mathcal{M}$ for $\not{\mathcal{M}}_{\mathbf{A}}$. Thus, we can now define the classes $\mathcal{S}^*(\Phi)$ and $\mathcal{C}(\Phi)$ on the similar lines of (1.2). We present below a function which belongs to the class $\not{\mathcal{M}}_{\mathbf{A}}$:

$$\psi(z) := 1 - \log(1+z) = 1 - z + \frac{z^2}{2} - \frac{z^3}{3} + \cdots. \tag{1.4}$$

Note that a function in $\not{\mathcal{M}}_{\mathbf{A}}$ reduces to its Ma-Minda counterpart by mere replacing z by $-z$, that is $\Phi(z) = \phi(-z)$. We list here below a few such examples:

$\mathcal{M}$	$\not{\mathcal{M}}_{\mathbf{A}}$
$\cos\sqrt{-z}$	$\cos\sqrt{z}$
$\sqrt{1+z}$	$\sqrt{1-z}$
$1-\log(1-z)$	$1-\log(1+z)$

Table 1 Ma-Minda type function and its counterpart non-Ma-Minda of type **A** function.

Distortion and Growth Theorems: Let us define the functions in a similar manner as that in [9]: $d_{\Phi n}(n = 1, 2, 3, \cdots)$ by $d_{\Phi n}(0) = d'_{\Phi n}(0) - 1 = 0$ and

$$1 + \frac{z d''_{\Phi n}(z)}{d'_{\Phi n}(z)} = \Phi(z^n), \tag{1.5}$$

which belongs to the class $\mathcal{C}(\Phi)$ and we write $d_{\Phi 1}$ as d_Φ. The structural formula of $d'_{\Phi n}$ is given by:

$$d'_{\Phi n}(z) = \exp \int_0^z \frac{\Phi(t^n) - 1}{t} dt, \tag{1.6}$$

which upon simplification, gives the structural formula of $d_{\Phi n}$. Similarly, we define $t_{\Phi n}(n = 1, 2, 3, \cdots)$ by $t_{\Phi n}(0) = t'_{\Phi n}(0) - 1 = 0$ and

$$\frac{z t'_{\Phi n}(z)}{t_{\Phi n}(z)} = \Phi(z^n),$$

which belongs to the class $\mathcal{S}^*(\Phi)$ and we write $t_{\Phi 1}$ as t_Φ. The structural formula for $t_{\Phi n}$ is given by:

$$t_{\Phi n}(z) = z \exp \int_0^z \frac{\Phi(t^n) - 1}{t} dt. \tag{1.7}$$

Note that $z d'_{\Phi n}(z) = t_{\Phi n}(z)$. Ma-Minda [9] proved the distortion and growth theorems for the classes $\mathcal{C}(\phi)$ and $\mathcal{S}^*(\phi)$ when $\phi \in \mathcal{M}$. Here, we prove that the result does not remain same in the case of functions in $\not{\mathcal{M}}_{\mathbf{A}}$. It is examined with an example and which is further generalized. For this, let us consider the class $\mathcal{C}(\psi)$, the structural formula, given in (1.6) yields:

$$d'_\psi(z) = \exp \sum_{k=1}^{\infty} \frac{(-z)^k}{k^2}.$$

A numerical computation shows that $d'_\psi(1/2) \approx 0.63864$ and $d'_\psi(-1/2) \approx 1.79004$. Let the function f be such that:

$$f'(z) = d'_{\psi 2} = \exp \sum_{k=1}^{\infty} \frac{(-1)^k (z)^{2k}}{2k^2},$$

clearly, $f \in \mathcal{C}(\psi)$. A numerical computation shows that $|f'(1/2)| \approx 0.88874$. Hence

$$d'_\psi(r) \leq |f'(z_0)| \leq d'_\psi(-r), \text{ for } z_0 = r = \frac{1}{2}.$$

Thus functions in $\mathcal{C}(\psi)$ violate distortion theorem, which shows that $\phi'(0) > 0$ is inevitable in obtaining the distortion theorem of [9] for functions in $\mathcal{M}$.

Remark 1.2. Since the images of the unit disk under the functions $\Phi(z)$ and $\phi(z)$ are identical, we have $\mathcal{C}(\Phi) = \mathcal{C}(\phi)$ and $\mathcal{S}^*(\Phi) = \mathcal{S}^*(\phi)$. Therefore to obtain distortion and growth theorems for functions in $\mathcal{C}(\Phi)$ and $\mathcal{S}^*(\Phi)$, where the function Φ belongs to $\not{\mathcal{M}}_{\mathbf{A}}$, it is sufficient to replace $\phi(z)$ by $\Phi(-z)$, where $\phi \in \mathcal{M}$ in the result [9, Corollary 1, p.159].

Theorem 1.3 (Distortion Theorem for $\mathcal{C}(\Phi)$)**.** *Suppose $f \in \mathcal{C}(\Phi)$ and $|z_0| = r < 1$. Then*

$$d'_\Phi(r) \leq |f'(z_0)| \leq d'_\Phi(-r).$$

Equality holds for some non zero z_0 if and only if f is a rotation of d_Φ, given in (1.5).

Proof. Since $\mathcal{C}(\Phi) = \mathcal{C}(\phi)$, where ϕ is in $\mathcal{M}$, from [9, Corollary 1], we get the following

$$d'_\phi(-r) \leq |f'(z_0)| \leq d'_\phi(r).$$

Now, (1.5) yields $d'_\Phi(z) = d'_\phi(-z)$, which establishes the desired result. ∎

Corollary 1.4 (Growth Theorem for $\mathcal{C}(\Phi)$)**.** *Suppose $f \in \mathcal{C}(\Phi)$ and $|z_0| = r < 1$. Then*

$$d_\Phi(r) \leq |f(z_0)| \leq -d_\Phi(-r).$$

Equality holds for some non zero z_0 if and only if f is a rotation of d_Φ, given in (1.5).

From Corollary 1.4, we get the following growth theorem:

Corollary 1.5 (Growth Theorem for $\mathcal{S}^*(\Phi)$)**.** *Suppose $f \in \mathcal{S}^*(\Phi)$ and $|z_0| = r < 1$. Then*

$$t_\Phi(r) \leq |f(z_0)| \leq -t_\Phi(-r).$$

Equality holds for some non zero z_0 if and only if f is a rotation of t_Φ, given in (1.6).

In order to prove distortion theorem for $\mathcal{S}^*(\Phi)$, we additionally assume Φ to be convex in the direction of imaginary axis on the similar approach as in case of [9].

Theorem 1.6 (Distortion Theorem for $\mathcal{S}^*(\Phi)$)**.** *Suppose $f \in \mathcal{S}^*(\Phi)$ and $|z_0| = r < 1$. Then*

$$t'_\Phi(r) \leq |f'(z_0)| \leq t'_\Phi(-r).$$

Equality holds for some non zero z_0 if and only if f is a rotation of t_Φ, given in (1.6).

The proof is much akin to proof of Theorem 1.3 and therefore omitted here.

We come across the following observations, enlisted below, while examining the geometry of a function defined on $\mathbb{D}$ in general, which are of great use in deriving our results:

(1) A function with real coefficients is always symmetric with respect to the real axis, but not conversely, for instance:

$$f_1(z) = iz,\ f_2(z) = 1 + iz,\ f_3(z) = \frac{1+iz}{1-z^2}.$$

The converse holds under special conditions, namely if f is symmetric with respect to the real axis, $f(0) = 0$ and $f'(0)$ be some non zero real number, then the function f has real coefficients. In fact the functions f_1, f_2 and f_3 are symmetric with respect to the real axis but do not have real coefficients as $f_i'(0)$ is not a real number for $(i = 1, 2, 3)$.

(2) Let $f(z)$ be an analytic function with real coefficients and $f(0) = 0$. Then f is typically real if and only if its first coefficient is positive. Thus $\phi'(0) > 0$ implies $\phi - 1$ is typically real whereas the function $\Phi - 1$ is non typically real due to $\Phi'(0) < 0$.

(3) Geometrically, it is evident that the real part of a function attains its maximum/ minimum value on the real line if and only if the function is symmetric with respect to the real axis and convex in the direction of imaginary axis.

We now introduce the following classes involving $\psi(z)$, a non-Ma-Minda of type **A** function:

$$\mathcal{S}^*_{\text{lg}} := \left\{ f \in \mathcal{S} : \frac{zf'(z)}{f(z)} \prec \psi(z) \right\} \text{ and } \mathcal{C}_{\text{lg}} := \left\{ f \in \mathcal{S} : 1 + \frac{zf''(z)}{f'(z)} \prec \psi(z) \right\},$$

where $\psi(z)$ is given in (1.4). By the structural formula (1.7), we get a function $f \in \mathcal{S}^*_{\text{lg}}$ if and only if there exists an analytic function q, satisfying $q(z) \prec \psi(z)$ such that

$$f(z) = z \exp\left(\int_0^z \frac{q(t) - 1}{t} dt \right). \tag{1.8}$$

Now, we give some examples of the functions in the class $\mathcal{S}^*_{\text{lg}}$. For this, let us assume

$$\psi_1(z) = 1 + \frac{z}{6},\ \psi_2(z) = \frac{4+z}{4-z},\ \psi_3(z) = 1 + z \sin\frac{z}{4},\ \psi_4(z) = \frac{8+2z}{8+z}.$$

A geometrical observation leads to $\psi_i(z) \subset \psi(z)$ $(i = 1, 2, 3, 4)$. Thus $\psi_i(z) \prec \psi(z)$. Now, the functions $f_i's$ belonging to the class $\mathcal{S}^*_{\text{lg}}$ corresponding to each of the functions $\psi_i's$ are determined by the structural formula (1.8) as follows:

$$f_1(z) = z\exp\left(\frac{z}{6}\right),\ f_2(z) = \frac{z}{(4-z)^2},\ f_3(z) = z\exp\left(4\left(1-\cos\frac{z}{4}\right)\right),\ f_4(z) = z + \frac{z^2}{8}.$$

In particular, for $q(z) = \psi(z) = 1 - \log(1+z)$, the corresponding function obtained as follows:

$$f_0(z) = z\exp\left(\int_0^z \frac{-\log(1+t)}{t}dt\right) = z - z^2 + \frac{3}{4}z^3 - \frac{19}{36}z^4 + \frac{107}{288}z^5 + \cdots, \tag{1.9}$$

acts as an extremal function in many cases for $\mathcal{S}^*_{\text{lg}}$.

Remark 1.7. The distortion and growth theorems for $\mathcal{C}_{\text{lg}}$ and $\mathcal{S}^*_{\text{lg}}$ can be obtained from that of $\mathcal{C}(\Phi)$ and $\mathcal{S}^*(\Phi)$, respectively, proved in Theorem 1.3-Theorem 1.6.

Here, we establish inclusion results, radius problems, majorization result and estimation of the Bloch function norm for the functions in the class $\mathcal{S}^*_{\text{lg}}$. Let $g(z) = z + \sum_{n=2}^{\infty} g_n z^n$ and $h(z) = z + \sum_{n=2}^{\infty} h_n z^n \in \mathcal{A}$, where $g_n,\ h_n > 0$ with $g_n - h_n > 0$. The convolution (Hadamard Product) of g and h is given by

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

In the coefficient bound section, we consider the class $\mathcal{M}_{g,h}(\phi)$:

$$\mathcal{M}_{g,h}(\phi) = \left\{ f \in \mathcal{A} : \frac{(f*g)(z)}{(f*h)(z)} \prec \phi(z),\ \phi \in \mathcal{M} \right\}. \tag{1.10}$$

This class is defined in [12] and authors have obtained Fekete-Szegö bound for the same. We determine the bounds of fourth coefficient $|a_4|$, second Hankel determinant $|a_2a_4 - a_3^2|$ and the quantity $|a_2a_3 - a_4|$ for the functions in the class $\mathcal{M}_{g,h}(\phi)$. The importance of this class lies in unification of various subclasses of $\mathcal{S}$, discussed in detail in the coefficient section. Some of our results reduce to many earlier known results of Lee et al. [7], Mishra et al. [11] and Singh [18]. In view of (1.10), we also consider the class $\mathcal{M}_{g,h}(\Phi)$ for Φ in $\mathcal{M}_{\mathbf{A}}$. Now, we introduce the class:

$$\mathcal{M}_{\alpha}(\Phi) = \left\{ f \in \mathcal{A} : \frac{zf'(z) + \alpha z^2 f''(z)}{\alpha z f'(z) + (1-\alpha) f(z)} \prec \Phi(z),\ (0 \le \alpha \le 1) \right\}.$$

Note that when $g(z) = (z(1 + (2\alpha - 1)z))/(1-z)^3$ and $h(z) = (z(1 + (\alpha - 1)z))/(1-z)^2$, we have $\mathcal{M}_{g,h}(\Phi) =: \mathcal{M}_{\alpha}(\Phi)$. Further, the power series expansion of g and h, respectively yield

$$g_2 = \ 2(1+\alpha),\ g_3 = 3(1+2\alpha),\ g_4 = 4(1+3\alpha)\ldots \tag{1.11}$$

and

$$h_2 = 1 + \alpha,\ h_3 = 1 + 2\alpha,\ h_4 = 1 + 3\alpha\ldots. \tag{1.12}$$

By setting $\mathcal{M}_{\alpha}(\psi) =: \mathcal{S}_{\text{lg}}(\alpha)$, then for $\alpha = 0$, $\mathcal{S}_{\text{lg}}(\alpha)$ reduces to $\mathcal{S}^*_{\text{lg}}$ and for $\alpha = 1$, it reduces to $\mathcal{C}_{\text{lg}}$. We obtain the sharp bounds of initial coefficients such as a_2, a_3, a_4 and a_5, Fekete-Szegö functional, second Hankel determinant for functions in $\mathcal{S}_{\text{lg}}(\alpha)$. Further, using these sharp bounds, we estimate the third Hankel determinant bound for the functions in $\mathcal{S}_{\text{lg}}(\alpha)$. The below stated lemmas are required in support of our results:

Lemma 1.8. *([9]) Let $p \in \mathcal{P}$ be of the form $1 + \sum_{n=1}^{\infty} p_n z^n$. Then*

$$|p_2 - vp_1^2| \leq \begin{cases} -4v + 2, & v \leq 0; \\ 2, & 0 \leq v \leq 1; \\ 4v - 2, & v \geq 1. \end{cases}$$

When $v < 0$ or $v > 1$, the equality holds if and only if $p(z)$ is $(1+z)/(1-z)$ or one of its rotations. If $0 < v < 1$, then the equality holds if and only if $p(z) = (1+z^2)/(1-z^2)$ or one of its rotations. If $v = 0$, the equality holds if and only if $p(z) = (1+\eta)(1+z)/(2(1-z)) + (1-\eta)(1-z)/(2(1+z))(0 \leq \eta \leq 1)$ or one of its rotations. If $v = 1$, the equality holds if and only if p is the reciprocal of one of the functions such that the equality holds in the case of $v = 0$. Though the above upper bound is sharp for $0 < v < 1$, still it can be improved as follows:

$$|p_2 - vp_1^2| + v|p_1|^2 \leq 2 \quad (0 < v \leq 1/2) \tag{1.13}$$

and

$$|p_2 - vp_1^2| + (1 - v)|p_1|^2 \leq 2 \quad (1/2 \leq v < 1).$$

Lemma 1.9. *([6]) Let $p \in \mathcal{P}$ be of the form $1 + \sum_{n=1}^{\infty} p_n z^n$. Then*

$$2p_2 = p_1^2 + x(4 - p_1^2)$$

and

$$4p_3 = p_1^3 + 2p_1(4 - p_1^2)x - p_1(4 - p_1^2)x^2 + 2(4 - p_1^2)(1 - |x|^2)y$$

for some x and y such that $|x| \leq 1$ and $|y| \leq 1$.

The following result is proved in [8]:

Lemma 1.10. *Let $p \in \mathcal{P}$ with coefficients p_n as above, then*

$$|p_3 - 2p_1p_2 + p_1^3| \leq 2 \tag{1.14}$$

and

$$|p_1^4 - 3p_1^2p_2 + p_2^2 + 2p_1p_3 - p_4| \leq 2. \tag{1.15}$$

2. Radius problems

In addition to a majorization result, this section is chiefly devoted to the estimation of various radius constants associated with $\mathcal{S}^*_{\mathrm{lg}}$. We begin with establishing the following bounds meant for $\mathcal{S}^*_{\mathrm{lg}}$:

Theorem 2.1. *Let $f \in \mathcal{S}^*_{\mathrm{lg}}$. Then we have for $|z| = r < 1$,*

$$1 - \log(1 + r) \leq \operatorname{Re} \frac{zf'(z)}{f(z)} \leq 1 - \log(1 - r) \tag{2.1}$$

and

$$\left| \operatorname{Im} \frac{zf'(z)}{f(z)} \right| \leq \tan^{-1} \left(\frac{r}{\sqrt{1 - r^2}} \right). \tag{2.2}$$

Proof. Since $f \in \mathcal{S}^*_{\mathrm{lg}}$, we have $zf'(z)/f(z) \prec 1 - \log(1+z)$. Thus, by the definition of subordination, we have

$$\frac{zf'(z)}{f(z)} = 1 - \log|q(z)| - i \arg(q(z)), \tag{2.3}$$

where $q(z) = 1 + \omega(z)$, ω be a Schwarz function satisfying $\omega(0) = 0$ and $|\omega(z)| \leq |z|$. Let $q(z) = u + iv$, then $(u-1)^2 + v^2 < 1$. For $|z| = r$, we have

$$|q - 1| \leq r. \tag{2.4}$$

Squaring both sides of the above equation yields

$$T : (u-1)^2 + v^2 \leq r^2. \tag{2.5}$$

Clearly T represents the equation of the disk with center: $(1, 0)$ and radius r, for which the point $(0, 0)$ lies outside the disk T. From (2.4) and (2.5), we have

$$|q(z)| \leq 1 + r \text{ and } |q(z)| \geq 1 - r.$$

Since, $\log x$ is an increasing function on $[1, \infty)$, we have

$$\log(1 - r) \leq \log|q(z)| \leq \log(1 + r).$$

Thus we get the desired result (2.1) by taking the real part in (2.3). If $v = au$, representing the equation of the tangent to the boundary of disk T, which passes through the origin O, then the tangent and the boundary of the disk have a common point, hence from (2.5), we obtain

$$(1 + a^2)u^2 - 2u + 1 - r^2 = 0.$$

By the definition of tangent, we get

$$1 - (1 + a^2)(1 - r^2) = 0, \text{ which yields } a = \pm\frac{r}{\sqrt{1 - r^2}}.$$

Therefore we have

$$\tan^{-1}\left(\frac{-r}{\sqrt{1-r^2}}\right) \leq \arg q(z) \leq \tan^{-1}\left(\frac{r}{\sqrt{1-r^2}}\right).$$

Thus we obtain the desired result (2.2). ∎

Theorem 2.2. *Let $f \in \mathcal{S}^*_{\mathrm{lg}}$. Then the following holds:*

(i) *f is starlike of order α in $|z| < \exp(1-\alpha) - 1$ whenever $1 - \log 2 \leq \alpha < 1$.*
(ii) *$f \in \mathcal{M}(\beta)$ in $|z| < 1 - \exp(1-\beta)$ whenever $\beta > 1$.*
(iii) *f is convex of order α in $|z| < \tilde{r}(\alpha) < 1$ whenever $0 \leq \alpha < 1$, where $\tilde{r}(\alpha)$ is the smallest positive root of the equation:*

$$(1-r)(1-\log(1+r))(1-\log(1+r)-\alpha) - r = 0, \tag{2.6}$$

for the given value of α.
(iv) *f is strongly starlike of order γ in $|z| < r(\gamma)$ whenever $0 < \gamma \leq \gamma_0 \approx 0.514674$, where*

$$r(\gamma) = \sqrt{2\left(1 - \frac{1}{\sqrt{1 + \tan^2\left(\tan\frac{\gamma\pi}{2}\right)}}\right)}. \tag{2.7}$$

(v) *f is $k-$starlike function in $|z| < r(k)$ whenever $k > 0$, where $r(k)$ is the smallest positive root of the equation*

$$1 + r - e(1-r)^k = 0, \tag{2.8}$$

for the given value of k. In particular, for $k = 1$, f is parabolic starlike in $|z| < \frac{e-1}{e+1}$.

Proof. (i) Since $f \in \mathcal{S}^*_{\mathrm{lg}}$, we obtain the following from Theorem 2.1:

$$\mathrm{Re}\,\frac{zf'(z)}{f(z)} \geq 1 - \log(1+r), \quad |z| = r < 1,$$

which yields the inequality

$$\operatorname{Re}\frac{zf'(z)}{f(z)} > \alpha, \quad \text{whenever } 1-\log 2 \leq \alpha < 1,$$

which holds true in the open disk of radius $\exp(1-\alpha)-1$. For the function f_0, given in 1.9 and $z_0 = \exp(1-\alpha)-1$, we have

$$\operatorname{Re}\frac{z_0 f_0'(z)}{f_0(z)} = \alpha.$$

Hence this result is sharp.
(ii) From Theorem 2.1, we get

$$\operatorname{Re}\frac{zf'(z)}{f(z)} \leq 1-\log(1-r), \quad |z| = r < 1,$$

which yields the following inequality

$$\operatorname{Re}\frac{zf'(z)}{f(z)} < \beta, \quad \text{for } \beta > 1,$$

which holds true in the open disk of radius $1-\exp(1-\beta)$. For the function f_0, given in 1.9 and $z_0 = \exp(1-\beta)-1$, we get

$$\operatorname{Re}\frac{z_0 f_0'(z)}{f_0(z)} = \beta.$$

Therefore the result is sharp.
(iii) Let $f \in \mathcal{S}^*_{\lg}$. Now, $f \in \mathcal{C}_{\lg}(\alpha)$, whenever

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

Since we have $zf'(z)/f(z) = 1-\log(1+\omega(z))$, where ω is a Schwarz function, we obtain

$$\operatorname{Re}\left(1+\frac{zf''(z)}{f'(z)}\right) = \operatorname{Re}(1-\log(1+\omega(z))) - \operatorname{Re}\left(\frac{z\omega'(z)}{(1+\omega(z))(1-\log(1+\omega(z)))}\right). \tag{2.9}$$

The function ω satisfies the following inequality, given in [13]

$$|\omega'(z)| \leq \frac{1-|\omega(z)|^2}{1-|z|^2}. \tag{2.10}$$

Using the above result, (2.9) reduces to

$$\operatorname{Re}\left(1+\frac{zf''(z)}{f'(z)}\right) \geq \frac{(1-r)(1-\log(1+r))^2 - r}{(1-r)(1-\log(1+r))} =: \nu(r).$$

To determine $\tilde{r}(\alpha)$, it suffices to choose

$$\tilde{r}(\alpha) = \max\{r \in [0,1) : \nu(r) - \alpha \geq 0\}, \text{ for the given } \alpha,$$

or equivalently,

$$\tilde{r}(\alpha) = \max\{r \in [0,1) : \nu(r,\alpha) \geq 0\}, \text{ for the given } \alpha,$$

where $\nu(r,\alpha) := (1-r)(1-\log(1+r))(1-\log(1+r)-\alpha)-r$. Clearly, $\nu(0,\alpha) > 0$ and $\nu(1,\alpha) < 0$ for all $\alpha \in [0,1)$. Thus, there must exist $\tilde{r}(\alpha)$ such that $\nu(r,\alpha) \geq 0$ for all $r \in [0,\tilde{r}(\alpha)]$, where $\tilde{r}(\alpha)$ is the smallest positive root of the equation (2.6). Hence the result.
(iv) Since $f \in \mathcal{S}^*_{\lg}$, we have

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

Now, $f \in \mathcal{SS}^*(\gamma)$ whenever

$$\left|\arg\left(1-\frac{1}{2}\log((1+x)^2+y^2)-i\arctan\frac{y}{1+x}\right)\right|<\gamma\frac{\pi}{2},$$

for $|z|=\sqrt{x^2+y^2}<1$. Consider the function f_0, given in (1.9) and let us assume

$$z_0=\frac{1}{\sqrt{1+A^2}}-1+i\frac{A}{\sqrt{1+A^2}},$$

where $A=\tan(\tan\gamma\pi/2)$. We have $|z_0|=\sqrt{2(1-1/(\sqrt{1+A^2})}<1$, whenever $\gamma\leq\gamma_0$ and

$$\begin{aligned}|\arg(1-\log(1+z_0))| &= \left|\arg\frac{z_0f_0'(z_0)}{f_0(z_0)}\right|\\ &= |\arg(1-i\arctan(\tan(\tan\gamma\pi/2)))|\\ &= |\arctan(-\arctan(\tan(\tan\gamma\pi/2)))|\\ &= \arctan(\arctan(\tan(\tan\gamma\pi/2)))\\ &= \gamma\pi/2.\end{aligned}$$

A geometrical observation reveals that $f\in\mathcal{SS}^*(\gamma)$ whenever $|z|<r(\gamma)$, where $r(\gamma)$ is given by (2.7) and therefore the result is sharp.
(v) Let $f\in\mathcal{S}^*_{\mathrm{lg}}$. Now, $f\in k-\mathcal{ST}$ whenever

$$\mathrm{Re}(1-\log(1+\omega(z)))>k|\log(1+\omega(z))|\ (z\in\mathbb{D}),$$

for some Schwarz function ω. Let $g(\omega(z)):=|\log(1+\omega(z))|=|\log|1+\omega(z)|+i\arg(1+\omega(z)|$, $\omega(z)=Re^{it}$, where $R\leq|z|=r<1$ and $-\pi<t<\pi$. Now, consider

$$g^2(R,t)=\left(\frac{1}{2}\log(1+R^2+2R\cos t)\right)^2+\left(\arctan\left(\frac{R\sin t}{1+R\cos t}\right)\right)^2,$$

which upon partially differentiation with respect to t, yields

$$h(R,t):=\frac{R\left(2(R+\cos t)\arctan\dfrac{R\sin t}{1+R\cos t}-\log(1+R^2+2R\cos t)\sin t\right)}{1+R^2+2R\cos t}.$$

Then clearly, the function $h(R,t)\geq0$ for $t\in[0,\pi)$ and $h(R,t)\leq0$ for $t\in(-\pi,0]$. Thus

$$\max_{-\pi<t<\pi}g(R,t)=\max\{g(R,-\pi),g(R,\pi)\},$$

which yields

$$|\log(1+\omega(z))|\leq|\log(1-R)|\leq|\log(1-r)|. \tag{2.11}$$

The inequality (2.11) and Theorem 2.1 reveal that the result follows at once by showing

$$1-\log(1+r)\geq k|\log(1-r)|,$$

whenever

$$m(r,k):=1+r-e(1-r)^k\leq0.$$

To determine $r(k)$, we choose

$$r(k):=\max\{r\in[0,1):m(r,k)\leq0\}.$$

Clearly, $m(0,k)<0$ and $m(1,k)>0$ for fixed value of k. Thus, there must exist $r(k)$ such that $m(r,k)\leq0$ for all $r\in[0,r(k)]$, where $r(k)$ is the smallest positive root of the equation (2.8). Hence the result. ∎

Our next result involves the concept of majorization. For any two analytic functions f and g, we say f is majorized by g in $\mathbb{D}$, denoted by $f \ll g$, if there exists an analytic function $\mu(z)$ in $\mathbb{D}$, satisfying

$$|\mu(z)| \leq 1 \text{ and } f(z) = \mu(z)g(z).$$

The following majorization result involves the class $\mathcal{S}^*_{\text{lg}}$:

Theorem 2.3. *Let $f \in \mathcal{A}$. Suppose that $f << g$ in $\mathbb{D}$, where $g \in \mathcal{S}^*_{\text{lg}}$. Then*

$$|f'(z)| \leq |g'(z)|, \text{ for } |z| \leq \tilde{r},$$

where $\tilde{r}$ is the smallest positive root of the following equation:

$$(1-r^2)(1-\log(1+r)) - 2r = 0. \tag{2.12}$$

Proof. Since $g \in \mathcal{S}^*_{\text{lg}}$, we have $\frac{zg'(z)}{g(z)} \prec 1 - \log(1+z)$. Then there exists a Schwarz function $\omega(z)$ such that

$$\frac{zg'(z)}{g(z)} = 1 - \log(1+\omega(z)). \tag{2.13}$$

Let $\omega(z) = Re^{it}$, $R \leq |z| = r < 1$ and $-\pi < t < \pi$. Consider the function $h(R,t)$ given by

$$|1 - \log(1+Re^{it})|^2 = \left(1 - \frac{1}{2}\log(1+R^2+2R\cos t)\right)^2 + \left(\arctan\frac{R\sin t}{1+R\cos t}\right)^2,$$

which upon differentiation with respect to t yields

$$h_t(R,t) = \frac{R\left(2(R+\cos t)\arctan\dfrac{R\sin t}{1+R\cos t} - (-2+\log(1+R^2+2R\cos t)\sin t\right)}{1+R^2+2R\cos t}.$$

Then clearly, the function $h_t(R,t) \geq 0$ for $t \in [0,\pi)$ and $h_t(R,t) \leq 0$ for $t \in (-\pi, 0]$. Thus

$$\min_{-\pi<t<\pi} h(R,t) = h(R,0),$$

which yields

$$|1 - \log(1+\omega(z))| \geq 1 - \log(1+R) \geq 1 - \log(1+r).$$

Further, condition (2.13) yields

$$\left|\frac{g(z)}{g'(z)}\right| = \frac{|z|}{|1-\log(1+\omega(z)|} \leq \frac{r}{1-\log(1+r)}. \tag{2.14}$$

By the definition of majorization, we get $f(z) = \mu(z)g(z)$, which upon differentiation, gives

$$f'(z) = \mu(z)g'(z) + g(z)\mu'(z) = g'(z)\left(\mu(z) + \mu'(z)\frac{g(z)}{g'(z)}\right). \tag{2.15}$$

The function μ satisfies the inequality (2.9), thus using (2.9) for μ and substituting the inequality (2.14) in (2.15), we get

$$|f'(z)| \leq K(r,\zeta)|g'(z)|,$$

where $K(r,\zeta) = \zeta + \frac{r(1-\zeta^2)}{(1-r^2)(1-\log(1+r))}$ for $|\mu(z)| = \zeta$ $(0 \leq \zeta \leq 1)$. Let us choose

$$\tilde{r} := \max\{r \in [0,1) : K(r,\zeta) \leq 1, \ \forall \zeta \in [0,1]\},$$

or equivalently

$$\tilde{r} = \max\{r \in [0,1) : \eta(r,\zeta) := (1-r^2)(1-\log(1+r)) - r(1+\zeta) \geq 0, \ \forall \zeta \in [0,1]\}.$$

For $\zeta = 1$, $\eta(r,\zeta)$ attains its minimum value, which is given by

$$\eta(r,1) =: \nu(r) = (1-r^2)(1-\log(1+r)) - 2r.$$

Clearly, $\nu(0) = 1 > 0$ and $\nu(1) = -2 < 0$. In view of these inequalities there must exist $\tilde{r}$ such that $\nu(r) \geq 0$ for all $r \in [0, \tilde{r}]$, where $\tilde{r}$ is the smallest positive root of the equation (2.12). Hence the proof is complete. ∎

3. Inclusion relations

In this section, we give inclusion relations between the classes $\mathcal{S}^*_{\text{lg}}$ and various other subclasses of starlike functions, namely $\mathcal{S}^*(\alpha)$, $\mathcal{SS}^*(\gamma)$, $\mathcal{ST}(1,\alpha)$ and $\mathcal{S}^*(q_C)$.

Theorem 3.1. *The class $\mathcal{S}^*_{\text{lg}}$ satisfies the following results:*

(i) $\mathcal{S}^*_{\text{lg}} \subset \mathcal{S}^*(\alpha) \subset \mathcal{S}^*$ *for* $0 \leq \alpha \leq 1 - \log 2$.
(ii) $\mathcal{S}^*_{\text{lg}} \subset \mathcal{SS}^*(\gamma) \subset \mathcal{S}^*$ *for* $2\tilde{f}(\theta_0)/\pi \leq \gamma \leq 1$, *where* θ_0 *is the smallest positive root of the equation* $-2 + \log(2(1+\cos\theta)) + \theta \tan\theta/2 = 0$ *and* $\tilde{f}(\theta) = \arg(1 - \log(1 + e^{i\theta}))$, $\theta \in [0, \pi)$.
(iii) $\mathcal{S}^*_{\text{lg}} \subset \mathcal{ST}(1,\alpha)$ *for* $\alpha \leq 1 - 2\log 2$.
(iv) $\mathcal{S}^*(q_C) \subset \mathcal{S}^*_{\text{lg}} \subset \mathcal{S}^*$ *for* $c \leq c_0$, *where* $c_0 = \log 2(2 - \log 2)$.

The above constants in each part is best possible. The pictorial representation of the result is depicted in the ***Figure 1****.*

Proof. (i) Since $f \in \mathcal{S}^*_{\text{lg}}$, we have $zf'(z)/f(z) \prec 1 - \log(1+z)$. Theorem 2.1 yields the following:

$$1 - \log 2 = \min_{|z|=1} \operatorname{Re}(1 - \log(1+z)) < \operatorname{Re} \frac{zf'(z)}{f(z)}.$$

Hence, the result follows.
(ii) Let $f \in \mathcal{S}^*_{\text{lg}}$. Thus, we have

$$\begin{aligned}\left|\arg \frac{zf'(z)}{f(z)}\right| &< \max_{|z|=1} |\arg 1 - \log(1+z)| \\ &= \max_{-\pi\leq\theta\leq\pi} \left|\arctan\left(\frac{-\theta}{2 - \log(2(1+\cos\theta))}\right)\right| \\ &=: \max_{-\pi\leq\theta\leq\pi} |\tilde{f}(e^{i\theta})|.\end{aligned}$$

Due to the symmetricity of the function $\tilde{f}(\theta)$, we consider $\theta \in [0, \pi]$ and $\tilde{f}'(\theta) = 0$ yields $-2 + \log(2(1+\cos(\theta_0))) + \theta_0 \tan(\theta_0/2) = 0$, where $\theta_0 \approx 1.37502$. A calculation shows that $\tilde{f}''(\theta) < 0$, which implies $\max_{0\leq\theta\leq\pi} \tilde{f}(\theta) = \tilde{f}(\theta_0) \approx 0.88329$. Thus $f \in \mathcal{SS}^*(\gamma)$, for the given range of γ.
(iii) Let us consider the domain $\Omega_\alpha := \{w \in \mathbb{C} : \operatorname{Re} w > |w - 1| + \alpha\}$, whose boundary represents a parabola, for $w = x + iy$, given by:

$$x = \frac{y^2}{2(1-\alpha)} + \frac{1+\alpha}{2},$$

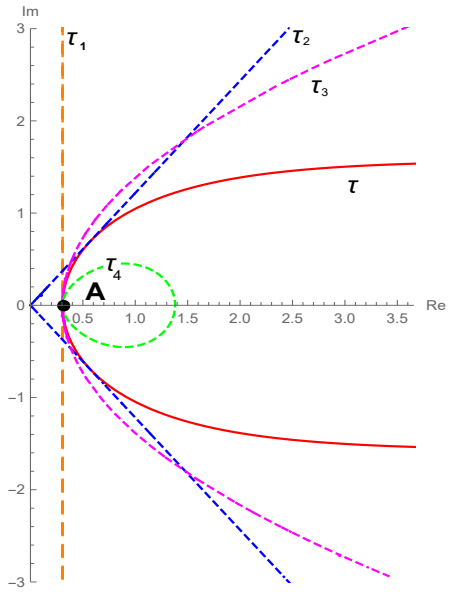
whose vertex is given by: $((1+\alpha)/2, 0)$. In order to prove the result, it suffices to show

$$\begin{aligned}h(\theta) &:= \operatorname{Re}(1 - \log(1+z)) - |\log(1+z)| \\ &= 1 - \frac{1}{2}\log(2(1+\cos\theta)) - \sqrt{\frac{1}{4}\log^2(2(1+\cos\theta)) + \frac{\theta^2}{4}} > \alpha,\end{aligned}$$

for $z = e^{i\theta}$. A numerical computation shows that $\min_{-\pi\leq\theta\leq\pi} h(\theta) = h(0) = 1 - 2\log 2$. Hence, the result.
(iv) Since $f \in \mathcal{S}^*(q_C)$, we have $zf'(z)/f(z) \prec \sqrt{1+cz}$ and

$$\sqrt{1-c} = \min_{|z|=1} \sqrt{1+cz} < \operatorname{Re} \frac{zf'(z)}{f(z)} < \max_{|z|=1} \sqrt{1+cz} = \sqrt{1+c}.$$

Similar analysis can be carried out for the imaginary part bounds and therefore by using Theorem 2.1, we get the result. ∎



$$\tau : |\exp(1-w)-1| = 1$$

$$\tau_1 : \operatorname{Re} w = 1 - \log 2$$

$$\tau_2 : |\arg w| = \frac{7029}{12500}\frac{\pi}{2}$$

$$\tau_3 : \operatorname{Re} w - |w-1| = 1 - 2\log 2$$

$$\tau_4 : |w^2(z) - 1| = \log 2(2 - \log 2)$$

$$A = 1 - \log 2$$

FIGURE 1 Boundary curves of best dominants and subordinant of $\psi(z) = 1 - \log(1+z)$.

4. Coefficient bounds

This section, deals with various coefficient related bound estimates. Here we need the function $H(q_1, q_2)$, given in [1, Lemma 3], to establish our results in what follows. Evidently the class $\mathcal{M}_{g,h}(\phi)$ unifies various subclasses of $\mathcal{S}$ for different choices of g and h. A few of the same are enlisted below for ready reference:

$$\frac{(f*g)(z)}{(f*h)(z)} = \begin{cases} \frac{zf'(z)}{f(z)}, & g(z) = \frac{z}{(1-z)^2},\ h(z) = \frac{z}{1-z}; \\ \frac{zf'(z)+\alpha z^2 f''(z)}{f(z)}, & g(z) = \frac{z(1+(2\alpha-1)z)}{(1-z)^3},\ h(z) = \frac{z}{1-z}; \\ \frac{(zf'(z)+\alpha z^2 f''(z))'}{f'(z)}, & g(z) = \frac{z(1-z^2+2\alpha z(2+z))}{(1-z)^4},\ h(z) = \frac{z}{(1-z)^2}; \\ \frac{2zf'(z)}{f(z)-f(-z)}, & g(z) = \frac{z}{(1-z)^2},\ h(z) = \frac{z}{(1-z^2)}; \\ \frac{(2zf'(z))'}{(f(z)-f(-z))'}, & g(z) = \frac{z(1+z)}{(1-z)^3},\ h(z) = \frac{z(1+z^2)}{(1-z^2)^2}; \\ \frac{zf'(z)+\alpha z^2 f''(z)}{\alpha z f'(z)+(1-\alpha)f(z)}, & g(z) = \frac{z(1+(2\alpha-1)z)}{(1-z)^3},\ h(z) = \frac{z(1+(\alpha-1)z)}{(1-z)^2}. \end{cases} \tag{4.1}$$

Remark 4.1. Murugusundaramoorthy et al. [12, Theorem 6.1] obtained the result of Fekete-Szegö functional bound for functions in the class $\mathcal{M}_{g,h}(\phi)$. Since $\mathcal{M}_{g,h}(\Phi(z)) = \mathcal{M}_{g,h}(\phi(z))$, we state below the parallel result for functions in the class $\mathcal{M}_{g,h}(\Phi(z))$ by simply replacing each B_i by $(-1)^i C_i$, the result is needed to prove our subsequent example.

Theorem 4.2. *For $f \in \mathcal{M}_{g,h}(\Phi)$, we have*

$$|a_3 - ta_2^2| \le \begin{cases} \frac{C_2}{g_3-h_3} - \frac{tC_1^2}{(g_2-h_2)^2} + \frac{(g_2h_2-h_2^2)C_1^2}{(g_3-h_3)(g_2-h_2)^2}, & t \le \kappa_1; \\ \frac{-C_1}{g_3-h_3}, & \kappa_1 \le t \le \kappa_2; \\ \frac{-C_2}{g_3-h_3} + \frac{tC_1^2}{(g_2-h_2)^2} - \frac{(g_2h_2-h_2^2)C_1^2}{(g_3-h_3)(g_2-h_2)^2}, & t \ge \kappa_2, \end{cases}$$

where

$$\kappa_1 = \frac{(g_2-h_2)^2(C_2+C_1)+h_2(g_2-h_2)C_1^2}{(g_3-h_3)C_1^2} \text{ and } \kappa_2 = \frac{(g_2-h_2)^2(C_2-C_1)+h_2(g_2-h_2)C_1^2}{(g_3-h_3)C_1^2}.$$

The result is sharp whenever f satisfies:

$$\frac{(f*g)(z)}{(f*h)(z)} = \begin{cases} \Phi(z), & t < \kappa_1 \text{ or } t > \kappa_2; \\ \Phi(z^2), & \kappa_1 < t < \kappa_2; \\ \Phi(\Psi(z)), & t = \kappa_1; \\ \Phi(-\Psi(z)), & t = \kappa_2, \end{cases}$$

where $\Psi(z) = \dfrac{z(z+\eta)}{1+\eta z} \quad (0 \le \eta \le 1)$.

Remark 4.3. We notice that the bound of Fekete-Szegö stated in [12, Theorem 6.1], namely $|a_3 - \mu a_2^2| \le B_1/2(g_3 - h_3)$ when $\sigma_1 \le \mu \le \sigma_2$, is wrong and should be $|a_3 - \mu a_2^2| \le B_1/(g_3 - h_3)$, which is appropriately corrected in Theorem 4.2.

In the following example, we establish a Fekete-Szegö result for the class $\mathcal{S}_{\mathrm{lg}}(\alpha)$:

Example 1. Let $f \in \mathcal{S}_{\mathrm{lg}}(\alpha)$. Then

$$|a_3 - ta_2^2| \le \begin{cases} \dfrac{3}{4(1+2\alpha)} - \dfrac{t}{(1+\alpha)^2}, & t \le \dfrac{(1+\alpha)^2}{4(1+2\alpha)} =: \kappa_1; \\ \dfrac{1}{2(1+2\alpha)}, & \dfrac{(1+\alpha)^2}{4(1+2\alpha)} \le t \le \dfrac{5(1+\alpha)^2}{4(1+2\alpha)}; \\ \dfrac{t}{(1+\alpha)^2} - \dfrac{3}{4(1+2\alpha)}, & t \ge \dfrac{5(1+\alpha)^2}{4(1+2\alpha)} =: \kappa_2. \end{cases} \tag{4.2}$$

The result is sharp.

Proof. Since $f \in \mathcal{S}_{\mathrm{lg}}(\alpha) = \mathcal{M}_\alpha(\psi(z))$, we have $C_1 = -1$, $C_2 = 1/2$ and $C_3 = -1/3$. The result follows from Theorem 4.2 by substituting the values of $g_i's$ and $h_i's$ from (1.11) and (1.12), respectively. Equality holds whenever f satisfies:

$$\frac{(f*g)(z)}{(f*h)(z)} = \begin{cases} 1 - \log(1+z), & t < \kappa_1 \text{ or } t > \kappa_2; \\ 1 - \log(1+z^2), & \kappa_1 < t < \kappa_2; \\ 1 - \log(1 + \frac{z(z+\eta)}{1+\eta z})), & t = \kappa_1; \\ 1 - \log(1 - \frac{z(z+\eta)}{1+\eta z}), & t = \kappa_2. \end{cases} \tag{4.3}$$

∎

Example 2. Let $f \in \mathcal{S}_{\mathrm{lg}}(\alpha)$. Then

$$\text{(i)} \quad |a_3 - a_2^2| \le \frac{1}{2(1+2\alpha)}, \tag{4.4}$$

$$\text{(ii)} \quad |a_3| \le \frac{3}{4(1+2\alpha)}. \tag{4.5}$$

These results are sharp.

The proof directly follows from Example 1.

Theorem 4.4. *Let $f \in \mathcal{M}_{g,h}(\phi)$ and either*

$$(g_3 - h_3)^2 \le L \tag{4.6}$$

or

$$L < (g_3 - h_3)^2 \le 2L, \tag{4.7}$$

where $L = (g_2 - h_2)(g_4 - h_4)$, *then*

(1)

$$|a_2a_4 - a_3^2| \leq \frac{B_1^2}{(g_3 - h_3)^2},$$

whenever B_1, M *and* T *satisfy the conditions*

$$|M| - B_1^2(g_2 - h_2)^4(g_4 - h_4) \leq 0 \tag{4.8}$$

and

$$|T| + B_1(g_3 - h_3)^2(g_2 - h_2) - 2B_1(g_2 - h_2)^2(g_4 - h_4) \leq 0. \tag{4.9}$$

(2)

$$|a_2a_4 - a_3^2| \leq \frac{|M|}{(g_2 - h_2)^4(g_3 - h_3)^2(g_4 - h_4)},$$

whenever B_1, M *and* T *satisfy the conditions*

$$|T| + B_1(g_3 - h_3)^2(g_2 - h_2) - 2B_1(g_2 - h_2)^2(g_4 - h_4) \geq 0$$

and

$$2|M| - B_1|T|(g_2 - h_2)^2 - B_1^2(g_3 - h_3)^2(g_2 - h_2)^3 \geq 0$$

or

$$|T| + B_1(g_3 - h_3)^2(g_2 - h_2) - 2B_1(g_2 - h_2)^2(g_4 - h_4) \leq 0$$

and

$$|M| - B_1^2(g_2 - h_2)^4(g_4 - h_4) \geq 0.$$

(3)

$$|a_2a_4 - a_3^2| \leq \frac{B_1^2}{(g_3-h_3)^2} - \frac{1}{(|M|-B_1|T|(g_2-h_2)^2-B_1^2(g_3-h_3)^2(g_2-h_2)^3+B_1^2(g_2-h_2)^4(g_4-h_4))} \times \frac{B_1^2(|T|+B_1(g_3-h_3)^2(g_2-h_2)-2B_1(g_2-h_2)^2(g_4-h_4))^2}{4(g_3-h_3)^2(g_4-h_4)},$$

whenever B_1, M *and* T *satisfy the conditions*

$$|T| + B_1(g_3 - h_3)^2(g_2 - h_2) - 2B_1(g_2 - h_2)^2(g_4 - h_4) > 0 \tag{4.10}$$

and

$$2|M| - B_1|T|(g_2 - h_2)^2 - B_1^2(g_3 - h_3)^2(g_2 - h_2)^3 \leq 0, \tag{4.11}$$

where

$$\begin{aligned} M = \; & B_1^4\bigg(- h_2^2(g_2 - h_2)^2(g_4 - h_4) + (g_3 - h_3)\bigg(g_2g_3h_2^2 - g_3h_2^3 + g_2^2h_2h_3 - 3g_2h_2^2h_3 + 2h_2^3h_3 \\ & + (g_3 - h_3)(-g_2h_2^2 + h_2^3)\bigg)\bigg) + B_1B_3(g_3 - h_3)^2(g_2 - h_2)^3 - B_2^2(g_2 - h_2)^4(g_4 - h_4) \\ & + B_1^2B_2\bigg((g_3 - h_3)(g_2 - h_2)^2\Big(g_3h_2 + g_2h_3 - 2h_2h_3 - 2h_2(g_2 - h_2)(g_4 - h_4)\Big)\bigg) \end{aligned} \tag{4.12}$$

and

$$\begin{aligned} T = \; & 2B_2(g_2 - h_2)^2(g_4 - h_4) + 2B_1^2h_2(g_2 - h_2)(g_4 - h_4) - B_1^2g_3h_2(g_3 - h_3) \\ & - B_1^2g_2h_3(g_3 - h_3) + 2B_1^2h_2h_3(g_3 - h_3) - 2B_2(g_3 - h_3)^2(g_2 - h_2). \end{aligned} \tag{4.13}$$

Proof. The series expansion of the functions f, g and h yields

$$\frac{(f*g)(z)}{(f*h)(z)} = 1 + a_2(g_2-h_2)z + (a_3(g_3-h_3) + a_2^2h_2(h_2-g_2))z^2 + \cdots . \tag{4.14}$$

Here, we define a function p in $\mathcal{P}$ as follows:

$$p(z) = \frac{1+\omega(z)}{1-\omega(z)} = 1 + p_1z + p_zz^2 + \cdots .$$

Then, we have

$$\omega(z) = \frac{p(z)-1}{p(z)+1}, \tag{4.15}$$

clearly, ω is a Schwarz function. Since $(f*g)(z)/(f*h)(z) \prec \phi(z)$, we get

$$\frac{(f*g)(z)}{(f*h)(z)} = \phi(\omega(z)). \tag{4.16}$$

Now, using (4.14), (4.15) and (1.1) in (4.16), we get

$$a_2 = \frac{B_1p_1}{2(g_2-h_2)},\ a_3 = \frac{B_2p_1^2(g_2-h_2) - B_1(p_1^2-2p_2)(g_2-h_2) + B_1^2p_1^2h_2}{4(g_2-h_2)(g_3-h_3)} \tag{4.17}$$

and

$$\begin{aligned} a_4 = \frac{1}{8(g_2-h_2)(g_3-h_3)(g_4-h_4)}\Bigg(& p_1(-2B_2p_1^2 + B_3p_1^2 + 4B_2p_2)(g_2-h_2)(g_3-h_3) \\ &+ B_1^3p_1^3h_2h_3 - B_1^2p_1(p_1^2-2p_2)\Big(g_3h_2 + (g_2-2h_2)h_3\Big) + B_1\Bigg(p_1^3\Big(g_2(g_3+(B_2-1)h_3) \\ &+ h_2\big((B_2-1)g_3 + h_3 - 2B_2h_3\big)\Big) - 4p_1p_2(g_2-h_2)(g_3-h_3) + 4p_3(g_2-h_2)(g_3-h_3)\Bigg)\Bigg). \end{aligned} \tag{4.18}$$

We assume $p_1 =: p \in [0,2]$ and upon substituting the values of p_2 and p_3, given in Lemma 1.9, in the expression $a_2a_4 - a_3^2$, we get

$$\begin{aligned} a_2a_4 - a_3^2 = \frac{1}{16(g_2-h_2)^4(g_3-h_3)^2(g_4-h_4)}\Bigg(& p^4M - p^2x(4-p^2)(g_2-h_2)^2B_1T \\ &- p^2x^2(4-p^2)B_1^2(g_3-h_3)^2(g_2-h_2)^3 - (4-p^2)^2x^2B_1^2(g_2-h_2)^4(g_4-h_4) \\ &+ 2B_1^2p(4-p^2)(g_3-h_3)^2(g_2-h_2)^3z(1-|x|^2)\Bigg), \end{aligned}$$

where M and T are given in (4.12) and (4.13), respectively. Applying triangular inequality in the above equation with the assumption that $|x| = \rho$, we get

$$\begin{aligned} |a_2a_4 - a_3^2| \le \frac{1}{16(g_2-h_2)^4(g_3-h_3)^2(g_4-h_4)}\Bigg(& |M|p^4 + p^2(4-p^2)\rho^2B_1^2(g_3-h_3)^2(g_2-h_2)^3 \\ &+ 2B_1^2(g_3-h_3)^2(g_2-h_2)^3p(4-p^2)(1-\rho^2) + B_1|T|p^2\rho(4-p^2)(g_2-h_2)^2 \\ &+ B_1^2(g_2-h_2)^4(g_4-h_4)(4-p^2)^2\rho^2\Bigg) =: G(p,\rho). \end{aligned}$$

The function $G(p,\rho)$ is an increasing function of ρ in the closed interval $[0,1]$, when (4.6) or (4.7) holds. Thus $\max_{0\le\rho\le1} G(p,\rho) = G(p,1) =: F(p)$. Further, a computation yields

$$\begin{aligned} F(p) &= \frac{1}{16(g_2-h_2)^4(g_3-h_3)^2(g_4-h_4)}\Bigg(\Big(|M| - B_1(g_2-h_2)^2\Big(B_1(g_2-h_2)^2(g_4-h_4)x - |T| \\ &\quad - B_1(g_2-h_2)(g_3-h_3)^2\Big)\Big)p^4 + 4B_1(g_2-h_2)^2\Big(|T| + B_1(g_2-h_2)(g_3-h_3)^2 \\ &\quad - 2B_1(g_2-h_2)^2(g_4-h_4)\Big)p^2 + 16B_1^2(g_2-h_2)^4(g_4-h_4)\Bigg) \\ &=: \frac{1}{16(g_2-h_2)^4(g_3-h_3)^2(g_4-h_4)}\Big(Ap^4 + Bp^2 + C\Big). \end{aligned} \tag{4.19}$$

We recall that

$$\max_{0\le t\le4}(At^2+Bt+C) = \begin{cases} C, & B\le0, A\le-\frac{B}{4}; \\ 16A+4B+C, & B\ge0,\ A\ge-\frac{B}{8} \quad \text{or} \quad B\le0,\ A\ge-\frac{B}{4}; \\ \dfrac{4AC-B^2}{4A}, & B>0,\ A\le-\frac{B}{8}. \end{cases} \tag{4.20}$$

From (4.19) and (4.20), we get the desired result. ∎

Remark 4.5. In view of the first case of (4.1), Theorem 4.4 reduces to the result obtained by Lee et al. [7] which gives the sharp second Hankel determinant bound for the class of starlike functions.

Remark 4.6. We notice that the bound evaluated in [7, Theorem 1] for the case 3, reproduced below:

$$|a_2a_4 - a_3^2| \le \frac{B_1^2}{12}\left(3|4B_1B_3 - B_1^4 - 3B_2^2| - 4B_1|B_2| + 4B_1^2 - |B_2|^2\right),$$

has an error and it should be:

$$|a_2a_4 - a_3^2| \le \frac{B_1^2}{12}\left(3|4B_1B_3 - B_1^4 - 3B_2^2| - 4B_1|B_2| - 4B_1^2 - |B_2|^2\right).$$

Remark 4.7. In view of the sixth case of (4.1) for, $\phi(z) = \frac{1+z}{1-z}$, Theorem 4.4 reduces to a result obtained in [18].

Remark 4.8. The second Hankel determinant bound for the functions in the class $M_{g,h}(\Phi)$ can be obtained from Theorem 4.4 by replacing each B_i by $(-1)^iC_i$.

Example 3. Let $f\in\mathcal{S}_{\mathrm{lg}}(\alpha)$. Then

$$|a_2a_4 - a_3^2| \le \begin{cases} \dfrac{1}{4(1+2\alpha)^2}, & 0\le\alpha\le\frac{2+\sqrt{15}}{11}; \\ \dfrac{31\alpha^4+136\alpha^3-14\alpha^2-24\alpha-3}{2(61\alpha^2-20\alpha-5)(1+\alpha)(1+3\alpha)(1+2\alpha)^2}, & \frac{2+\sqrt{15}}{11}\le\alpha\le1. \end{cases}$$

Proof. For $f\in\mathcal{M}_{g,h}(\psi)$, we have $C_1=-1$, $C_2=1/2$ and $C_3=-1/3$. Now, $\mathcal{M}_\alpha(\phi(z)) = \mathcal{M}_\alpha(\phi(-z)) = \mathcal{M}_\alpha(\Phi(z))$. Thus using Theorem 4.4 and Remark 4.8, we get

$$M = \frac{1}{12}(1+\alpha)^3(61\alpha^2-20\alpha-5) \text{ and } T = -(1+5a+17\alpha^2+13\alpha^3).$$

To get the desired estimate, we now consider the following cases:

(i) For $0\le\alpha\le(2+\sqrt{15})/11$, it is easy to verify that M and T satisfy the inequalities (4.8) and (4.9), respectively.

(ii) For $(2+\sqrt{15})/11 < \alpha \leq 1$, it is easy to verify that the inequalities (4.10) and (4.11) hold true for M and T.

Now, the assertion follows at once from Theorem 4.4. ∎

We recall that the function $f \in \mathcal{S}$ is in the class $\mathcal{S}_s^*(\phi)$ if it satisfies

$$\frac{2zf'(z)}{f(z)-f(-z)} \prec \phi(z),\ z \in \mathbb{D}$$

and is in the class $\mathcal{C}_s(\phi)$ if it satisfies

$$\frac{(2zf'(z))'}{(f(z)-f(-z))'} \prec \phi(z),\ z \in \mathbb{D}.$$

The following couple of corollaries can be obtained from Theorem 4.4, in view of fourth and fifth cases of (4.1) respectively.

Corollary 4.9. *Let $f \in \mathcal{S}_s^*(\phi)$. Then we have*

$$|a_2a_4 - a_3^2| \leq \begin{cases} B_1^2/4, & \text{when } A \text{ holds;} \\ |M|/256, & \text{when } B \text{ holds;} \\ \dfrac{B_1^2}{4} - \dfrac{B_1^2(|T|-24B_1)^2}{64(|M|-4B_1|T|+32B_1^2)}, & \text{when } C \text{ holds,} \end{cases}$$

where $M = 16B_1^2B_2 - 64B_2^2 + 32B_1B_3$, $T = 16B_2 - 4B_1^2$ and

$$\begin{aligned} A &: |M| - 64B_1^2 \leq 0 \text{ and } |T| - 24B_1 \leq 0. \\ B &: |M| - 2B_1|T| - 16B_1^2 \geq 0 \text{ and } |T| - 24B_1 \geq 0, \text{ or} \\ &\quad |M| - 64B_1^2 \geq 0 \text{ and } |T| - 24B_1 \leq 0. \\ C &: |M| - 2B_1|T| - 16B_1^2 \leq 0 \text{ and } |T| - 24B_1 > 0. \end{aligned}$$

Corollary 4.10. *Let $f \in \mathcal{C}_s(\phi)$. Then we have*

$$|a_2a_4 - a_3^2| \leq \begin{cases} B_1^2/36, & \text{when } A \text{ holds;} \\ |M|/147456, & \text{when } B \text{ holds;} \\ \dfrac{B_1^2}{36} - \dfrac{B_1^2(|T|-368B_1)^2}{2304(|M|-16B_1|T|+1792B_1^2)}, & \text{when } C \text{ holds,} \end{cases}$$

where $M = 128(9B_1^2B_2 - 32B_2^2 + 18B_1B_3)$, $T = 8(28B_2 - 9B_1^2)$ and,

$$\begin{aligned} A &: |M| - 4096B_1^2 \leq 0 \text{ and } |T| - 368B_1 \leq 0. \\ B &: |M| - 8B_1|T| - 1152B_1^2 \geq 0 \text{ and } |T| - 368B_1 \geq 0, \text{ or} \\ &\quad |T| - 368B_1 \leq 0 \text{ and } |M| - 4096B_1^2 \geq 0. \\ C &: |T| - 368B_1 > 0 \text{ and } |M| - 8B_1|T| - 1152B_1^2 \leq 0. \end{aligned}$$

Remark 4.11. When $\phi(z) = \frac{1+z}{1-z}$, the Corollaries 4.9 and 4.10 reduce to the results obtained in [11] for the classes $\mathcal{S}_s^*(\phi)$ and $\mathcal{C}_s(\phi)$ of starlike functions and convex functions with respect to symmetric points respectively.

Remark 4.12. Note that the second Hankel determinant bound for the functions in the classes $\mathcal{S}_s^*(\Phi)$ and $\mathcal{C}_s(\Phi)$ can be obtained from the Corollaries 4.9 and 4.10, respectively by replacing each B_i by $(-1)^iC_i$.

Expressing the fourth coefficient a_4 for the function $f \in \mathcal{M}_{g,h}(\phi)$ in terms of the Schwarz function $\omega(z) = 1 + \omega_1 z + \omega_2 z^2 + \cdots$, we obtain the bound of a_4 as follows:

$$|a_4| \leq \frac{B_1}{g_4 - h_4} H(q_1, q_2), \tag{4.21}$$

where

$$q_1 = \tfrac{1}{B_1(g_2-h_2)(g_3-h_3)}\Big(2B_2(g_2 - h_2)(g_3 - h_3) + B_1^2(g_3h_2 + g_2h_3 - 2h_2h_3)\Big) \tag{4.22}$$

and

$$q_2 = \tfrac{1}{B_1(g_2-h_2)(g_3-h_3)}\Big(B_3(g_2 - h_2)(g_3 - h_3) + B_1^3h_2h_3 + B_1B_2(g_3h_2 + g_2h_3 - 2h_2h_3)\Big). \tag{4.23}$$

Remark 4.13. In view of first case of (4.1), for $\phi(z) = \sqrt{1+z}$, the above result (4.21) reduces to the result obtained in [15].

Remark 4.14. Note that the bound for the fourth coefficient for the functions in the class $\mathcal{M}_{g,h}(\Phi)$ can be obtained from (4.21), (4.22) and (4.23) by replacing each B_i by $(-1)^i C_i$.

Example 4. Let $f \in \mathcal{S}_{\lg}(\alpha)$, then

$$|a_4| \leq \frac{19}{36(1+3\alpha)}.$$

The result is sharp.

Proof. For $f \in \mathcal{M}_{g,h}(\psi)$, we have $C_1 = -1$, $C_2 = 1/2$, $C_3 = -1/3$. Equations (4.22), (4.23) and Remark 4.14 yield $q_1 = -5/2$ and $q_2 = 19/12$. The result follows from (4.21) and extremal functions f, up to rotations can be obtained when f satisfies

$$\frac{zf'(z) + \alpha z^2 f''(z)}{\alpha z f'(z) + (1-\alpha) f(z)} = 1 - \log(1+z).$$

This completes the proof. ∎

Expressing the expression $a_2a_3 - a_4$ for the function $f \in \mathcal{M}_{g,h}(\phi)$ in terms of the Schwarz function $\omega(z) = 1 + \omega_1 z + \omega_2 z^2 + \cdots$, we obtain the bound as follows:

$$|a_2a_3 - a_4| \leq \frac{B_1}{g_4 - h_4} H(q_1, q_2),$$

where

$$q_1 = \frac{2B_2(g_2 - h_2)^2(g_3 - h_3) + B_1^2(g_2 - h_2)(-g_4 + g_3h_2 + g_2h_3 - 2h_2h_3 + h_4)}{B_1(g_2 - h_2)^2(g_3 - h_3)} \tag{4.24}$$

and

$$\begin{aligned} q_2 =& \frac{1}{B_1(g_2 - h_2)^2(g_3 - h_3)}\Big(B_3(g_2 - h_2)^2(g_3 - h_3) + B_1B_2(g_2 - h_2)(-g_4 + g_3h_2 + g_2h_3 \\ & - 2h_2h_3 + h_4) + B_1^3h_2(-g_4 + g_2h_3 - h_2h_3 + h_4)\Big). \end{aligned} \tag{4.25}$$

Example 5. Let $f \in \mathcal{S}_{\lg}(\alpha)$. Then

$$|a_2a_3 - a_4| \leq \frac{1}{3(1+3\alpha)}.$$

The result is sharp.

Proof. Here, we have $\Phi(z) = 1 - \log(1+z)$, $C_1 = -1$, $C_2 = 1/2$ and $C_3 = -1/3$. Upon replacing each B_i by $(-1)^i C_i$ in (4.24) and (4.25), we get

$$q_1 = -\frac{5\alpha^2 + 3\alpha + 1}{(1+\alpha)(1+2\alpha)} \text{ and } q_2 = \frac{19\alpha^2 - 12\alpha - 4}{6(1+\alpha)(1+2\alpha)}.$$

Here we observe that q_1 and q_2 belong to D_2, which is given in [1, Lemma 3]. Therefore the extremal functions f, up to rotations can be obtained when f satisfies

$$\frac{zf'(z) + \alpha z^2 f''(z)}{\alpha z f'(z) + (1-\alpha) f(z)} = 1 - \log(1+z^3).$$

Thus the desired result follows now. ∎

Theorem 4.15. *Let $f \in \mathcal{S}_{\lg}(\alpha)$. Then, we have*

$$|a_5| \leq \frac{107}{288(1+4\alpha)}.$$

The result is sharp.

Proof. The equations (1.11), (1.12), (4.14), (4.15) and (4.16) with $\psi(z)$ in place of $\phi(z)$, yield a_5 in terms of p_1, p_2, p_3 and p_4 as follows

$$\begin{aligned}
|a_5| &= \frac{1}{4(1+4\alpha)} \left| \frac{695}{1152} p_1^4 - \frac{27}{16} p_1^2 p_2 + \frac{1}{2} p_2^2 + \frac{13}{12} p_1 p_3 - \frac{1}{2} p_4 \right| \\
&=: \frac{1}{8(1+4\alpha)} \left| P + \frac{1}{6} p_1 Q - \frac{1}{24} p_1^2 R \right| \\
&\leq \frac{1}{8(1+4\alpha)} \left(|P| + \frac{1}{6}|p_1||Q| + \frac{1}{24}|p_1|^2 |R| \right),
\end{aligned}$$

where $P = p_1^4 - 3p_1^2 p_2 + p_2^2 + 2p_1 p_3 - p_4$, $Q = p_3 - 2p_1 p_2 + p_1^3$ and $R = p_2 - (23/24)p_1^2$. Since $|P| \leq 2$ from (1.15), $|Q| \leq 2$ from (1.14) and $|R| \leq 2$ from (1.13), we get

$$|a_5| \leq \frac{1}{8(1+4\alpha)} \left(2 + \frac{2}{3} + \frac{1}{12}|p_1|^2 - \frac{1}{576}|p_1|^4 \right).$$

Let us assume $G(p_1) := |p_1|^2/12 - |p_1|^4/576$. Then, the formula given in (4.20) yields the bound when $A = -1/576$, $B = 1/12$ and $C = 0$. Letting $p_1 = 1$, $p_2 = 2$, $p_3 = -2/3$ and $p_4 = -1/64$ shows that the result is sharp. ∎

Recall that $|H_3(1)| \leq |a_3||a_2 a_4 - a_3^2| + |a_4||a_4 - a_2 a_3| + |a_5||a_3 - a_2^2|$. Using Examples 2, 3, 4, 5 and Theorem 4.15, we can estimate the bound for $H_3(1)$ for the class $\mathcal{S}_{\lg}(\alpha)$, which is stated below in the following theorem:

Theorem 4.16. *Let $f \in \mathcal{S}_{\lg}(\alpha)$. Then*

$$|H_3(1)| \leq g(\alpha),$$

where

$$g(\alpha) = \frac{949 + 11388\alpha + 52493\alpha^2 + 114974\alpha^3 + 117180\alpha^4 + 42568\alpha^5}{1728(1+4\alpha)(1+3\alpha)^2(1+2a)^4},$$

when $0 \leq \alpha \leq \frac{2+\sqrt{15}}{11}$ and

$$\begin{aligned}
g(\alpha) =& \frac{1}{1728(1+\alpha)(1+4\alpha)(1+3\alpha)^2(1+2\alpha)^3(61\alpha^2 - 20\alpha - 5)} \Big(-5069 - 76035\alpha \\
& - 385994\alpha^2 - 619570\alpha^3 + 831511\alpha^4 + 3545777\alpha^5 + 3327024\alpha^6 + 1298324\alpha^7 \Big),
\end{aligned}$$

when $\frac{2+\sqrt{15}}{11} \leq \alpha \leq 1$.

Remark 4.17. Taking $\alpha = 0$ and 1, we get all the above bounds for the classes $\mathcal{S}^*_{\mathrm{lg}}$ and $\mathcal{C}_{\mathrm{lg}}$, respectively.

As we know the function z is univalent in $\mathbb{D}$, we have $z^n \prec z$, which further implies $1+z^n \prec 1+z$ $(n \geq 1)$. Now, there exists a Schwarz function $\omega(z)$ such that

$$1 + z^n = 1 + \omega(z).$$

Since $|z| < 1$ and $|\omega(z)| < 1$, we can view $1 + z$ and $1 + \omega(z)$ as a shifted unit disk. Thus the branch of the log function is well defined and we can write:

$$1 - \log(1 + z^n) = 1 - \log(1 + \omega(z)).$$

Hence $1 - \log(1 + z^n) \prec 1 - \log(1 + z)$ for all $n \geq 1$. Let us define a function f_n in the class $\mathcal{A}$ as:

$$f_n(z) = z + a_{2,n}z^2 + a_{3,n}z^3 + \cdots = z + \sum_{m=2}^{\infty} a_{m,n}z^m.$$

We consider the subclass $\mathcal{S}^*_{\mathrm{lg},n}$ of $\mathcal{S}^*_{\mathrm{lg}}$ consisting of the functions f_n satisfying

$$\frac{zf_n'(z)}{f_n(z)} = 1 - \log(1 + z^n) \quad (n \geq 1),$$

which upon simplification yields

$$zf_n'(z) = f_n(z)(1 - \log(1 + z^n)).$$

Further, we have

$$\sum_{j=1}^{\infty}\left(\sum_{k=1}^{\infty}(-1)^k\frac{a_j}{k}z^{nk+j}\right) = \sum_{s=1}^{\infty}(s-1)a_sz^s.$$

On comparing the coefficients of like power terms on either side of the above equation, we get a special pattern due to which we conjecture the following:

Conjecture 1. Let $f_n \in \mathcal{S}^*_{\mathrm{lg},n}$. Then

$$|a_{m,n}| \leq |a_{m,1}|.$$

5. Further results

We recall that the set $\mathcal{B}$ is the space of all Bloch functions. An analytic function f is said to be a Bloch function if it satisfies

$$\kappa_{\mathcal{B}}(f) = \sup_{z\in\mathbb{D}}(1 - |z|^2)|f'(z)| < \infty. \tag{5.1}$$

Also, $\mathcal{B}$ is a Banach space with the norm $||.||_{\mathcal{B}}$ defined by

$$||f||_{\mathcal{B}} = |f(0)| + \kappa_{\mathcal{B}}(f),\ f \in \mathcal{B}. \tag{5.2}$$

Now, we give below a result involving Bloch function norm for the functions in the class $\mathcal{S}^*_{\mathrm{lg}}$:

Theorem 5.1. *The set* $\mathcal{S}^*_{\mathrm{lg}} \subseteq \mathcal{B}$. *Further, if* $f \in \mathcal{S}^*_{\mathrm{lg}}$, *then* $||f||_{\mathcal{B}} \leq x \approx 1.27429$.

Proof. If $f \in \mathcal{S}^*_{\mathrm{lg}}$, then $zf'(z)/f(z) = 1 - \log(1 + \omega(z)) =: g(z)$. By the structural formula, given in (1.8), we get

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

Upon differentiating f and further considering the modulus, we obtain

$$|f'(z)| = |g(z)| \left|\exp \int_0^z \frac{g(t)-1}{t} dt\right| \leq |1 - \log(1+\omega(z))| \exp\left(\int_0^z \frac{|\log(1+\omega(t))|}{|t|} dt\right). \quad (5.3)$$

Let $t = re^{i\theta_1}$ and $\omega(t) = Re^{i\theta_2}$, where $R \leq r = |t| < 1$ and $-\pi < \theta_1, \theta_2 < \pi$. Now, by using the similar analysis carried out in the proof of part (v) of Theorem 2.2 and in Theorem 2.3, equation (5.3) reduces to

$$|f'(z)| \leq (1 - \log(1-R)) \exp\left(|\log(1-R)| \int_{-\pi}^{\pi} e^{i\theta_1} d\theta_1\right) \leq 1 - \log(1-r).$$

Thus we have

$$(1-|z|^2)|f'(z)| \leq (1-r^2)(1-\log(1-r)) =: g(r),$$

which upon differentiation, gives

$$g'(r) = 1 - r + 2r\log(1-r).$$

Taking $g'(r) = 0$, yields $r_0 \approx 0.453105$. Since $g''(r_0) < 0$, we have $\max_{0\leq r<1} g(r) = g(r_0) \approx 1.27429 < \infty$. Using (5.1), we obtain $\mathcal{S}^*_{\text{lg}} \subseteq \mathcal{B}$. We can now, estimate the norm $||f||_{\mathcal{B}}$ for the functions in the class $\mathcal{S}^*_{\text{lg}}$. Now, by using the definition of norm, given in (5.2), we have

$$||f||_{\mathcal{B}} \leq f(0) + 1.27429.$$

By the normalization of the function f, the result follows now at once. ∎

The following theorem gives the sufficient condition for the given function g to belong to the class $\mathcal{S}^*_{\text{lg}}$.

Theorem 5.2. *Let m, $n \geq 1$ and $0 \leq \lambda \leq 1$. Then, $g(z) = z\exp(\alpha) \in \mathcal{S}^*_{\text{lg}}$, where*

$$\alpha = \sum_{k=1}^{\infty} \frac{1}{k^2}\left(\lambda\left(\frac{(-z)^{nk}}{n} - \frac{(-z)^{mk}}{m}\right) + \frac{(-z)^{mk}}{m}\right).$$

Proof. For the given α, we have

$$g(z) = z\left(\exp\left(\frac{(\lambda-1)}{m} z^m \left(1 - \frac{z^m}{4} + \frac{z^{2m}}{9} - \cdots\right) - \frac{\lambda}{n} z^n \left(1 - \frac{z^n}{4} + \frac{z^{2n}}{9} - \cdots\right)\right)\right).$$

Then, we have

$$\begin{aligned}\frac{zg'(z)}{g(z)} &= 1 + (\lambda-1)z^m\left(1 - \frac{z^m}{2} + \frac{z^{2m}}{3} - \cdots\right) - \lambda z^n\left(1 - \frac{z^n}{2} + \frac{z^{2n}}{3} - \cdots\right) \\ &= \lambda(1 - \log(1+z^n)) + (1-\lambda)(1-\log(1+z^m)).\end{aligned}$$

We observe that $1 - \log(1+z^t) \prec 1 - \log(1+z) =: \psi(z)$ for all $t \geq 1$ and the function ψ is convex in $|z| < 1$. Thus the result follows at once when $0 \leq \lambda \leq 1$. ∎

When $m = n$, the above Theorem yields the following result:

Corollary 5.3. *Let $n \geq 1$ and $g(z) = z\exp(\alpha)$, where*

$$\alpha = \frac{1}{n}\left(\sum_{k=1}^{\infty} \frac{(-z)^{nk}}{k^2}\right).$$

We have $g \in \mathcal{S}^*_{\text{lg}}$.

Theorem 5.4. *The class $\mathcal{S}^*_{\text{lg}}$ is not a vector space.*

Proof. For if, the class $\mathcal{S}^*_{\text{lg}}$ is a vector space, then the class preserves additive property, that is, whenever two functions belong to the class $\mathcal{S}^*_{\text{lg}}$, then their sum also belongs to the class $\mathcal{S}^*_{\text{lg}}$. Let f_1 and $f_2 \in \mathcal{S}^*_{\text{lg}}$. Then, using (1.6), we obtain

$$f_1(z) = z\exp\left(\int_0^z \frac{-\log(1+\omega_1(t))}{t}dt\right) \text{ and } f_2(z) = z\exp\left(\int_0^z \frac{-\log(1+\omega_2(t))}{t}dt\right),$$

for some Schwarz functions ω_1 and ω_2. Thus, the sum of the functions, $f_1 + f_2$ to be in $\mathcal{S}^*_{\text{lg}}$, there should exist some Schwarz function $\omega(z)$ such that

$$\omega(z) = \frac{\exp(-z(A'(z)\exp A(z) + B'(z)\exp B(z)))}{\exp A(z) + \exp B(z)} - 1,$$

where

$$A(z) = \int_0^z \frac{-\log(1+\omega_1(t))}{t}dt \text{ and } B(z) = \int_0^z \frac{-\log(1+\omega_2(t))}{t}dt.$$

Then $\omega(0) = 0$ and $|\omega(z)| < 1$ for all A and B. Let $\omega_1(z) = z$ and $\omega_2(z) = z^2$. We observe that $\omega(z) \approx 1.03053$ at $z = -(\frac{1}{2} + i\frac{2}{3})$, which contradicts the existence of Schwarz function $\omega(z)$ with $|\omega(z)| < 1$. Hence the assertion follows. ∎

The following theorem is an immediate consequence of the growth Theorem of $\mathcal{S}^*(\Phi)$.

Theorem 5.5. *Let $f \in \mathcal{S}^*_{\text{lg}}$, then we have*

$$|f(z)| \leq |z|\exp\left(\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2}\right) = |z|L \quad (z \in \mathbb{D}),$$

where $L \approx 0.822467$.

Proof. In view of the Remark 1.7, we get

$$t_\Psi(r) \leq |f(z)| \leq -t_\Psi(-r),$$

For $|z| = r$, we have

$$\log\left|\frac{f(z)}{z}\right| \leq \int_0^r \frac{\log(1+t)}{t}dt \leq \int_0^1 \frac{\log(1+t)}{t}dt = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n^2}.$$

The convergent nature of the series on the right side of the above equality yields the desired result at once. ∎

Acknowledgements. The work presented here was supported by a Research Fellowship from the Department of Science and Technology, New Delhi.

References

[1] R. M. Ali, V. Ravichandran and N. Seenivasagan, Coefficient bounds for p-valent functions, Appl. Math. Comput. **187** (2007), no. 1, 35–46.
[2] M. K. Aouf, J. Dziok and J. Sokół, On a subclass of strongly starlike functions, Appl. Math. Lett. **24** (2011), no. 1, 27–32.
[3] S. Banga and S. S. Kumar, Applications of differential subordinations to certain classes of starlike functions, J. Korean Math. Soc. **57** (2020), no. 2, 331–357.
[4] D. A. Brannan and W. E. Kirwan, On some classes of bounded univalent functions, J. London Math. Soc. **1** (1969), no. 2, 431–443.
[5] P. Goel and S. S. Kumar, Certain class of starlike functions associated with modified sigmoid function, Bull. Malaysian Math. Sci. Soc. **43** (2020), no. 1, 957–991.
[6] U. Grenander and G. Szegö, *Toeplitz forms and their applications*, California Monographs in Mathematical Sciences, University of California Press, Berkeley, 1958.

[7] S. K. Lee, V. Ravichandran and S. Supramaniam, Bounds for the second Hankel determinant of certain univalent functions, J. Inequalities Appl. **2013** 2013:281, 17 pp.
[8] R. J. Libera and E. J. Złotkiewicz, Early coefficients of the inverse of a regular convex function, Proc. Am. Math. Soc. **85** (1982), no. 2, 225–230.
[9] W. C. Ma and D. Minda, A unified treatment of some special classes of univalent functions, in *Proceedings of the Conference on Complex Analysis (Tianjin, 1992)*, 157–169, Conf. Proc. Lecture Notes Anal., I, Int. Press, Cambridge, MA.
[10] R. Mendiratta, S. Nagpal and V. Ravichandran, On a subclass of strongly starlike functions associated with exponential function, Bull. Malaysian Math. Sci. Soc. **38** (2015), no. 1, 365–386.
[11] A. K. Mishra, J. K. Prajapat and S. Maharana, Bounds on Hankel determinant for starlike and convex functions with respect to symmetric points, Cogent Math. **6** (2016), no. 1, 1–9.
[12] G. Murugusundaramoorthy, S. Kavitha and T. Rosy, On the Fekete-Szegö problem for some subclasses of analytic functions defined by convolution, Proc. Pakistan Acad. Sci. **44** (2007), no. 4, 249–254.
[13] Z. Nehari, *Conformal mapping*, McGraw-Hill Book Co., Inc., New York, Toronto, London, 1952.
[14] V. Ravichandran and S. Verma, Bound for the fifth coefficient of certain starlike functions, C. R. Acad. Sci. Paris, Ser. I. **353** (2015), no. 6, 505–510.
[15] M. Raza and S. N. Malik, Upper bound of the third Hankel determinant for a class of analytic functions related with lemniscate of Bernoulli, J. Inequal. Appl. **2013**, 2013:412, 8 pp.
[16] M. S. Robertson, Certain classes of starlike functions, Michigan Math. J. **32** (1985), no. 2, 135–140.
[17] Y. J. Sim, O. S. Kwon, N. E. Cho and H. M. Srivastava, Some classes of analytic functions associated with conic regions, Taiwan. J. Math. **16** (2012), no. 4, 387–408.
[18] G. Singh, Hankel determinant for a new subclass of analytic functions, Sci. Magna. **8** (2012), no. 4, 61–65.
[19] H. M. Srivastava, A. K. Mishra and M. K. Das, The Fekete-Szegö Problem for a Subclass of Close-to-Convex Functions, Complex Variables Theory Appl. **44** (2001), no. 2, 145–163.

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

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

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Optimization of Neural Network Architecture using Lions Optimization Algorithm

Kaptan Malik
Department of Computer Science
Delhi Technological University
Delhi, India
malikhoney232@gmail.com

Kanishk Srivastava
Department of Computer Science
Delhi Technological University
Delhi, India
kanishksrivastava21@gmail.com

Krishna Sharma
Department of Computer Science
Delhi Technological University
Delhi, India
krishna858606@gmail.com

***Abstract*—For several years, humans have observed nature to find patterns and similarities between their problems and natural occurrences. Due to this, there has been a surge of research and algorithms regarding our environment. It is usually a debilitating assignment to structure and manufacture an ANN for a selected issue such as nature. There is a pair of thorough venture rules that are accessible to characterize the ANN and, in any case, numerous parameters to switch. The productive techniques are an option in contrast to counterfeit neural systems planning; however, the computational cost required for specific applications is restrictive. Accordingly, the critical point here is to contemplate progressed developmental calculation strategies, which likewise advance accomplishing ANN structures that commit to the task efficiently, explicit to each issue using Lion's Optimization Algorithm (LOA). Pima Indians Diabetes Database used as a model for the optimization of ANN. The results show that subsequent generations of modeled lions provide better accuracy and efficiency to tend towards the highly optimized solution.**

Keywords—Nature-inspired; ANN; Lion's Optimization Algorithm;

I. INTRODUCTION

Research related to nature [1] expects to require care of complex issues with vulnerability, high level of detail, loose and steep clashes, just smitten by personal motivation. Biological processing begins to be more prevalent in understanding genuine life issues [2]. For an extended time, humans have used the direction of nature to find the foremost proper accountability issues. Subsequently, during the recent leading decades, there has been a developing endeavor in creating algorithms roused naturally [3]. For example, the Genetic Algorithm mimics Darwinian advancement ideas. Counterfeit Immune Systems, reenact organic insusceptible frameworks for improvement. Ant Colony Optimization is inspired by ants and their methods to scourge for food. Particle Swarm Optimization talks about how offspring of the creature's herd together, how their movement gives an insight about optimization. Bee's Algorithm impersonates how bees behave and copulate in their colonies. Some algorithms provide a better account than others. Therefore, seeking for new optimization procedures is an open issue during this paper [4]. Optimization of the neural network structure of the Lion's Optimization Algorithm is employed.

II. LITERATURE REVIEW

A. Lion's Social Behavior

Lions, if compared to their counterparts, show a more varied level of companionship along with hatred towards others. This contrasting behavior makes lions the center of attention for many fields of research. The social structure for lions comprises of two halves. Nomads and Pride. As the name suggests, prides include one or more resident males accompanied by four to five resident females. The other group comprises of Nomads, which are isolated male or female lions traversing the land [4]. Young ones of lions reach a certain age in their prides, and after maturing, they either challenge the resident head or leave the pride as a nomad. Once they have left their respective prides, the newly formed nomads can be present in pairs as well, observed more commonly in males who have left a pride together as opposed to males from different prides. It is interesting to note that these lifestyles can be temporary, and a nomad may join a pride and vice versa. Regarding hunting, a group of female lions ambushes the prey by approaching it from different standpoints and stunt the victim rapidly. The success of a lion hunt is directly proportional to how managed the group is while hunting. Only a small portion of the pride engages in killing while others rest. The timing of lions engaging in mating is prevalent all year, due to which the females can bear cubs anytime. The number of partners associated with a lioness can be more than one. Urine spray for marking the pride's geographical monopoly is typical behavior. [5][6].

B. Meta-Heuristic Algorithms for Optimization

1) Ant Colony Optimization: This optimization is based on the nature of hoarding and collecting food from different sources. Once an ant leaves from the colony, it searches for food while leaving a scent behind. This scent comes from the pheromones. Pheromones are used as a guideline for other working ants to follow the same path. Ant colonies formalized themselves before proceeding further to use this optimization. If a task T_i has to be performed by an ant, the total number of ants must be high enough to match the number of tasks. For the given R_i assets for completing a task, execution begins, and resources utilized. This kind of behavior is useful for

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:23 UTC from IEEE Xplore. Restrictions apply.

encountering point-to-point problems faced in many mathematical and computing functions. Some examples are scheduling algorithms in operating systems, Knapsack problem, traveling salesperson problem [7].

2) Genetic Algorithms: This Algorithm is based on Darwinian concepts as evolution plays a significant role in it. The DNA of every living organism contains chromosomes. A chromosome contains vital information about the next generation. This information is optimized by nature through processes like mutation, crossovers, diversity. In this Algorithm, a genome is studied taken from every age and optimized through the methods above. A population consists of chromosomes in a large amount; a fitness function calculates the quality of the next generation chromosome using operations such as crossover and mutation. Hence the offspring obtained is optimized, and this process continues until the requirement of offspring is satisfied. The parameters for determining the quality of an offspring may vary. It can be dependent on the cost or flowtime used. Examples include game theory applications and tournament selection [8].

3) Particle Swarm Optimization: The early stages of the offspring of many animals which lay eggs exhibit behavior which can be explained and utilized using this Algorithm. A swarm, indicating small singular particles, is a field of ever-growing interest. It is applicable in swarm intelligence-related research. This Algorithm has also been merged with other optimization algorithms to give different results. In this optimization, for observation of pathways of single particles, a real-time function is implemented, for illustration of these pathways, vectors dependent on position and time are present. In a system, these particles tend to move around randomly along with the tendency to move the next best position P. The updating process takes place when a more suitable location is found by particle i. There exists an optimum site for each particle at any given time t. Our goal is to present the most optimal solution until the set of repetitions is complete [9].

4) Bee's Algorithm: The bee's Algorithm is dependent on factors such as population and density [10]. Bees in nature gather their food by searching their environment. This algorithm copies that behavior. The Algorithm uses a combination of two methods, which are a generalized search and a nearby search. The sole condition for applying the bee's Algorithm is that some measure of the distance between the solutions is defined. Several studies have demonstrated this Algorithm in practical use.

C. Lion's Optimization Algorithm

Based on the lion's behavior and social organization [4]. Lions are incredibly diverse creatures; they show animosity and companionship. The appearance of their sexual counterparts also differs significantly. Due to these reasons, Lions in this Algorithm is the subject of our study.

1) Hunting: The method of hunting in lions mostly depends on the females while the resident male's rest—the method of surrounding and ambushing the prey. Pictorially depicted in Fig. 1 [11]. They have unique strategies to gather food for their

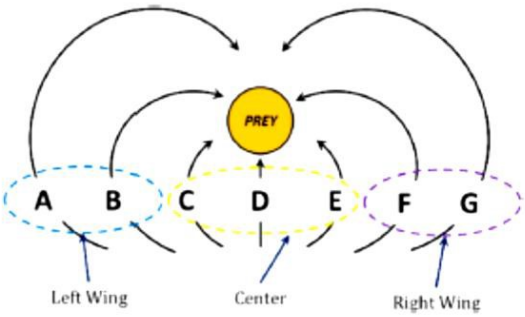


Fig. 1. Generalized lion hunting behavior

pride. Every lion has a particular in a hunt, and cooperation is key for a successful hunt [4][5].

2) Roaming: In a given territory of a lion in a pride, it is typical behavior for them to roam. In this Algorithm, we randomly select a certain percentage of the pride as lions who will walk around. During this strolling session, if a lion comes upon a healthier position than its previous one, that particular lion's best-visited position is changed to the current position. This roaming helps the Lion's Optimization Algorithm to increase its range for the search for a better position or a better solution [5].

3) Mating: For the survival of their species, diversification of genes and information is necessary, precisely what mating incorporates. In the Algorithm, males and females are selected randomly [4]. The offspring is a combination of the traits and information of their parents. The case of nomad lions slightly differs from lions in pride when it comes to mating. The frequency is lower and generated at random.

4) Defense: In an average pride, male lions upon reaching maturity, become aggressive, and fight other males in their pride. The male lions who are defeated in internal or external conflict eventually become a nomad. Contrary to this, internal conflict can also occur. In this, the recently matured lion can take over the pride. The defeated lion essentially becomes a nomad [5]. The variable used in the function for Defense contains these instructions:

1) Protection against matured cubs

2) Protection against outcast males

5) Migration: This is a measure to increase the range of our search to find the best position inspired by the movement of any particular lion or pride. Lions have a lifestyle comprising of not being stationary, and that is incorporated in LOA [5].

6) Convergence: Convergence is indicative of the endpoint of the Algorithm. The point of using different prides in LOA is for increasing randomness as each pride will behave and move around differently, which in turn helps the Algorithm to come to an optimal solution [5].

7) Implementation: Describing the relation of the above processes with the optimization technique in the Algorithm [4].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:23 UTC from IEEE Xplore. Restrictions apply.

- Prides act as small hubs of search spaces that spread around the region. Also, it facilitates the optimization of multi-modal problems.
- Dynamic of prides and nomad lions assist LOA to increase its range in the solution space. Roaming, especially in the case of nomad lions, provides more abstractness to the search.
- Mating provides diversity and traits from previous generations to help in the betterment of optimal solutions.
- Hunting comprises surrounding the prey, which can be as approaching the nearest solution from different angles.
- In the event of pride being takeover by a nomad lion or a mature male within the pride, the outcome gives the best solution within that region.

III. RESEARCH METHODOLOGY

The algorithms mentioned in the prior have been useful in different fields of research and development, but no algorithm exists, which provides the most desired solution for the optimizations above. Hence, looking for better optimization techniques is still a challenge [9].

A. Training of Network

Parameters (weights) used here help in finding the local minima. In an artificial neural network, weights used to maximize the equation provide us with the method to choose a specific solution. LOA takes lions in a pride as individuals and weights associated act as genes or chromosomes for a particular lion. Based on the results, the best-valued solution will be chosen for mating and repeating the cycle [14].

Consider (1),

$$Y = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_6 \quad (1)$$

Equation (1), consists of 6 inputs (x_1 to x_6) and six weights (w_1 to w_6), such that x_i represents our input and w_i are the weights for some $i \in [1,6]$. In each generation, we will try to optimize weights (w) of our network as per our Algorithm. Such a network will have a more generalized weight, trained on training data [4].

Pseudo-code for Training:

1) Create initial lion population

2) For each generation

- Based on the fitness function, discard weak lions.
- Update the new values of lions in the population.

B. Fitness Function

Fitness Function (otherwise called the Evaluation Function) is a method of quantitatively comparing how close the output is to the real value. Also useful for illustration purposes [10].

Pseudo-code for fitness function:

1) Create an empty list of fitness.

2) For each lion:

- Convert lion to weights and biases as per our Neural Network architecture using no. of neurons in each layer.
- Do forward propagation in the network.
- Find the prediction value
- Compute the accuracy using predicted and target value
- Append accuracy in the fitness list.

C. Algorithm Setup

For our analysis, we have taken the Pima Indians Diabetes dataset, which indicates that if a patient is suffering from diabetes or not. The dataset contains several parameters, such as insulin level, BMI, and age. The size of the dataset is 768x8, the eight columns present in the dataset are analogous to the parameters used in LOA. Population size and the number of lions present in a region are compared [12].

The dimension of each data is 8, which consists of:

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
- Blood Pressure: Diastolic blood pressure (mm Hg)
- Skin Thickness: Triceps skinfold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in kg/ (height in m) ^2)
- Diabetes Pedigree Function: Diabetes pedigree function
- Age: Age (years)

Iterations, after there is negligible change in error, is the stopping condition in the analysis.

The other parameters of the Algorithm are described in Table I.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:23 UTC from IEEE Xplore. Restrictions apply.

TABLE I. ALGORITHM PARAMETERS

Parameters	Values
Number of Pride	4
Percent of Nomad lion	0.2
Roaming Percent	0.2
Mutate Probability	0.2
Sex Rate	0.8
Maturing Probability	0.3
Immigration rate	0.4

IV. CASE ILLUSTRATION

Optimization of neural network architecture using LOA. The following steps followed to achieve a solution to the problem.

NOTE:

- All experiments were done in python
- Illustrations and graphs obtained from Jupyter Notebook

A. Reading and Training of data

First, the diabetes database imported with the help of python provides insight. The database, then split, so one-third of the data can be used for training, the next third used for testing, and the last third is used to verify our results.

From Fig. 2, it can be scrutinized that, on the 1000th generation, training accuracy went to 70% from 65% in the 0th generation.

From Fig. 3, the model is performing and giving good results (close to ~70%) on data that it has not seen.

From Fig. 4, the difference in the plot of both training and validation accuracy is almost constant over 1000 generations. Thus, the network generalizes more on the data. Therefore, it concludes that it works reasonably well on unseen data.

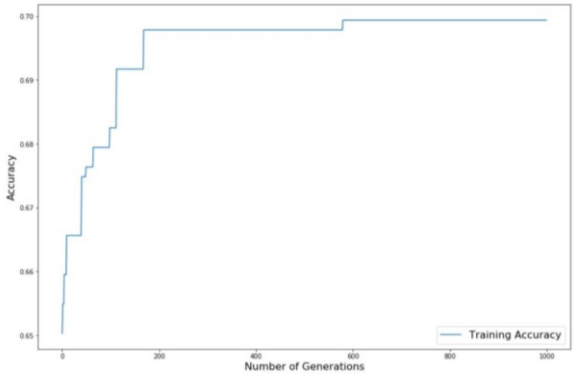


Fig. 2. Training Graph

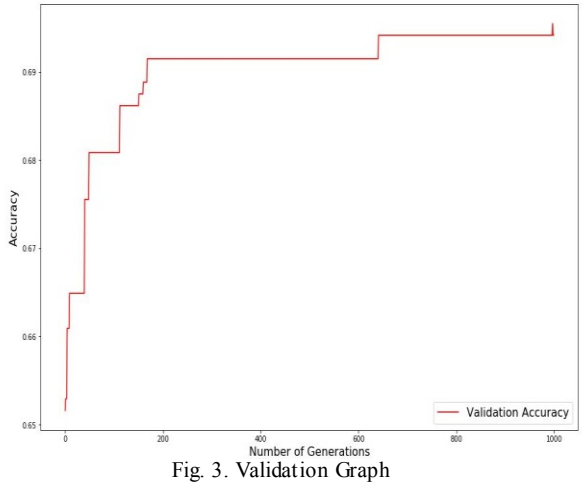


Fig. 3. Validation Graph

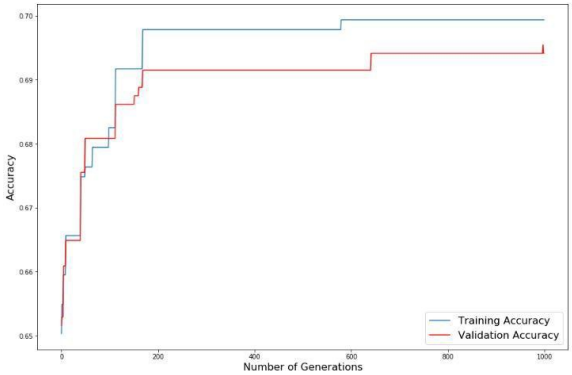


Fig. 4. Training versus Validation graph

V. CONCLUSION

The neural network trained with the help of Lions Algorithm as optimization for weights on the dataset of size 768x8 can be seen. The result from training concludes:

- On increasing the number of generations, accuracy increase significantly.
- It takes comparably lesser time for one generation (0.02s approx.) as compared to one epoch in the backpropagation network.
- On choosing the proper fitness function, In lesser generations, global minima were achieved.
- On deepening the network, i.e., to increase the number of hidden layers, accuracy increases.

In the future, this paper is to proceed with the vision of improving computing efficiency and time taken by processes like training and validation, all while researching other such genetic algorithms used for the optimization of neural network architectures [13].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:23 UTC from IEEE Xplore. Restrictions apply.

REFERENCES

[1] M. Dorigo, "Optimization, learning and natural algorithms," *Politecnico di Milano, Italy*, 1992.

[2] P. Favi, M. Yi, S. Lenaghan, S. C., L. Xia & M. Zhang, "Inspiration from the natural world: from bio-adhesives to bio-inspired adhesives," *Journal of Adhesion Science and Technology*, 28(3-4), p. 290-319, 2004.

[3] Oftadeh, R. Mahjoob, M.J., and M. Shariatpanahi, "A novel meta-heuristic optimization algorithm inspired by group hunting of animals: Hunting search," *Computers & Mathematics with Applications,* 60(7), p.2087-2098, 2010.

[4] M. Yazdani and F. Jolai, "Lion optimization algorithm (LOA): a nature-inspired metaheuristic algorithm," *Journal of computational design and engineering,* 3(1), p.24-36, 2016.

[5] B.R. Rajakumar, "The Lion's Algorithm: a new nature-inspired search algorithm," *Procedia Technology*, 6, p.126-135, 2012.

[6] H. Bauer, N. Vanherle, I. Di Silvestre and H.H De Iongh, "Lion—prey relations" in *West and Central Africa. Mammalian Biology*, 73(1), p.70-73, 2008.

[7] C. Blum, "Ant colony optimization: Introduction and recent trends," *Physics of Life Reviews*, 2(4), p.353-373, 2005.

[8] J. Kennedy and R. Eberhart, "Particle swarm optimization," *In Proceedings of ICNN'95-International Conference on Neural Networks*, 4, p. 1942-1948, 1995.

[9] DH Wolpert and WG Macready, "No free lunch theorems for optimization," *Evolut. Comput. IEEE*, 1(1), p. 67–82, 1997.

[10] V. Mallawaarachchi, "How To Define A Fitness Function In A Genetic Algorithm?," [online] Medium. Available at: <https://towardsdatascience.com/how-to-define-a-fitness-function-in-a-genetic-algorithm-be572b9ea3b4>, 2020

[11] D. Scheel, C. Packer, "Group hunting behavior of lions: a search for cooperation," *Anim. Behav*,41(4),p. 697–709, 1997.

[12] Kaggle.com, "Pima Indians Diabetes Database" [online] Available at: https://www.kaggle.com/uciml/pima-indians-diabetes-database, 2020.

[13] A Bashar, "SURVEY ON EVOLVING DEEP LEARNING NEURAL NETWORK ARCHITECTURES," *Journal of Artificial Intelligence,* 1(02), p. 73-82, 2019.

[14] A. Gad, "Genetic Algorithm Implementation In Python." [online] Medium. Available at: https://towardsdatascience.com/genetic-algorithm-implementation-in-python-5ab67bb124a6, 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:37:23 UTC from IEEE Xplore. Restrictions apply.

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Optimization of the process parameters for hydrotreating used cooking oil by the Taguchi method and Fuzzy logic

Naveen Kumar
Centre for Advanced Studies and Research in Automotive Engineering,
Delhi Technological University,
Bawana Road, Delhi, 110042, India
naveenkumardce@gmail.com

Ankit Sonthalia[1]
Department of Automobile Engineering, SRM IST
SRM IST, NCR Campus, Ghaziabad, Uttar Pradesh – 201204 India
ankit_sont@yahoo.co.in

Rashi Koul
Department of Mechanical Engineering, Amity University,
Amity Education Valley, Gurugram, Manesar, Panchgaon, Haryana - 122413 India
rashikoul@gmail.com

[1]Corresponding author: ankit_sont@yahoo.co.in

ABSTRACT

Hydrotreating process is an alternate approach for producing diesel hydrocarbons from the biomass based oils. In the present study, used cooking oil was selected for the hydrotreating process due to its high abundance. A batch reactor was used for carrying out the experiments. To increase the reaction rate a manganese, cerium promoted ruthenium based catalyst supported on Al_2O_3 was used. Design of experiments was used for optimizing the process parameters. The Taguchi method was selected as it reduces the number of experiments which saves time and money. The study was aimed at increasing the conversion percentage and diesel selectivity and reducing the naphtha selectivity. Since multi-objective optimization was required, Fuzzy logic was incorporated which utilizes the human thought logic. The analysis of variance shows that the reaction temperature and reaction pressure significantly affect the output parameters.

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Higher temperature leads to cracking of the oil resulting in the formation of large amount of lower carbon chains. Moreover, high hydrogen pressure results in increase in the hydrogenation process thereby increasing the diesel selectivity. The optimized parameters obtained from the study were 360°C reaction temperature, 40bar initial reaction pressure and 200min reaction time. Confirmation experiment was carried out using these parameters and the conversion efficiency and diesel selectivity was 89.7% and 88.2%, respectively. The study shows that the combination of Taguchi and Fuzzy logic is an effective method for optimizing the process parameters of the hydrotreating process.

INTRODUCTION

Energy security, concerns over the climate change and rapid depletion of petroleum reserves has created the need to move towards alternate sources of energy which are renewable in nature [1,2]. Biomass is considered to be the best resource which is renewable in nature and is a plausible alternative to conventional fuels [3]. Biofuels derived from biomass are environment friendly, non- toxic and biodegradable. On the basis of feedstock and nature of production, the biofuels are divided into first, second, third and fourth generation [4]. The first generation biofuel is produced via the transesterification reaction and is known as biodiesel or fatty acid methyl ester (FAME). Most of the vegetable oils and animal fats are used for producing the biodiesel [5,6]. Over the years this technology has matured and is already being used for commercial production of biodiesel. The biodiesel has high amount of oxygenated compounds which inherently lowers its heating value and decreases its oxidation stability [7]. Many countries have allowed blending of biodiesel with diesel so that their dependence on petroleum diesel decreases. However, compatibility issues with some of vehicle's components still remain. Moreover, a lot of waste water is generated during transesterification which is a concern for the environment [8]. Therefore, researchers are

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

shifting their focus towards upgrading the bio-oils to more stable and high quality hydrocarbons. These biofuels are produced via hydro-decarboxylation (HDCX), hydro-decarbonylation (HDCN) or hydro-deoxygenation (HDO) route using specific catalysts [9–11]. The prepared biofuels are linear with little nitrogen, sulfur and oxygen content. Also their cetane number is high with good cold flow properties, storage stability and high heating value. The fuels are compatible with the existing fossil fuel infrastructure and is a promising fuel for large scale application [12].

For the hydrotreating process conventional catalysts such as sulfided nickel-molybdenum and cobalt-molybdenum are used. However, the reactions take place under harsh conditions and the catalyst requires to be leached with sulfur which hinders their applications [13–19]. For sulfidation, hydrogen sulfide gas [20,21] can be passed over the catalyst or elemental sulfur or a sulfur rich compound, such as dimethyl-disulfide (DMDS) can be added to the feed [19,22]. However, the catalyst activity reduces over a period of time due to sulfur leaching and also the product stream is contaminated with sulfur [23,24]. During sulfidation, there is a possibility of sulfur being released into the atmosphere, which is detrimental for the atmosphere. These drawbacks of catalyst sulfidation can be circumvented by using catalyst which do not require sulfidation pre-treatment. Several noble metals such as palladium (Pd), platinum (Pt) and ruthenium (Ru) on different supports have been used in the past for hydrotreatment and they have shown high activity in converting vegetable oil into alkanes [25]. Generally, Pt catalysts support decarbonylation reaction, whereas Pd catalyst support decarboxylation reaction process [26]. Moreover, Ru based catalyst are also active for HDO reactions [27,28].

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Sunflower oil was hydroprocessed by Kikhtyanin et al. [29] using Pd/SAPO-31 catalyst. As the authors increased the temperature from 320 to 360°C the C_{17}/C_{18} ratio increased from 1-4 to 8-10 indicating the dominance of HDCN/HDCX reaction over HDO reaction process. The cloud and pour points of the produced fuel was less than -50°C. Liu et al. [23] hydroprocessed waste cooking oil using Ru/Al13-montmorillonite catalyst. The triglycerides and free fatty acids (FFA) containing high acid value cooking oil was deoxygenated to form hydrocarbon products. The authors observed the pour point to be 20°C with SiO_2 support whereas with montmorillonite support the pour point reduced to -15°C. Murata et al. [20] also hydrotreated jatropha oil using Pt-Re catalyst supported on HZSM-5. With a 10, oil to catalyst ratio, the optimum Re-Pt/HZSM-5 catalyst resulted in C_{15}-C_{18} yield of 67%. The authors found that initially the cleavage of double bond occurs followed by HDO, HDCN and HDCX reaction to form C_{15}-C_{18} alkane mixture. Jatropha oil and a model compound (stearic acid) was hydrotreated under mild conditions by Guo et al. [30] using ruthenium with different supports. Ru supported on $La(OH)_3$ resulted in 100% conversion of jatropha with 98% long chain selectivity at 200°C and 4MPa. The authors also checked the recyclability of the catalyst and found that up to four times nearly 70% alkanes were obtained under the same conditions. The effect of addition of cerium to palladium and ruthenium with $La(OH)_3$ support on the hydrocarbon yield was studied by Patil and Vaidya [31]. The authors in a fixed-bed reactor deoxygenated jatropha oil. At optimized reaction conditions, the yield and oil conversion for Pd-Ce and Ru-Ce were 81 & 97.3% and 75.3 & 91.8%, respectively. The catalyst efficiency was not affected for more than 40 hours of on-stream time. In another study [32], the same

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

authors used hydrotalcite-like (HTlc) supports for palladium and ruthenium catalyst and they studied the effect of various parameters such as pressure, temperature and H_2/oil ratio. At optimum reaction conditions, the yield with Pd is 91.2% and with Ru the yield is 85.6%. Both the catalyst were stable for over 30 hours of on-stream time.

Waste cooking oil (WCO) was used as the feedstock, in this study. WCO is generated in restaurants, fast-food centres, households and food processing industry. The annual generation of waste cooking oil in Europe is 700,000 to 1,000,000 tons, in the US nearly 1.2 billion to 3 billion gallons is generated whereas in Japan alone nearly 400,000-600,000 tons is generated [33]. Its storage and disposal is costly. In most of the developing countries, the used oil is poured into the sewer system which ends up polluting the underground water, rivers and oceans. The expense of treating the wastes containing the used oil can be minimized if the oil is reused such as converting it into fuel. The FFA content of the oil is usually high which increases the production cost of biodiesel. Base catalysts can easily trans-esterify the triglycerides having low fatty acid content but triglycerides with high fatty acid will form soap. Acid catalysed transesterification is slow in converting triglycerides with fatty acids [34]. Therefore, the high-acid value WCO is converted into biodiesel by a two-step process. First acid catalyst is used to remove the free fatty acids and then the base catalyst converts the triglycerides. The hydrotreatment process in converting WCO is economical since both triglycerides and free fatty acids are deoxygenated in a single step [35]. The waste cooking oil has been hydrotreated by many researchers.

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Bezergianni et al. [36] used a commercial nickel-molybdenum catalyst to study the effect of hydrogen to oil ratio, pressure, and liquid hourly space velocity on the production of diesel like fuel from waste cooking oil. The authors observed favourable fuel production with low pressure as high pressure promotes cracking. With high hydrogen to oil ratio cracking reactions are triggered resulting in reduction of WCO conversion to diesel like fuel. Higher space velocity promotes the production of diesel like fuel. Sulphur and nitrogen removal efficiency is also high. The effect of temperature alone was also studied by Bezergianni et al. [37] keeping the other conditions same. The authors studied the yield, selectivity and saturation of double bonds using a commercial hydrotreating catalyst. It was observed that lower temperature favours diesel like fuel production, whereas high temperature favours gasoline like fuel production. Moreover, the increase in temperature also favours the saturation of double bonds. Three types of catalyst namely a severe hydrocracking catalyst, a mild-hydrotreating catalyst, and a hydrocracking catalyst was used by Bezergianni et al. [38] at various operating conditions. The authors observed high conversion and diesel selectivity with all the catalysts. Hydrotreating catalyst resulted in highest removal of nitrogen and sulphur and 99% saturation of the fuel.

The literature reveals that different catalysts can be used for hydrotreating waste cooking oil. However, mostly nickel or cobalt based catalysts has been used by the researchers and few studies have used noble metals based catalyst for the hydrotreating of waste cooking oil. Moreover, sulphidation is necessary for the former catalysts which is carried out by using H_2S gas or by adding dimethyl disulphide (DMDS), whereas no

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

suplidation is required for noble metals. Also the energy requirement is less with the noble metals. Therefore, in the present study a cerium and manganese promoted catalyst containing ruthenium supported on Al_2O_3 was used for hydrotreating the waste cooking oil. To reduce the time and cost of the experimentation Taguchi and Fuzzy logic was used for optimizing the process parameters. The used cooking oil was collected from cafeteria of Delhi Technological University.

Experimental setup

In the present work, the used cooking oil was hydroprocessed in a 100ml batch reactor manufactured by Amar Equipments, India. The temperature of the system was controlled by an electric heating mantle along with water flowing through a cooling spiral. A motor operated impeller was used for stirring the reactor contents, its speed was fixed at 1200rpm. The reactor temperature was measured by a thermocouple and a PID programmable controller was used to control the temperature. The experimental setup schematic is shown in Fig. 1.

The effect of reaction temperature, hydrogen pressure, and reaction time on the conversion efficiency and diesel selectivity was studied. The range of the parameters used in this study are given in Table 1 and they were selected in accordance to Guo et al. [30] and Mansur et al. [39]. For carrying out the hydrogenation process, the reactor was filled with 20g of WCO and the reduced catalyst was added. The reactor was then closed tightly and nitrogen was purged through the reactor. Hydrogen was then filled in the reactor at room temperature. The reactor temperature was raised at the rate of 18°C per minute. When the pressure in the reactor fell below the required pressure, hydrogen was fed to

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

the reactor so that its pressure was constant throughout the experiment. The reaction time was considered from the moment the reactor temperature reached the required reaction temperature. After the stipulated reaction time, the reactor was allowed to cool to ambient conditions. The gases were vented to the atmosphere by bubbling through water. The liquid product was recovered using a syringe and weighed. Acetone was used for the rinsing of the reactor. The acetone containing the solid products was filtered and then the solid products were dried. The solids were then weighed to find out if any other solid apart from the catalyst was formed during the process. The reaction's performance was evaluated by studying the conversion percentage, diesel selectivity and naphtha selectivity, which was calculated using equation (1), (2) and (3).

$$\text{Coversion } (\%) = \frac{\text{Initial Oil}_{360+} - \text{Final Product}_{360+}}{\text{Initial Oil}_{360+}} * 100 \quad (1)$$

$$\text{Diesel Selectivity } (\%) = \frac{\text{Final Product}_{180-360} - \text{Initial Oil}_{180-360}}{\text{Initial Oil}_{360+} - \text{Final Product}_{360+}} * 100 \quad (2)$$

$$\text{Naphtha Selectivity } (\%) = \frac{\text{Final Product}_{40-200} - \text{Initial Oil}_{40-200}}{\text{Initial Oil}_{360+} - \text{Final Product}_{360+}} * 100 \quad (3)$$

Initial oil_{360+} and final $product_{360+}$ are the liquid feed and the product percentage having boiling point higher than 360°C. Initial $oil_{180\text{-}360}$ and final $product_{180\text{-}360}$ are the feed and product percentage having boiling point in the range of 180-360°C. Similarly, Initial $oil_{40\text{-}200}$ and final $product_{40\text{-}200}$ are the feed and product percentage having boiling point in the range of 40-200°C.

It is seen that the quantum of experiments required for finding the optimized reaction parameters are huge. To lower the number of experiments, Taguchi designed orthogonal arrays were used. These arrays help in saving time and experimental costs.

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Based upon the number of degrees of freedoms, the orthogonal array was selected. The degree of freedom is determined by the number of factors and the level of each factor. On the basis of Table 1, three factors and three levels of each factor were considered. Table 2 shows the factors and their levels. An L_9 orthogonal array was selected on the basis of the factors and their levels, shown in Table 3. The testing was carried out randomly on the basis of the table 3 by placing the main factors and their levels in the designated columns.

Catalyst Preparation and Characterization

The catalyst was prepared by a wet impregnation method. An aqueous solution containing 0.62gm cerium nitrate hexahydrate and 15ml of distilled water was prepared and it was slowly dropped with the help of a dropper on a 5gm calcined alumina support till it was just wet. The process was repeated for impregnating an aqueous solution of 0.274gm manganese acetate and 0.248gm ruthenium chloride. A solid sample was then obtained by drying the support at 100°C and calcined at 500°C for 3 hours. The support contained 5% of cerium and 5% of magnesium, whereas the ruthenium loading was 2% of the catalyst weight. The prepared catalyst, before every reaction, was reduced in a hydrogen (2MPa) environment at 220°C for 3 hours.

Fig. 2 shows the XRD pattern of the catalyst. The visible peaks at 2θ=38.5°, 46.7° and 66.9° are typical of Al_2O_3 which is prepared from microcrystalline boehmite. Similar peaks were observed by Sankaranarayanan et al. [8]. The pattern reveals that the metals are highly dispersed on the support surface and due to lower percentage of the metals their individual peaks are not visible [40]. It also indicates that the nature of Al_2O_3 is not

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

affected significantly by the method used for the impregnation of the metals. Even with higher percentage of ruthenium loading similar XRD pattern was observed by Mishra et al. [41].

The scanning electron microscope was used to study the morphology of the freshly prepared catalyst. The SEM analysis, shown in Fig. 3, was carried out before the catalyst reduction. The EDS analysis confirms the presence of ruthenium, cerium and manganese in the catalyst. The mass percentage of ruthenium, cerium and manganese in the catalyst are 6%, 2% and 0.03%, respectively.

Results and Discussion

Composition of used cooking oil

GC-MS (HP-624 column) analysis was carried out for determining the composition of used cooking oil. The results show that the used cooking oil contains palmitic acid (17.47%), linoleic acid (58.51%), oleic acid (16.75%), and stearic acid (4.85%). The unsaturated fatty acids in the used cooking oil were palmitoleic, oleic, linoleic, and linolenic acids whereas the saturated acids were palmitic acid, stearic acid. The saturated fatty acids were about 22.47% whereas the unsaturated fatty acids were 77.53% in the total acids of used cooking oil. All the fatty acids had even carbon atoms in the carbon chain i.e. C_{16}, C_{18} and C_{20}. The sixteen chain carbon atoms occupy 17.5%, eighteen chain carbon atoms occupy 80.7% whereas carbon chains with twenty carbon atoms occupy 1.79% of the fatty acids of used cooking oil [42].

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Hydrotreating of used cooking oil

During the hydroprocessing of unsaturated fatty acid methyl ester, complicated reactions are involved such as dehydration, cracking, hydrogenation, decarbonylation or decarboxylation and carbon deposits. These reactions may occur simultaneously or consecutively. The intermediate products formed are alcohols, relative fatty acids and olefins and the final products are carbon monoxide, carbon dioxide and alkanes [8,43]. The $C_1 - C_4$ alkanes are in gaseous form whereas alkanes containing greater than five number of carbon atoms are in liquid form and may be normal or isomeric in nature. The reaction can take place via hydrodeoxygenation route with water as by-product and through decarbonylation or decarboxylation route with CO or CO_2 as by-product [44]. The carbon chain has same number of atoms with hydrodeoxygenation route whereas with decarbonylation or decarboxylation route, the carbon chain reduces by one carbon atom [11]. The selectivity of diesel and the conversion percentage for all the tests cases is depicted in Fig. 4.

It is seen that all the three parameters affect the overall conversion efficiency as well as the diesel selectivity. Moreover, the reaction temperature has a stronger influence in comparison to the other two parameters. The general trend shows a decrease in conversion efficiency with high temperatures. The highest conversion efficiency was observed at 330°C and the lowest at 390°C. The selectivity of diesel was higher than naphtha at all the temperatures, and at 330°C the diesel selectivity was highest. With increase in temperature the diesel selectivity decreases. Moreover, as the temperature

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

increases the naphtha selectivity increases and it is highest at 390°C since cracking of the product takes place at higher temperatures and lighter molecules are formed [45].

The hydrogen pressure determines the pathway of the used cooking oil conversion into alkanes. As the hydrogen pressure increases the conversion efficiency increases because the acid-catalyzed reaction rate increases as more hydrogen is adsorbed on the catalyst surface thus producing more acid sites for the reaction [46]. This facilitates triglycerides deoxygenation along with the breaking of the glycerol backbone. Fig. 4 shows that the pressure and reaction time both affect the conversion percentage as well as the diesel selectivity. It is seen that at 330°C as the pressure and reaction time increases the conversion percentage increases but the diesel selectivity decreases. At 330°C, 40 bar pressure and 140 minutes reaction time, the conversion efficiency increases from 88.6% to 89.9% at 50 bar pressure and 200 minutes. Whereas the diesel selectivity decreases from 86.7% to 84.2% as the amount of time available for the reaction increases and also more amount of triglycerides are converted into alkanes. Also as large amount of hydrogen is present in the reactor, the formation of lighter products (water, propane, and methane) is enhanced. The energy consumption is also increased by increasing the hydrogen pressure [19,47]. At 390°C, 30bar and 200 minutes, the conversion efficiency is 78.2% and diesel selectivity is 67.8% which increases to 80.9% and 71.1% respectively at 40bar and 80 minutes reaction time. As the reaction pressure increases and time decreases, the relative conversion of waste cooking oil into lighter carbon chains and gas reduces and high amount of hydrogen is utilized in only forming long chain alkanes. Whereas at 50bar and 140 minutes reaction time, the diesel selectivity and the conversion

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

efficiency is 66.5% and 76.6%, respectively. Since at high hydrogen pressure and more reaction time, the longer chain carbon atoms may have cracked to smaller chain atoms.

To understand the reaction route, it is necessary to estimate the ratio of the amount of odd number of carbon atoms (C_{17}) to even number of carbon atoms (C_{18}) in the liquid product. The C_{17}/C_{18} ratio, shown in Fig. 5, indicates the relative activity as the long-chained used cooking oil consists mainly of C18 carbon atoms. It is seen that high amount of heptadecane was formed with the increase in temperature. This is most likely because of endothermic nature of the decarbonylation and decarboxylation reactions whereas the hydrodeoxygenation reaction is exothermic in nature [26,45]. At 330°C, with the increase in reaction time and pressure the amount of odd number of carbon atoms decreases. The primary reason for the trend is the increase in the adsorption of hydrogen on the catalyst's active sites which increases the deoxygenation activity and reduces the decarboxylation and decarbonylation reaction [48]. Similar trend is observed at 360°C and 390°C. Moreover, the reaction time has little effect on the heptadecane formation as the effect of hydrogen adsorption on active sites is more dominant.

Optimization of process parameters using fuzzy logic and Taguchi

Taguchi's orthogonal array emphasizes on signal-to-noise (S/N) ratio for handling the data. The ratio measures the deviation in parameters from their expected values. For the three variables use in this study namely conversion percentage, diesel selectivity and naphtha selectivity, found from the experiments, S/N ratio was calculated. For each parameter the ratio is characterized as the nominal-the-better (NTB), the smaller-the-better (STB) and the larger-the-better (LTB) [49,50]. The S/N ratio for STB and LTB is

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

calculated using equation 4 and 5, respectively. The optimal conditions and the influence of individual factors can be studied using the S/N ratio. For conversion percentage and diesel selectivity, the larger-the-better characteristic was used. Whereas, for naphtha selectivity, the smaller-the-better characteristic was used. Since the aim of the study is to increase the conversion efficiency and diesel selectivity and reduce naphtha selectivity. The S/N ratio for the three variables is given in Table 5.

$$\frac{S}{N} = -10log_{10}\left[\frac{1}{n}\sum_{i=1}^{n} y_i^2\right] \quad (4)$$

$$\frac{S}{N} = -10log_{10}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{y_i^2}\right] \quad (5)$$

Where y_i is the value of the performance characteristic measured during the trial, n is the number of tests in a trial and the S/N ratio has a unit of decibel (dB).

It is seen that using the Taguchi method only a single parameter can be optimized, whereas three parameters needs to optimized in the present study. Therefore, other theories have to be combined with the Taguchi method for optimizing multiple performance characteristic. Some of the other theories are desirability function [51], utility theory [52] and grey relation theory [53,54]. But in these theories, each response is allocated with weights. Since these weights are assigned on the basis of some assumptions that results in uncertainty and indistinctness in the solution. Therefore fuzzy based Taguchi method can be used for optimization of multiple objectives.

In fuzzy control, the logic of human thought in controlling a process is represented as fuzzy rules or relations. The fuzzy logic unit consists of a fuzzifier, membership functions, a fuzzy rule base, an inference engine and a defuzzifier [55]. The S/N ratios

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

obtained from the Taguchi method are normalized in the range 0-1 so that the quality characteristic of the performance parameter with smaller S/N ratio is present. The normalization of the S/N ratio is carried out using equation 6.

$$X_{normalized} = (X_i - X_{min})/(X_{max} - X_{min}) \quad (6)$$

Where *i* varies from 1,2,3,..etc. X_i denotes the S/N ratio at the i^{th} position, X_{min} is the minimum S/N ratio and X_{max} is the maximum S/N ratio.

The fuzzifier then fuzzifies the normalized S/N ratio using the membership functions. The input and the output variables were divided using the triangular membership function, which were used for finding the multi performance characteristics index (MPCI). The three input variables namely, reaction temperature, reaction pressure, and reaction time have three levels, small (S), medium (M), large (L), shown in Fig. 6. The output MPCI has seven levels namely, very-very small (VVS), very small (VS), small (S), medium (M), large (L), very large (VL), very-very large (VVL), shown in Fig. 7. The inference engine generates a fuzzy value by applying fuzzy reasoning using the fuzzy rules. Table 4 shows the fuzzy rules for all the inputs and the outputs. Lastly, the fuzzified value is defuzzified and converted into MPCI. The defuzzification is carried out using the centre of gravity method. Table 5 shows the inferred MPCI using the experimental combinations given in Table 3.

Since the design of experiments is orthogonal, the effect of all variables at different levels on the MPCI can be studied. The mean of the MPCI for each process parameter level was calculated and given in Table 6. The mean of MPCI for the reaction temperature at first level is calculated by averaging the mean performance characteristic

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

indexes for the test cases 1-3. Similarly, the mean MPCI for each level of other variables can be found. The total mean of the test cases is also given in Table 6. The effect of individual parameters is shown in the MPCI graph (Fig. 8), and the total mean of MPCI is indicated by the dashed line. Basically, the highest value of MPCI dictates the best parameters for the process as the variance of performance characteristics around the required value is small. However, it is still required to know the relative importance of each parameter such that the optimized combinations of the variable's level can be accurately determined.

Analysis of Variance (ANOVA)

To identify the most significant process variables and their effect on the output parameters the analysis of variance was conducted. By summing up and squaring the deviations from the total mean of multiple performance characteristics indexes, the variability of MPCI can be separated into contribution by each variable and error. Equation 7 is used to calculate the total sum of squared deviations (SD_T).

$$SD_T = \sum_{i=1}^{q} (\eta_i - \eta_m)^2 \quad (7)$$

Where q is the number of test cases, η_m is the mean of MPCI and η_i is the mean of MPCI for i^{th} test case. The total sum of squared deviation consists of two parts: one the sum of squared deviation (SD_D) due to each variable and second the sum of squared error (SS_E). Table 7 tabulates the results after applying analysis of variance. The percentage contribution of each variable in the total sum of squared deviations is calculated by equation 8. The percentage contribution signifies the importance of change in each variable on the output. In addition F-test with 95% confidence was used to determine the

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

significant contribution of each variable on the output. Large F value signifies that the change in a particular process parameter significantly affects the output.

$$\text{Contribution } (\%) = \frac{\text{SD}_\text{D}(\text{for each process parameter})}{\text{SD}_\text{T}} \quad (8)$$

The F-test and the contribution percentage shows that MPCI is significantly affected by the reaction temperature, followed by reaction pressure with reaction time having the least effect on the MPCI. The outcomes are reasonable enough as the increase in reaction temperature decreases the conversion efficiency and diesel selectivity and increases naphtha selectivity as the long chain carbon atoms are broken into small chain carbon atoms. Based on the discussions, the optimal process parameters are 330°C reaction temperature, 40bar reaction pressure and 200minutes of reaction time.

Confirmation experiment

After fixing the optimal process parameters, prediction and verification of the improvement in the performance characteristics is necessary. For the optimized process parameters, the MPCI is predicted using equation 9.

$$\eta = \eta_\text{m} + \sum_{\text{i}=1}^{\text{q}}(\eta_\text{i} - \eta_\text{m}) \quad (9)$$

where η is the estimated MPCI, ηm is the total mean MPCI, ηi is the mean of MPCI at the optimal level and q is the number of parameters that affect the MPCI. With the optimal process parameters, the hydrotreating of used cooking oil was carried out and the conversion efficiency, diesel selectivity and naphtha selectivity was found. The predicted MPCI and the experimental results with the optimal process parameters is given in Table 8. As compared to the initial process parameters, the conversion efficiency, diesel

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

selectivity has increased whereas the naphtha selectivity decreased. The signal-to-noise ratio and the MPCI has also improved for the optimized process parameters.

Conclusions

In the present study used cooking oil was hydrotreated using a ruthenium based catalyst. The reaction temperature, pressure and reaction time were varied and their effect on conversion efficiency, diesel selectivity and naphtha selectivity was observed. The results show that as the reaction temperature increases the conversion efficiency reduces since less amount of liquid is formed. The diesel selectivity also decreases since the long chain carbon atoms are broken into smaller chain atoms. The reaction pressure has a high influence on the conversion efficiency as on the active sites of the catalyst more amount of hydrogen is present for the hydrogenation. By using Taguchi method, the number of experiments were reduced and the process parameters were optimized by applying fuzzy logics to the Taguchi method. The study shows that the reaction temperature, pressure and reaction time can be optimized using this approach. The analysis of variance shows that the conversion efficiency and diesel selectivity is significantly affected by the reaction temperature, whereas reaction time has the least influence on output parameters. The optimized reaction temperature was 330°C, reaction pressure was 40bar and time was 200minutes. Using these parameters, hydrotreating of used cooking oil was carried out for confirmation and the results were found to be better than the experiments carried out earlier.

Acknowledgement

The authors are thankful to Dr. B. P. Kiran from Synthesis with Catalysts Pvt. Ltd. for providing the catalyst raw materials.

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

NOMENCLATURE

Ce	Cerium
CO	Carbon monoxide
CO_2	Carbon dioxide
DMDS	Dimethyl-disulfide
FAME	Fatty acid methyl ester
FFA	Free fatty acid
GS-MS	Gas chromatography, mass spectroscopy
HDCN	Hydro-decarbonylation
HDCX	Hydro-decarboxylation
HDO	Hydro-deoxygenation
HTlc	Hydrotalcite
LTB	Larger-the-better
NTB	Nominal-the-better
Pd	Palladium
Pt	Platinum
Ru	Ruthenium
STB	Smaller-the-better
S/N	Signal to noise ratio

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

MPCI	Mean performance characteristics index
WCO	Waste cooking oil

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

REFERENCES

1 Huang, Y., Wei, L., Julson, J., Gao, Y., Zhao, X., 2015. "Converting pine sawdust to advanced biofuel over HZSM-5 using a two-stage catalytic pyrolysis reactor". J. Anal. Appl. Pyrolysis. **111**, pp. 148–55.

2 Huang, Y., Wei, L., Crandall, Z., Julson, J., Gu, Z., 2015. "Combining Mo–Cu/HZSM-5 with a two-stage catalytic pyrolysis system for pine sawdust thermal conversion". Fuel. **150**, pp. 656–63.

3 Geo, V.E., Sonthalia, A., Nagarajan, G., Nagalingam, B., Aloui, F., 2019. "Effect of Port Premixed Liquefied Petroleum Gas on the Engine Characteristics". J. Energy Resour. Technol. Trans. ASME. **141**(11), pp. 1–8.

4 Demirbas, A., 2011. "Competitive liquid biofuels from biomass". Appl. Energy. **88**(1), pp. 17–28.

5 Naik, S.N., Goud, V. V., Rout, P.K., Dalai, A.K., 2010. "Production of first and second generation biofuels: A comprehensive review". Renew. Sustain. Energy Rev. **14**(2), pp. 578–97.

6 Thushari, P.G.I., Babel, S., 2017. "Biodiesel Production From Waste Palm Oil Using Palm Empty Fruit Bunch-Derived Novel Carbon Acid Catalyst". J. Energy Resour. Technol. **140**(3), pp. 32204–10.

7 Li, X., Luo, X., Jin, Y., Li, J., Zhang, H., Zhang, A., Xie, J., 2018. "Heterogeneous sulfur-free hydrodeoxygenation catalysts for selectively upgrading the renewable bio-oils to second generation biofuels". Renew. Sustain. Energy Rev. **82**(3), pp. 3762–97.

8 Sankaranarayanan, T.M., Banu, M., Pandurangan, A., Sivasanker, S., 2011. "Hydroprocessing of sunflower oil-gas oil blends over sulfided Ni-Mo-Al-zeolite beta composites". Bioresour. Technol. **102**(22), pp. 10717–23.

9 Saidi, M., Samimi, F., Karimipourfard, D., Nimmanwudipong, T., Gates, B.C., Rahimpour, M.R., 2014. "Upgrading of lignin-derived bio-oils by catalytic hydrodeoxygenation". Energy Environ. Sci. **7**(1), pp. 103–29.

10 Mortensen, P.M., Grunwaldt, J.-D., Jensen, P.A., Knudsen, K.G., Jensen, A.D., 2011. "A review of catalytic upgrading of bio-oil to engine fuels". Appl. Catal. A Gen. **407**(1–2), pp. 1–19.

11 Sonthalia, A., Kumar, N., 2019. "Hydroprocessed vegetable oil as a fuel for transportation sector: A review". J. Energy Inst. **92**(1), pp. 1–17.

12 Bhuiya, M.M.K., Rasul, M.G., Khan, M.M.K., Ashwath, N., Azad, A.K., Hazrat, M.A., 2016. "Prospects of 2nd generation biodiesel as a sustainable fuel – Part 2: Properties, performance and emission characteristics". Renew. Sustain. Energy Rev. **55**, pp. 1129–46.

13 Kubička, D., Bejblová, M., Vlk, J., 2010. "Conversion of Vegetable Oils into Hydrocarbons over CoMo/MCM-41 Catalysts". Top. Catal. **53**(3), pp. 168–78.

14 Kubička, D., Šimáček, P., Žilková, N., 2009. "Transformation of Vegetable Oils into Hydrocarbons over Mesoporous-Alumina-Supported CoMo Catalysts". Top. Catal. **52**(1), pp. 161–8.

15 Tiwari, R., Rana, B.S., Kumar, R., Verma, D., Kumar, R., Joshi, R.K., Garg, M.O.,

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Sinha, A.K., 2011. "Hydrotreating and hydrocracking catalysts for processing of waste soya-oil and refinery-oil mixtures". Catal. Commun. **12**(6), pp. 559–62.

16 Šimáček, P., Kubička, D., Šebor, G., Pospíšil, M., 2009. "Hydroprocessed rapeseed oil as a source of hydrocarbon-based biodiesel". Fuel. **88**(3), pp. 456–60.

17 Kubička, D., Horáček, J., 2011. "Deactivation of HDS catalysts in deoxygenation of vegetable oils". Appl. Catal. A Gen. **394**(1–2), pp. 9–17.

18 B. Donnis, R.G. Egeberg, P. Blom, K.G.K., 2009. "Hydroprocessing of bio-oils and oxygenates to hydrocarbons. Understanding the reaction routes". Top. Catal. **52**, pp. 229–40.

19 Sotelo-boy, R., Liu, Y., Minowa, T., Sotelo-Boyás, R., Liu, Y., Minowa, T., 2011. "Renewable Diesel Production from the Hydrotreating of Rapeseed Oil with Pt/Zeolite and NiMo/Al2O3 Catalysts". Ind. Eng. Chem. Res. **50**(5), pp. 2791–9.

20 Murata, K., Liu, Y., Inaba, M., Takahara, I., 2010. "Production of Synthetic Diesel by Hydrotreatment of Jatropha Oils Using Pt−Re/H-ZSM-5 Catalyst". Energy & Fuels. **24**(4), pp. 2404–9.

21 Knothe, G., 2010. "Biodiesel and renewable diesel: A comparison". Prog. Energy Combust. Sci. **36**(3), pp. 364–73.

22 Liu, Y., Sotelo-Boyás, R., Murata, K., Minowa, T., Sakanishi, K., 2009. "Hydrotreatment of Jatropha Oil to Produce Green Diesel over Trifunctional Ni–Mo/SiO2 –Al2O3 Catalyst". Chem. Lett. **38**(6), pp. 552–3.

23 Liu, Y., Sotelo-Boyás, R., Murata, K., Minowa, T., Sakanishi, K., 2012. "Production of Bio-Hydrogenated Diesel by Hydrotreatment of High-Acid-Value Waste Cooking Oil over Ruthenium Catalyst Supported on Al-Polyoxocation-Pillared Montmorillonite". Catalysts. **2**(4), pp. 171–90.

24 Liu, Y., Sotelo-Boyás, R., Murata, K., Minowa, T., Sakanishi, K., 2011. "Hydrotreatment of Vegetable Oils to Produce Bio-Hydrogenated Diesel and Liquefied Petroleum Gas Fuel over Catalysts Containing Sulfided Ni–Mo and Solid Acids". Energy & Fuels. **25**(10), pp. 4675–85.

25 Zhao, C., Brück, T., Lercher, J.A., 2013. "Catalytic deoxygenation of microalgae oil to green hydrocarbons". Green Chem. **15**(7), pp. 1720–39.

26 Snåre, M., Kubičková, I., Mäki-Arvela, P., Eränen, K., Murzin, D.Y., 2006. "Heterogeneous Catalytic Deoxygenation of Stearic Acid for Production of Biodiesel". Ind. Eng. Chem. Res. **45**(16), pp. 5708–15.

27 Osaka, Y., Ikeda, Y., Hashizume, D., Iwamoto, M., 2013. "Direct hydrodeoxygenation of cellulose and xylan to lower alkanes on ruthenium catalysts in subcritical water". Biomass and Bioenergy. **56**, pp. 1–7.

28 He, L., Wu, C., Cheng, H., Yu, Y., Zhao, F., Preparation, C., Test, A., 2012. "Supplementary Information Highly selective and efficient catalytic conversion of ethyl stearate into liquid hydrocarbons over Ru/TiO2 catalyst under mild conditions". Catal. Sci. Technol. **2**(7), pp. 2–4.

29 Kikhtyanin, O. V., Rubanov, A.E., Ayupov, A.B., Echevsky, G. V., 2010. "Hydroconversion of sunflower oil on Pd/SAPO-31 catalyst". Fuel. **89**(10), pp. 3085–92.

30 Guo, J.H., Xu, G.Y., Shen, F., Fu, Y., Zhang, Y., Guo, Q.X., 2015. "Catalytic

Preprint Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

conversion of Jatropha oil to alkanes under mild conditions with a Ru/La(OH)<inf>3</inf> catalyst". Green Chem. **17**(5), pp. 2888–95.

31 Patil, S.J., Vaidya, P.D., 2017. "Catalytic Hydrotreatment of Jatropha Oil over Lanthanum Hydroxide Supported Noble Metals: Effect of Promotion with Cerium". ChemistrySelect. **2**(35), pp. 11918–25.

32 Patil, S.J., Vaidya, P.D., 2018. "On the production of bio-hydrogenated diesel over hydrotalcite-like supported palladium and ruthenium catalysts". Fuel Process. Technol. **169**(September 2017), pp. 142–9.

33 Hribernik, A., Kegl, B., 2009. "Performance and Exhaust Emissions of an Indirect-Injection (IDI) Diesel Engine When Using Waste Cooking Oil as Fuel". Energy & Fuels. **23**(3), pp. 1754–8.

34 Cao, F., Chen, Y., Zhai, F., Li, J., Wang, J., Wang, X., Wang, S., Zhu, W., 2008. "Biodiesel production from high acid value waste frying oil catalyzed by superacid heteropolyacid". Biotechnol. Bioeng. **101**(1), pp. 93–100.

35 Guzman, A., Torres, J.E., Prada, L.P., Nuñez, M.L., 2010. "Hydroprocessing of crude palm oil at pilot plant scale". Catal. Today. **156**(1–2), pp. 38–43.

36 Bezergianni, S., Dimitriadis, A., Kalogianni, A., Knudsen, K.G., 2011. "Toward Hydrotreating of Waste Cooking Oil for Biodiesel Production, effect of pressure, H2/Oil Ratio, and Liquid Hourly Space Velocity". Ind. Eng. Chem. Res. **50**(7), pp. 3874–9.

37 Bezergianni, S., Dimitriadis, A., Kalogianni, A., Pilavachi, P.A., 2010. "Hydrotreating of waste cooking oil for biodiesel production. Part I: Effect of temperature on product yields and heteroatom removal". Bioresour. Technol. **101**(17), pp. 6651–6.

38 Bezergianni, S., Kalogianni, A., Dimitriadis, A., 2012. "Catalyst evaluation for waste cooking oil hydroprocessing". Fuel. **93**, pp. 638–41.

39 Mansur, D., Ruliana, Rustana, C.E., "Hydroprocessed Calophyllum inophyllum Oil for Linear Bio-alkane Fuel Production" Proceedings of the Conference on the Industrial and Commercial Use of Energy, ICUE. Asian Institute of Technology. pp 1–6.

40 Mishra, D.K., Dabbawala, A.A., Hwang, J.S., 2013. "Ruthenium nanoparticles supported on zeolite y as an efficient catalyst for selective hydrogenation of xylose to xylitol". J. Mol. Catal. A Chem. **376**, pp. 63–70.

41 Mishra, D.K., Dabbawala, A.A., Park, J.J., Jhung, S.H., Hwang, J.S., 2014. "Selective hydrogenation of d-glucose to d-sorbitol over HY zeolite supported ruthenium nanoparticles catalysts". Catal. Today. **232**, pp. 99–107.

42 Sonthalia, A., Kumar, N., 2019. "Comparison of fuel characteristics of hydrotreated waste cooking oil with its biodiesel and fossil diesel". Environ. Sci. Pollut. Res. .

43 Ruinart de Brimont, M., Dupont, C., Daudin, A., Geantet, C., Raybaud, P., 2012. "Deoxygenation mechanisms on Ni-promoted MoS2 bulk catalysts: A combined experimental and theoretical study". J. Catal. **286**, pp. 153–64.

44 Liu, Q., Zuo, H., Wang, T., Ma, L., Zhang, Q., 2013. "One-step hydrodeoxygenation of palm oil to isomerized hydrocarbon fuels over Ni supported on nano-sized

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

SAPO-11 catalysts". Appl. Catal. A Gen. **468**, pp. 68–74.

45 Chen, L., Li, H., Fu, J., Miao, C., Lv, P., Yuan, Z., 2016. "Catalytic hydroprocessing of fatty acid methyl esters to renewable alkane fuels over Ni/HZSM-5 catalyst". Catal. Today. **259**, pp. 266–76.

46 Nakagawa, Y., Liu, S., Tamura, M., Tomishige, K., 2015. "Catalytic Total Hydrodeoxygenation of Biomass-Derived Polyfunctionalized Substrates to Alkanes". ChemSusChem. **8**(7), pp. 1114–32.

47 Nimkarde, M.R., Vaidya, P.D., 2016. "Toward Diesel Production from Karanja Oil Hydrotreating over CoMo and NiMo Catalysts". Energy and Fuels. **30**(4), pp. 3107–12.

48 Yang, Y., Wang, Q., Zhang, X., Wang, L., Li, G., 2013. "Hydrotreating of C18 fatty acids to hydrocarbons on sulphided NiW/SiO2–Al2O3". Fuel Process. Technol. **116**, pp. 165–74.

49 Lu, S.-M., Li, Y.-C.M., Tang, J.-C., 2003. "Optimum design of natural-circulation solar-water-heater by the Taguchi method". Energy. **28**(7), pp. 741–50.

50 Wu, H.-W., Wu, Z.-Y., 2012. "Combustion characteristics and optimal factors determination with Taguchi method for diesel engines port-injecting hydrogen". Energy. **47**(1), pp. 411–20.

51 Datta, S., Bandyopadhyay, A., Pal, P.K., 2006. "Desirability Function Approach for Solving Multi-Objective Optimization Problem in Submerged Arc Welding". J. Manuf. Sci. Prod. **7**, pp. 127.

52 Kumar, P., Barua, P.B., Gaindhar, J.L., 2000. "Quality optimization (multi-characteristics) through Taguchi's technique and utility concept". Qual. Reliab. Eng. Int. **16**(6), pp. 475–85.

53 Datta, S., Bandyopadhyay, A., Pal, P.K., 2008. "Grey-based taguchi method for optimization of bead geometry in submerged arc bead-on-plate welding". Int. J. Adv. Manuf. Technol. **39**(11), pp. 1136–43.

54 Yao, A.W.L., Chi, S.C., Chen, C.K., 2005. "Development of an integrated Grey–fuzzy-based electricity management system for enterprises". Energy. **30**(15), pp. 2759–71.

55 Lin, J., Wang, K., Yan, B., Tarng, Y., 2000. "Optimization of the electrical discharge machining process based on the Taguchi method with fuzzy logics". J. Mater. Process. Technol. **102**(1), pp. 48–55.

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Figure Captions List

Fig. 1 Schematic of the setup for fuel production

Fig. 2 XRD pattern of the catalyst

Fig. 3 SEM images of the catalyst at (a) 20µm, (b) 10µm, (c) 5µm and (d) EDS analysis

Fig. 4 Conversion and Diesel yield at various test conditions

Fig. 5 Molar ratio of C_{17}/C_{18} at different test conditions

Fig. 6 Membership function for conversion

Fig. 7 Membership function for multi performance characteristics index

Fig. 8 MPCI graph

Table Caption List

Table 1 Process Parameters and their range

Table 2 Factors and their levels

Table 3 Experimental layout using an L_9 orthogonal array

Table 4 Fuzzy rule matrix

Table 5 S/N ratio for the three variables and the inferred MPCI

Table 6 MPCI Table

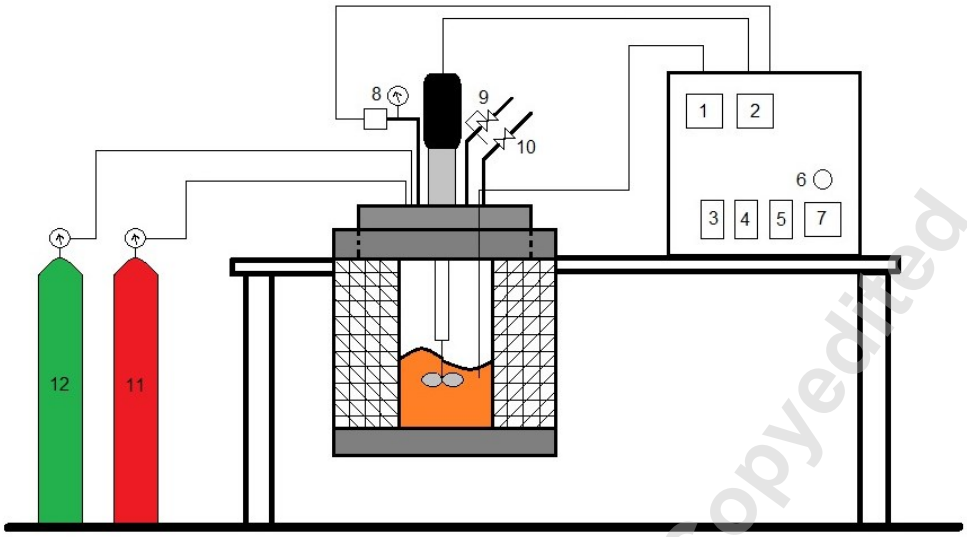
Table 7 ANOVA results

Table 8 Result of confirmation experiment

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME



(1) Temperature indicator and controller; (2) Pressure Indicator; (3) Heater Switch; (4) Motor Switch; (5) coolant switch; (6) motor speed controller; (7) motor speed display; (8) pressure sensor; (9) safety valve; (10) purging /gas collection line; (11) hydrogen cylinder; (12) nitrogen cylinder

Fig. 1 Schematic of the setup for fuel production

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

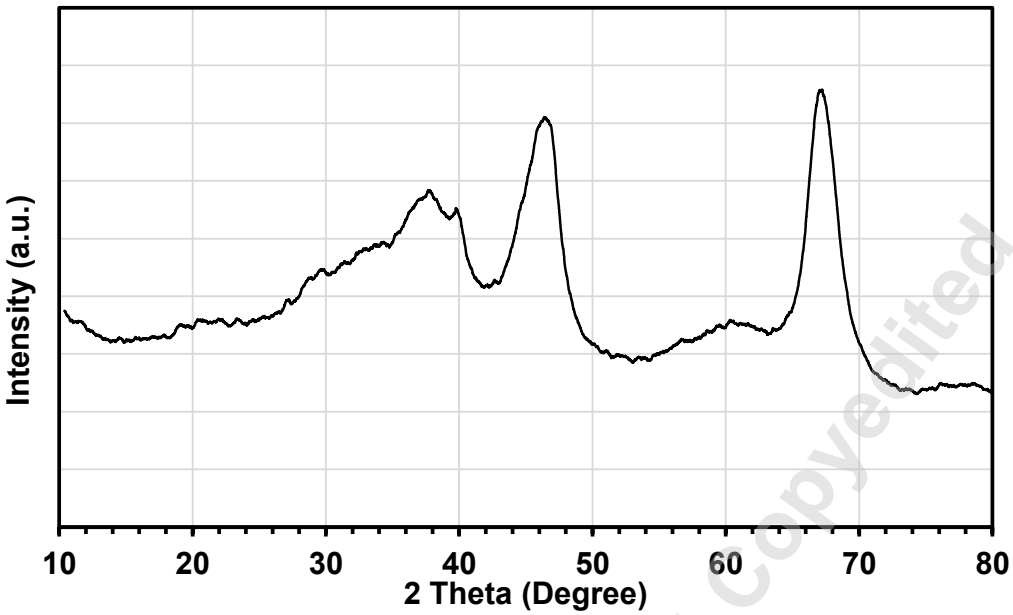


Fig. 2 XRD pattern of the catalyst

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

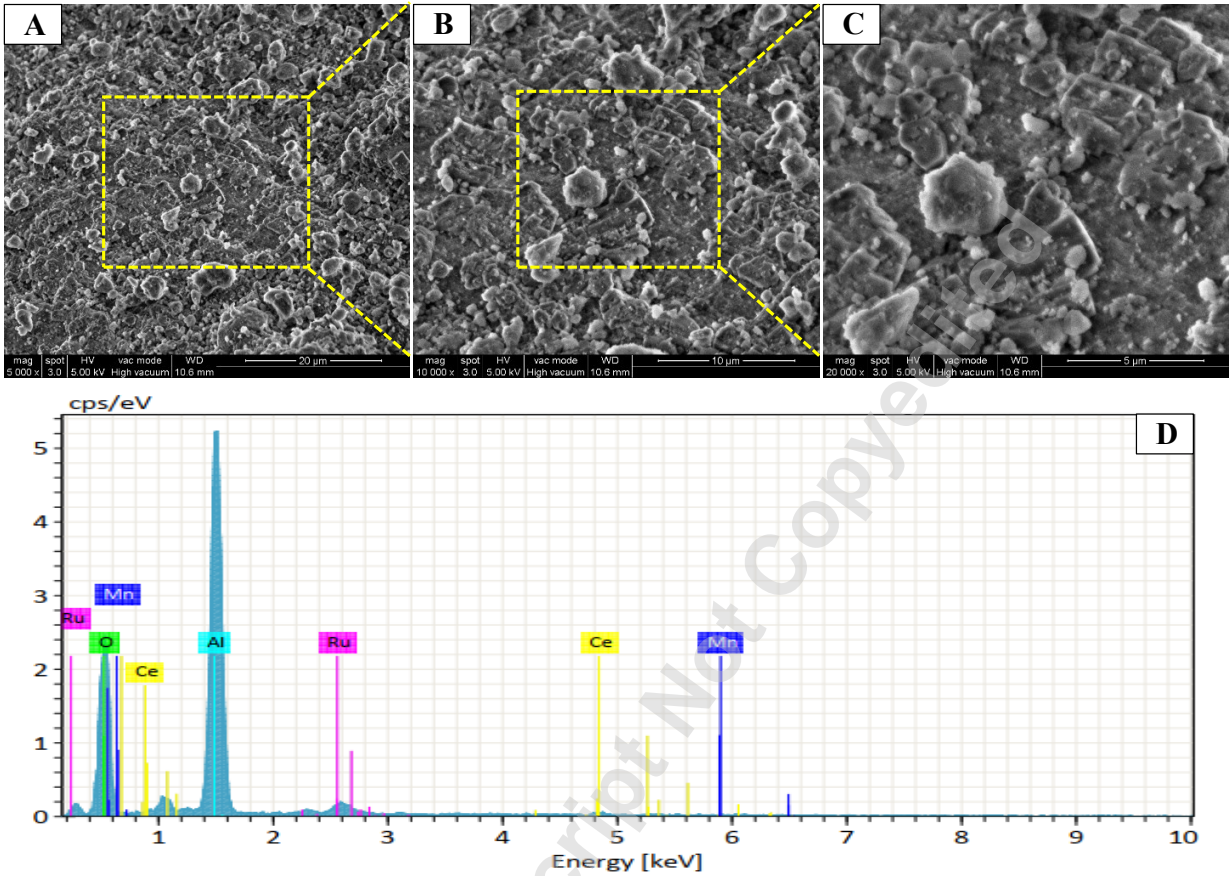


Fig. 3 SEM images of the catalyst at (a) 20µm, (b) 10µm, (c) 5µm and (d) EDS analysis

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

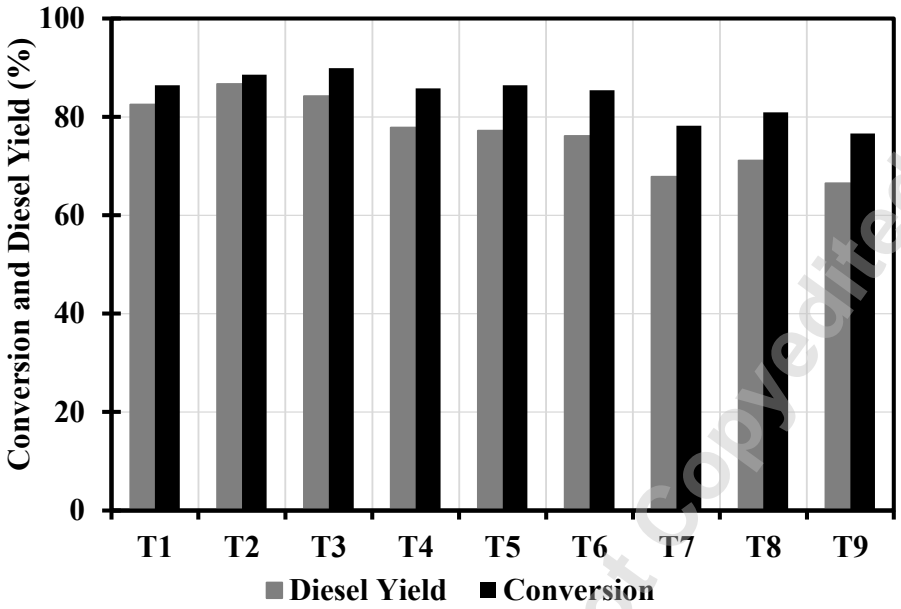


Fig. 4 Conversion and Diesel yield at various test conditions

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

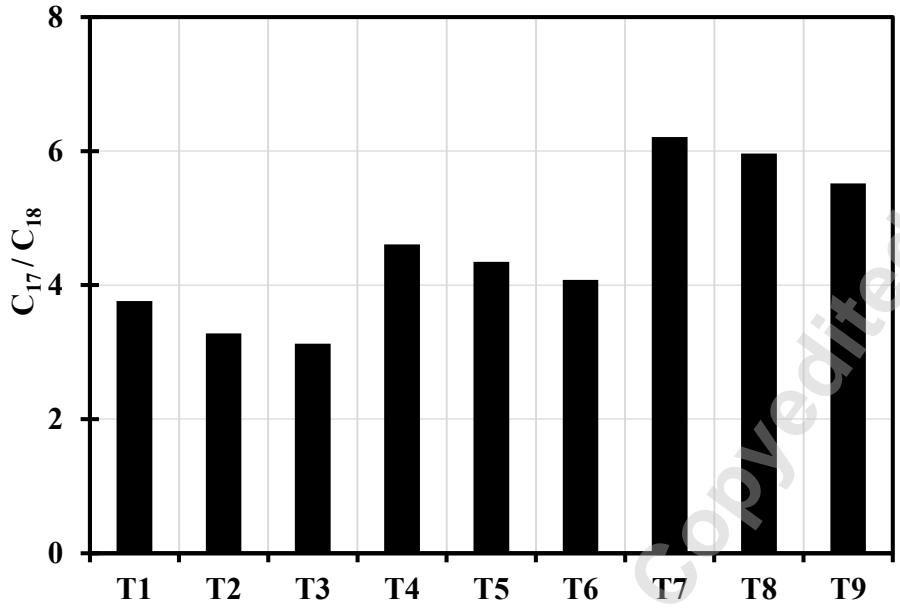


Fig. 5 Molar ratio of C_{17}/C_{18} at different test conditions

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

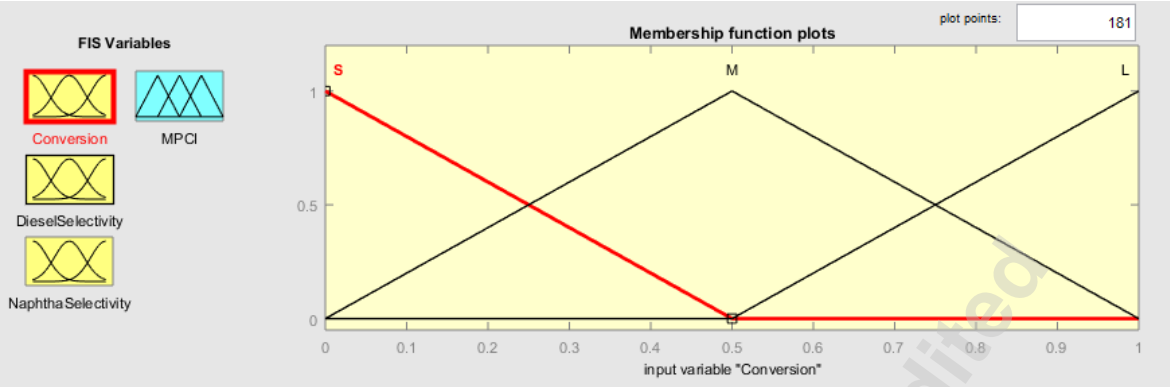


Fig. 6 Membership function for conversion

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

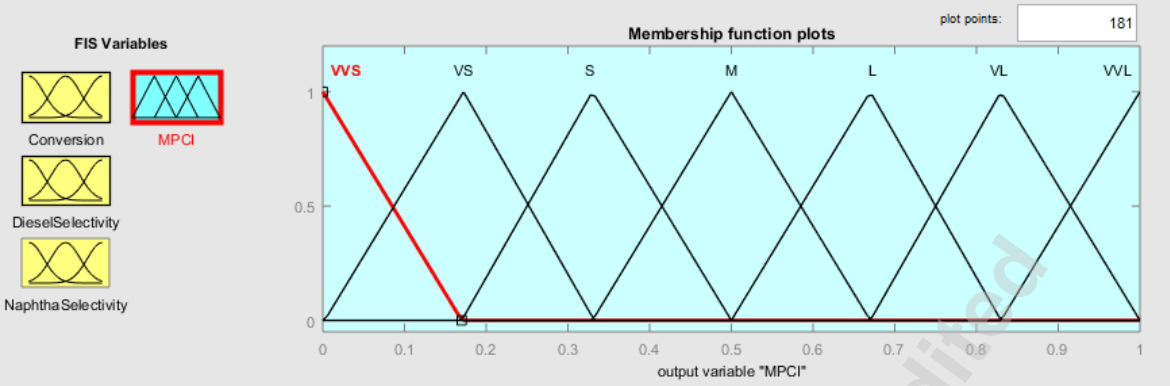


Fig. 7 Membership function for multi performance characteristics index

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

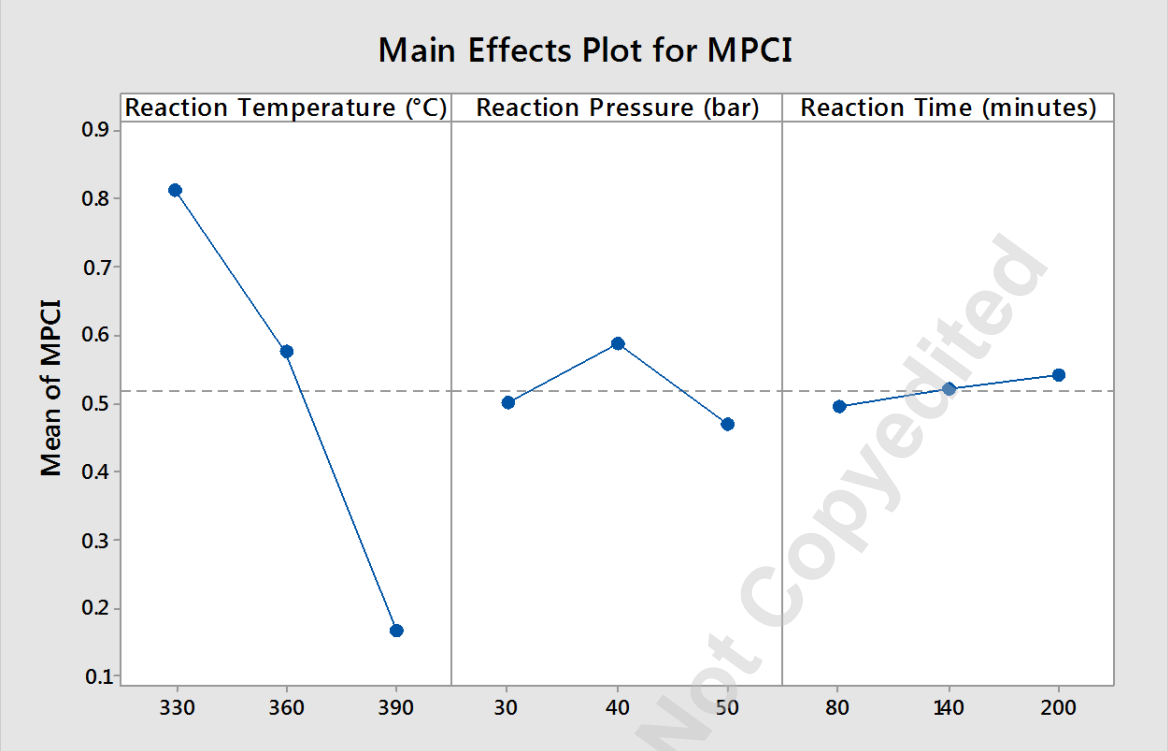


Fig. 8 MPCI graph

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 1. Process Parameters and their range

Process Parameter	**Range**
Reaction Temperature (°C)	330 – 390
Reaction Pressure (bar)	30 – 50
Reaction Time (minutes)	80 – 200

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 2. Factors and their levels

Factors	**Level 1**	**Level 2**	**Level 3**
Reaction Temperature (°C)	330	360	390
Reaction Pressure (bar)	30	40	50
Reaction Time (minutes)	80	140	200

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 3. Experimental layout using an L_9 orthogonal array

Test Case	Reaction Temperature (°C)	Reaction Pressure (bar)	Reaction Time (minutes)
T1	1	1	1
T2	1	2	2
T3	1	3	3
T4	2	1	2
T5	2	2	3
T6	2	3	1
T7	3	1	3
T8	3	2	1
T9	3	3	2

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 4. Fuzzy rule matrix

S. No.	Conversion (%)	Diesel Selectivity (%)	Naphtha Selectivity (%)	Output
1	S	S	S	VVS
2	S	S	M	VS
3	S	S	L	S
4	S	M	S	VS
5	S	M	M	S
6	S	M	L	M
7	S	L	S	S
8	S	L	M	M
9	S	L	L	L
10	M	S	S	VS
11	M	S	M	S
12	M	S	L	M
13	M	M	S	S
14	M	M	M	M
15	M	M	L	L
16	M	L	S	M
17	M	L	M	L
18	M	L	L	VL
19	L	S	S	S
20	L	S	M	M
21	L	S	L	L
22	L	M	S	M
23	L	M	M	L
24	L	M	L	VL
25	L	L	S	L
26	L	L	M	VL
27	L	L	L	VVL

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 5. S/N ratio for the three variables and the inferred MPCI

Test Case	Reaction Temperature (°C)	Reaction Pressure (bar)	Reaction Time (minutes)	Conversion (%)	S/N Ratio (Conversion)	Diesel Selectivity (%)	S/N Ratio (Diesel selectivity)	Naphtha Selectivity (%)	S/N Ratio (Naphtha selectivity)	MPCI
T1	330	30	80	86.4	38.73	82.5	38.33	1.8	-5.11	0.704
T2	330	40	140	88.6	38.95	86.7	38.76	1.2	-1.58	0.899
T3	330	50	200	89.9	39.08	84.2	38.51	1.4	-2.92	0.838
T4	360	30	140	85.8	38.67	77.8	37.82	3.1	-9.83	0.512
T5	360	40	200	86.4	38.73	77.2	37.75	3.3	-10.37	0.500
T6	360	50	80	85.4	38.63	76.1	37.63	4.1	-12.26	0.518
T7	390	30	200	78.2	37.86	67.8	36.62	8.0	-18.06	0.187
T8	390	40	80	80.9	38.16	71.1	37.04	7.6	-17.62	0.163
T9	390	50	140	76.6	37.68	66.5	36.46	8.4	-18.49	0.153

Accepted Manuscript Not Copyedited

Downloaded from http://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 6. MPCI Table

Process Parameter	MPCI			
	Level 1	**Level 2**	**Level 3**	**Max-Min**
Reaction Temperature (°C)	0.814	0.577	0.167	0.647
Reaction Pressure (bar)	0.501	0.587	0.469	0.118
Reaction Time (minutes)	0.495	0.521	0.542	0.047
	Mean Value of MPCI = 0.519			

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 7. ANOVA results

Process Parameter	Degrees of Freedom	Sum of Squares	Mean Squares	F	Contribution (%)
Reaction Temperature (°C)	2	0.6407	0.32035	28.89	93.1
Reaction Pressure (bar)	2	0.0222	0.0111	1	3.23
Reaction Time (minutes)	2	0.0032	0.0016	0.15	0.46
Error	2	0.0221	0.01105		3.21
Total	8	0.6882			100

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Journal of Energy Resources Technology. Received November 26, 2019;
Accepted manuscript posted May 02, 2020. doi:10.1115/1.4047405
Copyright © 2020 by ASME

Table 8. Result of confirmation experiment

	Initial Process Parameters		**Optimal Process Parameters**			**Improvement in S/N Ratio**	**Improvement in MPCI**
			Prediction	**Experiment**			
Setting Level	390°C, 50bar & 140minutes	S/N ratio	330°C, 40bar & 200minutes	330°C, 40bar & 200minutes	S/N ratio		
Conversion (%)	76.6	37.68		89.7	39.056	1.376	
Diesel Selectivity (%)	66.5	36.46		88.2	38.91	2.45	
Naphtha Selectivity (%)	8.4	-18.49		1.1	-0.82	17.67	
MPCI	0.053		0.904	0.940			0.887

Accepted Manuscript Not Copyedited

Downloaded from https://asmedigitalcollection.asme.org/energyresources/article-pdf/doi/10.1115/1.4047405/6539622/jert-19-1815.pdf by Delhi Technological University user on 08 June 2020

Plasma Chemistry and Plasma Processing
https://doi.org/10.1007/s11090-020-10090-2

ORIGINAL PAPER

Parametric Study of Plasma Characteristics and Carbon Nanofibers Growth in PECVD System: Numerical Modeling

Ravi Gupta[1] · **Suresh C. Sharma**[1]

Received: 16 December 2018 / Accepted: 21 May 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

The aim of the present work is to develop a numerical model to understand and optimize the process parameters for the growth of carbon nanofibers (CNFs) inside the plasma enhanced chemical vapor deposition system containing acetylene, hydrogen, and argon gases. Two-dimensional axis-symmetrical inductive couple plasma module is implemented using COMSOL Multiphysics 5.2 simulation software to analyze the density profiles and temperatures of electrons, ions, and neutral species in the plasma at different gas pressures and input plasma powers. The outcomes of the COMSOL computational model show that the electron density in the plasma decreases with increase in gas pressure and increases with increase in plasma power. Other than the computational model, an analytical model is developed in the present paper that accounts the plasma sheath equations to study the fluxes and energies of the plasma species. The results obtained from the plasma sheath model at the catalyst-substrate surface boundary are fed as the input parameters to surface deposition model to investigate the growth characteristics of carbon nanofibers, i.e., poisoning of the catalyst nanoparticle, height, and diameter of carbon nanofiber at different gas pressures and input plasma powers. It is found that electron density decays at the faster rate when gas pressure is increased and decays at slower rate when input plasma power is raised. Moreover, it is also found that growth rate of CNFs increases with increase in gas pressure and plasma power. However, the significant drop in CNF growth rate is observed when the gas pressure is high enough (above and around 50 Torr). From the results obtained, it can be concluded that the CNFs having good growth characteristics can be obtained at some optimum pressure range, i.e., one order of the magnitude in the units of Torr. A good comparison between numerical simulation results and analytical results with each other and with existing experimental results confirms the adequacy of the computational and analytical approach.

Keywords Carbon nanofibers · PECVD · Process parameters · Pressure · Plasma power · COMSOL

✉ Suresh C. Sharma
suresh321sharma@gmail.com

[1] Department of Applied Physics, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Delhi 110042, India

Published online: 30 May 2020

Introduction

Carbon nanofibers (CNFs) and carbon nanotubes (CNTs) characterize the unique forms of carbon in nanometer regime. Their exceptional properties such as high aspect ratios, high mechanical strength, and high chemical stability make them to be promising as field emission devices [1–5]. To obtain the enhanced field emission characteristics, the vertically oriented CNFs/CNTs are grown on catalyst substrate surface by plasma enhanced chemical vapor deposition (PECVD) technique [6, 7]. PECVD process uses electrical energy to generate a glow discharge (plasma) in which the energy is transferred into a gas mixture which transforms the gas mixture into reactive radicals, ions, neutral atoms, molecules, and other highly excited species. The areas with uncompensated surface charges, i.e., plasma sheaths are formed near the boundaries of the plasma exposed surfaces (i.e., catalyst-substrate surface and chamber walls). Therefore, nanostructures are grown in the plasma sheath region rather than bulk plasma. The electric field due to the plasma sheath near the catalyst-substrate surface has the significant role in the growth process. The strength of the electric field at the surface controls the kinetics of the ionized species and their deposition on the catalyst surface. The process parameters significantly affect the sheath dynamics which consequently affects the electric field strength [8–11].

The growth characteristics, i.e., nanostructure height and diameter and resulting structure of CNFs/CNTs depend on the fluxes and the energies of the radicals, ions, and neutral species from the bulk plasma to the catalyst nanoparticle surface. These fluxes and energies can be controlled by many operating parameters such as; plasma power, gas pressure, gas ratio, substrate temperature, and substrate bias, etc. Among them, gas pressure and plasma power directly influence the plasma sheath width, the strength of the electric field in the sheath region, densities, temperature, and energies of plasma species, and collision between plasma species, which in turn affects the growth characteristics of the CNFs and CNTs [7, 12–29].

In our previous works, we developed the analytical model and focused on the nucleation and growth mechanism of the CNFs in the PECVD process and studied the effect of metal catalyst, nitrogen doping, and presence of water molecules and oxygen in the plasma on the nucleation and growth of the CNFs [30–33]. However, in the present paper, we performed the simulations using COMSOL Multiphysics 5.2 simulational software to get deep insights into the plasma modifications, i.e., change in number densities of various plasma species and their temperatures due to the change in PECVD process parameters particularly input plasma power and gas pressure. Simultaneously, the gas discharge model accounting the sheath equations and volumetric reactions in the bulk plasma is also developed for the above mentioned purpose to verify whether the simulational and analytical model give the similar results or not. Moreover, the effects of the plasma modifications such as change in number density and temperature of various plasma species due to change in input plasma power and gas pressure on the growth characteristics of the CNFs grown over the nickel nanoparticles placed over the silicon substrate surface in inductively coupled plasma (ICP) chamber reactor have been studied using the surface deposition model developed in the present paper that accounts all the surface reactions of various gaseous species on the catalyst surface. The simulation results and model validation was conducted against the experimental observations available in the literature.

Model

The plasma enhanced chemical vapor deposition of the CNFs onto the nickel catalyst nanoparticles placed over silicon substrate via acetylene, hydrogen, and argon in the plasma reactor—has been investigated through the computational as well as analytical model. The ionized and neutral species accounted in the present model on the basis of experimentally specified plasma chemistry of the acetylene, hydrogen, and argon gas mixture [19, 34–39] are presented in Table 1. The volumetric reactions in the bulk plasma between various plasma species (considered in the present model) are obtained from the Refs. [19, 39].

Computational Approach

The computational model to simulate the plasma modifications was constructed in the COSMSOL Multiphysics modeling suite. The detailed approach used in the present paper are;

1. *Magnetic field interface* When time varying electric current, high power radio frequency (RF) waves having frequency of 13.6 MHz, is passed through the coils, the time varying magnetic field is generated in the chamber. Due to induction effect, time varying current is passed through the gases present in the chamber and initiate the generation of plasma. The antenna configuration of coils is taken into account to generate low temperature plasma. This interface gives the strength of the magnetic field as the function of the space coordinates in the plasma chamber on accounting the transformer effect of inductions coils using Maxwell's and other equations of the electro-magneto dynamics.
2. *Plasma interface (drift–diffusion interface)* Due to the transformer action of the coils, highly energetic electrons are generated in the plasma interact with gases via various volumetric reactions in the plasma and generate various ions and radicals. The characteristics of plasma, i.e., electron density, electron temperature, and electron potential can be evaluated using drift–diffusion interface. Loss of hydrogen due to interaction with chamber walls and generation of electrons due to ion impact reactions have neglected in the present model.
3. *Heavy Species Transport Interface* This interface involves the generation of the non-electron species such as ions and neutrals of the hydrocarbons and hydrogen in the plasma due to the various reactions in the plasma that takes place experimentally. This interface also considers the interaction of non-electron species from the plasma to the substrate surface. The nickel coated silicon (100) in pre-plasma condition is taken into account in present model. It is assumed that ions get completely neutralized or get absorbed when interact with the chamber walls. However, it is assumed that when excited hydrogen and argon species interact with chamber walls they revert back to their ground state.

Table 1 Species considered in the present model

Type	Neutrals	Positively charged species
X	C_2H, C_2H_2, C_2H_3, C_4H, C_4H_2, C_4H_3 C_6H, C_6H_2, C_6H_4, C_8H, C_8H_2, C_8H_4, $C_{10}H$, $C_{10}H_2$	C_2H^+, $C_2H_2^+$, $C_2H_3^+$, C_4H^+, $C_4H_2^+$, $C_4H_3^+$ C_6H^+, $C_6H_2^+$, $C_6H_4^+$, C_8H^+, $C_8H_2^+$, $C_8H_4^+$, $C_{10}H^+$, $C_{10}H_2^+$
Y	H_2, H	H_2^+, H^+

4. *Substrate Bias and laminar flow* This involves the direction of flow and fluxes of the ionic species from the bulk plasma region to catalyst substrate surface under the influence of electric field that exist between bulk plasma and negatively biased substrate surface. The convection of heavy species has been neglected.
5. *Electrostatics interface* This involves the electric field, space charge density, and induction current in the plasma chamber.

These interfaces are coupled in the 2D axisymmetric inductively coupled plasma (ICP) module in a frequency transient domain sketched schematically in Fig. 1. To capture the space charge region between electrons and ions near the chamber walls is important step in plasma computation. Therefore, the mesh with **Finer** standard size settings consisting quadratic triangle elements is adopted. 5 boundary layer with 1.4 stretching factor is taken into account. All the input parameters feed to the computational domain are listed in Table 2. The interfaces considered in the present computation involves multiple surface and volumetric reactions that takes place experimentally and involves several intermediate species. The COMSOL provides the inbuilt solver and user friendly graphical interface to solve the complex equations based on the above mentioned physical interfaces. The reduced transport properties are used to evaluate electron transport properties, i.e., $\mu_e = \frac{\mu_{rd}}{N}$, where N is the number density and μ_{rd} is the reduced mobility of electrons. The Maxwellian electron energy distribution function is adopted for the source coefficients, i.e., $f(\varepsilon) = \varphi^{-3/2}\beta_1 \exp(-(\frac{\varepsilon\beta_2}{\varphi}))$, where $\beta_1 = \Gamma\left[\frac{5}{2}\right]^{3/2}\Gamma\left[\frac{3}{2}\right]^{-5/2}$, $\beta_2 = \Gamma\left[\frac{5}{2}\right]\Gamma\left[\frac{3}{2}\right]^{-1}$, φ is the mean electron energy in eV, ε is the electron energy in eV, and Γ is gamma function.

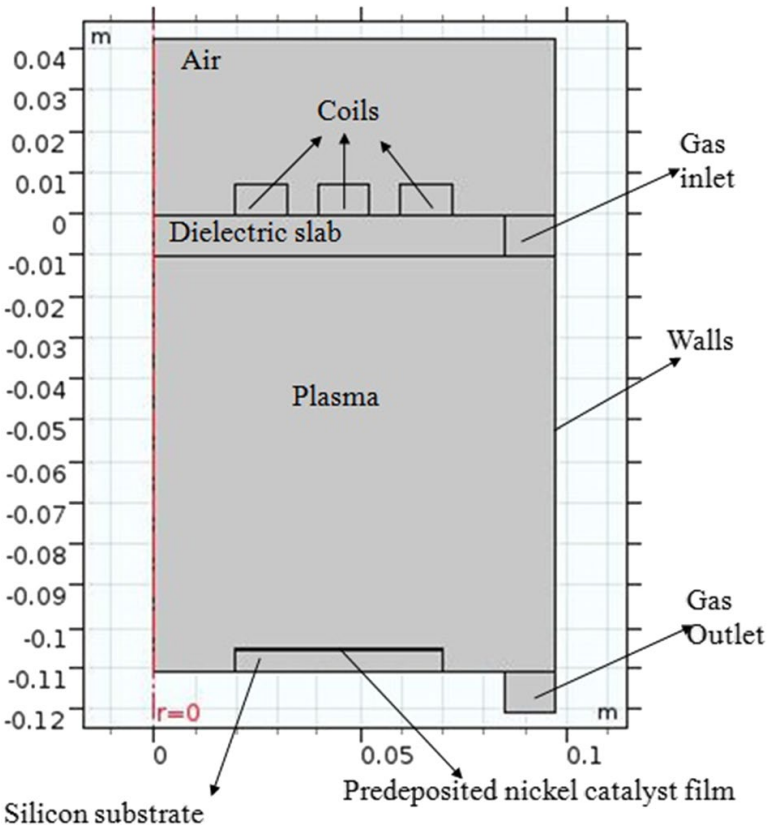


Fig. 1 Geometry of the model constructed in the COMSOL modeling suite which includes five turn inductive coils (source of inductively coupled plasma) connected to 13.56 MHz generator. A dielectric material is used between the coils and plasma reactor. The applied power to the coils reaches to the plasma reactor through this dielectric material. The pre-deposited nickel catalyst film over the silicon substrate is placed at the bottom of the plasma reactor

Table 2 Parameters fed into the computational model developed in the COMSOL Multiphysics modeling suite

Parameter	Value
Relative permittivity of vacuum	1
Relative permittivity of dielectric material	4.2
Relative permittivity of silicon substrate	12
Relative permeability of vacuum	1
Electrical conductivity of vacuum	0
Electrical conductivity of coils	6×10^{7} S m^{-1}
Electrical conductivity of dielectric material	0
Gas temperature	500 K
Reduced electron mobility	4×10^{24} V^{-1} m^{-1} s^{-1}
Substrate Bias	− 300 V

$$\frac{\partial(n_e)}{\partial t} + \nabla.\Gamma_e = \mathrm{R}_e, \tag{1a}$$

$$\frac{\partial(n_\varepsilon)}{\partial t} + \nabla.\Gamma_\varepsilon + E.\Gamma_e = \mathrm{R}_\varepsilon, \tag{2a}$$

$$\Gamma_e = -n_e(\mu_e E) - D_e \nabla n_e, \tag{3a}$$

$$\Gamma_\varepsilon = -n_\varepsilon(\mu_\varepsilon E) - D_\varepsilon \nabla n_\varepsilon, \tag{4a}$$

$$\mathrm{R}_e = \sum_{j=1}^{M} x_j k_j N_n n_e, \tag{5a}$$

$$\mathrm{R}_\varepsilon = \sum_{j=1}^{P} x_j k_j N_n n_e \Delta\varepsilon_j, \tag{6a}$$

$$k_j = \gamma \int_0^\infty \varepsilon \sigma_j(\varepsilon) f(\varepsilon) d\varepsilon, \tag{7a}$$

$$\gamma = (2q/m)^{1/2}, \tag{8a}$$

where Γe is the flux of diffusing electron termed as Fick's law of diffusion, where $\mathrm{D_e}$ is the diffusion constant. Re is the source term, where k_j is the rate coefficient for reaction j, $\Delta\varepsilon_j$ is the energy loss from reaction j, N_n is the number density of reactants, E is the electric field in the chamber, n_e is the electron density, m is the electron mass, ε is energy, σ_j is the collision cross section, x_j is the mole fraction for the target species of reaction j, and f is the electron energy distribution function. Poisson's equation $\left(-\nabla.\varepsilon_o\varepsilon_r\nabla V = \rho\right)$ is used to compute plasma density and plasma space charge density can be computed using equation $\rho = q(\sum_{k=1}^{M} Z_k n_k - n_e)$. The rate coefficients for surface reactions is $K_{f,i} = (\frac{\gamma_i}{1-\gamma_i/2})(\frac{1}{(\Gamma_{tot})^m})(\frac{1}{4})\sqrt{\frac{8RT_e}{\pi M_n}}$, where $\gamma_i = a_1 T_e^{b_1} \exp(-\frac{e_i}{RT_e})$ is the sticking coefficient of the ith species, $\mathrm{M_n}$ is the mean molecular mass of the gas mixture, R is the gas constant, $\mathrm{T_e}$

is the electron temperature, and a_1 and b_1 are constants. The velocity with which building units (carbon species) diffuse along the substrate interface is $v = \frac{\vec{j}_m.\hat{n}}{\rho}$, where v is the interface velocity and $\vec{j}_m.\hat{n}$ is the mass flux in vertical downward direction and ρ is density of species.

Using the above plasma module, simulations are performed for $Ar/C_2H_2/H_2$ plasma. The plasma species and chemical reactions considered in the present simulation are justified in order to generate 2D analysis of the number densities and fluxes of various species involved which are vital for nano assembly of selected carbon-based nanostructures. The limited species have been accounted to reduce the computational coast and duration.

Analytical Approach

In the PECVD technique, the CNFs and other carbon nanostructures are grown in the plasma sheath region formed between the bulk plasma and the substrate surface. The processes in the bulk plasma and on the catalyst substrate surface affect the nanostructure nucleation and formation. The sheath equations are used to find the number densities, temperatures, and energies of the plasma species in the plasma sheath region. However, the interaction of the plasma species with the catalyst substrate surface and the growth of CNFs have been accounted in the surface deposition model. The outcomes of sheath equations serve the initial values for the surface deposition model. The sheath equations account the continuity equation (Eqs. 1 and 2), momentum equation (Eqs. 3 and 4), and Poisson's equation (Eq. 5) to investigate the dynamics of the plasma species [8, 40],

$$\frac{d}{dz}(n_e u_e) = v_i n_e, \tag{1}$$

$$\frac{d}{dz}(n_i u_i) = v_i n_i, \tag{2}$$

where n_e and n_i are the number densities of electrons and positively charged species considered (assumed to be singly charged) in the present model, respectively, u_e and u_i are the velocities of electrons and positively charged species along the plasma sheath, respectively, $v_i (= \varepsilon_i r_i n_n)$ is the ionization frequency of the ions, ε_i is the ionization potential of the positively charged species, r_i is the jth gas pressure to total gas pressure ratio $\left(\frac{p_j}{p_T}\right)$ or ith type positively charged species number density to electron density ratio $\left(\frac{n_i}{n_e}\right)$, and n_n is the neutral atoms number density considered in the present model.

$$\frac{d}{dz}(u_e) = -\frac{e}{u_e}\frac{d\phi}{dz} - \frac{T_e}{n_e u_e}\frac{dn_e}{dz} - m_e v_{en}, \tag{3}$$

$$\frac{d}{dz}(u_i) = \frac{e}{u_i}\frac{d\phi}{dz} - \frac{T_i}{n_i u_i}\frac{dn_i}{dz} - m_i v_{in}, \tag{4}$$

$$\frac{d^2}{dz^2}(\phi) = 4\pi e(n_e - \sum r_i n_i), \tag{5}$$

where $v_{en}\left(=\sigma n_n \mathrm{v}_e\right)$ and $v_{in}\left(=\sigma n_n \mathrm{v}_i\right)$ are the collision frequencies of the electrons and ions with the neutrals respectively, σ is the collision cross-section, $\mathrm{v}_e\left(=\sqrt{\frac{8k_BT_e}{m_e\pi}+u_e^2}\right)$ and $\mathrm{v}_i\left(=\sqrt{\frac{8k_BT_i}{m_i\pi}+u_i^2}\right)$ are the mean velocities of electrons and ions, ϕ is the plasma sheath potential. The above-mentioned sheath equations have been solved for appropriate boundary conditions at the sheath edge such as; $u_e = 0$, $u_i\left[=\sqrt{\frac{k_BT_e}{m_i}}\right]$ is the velocity with which ions enter in the plasma sheath with the bulk plasma, $n_e = \sum r_i n_i$ (quasi-neutrality condition), $\sum r_i = 1$, $0 < r_i < 1$, $\phi = 0$, $\frac{d\phi}{dz} = -\frac{T_e}{e\lambda_i}$, $\lambda_i\left(=\frac{T_n}{\sigma p_T}\right)$ is the mean free path of the positively charged species.

The synthesis of the CNFs involves three main steps; (1) nucleation stage, in which carbon-containing species from plasma dissociates on the catalyst front surface that is exposed to plasma, (2) surface and bulk diffusion of carbon species generated on the catalyst surface, and (3) precipitation of diffused carbon species in the form of graphitic shells at the catalyst interface [30]. The CNF growth model incorporates numerous terms and equations corresponding to various complex processes and growth stages. The flux of the hydrocarbon and hydrogen species (ions and neutrals) from the plasma sheath on the catalyst surface generates carbon species and hydrogen radicals on the catalyst surface via various complex processes. The Eq. (6) corresponds to the number of carbon species per unit area per unit time generated on the catalyst active surface due to the hydrocarbon neutrals adsorption on the catalyst active surface (first term), thermal dissociation of adsorbed hydrocarbon neutrals on the catalyst active surface (second term), ion induced dissociation of hydrocarbon neutrals (third term), dissociation of incoming ions of the plasma sheath (fourth term), interaction of hydrocarbon ion and hydrogen ion (fifth term), loss of adsorbed hydrocarbons due to interaction with incoming hydrogen from the plasma (sixth term), carbon species evaporation from the catalyst surface (seventh term), desorption of hydrocarbon neutrals from catalyst surface (eighth term), and interaction between carbon and hydrogen species generated on the catalyst nanoparticle surface (ninth term) [9–11].

The number of carbon species generated on catalyst surface is one of the growth rate determining factors. A possible explanation for the change in growth rate is the correlation between the generations of the carbon species to the growth of amorphous carbon layer on the active surface of the catalyst. The rapid increase in number of carbon species enhances the growth of amorphous carbon layer which ultimately poisons the catalyst and slows down the dissociation rate of incoming hydrocarbon species. The term $\frac{d\Omega_c}{dt}\left(=1-n_c\pi D^2\right)$ in Eq. (6) denotes the growth rate of the amorphous carbon layer on the catalyst surface.

$$\begin{aligned}\frac{\partial n_C}{\partial t} = \Bigg[& \sum_X J_X(1-\theta_t) + \sum_X n_{sX}\upsilon \exp\left(\frac{-\varepsilon_{td}}{k_BT_S}\right) + \sum_{iX}\left(\sum_X \frac{n_{sX}y_d}{\upsilon_0}\right)J_{iX} + \sum_{iX} J_{iX} \\ & + \sum_{iY}\left(\sum_{iX}\frac{J_{iX}\sigma_{ads}}{\upsilon}\right)J_{iY} - \sum_Y\left(\sum_X n_{sX}\sigma_{ads}\right)J_Y - n_C\upsilon\exp\left(\frac{-\varepsilon_{ev}}{k_BT_S}\right) \\ & - \sum_X n_X\upsilon\exp\left(\frac{-\varepsilon_{dhc}}{k_BT_S}\right) - n_C\sigma_{ads}\upsilon n_H\Bigg] \times \frac{d\Omega_C}{dt},\end{aligned} \tag{6}$$

where n_C is the surface concentration of carbons, J_j and J_{ij} are the neutrals and ions flux obtained from the plasma sheath equations, respectively [10] (where j refers to the species of type X and Y), $n_{sj}\left(=\theta_j\upsilon_0\right)$ is the surface concentration of neutral species adsorbed on the catalyst surface from the plasma sheath [9–11], θ_j is the surface coverage by jth

species, υ_0 is the adsorption sites per unit area, υ is the thermal vibration frequency, $\sigma_{ads}\left(=6.8\times10^{-16}\ \mathrm{cm}^2\right)$ is the cross-section for the interaction among various species [10], $\varepsilon_{td}(=2.1\ \mathrm{eV})$ is the hydrocarbons dissociation energy, $\varepsilon_{\mathrm{ev}}(=1.8\ \mathrm{eV})$ carbon evaporation energy, $\varepsilon_{dhc}(=1.8\ \mathrm{eV})$ is the hydrocarbon desorption energy [41], and D is the diameter of the catalyst nanoparticle.

The relative number of hydrogen radicals generated on the catalyst surface is also an important parameter to change the growth process. These highly reactive hydrogen radicals interact with the adsorbed hydrocarbons and carbon species and prevent the early poisoning of the catalyst by slowing down the formation of the amorphous carbon layer. The Eq. (7) is indicative to the number of hydrogen radicals per unit time generated on the catalyst surface due to the adsorption of hydrogen species on the catalyst surface from the plasma [9] (first term), ion-induced dissociation of hydrogen (second term), generation of hydrogen radicals due to the thermal dissociation of hydrocarbons (third term) [41], direct hydrogen ion flux (fourth term), desorption of hydrogen radicals from catalyst (fifth term), interaction of hydrogen radicals with carbon species generated on the catalyst surface (sixth term), and interaction of adsorbed hydrogen radicals with influx of hydrogen neutrals and ions from the plasma (seventh and eight term) [41].

$$\begin{aligned}\frac{\partial n_H}{\partial t} &= \sum_Y J_Y(1-\theta_t)+\sum_{iX}\left(\sum_X \frac{n_{sX}y_d}{\upsilon_0}\right)J_{iX}+\sum_Y n_{sY}\upsilon\exp\left(\frac{-\varepsilon_{td}}{k_BT_S}\right)+\sum_{iY}J_{iY}\\ &\quad-\sum_Y n_{sY}\upsilon\exp\left(\frac{-\varepsilon_{dH}}{k_BT_S}\right)-n_C\sigma_{ads}\upsilon n_H-\sum_Y n_{sY}\sigma_{ads}J_Y-\sum_Y n_{sY}\sigma_{ads}J_{iY},\end{aligned} \tag{7}$$

where n_H is the surface concentration of hydrogen radicals, $\delta E_{dH}(=1.8\ \mathrm{eV})$ is the hydrogen atom desorption energy [41].

After the nucleation stage, the growth of the CNF is initiated that includes various processes which are clearly explained in volumetric growth rate equation of CNF (Eq. 8). The carbon species n_C dissolve in and diffuse through catalyst nanoparticle (first and second term), and then carbon layers precipitate from the rear sides of the catalyst nanoparticle to from inclined graphitic shells (third term). These graphitic shells, i.e., precipitated carbon species exert varying compressive stress on the catalyst nanoparticle due to which it undergoes shape deformation [42, 43]. The restructuring of the catalyst particle during the CNF growth is the crucial phenomenon this is because the extents up to which catalyst particle can reconstruct itself determine the diameter (d) of the CNF and number of graphitic shells in the CNF. Therefore, the first three terms of Eq. (8) are multiplied by the stress (P) applied by the growing graphitic shells. Moreover, self-diffusion of atoms of metal catalyst nanoparticle along the growth direction also makes catalyst to reconstruct itself (fourth term). Apart from the carbon deposition process, the etching by hydrogen is also an important step in CNF growth (last term). The hydrogen radicals etch carbons atoms present at the edges of the shells; however, carbon atoms under the geometrical shadow region of the catalyst particle contribute to the vertical growth of the graphitic shells.

$$\frac{d}{dt}\pi(D^2-d^2)h=\left[D_{b0}\exp\left(-\varepsilon_b/k_BT_S\right)+D_{S0}\exp\left(-\varepsilon_S/k_BT_S\right)+A_k\exp\left(-\delta\varepsilon_{inc}/k_BT_S\right)\nu D\right]\times\frac{P(4\pi D)n_C}{\nu\rho_{cnf}I_{ix}}+\left[\frac{D_m}{\pi D^2}+n_H(\pi D^2)\sum_Y I_{iY}\right]\times\frac{m_{cat}}{\rho_{cat}},\tag{8}$$

where D and h are the diameter of the catalyst nanoparticle and height of the CNF, respectively, $D_{b0}=\nu D/2\pi$, ε_b (=1.6 eV) is the energy barrier for carbon species bulk diffusion through catalyst nanoparticle, $D_{S0}=a_0^2\nu$, a_0(= 0.34 nm) is the carbon atoms inter-atomic distance, ε_S(=0.3 eV) is the energy barrier for carbon species diffusion through the catalyst nanoparticle surface [9–11], $A_k=a_0\nu$, $\delta\varepsilon_{inc}$(= 0.4 eV) is the carbon species energy barrier to diffuse along the nanofiber-catalyst interface [44], P (≈ 20 GPa) is the pressure exerted by graphene layers on catalyst nanoparticle [45], ρ_{cnf} is the nanofiber density, m_{cat} is the mass of the metal catalyst particle, and ρ_{cat} is the density of metal catalyst nanoparticle. $D_m=D_{m0}\exp\left(-\varepsilon_{SD}/k_BT_S\right)$ is the metal atoms diffusion coefficient [46], D_{m0} is the constant, ε_{SD} is the metal atoms self-diffusion activation energy.

Due to the continuous accumulation of positively charged species (first and second term) and electrons (last term) on the growing CNF surface over the negatively biased substrate surface, a charge is developed with growth time on the CNF surface that is represented by the Eq. (9).

$$\frac{dZ}{dt}=\sum_j I_{tip_ij}+\sum_j I_{cur_ij}-\gamma_e\left(I_{tip_e}+I_{cur_e}\right),\tag{9}$$

where Z is the charge number, I_{tip_ij} and I_{cur_ij} are the ions collection current due to jth species at the tip as well as the curved surface of the CNF, respectively [30]. I_{tip_e} and I_{cur_e} are the electron collection current at the tip as well as the curved surface of the CNF, respectively [30], and γ_e is the sticking coefficient of electrons.

The plasma species (electrons, ions, and neutrals) from the plasma sheath, accumulate and deposit on the catalyst surface and growing CNF surfaces, however, some of the species desorbs from the surface and move in the plasma again. Thus, it is also important to evaluate how the number densities of plasma species decay in plasma with growth time.

$$\frac{dn_e}{dt}=\sum_j\beta_j n_j-\sum_j\alpha_j n_e n_{ij}-\gamma_e\frac{n_{cnf}}{\lambda_d}\left(I_{\mathrm{tip_e}}+I_{\mathrm{cur_e}}\right),\tag{10}$$

$$\frac{dn_{ij}}{dt}=\beta_j n_j-\alpha_j n_e n_{ij}-\frac{n_{cnf}}{\lambda_d}\left(I_{\mathrm{tip_}ij}+I_{cur_ij}\right)+\sum k_\ell n_j n_{ij'}-\sum k_Q n_{j'} n_{ij},\tag{11}$$

$$\frac{dn_j}{dt}=\alpha_j n_e n_{ij}-\beta_j n_j+\frac{n_{cnf}}{\lambda_d}\left(1-\gamma_{ij}\right)\left(I_{tip_ij}+I_{cur_ij}\right)-\frac{n_{cnf}}{\lambda_d}\gamma_j\left(I_{tip_j}+I_{cur_j}\right)+IF_j-OF_j-J_{ai_j}+J_{desp_j}+J_{thd}-\sum k_Q n_j n_{ij'}+\sum k_\ell n_{j'} n_{ij},\tag{12}$$

where n_e, n_{ij}, n_j are the number density of electrons, positively charged ions, and neutrals of jth species, respectively. β_j is the ionization coefficient of neutral atoms [47], α_j is the

recombination coefficient of electrons and positively charged ions [48], n_{cnf} is the surface density of CNFs which are equal to the surface concentration of the catalyst nanoparticles as we have considered that each CNF nucleates over the single catalyst nanoparticle. $\lambda_d\left(=3\times10^{-3}\text{cm}\right)$ is the plasma sheath width, γ_{ij} is the sticking coefficient of positively charged ions, $J_{\text{a_j}}$ is the adsorption flux of neutrals on the catalyst surface, $J_{\text{desp_j}}$ is the desorption flux of neutrals from the catalyst substrate surface, J_{thd} is the hydrogen flux from the catalyst surface due to thermal dehydrogenation, $IF_j\left(=4.4\times10^{17}\times f_j\times V^{-1}\right)$ and $OF_j\left(=R_p\times n_j\times V^{-1}\right)$ are the jth neutrals species inflow rate and outflow rate into/from the plasma chamber, respectively. f_j is the inlet flow of corresponding gas, $V\left(\approx2\times10^4\ \text{cm}^3\right)$ is the volume of the chamber, and R_p is the pumping rate [35], k_j is the reaction rate coefficient of the reactions that are accounted in the present model.

The Eq. (10) represents the decay rate of electron number density in the bulk plasma due to the gain in electron number density per unit time due to ionization of neutral species (first term), decay rate due to electron–ion recombination and electron collection current at the CNFs surface (second and last term). The Eq. (11) represent the decay rate of positively charged species in the bulk plasma due to ionization of neutral species (first term), electron–ion recombination (second term), ion collection current at the CNF surface (third term), and reactions between different plasma species at different reaction rates (fourth and fifth terms). The Eq. (12) represent the decay rate of neutral species in the plasma on account of various processes such as; electron–ion recombination (first term), decrease in number density due to ionization of neutral species (second term), increase in number density due to neutralization of ions collected at the CNF surface (third term), decrease in number density due to the accretion of neutrals at the CNF surface (fourth term), inflow and outflow rate of neutral species from plasma chamber (fifth and sixth term), adsorption of neutrals to the catalyst surface (seventh term) and desorption of neutrals from the catalyst surface (eighth term). The ninth and eight terms represent the increase and decrease in neutral atoms number density due to the various reactions considered in the present model [19, 39]. The last term of Eq. (12) is corresponds to gain in hydrogen ion number density in the plasma due to thermal dehydrogenation of hydrocarbons on the catalyst surface [10].

The all above equations have been solved simultaneously for appropriate experimentally determined boundary conditions, i.e., operating parameters listed in Table 3 and initial number densities of various plasma species accounted in the present paper given in Table 4, obtained from Refs. [8, 19].

Results and Discussion

The effect of input process parameters such as gas pressure and input plasma power on the plasma characteristics have been investigated in the present study using the computational approach as well as analytical approach and their repercussions on the growth characteristics of the CNFs have been studied. The computational results of electron density and temperature at different gas pressures are shown in Figs. 2 and 3, respectively. At low pressure, i.e., about 50 mTorr, the electron density in the plasma and near the substrate was found to be $3.94\times10^{17}\ \text{m}^{-3}$ and $1.87\times10^{17}\ \text{m}^{-3}$, respectively. The electron temperature in the plasma and near the substrate was found to be 3.6 eV and 2.74 eV, respectively. When gas pressure is increased up to 5 Torr, the electron density in the plasma and near the substrate is found to be reduced to $2.67\times10^{16}\ \text{m}^{-3}$ and $0.04\times10^{16}\ \text{m}^{-3}$, respectively, and the electron temperature in the plasma and near the substrate is found to be reduced to 2.64 eV

Table 3 Initial conditions fed into the analytical model

Parameter	Description	Initial value
D	Diameter of the catalyst nanoparticle	40 nm
T_e	Electron temperature	3 eV
T_i	Ion temperature	0.17 eV
T_n	Neutral atom temperature	0.17 eV
ρ_{ct}	Density of nickel catalyst	8.96 g cm^{-3}
$\alpha_{10} \approx \alpha_{20}$	Electron–ion recombination coefficient	1.12×10^{-7} cm^3 s^{-1}
T_s	Substrate temperature	550 °C
γ_e	Electron sticking coefficient	0.99
γ_{ij}	Ion sticking coefficient	0.99
γ_j	Neutral atom sticking coefficient	0.99
θ_t	Total surface coverage	0.01
$f_{C_2H_2}$	C_2H_2 gas flow rate	50 sccm
f_{H_2}	H_2 gas flow rate	200 sccm
U_s	Substrate potential	− 300 V

and 2.34 eV, respectively. Moreover, the temporal variation of electron density and electron temperature obtained from the analytical model incorporated in the Sect. 2 are shown in Fig. 4a, b, respectively. It can be seen from Fig. 4 that electron density and electron temperature decreases with time and with gas pressure. The analytical plots are different from the computational plots. This is because computational plots show the variation of density and temperature of plasma species when stable plasma is achieved whereas analytical plots show the decay of density and temperature of species during the CNF growth. But the two types of plots (computational and analytical) show the same trend with the pressure variation in the plasma. With the increase in gas pressure, the mean free path of electrons decreases which in turn increase the collision between electrons and neutral species and make electrons to lose their energy faster and leads to decrease in the electron density in the plasma. The reduction in electron density decreases the electric field in the plasma, thereby decreasing the electron temperature. Moreover, due to the continuous collision between electrons and other species, the energy and number of electrons decay with space and time. Harilal et al. [20] experimentally found the decrease in electron density and temperature with time in the laser produced carbon plasma. Many studies also report the reduction in electron density and temperature when the gas pressure in the plasma is raised [17, 19, 21].

The number density distributions of the various neutral species in the bulk plasma region and substrate surface obtained from the computational model at different gas pressures are shown in Fig. 5. The number density of neutral species drops as one go from bulk plasma region to the substrate surface due to the ionization of neutral species, dissociation of the neutral species and various reactions between ion and neutral species. From Fig. 5, one can see that the number density of carbon-bearing species near the substrate surface is related to the gas pressure [22] keeping gas flow rate constant; when gas pressure is raised, the increment in the number densities of neutral species in the plasma is observed from the computational model and decreased decay rate of neutral species during the CNF growth is observed from the analytical model. This can be ascribed to the fact that with an increase in pressure, plasma becomes more collisional due to which species lose their

Table 4 Initial number density of various plasma species considered in the present model have been obtained from Refs. [8, 19]

Species considered	Initial values (cm^{-3})
n_e (electron number density)	4.18×10^{11}
n_{ij} (ion number density)$\times10^{11}$ cm^{-3}	
C_2H^+	0.002
$C_2H_2^+$	0.95
$C_2H_3^+$	0.084
C_4H^+	0.91×10^{-3}
$C_4H_2^+$	0.66
$C_4H_3^+$	0.54
C_6H^+	0.02×10^{-3}
$C_6H_2^+$	0.008
$C_6H_4^+$	0.097
C_8H^+	0.005×10^{-4}
$C_8H_2^+$	1.55×10^{-3}
$C_8H_4^+$	0.61×10^{-3}
$C_{10}H^+$	0.008
$C_{10}H_2^+$	0.56×10^{-3}
H_2^+	1.45
H^+	0.38
n_j (neutral species number density)	
C_2H	2.25×10^{11}
C_2H_2	5.15×10^{14}
C_2H_3	6.06×10^{13}
C_4H	9.12×10^{9}
C_4H_2	9.22×10^{12}
C_4H_3	1.85×10^{12}
C_6H	4.55×10^{9}
C_6H_2	3.42×10^{12}
C_6H_4	2.48×10^{13}
C_8H	1.11×10^{9}
C_8H_2	8.33×10^{11}
C_8H_4	4.24×10^{10}
$C_{10}H$	7.01×10^{8}
$C_{10}H_2$	6.14×10^{11}
H_2	1.48×10^{15}
H	2.54×10^{14}

energy faster and, therefore, temperature and energy of the neutral species decrease. The rates of reactions between ions and neutrals also depend on the temperature of the respective species. A large number of collisions at higher pressure decreases the neutral temperature, and thereby, decreasing the reaction rates. The reduction in electron number density with increase in pressure also leads to the lower ionization of neutral species. Thus, the concentration of neutral species increases in the plasma as gas pressure is raised.

In Fig. 6a, b, the temporal variation of CNF height and diameter, respectively, has been carried out at different gas pressures in the plasma. From Fig. 6c, d, one can see that the

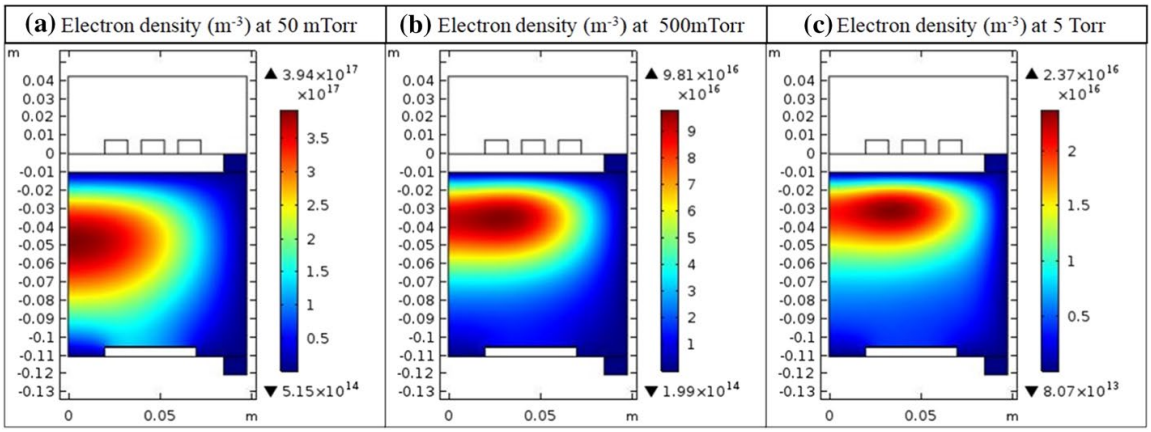


Fig. 2 Electron density spatial distribution in the plasma obtained from the computational model for different gas pressures, i.e., **a** 50 mTorr, **b** 500 mTorr, and **c** 5 Torr

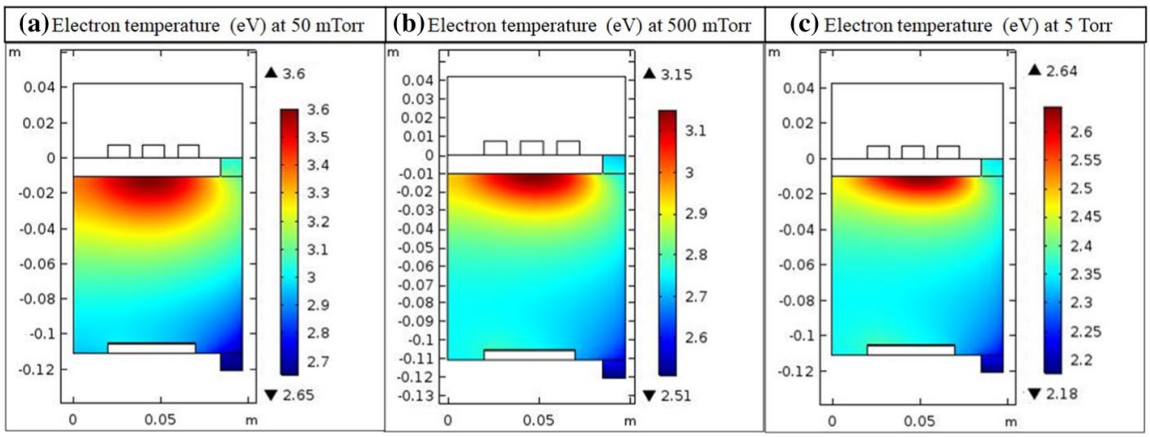


Fig. 3 Electron temperature spatial distribution in the plasma obtained from the computational model for different gas pressures, i.e., **a** 50 mTorr, **b** 500 mTorr, and **c** 5 Torr

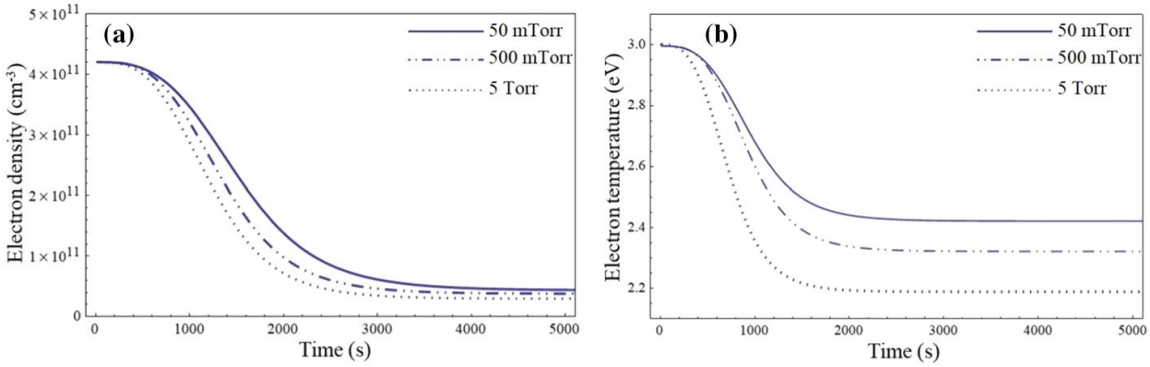


Fig. 4 Variation of **a** electron density and **b** electron temperature with growth time in the plasma for different gas pressures, i.e., 50 mTorr, 500 mTorr, and 5 Torr keeping other parameters constant that are given in the text

decomposition of various carbon-bearing species on the catalyst surface depends on the gas pressure. Moreover, the surface concentration of the carbon species over the catalyst surface changes with gas pressure. With increase in gas pressure, the neutral density in the plasma increases (see Fig. 5) due to which flux of hydrocarbon neutrals and hydrogen

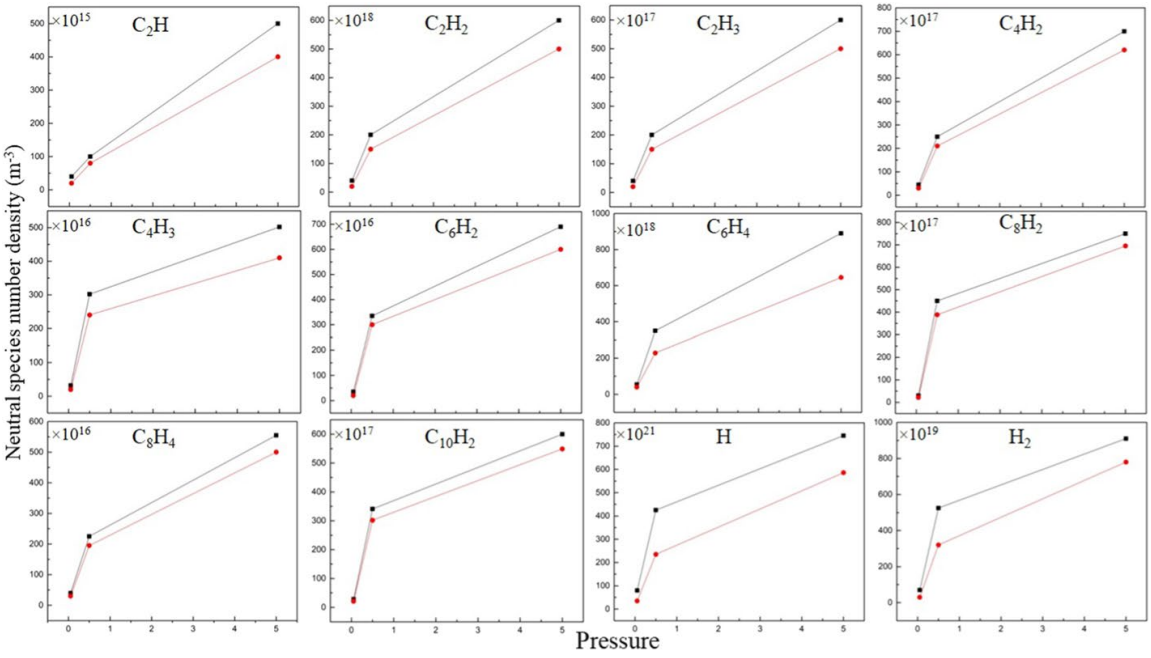


Fig. 5 Density distribution of the neutral species in the bulk plasma region (black dots) and near the substrate surface (red dots) obtained from the computational model for different gas pressures i.e., 50 mTorr, 500 mTorr, and 5 Torr

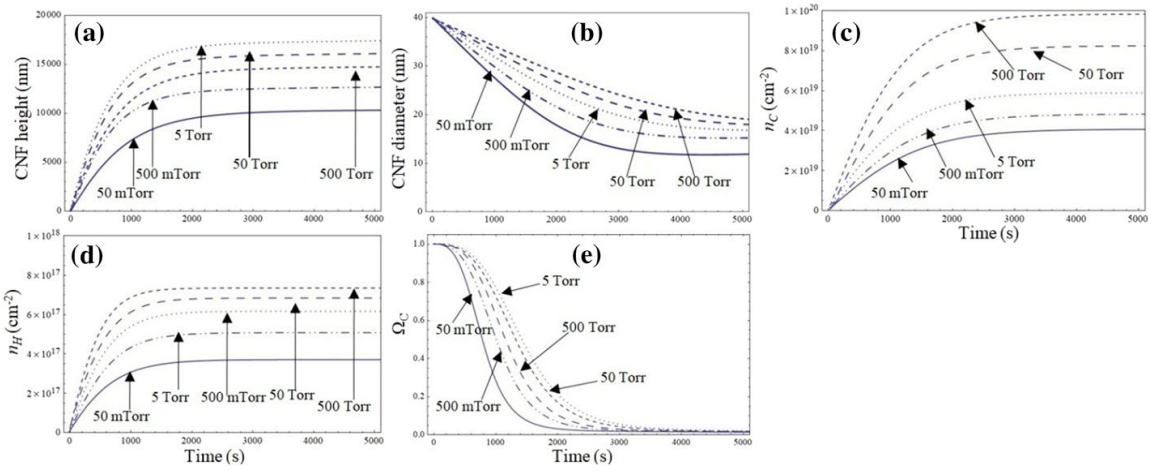


Fig. 6 Temporal variation of the **a** CNF height, **b** CNF diameter, **c** surface concentration of carbon species on the catalyst surface n_C, **d** surface concentration of hydrogen radicals on the catalyst surface n_H, and **e** poisoning rate of the catalyst for different gas pressures i.e., 50 mTorr, 500 mTorr, 5 Torr, 50 Torr, and 500 Torr keeping other parameter constant that are given in the text

neutrals on the catalyst nanoparticle surface increases. Which leads to more effective generation of carbon species and hydrogen radicals on the catalyst active surface (see Fig. 6c, d) and thus, enhance the CNF growth (height per unit deposition time) whereas, at sufficiently high pressure (around and above 50 Torr), CNF growth begins to drop significantly. This is because, at much higher pressure plasma is not sufficient to dissociate hydrocarbons neutrals but able to dissociate hydrogen molecules since they require less dissociation energy as compared to the hydrocarbons neutrals. Due to which hydrocarbons neutral flux and number of carbon species on the catalyst surface increases rapidly leading to the early poisoning of the catalyst. To make it clear, the poisoning rate (variation of blocking

function with the growth time) of the catalyst nanoparticle at different gas pressure is plotted in Fig. 6e which shows that catalyst nanoparticle maintains its activity for quite long time when gas pressure is increased up to 5 Torr. This is because of the removal of amorphous carbon layer due to the presence of sufficient number of hydrogen radicals. However, at high pressure, the rapid growth in surface density of carbon species is obtained (see Fig. 6c) due to which interaction between carbon species and hydrogen radicals becomes less effective. This leads to the rapid growth of the amorphous carbon layer over the catalyst active surface which in turn leads to the quick poisoning of the catalyst activity. The Fig. 6b shows the temporal variation of the CNF diameter at different gas pressures. The initial diameter (at time $t=0$) of the CNF is assumed to be same as the diameter of the catalyst nanoparticle and reduction of CNF diameter with growth time is observed due to the compressive stress exerted by the growing graphitic shells and also due to the etching of CNF side walls by energetic hydrogen species (radicals and ions). Moreover, it can also be seen that CNFs having large diameters are grown when gas pressure in the plasma is increased which can be explained by two ways; (1) the rate of diffusion and precipitation of carbon species over the growing walls of the CNFs increases due to the availability of large number of carbon species, (2) at higher pressure, ions lose their energy due to large of collisions and thus, fewer energetic ions reach near the substrate surface and, consequently, etching is reduced [22]. The results obtained in Fig. 6a, b are in good agreement with the available experimental observations discussed as follows; Wei et al. [7] experimentally investigated the effect of pressure in the plasma on the CNF length and observed that the length of the CNF increases with increase in pressure (0–30 Torr) and decrease when pressure is increased further (cf. Fig. 6a, Ref. [7]). Wang et al. [16] investigated the growth of CNTs via PECVD technique at two different gas pressures i.e., 15 Torr, and 30 Torr and found that the average length of the CNF grown at 30 Torr is smaller than that at pressure 15 Torr. Li et al. [23] studied the effect of gas pressure on the structure and growth of the CNTs via chemical vapor deposition (CVD) technique and observed that CNTs grow with shorter in length and larger in diameter at high pressure (50–760 Torr). Ganjipour et al. [14] observed the increase in CNTs diameter when gas pressure in the PECVD system is increased (0–500 Torr). Pint et al. [24] investigated the effect of pressure on the CNT growth via CVD process and found the increase in CNT height with increase in pressure (0–15 Torr). Tsakadze et al. [25] studied the effect of the pressure on the field emission properties of the CNTs grown via thermally CVD system and observed the length of the CNTs is decreased at high pressure range (300–700 Torr). Tanemura et al. [26] have tried to optimize the PECVD process parameters to enhance the growth of CNTs and found that diameter of the CNTs increases when pressure of C_2H_2 gas is raised in the growth environment. Chhowalla et al. [12] investigated the PECVD growth process conditions of the vertically aligned CNTs and observed the increase in the length of the CNTs with increase in the pressure (0–10 Torr). Hinkov et al. [27] found the increase in the growth rate of plasma grown single-walled CNTs with increase in the pressure (0–1000 mbar). Cruden et al. [18] reported the increase in the CNFs length with increase in pressure (2 Torr, 3 Torr, and 8 Torr) in the DC PECVD growth process of CNFs. From the results obtained in the present study and from the above specified experimental literature, one can conclude that CNFs grow at slower rate when pressure is very low and very high. However, the optimum pressure range for the enhanced CNFs growth is about the one order of the magnitude in the units of Torr.

Figures 7, 8 and 9 show the computational results of the electron number density, electron temperature, number densities of positively charged species in the bulk plasma region and near the substrate surface at different input plasma powers, respectively.

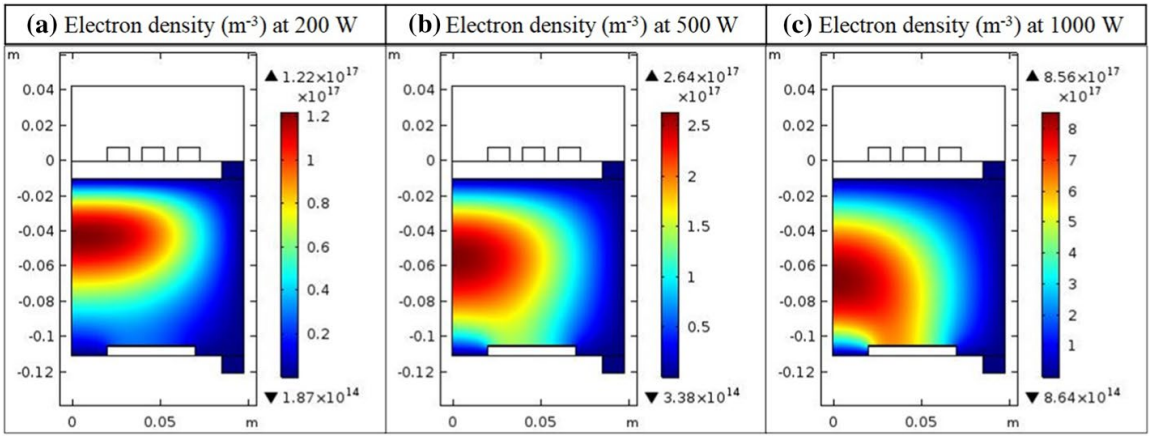


Fig. 7 Electron density distribution in the plasma obtained from the computational model for different input plasma powers, i.e., **a** 200 W, **b** 500 W, and **c** 1000 W

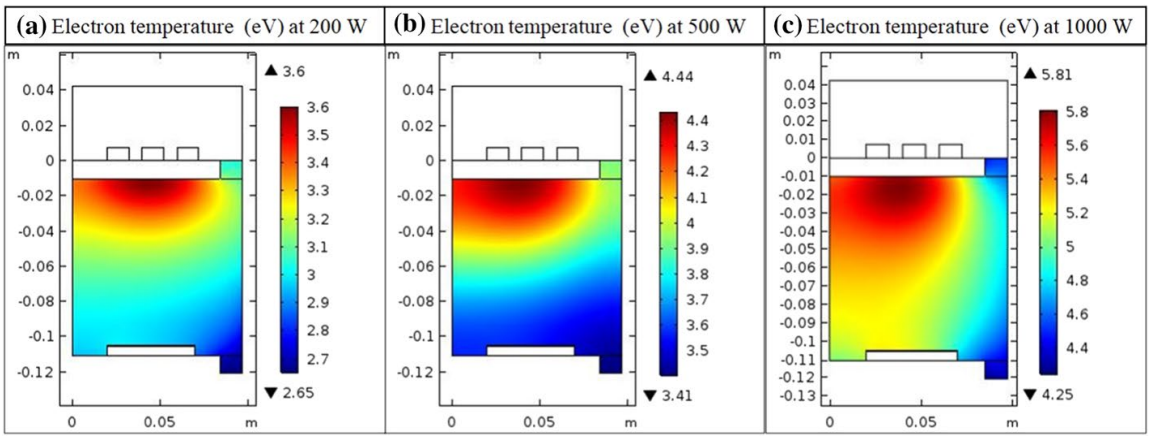


Fig. 8 Electron temperature spatial distribution in the plasma obtained from the computational model for different plasma powers, i.e., **a** 200 W, **b** 500 W, and **c** 1000 W

From the computational plots obtained at 200 W input plasma power, the electron density and temperature in the bulk plasma region is found to be 1.22×10^{17} m^{-3} and 3.6 eV, respectively, which reduced to 0.42×10^{17} m^{-3} and 3.07 eV, respectively, near the substrate surface. However, at 1000 W, the electron density and temperature are found to be 8.56×10^{17} m^{-3} and 5.81 eV, respectively, in the bulk plasma region and 5.89×10^{17} m^{-3} and 5.24 eV, respectively near the substrate surface. Moreover, from the spatial distribution of the densities of various species, we have obtained the reduction in the ion densities from bulk plasma region to substrate surface due to the collisions with the neutral species in the plasma as discussed before. The Fig. 10a, b shows the analytical results of the electron number density and electron temperature with the growth time in the plasma at different input plasma powers, respectively. When input plasma power is increased, the volume of the plasma region increases which goes into the increase in electric field in the plasma, and consequently, electron number density and electron temperature in the plasma increases. Collision et al. [21] experimentally observed the increase in the electron density and temperature when plasma power in the plasma system is increased. The increase in the electron density with input plasma power also leads to more ionization of the neutral species, thereby, increasing the ion densities in the plasma [see Fig. 9]. Therefore, ion densities decay at the slower rate in the plasma with the increase in input plasma power. Thus,

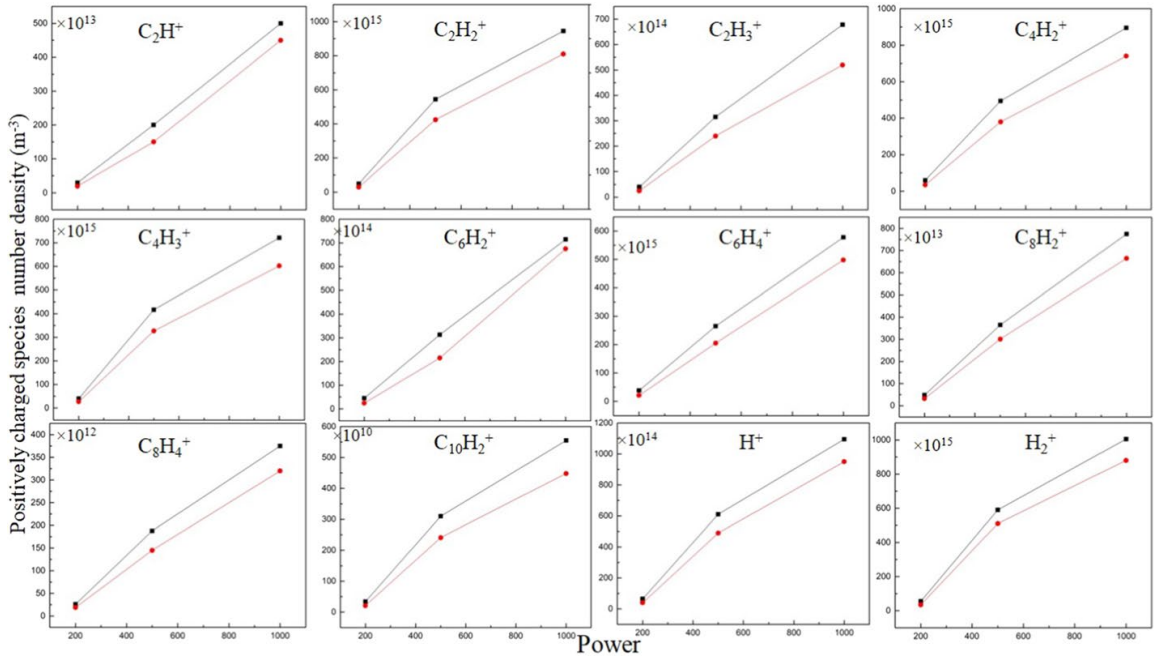


Fig. 9 Density distribution of the positively charged species (ions) in the bulk plasma region (black dots) and near the substrate surface (red dots) obtained from the computational model for different input plasma powers, i.e., 200 W, 500 W, and 1000 W

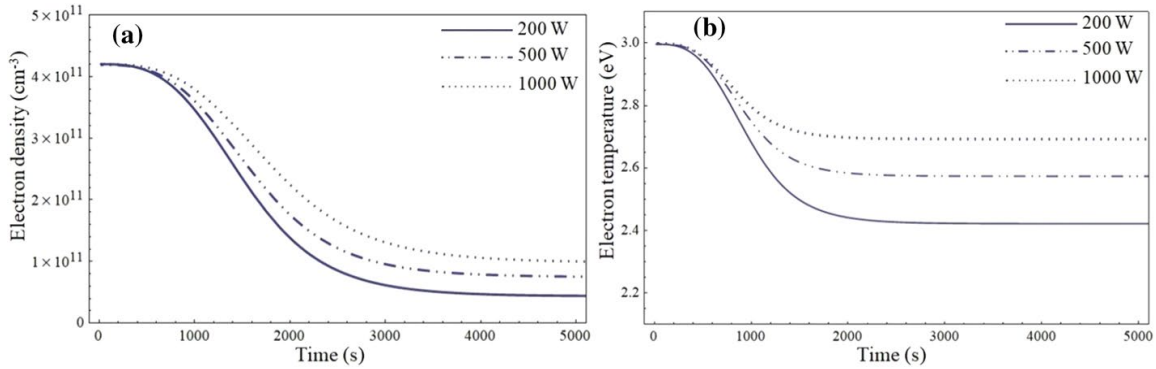


Fig. 10 Variation of **a** electron density and **b** electron temperature with growth time in the plasma for different gas pressures i.e., 50 mTorr, 500 mTorr, and 5 Torr keeping other parameters constant that are given in the text

ion-induced decomposition process and ion dissociation process on the catalyst nanoparticle surface becomes dominant due to which large number of carbon species and hydrogen radicals are produced on the active catalyst surface as shown in Fig. 11a, b, respectively. To understand the effect of power on the distribution of plasma species and their repercussion on the growth of CNFs, blocking function versus growth time curve is plotted for different input plasma powers in Fig. 11c. As input plasma power is increased, poisoning rate of the catalyst becomes slower which is somewhat surprising because one would think of faster poisoning of the catalyst due to the availability of a large number of carbon species on the catalyst surface. However, reverse trend is followed in reality, this is because a large number of hydrogen radicals are also produced at the same time on the catalyst surface and interacts with amorphous carbons and preserve the activity of the catalyst for quite longer time, thereby, increasing the growth of the CNF (see Fig. 11d). Moreover, at high input plasma power, the highly energetic hydrogen ions and radicals also etch walls of

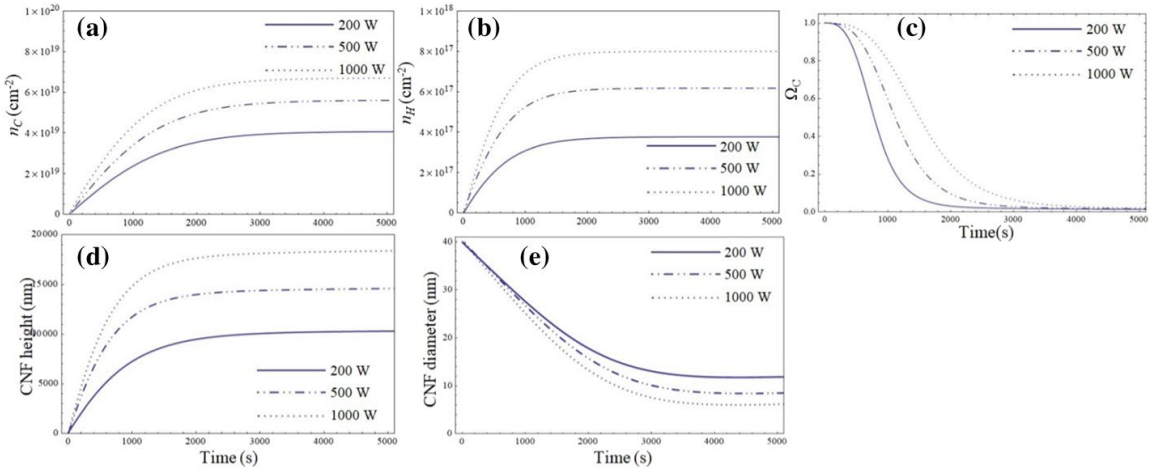


Fig. 11 Temporal variation of the **a** surface concentration of carbon species on the catalyst surface n_C, **b** surface concentration of hydrogen radicals on the catalyst surface n_H, **c** poisoning rate of the catalyst, **d** CNF height, and **e** CNF diameter for different input plasma powers i.e., 200 W, 500 W, and 1000 W

the CNF which leads to decrease in CNF diameter (see Fig. 11e). The results obtained in the Fig. 11d, e are in line with the available experimental observations as follows; Chang et al. [28] synthesized the CNTs at low temperature using microwave plasma and found the considerable reduction in the diameter of the CNT when input plasma power is raised (700–1000 W). Merkulov et al. [29] found the reduction in the diameter of the CNFs when plasma power of the DC PECVD system is increased. Wei et al. [7] also observed the increase in the CNF length with an increase in the plasma power (0–2000 W).

Conclusion

In the present paper, the numerical simulations have been carried out using COMSOL Multiphysics 5.2 simulational software to investigate the properties of $Ar + H_2 + C_2H_2$ reactive plasma as the function of process parameters. Moreover, the impact of plasma parameters on the growth dynamics of the CNFs have been studied using analytical model. From the numerical data obtained, we found that temperature and number density distribution of the electrons and other species (neutrals and positively charged) in the plasma is non-uniform i.e., densities of each species decreases from bulk plasma region to substrate surface. From the analytical model it is found that the temperature and density of electrons and other species decays in the plasma due to collisions and dissociation/ionization reactions among various species. Moreover, we have also found that the distribution of electron density and temperature in the plasma decreases when gas pressure is raised, and increases when input plasma power is raised and the densities of neutral and charged species are related to the concentration and energy of the electrons in the plasma.

From the surface deposition model, we have obtained that the dissociation rate of the carbon bearing species on the catalyst active surface to generate carbon–hydrogen species are related to the gas pressure and input plasma power; when gas pressure is raised (50 mTorr to 5 Torr), the growth rate of CNF increases. However, at very high pressure, i.e., above and around 50 Torr, the growth of the CNF decreases due to the quick poisoning of the catalyst whereas, the diameter of the CNF increases within the entire pressure range. Thus, we can conclude that the optimum pressure for the enhanced growth of the CNFs is

about one order of the magnitude in the units of Torr. Moreover, when input plasma power is raised (200–1000 W), the growth rate of the CNF increases, however, the CNF diameter decreases due to more effective etching of the energetic ions.

The present study can be used for the better optimization of the PECVD process parameters for the plasma assisted the catalyzed growth of the CNFs to fabricate the highly desirable field emitters based on the CNFs/CNTs. More importantly, the conclusions made in the present study are not only restricted to the growth of one-dimensional carbon nanostructures and may extend to the plasma-assisted growth of other carbon nanostructures at low pressure and low temperature.

References

1. Choi WB, Jin YW, Kim HY, Lee SJ, Yun MJ, Kang JK, Coi YS, Park NS, Lee NS, Kim JM (2001) Appl Phys Lett 78:1547–1549
2. Niu C, Sichel EK, Hoch R, Moy D, Tennent H (1997) Appl Phys Lett 70:1480–1482
3. Saito R, Dresselhaus G, Dresslhaus MS (1998) Physical properties of carbon nanotubes. Imperial College Press, London
4. Darkrim FL, Malbrunot P, Tartaglia GP (2002) Int J Hydrog Energy 27:193–202
5. Guillorn MA, Melechko AV, Merkulov VI, Ellis ED, Simpson ML (2001) J Vac Sci Technol B 19:2598–2601
6. Merkulov VI, Melechko AV, Guillorn MA, Simpson ML, Lowndes DH, Whealton JH, Raridon RJ (2002) Appl Phys Lett 80:4816–4818
7. Wei HW, Leou LC, Wei MT, Lin YY, Tsai CH (2005) J Appl Phys 98:044313
8. Mehdipour H, Ostrikov K, Rider AE (2010) Nanotechnology 21:455605
9. Denysenko I, Azarenkov NA (2011) J Phys D Appl Phys 44:174031
10. Denysenko I, Ostrikov K (2009) J Phys D Appl Phys 42:015208
11. Denysenko I, Ostrikov K, Yu MY, Azarenkov NA (2007) J Appl Phys 102:074308
12. Chhowalla M, Teo KBK, Ducati C, Rupessinghe NL, Amaratunga GAJ, Ferrari AC, Roy D, Robertson J, Milne WI (2001) J Appl Phys 90:5308–5317
13. Catherine Y, Couderc P (1986) Thin Solid Films 144:265–280
14. Ganjipour B, Mohrajerzadeh S, Hesamzadeh H, Khodadadi A (2005) Fuller Nanotubes Carbon Nanostrcut 13:365–373
15. Delzeit L, McAninch I, Cruden BA, Hash D, Chen B, Han J, Meyyappan M (2002) J Appl Phys 91:6027–6033
16. Wang BB, Lee S, Xu XZ, Choi S, Yan H, Zhang B, Hao W (2004) Appl Surf Sci 236:6–12
17. Shivkumar G, Tholeti SS, Alrefae MA, Fisher TS, Alexeenko AA (2016) J Appl Phys 119:113301
18. Cruden BA, Cassell AM, Hash DB, Meyyappan M (2004) J Appl Phys 96:5284–5292
19. Mao M, Bogaerts A (2010) J Phys D Appl Phys 43:205201
20. Harilal SS, Bindhu CV, Issac RC, Nampoori VPN, Vallabhan CPG (1997) J Appl Phys 82:2140–2146
21. Collison WZ, Ni TQ, Barnes MS (1998) J Vac Sci Technol A 16:100–107
22. Melechko AV, Merkulov VI, McKnight TE, Guillorn MA, Klein KL, Lowndes DH, Simpson ML (2005) J Appl Phys 97:041301
23. Li WZ, Wen JG, Tu Y, Ren ZF (2001) Appl Phys A 73:259–264
24. Pint CL, Nicholas N, Pheasant ST, Duque JG, Vasquez ANGP, Eres G, Pasquali M, Hauge RH (2008) J Phys Chem C 112:14041–14051
25. Tsakadze ZL, Ostrikov K, Sow CH, Mhaisalkar SG, Boey YC (2010) J Nanosci Nanotechnol 10:6575–6579
26. Tanemura M, Iwata K, Takahashi K, Fujimoto Y, Okuyama Y, Sugie H, Filip V (2001) J Appl Phys 90:1529–1533
27. Hinkov I, Farhat S, Scott CD (2005) Carbon 43:2453–2462
28. Chang SC, Lin TC, Pai CY (2007) Microelectron J 38:657–662
29. Merkulov VI, Hensely DK, Melechko AV, Guillorn MA, Lowndes DH, Simpson ML (2002) J Phys Chem B 106:10570–10577
30. Gupta R, Sharma SC, Sharma R (2017) Plasma Sources Sci Technol 26:024006
31. Gupta R, Gupta N, Sharma SC (2018) Phys Plasmas 25:043504
32. Gupta R, Sharma SC (2018) Contrib Plasma Phys. https://doi.org/10.1002/ctpp.201700138

33. Gupta R, Sharma SC (2017) Phys Plasma 24:073504
34. Dagel DJ, Mallouris CM, Doyle JR (1996) J Appl Phys 79:8735–8747
35. Denysenko IB, Xu S, Long JD, Rutkevych PP, Azarenkov NA, Ostrikov K (2004) J Appl Phys 95:2713–2724
36. Deschenaux C, Affolter A, Magni D, Hollenstein C, Fayet P (1999) J Phys D Appl Phys 32:1876
37. Benedikt J (2010) J Phys D Appl Phys 43:043001
38. Doyle JR (1997) J Appl Phys 82:4763–4771
39. Ostrikov K, Yoona HJ, Rider AE, Vladimirov SV (2007) Plasma Process Polym 4:27–40
40. Mehdipour H, Ostrikov K, Rider AE, Han Z (2011) Plasma Process Polym 8:386–400
41. Marvi Z, Xu S, Foroutan G, Ostrikov K (2015) Phys Plasmas 22:013504
42. Cantoro M, Hofmann S, Pisana S, Ducati C, Parvez A, Ferrari AC, Robertson J (2006) Diam Relat Mater 15:1029–1035
43. Hofmann S, Cantoro M, Kleinsorge B, Casiraghi C, Parvez A, Robertson J, Ducati C (2005) J Appl Phys 98:034308
44. Louchev OA, Laude T, Sato Y, Kanda H (2003) J Chem Phys 118:7622–7634
45. Yudasaka M, Kikuchi R, Matsui T, Ohki Y, Yoshimura S, Ota E (1995) Appl Phys Lett 67:2477–2479
46. See (http://matter.org.uk/matscicdrom/manual/df.html) for metal atom diffusion coefficient
47. Sodha MS, Misra S, Mishra SK (2010) Phys Plasmas 17:113705
48. Sodha MS, Misra S, Misra SK, Srivastava S (2010) J Appl Phys 107:103307

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

Laser Physics Letters

Astro Ltd.

LETTER

Plasma bubble evolution in laser wakefield acceleration in a petawatt regime

To cite this article: Monika Yadav *et al* 2020 *Laser Phys. Lett.* **17** 076001

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.

This content was downloaded from IP address 14.139.251.100 on 08/06/2020 at 10:27

IOP Publishing | Astro Ltd Laser Physics Letters
Laser Phys. Lett. **17** (2020) 076001 (8pp) https://doi.org/10.1088/1612-202X/ab8fa9

Letter

Plasma bubble evolution in laser wakefield acceleration in a petawatt regime

Monika Yadav[1], Devki Nandan Gupta[2], Suresh C Sharma[1] and Hyyong Suk[3]

[1] Department of Applied Physics, Delhi Technological University, Delhi 110 042, India
[2] Department of Physics and Astrophysics, University of Delhi, Delhi 110 007, India
[3] Department of Physics and Photon Science and APRI, Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea

E-mail: dngupta@physics.du.ac.in

Received 14 October 2019
Accepted for publication 25 April 2020
Published 3 June 2020



Abstract

The bubble regime of the laser wakefield acceleration is one of the most recent and promising mechanisms for generating quasi-monoenergetic electron beams. In this work, we propose to study the dynamics of bubble (a wake structure devoid of electrons created in underdense plasma by a highly-intense ultrashort laser pulse) in a petawatt regime. The dependence of the electron beam energy and the quality of the electron beam on the shape of the bubble is the main motivation behind this work. The bubble length as well as bubble shape have been investigated using two-dimensional particle-in-cell simulations. The evolution of the bubble with time, and the correlation of bubble length (longitudinal and transverse radius) with the intensity of laser pulse have been revealed in this study. The change of bubble dimensions can be estimated by various determining factors such as the laser pulse focusing, the beam loading, the residual electrons, and the bubble velocity. It has also been confirmed that the shape of the bubble cannot be predicted using fixed shape models as spherical or elliptical. Simulations unveil that as the bubble traverses in plasma, it evolves from spherical shape to the highly elliptical shape. And, as it approaches the dephasing length, the eccentricity decreases further. Consequently, the self-injection of plasma electrons in the bubble is seriously affected by the bubble evolution. Comparison of the electron energy gain at different intensities of laser pulse has also been provided. Various scaling laws for electron beam energy estimations are predicted in this investigation. High quality electron beam can be obtained by controlling the bubble evolution, which may have significant applications in future coherent light sources, biomedical, condensed matter physics, and x-ray generation by table-top FEL.

Keywords: petawatt laser, wakefield acceleration, plasma simulation

(Some figures may appear in colour only in the online journal)

1. Introduction

Laser wakefield ccceleration (LWFA) has been one of the most promising acceleration mechanisms at present since it was first proposed in 1979 by Tajima and Dawson [1]. It is due to their ability to generate quasi-monoenergetic electron bunches with up to GeV energies in millimetre/centimetre length plasmas [2–5]. The development in this field has been made possible by the progress in ultra-intense laser technology. Currently, based on chirped pulse amplification technique, ultra-high intensity lasers with peak powers in hundreds of terrawatt or even pett-awatt (PW) regimes [6, 7] have been demonstrated all over

 © 2020 Astro Ltd Printed in the UK

the world. A PW laser system is able to generate a focused intensity of 10^{20} to 10^{22} W cm^{-2}. A compressed pulse with pulse duration shorter than 20 fs and peak power of approximately 5 PW has been achieved on a multi-petawatt 3-cascaded all-optical parametric chirped-pulse amplification laser facility [8]. Due to these great achievements in laser technology, it has become convenient to carry out research in LWFA in highly non-linear regime (bubble or blow-out regime) [9–12]. The bubble is a wake structure formed when the driver pulse is intense enough to expel out all the electrons radially outward leaving a cavitated region behind the driver, surrounded by a sheath of relativistic high density electrons. It propagates behind the intense and tightly-focused laser pulse with nearly relativistic speed. The optimum conditions for the realization of the bubble is $k_pR \approx 2\sqrt{a_0}$, where $k_p = \omega_p/c$, k_p is the plasma wave vector, ω_p is the plasma frequency, c is the speed of light, R is the focal spot size, and $a_0 \gg 1$ (a_0 is the peak normalized vector potential of the laser) [13].

It has been reported in various previous papers that sometimes electrons from the background plasma can be self-injected at the back of the bubble and get accelerated by the longitudinal space charge field of the bubble. Although, several techniques have been developed to induce the electron injection in the bubble such as colliding pulse injection [14, 15], ionization induced injection [16, 17], and density transitions [18, 19] etc. But it is highly desirable to get the electrons self-injected in the bubble, realizing the experimental challenges related to small spatial and temporal scales of laser wakefield acceleration. The evolution of the driver pulse [20] causes various changes in the bubble geometry and wake potential [21]. Previously, it has been suggested that the bubble formation is not enough for the generation of mono-energetic beam. The collimation of the electron bunch has a dependence on the variation of the geometry of the bubble during its propagation. Therefore, understanding the mechanism of self-injection and its relation with the bubble dynamics is of key importance to control and improve the performance of laser wakefield acceleration in terms of beam energy and beam quality. Kostyukov *et al* [22] first proposed the theory for electromagnetic fields in the bubble and its geometry. The bubble shape was assumed to be spherical in their study. The slower bubble velocities and the impact of residual electrons were ignored by Kostyukov *et al.* Sadighi-Bonabi *et al* [23] presented the ellipsoidal cavity model and found it more consistent to explain the monoenergetic electron trajectory in a relativistic regime. Wu *et al* [24] modified the theory by taking only residual electrons into account. The change in the bubble shape to the ellipsoid was shown in the study. Kalmykov *et al* [25, 26] discussed that the self-injection mechanism is majorly governed by the driver evolution. Hence, it may cause uninterrupted self-injection causing degraded electron beam quality. Benedetti *et al* [27] presented a study of self-injection process, providing a better insight into the relation between self-injection and wake properties like wake velocity and wake amplitude etc. In 2015, Li *et al* [28] presented the solution for the electromagnetic fields in the non-linear bubble wake and bubble shape, taking the ellipsoidal assumption. Slower bubble velocities and residual electrons were taken into account while obtaining the general solution.

In our study, bubble dynamics has been studied in a systematic way. Quantitative information has been provided for the bubble geometry using two-dimensional particle-in-cell (2D-PIC) simulations [29, 30]. It has been shown that self-injection cannot be studied by taking a fixed bubble geometry, either spherical or ellipsoidal as the bubble geometry cannot be considered a constant in time (non-evolving). The bubble continuously changes depending on the wake properties such as wake velocity, wake amplitude and residual electrons etc. As the wake properties vary with time, the bubble also evolves. The electron beam quality is associated with the self-injection of electrons, and hence the bubble dynamics. Plasma parameter (density) has been used as a controlling parameter to produce a high-quality electron beam. The improvement in beam properties resulting from this is likely to be very significant in biomedical imaging, high energy density science, and ultrafast condensed matter [31, 32].

This paper has been divided into four sections. In section 1, the bubble geometry (shape/size) has been discussed. The particle-in-cell simulation method has been used for this study. The longitudinal and transverse bubble radius evolution with time and its dependence on laser pulse evolution have been shown. A qualitative discussion on the influence of laser strength parameter (a_0) on bubble radii and the maximum energy gain has been presented. In section 2, the varying bubble shape estimation has been reported. The self-injection of electrons in the bubble as a function of bubble size and shape has been discussed. The controlling and improving of the electron beam quality by optimizing the plasma density has been reported in the last section.

2. 2D-PIC simulation results

In this section, first, we discuss the geometry (shape/size) of the bubble wake structure and its evolution while it propagates in plasmas. In order to observe the shape and size of the bubble, 2D-PIC simulations have been performed on a relativistic and self-consistent plasma simulation VORPAL code. A Gaussian laser pulse, linearly polarized in the *y*-direction, of pulse duration $\tau = 20$ fs, laser spot size $r_0 = 12$ μm, laser wavelength $\lambda = 1$ μm, and power in petawatt regime is propagated in an underdense plasma. The simulation is run for approximately 1.2 mm propagation length. The laser pulse is launched in the *x*-direction and is channel-guided. The plasma density considered in the simulations is 10^{19} cm^{-3}. The longitudinal and transverse resolution of the simulation are $\lambda/30$ and $\lambda/8$, respectively. A moving window of size 100×65 μm is considered moving with the velocity of light. 3000×292 cells have been used with two particles per cell. And the background of ions is considered immobile in the simulations.

As the process of self-injection of electrons in the bubble wake is closely related to the evolution of the bubble, thus it is crucial to understand the dynamic process of bubble

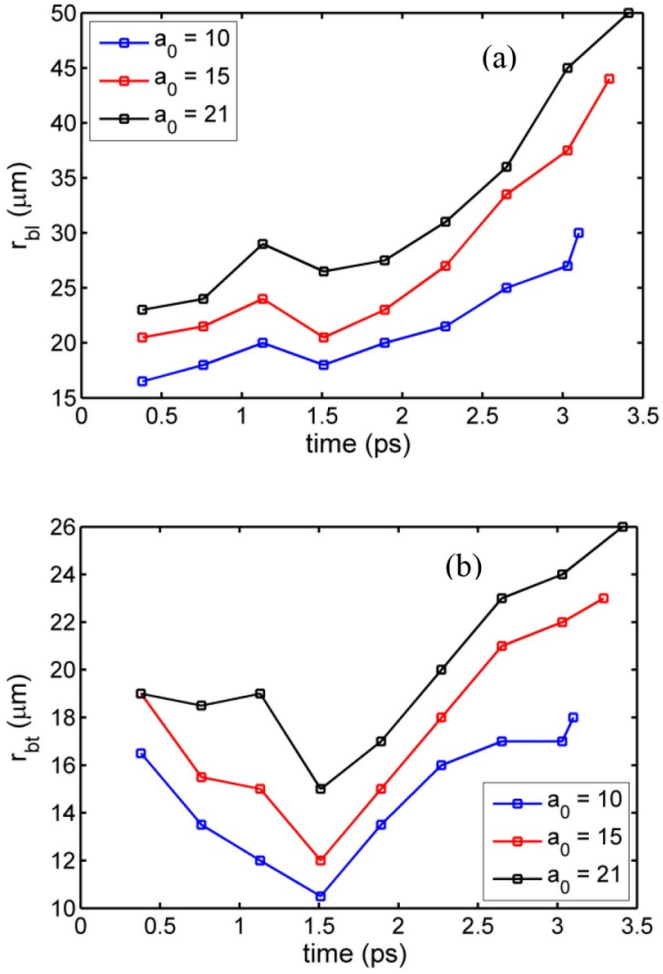


Figure 1. (a) Variation of longitudinal bubble radius (r_{bl} in micron) with time (in picosecond) for different values of normalized laser strength parameter $a_0 = 10, 15, 21$ and (b) variation of transverse bubble radius (r_{bt} in micron) with time (in picosecond) for different values of normalized laser strength parameter a_0 as used for estimation of the longitudinal bubble radius. Other laser-plasma parameters considered in these simulations are as follows: laser pulse duration $\tau = 20$ fs, laser spot size $r_0 = 12$ μm, laser wavelength $\lambda = 1$ μm, and plasma density of 10^{19} cm^{-3}.

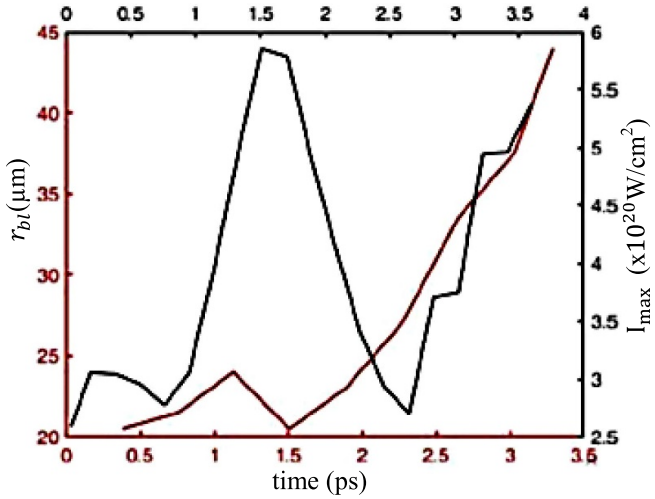


Figure 2. The black curve represents the variation in peak laser intensity (Imax. in W cm^{-2}) and the red curve represents the variation in bubble length (in micron) estimated with the propagation time (in picosecond).

lengthening. Bubble shape is basically characterized by longitudinal (r_{bl}) and transverse radii (r_{bt}). The longitudinal radius can be measured as the length of the accelerating part of the wake. The transverse radius can be calculated from the center of the bubble (where the electron density is almost zero) to the electron sheath (where the electron density equals the background density).

The bubble radii have been estimated from the density plots obtained from the PIC simulations. The bubble radii (longitudinal and transverse) as taken from the simulations at different times during the evolution are shown in figures 1(a) and (b). The bubble evolution has been studied for three different values of a_0 (i.e. $a_0 = 10, 15, 21$ for the corresponding laser intensities $1.37 \times 10^{20}, 3.1 \times 10^{20}, 6.4 \times 10^{20}$ W cm^{-2}, respectively) to confirm the consistency of the results. The simulations show that the trend of the evolution of radius reported for three different laser intensity parameters is identical. The change in bubble wake dimensions is primarily dependent on the evolution of laser pulse in plasmas. Figure 1(a) shows the bubble longitudinal radius with the propagation time during bubble evolution for different laser intensity parameters. Initially, the longitudinal bubble radius increases with time and starts to fall at a time of t = 1.2 ps. The initial lengthening of the bubble is due to the intensity amplification of the pulse in plasmas. Pulse compression and self-steepening of the laser pulse are the responsible factors for laser pulse intensity enhancement. The longitudinal growth of the bubble is followed by the self-compression of the laser pulse. Physically, it can easily be understood as follows: as the plasma electrons are pushed outwards with high velocities to accelerate by the ponderomotive force of the compressed pulse front, a charge separation is created just behind the pulse. And a large electric field is created. As the electrons behind the pulse receive a larger kick due to the compressed pulse front. Consequently, the electrons of the sheath take longer duration to return to the axis. This describes the bubble elongation process. The bubble wake evolution is therefore associated with the laser pulse evolution. The results show that the bubble longitudinal radius gets minima around t = 1.5 ps. As the laser pulse stars to focus beyond this time, the wakefield bubble evolution also starts rapidly to increase the bubble radius. The enhanced intensity of the focused laser pulse assists in bubble evolution. The transverse radius of the bubble shows the similar behaviour as shown in figure 1(b). However, the transverse radius gets minima at larger propagation time. It is obvious due to the bubble evolution corresponding to the wakefield excitation by a Gaussian laser pulse in plasmas. This may be understood using the laser intensity curve in figure 2. Figure 2 shows variation of laser peak intensity and bubble radius with propagation time. It corresponds to the oscillation of the laser longitudinal spot size. The laser intensity falls leading to the shrinkage of the bubble. The laser intensity falls and then gradually rises, leading to

further elongation of bubble longitudinal radius after 1.5 ps. Further increase in the length of the bubble might be due to other factors such as the beam loading. The transverse bubble radius evolution shows a slightly different trend from that of the longitudinal bubble radius evolution. This might be due to the evolution of transverse spot size. The laser evolves differently in the transverse direction leading to a different evolution pattern of transverse bubble radius.

These findings show that the bubble radius is strongly dependent on laser intensity. This may be confirmed from figure 3, where the longitudinal bubble radius is plotted with the laser intensity parameters for different propagation time. The bubble radius increased linearly with the laser intensity parameter (a_0). This effect becomes more profound at larger propagation time due to strong wakefield evolution. This linear dependence of longitudinal radius on laser intensity parameter (a_0) can also be deduced using the theoretical models proposed by Benedetti *et al* [27] and Tossi *et al* [33]. We validate our results using the theory proposed in these references as follows: assuming a spherical ion cavity of radius r_{bl}, the longitudinal wakefield in the plasma wake is given by: $E_z \simeq (k_p\zeta/2)\,E_0$, where ζ is measured from the centre of the bubble. The longitudinal field is maximum for $\zeta = r_{bl}$, where r_{bl} is given by $(2/k_p)\sqrt{a_0}$. The diameter of the bubble is approximately equal to the plasma wavelength in a non-linear regime i.e. $\lambda_{nl} \simeq (2\sqrt{a_0}/\pi)\,\lambda_p$, which concludes that $r_{bl} \simeq \lambda_{nl}/2$ (i.e. the longitudinal bubble radius is proportional to a_0). Therefore, the maximum field at the back of the bubble is given by $E_{zmax.} \simeq r_{bl}/2$. Also, for $a_0 \geq 1$, the maximum electric field of the wake is $E_{zmax.}a_0$. However, our simulation results shows that r_{bl} has a linear dependence on a_0. The results shown in figure 3 are corresponding to this outcome. The bubble radius increases with the laser intensity parameter a_0. This behavior is more severe for large propagation time, which is obvious due to strong wakefield evolution. The results have been compared with some previous studies and are found in agreement with the 2D simulation results obtained by Benedetti *et al* [27] and Toosi *et al* [33]. As the bubble expands, the electrons trapped at the back of the bubble experience a strong field gradient and get accelerated. A maximum energy is attained by the electrons before entering the decelerating phase. As observed previously that the bubble radius is strongly sensitive with the laser strength parameters. Thus the maximum electron energy gain should be laser intensity dependent.

To show these findings, we obtained the maximum electron energy ($\gamma_{max.}$) for different values of normalized laser strength parameter (a_0) in figure 3 in the red color curve. Five sets of simulation have been performed for different laser intensity parameters ($a_0 = 10$, 12.5, 15, 18 and 21). The electrons gain maximum energy about 0.65 GeV, 0.7 GeV, 0.79 GeV, 0.85 GeV, and 0.9 GeV for the corresponding laser intensity parameters of $a_0 = 10$, 12.5, 15, 18 and 21, respectively. The energy gain in LWFA depends on the maximum wakefield amplitude and the accelerating length. As one sees in figure 3 that the longitudinal bubble radius increases with a_0, which indicates the larger accelerating phase for electron acceleration. As dephasing length is proportional to $\sqrt{a_0}$, therefore, the trapped electrons traverse larger distance before getting dephased. A systematic tabulation of simulation parameters for bubble evolution is reported in table 1. Our results clearly show that, by varying the laser intensity parameter, the bubble evolution can be examined for electron energy gain estimation in LWFA. Using these simulation results, one may predict the electron energy gain from LWFA in the petawatt regime, where the bubble evolution is crucial in the determination of wakefield strength.

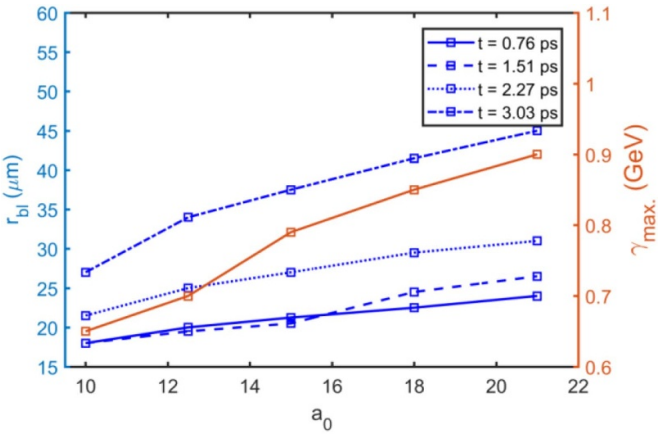


Figure 3. Scaling of the longitudinal bubble radius (r_{bl}) with normalized strength parameter (a_0). Different line styles show the variation of r_{bl} with a_0 for different propagation times (t = 0.76, 1.51, 2.27, and 3.03 ps) and maximum electron energy gain with a_0 is shown in red color. All other laser plasma parameters used in these simulations are same as used in figure 1.

Table 1. A systemic tabulation of simulation results of LWFA for $a_0 = 10$, 15, 21.

a_0	I_0 ($\times 10^{20}$ W cm^{-2})	K.E$_{(max)}$ (GeV)	$r_{bl(max)}$ (μm)	t (ps) for $r_{bl(max)}$
10	1.37	0.65	29	3.10
15	3.1	0.79	44	3.29
21	6.4	0.9	50	3.41

The bubble radii and their evolution have already been discussed. But in order to explore the shape of the bubble, both longitudinal and transverse radii have to be considered collectively. Hence, the ratio of longitudinal to transverse radius (r_{bl}/r_{bt}) has been evaluated to study the bubble geometry. The evolution of the ratio of longitudinal to transverse bubble radius for various laser intensity parameters at different propagation time is shown in figure 4. Initially, the bubble shape is almost spherical. But as the bubble traverses behind the laser in plasmas, its shape starts evolving. Shape gradually changes from spherical to elliptical (longitudinally). This can easily be understood using the results reported in figure 1. The longitudinal radius increases till 1.2 picosecond, while the transverse radius continues to fall till 1.5 picosecond. This indicates the enhanced longitudinal to transverse radius ratio (r_{bl}/r_{bt}). Eccentricity is maximum at 1.51 picosecond time. After this point, the ratio of longitudinal to transverse radius starts decreasing because of the increasing trend of transverse

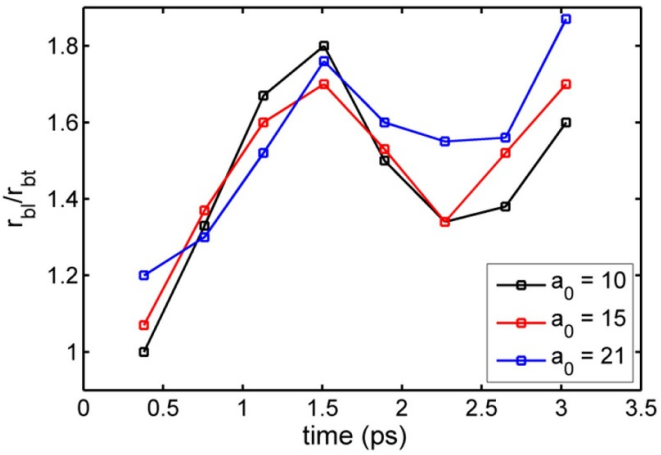


Figure 4. Variation of ratio of longitudinal to transverse bubble radius (r_{bl}/r_{bt}) at different propagation time (in picosecond) for three different values of a_0 ($a_0 = 10, 15, 21$). Other laser-plasma parameters are the same as used in figure 1.

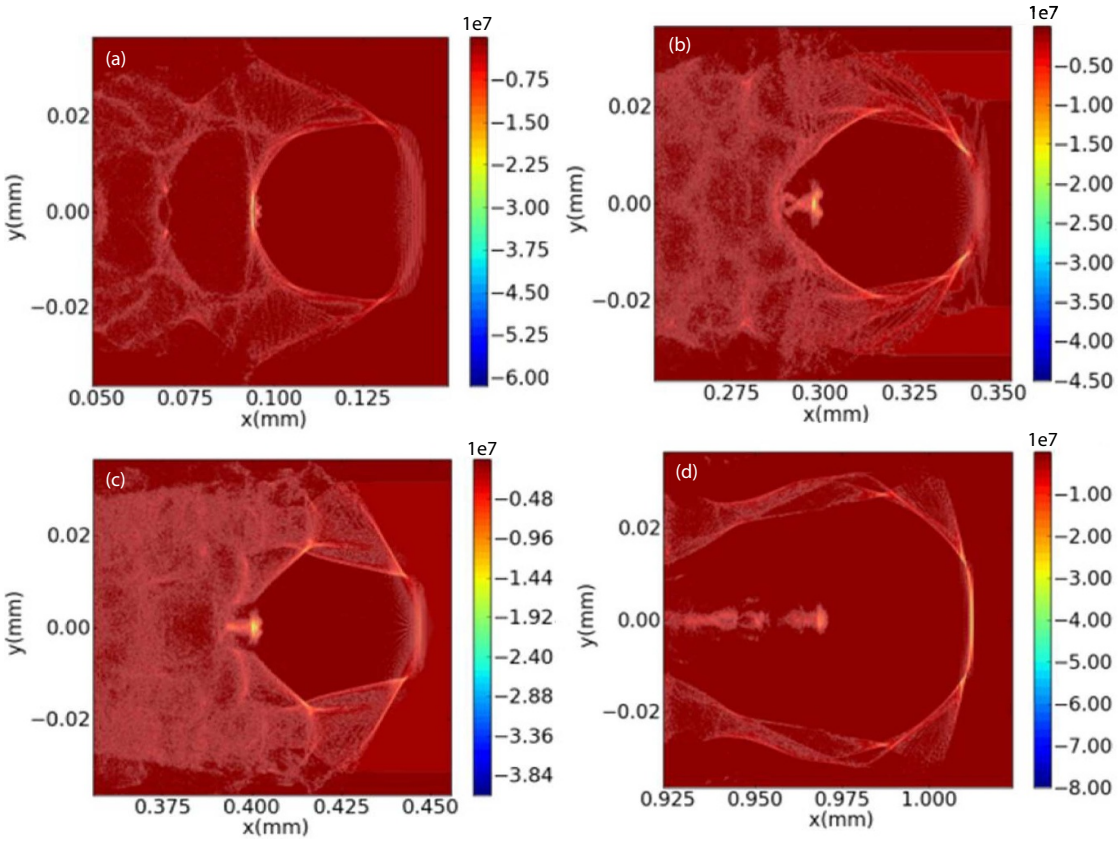


Figure 5. Electron density distribution plots derived from PIC simulation at different propagation time (a) t = 0.5 ps, (b) t = 1.2 ps, (c) t = 1.5 ps, and (d) t = 3.4 ps. The laser strength parameter a_0 considered for this simulation is 21.

bubble radius beyond 1.5 picosecond time. After that the transverse radius increases slowly compared with the longitudinal radius and the bubble remains more elliptical in shape. The bubble shape and size both depend on laser pulse intensity and plasma density.

Our simulation results are in reasonable agreement with the previously proposed outcomes [13, 27, 33]. The bubble becomes less longitudinal ellipsoidal from its previous shape (which was a more longitudinal ellipsoidal) due to increase in residual electrons and slower bubble velocity. As the laser intensity curve shows in figure 2, the laser intensity first increases up to 1.5 ps and then oscillates. The residual electrons inside the bubble alter the electromagnetic fields in the bubble as they carry certain charge and current densities. The slope of the transverse field was found to be reduced, while the slope of longitudinal field remained same. It leads to the increase in the longitudinal radius as compared to the transverse one. In this paper, our results show that the residual

electrons decreases initially due to the increase in laser intensity as a result of their inverse relation. Therefore, the residual electron decreases in the beginning, which consequently leads to increase in the eccentricity of the ellipsoidal bubble up to 1.51 picosecond propagation time. Furthermore, it has been observed that the velocity of bubble is not equal to c (velocity of light) during the entire propagation length. As the laser pulse excites the plasma wake, it starts losing energy to the plasma wave. The front of the laser pulse etches back due to pump depletion with an etching velocity. Hence, the phase velocity of the wake is modified. It is less than the speed of light and also varies with time [34]. Therefore, this may also be a responsible factor for bubble shape evolution. And, of course, this effect may be crucial in petawatt regime. Thus, understanding the evolution of bubble provides an opportunity to explore various ways to control the bubble geometry and, therefore, improve the energy gain and the beam quality in LWFA. So, unlike previous studies, the bubble shape cannot be considered to be constant. It continuously evolves with time, as shown in figure 4. Hence, the results depicted in figure 4 can be used for the scaling of the bubble dimension with the laser intensity for controlling and optimization of electron beam quality in LWFA. The previous studies involving bubble dimensions might get affected because of the evolving bubble shape. Hence, considering the variable shape of the bubble in further blow-out region studies is highly desirable to obtain more realistic results.

Details of bubble evolution describes the mechanism of self-injection, which in turn describes key properties of electron beam such as energy spread, angular divergence, emittance etc. The density plots shown in figure 5 shows that the self-injection of plasma electrons in the bubble is seriously affected by the bubble evolution. It can be observed from figures 1 and 5 that no substantial injection of charge is visible as the bubble has not started to expand yet at 0.5 picosecond time. As the longitudinal radius starts to grow due to laser self-focusing, the self-injection of charge becomes noticeable at a propagation time of 1.2 ps. At t = 1.5 ps, the longitudinal bubble radius falls and the corresponding injected charge in the bubble reduces. This is due to the fact that as the bubble size decreases the injected electrons gets de-trapped from the bubble. A significant amount of charge is injected into the wake only after the bubble radius has started to increase significantly (for example after the time of 1.5 ps in our case). Beyond the propagation time of 1.5 ps, however, the bubble radius continues to increase due to beam loading effect. The simulation results also show that the self-injection of electrons in bubble expansion.

High energy electron beam can be obtained by running the simulations up to the dephasing limit but the beam quality and charge is compromised due to uncontrolled injection. Understanding the mechanism of electron injection in the bubble is crucial to control the beam quality by controlling the injection. This undesirable effect can be avoided if injection is made a one-time event. In this particular paper, the main focus was to investigate the evolution of bubble (blow-out regime) structure in laser wakefield acceleration scheme.

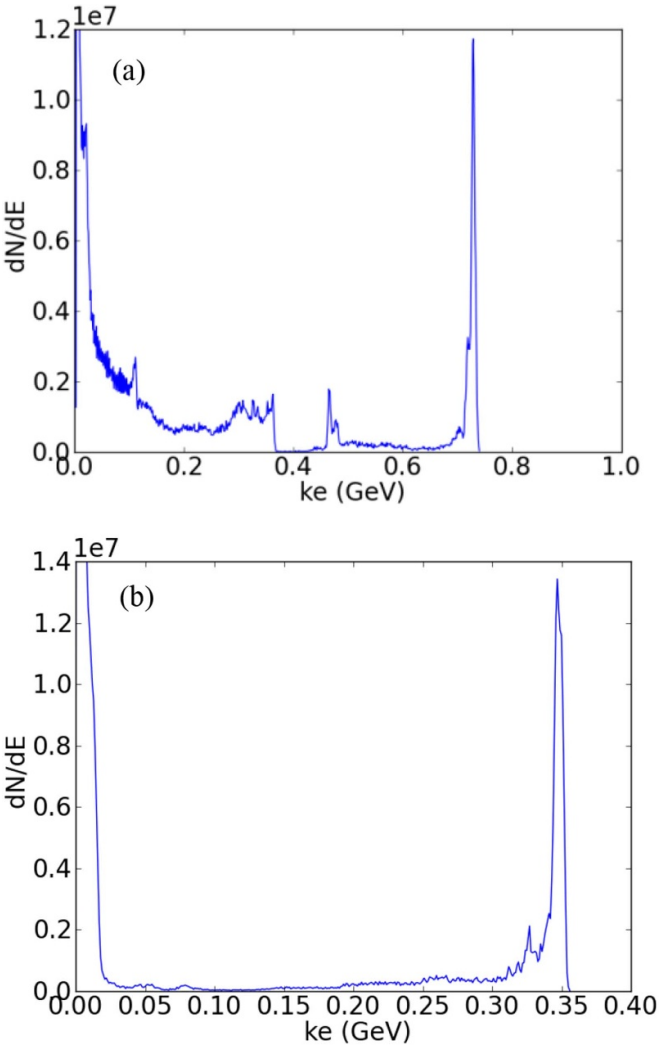


Figure 6. Electron energy spectra for (a) plasma density 6×10^{18} cm^{-3} and (b) 4×10^{18} cm^{-3} for laser intensity parameter $a_0 = 21$. Other laser parameters used are: laser pulse duration $\tau = 20$ fs, laser spot size $r_0 = 12$ μm, and laser wavelength $\lambda = 1$ μm.

Certain optimized parameters were chosen to run the simulation. The plasma density equal to 5×10^{18} cm^{-3} was used to observe the blow-out regime. The injection of electrons in the bubble and formation of electron bunch was noticeable in the simulation results. But, extending our paper to obtain even better quality and higher charged electron bunch, further simulations were carried out by varying plasma density, keeping all other parameters the same. The value of plasma density is optimized to obtain a better quality quasi-monoenergetic electron beam, as shown in figure 6. We obtain the electron energy spectra for different plasma densities ($n = 6 \times 10^{18}$ and 4×10^{18} cm^{-3}). It has been observed that lowering the plasma density reduces the electron energy gain but assists in electron number enhancement. It is due to the reason that the self-injection terminates at a point of shrinkage of bubble and the bunch shortens. The electrons injected later are closer to the bubble axis and are, therefore, seen at a higher accelerating gradient. The back part of the bunch matches up with

the front part in energy and thus, reducing the energy spread. The final energy of the electron bunch can be estimated about 1.3 GeV approximately over a length of 2.5 mm (considering the accelerating gradient can be preserved over the dephasing length).

We predict that the plasma density of 4×10^{18} cm^{-3} produces a highly mono-energetic beam and no secondary injections are observed. A monoenergetic electron bunch obtained of peak energy 0.72 GeV and 0.35 GeV has been obtained for plasma density $=6 \times 10^{18}$ cm^{-3} and 4×10^{18} cm^{-3}, respectively. The computed charge of the electron bunch for density 6×10^{18} cm^{-3} is approximately 294 pC corresponding to the peak energy. This charge estimation for plasma density 4×10^{18} cm^{-3} is about 356 pC. However, the energy spread in both the cases is calculated to be less than 1%. The electron bunch charge delivered by the lower plasma density case is found to be greater than the higher plasma density case. For $a_0 = 21$, the rms normalized transverse emittance of the electron beam computes to about 12.62 and 11.41 π mm mrad for plasma densities 4×10^{18} cm^{-3} and 6×10^{18} cm^{-3}, respectively. Usually, the beam emittance varies with both the bubble size and plasma density. The transverse emittance of the beam is larger for lower plasma density, which is justified with the scaling laws provided by Kneip *et al* [35]. Such electron beams can be of great use for a number of applications requiring a highly charged electron beam like in seeding high-flux γ-ray, electron radiography, radiobiology and in the fast ignition scheme and so on.

3. Conclusions

In this paper, the bubble evolution in laser wakefield acceleration has been predicted via 2D-PIC simulations. The evolution of the bubble has been shown to be dependent on driver evolution. The self-injection dynamics and its relation with the bubble evolution have been discussed. Injection is facilitated by bubble expansion and contraction suppresses it. As beam energy and the beam quality is determined by the electron injection, hence it is necessary to understand the bubble evolution. It has also been demonstrated that continuous injection sometimes degrades the beam quality, therefore controlling injection is fundamental. Plasma density optimization has been done to obtain the improved beam quality. It has been presented that low plasma density leads to the increase in the bunch charge but at the cost of beam energy. It may be quite significant for various applications demanding high charge and a low energy beam such as non-destructive-testing, condensed matter physics, radiolysis and radiotherapy etc.

Acknowledgments

One of the authors (MY) is grateful to University Grant Commission (UGC), Govt. of India, for financial support to carry out this work. This work was supported by the Department of Science and Technology, Govt. of India, and National Research Foundation of Korea (Grant No. 2017R1A2B3010765).

Disclosures

The authors declare no conflits of interest. The data that support the findings of this study is available from the corresponding author upon reasonable request.

References

[1] Tajima T and Dawson J M 1979 *Phys. Rev. Lett.* **43** 267
[2] Faure J, Glinec Y, Pukhov A, Kiselev S, Gordienko S, Lefebvre E, Rousseau J P, Burgy F and Malka V 2004 *Nature* **431** 541
[3] Geddes C G R, Toth C, Tilborg J V, Esarey E, Schroeder C B, Bruhwiler D, Nieter C, Cary J and Leemans W P 2004 *Nature* **431** 538
[4] Leemans W P, Nagler B, Gonsalves A J, Toth C, Nakamura K, Geddes C G R, Esarey E, Schroeder C B and Hooker S M 2006 *Nat. Phys.* **2** 696
[5] Gopal K and Gupta D N 2017 *Phys. Plasmas* **24** 103101
[6] Zeng X *et al* 2014 *Opt. Lett.* **42** 2014
[7] Yu T J, Lee S K, Sung J H, Yoon J W, Jeong T M and Lee J 2012 *Opt. Exp.* **20** 10807
[8] Sung J H, Lee H W, Yoo J Y, Yoon J W, Lee C W, Yang J M, Son Y J, Jang Y H, Lee S K and Nam C H 2017 *Opt. Lett.* **42** 2058
[9] Pukhov A and Meyer-ter-vehn J 2002 *Appl. Phys.* B **74** 355
[10] Rosenzweig J B, Breizman B, Katsouleas T and Su J J 1991 *Phys. Rev.* A **44** R6189
[11] Mora P and Antonsen T M 1996 *Phys. Rev.* E **53** R2068
[12] Gordienko S and Pukhov A 2005 *Phys. Plasmas* **12** 043109
[13] Lu W, Huang C, Zhou M, Tzoufras M, Tsung F S, Mori W B and Katsouleas T 2006 *Phys. Plasmas* **13** 056709
[14] Esarey E, Hubbard R F, Leemans W P, Ting A and Sprangle P 1997 *Phys. Rev. Lett.* **79** 2682
[15] Faure J, Rechatin C, Norlin A, Liftschitz A, Glince Y and Malka V 2006 *Nature* **444** 737
[16] Bulanov S, Naumova N, Pegoraro F and Sakai J 1998 *Phys. Rev.* E **58** R5257
[17] Geddes C G R, Nakamura K, Plateau G R, Toth C, Cormier-Michel E, Esarey E, Schroeder C B, Cary J R and Leemans W P 2008 *Phys. Rev. Lett.* **100** 215004
[18] Gonsalves A J *et al* 2011 *Nat. Phys.* **7** 862
[19] Kalmykov S, Yi S A, Khudik V and Shvets G 2009 *Phys. Rev. Lett.* **103** 135004
[20] Savert A *et al* 2015 *Phys. Rev. Lett.* **115** 055002
[21] Xie B S, Wu H C, Wang H, Wang N Y and Yu M Y 2007 *Phys. Plasmas* **14** 073103
[22] Kostyukov I, Pukhov A and Kiselev S 2004 *Phys. Plasmas* **11** 5256
[23] Sadighi-Bonabi R and Rahmatollahpur S H 2010 *Phys. Plasmas* **17** 033105
[24] Wu H C, Xie B S, Zhang S, Hong X R, Zhao X Y and Liu M P 2010 *Phys. Plasmas* **17** 113103
[25] Kalmykov S Y, Beck A, Yi S A, Khudik V N, Downer M C, Lefebvre E, Shadwick B A and Umstadter D P 2011 *Phys. Plasmas* **18** 056704
[26] Kalmykov S Y *et al* 2010 *High Energy Density Phys.* **6** 200
[27] Benedetti C, Schroeder C B, Esarey E, Rossi F and Leemans W P 2013 *Phys. Plasmas* **20** 103108
[28] Li X F, Yu Q, Gu Y J, Huang S, Kong Q and Kawata S 2015 *Phys. Plasmas* **22** 083112
[29] Birdsall C K, Langdon A B, Vehedi V and Verboncoeur J P 2004 *Plasma Physics via Computer Simulation* (Boca Raton, FL: CRC Press)
[30] Dawson J M 1983 *Rev. Mod. Phys.* **55** 403
[31] Malka V, Faure J, Gauduel Y A, Lefebvre E, Rousse A and Phuoc K T 2008 *Nat. Phys.* **4** 447

[32] Albert F and Thomas A G R 2016 *Plasma Phys. Control. Fusion* **58** 103001
[33] Toosi E S, Mirzanejhad S and Dorranian D 2016 *Laser Part. Beams* **34** 193
[34] Lu W, Tzoufras M, Joshi C, Tsung F S and Mori W B 2007 *Phys. Rev. Spec. Top. Accelerators Beams* **10** 061301
[35] Kneip S *et al* 2012 *Phys. Rev. Spec. Top. Accelerators Beams* **15** 021302

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Real-time Pedestrian Tracking based on Deep Features

Geetanjali Bhola
Department of Information Technology, Delhi Technological University
Delhi, India
geetanjali.getz@gmail.com

Akhil Kathuria
Department of Information Technology Delhi Technological University
Delhi, India
akhilkathuria61@gmail.com

Deepak Kumar
Department of Information Technology, Delhi Technological University
Delhi, India
dgupta330@gmail.com

Chandan Das
Department of Information Technology, Delhi Technological University
Delhi, India
Chandandas.dtu@gmail.com

Abstract—Multi-object tracking is a significant domain in Computer vision applications, which involves giving unique identities to different objects, and maintaining the association between them, for real-time applications. However, most of the trackers fail to achieve decent levels of accuracy as well as speed. In this paper, we propose a tracking technique, which utilizes both high-speed detections from Yolo, as well as deep feature extraction, from a convolutional neural network. The extracted features, along with position vectors and color histograms, are matched between corresponding frames, to develop an association between pedestrians. It can face issues like slight changes in object appearance in continuous frames, such as shape, size or illumination changes, partial occlusions, or re-identification of pedestrians, on re-entering the view, or after being occluded, for a certain length of frames. We have used the Yolo framework for fast object detection, a MobileNet architecture based custom CNN for feature extraction, and a set of algorithms to generate associations between frames. On the publically available Town-centre dataset, our framework can reach a MOTA of 93.2%.

Keywords: Object tracking, Convolutional Network, Feature Extraction

I. INTRODUCTION

Pedestrian tracking has become a major field of interest for computer vision researchers, due to the high amount of data acquired by CCTV cameras, in almost every field of work and the need to extract & analyze object features or human actions. Also, since the manual analysis of people in huge crowds is not feasible, there is a need for automated systems to do the work, with decent accuracy and reliability. Some of the major areas where this is highly necessary and useful involve, tracking of people or live objects, in remote locations, infrastructure planning, based on analyzing the crowd, at different times throughout the day, traffic analysis, such as counting the number of vehicles or pedestrians crossing a particular lane, security purposes, at shopping malls, public places, railway stations, banks, etc, sports, such as to track players or ball, in a football match.

There have been several techniques discovered and employed, to achieve the task of object tracking, with sufficient accuracy and at video frame rates. Most of them can be divided into two domains. The first domain involves techniques based on correlation filters [14] & static object characteristics [12]. Although they can achieve high frame rates, their drawback involves less accuracy, due to their disability to track the high amount of variations in the object's characteristics. The second domain involves techniques using complex convolutional networks [1,2,9,13], which can extract deep and non-linear features of objects, susceptible to variations, such as illumination or size changes. Although these can give high detection and tracking accuracy, they require high computation cost, which proves unaffordable and inefficient, for real-time purposes.

In this work, we have employed a technique that takes advantage of the recent advancements in the field of object detection by neural networks, to track pedestrians frame-by-frame in CCTV footage. We have used YOLO [7], which unlike R-CNNs [10] uses a single-stage detection strategy. It can achieve decent accuracy while running at video frame rates. After extracting the detections, we use a custom CNN model, which can extract deep features, for pedestrians in an automated manner. The model is created by removing the last fully-connected or classification layers and using the late layers, which can extract more generic deep features for objects.

As an improvement to the technique used in [3], along with the deep features extracted by the pre-trained CNN model, we have also combined it with color-histograms, position vectors, and other spatial features, that are used to create associations between detected pedestrians, across adjacent frames. As a result, our tracker has been able to reach an accuracy of 93.2%. The overall framework provides an end-to-end way to detect and track pedestrians, having the accuracy from deep features, and yet speedy enough to be able to process frames in real-time.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:24 UTC from IEEE Xplore. Restrictions apply.

II. RELATED WORK

- The work was done in SORT [1] focused on using a combination of principle techniques like Kalman filtering and Hungarian algorithm, for the association between identities. Still, it was able to reach a MOTA of 33% on MOT benchmark sequences. It also showed that the tracking performance depended heavily on the detection quality too.
- DeepSort [2], as an extension to SORT [1], used a deep-learning metric, based on offline pre-training, on a person re-identification dataset. This reduced the number of mismatches and false positives, by a large amount, depicting the significance of deep features, for tracking purposes.
- Real-time pedestrian tracking in [3] used a well-developed framework for multi-object tracking. However, the features extracted through the pre-trained CNN model were unable to differentiate between similar people, and it could generate results at four frames per second only. It used Fr-CNN [10] framework for detection, which is slower than the newly developed frameworks like Yolo [7].
- The proposed scheme in [9] used a three-fold technique to enhance the object appearance model. It focused on the techniques of transfer learning, feature extraction, and using both online image observations, as well as ground truth data, in initial frames.
- Online Boosting [12] used a feature selection algorithm, which allowed the classifier to adapt online to the background and object appearance changes. It showed how simple, fast-computable features (like color or orientation histograms) can be combined, to form a good real-time tracker.
- GoTurn [13] is one of the few trackers, to be completely based on the offline-training of neural networks. The network learns the relationship between object motion and appearance. It could run at 100fps. However, tracking novel objects with good accuracy would require some kind of online learning too.

III. METHODOLOGY

Our end-to-end framework takes a video input, or a sequence of frames, to create and track identities across the frames. The input is passed onto the Yolo framework [7], which is a state-of-the-art object detection system. It can work in real-time (on Pascal Titan, it processes images at 30fps) along with an mAP of 58% on the COCO test dataset.

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1

where,
t: expansion factor
c: number of output channels
n: repeating number
s: stride
3×3 kernels are used for spatial convolution.

Fig.1 CNN architecture (after MobileNetV2 [6], 2019). The final classification layer has been chopped and so CNN acts as a Feature extractor.

The output from YOLO consists of the bounding boxes for the detections, with specified confidence values. For performing feature extraction, Yolo uses Darknet-53 [7] network, which has 53 convolutional layers and independent logistic classifiers for the class predictions. The detections are then sent to our pre-trained CNN (see Fig.1) which is based on MobileNet [6] architecture (on ImageNet [11] dataset) and it outputs the corresponding features. Then the association of detections with the current pool of identities is done based on the extracted features, position vectors, and color-histograms, with specified threshold and weight values.

The step-by-step sequence for our framework is shown in Fig.2. Frames are extracted from the video input, sequentially, preprocessed, and then sent as input to Yolo detector [7] one-by-one. The pedestrian detections are sent into the pre-trained CNN for feature extraction, their position vectors are stored, and also color-histograms are calculated for r-g-b channels and then flattened and stored. Then they are matched according to a specific algorithm, with the stack of global identities, detected till now, to create

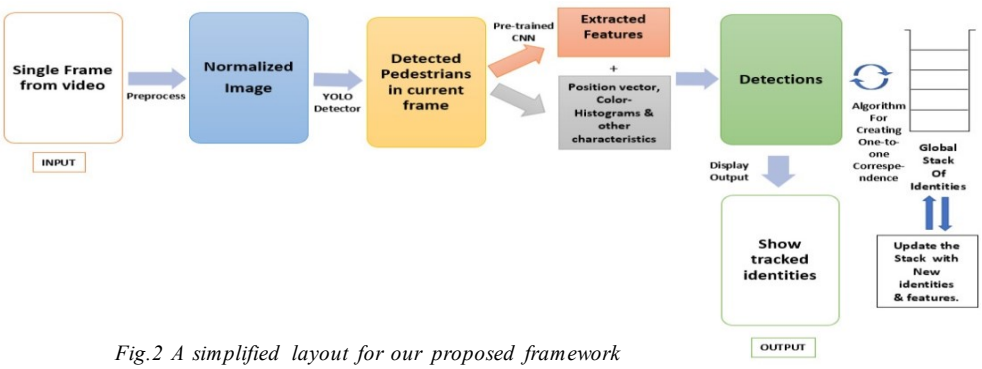


Fig.2 A simplified layout for our proposed framework

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:24 UTC from IEEE Xplore. Restrictions apply.

correspondence among them. At last, global identities and their values are updated, according to the current frame.

A. Datasets employed

For object detection, we have used the Yolo detector [7] which is trained on COCO dataset [17], which is a large scale dataset for purposes like object detection, recognition, or segmentation. It has around 200K labeled images with around 80 object categories and multiple captions per image.

For our pre-trained CNN, we used the ImageNet dataset [11], which is a huge database of images, with more than 14 million hand-annotated images, with many of them also having bounding boxes. It has around 20,000 categories, for which the annotation database is freely available.

B. Features used for matching

- The extracted features from the pre-trained CNN are of the shape (7,7,1280) which are flattened and stored as a 1-D vector. Features are stored as shown in Fig.3(i).
- The coordinates of the bounding boxes of the detections are stored as shown in Fig.3(iii).
- Color histogram for detection is obtained by calculating histograms with 8-bins for R-G-B channels separately, and then flattening them into a single vector of length 512 (8*8*8), as shown in Fig.3(ii).

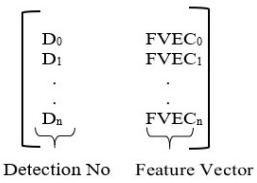


Fig. 3 (i) Feature Matrix

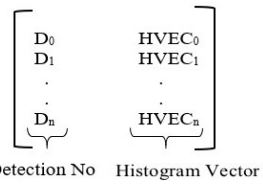


Fig. 3 (ii) Histogram Matrix

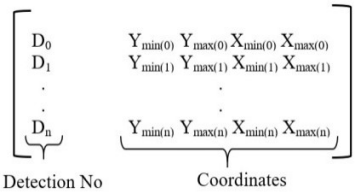


Fig. 3 (iii) Coordinate Matrix

Between any two detections, feature distance (FD), pixel distance (PD), and histogram distance (HD) are calculated as follows:

$$FD = Vec0 - Vec1 \quad (1)$$

$$PD = Euclidean\ distance\ between\ C1\ and\ C2 \quad (2)$$

$$HD = 1 - Histogram\ Intersection\ (H1,H2) \quad (3)$$

Where '–' represents numpy subtraction between two vectors, C1 and C2 are centroids of respective bounding boxes and H1 and H2 are respective color-histogram vectors.

All the above distances are then normalized and assigned specific weight values (w_i) (which are decided based on observations). Individual thresholds for the above distances are calculated according to the difference between the current frame no and the last-seen frame value of the identity in the global stack. Finally, Total distance, TD is calculated as follows:

$$TD = w1.FD + w2.PD + w3.HD \quad (4)$$

where TD should be <= total threshold value and w1,w2,w3 are corresponding weights.

C. Association Algorithm

The ground truth is obtained from the first frame, by the manually marked bounding boxes for pedestrians in the ground truth file. Random identities are given to them, which are stored in the Global identities stack, along with their feature vectors, coordinates, and flattened color-histogram vectors. One more field stored along with the identities is their Last-seen frame value, which is used to adjust the threshold values for matching with later detections, and maintain the order of the identities in the stack (most recent first).

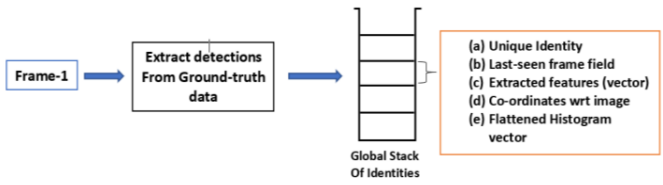


Fig.4(i) Extracting ground-truth from the first frame

If detection is matched with any identity, up to a pre-specified last-seen frame value, then the field for that identity is updated with the latest feature vector (or by averaging it with the previous one), histogram vector, the current coordinate values, and the current last-seen frame value. Otherwise, a new identity is created with the current features and values and stored at the top of the global identities stack.

The matching between a global-stack identity and a current detection involves:

- Calculation of feature distance, pixel distance, and histogram intersection values
- Calculation of individual thresholds according to the difference between the current frame no and the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:24 UTC from IEEE Xplore. Restrictions apply.

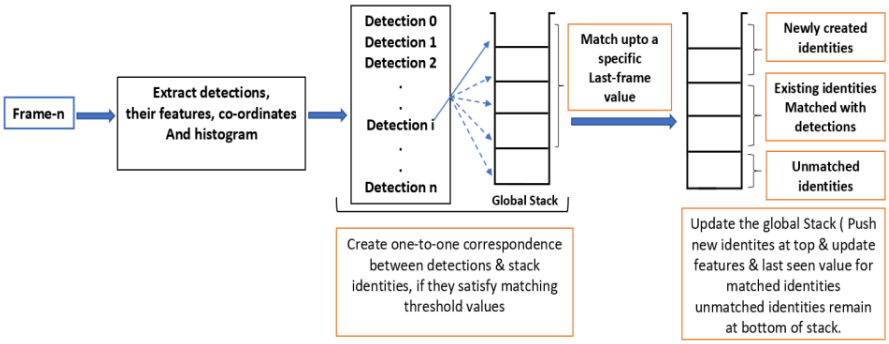


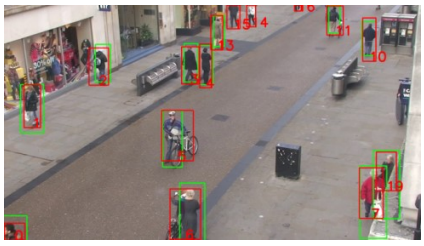
Fig.4(ii) Steps for matching detections with current global identities

last-seen frame value of the identity

- Choosing the closest detection to the global identity, provided it satisfies individual as well as global threshold values.
- Updating the global identity (if matched) with current frame-no as last-seen value and all feature vectors. Otherwise, creating a new one.

IV. RESULTS AND ANALYSIS

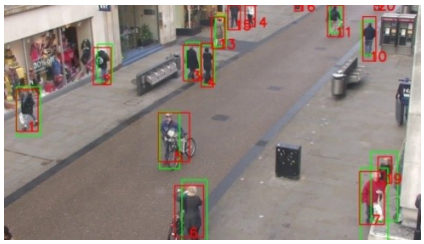
The First 250 frames from the CCTV footage of Town-Centre [15] dataset are extracted and used for tracking analysis. Accuracy and precision are calculated using the identity no of the pedestrians and the bounding boxes marked in the ground truth file.



(i)



(ii)



(iii)



(iv)

Fig.5 Detections with unique identities in three consecutive frames
(i) Identity-19, partially occluded, but still detected
(ii) Identity-19, goes undetected
(iii) Identity-19 re-identified
(iv) Almost all pedestrians detected, despite occlusion

A. Detector Accuracy:

Precision tells what proportion of identifications is correct, whereas Recall tells what proportion of actual positives was identified correctly. Yolo framework [7] is used with a confidence value of 0.5 and with the category of 'person'. The high confidence value along with the specified category results in negligible false positives from other object categories, and hence gives a precision of well above 0.95 in all frames. However, the recall across the frames is shown in Fig.6.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:24 UTC from IEEE Xplore. Restrictions apply.

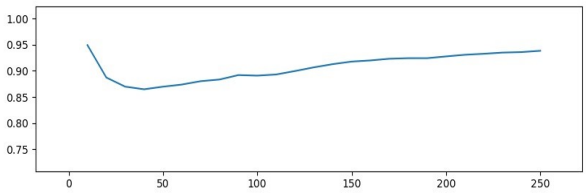


Fig.6 Variation of recall (y-axis) with frames (x-axis)

B. Tracker Accuracy:

The two performance measures used for the evaluation of the algorithm are:

Multiple-object tracking precision (MOTP):-

It gives a measure of the total error in estimated positions, between detections and the ground-truth bounding boxes. It shows the ability of tracker to estimate precise object positions. It is given by:

$$\text{MOTP} = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \tag{5}$$

where d_t^i denotes the distance between centroids of detection and ith pedestrian in ground-truth,
C_t denotes no of matches found in time t.

Experimental results for the Town-centre dataset [15], for the first 250 frames, using this framework gives a MOTP of **12 pixels**.

Multiple-object tracking accuracy (MOTA):-

A higher value of MOTA represents the tracker's ability to consistently recognize and label correct objects over time. It is given by:

$$\text{MOTA} = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \tag{6}$$

where m_t denotes no of misses in detections,
fp_t denotes no of false positives,
mme_t denotes no of mismatches,
g_t denotes total no of objects present, in time t.

Experimental results for the Town-centre dataset [15], for the first 250 frames, using this framework give a MOTA of **93.2%**.

Table 1 Tracking results on town-centre dataset (adapted from [19])

	MOTA	MOTP	MODP	MODA
GMCP [19]	75.59	71.93	72.01	75.71
Benfold et al. [20]	64.9	80.4	80.5	64.8
Zhang et al. [21]	65.7	71.5	71.5	66.1
Pellegrini et al. [22]	63.4	70.7	70.8	64.1
Yamaguchi et al. [23]	63.3	70.9	71.1	64.0
Leal-Taixe et al. [24]	67.3	71.5	71.6	67.6
Proposed method	**93.2**	**12.0**	**75.1**	**91.4**

V. CONCLUSION

In this work, we have presented an efficient multi-object tracking algorithm, which utilizes both high-speed detections from the Yolo framework, as well as deep feature extraction, from a convolutional neural network. Our analysis shows that combining both shallow and deep features, for visual object tracking, can provide a balanced framework, with decent accuracy and ability to work in real-time. Our tracker can achieve an accuracy of 93.2% and a tracking precision of 12px. It can face situations like slight variations in object appearance in continuous frames, partial occlusions, noisy images as well as re-identification of pedestrians (as shown in Fig.5), when they get back in the view, after being occluded, or being out of the view, which has been a drawback of many of the previous algorithms developed in this field (such as in [3]). Also, since we have employed a better CNN model, the mismatching of identities between different pedestrians, has been greatly reduced.

For cases of total occlusion, one can use a position prediction algorithm, for estimating pedestrian movements based on previous trajectories of motion.

REFERENCES

[1] Bewley, Alex and Ge, Zongyuan and Ott, Lionel and Ramos, Fabio and Upcroft, Ben: Simple online and realtime tracking. In: IEEE ICIP (2016)
[2] N. Wojke, A. Bewley and D. Paulus: Simple Online and Real time tracking with a Deep Association Metric. In: IEEE ICIP, Beijing (2017)
[3] Debaditya Acharya, Kourosh Khoshelham, Stephan Winter: Real-time detection and tracking of pedestrians in CCTV images using a deep convolutional neural network. In: Proc. of the 4th Annual Conference of Research@Locate, Australia (2017)
[4] Peter Karkus, David Hsu, Wee Sun Lee: Particle Filter Networks with Application to Visual Localization. In: Proceedings of The 2nd Conference on Robot Learning, PMLR 87:169-178 (2018)
[5] Susan, Seba, and Seema Chandna:. Object Recognition from Color Images by Fuzzy Classification of Gabor Wavelet Features. In: Computational Intelligence and Communication Networks (CICN), 2013 5th International Conference on, pp. 301-305. IEEE (2013)
[6] Mark Sandler, Andrew Howard, Menglong Zhu: MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: IEEE-CVPR (2018)
[7] Joseph Redmon, Ali Farhadi: YOLOv3: An Incremental Improvement. In: arXiv:1804.02767 [cs.CV] (2018)
[8] Keni Bernardin, Alexander Elbs, Rainer Stiefelhagen: Multiple Object Tracking Performance Metrics and Evaluation in a Smart Room Environment. In: IEEE-Visual Surveillance (2006)
[9] Chen, Y., X. Yang, B. Zhong, S. Pan, D. Chen, and H. Zhang: Cnntracker: Online discriminative object tracking via deep convolutional neural network. In: Applied Soft Computing 38, 1088 – 1098 (2016)
[10] S Ren: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: Advances in Neural Information Processing Systems 28, pp. 91–99 (2015)
[11] ImageNet dataset available at, https://www.image-net.org
[12] Grabner, H., Grabner, M., Bischof, H.: Real-Time Tracking via On-line Boosting. In: Proc. British Machine Vision Conf. (2006)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:24 UTC from IEEE Xplore. Restrictions apply.

[13] D Held: Learning to Track at 100 FPS with Deep Regression Networks. In: Leibe B., Matas J., Welling M. (eds) Computer Vision – ECCV (2016)
[14] Qiang Wang, Jin Gao, Junliang Xing, Mengdan Zhang, Weiming Hu: DCFNet: Discriminant Correlation Filters Network for Visual Tracking. In: arXiv:1704.04057 [cs.CV] (2017)
[15] Town Centre dataset available at, https://www.robots.ox.ac.uk/ActiveVision/Research/Projects/2009bbenfold_headpose/project.html
[16] Tripathi, G., Singh, K. & Vishwakarma, D.K.: Convolutional neural networks for crowd behaviour analysis: a survey. In: Vis Comput 35, 753–776 (2019)
[17] COCO dataset available at, https://www.cocodataset.org
[18] Vijayakumar, T.: COMPARATIVE STUDY OF CAPSULE NEURAL NETWORK IN VARIOUS APPLICATIONS. Journal of Artificial Intelligence, 1(01), 19-27 (2019)
[19] Dehghan A., Idrees H., Zamir A.R., Shah M. (2014) Automatic Detection and Tracking of Pedestrians in Videos with Various Crowd Densities. In: Weidmann U., Kirsch U., Schreckenberg M. (eds) Pedestrian and Evacuation Dynamics 2012. Springer, Cham
[20] Benfold, B., Reid, I.: Stable multi-target tracking in real time surveillance video. In: CVPR.(2011)
[21] Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using networkflows. In: CVPR. (2008)
[22] Pellegrini, S., Ess, A., Van Gool, L.: Improving data association by joint modeling of pedestrian trajectories and groupings. In: ECCV. (2010)
[23] Yamaguchi, K., Berg, A., Ortiz, L., Berg, T.: who are you with and where are you going? In: CVPR. (2011)
[24] Leal-Taixe, L., Pons-Moll, G., Rosenhahn, B.: Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In: ICCV Workshops. (2011)

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:32:24 UTC from IEEE Xplore. Restrictions apply.

International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 11, Issue 5, May 2020, pp. 409-419, Article ID: IJARET_11_05_043
Available online athttp://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=5
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
DOI: 10.34218/IJARET.11.5.2020.043

© **IAEME** Publication **Scopus** Indexed

REFINING YOLOV4 FOR VEHICLE DETECTION

Pooja Mahto, Priyamm Garg, Pranav Seth and J Panda
Department of Electronics & Communication Engineering,
Delhi Technological University, India

ABSTRACT

Real-time vehicle detection is a technology employed in applications like self-driving cars, traffic camera surveillance. Every year we see better and updated state-of-the-art (SOTA) object detectors, but as those are trained on general-purpose datasets (like MS COCO), we miss out on targeted model improvements for vehicular data. The aim of this paper is to improve the newly released, YOLOv4 detector, specifically, for vehicle tracking applications using some existing methods such as optimising anchor box predictions by using k-means clustering. We also carefully hand-pick and verify some key techniques mentioned in the original paper, to optimise YOLOv4 as per the requirements of our dataset (UA-DETRAC).

Our fine-tuned model is also compared with the existing models on a number of performance metrics such as - precision, recall, F1 score, mean average precision, and the average IoU. Our experimental results show that the SOTA model which already has real-time object detection capabilities can be further improved for highly targeted use cases. We urge the readers to expand the scope of the paper (and the original model) to other specific situations as well.

Key words: object detection, convolutional neural network (CNN), YOLOv4, vehicle detection, k-means clustering

Cite this Article: Pooja Mahto, Priyamm Garg, Pranav Seth and J Panda, Refining Yolov4 for Vehicle Detection, *International Journal of Advanced Research in Engineering and Technology (IJARET)*, 11(5), 2020, pp. 409-419.
http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=5

1. INTRODUCTION

Object Detection is a sub-field of computer vision that is revolutionizing the way we identify things. Object detection and tracking are important and challenging tasks in many computer vision applications such as surveillance, vehicle detection and autonomous robot navigation. Object detection differs from object classification in terms of bounding box formation. We try to draw a bounding box around the specific object in detection algorithms. There can be multiple bounding boxes in an image, detecting different objects in the image.

There are two approaches to object detection: 1.) Machine learning-based approach, and 2.) Deep learning-based approach.

For the machine learning approach, there is a necessity of defining features, then using techniques for classification. Viola, Jones at al proposed the Viola-Jones algorithm using a boosted cascade of simple Haar features [1]. Scale-invariant feature transform (SIFT) [2] and Histogram of oriented gradients (HOG) [3] are also feature extraction techniques. Support Vector Machines (SVM) are used for classification.

For the deep learning approach, defining features are not important, as these perform end-to-end object detection. Convolutional Neural networks (CNN) is a class of deep neural networks, which find application mainly in image visualisation. Zhao et al provided a review of object detection using a deep learning approach [4]. Region-based CNN algorithms (R-CNN, Fast R-CNN, Faster R-CNN), Single Shot MultiBox Detector (SSD), YOLO are some deep learning-based approaches to object detection.

A standard CNN cannot be used for object detection, due to the variable length of the output layer, as there can be multiple occurrences of the same object in an image. One way to approach the problem of multiple occurrences of an object in an image is the division of the image into different regions of interest and using a CNN for classification of an object's presence in that region. But the problem with this approach is that the objects of interest tend to have different spatial locations in the image. This results in high computation due to a large number of regions of interest.

R-CNN algorithm was proposed by Girschick et al [5] as a solution to the problem of a large number of regions of interest. They proposed a method of extracting 2000 regions from the image, called as 'region proposals', using selective search, and classifying in those region proposals only instead of a large number of regions. Fast R-CNN improvised the previous R-CNN algorithm by generating a convolutional feature map by directly feeding the image to the CNN [6]. Different region proposals are identified from the feature map. This greatly reduced the time to test an image compared to the R-CNN. Shaoqing et al further developed the Fast R-CNN, by removing the selective search and allowing the network to learn the region proposals. They named it Faster R-CNN, as it takes even less time than the previous Fast R-CNN to test the image [7].

A typical object detector architecture consists of four components- the input, the backbone, the head, and the neck. The backbone is the pre-trained network taking the input image and does feature extraction. The head predicts the classes and bounding boxes of objects. The neck serves the purpose of increasing robustness by collecting feature maps from intermediate stages.

These algorithms can be trained and optimized for the problem of vehicle detection. In this paper, we have used the object detection algorithm YOLOv4 and optimised it for vehicle detection. YOLOv4 provides higher accuracy and faster results so as to implement real-time vehicle detection.

The rest of the paper is organized as follows. Section 2 provides the development of the YOLO algorithm, talking about its two latest versions, v3 and v4. Section 3 shows our proposed refinements in the YOLOv4 algorithm to optimize it for more accurate real-time vehicle detection. Section 4 and 5 focus on the implementation and the comparisons with previous techniques, respectively.

2. EXISTING WORK (YOLO)

Joseph Redmon et al formed a new algorithm in 2015 called You Only Look Once (YOLO) [8]. In YOLO, a single convolutional network predicts the classes and bounding boxes for the whole image in one run of the algorithm, unlike the region-based classification algorithms.

The input image is divided into an S x S grid, and each grid is responsible for the detection of an object. In each grid, we take m bounding boxes, and for each bounding box, the network provides an offset value for bounding box and class probability. The bounding boxes which have a class probability above a specific threshold are selected and are further used to locate the object in the image.

2.1. YOLOv3

In their paper for YOLOv3, Redmon et al showed an incremental improvement over YOLOv2 [9][10]. YOLOv3 improvised over its predecessor in various techniques. Firstly, it improved the bounding box prediction by using dimension clusters as anchor boxes. The network then uses logistic regression to predict an objectness score for each bounding box. There is a threshold for each classifier, and classes with higher scores than the threshold are assigned to the bounding box. Secondly, independent logistic classifiers replaced the softmax for better class prediction. Each box then predicts the classes the bounding box may have by multiclass classification. Thirdly, increasing the number of convolutional layers to 53, using the Darknet-53 framework, which is the backbone behind YOLOv3. The network is a hybrid of YOLOv2, Darknet-19, and consists of 3 x 3 and 1 x 1 filters with shortcut connections. Lastly, it makes predictions across 3 different scales, which is similar to the Feature Pyramid Network (FPN) [11]. It adds several convolutional layers from the base feature extractor. At each scale, 3 bounding boxes are predicted. This ultimately gives better features from early in the network, and the predictions at the last layers are benefited from the early computations.

Lecheng Ouyang et al implemented vehicle target detection based on YOLOv3 in complex scenes and showed advantages over traditional target detection algorithms in accuracy and speed [12]. They showed how YOLOv3 can be used for vehicle detection, and it gives an accuracy of 89.16% at 21fps on the VOC dataset. YOLOv3 is extremely fast, running in 22ms at 320 x 320 with 28.2 mAP. Despite being 3 times faster than RetinaNet [13], it has lower accuracy.

2.2. YOLOv4

Bochkovskiy et al proposed the YOLOv4 algorithm with significant changes from the previous version, and much better accuracy [14]. Published in April 2020 it is the latest, most advanced iteration of YOLO, and the first developed by the original authors (Redmon et al). To achieve better accuracy, they designed a deeper and complex network, where they used Dense Block. It contains multiple convolutional layers, with batch normalization, ReLU, after which convolution takes place. Further, a Dense Net is formed consisting of multiple Dense Blocks with transition layers in between [15]. Cross-Stage Partial connections (CSP) then separate the input feature maps of Dense Blocks into 2 parts [16]. One part goes directly to the input to the next transition layer, and the other part will go through the Dense Block. This lowers the computational needs as only one part is going through the Dense Block. For the backbone of the feature extraction, they used the CSPDarknet-53, which uses the CSP connections along with Darknet-53 from the previous YOLOv3. Spatial Pyramid Pooling (SPP) is used as the neck over CSPDarknet-53 as it increases the receptive field, differentiates the most significant feature and does not cause a reduction in speed [17]. In place of the Feature Pyramid Network (FPN) in YOLOv3, here they used Path Aggregation Network (PANet) [18]. For the head, they used the original YOLOv3 network.

Apart from the architecture, it consists of a training strategy to get better accuracy without the extra cost to hardware, which is termed as *Bag of Freebies*. With these, we can get better performance for 'free'. Another set of strategies which give better results at low cost, but not completely free, were termed as the *Bag of Specials*.

There are some additional improvements on design for a robust training on a single GPU like the introduction of new data augmentation methods like CutMix, Mosaic, etc., selection of hyper-parameters while applying general algorithms, and modifications to existing methods to make their design suitable for better detection and training. The creators of YOLOv4 have chosen *CSP darknet 53* backbone, *SPP additional module, PANet path-aggregation* neck, *Greedy NMS and YOLOv3* (anchor-based) head as the architecture of YOLOv4.

3. PROPOSED REFINEMENTS IN YOLOV4

We propose the following *additions to YOLOv4* and used a few techniques from the *Bag of Freebies/Specials* to refine vehicle detection on the UA-DETRAC Dataset:

- Anchor box optimisation using k-means clustering (ABK)
- Non-maximum suppression using distance-IoU (DIoU-NMS)
- Spatial Attention Module (SAM)
- Self-adversarial training (SAT)

Anchor boxes (or priors) are used extensively in single-layer object detection algorithms to set the initial dimensions of bounding boxes, as they are better than other unsupervised learning algorithms which are biased towards bounding boxes with large dimensions. These initial dimensions are then corrected, and resized to the ground truth box dimensions, based on some feedback from the neural network. This feedback can be provided by using the K-means clustering technique [26], which initialises the normalisation (correction) process by taking 'K' random boxes (aka means or centroids) as cluster heads. Clusters are then repeatedly assigned around the nearest centroid, and updated based on a certain *threshold* value, until convergence. This threshold can be taken in terms of the Euclidean distance or the Jaccard Index (IoU value), but the latter turns out to be more accurate. Intersection over Union is the overlap between the anchor and the ground truth box, divided by the non-overlap. If the highest IoU is greater than 50%, the anchor box is said to detect the object. Otherwise, it is said to be ambiguous or unlearned.

To ascertain the number of centroids (the k-value) we need to look at the trade-off between the ideal overlap value and the tolerable training time [9]. Increasing the number of centroids will definitely increase the IoU (overlap) but it will also substantially increase the computational overhead, as there is a linear increment in the number of convolution filters resulting in increased training time. Experimentally, we found the average IoU value to be 78.30% for K=12, (as compared to 75.83% for K=9) to be the most optimal. Subsequent increments in the K value didn't translate into meaningful results. As will be seen later, ABK yields a 4%-point increase in mean average precision (mAP@.50), at K=12, by itself.

Bounding box regression is used to optimise the loss function. Zheng et al [19], proposed using distance-based IoU loss instead of generalised IoU (GIoU) over problems of slow convergence. DIoU uses normalised distances between the ground truth (gt) and the bounding box. Non-maximum suppression (NMS) is used to ensure that an object appearing in multiple boxes in the grid is calculated only once. Using IoU as the criterion to eliminate redundant detection boxes, doesn't work well on cases with object occlusion. Hence using DIoU with NMS [20] makes the system less susceptible to occlusion because of the consideration for central distance along with overlap area.

Convolution Block Attention Module (CBAM) is an attention module for feed-forward convolutional neural networks. Here Woo et al [21], propose a simultaneous application of Max Pool and the Average Pool layer to generate a spatial attention map, by concatenating them and then convolving them. Such a Spatial Attention Module (SAM) causes adaptive feature refinement. In YOLOv4, the authors modify SAM to focus on point-wise attention, effectively removing the maximum and average pooling layers. We used SAM at down-sampling points (encoder) and up-sampling points in the detector (decoder). As will be observed later, SAM yields a 12%-point increase in average IoU, by itself.



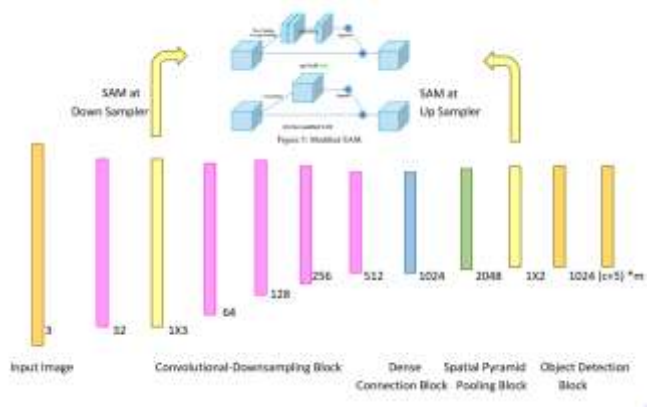
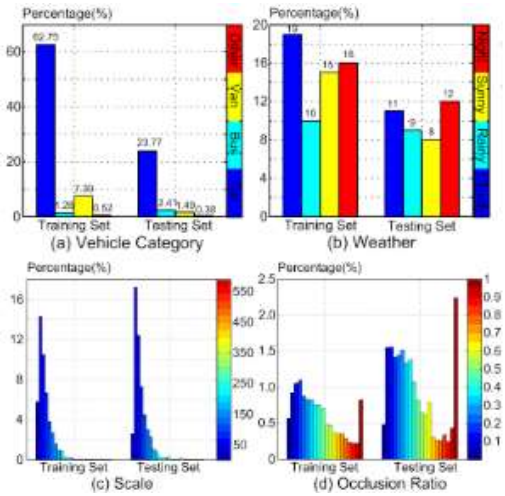


Figure 1. Predicted Anchor Boxes **Figure 2**. Refined YOLOv4 with Modified SAM

Self-Adversarial Training (SAT) is a data augmentation technique using back-propagation [22] with two iterations of forward-backwards passes. On the first backward pass, to minimize the cost function, we alter the original image instead of the network weights. Contrary to the usual method, this actually degrades the network performance, or simply put, the model performs an "adversarial attack" on itself. Now, this modified image is used to train the network on the second iteration. This way we are able to reduce overfitting and make the model more universal [14]. As will be observed later, SAT yields an 8.2%-point increase in mean average precision (mAP@.50), by itself.

4. DATASET

For training and testing, we use the UA-DETRAC Benchmark dataset [23]. We owe our debt of gratitude to the authors for publicly providing this diverse dataset on vehicle applications for academic research. It consists of 100 video sequences, shot at 25fps at a resolution of 960X540, divided in a 60-40 ratio for training and testing use cases. To avoid overfitting, both kinds of videos are taken at different locations but share similar attributes and traffic conditions. The test set is further divided into *easy*, *medium* and *hard* levels of difficulty based on the Edge Box detection rate [24]. Certain areas of low-resolution are also marked as *ignore-regions* to lower the computational overhead. Over 140,000 manually annotated frames of the dataset are each annotated, using attributes like vehicle category (Bus, Car, Van, Others), occlusion, weather (Sunny, Rainy, Cloudy, Night), truncation, and scale. The dataset specifications and sample data are shown below:



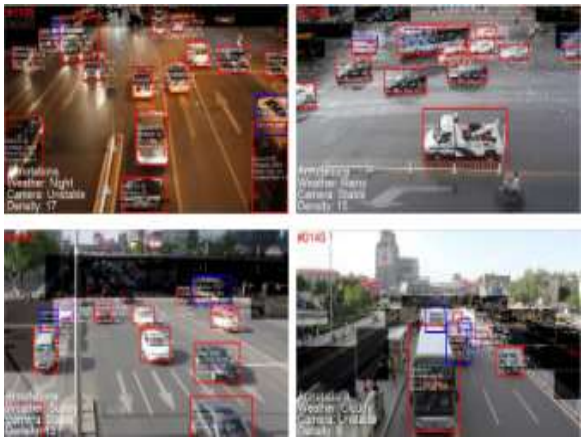


Figure 3. UA-DETRAC Classification **Figure 4**. Annotated Frames from Dataset

The data annotation files for DETRAC-Training set were in the XML format and had to be first converted into the VOC format. Then the corresponding images are migrated into the target directory. After this, the data is converted into separate TXT formats suitable for working with YOLO.

5. EXPERIMENTATION

We test the efficacy of various improvement techniques on UA-DETRAC dataset in order to refine and tune YOLOv4 for better Vehicle Detection.

5.1. Experimental Setup

We set the hyper-parameters of YOLOv4 as follows: the training steps are 8000; the step decay learning rate scheduling strategy is adopted with initial learning rate 0.013 and multiplied with a factor 0.1 at 6400 steps and 7200 steps, respectively; The momentum and weight decay are respectively set as 0.9 and 0.0005. All architectures use a single GPU to execute multi-scale training in the batch size of 64 and mini-batch size of 64. We adopt default momentum 0.949, IoU threshold for assigning ground truth 0.213, and loss normalizer 0.07 for genetic algorithm experiments as proposed by authors of YOLOv4. We experimented with a large number of BoF, provided by YOLOv4 including data augmentation using flip, mosaic data augmentation, class label smoothing, self-adversarial training, Optimized Anchors, different kinds of IoU losses. We also conducted experiments on various BoS, including anti-aliasing, assisted excitation, NMS techniques. For all experiments, we used Google's Jupyter notebook Colab with one GPU (1xTesla K80, compute 3.7, 2496 CUDA cores, 12GB GDDR5 VRAM) for training. Therefore, techniques such as *syncBN* which are used to optimize the use of multiple GPUs were not deployed.

5.2. Methodology

We started with a comparison of the default YOLOV3 model with the default YOLOV4 model. Based on the results we shortlisted techniques from BoF and BoS which could be used to further enhance the performance of YOLOv4 for UA-DETRAC data set. From BoF apart from cut-mix and mosaic data augmentation technique and label smoothing we use flip data augmentation technique, self-adversarial training technique. From BoS, we use DIoU-NMS and SAM. We also use 12 customised anchors generated by K-mean Clustering, instead of the 9 standard Anchors of YOLOv4.

6. RESULTS AND COMPARISONS

We use the following performance metrics to compare the various models and iterations. (Assuming TP, TN, FP, FN denote the true positives, true negatives, false positives and false negatives, respectively)

6.1. Average IoU

Intersection over Union is the ratio of area common [intersection] between the ground-truth box (B_{gt}) and the predicted bounding box (B_p), and the total area by these two boxes [union]. Mathematically,

$$IoU = area(B_p \cap B_{gt}) / area(B_p \cup B_{gt}) \; 0 \leq IoU \leq 1$$

Average IoU is the IoU value averaged over the entire dataset.

6.2. Precision

Precision is a measure of what proportion of predicted positives were truly positive. Mathematically,

$$Precision = TP / (TP + FP)$$

6.3. Recall

Recall is a measure of what proportion of actual positives were classified correctly.

$$Recall = TP / (TP + FN)$$

6.4. F1 Score

F1 score [25] is the harmonic mean of recall and precision. It maintains a balance between recall and precision, as there is a trade-off between the two (one value is improved at the expense of the other). Mathematically,

$$F1\ Score = 2 \times (Precision.\ Recall) / (Precision + Recall)$$

6.6. Average Precision

Average precision is a measure that combines recall and precision for ranked retrieval results. For one information need, the average precision is the mean of the precision scores after each relevant document is retrieved.

6.6. Mean Average Precision (mAP [0.50])

Mean of the Average Precisions calculated over multiple classes (here 4), using single IoU threshold of 0.5.

We tested our models for 40 videos divided into 56,167 frames consisting of four classes i.e., Bus, Car, Van, Others. In comparison between the default YOLOv3 model and the default YOLOv4 model, we found that YOLOv3 outperforms YOLOv4 for the *Bus* class. And performs almost the same for the *Van* class. While analysing weather conditions we noticed YOLOv3 performs much better than YOLOv4 for night condition with an 8%-point higher mAP(@.50). After analysing the architecture of both the default models we reached to the following conclusions:

- Higher AP of the *bus* class indicates the anchor size of YOLOv3 is more suitable for detection of buses than anchor sizes of YOLOv4.
- Poor performance of YOLOv4 during the *night* is a result of the mosaic data augmentation technique used, as mosaic combines four images together. But due to

the difference in background intensity levels, the model is not able to adapt to the changes.

- Low precision indicates that YOLOV4 detected lesser TP than YOLOv3.
- Average IoU of YOLOv4 is 10% lesser than average IoU of YOLOv3 which again indicates incompatible anchor boxes.
- Low AP of the *others* class indicates that both the models were unable to detect vehicles of this class properly, as they were not well-defined and hence, detected as noise by the system. *Others* include bikes, trucks and other vehicles.



Figure 5 Detection on the Test data

In *YOLOv4+ABK*, we used 12 anchor boxes generated by K- Means Clustering Method. AP of class bus, van, others are increased by a significant amount of 10.72% , 8.35%, 1.23% respectively whereas we notice a drop in AP of class cars. Average IoU has increased by 15% and we notice a significant increase in overall the model with a 4%-point increment in mAP(@.50).

In *YOLOv4+SAT*, we use Self Adversarial Training which adds noise while training the model to make the model robust to noises in test data and adversarial attacks. YOLOv4+SAT increments the AP of class Others by a point of 17.72% and bus by 12.14% as compared to the default YOLOV4 model. Increment of 8.6%-point in mAP(@.50) and 14%-point in Average IoU is observed.

YOLOv4+SAM uses Spatial Attention Module which tells the model where to focus while training and detection, basically where to look for the features in a given image. This helps the model to overcome the intensity problem faced during the *night* condition. As a result, *average IoU* is increased by 13%-point and an increment of 0.5%-point in mAP(@0.50) is noticed.

Refined YOLOv4 uses all of the above-mentioned techniques i.e. ABK, SAT, SAM and uses DIoU NMS for non-maximal suppression which results in an AP of 85.11%, 76.57%, 59.67%, 49.25% for bus, car, van and others. High Precision of 77%, high recall of 71% and an F1-score of 74% is observed. Average IoU was found to be 65.7%. And an overall increment of 10%-point in mAP(@.50) at 67.7% is observed. **Refined YOLOv4: YOLOv4+ABK+SAT+SAM*

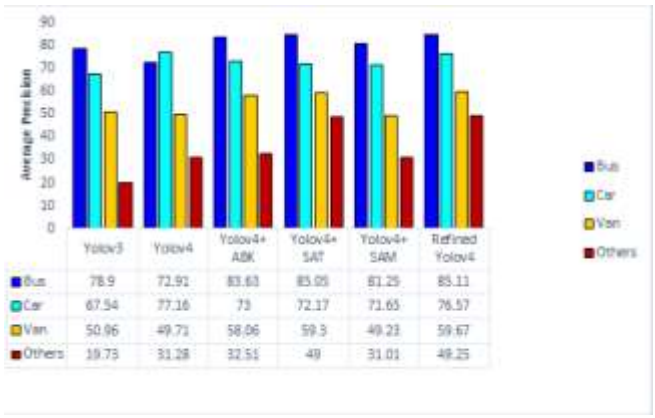


Figure 6: Influence of different YOLO models on AP of different classes

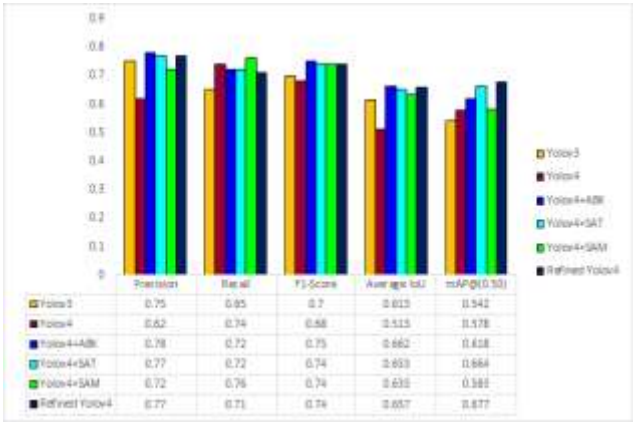


Figure 7: Influence of different YOLO models on various performance metrics.

7. CONCLUSION

In this paper, we fine-tuned the new state-of-the-art object detection algorithm YOLOv4, to specifically suit the needs of vehicle detection, using some pre-existing tweaks and techniques. Wherever possible we have experimentally verified the results and presented the same in the paper. The refined YoloV4 model boosted the performance in terms of each individual aforementioned metric. This final combined model gives benchmark results with an mAP of 67.7% (10%-point higher than base model) on the DETRAC-test dataset. We also observed YOLOv4 to be definitely faster than the previous iterations. In our calculations it was able to test around 57K frames in 25 minutes on an average, translating into a frame rate of around 38 FPS, enabling real-time computation on a single GPU. The results offer a promising scope of application of varied techniques and tweaks to optimise YOLOv4 for other specific situations (IC detection, Crack detection, face detection etc).

REFERENCES

[1] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*(2), 137-154.

[2] Lowe, David G. (1999). Object recognition from local scale-invariant features. *Proc. 7th International Conference on Computer Vision (ICCV'99)* (Corfu, Greece): 1150-1157. doi:10.1109/ICCV.1999.790410.

[3] Navneet Dalal, Bill Triggs. (2005), Histograms of Oriented Gradients for Human Detection. *International Conference on Computer Vision & Pattern Recognition (CVPR '05)*, Jun, San Diego, United States. pp.886–893, ff10.1109/CVPR.2005.177ff. ffinria-00548512f

[4] Zhao, Zhong-Qiu & Zheng, Peng & Xu, Shou-Tao & Wu, Xindong. (2019). Object Detection with Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems*. PP. 1-21. 10.1109/TNNLS.2018.2876865.

[5] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[6] Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448).

[7] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91-99).

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779-788).

[9] Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263-7271).

[10] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

[11] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).

[12] Ouyang, L., & Wang, H. (2019, July). Vehicle target detection in complex scenes based on YOLOv3 algorithm. In *IOP Conference Series: Materials Science and Engineering* (Vol. 569, No. 5, p. 052018). IOP Publishing.

[13] T. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, (2017), "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 2999-3007, doi: 10.1109/ICCV.2017.324.

[14] Bochkovskiy, A., Wang C., Mark Liao, H. (2020). "YOLOv4: Optimal Speed and Accuracy of Object Detection". arXiv:2004.10934v1 [cs.CV] 23 Apr 2020

[15] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

[16] Wang, C., Liao, H.M., Yeh, I., Wu, Y., Chen, P., & Hsieh, J. (2019). CSPNet: A New Backbone that can Enhance Learning Capability of CNN. *ArXiv, abs/1911.11929*.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. (2015), Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 37(9):1904–1916.

[18] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. (2018), Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 8759–8768.

[19] Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. (2020). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. AAAI 2020

[20] Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). Soft-NMS--improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision* (pp. 5561-5569).

[21] Woo, S., Park, J., Lee, J., & Kweon, I. (2018). CBAM: Convolutional Block Attention Module. *ArXiv, abs/1807.06521*.

[22] Hecht-Nielsen, R. (1992). Theory of the backpropagation neural network. In *Neural networks for perception* (pp. 65-93). Academic Press.

[23] Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M. C., Qi, H., & Lyu, S. (2015). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint arXiv:1511.04136*.

[24] Zitnick, C. L., & Dollár, P. (2014). Edge boxes: Locating object proposals from edges. In *European conference on computer vision* (pp. 391-405). Springer, Cham.

[25] Sasaki, Yutaka. (2007). The truth of the F-measure. Teach Tutor Mater.

[26] Jin X., Han J. (2017) K-Means Clustering. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning and Data Mining. Springer, Boston, MA

[27] Suraj Nair Aiyappa, Ramamurthy B, (2018), An Efficient Approach Towards Clustering Using K-Means Algorithm, International Journal of Civil Engineering and Technology, 9(2), pp. 705–714

[28] Sasikaladevi V and Mangai V, (2018), Colour Based Image Segmentation Using Hybrid K means with Watershed Segmentation, *International Journal of Mechanical Engineering and Technology*, 9(8), pp. 1367–1377.

[29] Ch. Ramesh, Dr. N.B. Venkateswarlu, Dr. J.V.R. Murthy, (2012), A Novel K-Means Based Jpeg Algorithm for Still Image Compression, *International Journal of Computer Engineering and Technology*, 3(1), pp. 339–354

[30] Chandra Das, Shilpi Bose, Matangini Chattopadhyay, Samiran Chattopadhyay, (2014), A Novel Distance Based Modified K-Means Clustering Algorithm for Estimation of Missing Values In Micro-Array Gene Expression Data, *International Journal of Information Technology & Management Information System*, 5(3), pp. 1–13

Silicon
https://doi.org/10.1007/s12633-020-00531-0

ORIGINAL PAPER

Reliability of Sub-20 nm Black Phosphorus Trench (BP-T) MOSFET in High-Temperature Harsh Environment

Ajay Kumar[1,2] · Neha Gupta[3] · Rishu Chaujar[4]

Received: 15 February 2020 / Accepted: 13 May 2020
© Springer Nature B.V. 2020

Abstract
In this work, the high-temperature reliability of the Black Phosphorus Trench (BP-T) MOSFET device has been analyzed. When the temperature is very high (500 K), the proposed device (BP-T-MOSFET) shows very high reliability with less effects in the device performance. The device is also calibrated with the experimental data. All the simulation has been performed using ATLAS-3D device simulator. BP-T-MOSFET shows high-temperature reliability in terms of reduced sub-threshold current, hot-electron injection gate current, and impact ionization substrate current. Further, higher switching ratio (I_{on}/I_{off}), electron mobility, and electron velocity have also been observed owing to the electrical properties of black phosphorus used in the trench gate which shows the potential in a high-temperature environment. Owing to trench structure and electrical properties of black phosphorus, BP-T-MOSFET has significant device efficiency and low leakage currents at high temperatures which prove as highly shrieked MOS device applications in harsh temperature environment.

Keywords BP-T-MOSFET · High-temperature · Mobility · Reliability · Switching ratio

1 Introduction

High-temperature analysis of a semiconductor device is a vital reliability issue for Si-based transistor technology. Higher temperatures may damage the device or distract the device performance when the device is in the sub-20 nm regime. For high-temperature applications, the device, material plays a vital role and decided the reliability [1] duration of the device. In nano-electronics, for a high-performance application of a transistor black phosphorus (BP) is frequently used owing to the layered structure and high mobility (~1000 cm^2V^{-1} s^{-1}) [2]. BP has a variable bandgap from 0.3 eV to 1.0–1.5 eV (bulk to a single layer) [3]. It has many more excellent electrical properties and frequently used in nanodevices as compared to other materials [4–6]. Therefore, in conventional recessed channel MOSFET, the gate material is replaced by BP to augment the performance of the device [7].

When the gate length shrinks at sub-20 nm of a MOSFET, short channel effects (SCEs) [8, 9] are created and hinders the performance of the device. Due to the SCEs, such as hot carrier effects (HCEs), gate tunneling current, threshold voltage roll-off, punch-through [10] and drain induced barrier lowering (DIBL) [11–13], problems are created. Groove (recessed) channel device [14–17] is one of the promising devices to overcome these SCEs. A recessed channel device [18, 19] has been introduced in which a metal gate is used. Recessed channel device also has some drawback such as corner effects thus; to diminish these effects BP is amalgamated in the recessed gate as BP is the most suitable material in semiconductor devices. In this work, the proposed BP-T device has been used in harsh temperature environment and the electrical characteristics of a device are not altered and give high performance. The simulation has been performed using

✉ Neha Gupta
gupta.neha@niecdelhi.ac.in

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Ajay Kumar
ajaykumar@dtu.ac.in

[1] ECE Department, Jaypee Institute of Information Technology, Noida, India

[2] Electrical Engineering Department, Delhi Technological University, Delhi, India

[3] Department of Applied Science and Humanities, ADGITM, New Delhi, India

[4] Applied Physics Department, Delhi Technological University, Delhi, India

Published online: 01 June 2020

TCAD device simulator. The temperature reliability has been explored in terms of reduced sub-threshold current, hot-electron injection gate current, and impact ionization substrate current along with higher I_{on}/I_{off} ratio, electron mobility, and electron velocity; owing to the electrical properties of black phosphorus. Results also reflect that for future nanotechnology, black phosphorus may use as a promising semiconductor device material [20]. Owing to its high-temperature reliability, the proposed device can be also suitable for space applications.

2 Device Design Parameters

The schematic device design of BP-T-MOSFET is reflected in Fig. 1 and its parameters are given in Table 1. For high-temperature reliability assessment highly doped source and drain have been considered while the lightly doped substrate is considered. In the given architecture, since gate length (L_G) and channel length ($L_{CH} = L_G + 2(NJD + t_{ox})$) are different. Gate length (L_G) is 20 nm however channel length (L_{CH}) is equal to $L_G + 2(NJD + t_{ox})$ [$(20 + 2(10 + 2)) = 44$ nm] as clearly reflected in Fig. 1.

3 Simulation Methodology and Calibration

The proposed device in this paper is BP-T-MOSFET as shown in Fig. 1. For the performance analysis of the BP-T device, the ATLAS simulator has been used [21]. In this simulation mobility and recombination, models have been used with a gate voltage of 1.1 V and the constant drain voltage of 0.2 V. The simulation models have been

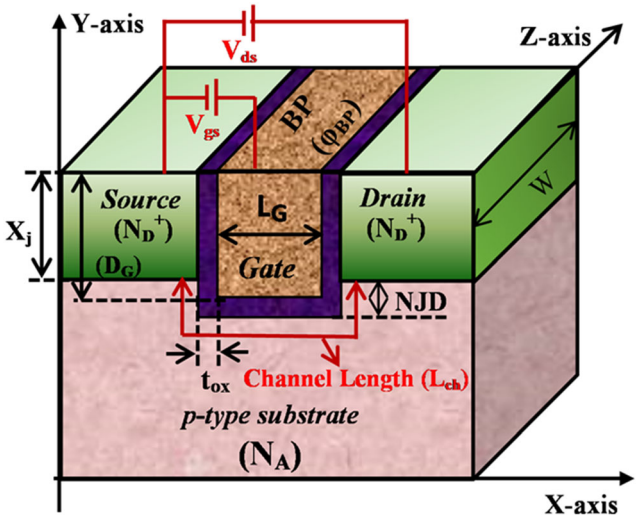


Fig. 1 Device design of Black Phosphorus Trench (BP-T) MOSFET

Table 1 Design parameters OF BP-T-MOSFET

Device parameters	BP-MOSFET
Gate Length (L_G)	20 nm
Channel Length (L_{ch})	44 nm
Source and Drain Region Length (X_i)	30 nm
Device Width (W)	200 nm
Groove Depth (D_G)	38 nm
Negative Junction Depth (NJD)	10 nm
Substrate Doping (N_A)	5×1017 cm^{-3}
Source/Drain Doping (N_D^+)	5×10^{19} cm^{-3}
Physical Oxide Thickness (t_{ox})	2 nm
Work Function for TGRC-MOSFET ()	5.16 eV
Gate to Source Voltage (V_{gs})	1.1 V
Drain to Source Voltage (V_{ds})	0.2 V

utilized by following equations [21] which are listed below-.

1. Poisson's equation

$$div(\varepsilon \nabla \varphi) = -\rho \tag{1}$$

Where is ε represents permittivity, ρ is space charge density and φ is potential.

2. Current continuity equation

For Holes and electrons are-

$$\frac{\partial p}{\partial t} = \frac{1}{q} div \vec{J}_p + G_p - R_p \tag{2}$$

$$\frac{\partial n}{\partial t} = \frac{1}{q} div \vec{J}_n + G_n - R_n \tag{3}$$

Where, holes and electrons current densities are represented by $\vec{J}_p$ and $\vec{J}_n$ are; however, G_p and G_n indicated the generation rate for holes and electrons; recombination rate for holes and electrons are R_p and R_n respectively.

3. Recombination

SRH (Shockley Read Hall)

$$R_{SRH} = \frac{pn - n_{ie}^2}{\tau_p[n + n_{ie}e^{(-(E_{TRAP}/kT_L))}] + \tau_n[p + n_{ie}e^{(-(E_{TRAP}/kT_L))}]} \tag{4}$$

Where, change in the intrinsic Fermi level and the trap energy level represented by E_{TRAP}; τ_n and τ_p, are the electron and hole lifetimes.

Auger Recombination

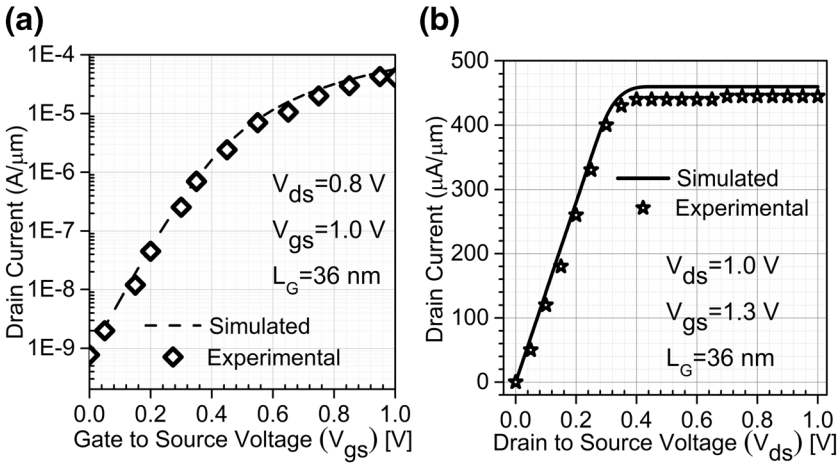


Fig. 2 Model calibration with experimental data for recessed channel MOSFET with **a** Transfer and **b** Output characteristics

$$R_{Auger} = AUGN\left(pn^2 - nn_{ie}^2\right) + AUGP\left[np^2 - pn_{ie}^2\right] \quad (5)$$

AUGP and AUGN are user-defined where AUGP = 1.8×10^{-31} cm^6 /s and AUGN = 8.3×10^{-32} cm^6/s.

4. Parallel Electric-Field-Dependent Mobility

$$\mu(\mathrm{E}) = \mu_0 \left[\frac{1}{1 + (\mu_0 E/\upsilon_{sat})^{\beta}}\right]^{1/\beta} \quad (6)$$

For the validation of simulation models, the experimental data [22] of a fabricate trench MOSFET (36 nm gate length) taken and plotted for the comparison with simulation data in terms of transfer characteristics and output characteristics as shown in Fig. 2a, b respectively. Figure 2 shows that the experimental results and simulated results are well-calibrated, thus showing the validity of simulation models. Furthermore, the fabrication feasibility of BP-T-MOSFET has been shown by the process flow chart as shown in Fig. 3. BP devices are significantly affected by the ambient conditions because BP reacts easily with water and oxygen molecules present in the

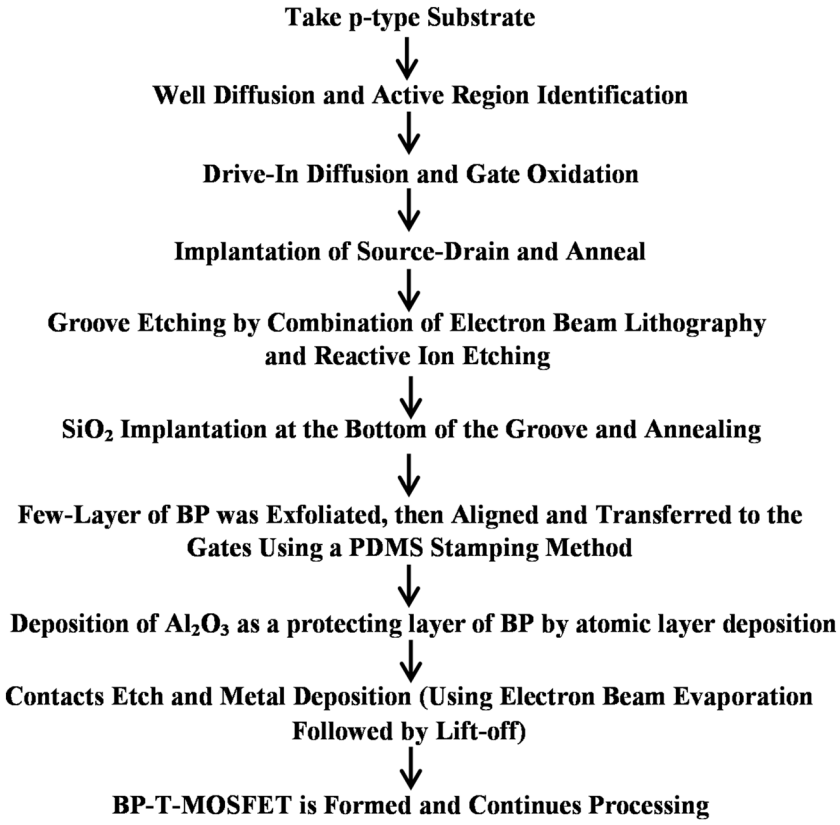


Fig. 3 Fabrication flowchart of BP-T-MOSFET

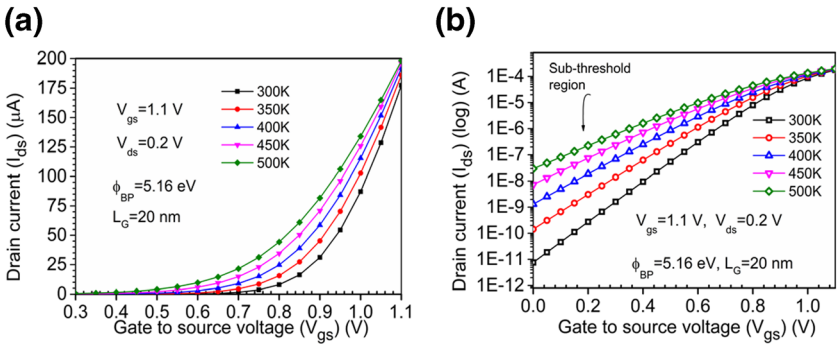


Fig. 4 **a** Drain current and **b** Sub-threshold current; for BP-T structure at various temperatures (300 K–500 K)

environment. Therefore to improve the device feasibility, Al_2O_3 is deposited by atomic layer deposition to protect BP from ambient degradation [23].

4 Results and Discussions

Figure 4a, b reflects the transfer characteristics of the device, and the drain current slightly reduces when the temperature increases from 300 K to 500 K in the BP-T device (as reflected in Fig. 4a) due to amalgamation of BP on to the trench region which shows the reliability of the proposed device at higher temperatures. Figure 4b reflects the device behavior in the sub-threshold region at high temperatures. Below threshold voltage, the off-current (leakage current) increases with the temperature due to the negative temperature coefficient (NTC) [24] behavior of semiconductors. When the temperature increases from 300 K to 500 K, resistivity reduces and conductivity increases due to the NTC nature of semiconductor, and hence on-resistance reduces below threshold voltage as shown in Fig. 5.

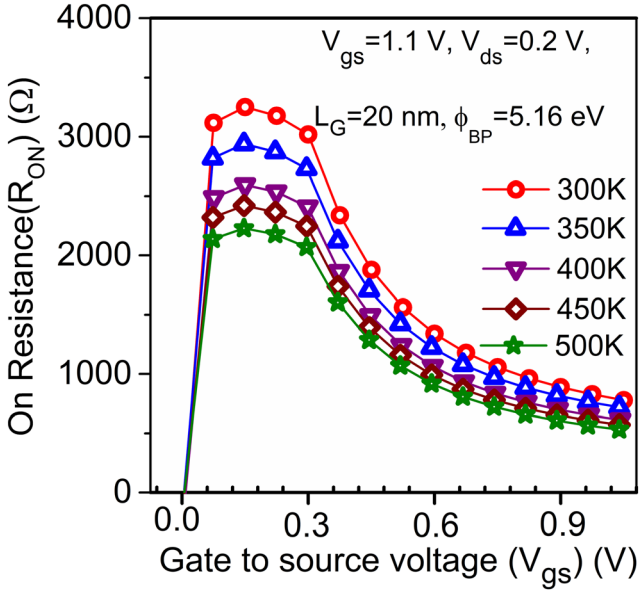


Fig. 5 On-resistance (R_{ON}); for BP-T structure at various temperatures (300 K–500 K)

Further, it is precisely detected that a significant the difference in I_{off} (10^{-12} A and 10^{-8} A), while I_{on} (7.09×10^{-4} A and 5.13×10^{-4} A) are nearly equal between 300 K to 500 K and due to this, the I_{on}/I_{off} ratio (switching ratio) of BP-T device, is higher at 300 K as evident in Fig. 6a. The switching capability and gate controllability are achieved in BP-T structure, at the drain bias of 0.2 V, and gate bias of 1.1 V at high temperature by which BP-T structure emerged as an ultra-low-power device for high-temperature applications. Moreover, the sub-threshold slope is very low (67 mV/dec) at room temperature and it goes to 78 mV/dec even at the temperature is very high (500 K) as reflected in Fig. 6b.

Further, the surface potential φ_f has been evaluated for the proposed device by the Eq. (7):

$$\phi_f = \frac{kT}{q} \ln\left(\frac{n}{n_i}\right) \tag{7}$$

The deviation of n_i (the intrinsic carrier concentration) with temperature is given by Eq. (8):

$$n_i = \left(\frac{N_c}{N_v}\right)^{1/2} \exp\left(\frac{-E_g}{2kT}\right) \tag{8}$$

Figure 7a, b show the hot electron injection gate current (HEIGC) [25] and impact ionization substrate current (IISC) [26] respectively. In a MOSFET, due to an increase in temperature, the charge carriers gain sufficient kinetic energy to surmount the potential barrier necessary to break an interface state. Some of these charge carriers surmount the Si-SiO_2 interface barrier and some of them move towards substrate which causes HEIGC and IISC respectively. From Fig. 7a, it is observed that the HEIGC is very less ($\sim 10^{-22}$) at room temperature and also has less value ($\sim 10^{-18}$) for higher temperatures (500 K) as well. Similarly, IISC is also had very little ($\sim 10^{-16}$) at room temperature and also has less value ($\sim 10^{-14}$) for higher temperature (500 K) as shown in Fig. 7b. Thus, the BP-T device shows the thermal reliability even the temperature is very high.

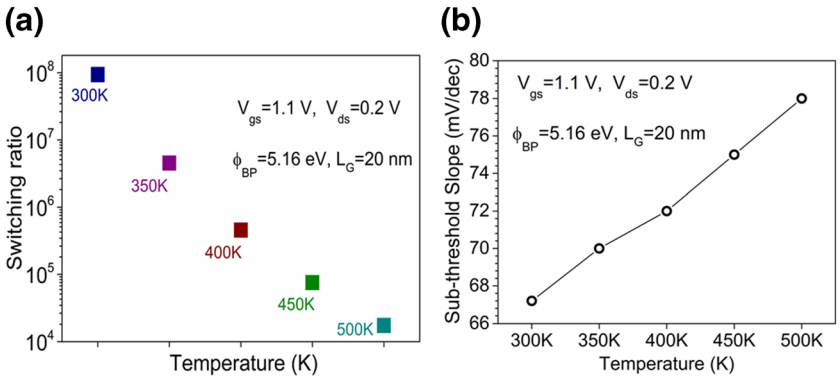


Fig. 6 **a** Switching ratio (I_{on}/I_{off}) and **b** Sub-threshold slope; of BP-T structure at various temperatures (300 K–500 K)

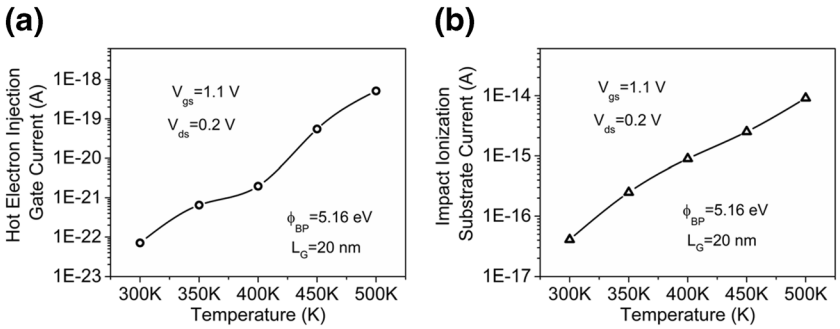


Fig. 7 **a** Hot-electron injection gate current and **b** Impact ionization substrate current of BP-T structure at various temperatures (300 K–500 K)

Further, Fig. 8 indicates the surface potential along the channel for BP-T structure at a very high temperature, and it observes that the potential is higher at 300 K and slightly reduces at a 500 K, which reflects that the proposed device gives better performance even the temperature is very high. Further, device efficiency [14] of the BP-T device has been evaluated at the various temperature, as shown in Fig. 9a. This result shows that BP-T-MOSFET is much more efficient at room temperature due to very high transconductance. However, the device efficiency has not changed significantly

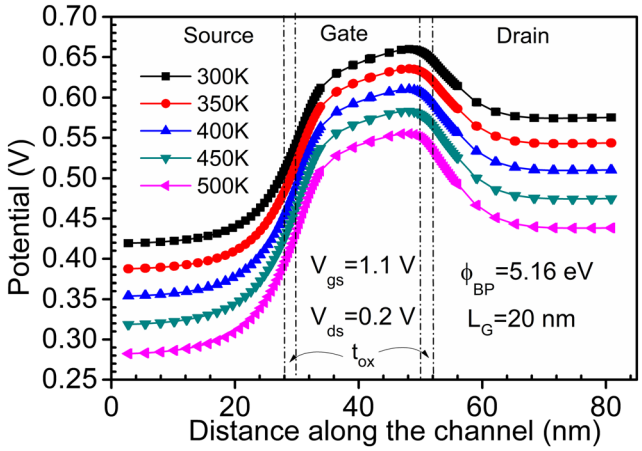


Fig. 8 The surface potential of BP-T structure at various temperatures (300 K–500 K)

at high temperatures (at the high gate to source voltage after the threshold voltage), which indicated that the device is more reliable even if the temperature is very high. Furthermore, the change in threshold voltage of the device has been observed with temperature as shown in Fig. 9b. With the increase in temperature, threshold voltage reduces due to the creation of more charge carriers at high temperature, resulting in the reduction of electron mobility (as shown in Fig. 10b). As we know that by increasing the temperature, Fermi potential decreases [27], hence the threshold voltage reduces with temperature. Therefore, the change in threshold voltage caused the change in drain current [28] as shown in Fig. 4a, b). It is also observed that the electric field near the drain corner reduces significantly (as shown in Fig. 10a) to mitigate the corner effects owing to the amalgamation of black phosphorus as the gate material. It is precisely evident from Fig. 10b that the enhancement in electron mobility (1000 cm^2V^{-1} s^{-1}) is found at a lower temperature as electron mobility is inversely proportional to temperature. In BP-T structure, electrons have very less amount of resistance to travel from drain to source; hence the mobility is higher at the source side, as shown in Fig. 10b.

Higher electron mobility also triggered the switching ratio (reflected in Fig. 4). The impact of higher temperature is less prominent in the channel and electron mobility has less variation with the temperature. However, electron mobility is reduced for higher temperatures as the

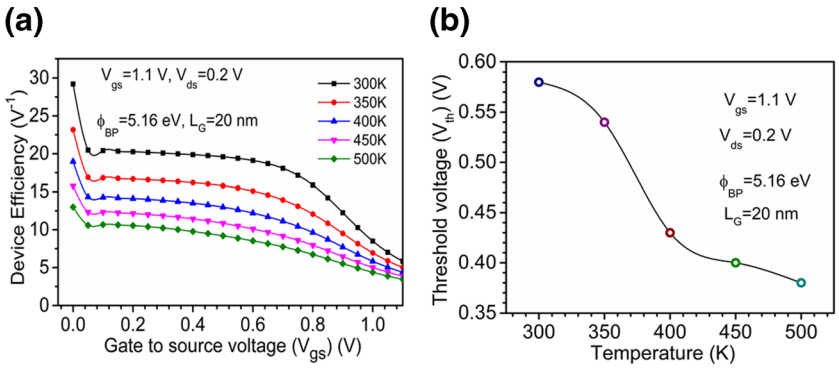


Fig. 9 **a** Device efficiency and **b** Threshold voltage of BP-T structure along the channel at various temperatures (300 K–500 K)

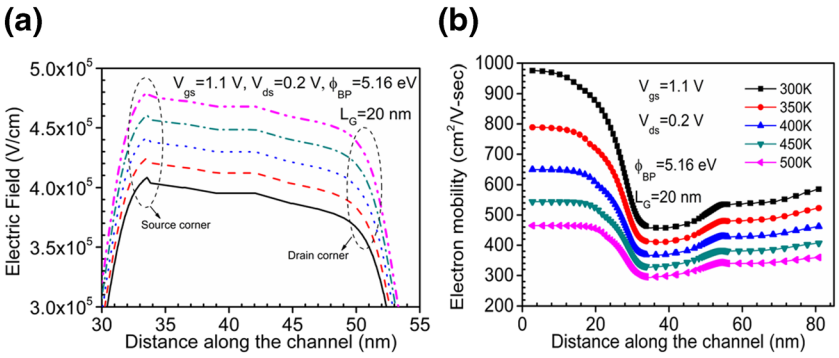


Fig. 10 **a** The electric field at the source and drain corners, and **b** Electron mobility of BP-T structure at various temperatures (300 K–500 K)

mobility has an inverse relation with temperature but directly proportional to the electron velocity hence, electron mobility is less and electron velocity is high for very high temperatures (as shown in Fig. 11a) while the overall impact of high temperature is very less as reflected in the results.

Further, the high-temperature analysis has been observed in terms of electron velocity and evaluated at a high temperature (300 K–500 K). It is observed that electron velocity (reflected in Fig. 11a) is slightly reduced when temperature enhances from 300 K to 500 K. Figure 11b demonstrates the impact of the high-temperature environment on electron temperature along the channel. From the contour profile (Fig. 11b), it is evident that in the BP-T device, the electron temperature variation is very less in the channel region and also when the electron travels from source to drain. As the device needs low electron temperature to eliminate (minimize) the leakage currents for

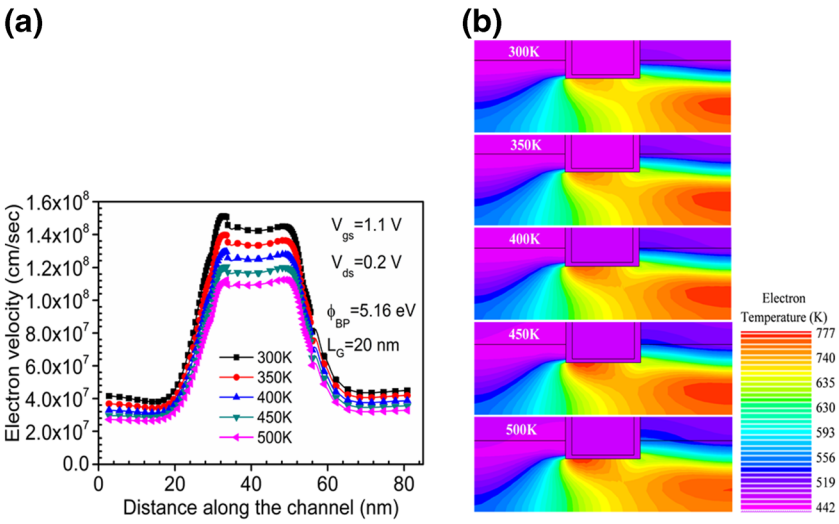


Fig. 11 **a** Electron velocity and **b** Contour plots of electron temperature of BP-T structure along the channel at various temperatures (300 K–500 K)

making it reliable for switching and ultra-low power applications. Thus, this work examines the reliability of BP-T structure at high temperatures. The results are obtained in terms of the electrical performance of the device and show that the proposed device structure is highly reliable for the very high-temperature applications.

5 Conclusion

This work examines the high temperature (300 K–500 K) reliability of BP-T-MOSFET. The electrical parameters such as switching ratio (I_{on}/I_{off}) (off-currents 10^{-12} A while on-currents 7.09×10^{-4} A), electron mobility (1000 $cm^2V^{-1}\ s^{-1}$), electron velocity (1.5×10^8 cm/s), hot electron injected gate current, impact ionization substrate current and transconductance (g_m) reflects that the proposed device gives good performance even in the very high temperature (500 K) at very low drain bias (0.2 V). Also, the device efficiency has not been altered in a harsh temperature environment as the device efficiency is high for BP-T-MOSFET. Thus, the analytical results show that the proposed device (BP-T-MOSFET) is highly reliable in the very high-temperature environment and space applications.

Acknowledgments The authors are thankful to DTU and JIIT for supporting this work.

References

1. ElKashlan RY, Samy O, Anis A, Ismail Y, Abdelhamid H (2020) Unified quantum and reliability model for ultra-thin double-gate MOSFETs. Silicon 12(1):21–28
2. Lam K-T, Dong Z, Guo J (2014) Performance limits projection of black phosphorous field-effect transistors. IEEE Electron Device Lett 35(9):963–965
3. Das S, Zhang W, Demarteau M, Hoffmann A, Dubey M, Roelofs A (2014) Tunable transport gap in phosphorene. Nano Lett 14(10): 5733–5739
4. Wang Y, Slassi A, Cornil J, Beljonne D, Samorì PJS (2019) Black phosphorus: tuning the optical and electrical properties of few-layer black phosphorus via physisorption of small solvent molecules (Small 47/2019). Small 15(47):1970252
5. Jia C, Zhao L, Cui M, Yang F, Cheng G, Yang G, Zeng Z (2019) Surface coordination modification and electrical properties of few-layer black phosphorus exfoliated by the liquid-phase method. J Alloys Compd 799:99–107
6. Xu Y, Shi X, Zhang Y, Zhang H, Zhang Q, Huang Z, Xu X, Guo J, Zhang H, Sun L (2020) Epitaxial nucleation and lateral growth of high-crystalline black phosphorus films on silicon. Nat Commun 11(1):1–8
7. Kumar A, Tripathi M, Chaujar R (2018) Comprehensive analysis of sub-20 nm black phosphorus based junctionless-recessed channel MOSFET for analog/RF applications. Superlattice Microst 116: 171–180
8. D'Agostino F, Quercia D (2000) Short-channel effects in MOSFETs. Introduction to VLSI design (EECS 467)
9. Bulucea C, Wang F-C, Chaparala P (2003) Field-effect transistor for alleviating short-channel effects. Google Patents
10. Nishizawa J-i, Ohmi T (1995) Short channel MOSFET with buried anti-punch through region. Google Patents
11. Kumar A, Gupta N, Chaujar R (2016) Analysis of novel transparent gate recessed channel (TGRC) MOSFET for improved analog behaviour. Microsyst Technol 22(11):2665–2671
12. Wang CH, Zhang P-F (1999) Three-dimensional DIBL for shallow-trench isolated MOSFET's. IEEE Trans Electron Devices 46(1):139–144
13. Sahay S, Kumar MJ (2017) Spacer design guidelines for nanowire FETs from gate-induced drain leakage perspective. IEEE Trans Electron Devices 64(7):3007–3015
14. Kumar A (2017) Effect of trench depth and gate length shrinking assessment on the analog and linearity performance of TGRC-MOSFET. Superlattice Microst 109:626–640
15. Kumar A, Gupta N, Chaujar R (2016) TCAD RF performance investigation of transparent gate recessed channel MOSFET. Microelectron J 49:36–42
16. Kumar A, Tripathi M, Chaujar R (2018) Reliability issues of In2O5Sn gate electrode recessed channel MOSFET: impact of interface trap charges and temperature. IEEE Trans Electron Devices 65:860–866
17. Kumar A, Tiwari B, Singh S, Tripathi MM, Chaujar R (2018) Radiation analysis of N-channel TGRC-MOSFET: an X-ray dosimeter. IEEE Trans Electron Devices 65(11):5014–5020
18. Chaujar R, Kaur R, Saxena M, Gupta M, Gupta R (2009) TCAD assessment of gate electrode workfunction engineered recessed channel (GEWE-RC) MOSFET and its multi-layered gate architecture, part II: analog and large signal performance evaluation. Superlattice Microst 46(4):645–655
19. Kumar A, Gupta N, Chaujar R (2016) Power gain assessment of ITO based transparent gate recessed channel (TGRC) MOSFET for RF/wireless applications. Superlattic Microst 91:290–301
20. Schwierz F (2010) Graphene transistors. Nat Nanotechnol 5(7): 487–496
21. Silvaco I (2011) ATLAS user's manual. Santa Clara, CA, Ver, vol 5
22. Appenzeller J, Martel R, Avouris P, Knoch J, Scholvin J, del Alamo JA, Rice P, Solomon P (2002) Sub-40 nm SOI V-groove n-MOSFETs. IEEE Electron Device Lett 23(2):100–102
23. Wei Q, He J, Yang S, Jia H, Liu Y, Liu W, Liu Y, Li T (2016) Investigation of black phosphorus field-effect transistors and its stability. Opt Quant Electron 48(6):344
24. Feteira A (2009) Negative temperature coefficient resistance (NTCR) ceramic thermistors: an industrial perspective. J Am Ceram Soc 92(5):967–983
25. Tam S, Ko P-K, Hu C (1984) Lucky-electron model of channel hot-electron injection in MOSFET's. IEEE Trans Electron Dev 31(9): 1116–1125
26. Kamata T, Tanabashi K, Kobayashi K (1976) Substrate current due to impact ionization in MOS-FET. Jpn J Appl Phys 15(6):1127
27. Elewa T, Balestra F, Cristoloveanu S, Hafez IM, Colinge J-P, Auberton-Herve A-J, Davis JR (1990) Performance and physical mechanisms in SIMOX MOS transistors operated at very low temperature. IEEE Trans Electron Dev 37(4):1007–1019
28. Hasanuzzaman M, Islam SK, Tolbert LM (2004) Effects of temperature variation (300–600 K) in MOSFET modeling in 6H–silicon carbide. Solid State Electron 48(1):125–132

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



Review Paper

Silicon–air batteries: progress, applications and challenges

Rishabh Bansal[1] · Prajwal Menon[1] · R. C. Sharma[1]

Received: 1 April 2020 / Accepted: 15 May 2020
© Springer Nature Switzerland AG 2020

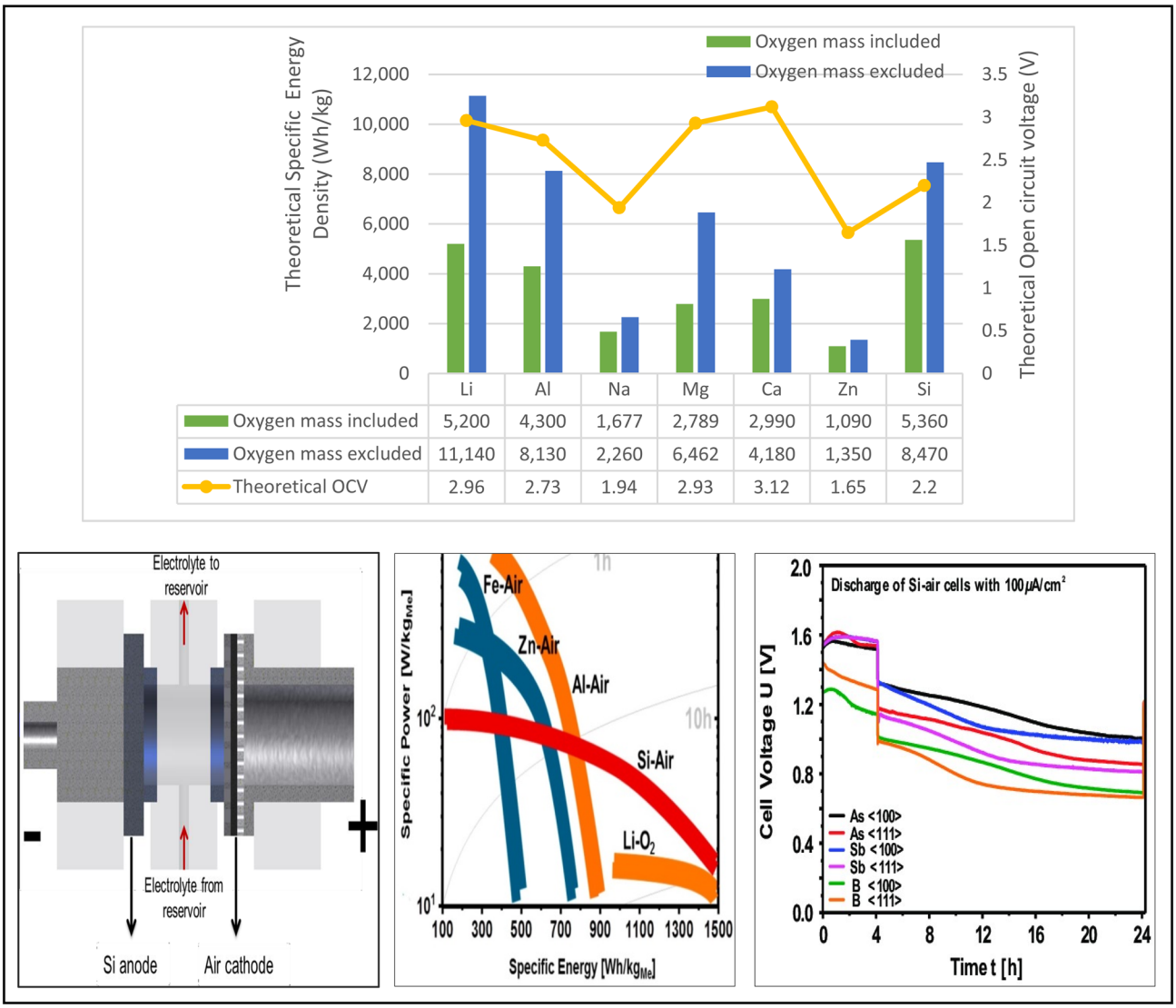
Abstract
Silicon–air battery is an emerging energy storage device which possesses high theoretical energy density (8470 Wh kg^{-1}). Silicon is the second most abundant material on earth. Besides, the discharge products of silicon–air battery are non-toxic and environment-friendly. Pure silicon, nano-engineered silicon and doped silicon have been found potential candidate for anode. Meso-porous meso-cellular carbon having optimal pore size and imbibed with α-MnO_2 nanowires catalyst is found the most promising cathode. Several technical, design and corrosion problems associated with Si–air battery system have to be resolved for its mass scale deployment. This review presents comprehensive information on Si–air battery technology.

✉ R. C. Sharma, rcsharma@dce.ac.in | [1]Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India.

SN Applied Sciences (2020) 2:1141 | https://doi.org/10.1007/s42452-020-2925-7

Published online: 29 May 2020

Graphic abstract



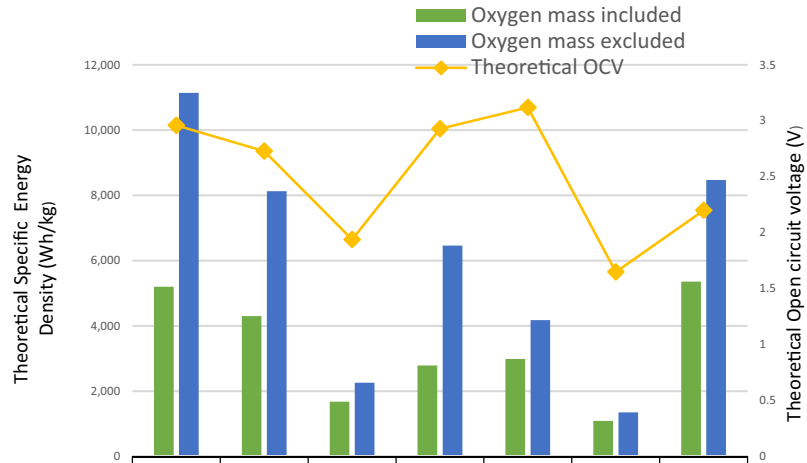
Keywords Silicon · Si–air battery · Ionic liquids · Gel polymer electrolyte · Oxygen reduction catalyst · Heteroatom doped carbons

1 Introduction

Isaac Newton and Michael Faraday have given us the intellectual tools to wheel the battery of innovations for clean energy and efficient techno-development of global community. Extensive usage of energy coupled with global climate change has necessitated the growth of renewable energy production and storage systems. World-wide interest in electric vehicles (EVs) has focused on miniaturization of light weight rechargeable batteries. Li-ion batteries with high energy efficiency and more cyclic capabilities has become the most promising candidate for EV applications. However, its low storage capacity (100–200 Wh kg^{-1}), safety issues and high cost has led to extensive research activities to find alternative energy storage technology. Metal–air battery systems such as Li–air, Mg–air, Zn–air, Al–air and Si–air are highly promising for powering automobiles, industrial equipment, computers, electronic device, hearing aids and a large number of utility items [1–8]. Apart from possessing high energy density (400–1700 Wh kg^{-1}), metal–air batteries are compact, low cost, lighter and environment friendly [9–15]. The reason

for high energy density possessed by metal–air batteries is its cathode which operates on oxygen present in the air during discharge, replacing costly chemicals used in present Li-ion batteries. Early work on metal–air batteries encountered the problems associated with the material at the air cathode surface, thermal management system along with various technical problems associated with anodes. Theoretical specific energy densities and open circuit voltages of metal–air battery systems are shown in Fig. 1.

Lithium metal is considered to be the best candidate as anode material because of its high theoretical specific energy (3860 Ah kg^{-1}), low density (0.59 g cm^{-3}) and the lowest electrode potential (3.045 V vs. SCE). Si–air battery has comparable specific energy density (5360 Wh kg^{-1}) with Li–air battery (5200 Wh kg^{-1}). Also, silicon (28.2%) is more abundant in earth's crust when compared with lithium (0.002%) and is also less expensive than lithium metal. The practical performances of some battery systems including metal air batteries have been represented in the form of Ragone plots [16] (Fig. 2). It may be noted from the Ragone plots that metal–air batteries deliver higher specific energy while Li–ion battery provides the best power capability. Also, Li–air and Si–air batteries provide specific energies of the same order but there is large difference in their specific power capabilities. Moreover, rechargeability of Si–air batteries is a challenging task.



	Li	Al	Na	Mg	Ca	Zn	Si
Oxygen mass included	5,200	4,300	1,677	2,789	2,990	1,090	5,360
Oxygen mass excluded	11,140	8,130	2,260	6,462	4,180	1,350	8,470
Theoretical OCV	2.96	2.73	1.94	2.93	3.12	1.65	2.2

Fig. 1 Comparison of theoretical specific energy densities and open circuit voltages (OCVs) of different metal–air batteries [10]

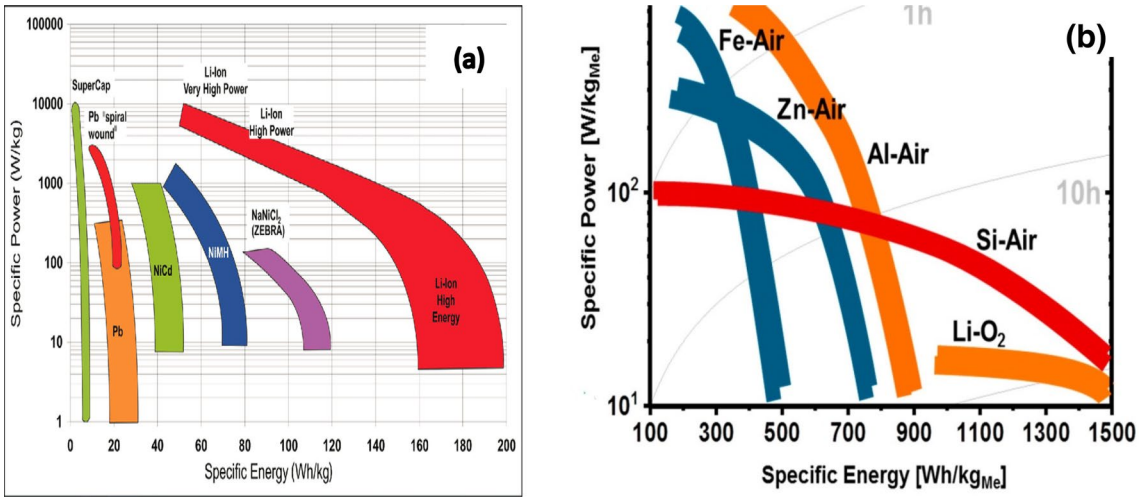


Fig. 2 **a** Ragone plot [16] for the important batteries systems; **b** the comparison of different metal air battery systems. The typical logarithmic axes of Ragone plot **a** is changed to logarithmic y and linear x in **b** in order to represent the differences between the metal–air batteries

Si–air battery research was initiated by Prof. Yair-Ein-Eli and team at the Technion–Israel Institute of Technology in 2009 [17]. This novel type of primary battery had an ideal specific energy of $E = 8470$ Wh kg^{-1} and a long-term stable voltage of 1.0–1.2 V. It consisted of doped silicon wafers as fuel, an air cathode, and a specially synthesized ionic liquid (EMI·(HF)$_2$·3F) as electrolyte.

2 Silicon–air cell construction

Yair-Ein-Eli et al. [17] used a silicon single-crystal anode, air cathode and specially synthesized room temperature ionic liquid (EMI·(HF)$_2$·3F) as electrolyte. Graphic representation of cell assembly is shown in Fig. 3. The cell is made up of three plastic plates. The first plate is made of multiple pieces of silicon wafers of size 1 cm x1cm which were pressed into a viton O-ring. The exposed area of silicon wafers is 0.5 cm^2. The second sheet consists of circular sheets of graphitic air electrodes having 0.5 cm^2 exposed area. The third sheet is employed to make an outer circle periphery which is used to make electrical contact. High grade copper wires are used for making terminal contacts. Si wafer anodes were pre-treated before cell construction; the treatment involved immersion of Si-wafers in HF solution (1HF:5H$_2$O) for 10 s to remove surface oxide layer followed by cleaning with de-ionized (DI) water and finally drying in a nitrogen stream. The anode-separator-cathode is sandwiched and 0.5 ml of the ionic liquid was added to complete the cell assembly. For adequate wetting of all the components, the cell assembly is maintained at OCV for 4 h during which no current was drawn from the cell assembly. The cell was later used to draw current to initiate the discharge process.

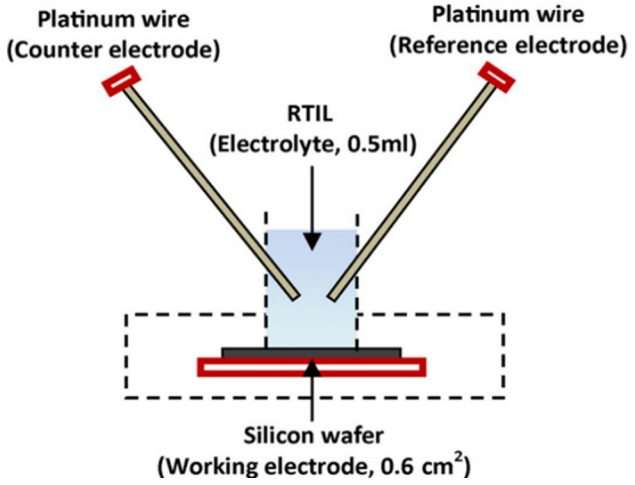


Fig. 3 Silicon–air cell assembly [17]

3 Anode for Si–air battery

3.1 Silicon wafers

Doped silicon wafers are widely used as anode material for Si–air batteries. Silicon <100> crystals, both moderately doped (n-type) and profoundly doped (n^{++}-type), are held with a screwed back stainless steel contact plate electroplated with thin layer of gold [18]. The corrosion rates for various types of silicon material as anode is shown in Fig. 4.

Silicon corrosion rates and polarization voltammograms form a basis for anode selection. On one hand, p^{++} silicon has lower corrosion rates and better cell voltage at high currents whereas on the other hand, n^{++} silicon has higher OCV and better cell voltages at low currents. Even though n^{++}doped Si exhibited greater corrosion rates, it has been a subject of interest due to its greater cell voltage at low values of current.

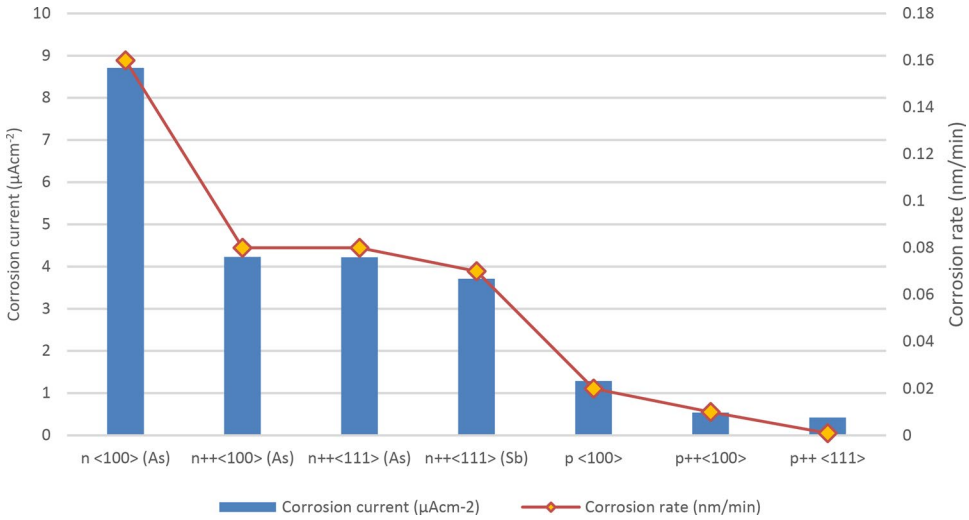


Fig. 4 Silicon electrode corrosion rates in EMI.(HF)$_2$.3F [19]

3.2 Nanostructured silicon

Nanostructured silicon/arrays of silicon nanowires or nanorods have been used as anode material in lithium-ion batteries to overcome cracking and pulverization [20, 21]. A robust electrical contact is maintained when silicon nanowires are used as anode material since each nanowire is directly connected to the current collecting substrate. Mingyuan Ge et al. [22] reported that porous silicon nanowires synthesized by direct etching of boron-doped silicon wafers possesses high porosity and electron conductivity. As anode material, these nanowires exhibit high structure stability, higher electrochemical performance and long cycle life. Even after 250 cycles, the capacity remains stable above 2000, 1600, and 1100 mAh g^{-1} at current rates of 2, 4, and 18 A g^{-1} respectively. Zhong et al. [23] used nanostructured silicon as the anode material in the design of Si–air battery with alkaline electrolyte. Nano-engineered silicon anodes have been found to increase the reversible charge capacity and longer cycle life for new generation batteries. The surface of the silicon wafer is reformed by metal-assisted electro less etching method to facilitate formation of micro-porous layer on the silicon wafers [24–27]. The native oxide layer is removed by immersing the clean silicon pieces into a buffered oxide etching (BOE) agent.

3.3 Doped silicon

Silicon doped with Arsenic(As), Antimony(Sb) and Boron(B) with different orientations, i.e. <100> and <111>, was used as anode by Durmus et al. [28, 29]. The following six types of silicon anodes were used.

i. As-doped [Crystal family type <100>] (Resistivity: 0.001–0.007 Ω cm)
ii. As-doped [Crystal family type <111>] (Resistivity: 0.001–0.010 Ω cm)
iii. Sb-doped [Crystal family type <100>] (Resistivity: 0.007–0.020 Ω cm)
iv. Sb-doped [Crystal family type <111>] (Resistivity: 0.022–0.028 Ω cm)
v. B-doped [Crystal family type <100>] (Resistivity: 0.001–0.005 Ω cm)
vi. B-doped [Crystal family type <111>] (Resistivity: 0.002–0.016 Ω cm)

Two stage surface treatment of the doped wafer is carried out with argon/oxygen plasma to remove organic contaminants. The native oxide layers are removed from the wafer surfaces with the help of argon/sulfur hexafluoride plasma. The cell assembly is fabricated and kept for 4 h in a climate chamber which is maintained at 25 °C and 50% relative humidity. During this period no current is drawn and the silicon–air cells are held at their open-circuit voltage (OCV) before carrying out discharge process.

The 3D images[1] (Fig. 5) revealed that the surface characteristics of <100> vs. <111> oriented Si anodes vary significantly at low discharge currents. With both, arsenic-doped and antimony-doped Si anodes having orientation <100>, homogeneous 3D images are observed while with <111> oriented Si anode, 3D images consists of polygons. When cells are discharged at 0.1 mA cm^{-2} for 20 h, it is observed that the surface characteristics for corresponding orientation at mesoscale are almost similar for As- and Sb-doped Si anodes. 3D images of B-doped (p-type) and B-doped (n-type) Si anodes, under similar conditions of discharge, exhibited large differences in their surface characteristics. B-doped Si anodes with <100> orientation presented a homogeneous surface with a few individual small pores, while <111> oriented anodes showed a grain like structure.

Thus, the surface characteristics of B-doped Si anodes differ significantly from As- and Sb- doped Si anodes, indicating possibly two types of discharge mechanism at the anode surfaces. The studies on different anodes for Si–air batteries recommend that n^{++} doped Si anode may be a better choice for higher cell voltage and low current applications while nano-engineered Si anode can be used in applications where higher charge capacity and longer cycle life is needed predominantly.

4 Cathode for Si–air battery

4.1 Air cathode

Air cathode composition and structure play an important role in Si–air battery performance. Porous carbon imbibed with a catalyst, and a polymer binder is used as the air cathode. The specific capacity of an air electrode depends on several factors. Cheng et al. [30] and Xiao et al. [31] studied the effect of surface areas of different graphitic materials on specific capacity of the battery and it was found to be proportional to the surface area. Yang et al. [32] investigated the effect of pore size on specific capacity of Si–air cell and reported that air cathode with meso-porous meso-cellular carbon foam have higher specific capacity. Instead of too large or too small, optimum pore size is recommended. The thickness of the air cathode also

[1] High resolution 3D images of silicon wafers may be taken using a confocal laser scanning microscope (OLS4100, Olympus Corp., Japan), having magnification (Optical and Digital) in the ranges from 108× to 17,280× which can resolve up to 10 nm in size in the z direction and 120 nm in the x–y plane.

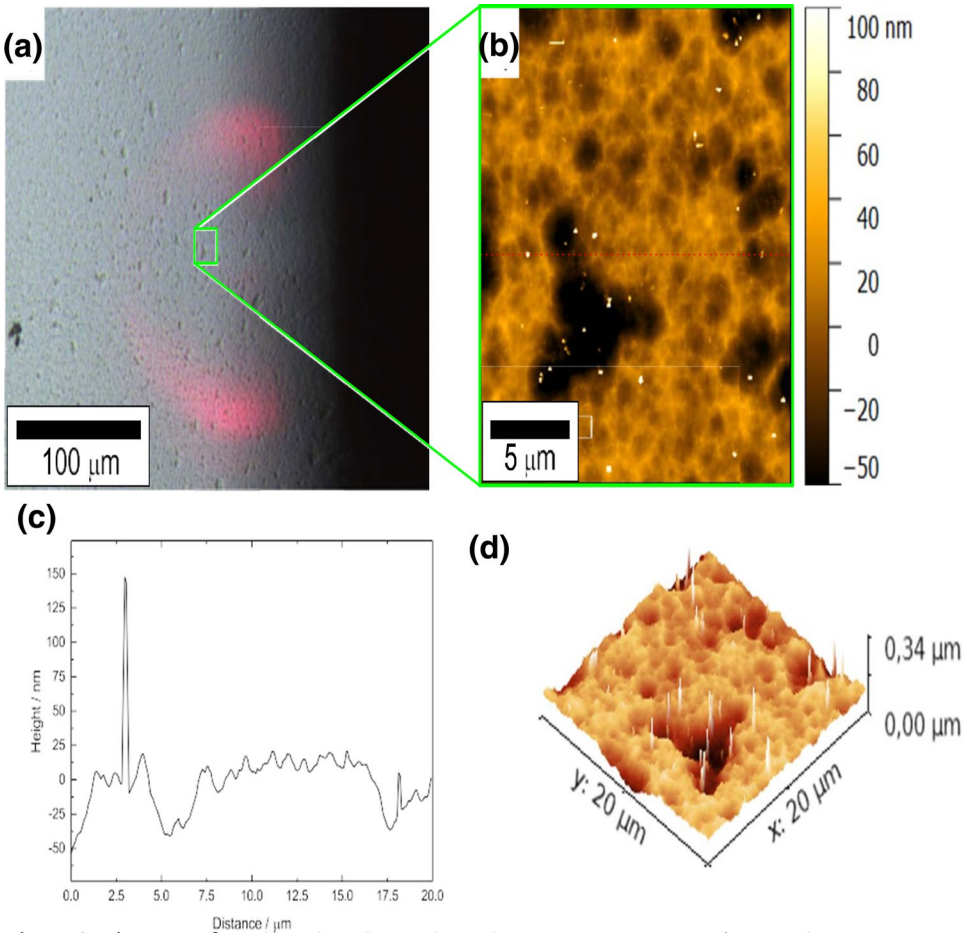


Fig. 5 3D images of As-doped Si electrode with <100> orientation. **a** optical microscopy image from the surface; **b** green square in **a** as imaged by AFM; **c** a line scan of green square position, as indicated in **b**; **d** 3D reconstruction of **b**

plays an important role. Zhang et al. [33] found that the cells with thicker electrodes have lower specific capacity, which is endorsed to slow oxygen diffusion through thicker air cathodes.

Beattie et al. [34] reported that specific capacity of the cells is inversely proportional to the amount of carbon loaded in the cathode. Another factor that affects the electrochemical performance of an air electrode is oxygen diffusion through the air electrode, which can be facilitated by increasing oxygen pressure. Read et al. [35], Yang and Xia [36] and Tran et al. [37] reported that the specific capacity linearly increases with the oxygen pressure. The porosities, wettability and electrical contact of electrodes are determined by the ratio of carbon, catalyst, and polymer, which ultimately affects the electrochemical performance of Si–air cells.

4.2 Catalysts for air cathode

4.2.1 Noble metals

Noble metals such as platinum and Pt-based alloys have been used as oxygen reduction reaction electro catalysts for many decades. The nanoparticles of PtAu exhibited high stability and exemplary catalytic activity for carbon air cathodes of fuel cells and battery systems. Several studies [38–41] reported that ORR properties of Pt-based catalysts can be further improved by changing the size of the particles to nano scale and changing the surface characteristics of Pt-based nanostructures. When alloyed with other noble or transition metals, the performance of Pt based catalysts can be improved along with a reduction in cost [42–47].

4.2.2 Carbonaceous materials

Graphene and carbonaceous materials have been preferred as electro catalysts for air cathode of Si–air battery systems. These have found application as metal free catalyst or catalyst support because of their superior properties which include low cost, widespread availability, low electrical resistivity, very big surface areas and noble stability under severe conditions [48–64]. However, it has been reported that the electrochemical performance of these types of electro catalysts is highly dependent on their structure, morphology, pore size, and wall thickness.

Appropriate pre-treatment of the surface of carbonaceous materials improves its functionality and enhance the durability of ORR catalysts.

4.2.3 Transition-metal oxides

High spin transition-metal complexes having spinels and perovskites morphology represents another class of cathode electro catalysts, highly useful in Si–air battery system [65–67]. These have shown excellent stability in alkaline medium whereas the stability in acidic electrolyte is relatively poor. Oxides of titanium as co-catalysts have further enhanced the activity of noble metals and its alloys with non-precious metals. Chen et al. [68] examined titanium nitrides and reported that these are highly efficient cathode electro catalysts. Transition metal carbides act as promising supports to promote the stability and activity of Pt- based ORR electro catalysts [69, 70]. The variable valence of transition metals such as Mn, Co, Fe and Ni gives rise to multiple structures to the oxides [71]. Mn-based oxides have been extensively studied as electro catalysts for air-electrodes. Electrochemically active manganese oxides such as MnO_2, Mn_2O_3, Mn_3O_4, and Mn_5O_8 exhibits high catalytic activity [72–75].

4.2.4 Inorganic–organic composites

The integration of organic matrix with nano-sized metal constitutes inorganic–organic composites. It is another class of materials that are used as noble metal-free electro catalysts for oxygen electrochemistry [76–84]. The non-precious Ni/Fe metals easily form composite with polyoxometalate (POM) and o-Anisidine (oA) and have exhibited high catalytic activity and stability towards oxygen evolution reaction (OER). A competition between metal cation takes place to make complex with composite. Nanostructured $Ni_xFe_y(OH)_2$ powders, coated onto various substrate with the aid of polymeric binders, have shown high catalytic activity in alkaline electrolytes. Such composites can be prepared by pyrolysis of metal, nitrogen and carbon as these is not stable in strong acidic or alkali media. The catalytic activities of these composites depend on several parameters which include the metal, the precursors, the substrate and the synthesis method [85]. Nano engineered metal composites with heterocyclic conjugated polymers (polyaniline, polypyrrole and poly(3,4-ethylenedioxithiophene) have also been used as electro catalysts [78, 86, 87]. These composites have advantage over other types of catalysts as the incorporation of metal ions in matrices creates large number of active sites [86]. Further, the bridges between matrix and metal promote physical and electrical contact [87]. Inorganic–organic composites such as FeCo-EDA catalysts have shown outstanding ORR activity in alkaline electrolytes and exhibits nearly 3 times higher mass activity than commercial Pt/C catalyst [88].

4.2.5 Hetero-atom doped carbons

Heteroatom (N & S) doped carbon nano-materials [89–92] are the most promising ORR catalyst as it markedly reduces the cost and increase the efficiency of metal–air battery systems/fuel cells. Vertically aligned nitrogen-doped carbon nanotube (VA-NCNT) arrays can act as a metal-free electrode to catalyze a $4e^-$ ORR process with a three times higher electro catalytic activity and better long-term stability than commercially available Pt/C electrodes in an alkaline cell [89]. These carbon-based metal-free ORR catalysts are also free from the CO poisoning and methanol crossover effects. Choi et al. [93, 94] synthesized nitrogen and sulfur-doped carbons and studied their electro catalytic activity towards oxygen reduction reactions (ORR). The carbon catalyst synthesized from cysteine and dual-doped with nitrogen and sulfur, exhibited the highest onset potential (0.55 V vs. Ag/AgCl) and electrochemical activity in acidic media, – 0.2 mA (at 0.2 V vs. Ag/AgCl), which is about 43% of that of commercial Pt/C (40 wt%). From XPS studies they showed that in the carbon catalyst, sulfur was doped as sulfate or sulfonate. It was concluded that not only nitrogen doping but also sulfur doping of carbon improve its electro catalytic activity towards ORR. Zhang et al. [95, 96]. synthesized graphene-based B/N co-doped carbon nano sheets (G-CBP). It was argued that high surface area, unique 2D sheet nanostructure, and high heteroatom-doping contents (5.4% B and 5.3% N) of G-CBP results into an excellent electrochemical performance for the oxygen reduction reaction under alkaline conditions (0.1 M KOH).

4.3 Oxygen reduction reaction (ORR) in air electrode

Polytetrafluoroethylene (PTFE) with activated carbon black (0.45–0.5 g cm^{-2} loaded) is used as air cathode [17]. The mixture is encapsulated onto a nickel 200 mesh under high pressure and manganese dioxide is added as cathode catalyst. The reduction of oxygen occurring at the air electrode is said to proceed either as 4 electrons or 2 electrons as shown:

$$O_2 + 4e^- \leftrightharpoons 2O^{2-} \quad (1)$$

$$O_2 + 2e^- \leftrightharpoons O_2^{2-} \quad (2)$$

The scanning electron microscopy (SEM) micrographs of air cathode revealed that activated carbon on cathode is covered with sedimentation during the battery discharge.

Peter Jakes et al. [18] studied the role of manganese oxide (MnO_2), which acts as an oxygen- reduction reaction (ORR) catalyst on Si–air battery performance. The chemical state of Mn ions in the MnO_2 catalysts was examined by XP and EPR spectroscopy. It was inferred from the results that during the discharge process, MnO_2 also undergo changes during an electrochemical reaction when the electrolyte has fluoride ions, commonly present in the RTIL.

$$MnO_2 + 8(HF)_3F + 2e^- \leftrightharpoons MnF_2 + 10(HF)_2F + 2H_2O \quad (3)$$

The formation of non-ORR catalyst MnF_2 results in the loss of available catalytic sites and also lowers the ionic conductivity of the electrolyte. The change of MnO_2 to MnF_2 greatly affects the Si–air battery performance by limiting the discharge capacity and inhibiting the cell discharge of Si–air battery. The electrochemical changes at air cathode are quite complex and complete understanding will require extensive studies with different catalyst and electrode/electrolyte systems. However, studies on new catalyst and new ionic solvent may lead to a suitable system to better understand the phenomenon.

5 Electrolytes for Si–air battery

5.1 Room temperature ionic liquid (RTIL)

A room temperature ionic liquid (RTIL), EMI.$(HF)_2$·3F is the most common type of electrolyte used in a Si–air battery. The EMI.$(HF)_2$·3F is synthesised by a reaction of 1-ethyl-3-methylimidazolium chloride and hydrogen fluoride by Hagiwara et al. [97]. The electrolyte has low viscosity and high conductivity amongst all RTILs with 100 mS cm^{-1}, besides chemical stability in air and high tolerance at extreme relative humidity conditions [98, 99]. The schematic of first experimental Si–Air Battery that utilised RTIL as the electrolyte is shown in Fig. 6.

Potentiodynamic polarization studies of the RTIL-silicon wafer couples (anodic half-cell) and of the IL–air electrode (cathodic half-cell) is used to evaluate EMI.$(HF)_2$.3F as a viable candidate for electrolyte in a Si–air battery is shown in Fig. 7.

The half-cell reactions, crystallization reaction and overall reaction when two anions of RTIL contribute to the oxidation of silicon at anode and reduction at air cathode are:

$$\text{Anode: } Si + 12(HF)_2F^- \leftrightharpoons 8(HF)_3F^-SiF_4 + 4e^- \quad (4)$$

$$\text{Cathode: } O_2 + 12(HF)_3F^- + 4e^- \leftrightharpoons 16\,(HF)_2F^- + 2H_2 \quad (5)$$

$$\text{Crystallization: } SiF_4 + 2H_2O + 4(HF)_2F^- \leftrightharpoons SiO_2 + 4(HF)_3F^- \quad (6)$$

$$\text{Overall Cell Reaction: } Si + O_2 \leftrightharpoons SiO_2 \quad (7)$$

5.1.1 Effect of water in the RTIL

Cohn et al. [100] examined the effect of addition of water in the RTIL electrolyte. It has been concluded that discharge performance of Si–air battery system is greatly obstructed since during cell discharge, the SiO_2 precipitates on air cathode's porous carbon. The precipitation of SiO_2 layer at a certain point of cell discharge prevents diffusion of oxygen which results into a short supply of oxygen at the electrolyte–electrode interface of air cathode. This results into lowering in discharge capacities of cathode than expected theoretically. Since water plays an important role in the formation of SiO_2, it was argued that the formation of SiO_2 may shift towards separator or anode by adding water into the electrolyte. If reaction is shifted to take place at the bulk electrolyte rather than at the electrolyte–electrode interface, an alternative location for the SiO_2 formation can be introduced causing suffocation to the air–cathode. This is due to the loss of available catalytic sites on the porous surface of the air cathode for oxygen reduction.

The Si–air cells under study are discharged at a constant current density of 0.3 mAh cm^{-2} and the cut-off potential of 0.5 V. Variation in water content in the electrolyte did not have any affect in open circuit potential (OCP) for the Si–air cell. The dependence of discharge capacity of the Si–air cell is studied as a function of water content of RTIL at a fixed discharge current of 0.3 mAh cm^{-2}. Initially, the discharge capacity increases with water content, reaching a maximum value of 72.5 mAh cm^{-2} at 15 vol% of water. On increasing water content beyond 15%, a sharp decrease (up to 40%) is observed in the cell discharge capacity up to 80% water content. Beyond 80% water content of RTIL, the cell could be operated as it did not display any discharge. From these results, it has been concluded that addition of water in RTIL may be having two opposing effect. One, it facilitates the formation of SiO_2 in the bulk electrolyte and not at the air electrode due to which porosity of air electrode remains more or less intact. Second, the activity of RTIL in the electrolyte decreases on increasing the content of water resulting into loss of its distinctive properties of facilitating cathode reaction. With increment in the percentage of water in the electrolyte, the fraction of RTIL in the mixture decreases, and therefore, the unique properties of the RTIL are reduced.

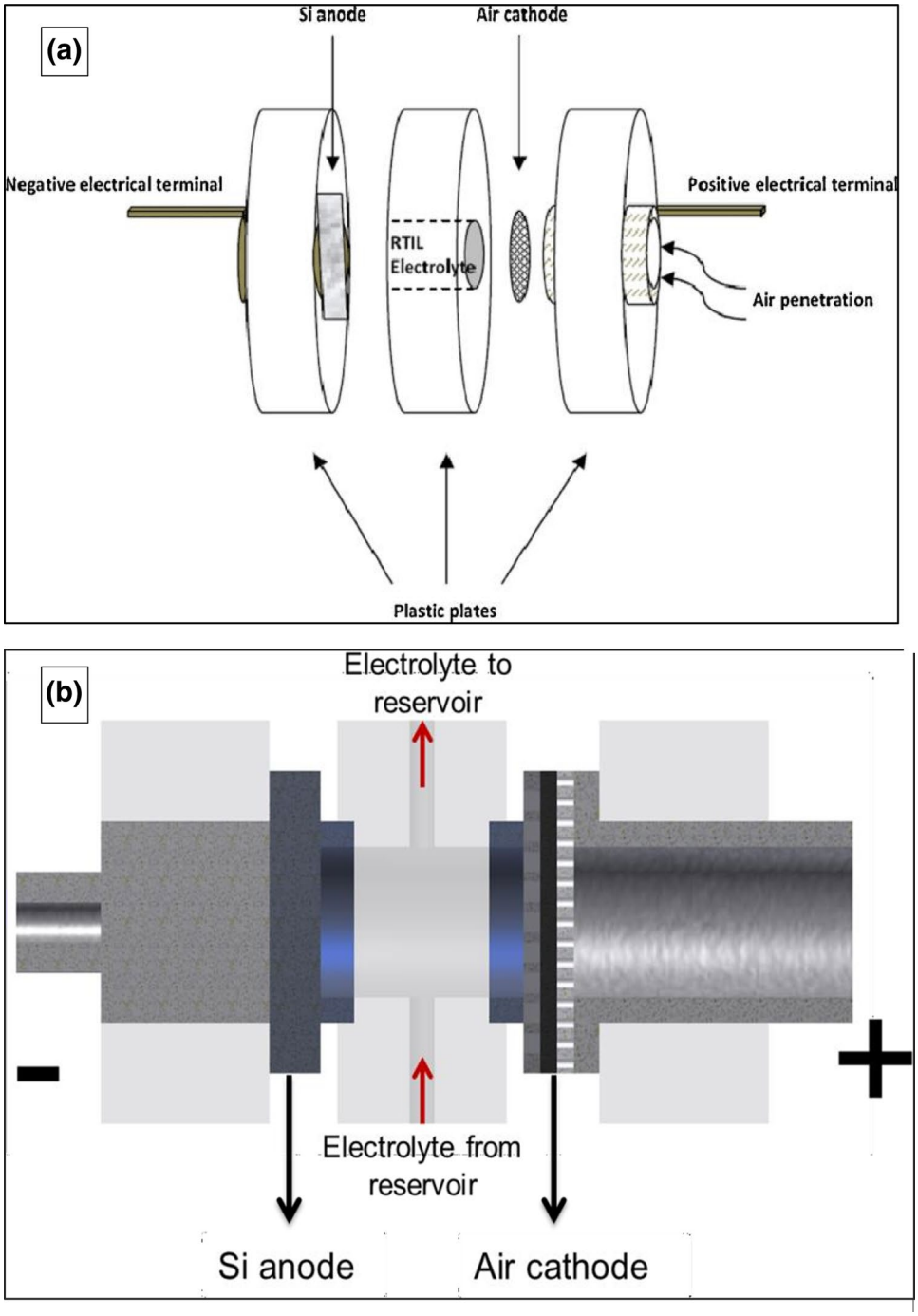


Fig. 6 Schematic diagram of **a** silicon–air battery [28, 29]; **b** A refill-type Si–air cell [28, 29]

5.1.2 Gel polymer electrolytes

Gel polymer electrolytes (GPEs) are formed by integrating a liquid electrolyte into a polymer matrix. Although GPEs are in solid state but the conduction in them is similar to a liquid electrolyte. The use of a gel electrolyte as a moisture barrier eliminates the need to handle liquid electrolyte and simplifies technical issues in cell architecture. Shape flexibility, mechanical stability, high safety and a modest loss in ionic conductivity during discharge process have made GPEs come into focus as electrolytes. They are commonly used in a variety of applications including Li-Ion batteries, solar cells and chemical sensors [101–104].

Yair-Ein-Eli et al. [105] explored the possibility of utilizing a composite polymer electrolyte in Si–air battery system. The gelled polymer electrolyte was derived using specially synthesized EMI.$(HF)_2$.3F ionic liquid and 2-hydroxyethyl methacrylate (HEMA) as the matrix. The combination of the RTIL and polymer composites are also studied extensively by Tsuda et al. [106].

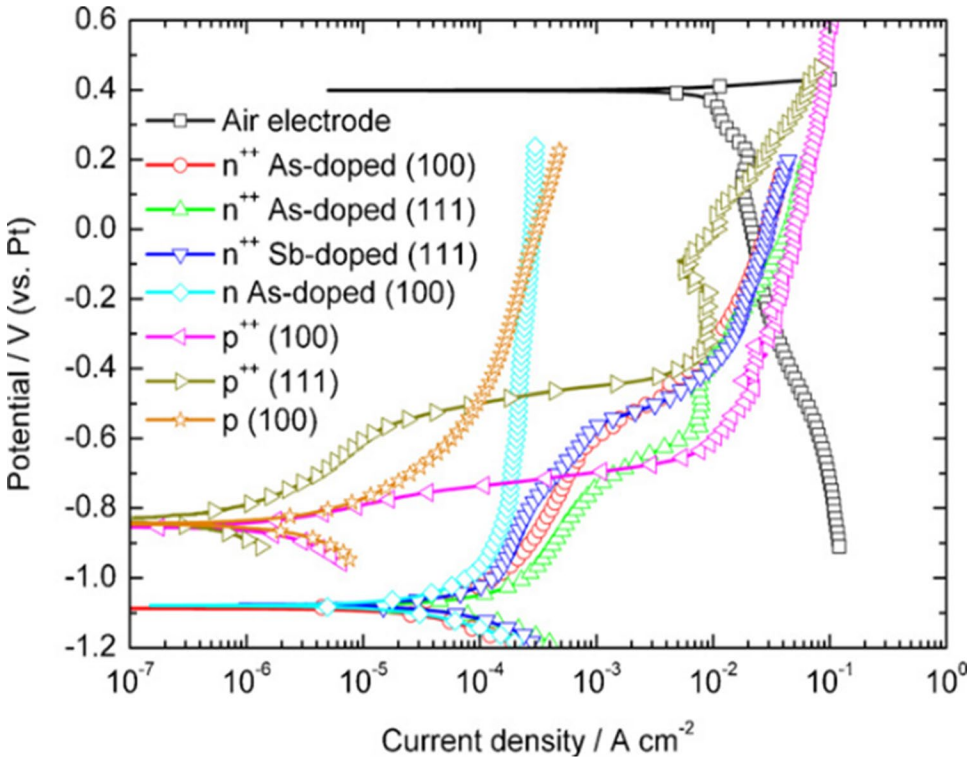


Fig. 7 Polarization voltammograms [19] recorded with RTIL [EMI·$(HF)_2$·3F] at 5 mV s^{-1} for various electrodes: silicon single-crystal anodes and air cathodes

When the battery using GPE of 50–70 mol% ionic liquid is discharged at a current density of 0.1 mA cm^{-2} in an ambient atmosphere, it's operating potential increases with increasing percentage of RTIL. It is observed that at 50 mol% RTIL, it is 0.4 V; at 60 mol% it is 0.5 V and at 70 mol% it is 0.6 V (discharge current was fixed at 0.1 mA cm^{-2}). When compared with discharge potentials of the cells using pure RTIL, the operating potential of GPEs is ~ 0.5 V lower. It has been argued that difference in ionic conductivity of the GPEs is responsible for this variation in operating potential. For instance, it is observed that the ionic conductivity of 60 mol% RTIL is 23 mS cm^{-1} which is lower than the ionic conductivity of pure RTIL (100 mS cm^{-1}).

Gelled electrolytes comprising of 40–70 mol% of EMI.$(HF)_2$·3F RTIL and HEMA polymer are found to be free standing and mechanically stable. These GPE can also be well integrated with both Si anode and air cathode. The cell comprising of these electrolytes exhibited a long discharge time of 850 h at 0.1 mA cm^{-2} operative discharge current density. However, these cells have lower cell voltage when compared with the cells using pure RTIL, and it has been attributed to higher ionic conductivity of the later. Thus, use of GPEs as electrolyte in Si–air battery is feasible for low power devices.

5.1.3 Corrosion analysis of silicon anode with RTIL electrolyte

Jakobi et al. [28, 29] studied corrosion mass losses of As-, Sb-, and B-doped Si anodes at different discharge rates and varying discharge time (up to 20 h) (Fig. 8). It was concluded that the electrochemical etching increases surface area significantly along with formation of oxide over layer. The corrosion mass of all types of Si anodes increases with the increase in current density. The electrochemical etching is capable of producing a strong non-porous Si electrode incorporating thick porous layers and tiny pores.

Such Si anodes shall be favoured in alkaline Si–air batteries due to their enriched discharge behaviour in comparison to flat Si anodes. Further these high surface area nonporous Si anodes in room temperature ionic liquids are capable of minimizing strong passivation and self-discharge of Si anodes, enhancing its extended utilization, resulting into development of Si–air batteries having stable voltage profile. The increase in corrosion mass loss for current densities (up to 0.3 mA cm^{-2}) is found to be linear for n-type Si anodes. The corrosion mass is significantly lower for n-type Si anodes with <111> orientation when compared with <100> orientation. Similar trend is found in the As- and Sb-doped Si anodes as well as in p-type Si

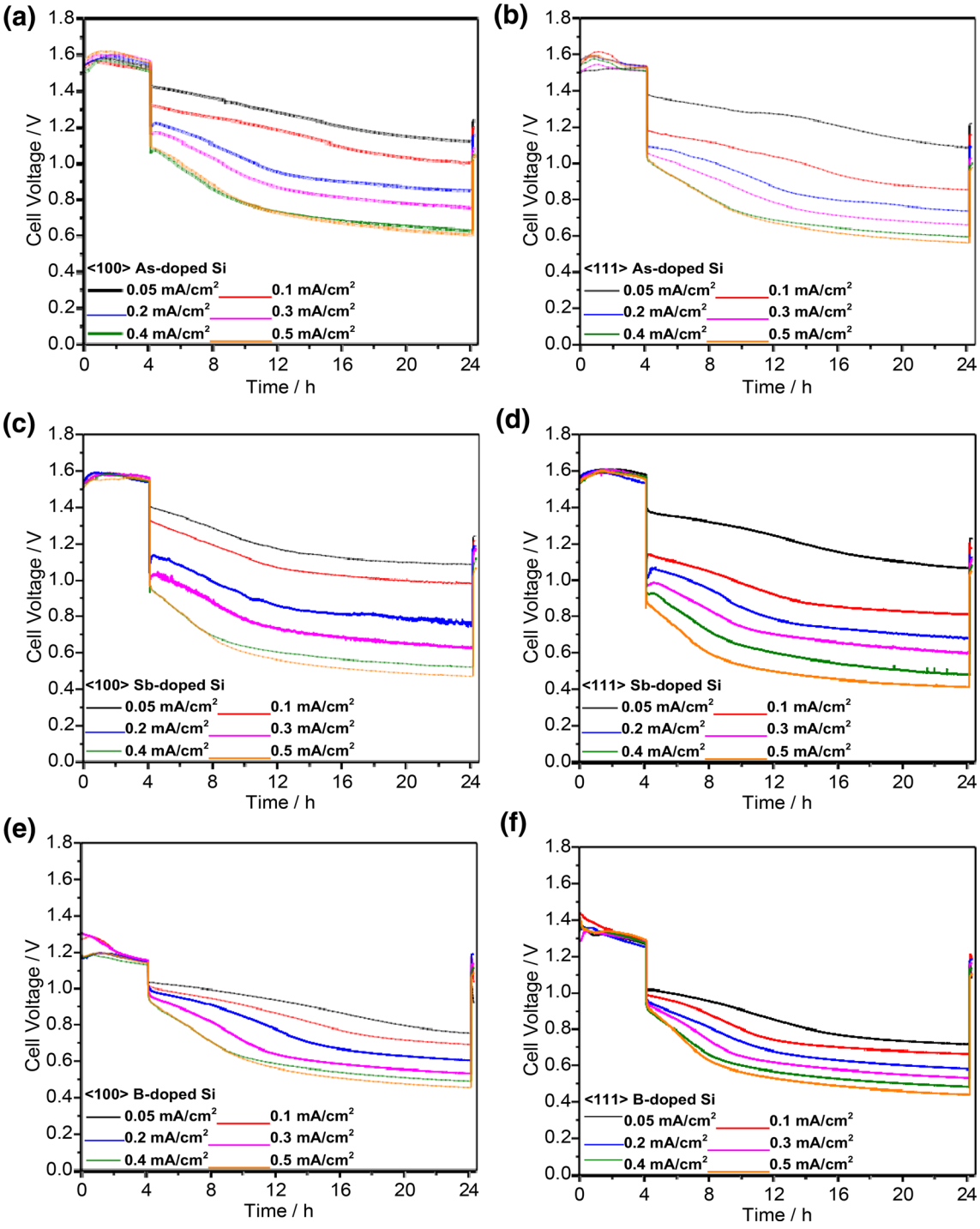


Fig. 8 Influence of dopant type and orientation of silicon anodes on corrosion of Si–air cells for **a** <100> oriented As-doped Si anode; **b** <111> oriented As-doped Si anode; **c** <100> oriented Sb-doped Si anode; **d** <111> oriented Sb-doped Si anode; **e** <100> oriented B-doped Si anode; **f** <111> oriented B-doped Si anode. Durmus et al. [28, 29]

anodes to n-type Si anodes. At higher current densities, the corrosion rate is significantly high. Discharge potentials for various dopants at varying current densities are found as detailed below.

(a) For low current densities: $As_{100} > As_{111} \approx Sb_{100} > Sb_{111} \gg B_{100} > B_{111}$

(b) For high current densities: $As_{100} > As_{111} \gg Sb_{100} > B_{100} > B_{111} > Sb_{111}$

The anode efficiency, which is dependent on corrosion of an anode, is an essential factor for screening of materials for Si–air batteries. Anode efficiency cannot be 100%; it is a factor that is dependent on the corrosion of the doped anode in the electrolyte. Anode mass conversion efficiency of differently doped Si-anodes for varying current densities are analysed and ranked as: *Anodic mass conversion efficiency rankings:* $B_{100} \approx B_{111} \approx As_{111} > Sb_{111} \gg Sb_{100} \approx As_{100}$

On the basis of these findings, it has been concluded that <111> oriented As-doped silicon is the first-rate choice as the anode in Si–air batteries, employing EMI·$(HF)_2$·3F as the electrolyte.

5.1.4 Mechanism of discharge termination in silicon–air batteries

The cell impedance during the discharge process of a Si–Air battery employing RTIL electrolyte has also been investigated [107]. Electrochemical Impedance Spectroscopy (EIS) is utilized to study the generation and development of Si anode and air cathode interfaces in the battery during storage and operation. The effect of exchanging the anode and cathode during discharge and at the end of the discharge is also studied. In order to determine which electrode is contributing to the cell impedance as the most dominant source, it is needed to obtain and analyse the impedance spectra of each electrode. The study of anode impedance exhibited two well-defined semicircles which correspond to high frequency and low frequency regions. The high frequency zone is attributed to space charge layer capacitance and a charge transfer resistance whereas the low frequency zone is attributed to the formation of thin oxide layer on the sub-micron pores during anodic reaction that adds to resistance and capacitance of the cells [108, 109]. Since the total impedance at the air cathode is considerably lower than the impedance at the anode, it can be neglected. Therefore, the overall cell impedance is originated from the silicon anode impedance.

The role of silicon in the potential drop of the battery during discharge is identified by employing a three-electrode configuration to the Si–air system. The initial drop of the air cathode is larger than the Si anode which can be attributed to dissolved oxygen in the electrolyte that is easily accessible and is consumed in the beginning of the operation but once more and more oxygen is consumed, the over potential of the air cathode increases and the voltage of the air cathode remains unmodified at 0.4 V. On the other hand, the over potential of the Si anode behaves differently. The over potential at electrodes increases with continuous discharge of the battery. The voltage profile of Si–air battery mimics the anode profile which clearly confirms the assumption that Si anode is the dominant electrode which controls and defines the discharge behaviour of a Si–air system. Hence, the capacity of the battery is limited by the Si anode rather than the air cathode.

Table 1 Specific capacities of silicon–air batteries

Discharge current density (ma cm^{-2})	KOH concentration (M)	Mass of silicon (mg)	Specific capacity (mAh g^{-1})
0.05	6	2.26	154.8
0.05	2	1.63	214.7
0.05	0.6	0.49	715.7
0.1	0.6	0.58	1206.0

The importance of the Si-anode with respect to discharge of Si–air battery is investigated by replacing the anode electrode during the discharge as well as at the end of the discharge. When the anode is replaced during the middle of the discharge, there is no effect on the discharge voltage. However, the discharge capacity almost doubles at the end of the operation of the second silicon anode. With each changing period, the capacity decreases to such an extent that after the fifth replacement, the increase in the discharge capacity is negligible as compared to previous exchanges. The reason why the fifth replacement does not give a significant increase on discharge capacity is because of the degradation of the air cathode and the electrolyte during the discharge process.

5.2 Alkaline electrolyte

A primary Si–Air battery employing nanostructured silicon as anode material with an alkaline solution based electrolyte was investigated by Zhong et al. [23]. Typical galvanostatic discharge of the cell exhibited that the cell can be used to draw current at 0.05 mA cm^{-2} over a period of 30 h at an operating potential of 1.2 V. In contrast, in a control experiment using an unmodified silicon wafer, the cell is completely discharged in less than 10 min at a potential of 1.1 V i.e. the potential quickly drops to zero. The electrode reactions are:

$$\text{Anode: } Si + 4OH^- \leftrightharpoons Si(OH)_4 + 4e^- E_o = 1.69\,V \quad (8)$$

$$\text{Cathode: } O_2 + 2H_2O + 4e^- \leftrightharpoons 4OH^- E_o = 0.40\,V \quad (9)$$

The main benefit of employing an alkaline solution is its ability to dissolve $Si(OH)_4$. The surface can be covered by silicon oxide if the dissolution rate of $Si(OH)_4$ is not fast enough. The formation of a porous surface structure increases the dissolution rate of $Si(OH)_4$ in the KOH electrolyte. This facilitates the effective removal of the oxide and reactivates the Si-surface. On the other hand, the

unmodified silicon surface become passive very quickly and results in a very short discharge time.

Various concentrations of the alkaline solution as electrolyte are studied. It is reported that the self-discharge rate is highly dependent on the electrolyte composition. It is observed that anodic dissolution potential is directly proportional to KOH concentrations (Table 1).

5.2.1 Corrosion analysis of silicon anode with alkaline electrolyte

Durmus et al. [110] studied the corrosion of As-doped (0.001–0.007 µcm) crystalline Si wafers in 0.5–12 M KOH solution. It is concluded that the chemical corrosion of Si is a surface reaction limited process. While in 0.3–5 M KOH, anisotropic etching leads to very rough surfaces and pyramidal hillocks, smoother surfaces at > 5 M KOH is attributed to polishing process. Hence, the anodic oxidation of As-doped Si does not depend on strength of KOH. Durmus et al. proposed that the electron transfer is the rate determining step of corrosion of As-doped Si using galvanostatic discharge data. They further suggested that the focus on the corrosion inhibition will help in improving the shelf-life and discharge performance of the primary Si–air batteries. It can be concluded that weak KOH enhances the shelf life while strong KOH will ensure continuous discharge at high current densities. The effect of KOH concentration on the corrosion rates of As-doped Si anodes can be understood by considering the following reaction mechanism:

$$Si + 2OH^- + 2H_2O \leftrightharpoons SiO_2(OH)_2^{2-} + 2H_2 \quad (10)$$

At 15 °C, Si undergoes corrosion @ 0.66 µm h^{-1}, which increases to 1.46 µm h^{-1} at 25 °C. Further increase in temperature showed exponential increase of the corrosion rates, observed to be 13.62 µm h^{-1} at 60 °C. This behaviour is in accordance with Arrhenius equation which states that etching rate increases with temperature.

The variation of KOH changes the relative concentration of OH^- and H_2O in the solutions, which results in a change in their activities. In case of hydroxide ions, the activity increases almost linearly beginning from 1 M of KOH whereas KOH activity decreases [111–113].

6 Conclusion and outlook

This mini review is an effort to summarise recent developments of Si-anode, air cathode, different electrolytes and their effect on the discharge mechanisms and working potentials in a Si–air battery system. Despite the intense research, there is still much scope for further advancement of this novel type of Si–air technology. Prospective directions for future research in Silicon–air batteries could focus on the following:

1. Pure silicon wafers are the most common material for Si anode but the drawback is their high corrosion rates. It is required to find a Si anode that is capable to sustain its capacity over hundreds of discharge and charge cycles. Therefore, nano-structured silicon and doped silicon can be used to improve the performance of the Si anode. Structural modifications such as electro deposition or chemical modification by the means of additives has proven to be an ideal solution to find such an anode material. The additives must be effective in small quantities since a large amount of additive can reduce the overall capacity of the silicon material. More studies are needed to properly understand the charge and discharge mechanisms for the optimisation of the discharge time. A more effective methodology is needed to find solution for corrosion or passivation in Si–air battery.
2. The air cathode is an essential part of any type of metal–air battery but its kinetics of the oxygen reduction reaction hinders its performance. Therefore, electro catalysts play an important role in the cathode. Noble metals such as Pt and Pt-alloy based catalysts have been extensively used as electro catalysts in air electrode. To further reduce the cost, different type of catalysts such as high spin transition metal complexes and inorganic–organic composites can be used as a highly efficient alternative. It has been observed that the specific capacity of an air electrode depends on various factors. Thus, a suitable choice for an air electrode would be to choose a very thin and highly porous graphite nano-sheet with optimal pore size which is impregnated with an appropriate electro catalyst. Further development of low cost catalysts with improved activity and durability is necessary to improve the sluggish kinetics of air electrode. One such example of an electro catalyst could be inexpensive bi-functional catalysts which is active to both oxygen evolution reaction and oxygen reduction reaction.
3. New highly stable electrolyte technologies that allow long-lasting operation of Si–air batteries are required for further advancement. At present, Ionic liquids are the most promising electrolytes for Si–Air batteries. Thermally and mechanically stable gel polymer electrolyte appears to be a viable candidate as electrolyte for future research. Solidification of the electrolyte system is important to maintain the safe operation & easy handling of the battery and structural integrity under various deformations. It is also highly important

to study the properties at the electrode–electrolyte interface of the battery.

4. Simulations are needed in order to provide an optimal design of Si–air batteries in terms of structural design of anode and cathode, as well as dispersion of electrolyte, electro catalyst, current collector and depolarizers in order to provide a corrosion-free contact between the electrodes and electrolyte. Moreover, the understanding of silicon dendritic growth in Si–air batteries is still in its infancy. Although there are many challenges faced, it is worth applying efforts to develop Si–air batteries for next-generation energy storage technology.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Zhang X, Wang XG, Xie Z, Zhou Z (2018) Recent progress in rechargeable alkali metal–air batteries. Green Energy Environ 1(1):4–17
2. Fu J, Cano ZP, Park MG, Yu A, Fowler M, Chen Z (2018) Electrically rechargeable zinc–air batteries: progress, challenges, and perspectives. Adv Mater 29(7):1604685
3. Rinaldi A, Wang Y, Tan KS, Wijaya O, Yazami R (2015) Lithium-air batteries for medium- and large-scale energy storage. In: Advances in batteries for medium and large-scale energy storage. pp 387–440
4. Zhang T, Tao Z, Chen J (2014) Magnesium–air batteries: from principle to application. Mater Horiz 1(2):196–206
5. Li Y, Lu J (2017) Metal–air batteries: will they be the future electrochemical energy storage device of choice? ACS Energy Lett 2(6):1370–1377
6. Hamlen RP, Hoge WH, Hunter JA, O'Callaghan WB (1991) Applications of aluminum–air batteries. IEEE Aerosp Electron Syst Mag 6(10):11–14
7. Peng H, Xu Y, Pan J, Zhao Y, Wang L, Shi X (2018) Flexible metal–air batteries: fundamentals and applications. In: Metal-Air Batteries. pp 367–396
8. Das SK, Lau S, Archer LA (2014) Sodium–oxygen batteries: a new class of metal–air batteries. J Mater Chem A 2(32):12623
9. Blurton KF, Sammells AF (1979) Metal/air batteries: their status and potential: a review. J Power Sources 4(4):263–279
10. Girishkumar G, McCloskey B, Luntz AC, Swanson S, Wilcke W (2010) Lithium–air battery: promise and challenges. J Phys Chem Lett 1(14):2193–2203
11. Kraytsberg A, Ein-Eli Y (2011) Review on Li–air batteries—opportunities, limitations and perspective. J Power Sources 196(3):886–893
12. Lee JS, Tai Kim S, Cao R, Choi NS, Liu M, Lee KT, Cho J (2010) Metal-air batteries with high energy density: Li–air versus Zn–air. Adv Energy Mater 1(1):34–50
13. Wang T, Kaempgen M, Nopphawan P, Wee G, Mhaisalkar S, Srinivasan M (2010) Silver nanoparticle-decorated carbon nanotubes as bifunctional gas-diffusion electrodes for zinc–air batteries. J Power Sources 195(13):4350–4355. https://doi.org/10.1016/j.jpowsour.2009.12.137
14. Li CS, Sun Y, Gebert F, Chou SL (2017) Current progress on rechargeable magnesium–air battery. Adv Energy Mater 7(24):1700869
15. Mohamad AA, Mohamed NS, Alias Y, Arof AK (2003) Mechanically alloyed Mg2Ni for metal-hydride–air secondary battery. J Power Sources 115(1):161–166
16. Durmus YE (2018) Investigation and development of a resource efficient metal–air battery: silicon–air. M.Sc. disssertation, RWTH Aachen University, Germany
17. Cohn G, Starosvetsky D, Hagiwara R, Macdonald DD, Ein-Eli Y (2009) Silicon–air batteries. Electrochem Commun 11(10):1916–1918
18. Jakes P, Cohn G, Ein-Eli Y, Scheiba F, Ehrenberg H, Eichel RA (2012) Limitation of discharge capacity and mechanisms of air–electrode deactivation in silicon–air batteries. Chemsuschem 5(11):2278–2285
19. Cohn G, Ein-Eli Y (2010) Study and development of non-aqueous silicon–air battery. J Power Sources 195(15):4963–4970
20. Teki R, Datta MK, Krishnan R, Parker TC, Lu TM, Kumta PN, Koratkar N (2009) Nanostructured silicon anodes for lithium ion rechargeable batteries. Small 5(20):2236–2242
21. Szczech JR, Jin S (2011) Nanostructured silicon for high capacity lithium battery anodes. Energy Environ Sci 4(1):56–72
22. Ge M, Rong J, Fang X, Zhou C (2012) Porous doped silicon nanowires for lithium ion battery anode with long cycle life. Nano Lett 12(5):2318–2323
23. Zhong X, Zhang H, Liu Y, Bai J, Liao L, Huang Y, Duan X (2011) High-capacity silicon–air battery in alkaline solution. Chem Sus Chem 5(1):177–180
24. Zhong X, Qu Y, Lin YC, Liao L, Duan X (2011) Unveiling the formation pathway of single crystalline porous silicon nanowires. ACS Appl Mater Interfaces 3(2):261–270
25. Xue M, Zhong X, Shaposhnik Z, Qu Y, Tamanoi F, Duan X, Zink JI (2011) pH-operated mechanized porous silicon nanoparticles. J Am Chem Soc 133(23):8798–8801
26. Qu Y, Liao L, Li Y, Zhang H, Huang Y, Duan X (2009) Electrically conductive and optically active porous silicon nanowires. Nano Lett 9(12):4539–4543
27. Qu Y, Zhong X, Li Y, Liao L, Huang Y, Duan X (2010) Photocatalytic properties of porous silicon nanowires. J Mater Chem 20(18):3590
28. Durmus YE, Jakobi S, Beuse T, Aslanbas Ö, Tempel H, Hausen F, Kungl H (2017) Influence of dopant type and orientation of silicon anodes on performance, efficiency and corrosion of silicon–air cells with EMIm(HF)2.3F electrolyte. J Electrochem Soc 164(12):A2310–A2320
29. Durmus YE (2018) Investigation and development of a resource efficient metal–air battery: silicon–air, M.S. dissertation
30. Cheng H, Scott K (2010) Carbon-supported manganese oxide nanocatalysts for rechargeable lithium–air batteries. J Power Sources 195(5):1370–1374
31. Xiao J, Wang D, Xu W, Wang D, Williford RE, Liu J, Zhang JG (2010) Optimization of air electrode for li/air batteries. J Electrochem Soc 157(4):A487
32. Yang X, He P, Xia Y (2009) Preparation of mesocellular carbon foam and its application for lithium/oxygen battery. Electrochem Commun 11(6):1127–1130
33. Zhang GQ, Zheng JP, Liang R, Zhang C, Wang B, Hendrickson M, Plichta EJ (2010) Lithium-Air Batteries Using SWNT/CNF Buckypapers as Air Electrodes. J Electrochem Soc 157(8):A953
34. Beattie SD, Manolescu DM, Blair SL (2009) High-capacity lithium–air cathodes. J Electrochem Soc 156(1):A44

35. Read J, Mutolo K, Ervin M, Behl W, Wolfenstine J, Driedger A, Foster D (2003) Oxygen transport properties of organic electrolytes and performance of lithium/oxygen battery. J Electrochem Soc 150(10):A1351
36. Yang X, Xia Y (2009) The effect of oxygen pressures on the electrochemical profile of lithium/oxygen battery. J Solid State Electrochem 14(1):109–114
37. Tran C, Yang XQ, Qu D (2010) Investigation of the gas-diffusion-electrode used as lithium/air cathode in non-aqueous electrolyte and the importance of carbon material porosity. J Power Sources 195(7):2057–2063
38. Yamamoto K, Imaoka T, Chun WJ, Enoki O, Katoh H, Takenaga M, Sonoi A (2009) Size-specific catalytic activity of platinum clusters enhances oxygen reduction reactions. Nat Chem 1(5):397–402
39. Lim B, Jiang M, Camargo PHC, Cho EC, Tao J, Lu X, Xia Y (2009) Pd-Pt bimetallic nanodendrites with high activity for oxygen reduction. Science 324(5932):1302–1305
40. Zhou ZY, Tian N, Li JT, Broadwell I, Sun SG (2011) Nanomaterials of high surface energy with exceptional properties in catalysis and energy storage. Chem Soc Rev 40(7):4167
41. Bing Y, Liu H, Zhang L, Ghosh D, Zhang J (2010) Nanostructured Pt-alloy electrocatalysts for PEM fuel cell oxygen reduction reaction. Chem Soc Rev 39(6):2184
42. Stamenkovic VR, Fowler B, Mun BS, Wang G, Ross PN, Lucas CA, Markovic NM (2007) Improved oxygen reduction activity on Pt3Ni(111) via increased surface site availability. Science 315(5811):493–497
43. Stamenkovic VR, Mun BS, Arenz M, Mayrhofer KJJ, Lucas CA, Wang G, Markovic NM (2007) Trends in electrocatalysis on extended and nanoscale Pt-bimetallic alloy surfaces. Nat Mater 6(3):241–247
44. Greeley J, Stephens IEL, Bondarenko AS, Johansson TP, Hansen HA, Jaramillo TF, Nørskov JK (2009) Alloys of platinum and early transition metals as oxygen reduction electrocatalysts. Nat Chem 1(7):552–556
45. Strasser P, Koh S, Anniyev T, Greeley J, More K, Yu C, Nilsson A (2010) Lattice-strain control of the activity in dealloyed core–shell fuel cell catalysts. Nat Chem 2(6):454–460
46. Kim J, Lee Y, Sun S (2010) Structurally ordered FePt nanoparticles and their enhanced catalysis for oxygen reduction reaction. J Am Chem Soc 132(14):4996–4997
47. Jong Yoo S, Kim SK, Jeon TY, Jun Hwang S, Lee JG, Lee SC, Lim TH (2011) Enhanced stability and activity of Pt–Y alloy catalysts for electrocatalytic oxygen reduction. Chem Commun 47(41):11414
48. Šljukić B, Banks CE, Mentus S, Compton RG (2004) Modification of carbon electrodes for oxygen reduction and hydrogen peroxide formation: the search for stable and efficient sono-electrocatalysts. Phys Chem Chem Phys 6(5):992–997
49. Shao Y, Liu J, Wang Y, Lin Y (2009) Novel catalyst support materials for PEMfuelcells: current status and future prospects. J Mater Chem 19(1):46–59
50. Banham D, Feng F, Fürstenhaupt T, Pei K, Ye S, Birss V (2011) Effect of Pt-loaded carbon support nanostructure on oxygen reduction catalysis. J Power Sources 196(13):5438–5445
51. Bao X, Von Deak D, Biddinger EJ, Ozkan US, Hadad CM (2010) A computational exploration of the oxygen reduction reaction over a carbon catalyst containing a phosphinate functional group. Chem Commun 46(45):8621
52. Von Deak D, Biddinger EJ, Ozkan US (2011) Carbon corrosion characteristics of CNx nanostructures in acidic media and implications for ORR performance. J Appl Electrochem 41(7):757–763
53. Zhou Y, Neyerlin K, Olson TS, Pylypenko S, Bult J, Dinh HN, O'Hayre R (2010) Enhancement of Pt and Pt-alloy fuel cell catalyst activity and durability via nitrogen-modified carbon supports. Energy Environ Sci 3(10):1437. https://doi.org/10.1039/c003710a
54. Pylypenko S, Queen A, Olson TS, Dameron A, O'Neill K, Neyerlin KC, O'Hayre R (2011) Tuning carbon-based fuel cell catalyst support structures via nitrogen functionalization. I. investigation of structural and compositional modification of highly oriented pyrolytic graphite model catalyst supports as a function of nitrogen implantation dose. J Phys Chem C 115(28):13667–13675. https://doi.org/10.1021/jp1122344
55. Oh H-S, Oh J-G, Lee WH, Kim H-J, Kim H (2011) The influence of the structural properties of carbon on the oxygen reduction reaction of nitrogen modified carbon based catalysts. Int J Hydrog Energy 36(14):8181–8186. https://doi.org/10.1016/j.ijhydene.2011.04.139
56. Liu S-H, Wu M-T, Lai Y-H, Chiang C-C, Yu N, Liu S-B (2011) Fabrication and electrocatalytic performance of highly stable and active platinum nanoparticles supported on nitrogen-doped ordered mesoporous carbons for oxygen reduction reaction. J Mater Chem 21(33):12489. https://doi.org/10.1039/c1jm11488c
57. Liang C, Ding L, Li C, Pang M, Su D, Li W, Wang Y (2010) Nanostructured WCx/CNTs as highly efficient support of electrocatalysts with low Pt loading for oxygen reduction reaction. Energy Environ Sci 3(8):1121. https://doi.org/10.1039/c001423k
58. Shao M, Merzougui B, Shoemaker K, Stolar L, Protsailo L, Mellinger ZJ, Chen JG (2011) Tungsten carbide modified high surface area carbon as fuel cell catalyst support. J Power Sources 196(18):7426–7434. https://doi.org/10.1016/j.jpowsour.2011.04.026
59. Zhang W, Sherrell P, Minett AI, Razal JM, Chen J (2010) Carbon nanotube architectures as catalyst supports for proton exchange membrane fuel cells. Energy Environ Sci 3(9):1286. https://doi.org/10.1039/c0ee00139b
60. Yuan Y, Smith JA, Goenaga G, Liu D-J, Luo Z, Liu J (2011) Platinum decorated aligned carbon nanotubes: electrocatalyst for improved performance of proton exchange membrane fuel cells. J Power Sources 196(15):6160–6167. https://doi.org/10.1016/j.jpowsour.2011.03.026
61. Zhang S, Shao Y, Yin G, Lin Y (2010) Carbon nanotubes decorated with Pt nanoparticles via electrostatic self-assembly: a highly active oxygen reduction electrocatalyst. J Mater Chem 20(14):2826. https://doi.org/10.1039/b919494k
62. Zhang Y, Hu Y, Li S, Sun J, Hou B (2011) Manganese dioxide-coated carbon nanotubes as an improved cathodic catalyst for oxygen reduction in a microbial fuel cell. J Power Sources 196(22):9284–9289. https://doi.org/10.1016/j.jpowsour.2011.07.069
63. Fujigaya T, Uchinoumi T, Kaneko K, Nakashima N (2011) Design and synthesis of nitrogen-containing calcined polymer/carbon nanotube hybrids that act as a platinum-free oxygen reduction fuel cell catalyst. Chem Commun 47(24):6843. https://doi.org/10.1039/c1cc11303h
64. Geng D, Chen Y, Chen Y, Li Y, Li R, Sun X, Knights S (2011) High oxygen-reduction activity and durability of nitrogen-doped graphene. Energy Environ Sci 4(3):760. https://doi.org/10.1039/c0ee00326c
65. Hayden BE, Pletcher D, Suchsland J-P, Williams LJ (2009) The influence of support and particle size on the platinum catalysed oxygen reduction reaction. Phys Chem Chem Phys 11(40):9141. https://doi.org/10.1039/b910110a
66. Wu G, Nelson MA, Mack NH, Ma S, Sekhar P, Garzon FH, Zelenay P (2010) Titanium dioxide-supported non-precious metal oxygen reduction electrocatalyst. Chem Commun 46(40):7489. https://doi.org/10.1039/c0cc03088k

67. Ho VTT, Pan C-J, Rick J, Su W-N, Hwang B-J (2011) Nanostructured Ti0.7Mo0.3O_2 support enhances electron transfer to pt: high-performance catalyst for oxygen reduction reaction. J Am Chem Soc 133(30):11716–11724. https://doi.org/10.1021/ja2039562
68. Chen J, Takanabe K, Ohnishi R, Lu D, Okada S, Hatasawa H, Domen K (2010) Nano-sized TiN on carbon black as an efficient electrocatalyst for the oxygen reduction reaction prepared using an mpg-C3N4 template. Chem Commun 46(40):7492. https://doi.org/10.1039/c0cc02048f
69. Hu Z, Chen C, Meng H, Wang R, Shen PK, Fu H (2011) Oxygen reduction electrocatalysis enhanced by nanosized cubic vanadium carbide. Electrochem Commun 13(8):763–765. https://doi.org/10.1016/j.elecom.2011.03.004
70. Yin S, Cai M, Wang C, Shen PK (2011) Tungsten carbide promoted Pd–Fe as alcohol-tolerant electrocatalysts for oxygen reduction reactions. Energy Environ Sci 4(2):558–563. https://doi.org/10.1039/c0ee00445f
71. Post JE (1990) Manganese oxide minerals: crystal structures and economic and environmental significance. Proc Natl Acad Sci 96(7):3447–3454. https://doi.org/10.1073/pnas.96.7.3447
72. Mathur A, Halder A (2019) One step synthesis of bifunctional iron doped manganese oxide nanorods for rechargeable zinc air battery. Catal Sci Technol. https://doi.org/10.1039/c8cy02498g
73. Roche I, Chaînet E, Chatenet M, Vondrák J (2006) Carbon-supported manganese oxide nanoparticles as electrocatalysts for the oxygen reduction reaction (ORR) in alkaline medium: physical characterizations and ORR mechanism. J Phys Chem C 111(3):1434–1443. https://doi.org/10.1021/jp0647986
74. Cheng F, Shen J, Ji W, Tao Z, Chen J (2009) Selective synthesis of manganese oxide nanostructures for electrocatalytic oxygen reduction. ACS Appl Mater Interfaces 1(2):460–466. https://doi.org/10.1021/am800131v
75. Lima FHB, Calegaro ML, Ticianelli EA (2007) Electrocatalytic activity of manganese oxides prepared by thermal decomposition for oxygen reduction. Electrochim Acta 52(11):3732–3738. https://doi.org/10.1016/j.electacta.2006.10.047
76. Jaouen F, Proietti E, Lefèvre M, Chenitz R, Dodelet J-P, Wu G, Zelenay P (2011) Recent advances in non-precious metal catalysis for oxygen-reduction reaction in polymer electrolyte fuel-cells. Energy Environ Sci 4(1):114–130. https://doi.org/10.1039/c0ee00011f
77. Chen Z, Higgins D, Yu A, Zhang L, Zhang J (2011) A review on non-precious metal electrocatalysts for PEM fuel cells. Energy Environ Sci 4(9):3167. https://doi.org/10.1039/c0ee00558d
78. Bezerra CWB, Zhang L, Lee K, Liu H, Marques ALB, Marques EP, Zhang J (2008) A review of Fe–N/C and Co–N/C catalysts for the oxygen reduction reaction. Electrochim Acta 53(15):4937–4951. https://doi.org/10.1016/j.electacta.2008.02.012
79. Easton EB, Yang R, Bonakdarpour A, Dahn JR (2007) Thermal evolution of the structure and activity of magnetron-sputtered TM–C–N (TM= Fe, Co) oxygen reduction catalysts. Electrochem Solid-State Lett 10(1):B6–B10. https://doi.org/10.1149/1.2363947
80. Easton EB, Bonakdarpour A, Yang R, Stevens DA, Dahn JR (2008) Magnetron sputtered Fe–C–N, Fe–C, and C–N based oxygen reduction electrocatalysts. J Electrochem Soc 155(6):B547. https://doi.org/10.1149/1.2899013
81. He Q, Yang X, Ren X, Koel BE, Ramaswamy N, Mukerjee S, Kostecki R (2011) A novel CuFe-based catalyst for the oxygen reduction reaction in alkaline media. J Power Sources 196(18):7404–7410. https://doi.org/10.1016/j.jpowsour.2011.04.016
82. Winther-Jensen B, Winther-Jensen O, Forsyth M, MacFarlane DR (2008) High rates of oxygen reduction over a vapor phase-polymerized PEDOT electrode. Science 321(5889):671–674. https://doi.org/10.1126/science.1159267
83. Trojánek A, Langmaier J, Šebera J, Záliš S, Barbe J-M, Girault HH, Samec Z (2011) Fine tuning of the catalytic effect of a metal-free porphyrin on the homogeneous oxygen reduction. Chem Commun 47(19):5446. https://doi.org/10.1039/c1cc11075f
84. Chen J, Zhang W, Officer D, Swiegers GF, Wallace GG (2007) A readily-prepared, convergent, oxygen reduction electrocatalyst. Chem Commun 32:3353. https://doi.org/10.1039/b707799h
85. Wu G, More KL, Johnston CM, Zelenay P (2011) High-performance electrocatalysts for oxygen reduction derived from polyaniline, iron, and cobalt. Science 332(6028):443–447. https://doi.org/10.1126/science.1200832
86. Bashyam R, Zelenay P (2006) A class of non-precious metal composite catalysts for fuel cells. Nature 443(7107):63–66. https://doi.org/10.1038/nature05118
87. Lee S, Zhu S, Milleville CC, Lee C-Y, Chen P, Takeuchi KJ, Marschilok AC (2010) Metal–air electrochemical cells: silver–polymer–carbon composite air electrodes. Electrochem Solid State Lett 13(11):A162. https://doi.org/10.1149/1.3479660
88. Chen Z, Choi J-Y, Wang H, Li H, Chen Z (2011) Highly durable and active non-precious air cathode catalyst for zinc air battery. J Power Sources 196(7):3673–3677. https://doi.org/10.1016/j.jpowsour.2010.12.047
89. Gong KP, Du F, Xia ZH, Durstock M, Dai LM (2009) Nitrogen-doped carbon nanotube arrays with high electrocatalytic activity for oxygen reduction. Science 323:760–764
90. Shui J, Karan NK, Balasubramanian M, Li SY, Liu DJ (2012) Fe/N/C composite in Li–O_2 battery: studies of catalytic structure and activity toward oxygen evolution reaction. J Am Chem Soc 134:16654–16661
91. Shui J, Wang M, Du F, Dai L (2015) N-doped carbon nanomaterials are durable catalysts for oxygen reduction reaction in acidic fuel cells. Sci Adv 1(1):e1400129
92. Choi CH, Chung MW, Kwon HC, Park SH, Woo S (2013) B, N- and P, N-doped graphene as highly active catalysts for oxygen reduction reactions in acidic medi. J Mater Chem A 1:3694–3699. https://doi.org/10.1039/C3TA01648J
93. Choi CH, Park SH, Woo S (2012) Phosphorus–nitrogen dual doped carbon as an effective catalyst for oxygen reduction reaction in acidic media: effects of the amount of P-doping on the physical and electrochemical properties of carbon. J Mater Chem 22:12107. https://doi.org/10.1039/c2jm31079a
94. Zhang S, Cai Y, He H, Zhang Y, Liu R, Cao H, Wang M, Liu J, Zhang G, Li Y, Liu H, Li B (2016) Heteroatom doped graphdiyne as efficient metal-free electrocatalyst for oxygen reduction reaction in alkaline medium. J Mater Chem A 4:4738–4744. https://doi.org/10.1039/c5ta10579j
95. Zhang J, Dai L (2015) Heteroatom-doped graphitic carbon catalysts for efficient electrocatalyst of oxygen reduction reaction. ACS Catal 5(12):7244–7253. https://doi.org/10.1021/acscatal.5b01563
96. Choi CH, Park SH, Woo S (2011) Heteroatom doped carbons prepared by the pyrolysis of bio-derived amino acids as highly active catalysts for oxygen electro-reduction reactions. Green Chem 13:406–412
97. Hagiwara R, Hirashige T, Tsuda T, Ito Y (2002) A highly conductive room temperature molten fluoride: EMIF 2.3HF. J Electrochem Soc 149(1):D1. https://doi.org/10.1149/1.1421606
98. Matsumoto K, Hagiwara R, Yoshida R, Ito Y, Mazej Z, Benkič P, Matsubara S (2004) Syntheses, structures and properties of 1-ethyl-3-methylimidazolium salts of fluorocomplex anions. Dalton Trans 1:144–149. https://doi.org/10.1039/b310162b
99. Hagiwara R, Nakamori Y, Matsumoto K, Ito Y (2005) The effect of the anion fraction on the physicochemical properties of

EMIm(HF)nF (n = 1.0–2.6). J Phys Chem B 109(12):5445–5449. https://doi.org/10.1021/jp047006l
100. Cohn G, Macdonald DD, Ein-Eli Y (2011) Remarkable impact of water on the discharge performance of a silicon–air battery. ChemSusChem 4(8):1124–1129. https://doi.org/10.1002/cssc.201100169
101. Long L, Wang S, Xiao M, Meng Y (2016) Polymer electrolytes for lithium polymer batteries. J Mater Chem A 4(26):10038–10069. https://doi.org/10.1039/c6ta02621d
102. Wang Y (2009) Recent research progress on polymer electrolytes for dye-sensitized solar cells. Solar Energy Mater Solar Cells 93(8):1167–1175. https://doi.org/10.1016/j.solmat.2009.01.009
103. Fergus JW (2010) Ceramic and polymeric solid electrolytes for lithium-ion batteries. J Power Sources 195(15):4554–4569. https://doi.org/10.1016/j.jpowsour.2010.01.076
104. Zewde BW, Carbone L, Greenbaum S, Hassoun J (2018) A novel polymer electrolyte membrane for application in solid state lithium metal battery. Solid State Ionics 317:97–102. https://doi.org/10.1016/j.ssi.2017.12.039
105. Cohn G, Altberg A, Macdonald DD, Ein-Eli Y (2011) A silicon–air battery utilizing a composite polymer electrolyte. ElectrochimicaActa 58:161–164. https://doi.org/10.1016/j.electacta.2011.09.026
106. Tsuda T (2002) A highly conductive composite electrolyte consisting of polymer and room temperature molten fluorohydrogenates. Solid State Ionics 149(3–4):295–298. https://doi.org/10.1016/s0167-2738(02)00399-5
107. Cohn G, Eichel RA, Ein-Eli Y (2013) New insight into the discharge mechanism of silicon–air batteries using electrochemical impedance spectroscopy. Phys Chem Chem Phys 15(9):3256. https://doi.org/10.1039/c2cp43870d
108. Raz O, Shmueli Z, Hagiwara R, Ein-Eli Y (2010) Porous silicon formation in fluorohydrogenate ionic liquids. J Electrochem Soc 157(3):H281. https://doi.org/10.1149/1.3273082
109. Raz O, Starosvetsky D, Tsuda T, Nohira T, Hagiwara R, Ein-Eli Y (2007) Macroporous silicon formation on n-si in room-temperature fluorohydrogenate ionic liquid. Electrochem Solid State Lett 10(3):D25. https://doi.org/10.1149/1.2409058
110. Durmus YE, Montiel GSS, Aslanbas Ö, Tempel H, Hausen F, de Haart LGJ, Kungl H (2018) Investigation of the corrosion behavior of highly As-doped crystalline Si in alkaline Si–air batteries. ElectrochimicaActa 265:292–302. https://doi.org/10.1016/j.electacta.2018.01.145
111. Akerlof GC, Bender P (1948) Thermodynamics of aqueous solutions of potassium hydroxide. J Am Chem Soc 70(7):2366–2369. https://doi.org/10.1021/ja01187a016
112. Balej J (1985) Water vapour partial pressures and water activities in potassium and sodium hydroxide solutions over wide concentration and temperature ranges. Int J Hydrog Energy 10(4):233–243. https://doi.org/10.1016/0360-3199(85)90093-x
113. Nickell RA, Zhu WH, Payne RU, Cahela DR, Tatarchuk BJ (2006) Hg/HgO electrode and hydrogen evolution potentials in aqueous sodium hydroxide. J Power Sources 161(2):1217–1224. https://doi.org/10.1016/j.jpowsour.2006.05.028

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

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Survey Paper:Introduction to Generative Adversarial Networks in Speech Imitation

Rajashree Guha
Delhi Technological University, Delhi, India
rajashree22guha@gmail.com

Vaibhav Sharma
Delhi Technological University, Delhi, India
vaibhavsharma1998@gmail.com

Yatin Yadav
Delhi Technological University, Delhi, India
yatinyadav1998@gmail.com

Dr. Aruna Bhat
Delhi Technological University, Delhi, India
aruna.bhat@dtu.ac.in

***Abstract*— Speech imitation, that is, the method of converting a source speaker's utterance in a manner such that it sounds as though it were uttered by the target has been an active and volatile area of research in current times. The paper focuses on the techniques of modifying sounds emitted by the original speaker (also called the source). Speech Imitation involves metamorphosing the source speaker's speech to present it in a way as though it were produced by the selected speaker (also called the target). Most systems specializing in speech imitation compulsorily necessitate the usage of parallel data to construct natural or unprocessed sounding speech. Examples of such systems would include RBM [1]. However, there are various obstacles associated with parallel data such as it being an arduous task to gather in real-time and the requirement for the vocabulary to be an exact match which perhaps might not be available to train the model. To remove such dependency on parallel data, a recently innovated model, GANs [3] are being explored.**

Keywords— ***CycleGAN, VoiceGAN, StarGAN, Generative Adversarial Network (GAN)***

I. INTRODUCTION

GAN is a category of frameworks belonging to the field of machine learning. Such a model was first introduced by Ian J. Goodfellow in 2014 [3]. The framework is an algorithmic organization that utilizes two architectures – a generator and a discriminator. As the name "Adversarial" suggests, the two systems are pitted against each other. An integral part of training is the simultaneous training of the generator model (G) and the discriminator model (D). G is modeled to replicate and record the probability distribution of training data (original data) and create data synonymous to it from the active noise which is given as input. D, on the other hand, is trained to figure out if the data under observation is a member of the set of original data or the set of data produced by the generator.

The organization employs the two-player minimax game method where the generator attempts to increase the probability of the discriminator making an error and vice versa.

II. RELATED WORK

Contemporarily, GANs are being comprehensively engaged in style transfer between images or pictures [2, 6]. The idea is analogously being extended to the area of speech imitation. Extending the concept from images to that of speech imitation is made possible because of spectrograms.

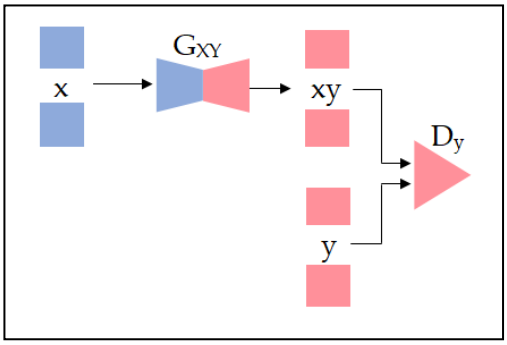


Fig. 1. Generative Adversarial Network Model

The two architectures work together to function as adversaries. Here the input x closely matches probability distribution P(x) (the original data) and y closely matches probability distribution P(y) (the original data). The minimax algorithm employed in the GAN has the discriminator (D) attempting to maximize the discriminator loss

$$L_D(XY) = E_{x \sim P(x)}[\log(1 - D_Y(G_{XY}(x)))] + E_{y \sim P(y)}[\log D_Y(y)] \quad (1)$$

Where G_{XY} converts x into xy and D_Y tries to classify the converted xy correctly. While the generator (G) attempts to minimize the generator loss

$$L_G(XY) = E_{x \sim P(x)}[\log(1 - D_Y(G_{XY}(x)))] \quad (2)$$

In this paper, 3 models of GAN have been discussed and compared.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:42 UTC from IEEE Xplore. Restrictions apply.

III. CYCLEGAN

Cycle-GAN has originally been introduced to be utilized in style transfer in images [5]. The concept has been extended to speech imitation [8] with the introduction of Gated CNN [4] and the addition of identity-mapping loss [9]. The working of cycleGAN is based on the concept that when the *x* (source data) is transformed to *xy* (x→y) and the newly acquired *xy* is re-converted to *xyx* (x→y→x) and *xyx* ought to resemble *x* for successful conversion. CycleGAN employs three loss functions in an attempt to mimic the target sound.

Adversarial loss: It measures the efficiency of the basic architecture of the GAN used in the Cycle-GAN while they execute their respecting functions. D attempts to maximize its loss function, while G attempts to minimize the same.

$$L_A(XY) = E_{x \sim P(x)}[\log(1 - D_Y(G_{XY}(x)))] + E_{y \sim P(y)}[\log D_Y(y)] \quad (3)$$

The methodology followed by cycleGAN is as: the generator converting *x* to *y*, G_{XY} captures the probability distribution of original data (*x*) and is a parameter for its input and produces a specimen of source data converted to *y* (*xy*) as output, which resembles *y*. The discriminator D_Y is responsible for reckoning *xy* as a fake sample, thus producing an output of a measure approximate to 0 and classifying *y* as a part of the original or source sample and hence producing a number approximate to 1 as output.

Cycle-consistency loss [7]: While the above adversarial loss involves the generator (G) converting *x* to *y* and the discriminator ensures that the generated *xy* resembles *y* as much as possible, but it does not ensure the preservation of linguistic content innate to the original distribution sample (*y*) in the converted distribution sample (*xy*). To circumvent the possible loss of linguistic content, the cycle-consistency loss, mapping *x* with *xyx*, is a beneficial introduction.

$$L_C = E_{y \sim P(y)}[\,|\,y - G_{XY}(G_{YX}(y))\,|\,] + E_{x \sim P(x)}[\,|\,x - G_{YX}(G_{XY}(x))\,|\,] \quad (4)$$

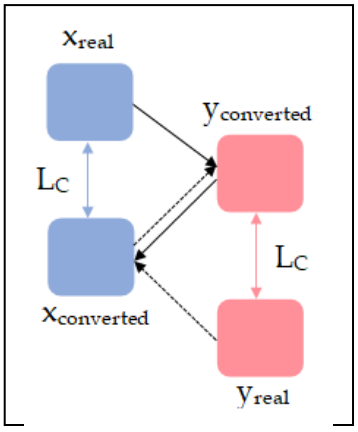


Fig. 2. Cycle Consistency Loss

The loss quantity is introduced in the final loss function by multiplying it with a regularization limit λ_C ranging between 0 and 1.

Identity-mapping loss: Its functionality has primarily been illustrated for style transfer in images as it successfully conserves the hue of the painting. Analogously, in case of sound, it is useful in retaining the linguistic features. The loss is multiplied with the regulation factor λ_I.

$$L_I = E_{x \sim P(x)}[\,|\,x - G_{YX}(x)\,|\,] + E_{y \sim P(y)}[\,|\,y - G_{XY}(y)\,|\,] \quad (5)$$

Therefore, the loss function for D and G of Cycle-GAN becomes from (3), (4), and (5):

$$L_{CYC\text{-}D} = L_D(XY) + L_D(YX) \quad (6)$$

$$L_{CYC\text{-}G} = L_G(XY) + L_G(YX) + \lambda_C \cdot L_{CYC} + \lambda_I \cdot L_I \quad (7)$$

Adding (6) and (7) gives us the total loss function,

$$L_{CYC\text{-}FULL} = L_A(XY) + L_A(YX) + \lambda_C \cdot L_{CYC} + \lambda_I \cdot L_I \quad (8)$$

IV. VOICEGAN

VoiceGAN [12] is built on the framework of DiscoGAN [11] which works over images for purpose of style-transfer. DiscoGAN employs two GAN models simultaneously for the transfer of *x*→*y* and *y*→*x*. G_{XY} converts source distribution *x* to the required converted *y* (*xy*) and G_{YX} converts source distribution *y* to the required converted *x* (*yx*). D_X calculates the similarity between the original data distribution *x* and the converted *x* (*yx*), whereas D_Y does the same for *y*. A considerable adaptation is that voiceGAN ensures that *x* and *xyx* (x→y→x) remain the same and so does *y* and *yxy* (y→x→y).

The first-hand problems faced by discoGAN are to retain the linguistic content of the audio and to tackle the problem of different sizes of audio used as input, which is not the case for images. For this purpose voiceGAN introduces an extra discriminator (style discriminator, D_S) that uses the concept of labeling; hence making sure that the converted data also contains the linguistic content of the original data distribution.

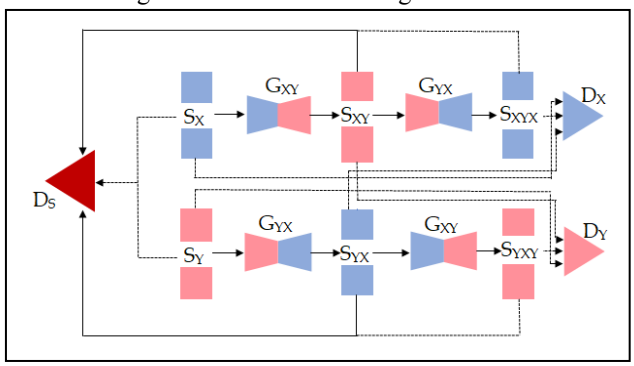


Fig. 3. VoiceGAN model

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:42 UTC from IEEE Xplore. Restrictions apply.

Reconstruction Loss: The original data x and the $x \rightarrow y$ are matched to see if any dissimilarity has crept in during the process. That is between S_X and S_{XYX}. This is multiplied by an adjusting factor 'α'. It also computes the difference between y and yx . That is between S_X and S_{XY}. This is multiplied by an adjusting factor 'β'.

$$L_R(XYX) = \alpha \cdot E_{x \sim P(x)} [\, |G_{YX}(G_{XY}(x)) - x| \,] + \beta \cdot E_{x \sim P(x)} [\, |G_{XY}(x) - x| \,] \quad (9)$$

Style Loss: The introduced Discriminator D_S will make sure that both the original and the obtained data fall under the same label. It has three components:

- It calculates the similarity between the $label_X$ and the original data X provided.
- Next, it calculates the similarity between the $label_Y$ and the converted data Y generated by the generator $G_{XY}(x)$.
- Next, it calculates the similarity between the $label_X$ and the reconverted data X generated by the generator $G_{YX}(G_{XY}(x))$.

$$L_S(XYX) = E_{x \sim P(x)} [d(label_X, D_S(x))] + E_{x \sim P(x)} [d(label_Y, D_S(G_{XY}(x)))] + E_{x \sim P(x)} [d(label_X, D_S(G_{YX}(G_{XY}(x))))] \quad (10)$$

Therefore, the overall loss function becomes:

$$L_{VOC\text{-}D} = L_D(XY) + L_S(XYX) + L_D(YX) + L_S(XYX) \quad (11)$$

$$L_{VOC\text{-}G} = L_R(XYX) + L_G(XY) + L_R(YXY) + L_G(YX) \quad (12)$$

V. STARGAN

StarGAN [10] is an advancement of Cycle-GAN. This is because of the following reasons:

- Star-GAN facilitates many-to-many mappings and without the aid of parallel-data, allowing it to accomplish what had been attempted in several methods previously [8, 14].
- Based on Cycle-GAN, Star-GAN presents attribute label 'c'- a succession of one-hot vectors representing several features of the audio. The attribute label constitutes of the value 1 at the index of its corresponding class and the value 0 for the rest of the indices.
- Star-GAN also introduces a domain classifier (C) and thus, a domain loss as well.

Star-GAN is thus capable of converting dissimilar voices into one another by using the concept of labels (c).

The domain classification loss is developed for C and G as:

$$L_{DC\text{-}C}(c, y) = - E_{c \sim p(c),\, y \sim p(y|c)} [\log p_C(c \mid y)] \quad (13)$$

$$L_{DC\text{-}G}(G) = - E_{x \sim p(x),\, c \sim p(c)} [\log p_C(c \mid G(x, c))] \quad (14)$$

Where, $p_C(c \mid y)$ is the probability that the original sample y, is associated with attribute 'c' and $p_C(c \mid G(x, c))$ is the probability that the converted y (xy) matches attribute 'c'.

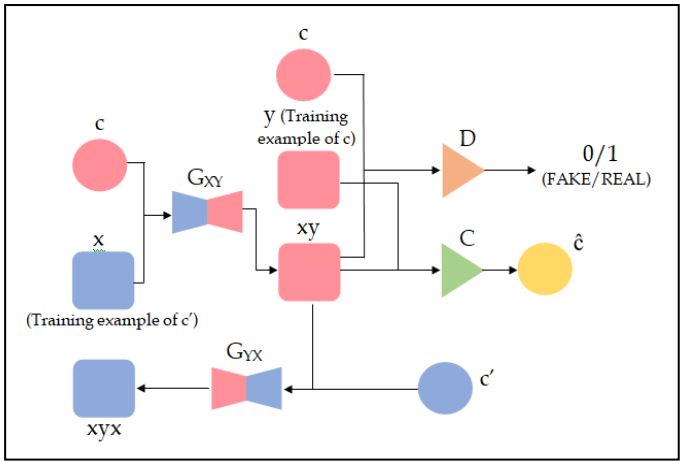


Fig. 4. StarGAN model

G and C attempt to reduce the loss function as much as possible, which would imply that C has correctly classified G(x, c) based on y ~ p(y|c), as belonging to attribute c.

Overall loss function becomes:

$$L_{STR\text{-}G} = L_A(XY) + L_A(YX) + \lambda_{DC} \cdot L_{DC\text{-}G}(G_{XY}) + \lambda_{DC} \cdot L_{DC\text{-}G}(G_{YX}) + \lambda_C \cdot L_C + \lambda_I \cdot L_{ID} \quad (15)$$

$$L_{STR\text{-}D} = L_D(XY) + L_D(YX) \quad (16)$$

$$L_{STR\text{-}C} = L_{DC\text{-}C}(c, y) + L_{DC\text{-}C}(c', x) \quad (17)$$

TABLE I. COMPARING STAR, VOICE, CYCLEGAN

	StarGAN	**VoiceGAN**	**CycleGAN-VC**
Loss function	Domain classification, Adversarial, Identity mapping, Cycle consistency	Style, Reconstruction	Adversarial, Identity mapping, Cycle consistency
Parallel data	No need	No need	No need
Model built upon	CycleGAN-VC	DiscoGAN	CycleGAN (for images)
Accuracy	Subjective : Preference score Objective : Mel - cepstral coefficients	Subjective : Visualization of spectrograms Objective : NIST - STNR test	Subjective : Mean opinion score Objective : Modulation Spectra [17] and Global

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:42 UTC from IEEE Xplore. Restrictions apply.

	StarGAN	VoiceGAN	CycleGAN-VC
			Variance[13]
Dataset	VCC 2018 [20]	TIDIGITS [15]	VCC 2016 [16]
Future Scope	Enhancing accuracy	Modify various style attributes (minimal linguistic modifications), Various variations	Using in STFT spectrograms [19]; Applying in general uses

VI. CONCLUSION

Voice, Cycle, and Star GAN's all take different loss functions as given in Table 1 but a basic similarity can be drawn. The loss function's input is shown in Figure 5. L_G and L_D are fundamental building components of a regular GAN both are combined as adversarial loss and are used in Voice-GAN. These components are also used in the other two GAN's as the adversarial loss. The cycle consistency loss takes *x* and *xyx* as input and identity mapping loss which takes *x* and *xy* as inputs are used in star and cycleGAN follow a similar pattern to the reconstruction loss as used in voiceGAN. There is one additional feature in Voice and Star-GAN that is they make use of labels. Star-GAN includes the label of *y* & *xy* although VoiceGAN includes a label for *x*, *xy* & *xyx*. Thus a similar outline can be drawn for all three GAN's. Till now the focus was on the structures of all 3 GANs.

To find some inference for the working of the models they are judged based on their results. The above assessment is carried on the basics of loss functions in Voice, Cycle, and Star-GAN's.

TABLE II. COMPARING VCC 2016, TIDIGITS, VCC 2018

	VCC 2016	TIDIGITS	VCC 2018
Author	Chen, Ling-Hui; Tomoki, Toda; Villavicencio, Fernado; Wu,Zhizheng; Wester, Mirjam; Yamagishi,junichi (2016)	R. Gary Leonard, George R. Doddington	Yamagishi, Junichi; Ling, Zhenhua; Kinninen,Tomi; Villavicencio, Fernando; Saito, Daisuke; Toda, Tomoki; Lorenzo-Trueba, Jaime. (2018)
Sentences	54 and 162 utterances for assessment and training respectively	Uttering 77 digit sequence	35 and 81 utterances for assessment and training respectively
Speaker	5-Males, 5-Females	111-Adult Male, 115-Adult Females, 50-51 boys and girls	6-Male and 6-Female
Frequency	16000 Hz	20000 Hz	22050 kHz

Voice conversion challenge-2016 dataset is derived from Data And Production Speech dataset [18]. TIDIGITS was derived for writing algorithms to find dissimilar digit patterns. VCC 2018 is an update of VCC 2016 after making three major changes which were.

- Spoke task was introduced which was new and in this task speech imitation is done using not parallel data after it we evaluate it against parallel voice data system.
- Data amount reduced to half to make it more practical
- An attempt was made to fill the gap between VC and ASV communities.

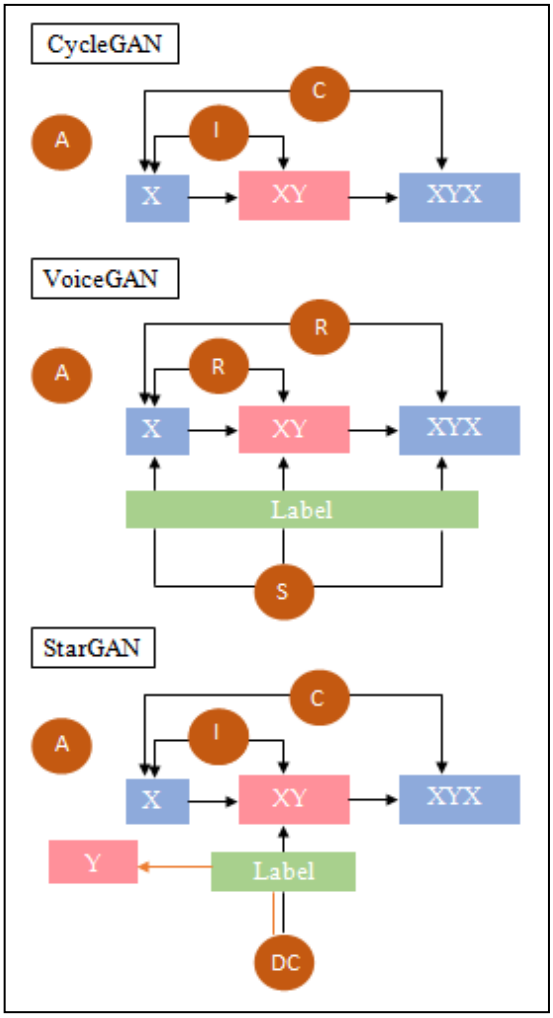


Fig. 5. Comparison of Loss Functions

VII. FUTURE WORK

All 3 GAN models, cycleGAN [8], voiceGAN [12], and starGAN [10] work on different loss functions but on comparison are inherently based on the concept of measuring the differences generated by the generator between the *x*, *xy*, *xyx*, *y*, *yx* and *yxy*. Labels introduced also help in retaining the linguistic content of the audio. Various adjustments and improvements such as STFT spectrograms, conversion between different languages, and mixing of two voices could be implemented.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:42 UTC from IEEE Xplore. Restrictions apply.

For verified benchmarks, various challenges such as Voice conversion Challenge [16, 20] are organized.

References

[1] NAKASHIKA, T., TAKIGUCHI, T. and ARIKI, Y., 2014. Voice Conversion Based on Speaker-Dependent Restricted Boltzmann Machines. *IEICE Transactions on Information and Systems*, E97.D(6), pp.1403-1410.

[2] Gatys, L., Matthias, B., Aaron, H., Eli, S., (2016). Preserving Color in Neural Artistic Style Transfer

[3] Goodfellow, Ian & Pouget-Abadie, Jean & Mirza, Mehdi & Xu, Bing & Warde-Farley, David & Ozair, Sherjil & Courville, Aaron & Bengio, Y.. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems. 3.

[4] Dauphin, Y. N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks, in Proc. ICML, 2017, pp. 933–941.

[5] Zhu, J.-Y., Park, T., Isola, P., Efros, A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks, in Proc. ICCV, 2017, pp. 2223– 2232.

[6] Gatys, L. A., Ecker, A. S., Bethge, M., Hertzmann, A. and Shechtman, E., "Controlling Perceptual Factors in Neural Style Transfer," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, 2017, pp. 3730-3738.

[7] Zhou, T., Krahenbuhl, P., Aubry, M., Huang, Q., Efros, A. A.: Learning dense correspondence via 3D-guided cycle consistency," in Proc. CVPR, 2016, pp. 117– 126.

[8] Kaneko, T., Kameoka, H.: Parallel-data-free voice conversion using cycle-consistent adversarial networks, arXiv:1711.11293 [stat.ML], Nov. 2017.

[9] Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation, in ICLR, 2017.

[10] Kameoka, H., Kaneko, T., Tanaka, K., Hojo, N.: StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks, (2018) 266-273.

[11] Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks, arXiv preprint arXiv:1703.05192, 2017.

[12] Gao, Y., Singh, R., Raj, B.: Voice Impersonation Using Generative Adversarial Networks, (2018) 2506-2510.

[13] Toda, T., Black, A. W., Tokuda, K.: Voice conversion based on maximumlikelihood estimation of spectral parameter trajectory, IEEE/ACM Trans. Audio Speech Lang. Process., vol. 15, no. 8, pp. 2222–2235, 2007.

[14] Saito, Y., Ijima, Y., Nishida, K., Takamichi, S.: Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5274–5278.

[15] Linguistic Data Consortium, "TIDIGITS," https://catalog.ldc.upenn.edu/LDC93S10, 1993.

[16] Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., Yamagishi, J.: The Voice Conversion Challenge 2016, in Proc. INTERSPEECH, 2016, pp. 1632–1636.

[17] Takamichi, S., Toda, T., Neubig, G., Sakti, S., Nakamura, S.: A postfilter to modify the modulation spectrum in HMM-based speech synthesis, in Proc. ICASSP, 2014, pp. 290–294.

[18] Mysore G. J.: Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? – a dataset, insights, and challenges, IEEE Signal Processing Letters, vol. 22, no. 8, pp. 1006–1010, Aug 2015.

[19] Kaneko, T., Takaki, S., Kameoka, H., Yamagishi, J.: Generative adversarial network-based postfilter for STFT spectrograms, in Proc. INTERSPEECH, 2017, pp. 3389–3393.

[20] Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., Ling, Z.: The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods, arXiv: 1804.04262 [eess.AS], Apr. 2018.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:42 UTC from IEEE Xplore. Restrictions apply.

The Astrophysical Journal, 896:89 (11pp), 2020 June 10 https://doi.org/10.3847/1538-4357/ab8c4a
© 2020. The American Astronomical Society. All rights reserved.

Testing Rotating Regular Metrics as Candidates for Astrophysical Black Holes

Rahul Kumar[1], Amit Kumar[2], and Sushant G. Ghosh[1,3]
[1] Centre for Theoretical Physics, Jamia Millia Islamia, New Delhi 110025, India; rahul.phy3@gmail.com
[2] Department of Applied Physics, Delhi Technological University, Delhi 110042, India
[3] Astrophysics and Cosmology Research Unit, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag 54001, Durban 4000, South Africa
Received 2020 February 1; revised 2020 April 9; accepted 2020 April 21; published 2020 June 16

Abstract

The Event Horizon Telescope, a global submillimeter wavelength very long baseline interferometry array, produced the first image of supermassive black hole M87* showing a ring of diameter $\theta_d = 42 \pm 3\ \mu$as, inferred a black hole mass of $M = (6.5 \pm 0.7) \times 10^9 M_\odot$, and allowed us to investigate the nature of strong-field gravity. The observed image is consistent with the shadow of a Kerr black hole, which according to the Kerr hypothesis describes the background spacetimes of all astrophysical black holes. The hypothesis, a strong-field prediction of general relativity, may be violated in the modified theories of gravity that admit non-Kerr black holes. Here, we use the black hole shadow to investigate the constraints when rotating regular black holes (non-Kerr) can be considered as astrophysical black hole candidates, paying attention to three leading regular black hole models with additional parameters g related to nonlinear electrodynamics charge. Our interesting results based on the systematic bias analysis are that rotating regular black holes shadows may or may not capture Kerr black hole shadows, depending on the values of the parameter g. Indeed, the shadows of Bardeen black holes ($g \lesssim 0.26M$), Hayward black holes ($g \lesssim 0.65M$) and non-singular black holes ($g \lesssim 0.25M$) are indistinguishable from Kerr black hole shadows within the current observational uncertainties, and thereby they can be strong viable candidates for the astrophysical black holes. Whereas Bardeen black holes ($g \leqslant 0.30182M$), Hayward black holes ($g \leqslant 0.73627M$), and non-singular black holes ($g \leqslant 0.30461M$), within the 1σ region for $\theta_d = 39\ \mu$as, are consistent with the observed angular diameter of M87*.

Unified Astronomy Thesaurus concepts: General relativity (641); Gravitational lensing (670); Black hole physics (159); Supermassive black holes (1663)

1. Introduction

The Kerr hypothesis states that the astrophysical black hole candidates are well described by the Kerr metric (Kerr 1963). Theoretically, this hypothesis is based on the uniqueness theorem that led to general relativity's no-hair theorem stating that the Kerr (1963) and Kerr–Newman (Newman et al. 1965) are the only stationary, axially symmetric, and asymptotically flat, respectively, vacuum and electro-vacuum solutions of the Einstein equations (Israel 1967, 1968; Carter 1971; Hawking 1972; Robinson 1975). But direct evidence of these solutions is still inconclusive and it may be difficult to rule out non-Kerr black holes (Ryan 1995; Will 2006; Bambi & Barausse 2011). With the current unprecedented observational techniques, it could become possible to probe deeper into the strong-field regime of gravity and eventually test the no-hair theorem. A rotating regular black hole, a prototype non-Kerr metric, has an additional parameter g due to magnetic charge arising from the nonlinear electrodynamics (NED) apart from mass M and rotation parameter a, and encompasses the Kerr black hole ($g = 0$) (Kerr 1963). Are such black holes candidates for the test of the no-hair theorem or the Kerr hypothesis? The idea of a regular black hole was realized by the Bardeen (1968); based on the idea of Sakharov (1966) and Gliner (1966) he presented the first-ever static spherically symmetric regular black hole model. Since then, invoking a suitable NED Lagrangian density, a large number of regular black holes have been constructed (Dymnikova 1992, 2004; Ayon-Beato & Garcia 1998, 1999; Bronnikov 2001; Burinskii & Hildebrandt 2002; Berej et al. 2006; Bronnikov & Fabris 2006; Junior et al. 2015; Sajadi & Riazi 2017), and more recently in Fan & Wang (2016), Bronnikov (2017), and Toshmatov et al. (2018). Subsequently, several rotating regular black holes also have been introduced (Bambi & Modesto 2013; Azreg-Ainou 2014; Toshmatov et al. 2014; Ghosh & Maharaj 2015; Ghosh 2015; Rodrigues & Junior 2017). Regular black holes and their extensions have been extensively studied in the wide context of astrophysical phenomena, namely, particle motion (Stuchlik & Schee 2014; Amir & Ghosh 2015; Garcia et al. 2015; Amir et al. 2016), and gravitational lensing and shadows (Eiroa & Sendra 2013; Li & Bambi 2014; Tsukamoto et al. 2014; Schee & Stuchlik 2015; Abdujabbarov et al. 2016; Amir & Ghosh 2016; Jusufi et al. 2018; Lamy et al. 2018; Kumar et al. 2019a; Ovgun 2019). Interestingly, the size and shape of rotating regular black hole shadows are affected such that the shadow radius decreases and distortion increases monotonically with the parameters g, and thus a violation of the no-hair theorem significantly alters the shadow shape and size (Bambi & Freese 2009; Johannsen & Psaltis 2010; Falcke & Markoff 2013; Johannsen 2013a, 2016). The black hole shadow relevance for testing the strong-field features of gravity, estimating black hole parameters and deducing any potential deviation from the Kerr geometry has resulted in comprehensive literature addressing shadows in both general relativity and modified gravities (Hioki & Maeda 2009; Amarilla et al. 2010; Amarilla & Eiroa 2012; Yumoto et al. 2012; Atamurotov et al. 2013; Grenzebach et al. 2014; Abdujabbarov et al. 2015; Ghasemi-Nodehi et al. 2015; Cunha et al. 2017; Amir et al. 2018; Ayzenberg & Yunes 2018; Perlick et al. 2018; Wang et al. 2018, 2019; Konoplya & Zhidenko 2019; Long et al. 2019; Mishra et al. 2019; Shaikh 2019).

Black hole shadows have become a physical reality with the recent detection of the horizon-scale image of the M87* black hole by the Event Horizon Telescope (EHT). Adopting a distance of $d = 16.8$ Mpc and fitting the geometric crescent models obtained from the general-relativistic magnetohydrodynamic simulations to the M87* observational visibility data the mass of the M87* black hole is estimated $M = (6.5 \pm 0.7) \times 10^9 M_\odot$ (Akiyama et al. 2019a, 2019b, 2019c). The bright sharp photon ring, a projection along the null geodesics of the photons orbiting around the black hole, encompasses the shadow and explicitly depends on the black hole parameters while largely remaining independent of the detailed accretion models (Beckwith & Done 2005; Johannsen & Psaltis 2010; Johannsen 2013b). The constraints on the compact emission region size with angular diameter $\theta_d = 42 \pm 3$ μas along with the central flux depression with a factor of $\geqslant 10$, which can be identified as the shadow, provide stringent evidence for the existence of the black hole. Though the observed shadow of the M87* black hole is found to be consistent with that for the Kerr black hole as predicted by general relativity, the current uncertainty in the measurement of spin angular momentum and the relative deviation of quadrupole moments do not eliminate non-Kerr black holes arising in modified gravities (Akiyama et al. 2019a, 2019b, 2019c; Cardoso & Pani 2019).

Using the M87* black hole shadow, one can investigate the viability of different black hole models in explaining the observational data and put constraints on the black hole parameters. The main aim of this paper is to investigate if black hole shadow can help us to probe whether the rotating regular black holes can be a candidate for astrophysical black holes. We made a systematic bias analysis between the model and injection shadows by considering the rotating regular black hole shadow as a model to fit with the Kerr shadow injections and for this purpose the shadow area A and oblateness D are used as observables to characterize the shadow (Tsupko 2017; Kumar & Ghosh 2020). The best-fit values of the parameters (a, g) for which rotating regular black holes well capture the given Kerr shadow are determined. It is found that as the parameters lie in a certain special range, the rotating regular black holes shadows may illustrate the resemblance with the Kerr black hole shadows and are indiscernible within the current observational uncertainties and even can affirm the apparent asymmetry and angular size of the observed M87* shadow.

The paper is organized as follows. Section 2 is devoted to the brief review of the null geodesic equations in general stationary, axially symmetric regular black hole spacetimes and the recipe to investigate the shadow. In Section 3 we discuss the systematic bias analysis and check the compatibility of rotating regular black holes to explain the Kerr black hole shadows in Section 4. In Section 5 we constrain the rotating regular black hole parameters with the aid of observed image of the M87* black hole. Finally, in Section 6 we summarize the obtained results and discuss the possibility of candidature of rotating regular black holes, with the shadow observations, for the astrophysical black holes.

2. Rotating Black Hole Shadow

A black hole surrounded by an optically thin emission region appears as a dark shadow in the observer's sky, which is an apparent cross-section of the gravitationally captured photon region confined by the innermost unstable photon orbits (Cunningham & Bardeen 1973). For this purpose, we study the photon geodesics around the rotating regular black hole spacetime, whose line element in Boyer–Lindquist coordinates (t, r, θ, ϕ) reads (Bambi & Modesto 2013; Toshmatov et al. 2014)

$$ds^2 = -\left(1 - \frac{2m(r)r}{\Sigma}\right)dt^2 - \frac{4am(r)r}{\Sigma}\sin^2\theta\, dt\, d\phi + \frac{\Sigma}{\Delta}dr^2 + \Sigma\, d\theta^2 + \left[r^2 + a^2 + \frac{2m(r)ra^2}{\Sigma}\sin^2\theta\right]\sin^2\theta\, d\phi^2, \quad (1)$$

and

$$\Sigma = r^2 + a^2\cos^2\theta, \quad \Delta = r^2 + a^2 - 2m(r)r, \quad (2)$$

where $m(r)$ is the mass function such that $\lim_{r\to\infty} m(r) = M$ and a are the spin parameter defined as $a = J/M$; J and M are, respectively, the angular momentum and ADM mass of rotating black hole. The metric (1) is Kerr (1963) and Kerr–Newman (Newman et al. 1965) spacetimes, respectively, when $m(r) = M$ and $m(r) = M - Q^2/2r$. In general metric (1) also describes various classes of rotating regular black holes when we choose $m(r)$ appropriately (Ghosh 2015; Abdujabbarov et al. 2016; Amir & Ghosh 2016). The photon motion in the spacetime (1) is determined by the corresponding geodesics equations obtained from the Hamilton–Jacobi equation (Carter 1968)

$$\frac{\partial S}{\partial \tau} = -\frac{1}{2}g^{\alpha\beta}\frac{\partial S}{\partial x^\alpha}\frac{\partial S}{\partial x^\beta}, \quad (3)$$

where τ is the affine parameter along the geodesics, and S is the Jacobi action, which reads as

$$S = -\mathcal{E}t + \mathcal{L}\phi + S_r(r) + S_\theta(\theta), \quad (4)$$

$S_r(r)$ and $S_\theta(\theta)$, respectively, are functions only of the r and θ coordinates. The metric (1) is time translational and rotational invariant, which leads to conserved quantities along geodesics, namely, energy $\mathcal{E} = -p_t$ and axial angular momentum $\mathcal{L} = p_\phi$, where p_μ is the photon's four-momentum. The Petrov-type D character of metric (1) ensures the existence of Carter's separable constant $\mathcal{K}$, which eventually leads to the following complete set of null geodesics equations in the first-order differential form (Carter 1968; Chandrasekhar 1985):

$$\Sigma\frac{dt}{d\tau} = \frac{r^2 + a^2}{\Delta}(E(r^2 + a^2) - aL) - a(aE\sin^2\theta - L), \quad (5)$$

$$\Sigma\frac{dr}{d\tau} = \pm\sqrt{\mathcal{V}_r(r)}, \quad (6)$$

$$\Sigma\frac{d\theta}{d\tau} = \pm\sqrt{\mathcal{V}_\theta(\theta)}, \quad (7)$$

$$\Sigma\frac{d\phi}{d\tau} = \frac{a}{\Delta}(\mathcal{E}(r^2+a^2) - a\mathcal{L}) - \left(a\mathcal{E} - \frac{\mathcal{L}}{\sin^2\theta}\right), \tag{8}$$

where $\mathcal{V}_r(r)$ and $\mathcal{V}_\theta(\theta)$, respectively, are related to the effective potentials for radial and polar motion and are given by

$$V_r(r) = ((r^2+a^2)E - aL)^2 - \Delta(K + (aE-L)^2), \tag{9}$$

$$\mathcal{V}_\theta(\theta) = \mathcal{K} - \left(\frac{\mathcal{L}^2}{\sin^2\theta} - a^2\mathcal{E}^2\right)\cos^2\theta. \tag{10}$$

The separability constant $\mathcal{K}$ is related to the Carter (1968) constant of motion $\mathcal{Q} = \mathcal{K} + (a\mathcal{E} - \mathcal{L})^2$, which is essentially a manifestation of the isometry of metric (1) along the second-order Killing tensor field. The black hole shadow boundary requires two dimensionless impact parameters (Chandrasekhar 1985)

$$\xi \equiv \mathcal{L}/\mathcal{E}, \qquad \eta \equiv \mathcal{K}/\mathcal{E}^2, \tag{11}$$

such that each geodesic is characterized by only these two parameters. Depending on the radial effective potential $\mathcal{V}_r(r)$ and the constants of motion, photons may follow scattering orbits, capturing orbits, and unstable orbits at the constant radii. These unstable orbits, constructing a photon region around the event horizon, delineate an apparent boundary separating the dark and bright regions on the observer's sky and account for the optical appearance of the black hole (Cunningham & Bardeen 1973). These orbits can be determined by the unique extremum of the potential in the exterior region to the event horizon r_+ i.e., $r > r_+$, such that

$$\mathcal{V}_r|_{(r=r_p)} = \frac{\partial\mathcal{V}_r}{\partial r}\bigg|_{(r=r_p)} = 0, \text{ and } \frac{\partial^2\mathcal{V}_r}{\partial \mathrm{r}^2}\bigg|_{(r=r_p)} > 0, \tag{12}$$

where r_p is the unstable photon orbit radius. Solving Equation (12) for Equation (9) results in the critical impact parameters (ξ_c, η_c) for the unstable orbits (Abdujabbarov et al. 2016; Kumar & Ghosh 2020)

$$\begin{aligned}\xi_c &= \frac{[a^2 - 3r_p^2]m(r_p) + r_p[a^2 + r_p^2][1 + m'(r_p)]}{a[m(r_p) + r_p[-1 + m'(r_p)]]},\\ \eta_c &= -\frac{r_p^3}{a^2[m(r_p) + r_p[-1 + m'(r_p)]]^2}[r_p^3 + 9r_p m(r_p)^2\\ &\quad +2[2a^2 + r_p^2 + r_p^2 m'(r_p)]r_p m'(r_p)\\ &\quad -2m(r_p)[2a^2 + 3r_p^2 + 3r_p^2 m'(r_p)]],\end{aligned} \tag{13}$$

where $'$ stands for the derivative with respect to the radial coordinate. If we consider the Kerr black hole $(m(r) = M)$, Equation (13) reduces to (Chandrasekhar 1985)

$$\begin{aligned}\xi_c &= \frac{M(r_p^2 - a^2) - r_p(r_p^2 + a^2 - 2Mr_p)}{a(r_p - M)},\\ \eta_c &= -\frac{r_p^3[4Ma^2 - r_p(r_p - 3M)^2]}{a^2(r_p - M)^2}.\end{aligned} \tag{14}$$

In particular, photons with $\eta_c = 0$ form planar circular orbits confined only to the equatorial plane, whereas $\eta_c > 0$ lead to three-dimensional spherical orbits (Chandrasekhar 1985). In rotating spacetimes, photons can either have prograde motion or retrograde motion, whose respective radii at the equatorial plane, r_p^- and r_p^+, can be identified as the real positive roots of $\eta_c = 0$ for $r_p \gtrless r_+$, and all other spherical photon orbits have radii $r_p^- < r_p < r_p^+$. Further, the maximum latitude of spherical orbits depends on the angular momentum of photons, viz., the smaller the angular momentum the larger the orbit latitude. Photons must have zero angular momentum to reach the polar plane of the black hole, whose orbit radius r_p^0, such that $r_p^- \leqslant r_p^0 \leqslant r_p^+$, can be determined by the zeros of $\xi_c = 0$. For the Kerr black hole, these photon orbit radii r_p are (Teo 2003)

$$\begin{aligned}r_p^- &= 2M\left[1 + \cos\left(\frac{2}{3}\cos^{-1}\left[-\frac{|a|}{M}\right]\right)\right],\\ r_p^+ &= 2M\left[1 + \cos\left(\frac{2}{3}\cos^{-1}\left[\frac{|a|}{M}\right]\right)\right],\\ r_p^0 &= M + 2\sqrt{M^2 - \frac{1}{3}a^2}\cos\left[\frac{1}{3}\cos^{-1}\left(\frac{M(M^2-a^2)}{\left(M^2 - \frac{1}{3}a^2\right)^{3/2}}\right)\right],\end{aligned} \tag{15}$$

which for the Schwarzschild black hole $(a = 0)$ takes the degenerate value $r_p^- = r_p^0 = r_p^+ = 3M$ and for the extremal Kerr black hole $(a = M)$ takes the values $r_p^- = M$, $r_p^0 = (1+\sqrt{2})M$ and $r_p^+ = 4M$. Thus the gravitationally lensed image of the photon region around the black hole yields the apparent shadow, whose boundary can be traced by the loci of critical impact parameters (ξ_c, η_c). For an observer at the position (r_o, θ_o), in the far exterior region of the black hole, the shadow boundary can be described in terms of the celestial coordinates (Bardeen 1973; Cunningham & Bardeen 1973)

$$\alpha = -r_o\frac{p^{(\phi)}}{p^{(t)}}, \qquad \beta = r_o\frac{p^{(\theta)}}{p^{(t)}}, \tag{16}$$

where $p^{(\mu)}$ is the photon four-momentum measured on an orthonormal-tetrad basis. On using geodesic Equations (5), (7), and (8), the celestial coordinates yield

$$\begin{aligned}\alpha &= -r_o\frac{\xi_c}{\sqrt{g_{\phi\phi}}(\zeta - \gamma\xi_c)}\bigg|_{(r_o,\theta_o)},\\ \beta &= \pm r_o\frac{\sqrt{\mathcal{V}_\theta(\theta)}}{\sqrt{g_{\theta\theta}}(\zeta - \gamma\xi_c)}\bigg|_{(r_o,\theta_o)},\end{aligned} \tag{17}$$

where

$$\zeta = \sqrt{\frac{g_{\phi\phi}}{g_{t\phi}^2 - g_{tt}g_{\phi\phi}}}, \qquad \gamma = -\frac{g_{t\phi}}{g_{\phi\phi}}\zeta. \tag{18}$$

The (α, β) in Equation (17), respectively, denote the apparent displacement along the perpendicular and parallel axes to the projected axis of the black hole symmetry. For an observer sitting in the asymptotically flat region $(r_o \to \infty)$, the celestial coordinates (Equation (17)) can be simplified as (Bardeen 1973)

$$\alpha = -\xi_c\csc\theta_o, \qquad \beta = \pm\sqrt{\eta_c + a^2\cos^2\theta_o - \xi_c^2\cot^2\theta_o}. \tag{19}$$

The celestial coordinates have the exact same form as that for a Kerr black hole but with different η_c and ξ_c. The photon

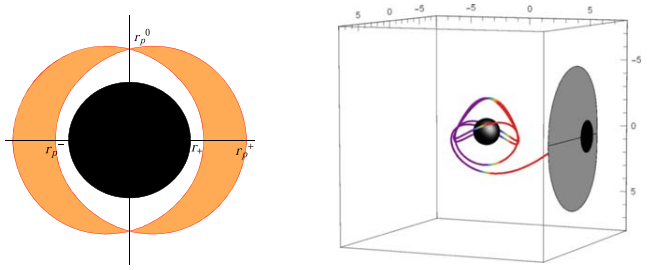


Figure 1. (Left panel) Photon region (shaded orange region) around a Kerr black hole (black disk). (Right panel) Photon orbit projection as a black hole shadow on the observer's image plane. In the projected image, the small dark black disk and gray deformed region, respectively, correspond to the black hole horizon and the shadow.

region and its projected image on the observer's plane for a Kerr black hole are shown in Figure 1. The parametric plot of Equation (17) or (19) in the (α, β) plane can cast a variety of black hole shadows for different choices of black hole mass function $m(r)$ (Abdujabbarov et al. 2016; Amir & Ghosh 2016; Kumar et al. 2019a). Therefore, it is pertinent to check whether shadow observations can be used as a tool to distinguish non-Kerr (rotating regular) black holes from the Kerr black hole or can place constraints on the deviation parameters. Despite that the rotating regular black holes are considerably different from the Kerr black hole, it may be difficult, at least in some cases, to differentiate some rotating regular black holes (e.g., rotating Bardeen black hole) from a Kerr black hole (Li & Bambi 2014; Tsukamoto et al. 2014). In what follows, we explore the possibilities of rotating regular black holes as astrophysical black hole candidates from the comparability of their shadows.

3. Systematic Bias Analysis

We consider the Kerr black hole shadows as the injection and simulated shadows of rotating regular black holes as models to fit with the injection. For a fixed observer position (r_o, θ_o) and a given black hole mass M, the injection solely depends on the black hole spin parameter a, whereas model shadows depend on both spin a and deviation parameter g. This deviation parameter can significantly alter the shape and size of shadows when compared with the Kerr black hole shadow (Abdujabbarov et al. 2016; Amir & Ghosh 2016; Kumar et al. 2019a). Consider two astronomical observables, namely, area A and oblateness D, respectively, given by (Tsupko 2017; Kumar & Ghosh 2020)

$$A = 2\int \beta(r_p) d\alpha(r_p) = 2\int_{r_p^-}^{r_p^+} \left(\beta(r_p)\frac{d\alpha(r_p)}{dr_p}\right) dr_p, \tag{20}$$

$$D = \frac{\alpha_r - \alpha_l}{\beta_t - \beta_b}. \tag{21}$$

Here, A and D, respectively, characterize the size and shape of shadow; a single observable induces degeneracy between the spin parameter and the deviation parameter (Tsukamoto et al. 2014). Oblateness D is defined as the ratio of horizontal and vertical diameters of the shadow as subscripts r, l, t, and b, respectively, standing for the right, left, top, and bottom of the shadow silhouette such that $D \neq 1$ refers to a rotating black hole (Tsupko 2017; Kumar & Ghosh 2020). For a given black hole model, these observables uniquely characterize the shadow over the black hole parameter space (a, g). For a fixed spin parameter, these observables can take considerably different values for model and injection shadows (Abdujabbarov et al. 2016; Amir & Ghosh 2016; Kumar et al. 2019a). It is also likely that a given set of shadow observables can correspond to more than one model shadows, though the values of black hole parameters are different (Kumar & Ghosh 2020). To explore this avenue, we make a systematic bias analysis by minimizing the $\chi^2_{\rm red}$ between the model and the injection to check their compatibility and to constrain the model parameter space (a, g). For fixed values of free parameters mass M, distance d, and inclination angle θ_o, we define the $\chi^2_{\rm red}$ as a function of the spin and deviation parameters

$$\chi^2_{\rm red}(a, g, a_K) = \frac{1}{2}\sum_{i=1}^{2}\left[\frac{\alpha^i(a, g) - \alpha^i_K(a_K)}{\sigma_i}\right]^2. \tag{22}$$

Here, α^i_K and α^i are, respectively, the injection and model shadow observables [D, A], where a_K and a are the injection and model spin parameters. In this model fitting procedure, the nonzero values of $\chi^2_{\rm red}$ quantify the departure of rotating regular black hole shadows from that of Kerr black hole shadows, such that for $\chi^2_{\rm red} \leqslant 1$ the rotating regular black hole shadow can capture the Kerr shadow and both are indistinguishable within the current observational uncertainties. Whereas if $\chi^2_{\rm red} > 1$, then the current shadow observations can easily discern two black holes and constraints can be placed on regular black hole parameters. For the fixed values of the injected spin parameter a_K and the deviation parameter g, the best-fit values of the model spin a by minimizing the $\chi^2_{\rm red}$ between the injection and model shadows are obtained. It can be expected without a priori that for the large values of the deviation parameter g, the extracted best-fit values of the model spin a will significantly deviate from the injected spins a_K. The constraints on the model parameter space (a, g) are obtained for which model shadows at best resemble the injected shadows. The standard deviation σ_i is assumed to be 10% of the range of each observable, which is the current uncertainty in the observational measurements of the EHT (Akiyama et al. 2019a, 2019b, 2019c).

4. Application to Regular Black Holes

The systematic bias analysis, introduced in the previous section, allows us to analyze whether the deviations of rotating regular black holes shadows from the Kerr shadows are large enough to be detectable with the current black hole shadow observations. We examine three well-known rotating regular black holes, viz., Bardeen, Hayward, and non-singular black holes. The shadows of these regular models have received significant attention and their shape and size are considerably different from those of the Kerr black hole shadows (Abdujabbarov et al. 2016; Amir & Ghosh 2016; Kumar et al. 2019a). Henceforth, for our purposes, we assume that the inclination angle is $\theta_o = \pi/2$ and consider the rotating regular black hole shadow as a model to fit with the Kerr black hole shadow injection.

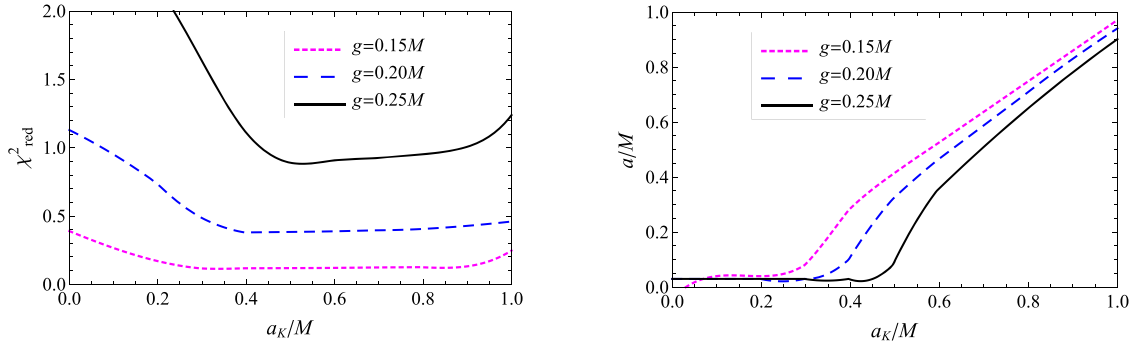


Figure 2. Minimized $\chi^2_{\rm red}$ (left panel) and extracted values of the rotating Bardeen black hole spin parameter a (right panel) for the best-fit model shadow as a function of the injected Kerr spin a_K.

4.1. Bardeen Black Holes

The Bardeen (1968) black hole metric is asymptotically ($r \to \infty$) flat, and near origin ($r \to 0$) behaves as the de-Sitter. The rotating regular Bardeen black hole (Bambi & Modesto 2013; Toshmatov et al. 2014) is described by metric (1) with the mass function (Bardeen 1968)

$$m(r) = M\left(\frac{r^2}{r^2 + g^2}\right)^{3/2}, \tag{23}$$

where the deviation parameter g can be identified as the magnetic monopole charge (Ayon-Beato & Garcia 2000). The rotating Bardeen black hole metric has been tested with the X-ray data from the disk around the black hole candidate in the Cygnus X-1 ($M = 14.8\,M_\odot$, $d = 1.86$ kpc, $\theta_o = 27^\circ.1$; Bambi 2014). Such that the 3σ bounds $a_K > 0.95M$ (Gou et al. 2011) and $a_K > 0.983M$ (Gou et al. 2014) for the Kerr metric infer a bound on Bardeen black hole parameters, respectively, $a > 0.78M$ and $g < 0.41M$, and $a > 0.89M$ and $g < 0.28M$ (Bambi 2014).

We begin with the calculations of the observables A and D for the rotating Bardeen black hole (model) and the Kerr black hole (injection) shadows and investigate whether given shadow observables for the injection shadow can also correspond to the modeled shadow. The model shadow area is found to decrease with both a and g, such that for $a = a_K$, the model shadow area is always smaller than that for the injection shadow area (Kumar & Ghosh 2020). It also turns out that the oblateness D decreases with both a and g (Kumar & Ghosh 2020). The behavior of $\chi^2_{\rm red}$ between the model and injection shadows with varying injected spin a_K is depicted in Figure 2. It is evident that $\chi^2_{\rm red}$ shows a non-monotonic behavior with increasing injected spin, such that $\chi^2_{\rm red}$ is relatively large for the small values of a_K then decreases with increasing a_K and again increases for the near-extremal value of a_K. For sufficiently smaller values of g, $\chi^2_{\rm red}$ is less than 1 for almost all values of injected spin a_K (see Figure 2). Whereas for higher values of g, model shadows could not substantially capture the injected shadows, as the $\chi^2_{\rm red} > 1$ for all a_K. This placed a bound on model parameter g, such that for $g \lesssim 0.26M$ model shadows are indistinguishable from the injected shadows. The best-fit values of the model spin parameter a for different values of g are extracted, which as a function of injected spin a_K are shown in Figure 2. The extracted model spin is biased from the injected spin and this disparity further increases with increasing g. The contour plots of $\chi^2_{\rm red}$ as a function of (a, g) for various values of a_K are shown in Figure 3. One can deduce that for the finite parameter space (a, g), model shadows are well consistent with the injected shadows within the current observational uncertainties. Figure 3 suggests that the parameter space viable for the indistinguishability of two shadows is bounded by $\chi^2_{\rm red} \leqslant 1$ and decreases with increasing a_K, viz., constraints on rotating Bardeen black hole parameters are substantially stronger for rapidly rotating Kerr black hole injections.

4.2. Hayward Black Holes

Another thoroughly studied regular black hole model was proposed by Hayward (2006). Besides the mass M it has one additional parameter ℓ, which determines the length associated with the region concentrating the central energy density, such that modifications in the spacetime metric appear when the curvature scalar becomes comparable with ℓ^{-2}. The spherically symmetric Hayward black hole model is identified as an exact solution of the general relativity minimally coupled to NED with magnetic charge g, where g is related to ℓ via $g^3 = 2M\ell^2$ (Fan & Wang 2016). The rotating Hayward black hole is also a Kerr-like black hole described by the metric (1) with the mass function (Hayward 2006; Bambi & Modesto 2013)

$$m(r) = \frac{Mr^3}{r^3 + g^3}. \tag{24}$$

The photon region and shadows of rotating Hayward black holes have been extensively discussed (Abdujabbarov et al. 2016; Kumar et al. 2019a; Liu et al. 2019).

Using Equation (24) in Equation (13), we construct the shadow and calculate the corresponding observables A and D. To see whether the rotating Hayward black hole (model) shadow can imitate the Kerr black hole (injection) shadow, the $\chi^2_{\rm red}$ is calculated between them. The minimized $\chi^2_{\rm red}$ and the extracted values of the best-fit model spin parameter a as a function of injected spin a_K and for different values of g are shown in Figure 4. The $\chi^2_{\rm red}$ varies nonuniformly with a_K, which infers the degree of similarities between the two shadows. Figure 4 deduces that the model shadow with the small values of g, can well capture the injection, such that $\chi^2_{\rm red} \approx 0$ and the extracted values of model spin are close to the injected spin values. Whereas for moderately high values of g ($g \sim 0.50M$), the model can fit with the injection only for the intermediate values of the injected spin, such that it cannot resemble the shadows of very slowly or rapidly rotating injections. Furthermore, for very large values of g ($g \gtrsim 0.70M$), the two shadows are clearly distinguishable as $\chi^2_{\rm red} > 1$ for all injected spin values ($0 \leqslant a_K \leqslant M$). The

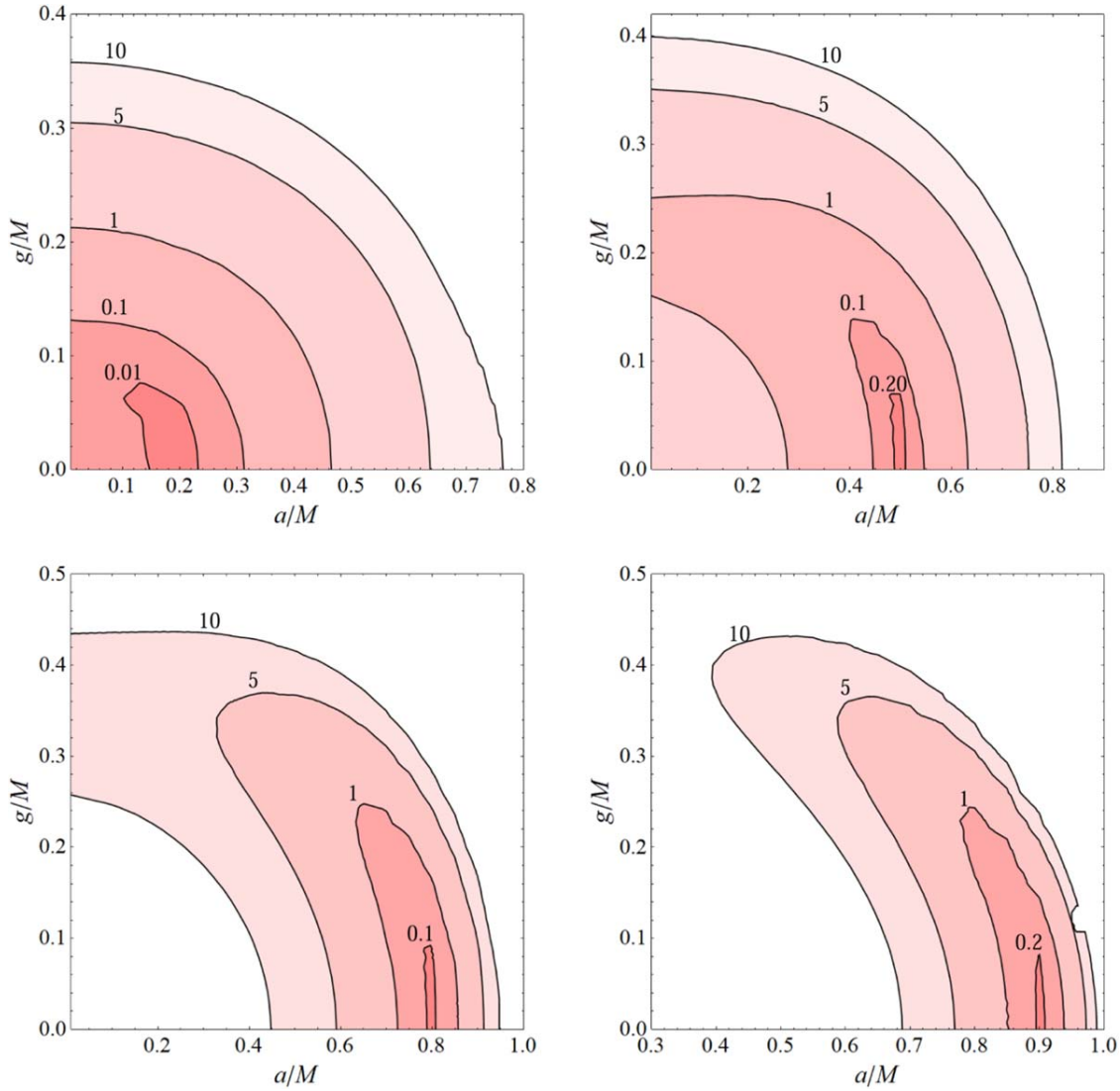


Figure 3. $\chi^2_{\rm red}$ between rotating Bardeen black hole and Kerr black hole shadows as a function of (a, g) for $a_K = 0.20M$, $0.50M$, $0.80M$, and $0.90M$.

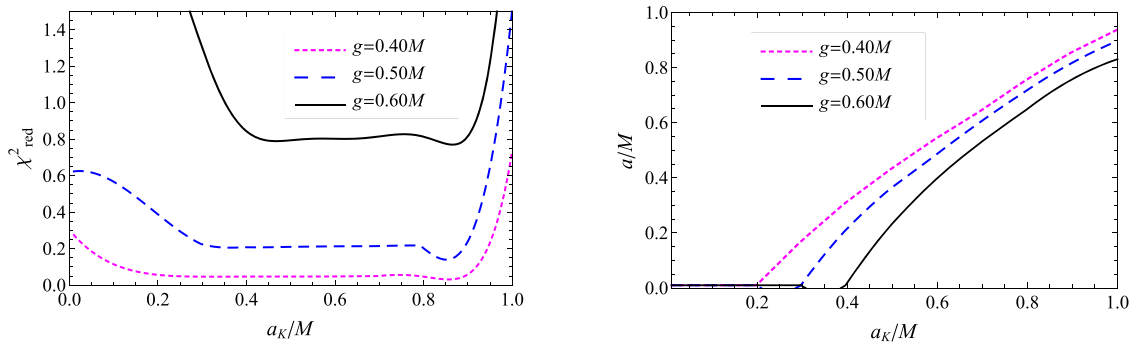


Figure 4. Minimized $\chi^2_{\rm red}$ (left panel) and extracted values of the rotating Hayward black hole spin parameter (right panel) for the best-fit model shadow as a function of Kerr spin a_K.

extracted best-fit values of the model spin are biased from the injected spins, and this disparateness is more prominent for higher values of g. In addition, constraints are placed on the parameters (a, g) for which the model can get fit with the given injection shadow. Figure 5 shows that for a given shadow injection, there is a finite parameter space for which the model shadow is consistent with the injected shadow. This parameter space decreases with increasing injected spin. Thus, the rapidly rotating Kerr black hole shadows can also be interpreted with the rotating Hayward black hole.

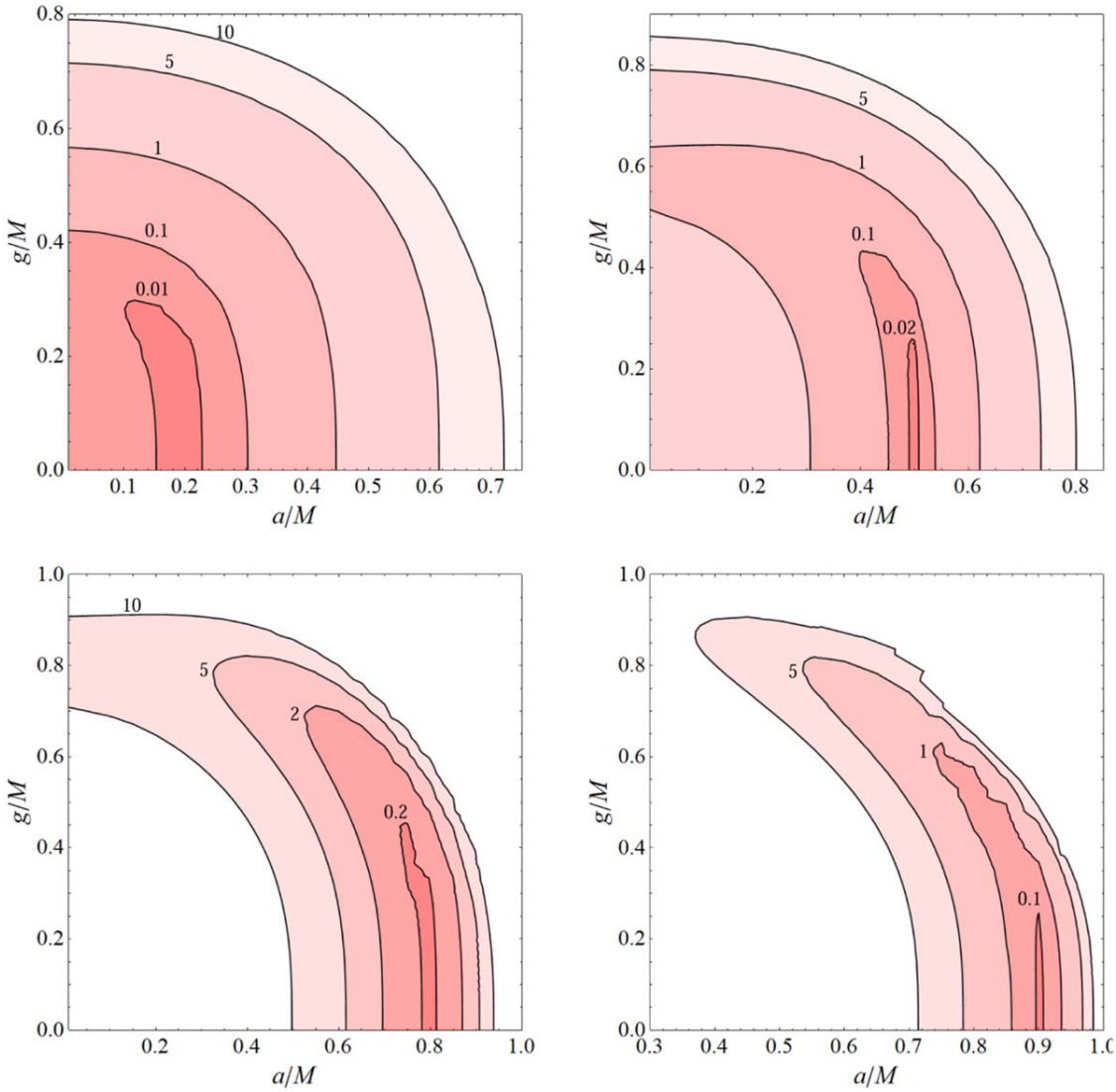


Figure 5. $\chi^2_{\rm red}$ between rotating Hayward black hole and Kerr black hole shadows as a function of (a, g) for $a_K = 0.20M$, $0.50M$, $0.80M$, and $0.90M$.

4.3. Non-singular Black Holes

Bardeen and Hayward regular black holes have an asymptotically de-Sitter core. The next regular model (Culetu 2015; Ghosh 2015), which we call non-singular for brevity, is a novel class of regular black hole with an asymptotic Minkowski core (Simpson & Visser 2020). While these non-singular models share many features with Bardeen and Hayward black holes, there are also notable differences, especially at the deep core (Simpson & Visser 2020). The mass function of rotating non-singular black hole reads (Ghosh 2015)

$$m(r) = Me^{-g^2/2Mr}, \tag{25}$$

where g is the NED charge. As per the previous section, we calculate the observables A and D for the shadows of rotating non-singular (model) and Kerr (injection) black holes.

In turn, the $\chi^2_{\rm red}$ between the model and injection shadows is minimized to determine the best-fit model spin. The behavior of $\chi^2_{\rm red}$ with varying a_K is similar to that for rotating Bardeen and Hayward black holes, as the most substantial gain in this model fitting occurs for the intermediate values of a_K (see Figure 6). The minimized $\chi^2_{\rm red}$ further increase with g. For instance, injected shadow with $a_K = 0.6M$, within the current observational uncertainties, is indiscernible from model shadows with spins $a = 0.542M$, $0.49M$, and $0.417M$, respectively, for $g = 0.15M$,$0.20M$, and $0.25M$. The best-fit model spins are lower from the injected spin, and the biasness is higher for large g. Moreover, the model parameter space (a, g), which minimizes the $\chi^2_{\rm red}$ for $a_K = 0.20M$, $0.50M$, $0.80M$, and $0.90M$, is shown in Figure 7; this infers that model shadows can resemble the injected shadows, i.e., $\chi^2_{\rm red} \leqslant 1$, over the finite parameter space. However, this parameter space (a, g) decreases with increasing injected spin a_K. Therefore, rapidly rotating Kerr black hole injected shadows placed stringent bounds on rotating non-singular black hole parameters, as the near-extremal values of a_K maximize the general-relativistic effects around the black holes.

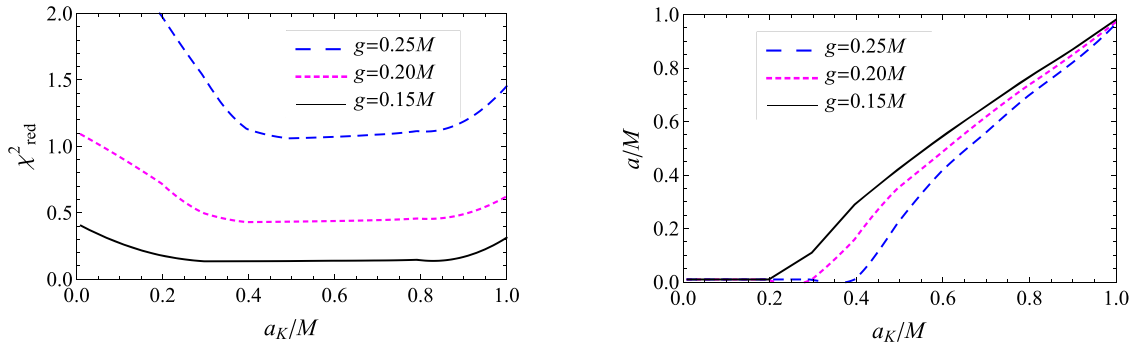


Figure 6. Minimized $\chi^2_{\rm red}$ (left panel) and extracted values of rotating non-singular black hole spin parameter a (right panel) for the best-fit model shadow as a function of Kerr spin a_K.

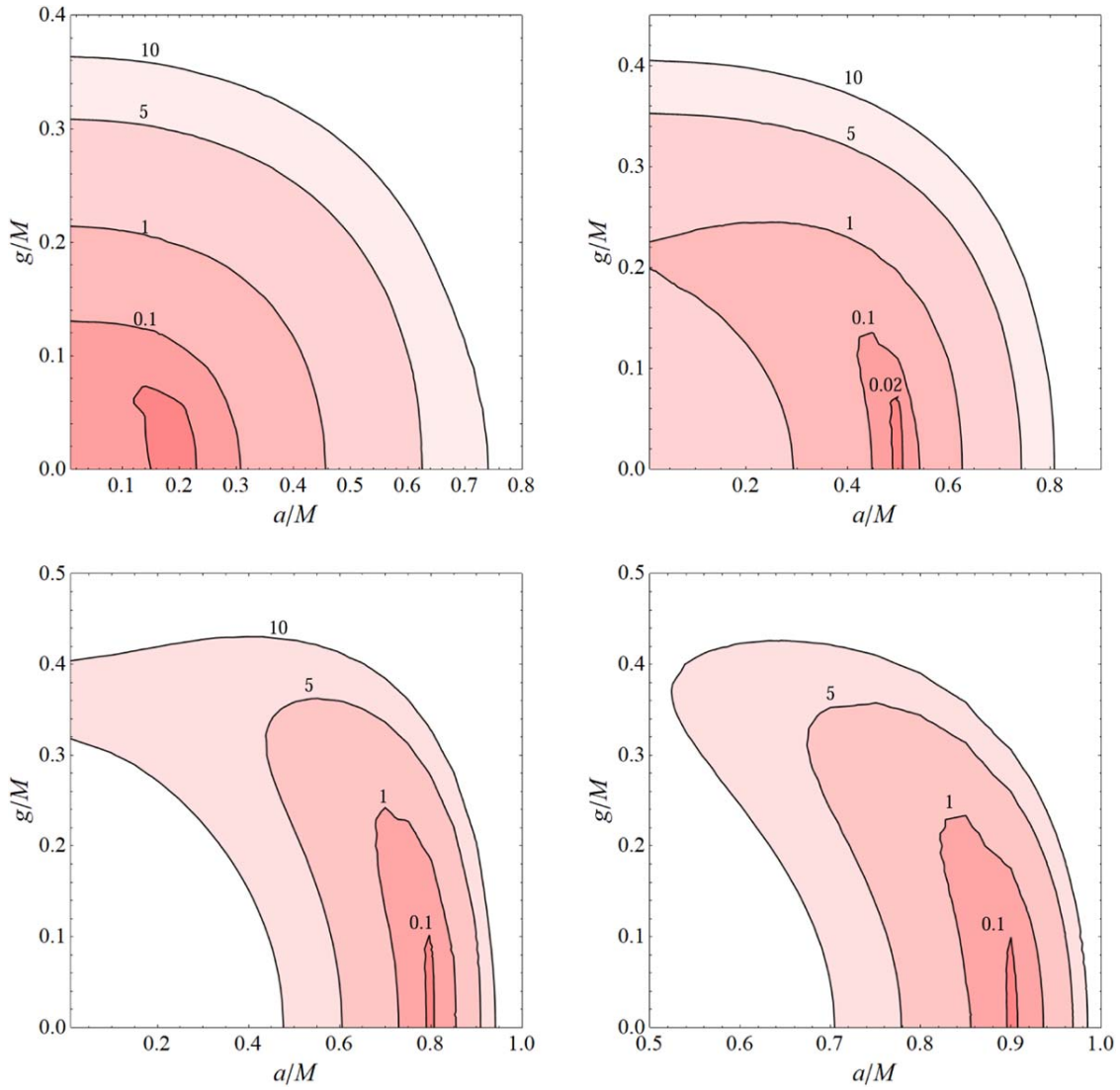


Figure 7. $\chi^2_{\rm red}$ between rotating non-singular black hole and Kerr black hole shadows as a function of (a, g) for $a_K = 0.20M$, $0.50M$, $0.80M$, and $0.90M$.

5. Observational Constraints from the M87* Black Hole Shadow

Recently, the EHT Collaboration, for the first time, has revealed the 1.3 mm image of the supermassive black hole M87* (Akiyama et al. 2019a, 2019b, 2019c). Even though the observed image of M87* black hole is consistent with the predicted shadow for the Kerr black hole, because of the uncertainty in the inferred black hole rotation parameter it may be difficult to ignore modified black hole theories (Akiyama et al. 2019a; Allahyari et al. 2020; Cunha et al. 2019; Kumar et al. 2019a, 2019b; Neves 2020; Vagnozzi & Visinelli 2019),

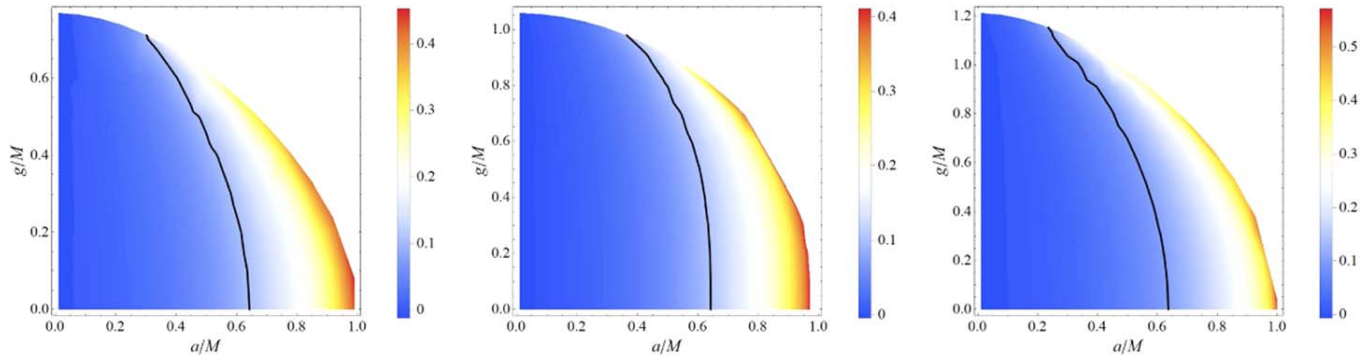


Figure 8. Circularity deviation ΔC as a function of (a, g) for rotating Bardeen (left), rotating Hayward (middle), and rotating non-singular (right) black holes. Black solid line in each plot corresponds to $\Delta C = 0.10$.

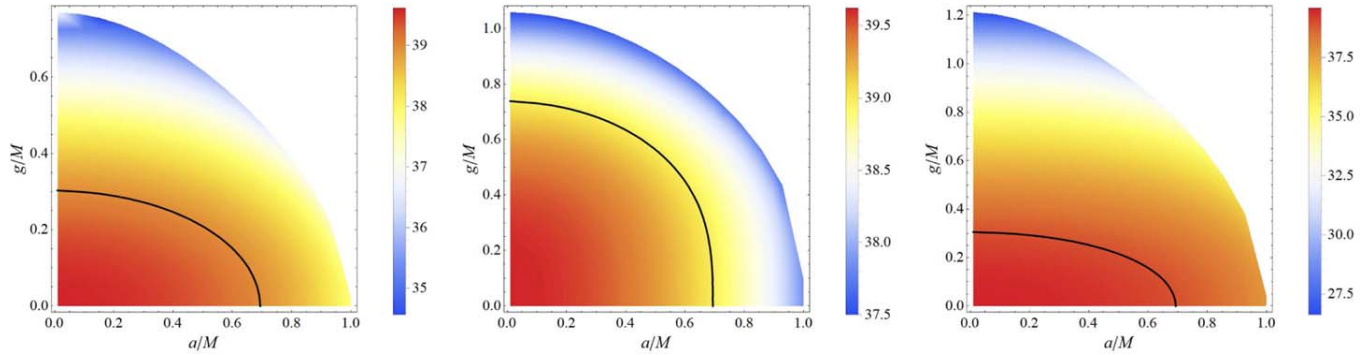


Figure 9. Shadow angular diameter θ_d as a function of (a, g) for rotating Bardeen (left), rotating Hayward (middle), and rotating non-singular (right) black holes. Black solid line in each plot corresponds to $\theta_d = 39\ \mu$as.

e.g., braneworld black holes with negative tidal charge (Banerjee et al. 2019), and superspinors within finite parameter space (Bambi et al. 2019) can explain the M87* black hole shadow observables. We consider the M87* black hole as the rotating regular black hole and used the shadow observables, namely, angular size and asymmetry, to put constraints on the parameter space (a, g) that at best can describe the observed shadow. The shadow is parameterized by $(R(\varphi), \varphi)$, where $R(\varphi)$ is the silhouette radial distance from the shadow center (α_C, β_C), and φ is the angular coordinate. The shadow average radius $\bar{R}$ is defined as (Johannsen & Psaltis 2010)

$$\bar{R} = \frac{1}{2\pi}\int_0^{2\pi} R(\varphi) d\varphi, \tag{26}$$

with

$$R(\varphi) = \sqrt{(\alpha - \alpha_C)^2 + (\beta - \beta_C)^2}, \quad \varphi \equiv \tan^{-1}\left(\frac{\beta}{\alpha - \alpha_C}\right),$$

and (α_C, β_C) can be interpreted as the shadow displacement from the black hole center (0, 0), such that due to the intrinsic axisymmetry the vertical displacement is zero, i.e., $\beta_C = 0$ and α_C is

$$\alpha_C = \frac{|\alpha_r + \alpha_l|}{2}. \tag{27}$$

Furthermore, we consider the circularity deviation ΔC as a measures of the shadow distortion from a perfect circle (Johannsen & Psaltis 2010; Johannsen 2013b)

$$\Delta C = 2\sqrt{\frac{1}{2\pi}\int_0^{2\pi} (R(\varphi) - \bar{R})^2 d\varphi}, \tag{28}$$

clearly $\Delta C = 0$ for non-rotating black holes that have circular shadows. The EHT data analysis reveals that the circularity deviation in the observed image of the M87* black hole is $\Delta C \leqslant 0.10$ (Akiyama et al. 2019a, 2019b, 2019c). The effects of the interplay between spin a and deviation parameter g on the ΔC for rotating regular black holes are analyzed and depicted in Figure 8.

It is clear from Figure 8 that the bound $\Delta C \leqslant 0.10$ merely constrains the rotating regular black hole parameter space (a, g), as all theoretically allowed values of g are permissible for the observed asymmetry of M87*, viz., $g \leqslant 0.7698M$ for the Bardeen black hole, $g \leqslant 1.0582M$ for the Hayward black hole, and $g \leqslant 1.2130M$ for the non-singular black hole. Further, the deduced angular diameter of the M87* black hole shadow can be used to constrain the parameters. The shadow areal radius $R_s = \sqrt{A/\pi}$ corresponds to the angular diameter θ_d in the sky

$$\theta_d = 2\frac{R_s}{d}, \tag{29}$$

which depends on the black hole mass $M = (6.5 \pm 0.7) \times 10^9 M_\odot$ and its distance from the Earth $d = 16.8$ Mpc. The emission ring diameter in the observed M87* black hole shadow is $\theta_d = 42 \pm 3\ \mu$as, though the emission region is not restricted to lie exactly at the photon ring and preferentially falls outside it. The geometric crescent model accounting for

the emission in the general-relativistic-magnetohydrodynamics simulated images and observational uncertainties lead to this offset between the two. The angular diameters of the rotating regular black hole shadows are calculated over the parameter space (a, g), and the results are illustrated in Figure 9. For the aforementioned black hole mass and distance, the non-rotating Schwarzschild black hole ($a = 0$, $g = 0$) cast the largest shadow with angular diameter $\theta_d = 39.6192\ \mu$as, which falls within the 1σ confidence region. The angular diameter θ_d as a function of (a, g) is presented in Figure 9 with the black solid line corresponding to $\theta_d = 39\ \mu$as. As shown in Figure 9, the angular diameter of the rotating regular black holes shadows over a finite parameter space is remarkably consistent with the observed angular diameter of the M87* black hole, within the 1σ confidence level. This in turn strongly constrains the parameter g, viz., $g \leqslant 0.30182M$ for the Bardeen black hole, $g \leqslant 0.73627M$ for the Hayward black hole, and $g \leqslant 0.30461M$ for the non-singular black hole. The rotating regular black holes with a very large value of g cannot account for the observed angular size of M87*. Thus, from Figures 8 and 9, it is evident that the M87* shadow shape and size could be explained with the rotating regular black holes, and they can be strong candidates for the astrophysical black holes.

6. Conclusion

In this paper, we present the systematic bias analysis for testing the rotating regular black hole (non-Kerr) metrics to analyze the deviation of their shadows from that of the Kerr black hole, which is detectable with the EHT observations. The shape and size of the black hole shadow can be very well determined only by the spacetime geometry, the observer's viewing angle, and by identifying the photons captured region, which ascertains that the shadow can be regarded as a potential tool for this. The size and shape of the black hole shadow are characterized by two observables, viz., area A and oblateness D (Kumar & Ghosh 2020). It turns out that the same shadow observable (A, D) can be associated with several black hole models with significantly different parameters (Kumar & Ghosh 2020). In particular, we have investigated whether the black hole shadow, using these observables, can determine if astrophysical black holes are indeed non-Kerr focusing on well-motivated three rotating regular black holes having an additional deviation parameter g due to the NED charge. We analyzed shadows produced by rotating regular black holes and compared them with those for the Kerr black holes using systematic bias analysis. We showed that rotating regular black holes, depending on the values of g, in some cases, cause shadows that are very similar to those produced by the Kerr black holes ($\chi^2_{\rm red} < 1$), but in other cases, the two would be clearly distinguishable ($\chi^2_{\rm red} > 1$). To illustrate on the above point, model shadows, viz., Bardeen black holes ($g \lesssim 0.26M$), Hayward black holes ($g \lesssim 0.65M$), and non-singular black holes ($g \lesssim 0.25M$), can well capture the Kerr black hole shadows within the current observational uncertainties (see Figures 3, 5, and 7). The spin a of the best-fit models, which minimized the $\chi^2_{\rm red}$, are found to be biased from the injected spin values a_K, and that becomes more biased with increasing g. In particular, model shadows with higher g resembled the injected shadows only for the intermediate a_K; therefore, the rotating regular black holes cannot mimic the slowly or rapidly rotating Kerr black hole shadows. Whereas for sufficiently large values of g, model shadows significantly differed from the injected shadows, and the current observational facilities can unambiguously discern ($\chi^2_{\rm red} > 1$) the model with injection shadows.

In turn, the M87* black hole shadow observables (ΔC, θ_d) are found to be consistent with the rotating regular black hole shadows within the finite particular parameter space (a, g) (see Figure 8 and 9). Further, according to our analysis, Bardeen black holes ($g \leqslant 0.30182M$), Hayward black holes ($g \leqslant 0.73627M$), and non-singular black holes ($g \leqslant 0.30461M$), within the 1σ region for $\theta_d = 39\ \mu$as, are consistent with the observed angular diameter of M87* black hole implying that the rotating regular black holes can be strong viable candidates for astrophysical black holes.

It is straightforward to employ the formalism to any other stationary, axisymmetric, and asymptotically flat black hole metric. In particular, it will be physically much more interesting to consider other non-Kerr black hole metrics, arising in modified gravities such as that proposed by Johannsen & Psaltis (2010), to compare with shadows produced by Kerr black holes for getting more insights.

S.G.G. would like to thank DST INDO-SA bilateral project DST/INT/South Africa/P-06/2016, SERB-DST for the ASEAN project IMRC/AISTDF/CRD/2018/000042, and also to IUCAA, Pune for the hospitality while this work was being done. R.K. would like to thank UGC for providing SRF.

References

Abdujabbarov, A., Amir, M., Ahmedov, B., & Ghosh, S. G. 2016, PhRvD, D93, 104004
Abdujabbarov, A., Atamurotov, F., Dadhich, N., Ahmedov, B., & Stuchlik, Z. 2015, EPJC, C75, 399
Akiyama, K., Albderdi, A., Alef, W., et al. 2019a, ApJL, 875, L1
Akiyama, K., Alberdi, A., Alef, W., et al. 2019b, ApJL, 875, L5
Akiyama, K., Alberdi, A., Alef, W., et al. 2019c, ApJL, 875, L6
Allahyari, A., Khodadi, M., Vagnozzi, S., & Mota, D. F. 2020, JCAP, 02, 003
Amarilla, L., & Eiroa, E. F. 2012, PhRvD, D85, 064019
Amarilla, L., Eiroa, E. F., & Giribet, G. 2010, PhRvD, D81, 124045
Amir, M., Ahmed, F., & Ghosh, S. G. 2016, EPJC, C76, 532
Amir, M., & Ghosh, S. G. 2015, JHEP, 07, 015
Amir, M., & Ghosh, S. G. 2016, PhRvD, D94, 024054
Amir, M., Singh, B. P., & Ghosh, S. G. 2018, EPJC, C78, 399
Atamurotov, F., Abdujabbarov, A., & Ahmedov, B. 2013, PhRvD, D88, 064004
Ayon-Beato, E., & Garcia, A. 1998, PhRvL, 80, 5056
Ayon-Beato, E., & Garcia, A. 1999, GReGr, 31, 629
Ayon-Beato, E., & Garcia, A. 2000, PhLB, B493, 149
Ayzenberg, D., & Yunes, N. 2018, CQGra, 35, 235002
Azreg-Ainou, M. 2014, PhRvD, D90, 064041
Bambi, C. 2014, PhLB, B730, 59
Bambi, C., & Barausse, E. 2011, ApJ, 731, 121
Bambi, C., & Freese, K. 2009, PhRvD, D79, 043002
Bambi, C., Freese, K., Vagnozzi, S., & Visinelli, L. 2019, PhRvD, D100, 044057
Bambi, C., & Modesto, L. 2013, PhLB, B721, 329
Banerjee, I., Chakraborty, S., & SenGupta, S. 2019, PhRvD, 101, 041301
Bardeen, J. 1968, in Proc. GR5, 174
Bardeen, J. 1973, in Black Holes, ed. C. DeWitt & B. S. DeWitt (New York: Gordon and Breach)
Beckwith, K., & Done, C. 2005, MNRAS, 359, 1217
Berej, W., Matyjasek, J., Tryniecki, D., & Woronowicz, M. 2006, GReGr, 38, 885
Bronnikov, K. A. 2001, PhRvD, D63, 044005
Bronnikov, K. A. 2017, PhRvD, D96, 128501
Bronnikov, K. A., & Fabris, J. C. 2006, PhRvL, 96, 251101
Burinskii, A., & Hildebrandt, S. R. 2002, PhRvD, D65, 104017
Cardoso, V., & Pani, P. 2019, LRR, 22, 4
Carter, B. 1968, PhRv, 174, 1559

Carter, B. 1971, PhRvL, 26, 331
Chandrasekhar, S. 1985, The Mathematical Theory of Black Holes (Oxford: Oxford Univ. Press)
Culetu, H. 2015, ITJP, 54, 2855
Cunha, P. V. P., Herdeiro, C. A. R., Kleihaus, B., Kunz, J., & Radu, E. 2017, PhLB, B768, 373
Cunha, P. V. P., Herdeiro, C. A. R., & Radu, E. 2019, Univ, 5, 220
Cunningham, C. T., & Bardeen, J. M. 1973, ApJ, 183, 237
Dymnikova, I. 1992, GReGr, 24, 235
Dymnikova, I. 2004, CQGra, 21, 4417
Eiroa, E. F., & Sendra, C. M. 2013, PhRvD, D88, 103007
Falcke, H., & Markoff, S. B. 2013, CQGra, 30, 244003
Fan, Z.-Y., & Wang, X. 2016, PhRvD, D94, 124027
Garcia, A., Hackmann, E., Kunz, J., Lämmerzahl, C., & Macias, A. 2015, JMP, 56, 032501
Ghasemi-Nodehi, M., Li, Z., & Bambi, C. 2015, EPJC, C75, 315
Ghosh, S. G. 2015, EPJC, C75, 532
Ghosh, S. G., & Maharaj, S. D. 2015, EPJC, C75, 7
Gliner, E. B. 1966, JETP, 22, 378
Gou, L., McClintock, J. E., Reid, M. J., et al. 2011, ApJ, 742, 85
Gou, L., McClintock, J. E., Remillard, R. A., et al. 2014, ApJ, 790, 29
Grenzebach, A., Perlick, V., & Lämmerzahl, C. 2014, PhRvD, D89, 124004
Hawking, S. W. 1972, CMaPh, 25, 152
Hayward, S. A. 2006, PhRvL, 96, 031103
Hioki, K., & Maeda, K.-i 2009, PhRvD, D80, 024042
Israel, W. 1967, PhRv, 164, 1776
Israel, W. 1968, CMaPh, 8, 245
Johannsen, T. 2013a, PhRvD, D87, 124017
Johannsen, T. 2013b, ApJ, 777, 170
Johannsen, T. 2016, CQGra, 33, 124001
Johannsen, T., & Psaltis, D. 2010, ApJ, 718, 446
Junior, E. L. B., Rodrigues, M. E., & Houndjo, M. J. S. 2015, JCAP, 1510, 060
Jusufi, K., Ovgun, A., Saavedra, J., Vasquez, Y., & Gonzalez, P. A. 2018, PhRvD, D97, 124024
Kerr, R. P. 1963, PhRvL, 11, 237
Konoplya, R. A., & Zhidenko, A. 2019, PhRvD, D100, 044015
Kumar, R., & Ghosh, S. G. 2020, ApJ, 892, 78
Kumar, R., Ghosh, S. G., & Wang, A. 2019a, PhRvD, D100, 124024
Kumar, R., Singh, B. P., & Ghosh, S. G. 2019b, arXiv:1904.07652
Lamy, F., Gourgoulhon, E., Paumard, T., & Vincent, F. H. 2018, CQGra, 35, 115009
Li, Z., & Bambi, C. 2014, JCAP, 1401, 041
Liu, C., Ding, C., & Jing, J. 2019, SCPMA, 62, 10411
Long, F., Wang, J., Chen, S., & Jing, J. 2019, JHEP, 10, 269
Mishra, A. K., Chakraborty, S., & Sarkar, S. 2019, PhRvD, D99, 104080
Neves, J. C. S. 2020, EPJC, 80, 343
Newman, E. T., Couch, R., Chinnapared, K., et al. 1965, JMP, 6, 918
Ovgun, A. 2019, PhRvD, D99, 104075
Perlick, V., Tsupko, O. Yu., & Bisnovatyi-Kogan, G. S. 2018, PhRvD, D97, 104062
Robinson, D. C. 1975, PhRvL, 34, 905
Rodrigues, M. E., & Junior, E. L. B. 2017, PhRvD, D96, 128502
Ryan, F. D. 1995, PhRvD, D52, 5707
Sajadi, S. N., & Riazi, N. 2017, GReGr, 49, 45
Sakharov, A. D. 1966, JETP, 49, 345
Schee, J., & Stuchlik, Z. 2015, JCAP, 1506, 048
Shaikh, R. 2019, PhRvD, D100, 024028
Simpson, A., & Visser, M. 2020, Univ, 6, 8
Stuchlik, Z., & Schee, J. 2014, IJMPD, D24, 1550020
Teo, E. 2003, GReGr, 35, 1909
Toshmatov, B., Ahmedov, B., Abdujabbarov, A., & Stuchlik, Z. 2014, PhRvD, D89, 104017
Toshmatov, B., Stuchlik, Z., & Ahmedov, B. 2018, PhRvD, D98, 028501
Tsukamoto, N., Li, Z., & Bambi, C. 2014, JCAP, 1406, 043
Tsupko, O. Yu 2017, PhRvD, D95, 104058
Vagnozzi, S., & Visinelli, L. 2019, PhRvD, D100, 024020
Wang, H.-M., Xu, Y.-M., & Wei, S.-W. 2019, JCAP, 1903, 046
Wang, M., Chen, S., & Jing, J. 2018, PhRvD, D98, 104040
Will, C. M. 2006, LRR, 9, 3
Yumoto, A., Nitta, D., Chiba, T., & Sugiyama, N. 2012, PhRvD, D86, 103001

Nanoscale Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Check for updates

Cite this: DOI: 10.1039/d0na00309c

The prospective application of a graphene/MoS_2 heterostructure in Si-HIT solar cells for higher efficiency†

Chandra Kamal Borah,[a] Pawan K. Tyagi[bc] and Sanjeev Kumar *[a]

The efficiency of a Si-HIT (heterojunction with intrinsic thin layer) solar cell based on a graphene/MoS_2 heterostructure has been optimized by varying the various parameters of graphene (Gr) as a transparent conducting electrode (TCE) and n-type molybdenum disulfide (n-MoS_2) as an emitter layer. The photovoltaic performance of a graphene/n-MoS_2/a-Si:H/p-cSi/Au single facial HIT solar cell has been studied using AFORS-HET v2.5 simulation software. A maximum output efficiency of 25.61% has been achieved. The obtained results were compared with the results from a commercially available a-Si:H layer and p-cSi wafer after simulation. Moreover, the dependence of the cell performance on changes in the TCE and the back contact materials has also been studied. Finally, it has been demonstrated that the graphene layer and n-MoS_2 layer could act as a TCE and an efficient emitter layer, respectively, in a n-MoS_2/p-cSi based HIT solar cell.

Received 18th April 2020
Accepted 8th June 2020

DOI: 10.1039/d0na00309c

rsc.li/nanoscale-advances

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

1. Introduction

The output efficiency (η) of a simple silicon (Si) heterojunction solar cell degrades owing to the presence of a large number of defects at the surface of the Si wafer. This deficiency of heterojunction solar cells can be overcome by the introduction of a thin intrinsic layer between the two doped layers of the heterojunction. Such a structure is known as a heterojunction with intrinsic thin layer (HIT) structure.[1] In a HIT structured solar cell, the intrinsic layer-like a-Si:H reduces the interfacial recombination state density to the minimum level between two doped layers[2] and reduces the leakage current.[3] Recently, SANYO Ltd reported attaining a 20% output efficiency in HIT solar cells based on n-type Si.[4] Furthermore, Panasonic has achieved the world's highest efficiency of 25.6% in a HIT cell based on an n-type c-Si.[5]

In a solar cell, a transparent conductive electrode (TCE) is required to ensure the easy transportation of carriers from the emitter layer to the external circuit. The low sheet resistance (10 Ω sq^{-1}) and high optical transparency (85%) mean that indium tin oxide (ITO) is a commonly used TCE in photovoltaic (PV) applications. However, the high cost, rarity and mechanical rigidity of ITO have limited its use for future flexible devices and these drawbacks initiated the search for new TCE materials.[6] Recently, pristine graphene has been reported to have an optical transparency as high as 97.7% (single layer graphene) with a work function of 4.31–4.5 eV (ref. 7 and 8) and a sheet resistance of 120 Ω sq^{-1} in few-layer graphene.[6] In addition, the intercalated graphene with ferric chloride is reported to have a high optical transparency of approximately 96% with a low sheet resistance of 8.8 Ω sq^{-1} and a high carrier density of 8.9×10^{14} cm^{-2} at room temperature and a high stability in air.[6] At room temperature, the carrier mobility in the graphene layer exceeds 10^5 cm^2 V^{-1} s^{-1}.[9] These extraordinary properties have led to the application of multilayer graphene as a TCE. Recently, Lazar *et al.*[7] reported that doping with boron (B) and $-NH_2$ can increase the work function of graphene (Φ_{Gr}) to 1.20 eV and 0.46 eV from 4.31 eV (pristine graphene). Kim *et al.*[8] demonstrated that, after being irradiated by an α beam the work function of graphene increases to 4.9 eV. Conversely, the low photoresponsivity, poor external quantum efficiency, and low absorption of light reduces the application of graphene as an emitter layer instead of the TCE layer in HIT solar cell.[10] Unlike graphene, another 2D material, molybdenum disulfide (MoS_2) has emerged as a complement 2D material for graphene as an emitter layer in HIT solar cells. MoS_2 has a mobility of 200 cm^2 V^{-1} s^{-1} in the monolayer and approximately 517 cm^2 V^{-1} s^{-1} in few layers, as well as a high on/off ratio in the order of 10^8 (ref. 11 and 12) and a fast photo-response time of 40 μs.[13] MoS_2 is reported to have an indirect bandgap of 1.2 eV in the bulk form and a direct bandgap of 1.9 eV in the single layer form.[14] Moreover, double layer MoS_2 has more than 85% optical transparency in the visible region.[15] Recently, it has been reported that the work function of MoS_2 changes with a change in the number of layers.[16] It has also been demonstrated that MoS_2

[a]Centre of Advanced Research, Department of Physics, Rajiv Gandhi University, Arunachal Pradesh-791112, India. E-mail: sanjeev.kumar@rgu.ac.in
[b]Department of Physics, Central University of Haryana, Haryana-123029, India
[c]Department of Applied Physics, Delhi Technological University, Delhi-110042, India

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0na00309c

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

can be used to form significant n-type doping with a Benzyl Viologen (BV) dopant[17] and polyethyleneimine (PEI)[18]. PEI reduces the sheet resistance of MoS_2 2.6 times.[18] These properties mean that MoS_2 can be utilized as an emitter layer in MoS_2/silicon HIT solar cells.

Experimentally and theoretically, it has been demonstrated that MoS_2 can act as an emitter layer in heterojunction silicon solar cells. Tsai *et al.*[19] fabricated a heterojunction solar cell of monolayer MoS_2 and p-cSi and achieved a power conversion efficiency of 5.23%. Hao *et al.*[20] also fabricated a bulk p–n junction after depositing a 40 nm thick layer of n-MoS_2 onto p-cSi substrates and a 1.3% photoconversion efficiency was achieved.

The high transparency, as well as the high mobility of the graphene layer with a tunable bandgap and work function, along with practical and theoretical demonstrations of graphene/MoS_2 and MoS_2/Si heterojunctions, motivated us to propose a single facial HIT solar cell structure of graphene/n-MoS_2/a-Si:H/p-cSi/Au. Here, we have used graphene as a TCE material. No reports have been published on the use of graphene as a TCE in HIT solar cells so far. In addition to graphene and MoS_2, there are a few interesting candidates such as carbon nanotubes, black phosphorous and MXene, that show potential for use in silicon solar cells. So far, a power conversion as high as 17.2% has been observed in carbon nanotube based silicon solar cells.[21–23] After proposing the structure of the cell, the influence of various parameters of the layers used on the cell performance were studied using AFORS-HET v2.5 simulation software. The main objective of the present simulation is to study the independent effects owing to the variation of the parameters of the layers on the cell performance and to demonstrate the prospective application of graphene as an effective TCE material in a n-MoS_2/p-cSi HIT solar cell.

2. Physical model, structural model and formation of the junctions

2.1 Physical model

AFORS-HET is solar cell simulation software, used to study and analyze the various properties of a heterojunction solar cell. For the proposed structure, this software solves Poisson's equation and the transport equation for electrons and holes in one dimension. These equations are as follows:

$$\frac{\varepsilon_0\varepsilon_r}{q}\frac{\partial^2\varphi(x,t)}{\partial x^2} = p(x,t) - n(x,t) + N_D - N_A + \sum_{\text{defects}}\rho_t(x,t)$$

$$-\frac{1}{q}\frac{\partial j_n(x,t)}{\partial x} = G_n(x,t) - R_n(x,t) - \frac{\partial}{\partial t}n(x,t)$$

$$\frac{1}{q}\frac{\partial j_p(x,t)}{\partial x} = G_p(x,t) - R_p(x,t) - \frac{\partial}{\partial t}p(x,t)$$

In which, n represents the electron density, p is the hole density, φ is the electric potential, q is the electron charge, $\varepsilon_0\varepsilon_r$ is the absolute/relative dielectric constant, and $N_{D/A}$ represents the donors/acceptors concentrations, which are assumed to be completely ionized. Here, n, p and φ are the independent variables for which the system of differential equations is solved.

The charge stored in the defects is described by a distribution function, f_t, specifying the probability that defects with a defect density N_t are at position E if the bandgap is occupied with electrons:

$$\text{Acceptor-type defect}: \quad \rho_t(x,t) = -\int dE f_t(E,x,t)N_t(E)$$

$$\text{Donor-type defect}: \quad \rho_t(x,t) = -\int dE[1 - f_t(E,x,t)]N_t(E)$$

The electron/hole currents ($j_{n/p}$) flow owing to the gradient of the corresponding quasi-Fermi energy ($E_{F_{n/p}}$). Within the semiconducting layer, this is equivalent to the sum of the diffusion and drift current with the corresponding mobility $\mu_{n/p}$.

$$j_n(x,t) = q\mu_n n(x,t)\frac{\partial E_{F_n}(x,t)}{\partial x} = -\frac{\mu_n kT}{q}\frac{\partial n(x,t)}{\partial x} + \mu_n n(x,t)\frac{\partial\varphi(x,t)}{\partial x}$$

$$j_p(x,t) = q\mu_p p(x,t)\frac{\partial E_{F_p}(x,t)}{\partial x} = -\frac{\mu_p kT}{q}\frac{\partial p(x,t)}{\partial x} - \mu_p p(x,t)\frac{\partial\varphi(x,t)}{\partial x}$$

If the heterostructure is illuminated (specifying the spectral distribution $\phi(\lambda)$ of the incoming photon flux), the super-bandgap optical generation rate $G_n = G_p\left(\text{for } \frac{hc}{\lambda} \geq E_g\right)$ from the valence band into the conduction band of the semiconductor layers can be evaluated by considering the Lambert–Beer absorption model by specifying the spectral absorption coefficient α, as well as specifying the dielectric properties (n,k) of each layer. An optical sub-bandgap generation $\left(\text{for } \frac{hc}{\lambda} < E_g\right)$ from a defect to the conduction/valence band can be defined by specifying the optical emission coefficients $e_n^o(E)$,$e_p^o(E) \neq 0$ for the defect state.

Recombination from the conduction band into the valence band may occur directly (band to band recombination, Auger recombination) and *via* trap states (Shockley–Read–Hall recombination, (SHR)).

$$R_{n,p}(x,t) = R_{n,p}^{\text{Band–Band}}(x,t) + R_{n,p}^{\text{Auger}}(x,t) + R_{n,p}^{\text{SRH}}(x,t)$$

The transportation across the heterojunction interface is modeled by drift-diffusion. In order to do so, an interface layer is considered. The specified interface defects are distributed homogeneously within this layer. The cell currents have to be

 This journal is © The Royal Society of Chemistry 2020

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

evaluated directly from the gradient of the corresponding quasi-Fermi energies. The electron affinity (χ), the bandgap (E_g) and the effective conduction/valence band density of states (N_C, N_V) will depend on the position x within the interface layer.

$$E_{F_n}(x) = -q\chi(x) + q\varphi(x) - kT \ln \frac{n(x)}{N_C(x)}$$

$$E_{F_p}(x) = -q\chi(x) + q\varphi(x) + kT \ln \frac{n(x)}{N_V(x)}$$

The electric potential is fixed to zero at one contact. At the second contact, a boundary condition has to be specified, which relates the external cell voltage/current to the internal quantities. Furthermore, the electron and hole currents moving into the metal contacts need to be modeled. The Schottky contact can either be voltage controlled, or current controlled.

2.2. Structural model

Layered structure MoS_2 can be directly deposited on Si substrates. However, owing to their structural mismatch (*i.e.* layered structure of MoS_2 and diamond-like structure of Si), large quantities of lattice defects are produced and this increases the recombination at the interface. Thus, the introduction of an intrinsic (i) buffer layer a-Si:H could be beneficial to balance the carrier injection and reduce the leakage current

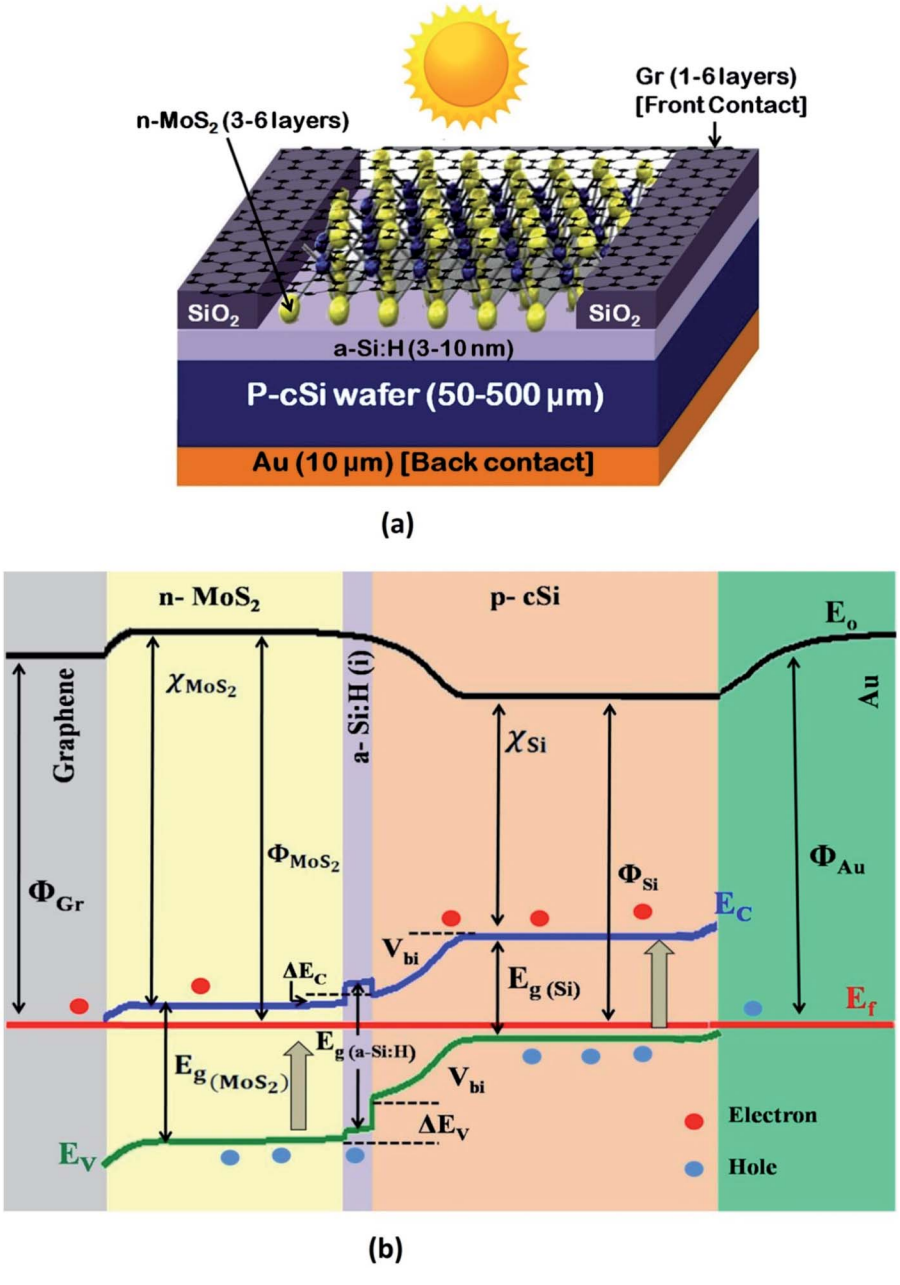


Fig. 1 (a) The proposed HIT solar cell structure, and (b) band diagram as generated using AFORS-HET software for the proposed structure.

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

in the cell.[3] Therefore, the as-synthesized MoS_2 is an n-type semiconductor[24] and p-type silicon is more popular in the PV industries.[4] Thus, in this simulation, we have used p-cSi as a base material. Moreover, the minority carrier diffusion length in p-cSi is also reported to be higher than n-cSi.[25] Furthermore, Fan *et al.*[26] have reported that ultrathin gold (Au) contacts provide a better electrical conductivity and optical reflectance-scattering to maintain the performance of the cell compared with ITO. Also, it has been demonstrated that an Au/Si interface formed by using novel transfer printing technology makes excellent electrical contact and exhibits a low contact resistance in comparison to that formed by vacuum deposition of Au on Si.[27] Therefore, the proposed structure of the single facial HIT solar cell is configured as graphene/n-MoS_2/a-Si:H(i)/p-cSi/Au, as shown in Fig. 1. No lattice mismatch between the layers has been considered for this simulation. Ideally, a semiconductor is called 2D if it satisfies the condition $\langle t \rangle \leq \lambda_F$, in which $\langle t \rangle$ is the average system of the 2D electron system, λ_F is the Fermi wavelength which is equal to $\sqrt{\frac{2\pi}{n_{2D}}}$ nm, and n_{2D} is the carrier density in cm^{-2}.[28] Following this condition as reported by Sarma *et al.*[29] and Suh *et al.*[30] even $n \sim 10^5$ cm^{-3} in graphene and $n \sim 1.8 \times 10^{14}$ cm^{-3} in MoS_2 cannot satisfy the above described conditions respectively. It has also been reported that doping in 2D materials increases the interlayer spacing.[31] Thus, we consider graphene and MoS_2 quasi 3D in nature instead of 2D. The quantum confinement in the structure was achieved by controlling the layer number of n-MoS_2. In the proposed HIT structure, as shown in Fig. 1a the carrier transport occurred along the *c*-axis. Details of the front contact and the back contact parameters of the proposed structure are given in Table 1[6,7,26,32–37] and the parameters, as well as the ranges of the n-MoS_2, a-Si:H and p-cSi are given in Table 2. The thickness of the single graphene was taken as 0.334 nm and the monolayer MoS_2 as 0.65 nm.[19,38,39] However, we started our simulation from five-layer (5L) graphene (1.67 nm) and six-layer (6L) MoS_2 (3.9 nm) by considering the high overlapping probability of the graphene and p-cSi. It was also reported that the Raman spectrum of 6L MoS_2 is very close to the bulk[37] and the optical transparency of 5L graphene (88.5%) is more than that of ITO (85%).[6] The parameters considered for optimization of the MoS_2 layer were the donor concentration (N_D), N_C/N_V, E_g, layer thickness, χ, dielectric constant (ε_r) and electron mobility (μ_n). The values of these parameters were kept within range and these and the results of the above mentioned previous literature are tabulated in Table 2.[3,40–47] To simulate and optimize the other parameters the bulk and interfacial defect distribution and recombination rates were kept constant in the software. Furthermore, the MS-Schottky contact model was selected for the graphene/n-MoS_2 and p-cSi/Au interfaces and the drift-diffusion model was selected for the n-MoS_2/a-Si:H and a-Si:H/p-cSi interfaces. The values and parameters of these interfaces[48] are listed in Table 3. The values of these parameters were varied accordingly under air mass 1.5 (AM1.5) illuminations under a power density of 100 mW cm^{-2} and a temperature of 300 K. In order to study the effect of the layer parameters on the cell performance, the open-circuit voltage (V_{OC}), short circuit current density (J_{SC}), fill factor (FF) and efficiency (η) were analyzed. After optimizing the parameters of n-MoS_2 further simulation was extended to the a-Si:H layer and the thickness and bandgap were optimized. The values and ranges of these parameters were obtained from previously published studies.[4,49–51] Further simulations were performed to optimize the parameters, such as the N_A, N_C/N_V, χ, and thickness for p-cSi. The values and the ranges of the parameters for the p-cSi layer were taken from the previously published research.[4,47*and*48] The defect properties were kept as the default. Finally, we optimized the parameters of the contact materials, that is graphene, Au, ITO, Ag, and so forth. The work function (Φ) and the number of layers were optimized and other parameters were fixed, as found in the literature.[7,8,34–37,52]

Table 1 Front and back contact parameters

	Front contact			Back contact	
Parameter	Initially	After optimization		Initially	After optimization
Material	Graphene	ITO	Al	Au	Ag
Thickness	1L–5L (1L = 0.334 nm)	80 nm	80 nm	10 μm	10 μm
Optical properties	$n = 2.7$ $k = 1.45$	ITO *nk* (default)	$n = 1.19$ $k = 7.05$	Au *nk* (default)	Ag *nk* (default)
Work function (eV)	4.31 4.50 4.54 4.77 4.90	4.4 4.45 4.5	4.06 4.20 4.26	5.4	4.64 4.74
Absorption loss	0.023 (constant)	ITO abs (default)	0.056	0	0
External reflection	1L = 0 2L = 0.023 3L = 0.046 4L = 0.070 5L = 0.092	ITO Ref (default)	0.868	0	0
Surface condition	Plane	Plane	Plane	Plane	Plane
Internal reflection	0	0	0	0	0

 This journal is © The Royal Society of Chemistry 2020

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Table 2 Parameters of the layers

Parameter	n-MoS_2	a-Si:H(i)	p-cSi
Thickness	3L–6L (1L = 0.65 nm)	3–10 nm	50–500 μm
Dielectric constant (ε_r)	4–14	11.9	11.9
Electron affinity χ (eV)	3.74–4.45	3.9	4.05
Band gap, E_g (eV)	1.41–1.48	1.6–2.0	1.12
Effective conduction band density, N_C (cm^{-3})	3×10^{18} to 9×10^{20}	2.5×10^{20}	3×10^{19} to 1×10^{21}
Effective valence band density, N_V (cm^{-3})	3×10^{18} to 9×10^{20}	2.5×10^{20}	3×10^{19} to 1×10^{21}
Effective electron mobility, μ_n ($cm^2\ V^{-1}\ s^{-1}$)	517	20	1104
Effective hole mobility, μ_p ($cm^2\ V^{-1}\ s^{-1}$)	8.5	5	420
Acceptor concentration N_A (cm^{-3})	0	0	10^{15} to 10^{17}
Donor concentration N_D (cm^{-3})	10^{12} to 10^{18}	0	0
Electron thermal velocity, V_e (cm s^{-1})	10^7	10^7	10^7
Hole thermal velocity, V_h (cm s^{-1})	10^7	10^7	10^7
Layer density, (g cm^{-1})	5.06	2.328	2.328
Auger electron recombination coefficient, ($cm^6\ s^{-1}$)	$\sim 10^{-24}$	0	2.2×10^{-31}
Auger hole recombination coefficient, ($cm^6\ s^{-1}$)	$\sim 10^{-24}$	0	9.9×10^{-32}
Band-to-band recombination coefficient, ($cm^3\ s^{-1}$)	$\sim 10^{-7}$	0	9.5×10^{-15}
Defect type	Single	Conduction tail	Single
Defect charge	Acceptor	Acceptor	Acceptor
Total defect density (cm^{-3})	10^{18}	6.4×10^{19}	10^{10}
Specific defect density ($cm^{-3}\ eV^{-1}$)	10^{18}	1.83×10^{21}	10^{10}
Defect level energy (eV)	0.6	0.035 (Urbach)	0.56
Electron and hole capture cross section (cm^{-2})	10^{-14} (default)	7×10^{-16}	10^{-14}
Optical properties	$n = 4.47$; $k = 1.01$	a-Si *nk* (default)	c-Si *nk* (default)

Table 3 Interfacial Parameters

Interface	Numerical model	Remarks
Graphene/n-MoS_2	MS-Schottky contact	Surface recombination velocity for electrons and holes = 10^7 cm s^{-1} (default)
p-cSi/Au		
n-MoS_2/a-Si:H	Drift diffusion model	Interface charge = 10^{11} cm^{-2}
a-Si:H/p-cSi		

2.3 Formation of the junctions

If n-MoS_2 and p-cSi are brought into contact, a heterojunction forms and bending of the energy bands occur close to the Fermi level. This built-in potential (V_{bi}) is developed at the MoS_2/p-cSi interface. This voltage facilitates the separation of photo-generated carriers in the solar cell. Generally, in the p–n junction solar cell, the interfacial defect density reduces the overall efficiency of the cell. Therefore, better surface passivation between the two layers is required to reduce the defect density at the interface. Here, an intrinsic layer a-Si:H is inserted between n-MoS_2 and p-cSi. This layer can be deposited on Si using the plasma-enhanced chemical vapor deposition method at a low temperature (<200 °C) without degrading the bulk properties of the Si substrate.[4] Also, it introduces band offsets between the active and emitter band which are also beneficial to obstructing the reverse flow of the charge carriers. The schematic energy band diagram of the proposed HIT solar cell is shown in Fig. 1b. The build-in potential (eV_{bi}) in the HIT solar cell is defined as:

$$\begin{aligned} eV_{bi} &= E_{g(MoS_2)} + \Delta E_V + kT \ln\left(\frac{N_A N_D}{N_{V(Si)} N_{C(MoS_2)}}\right) \\ &= E_{g(Si)} - \Delta E_C + kT \ln\left(\frac{N_A N_D}{N_{V(Si)} N_{C(MoS_2)}}\right) \end{aligned} \quad (1)$$

In which e, E_g, ΔE_C and ΔE_V, k and T are the electric charge, bandgap, the conduction band offset and valence band offset, Boltzmann constant, and the temperature in Kelvin, respectively. N_A and N_D represent the acceptor and donor concentrations and N_C and N_V are the conduction band and valence band density.[53] In this HIT cell, a barrier for hole transportation is built at the interface between n-MoS_2 and the pristine graphene corresponding to the potential energy of eV_{bi} which can be expressed using the following equation:[54]

$$eV_{bi} = e(\Phi_{MoS_2} - \Phi_{Gr}) \quad (2)$$

In which, Φ_{MoS_2} and Φ_{Gr} are the work functions of n-MoS_2 and pristine graphene respectively. As reported in previously published literature,[16and55] the Φ of a six layer MoS_2 (~5.4 eV) is greater than that of pristine graphene (~4.1 eV).

3. Results and discussion

3.1 Consequences of the n-MoS_2 layer

In this section, we have optimized the donor N_D of the n-MoS_2 layer in the range from 10^{12} to 10^{18} cm^{-3}, and other parameters such as the thickness, N_C/N_V, E_g, χ and ε_r are kept at feasible values as shown in Fig. 2a and b. In Fig. 2b, it can be seen that as

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

N_D was increased from 10^{12} cm^{-3} to 10^{16} cm^{-3}, the V_{OC} remained constant at the value of 636.7 mV. At $N_D > 10^{16}$ cm^{-3}, the V_{OC} increased to 646.1 mV as N_D was increased to 10^{18} cm^{-3}. At a low value of N_D, the consistency of V_{OC} indicates that it does not depend on N_D unless the condition $\Delta n_o \gg N_D$ is followed, as reported previously,[56] in which Δn_o represents the excess electron–hole pairs. At $N_D > 10^{16}$ cm^{-3}, the V_{OC} increases owing to enhancement in the eV_{bi} according to eqn (2). As the mobility of the photo-generated carriers remains almost constant even with an increase of N_D, this might keep J_{SC} constant at 32.33 mA cm^{-2}. Similarly, the FF was found to be constant at 76.71% until the N_D reached 10^{16} cm^{-3} and then the FF slightly reduced to 75.73%. This is due to an increase in the recombination rate at a higher N_D (10^{18} cm^{-3}). As we have observed, the N_D controls the V_{OC}, thus the η of the cell remained constant at 15.79% as the N_D increases, and a maximum of $\eta = 15.82\%$ was obtained at $N_D = 10^{18}$ cm^{-3}.

The N_C/N_V is related to the effective mass of the electrons (m_e^*) and holes (m_h^*) as follows: $N_{C/V} \propto 2m_{e/h}^*$. Furthermore, $m_{e/h}^*$ is related to the mobility ($\mu_{e/h}$) as follows: $\mu_{e/h} = \frac{e\tau}{m_{e/h}^*}$. From these two relationships, it can be concluded that with the change in mobility ($\mu_{e/h}$) the conduction and valence band density (N_C/N_V) also changes. It has been reported that the mobility of the charge carriers in MoS_2 changes with the change in the number of layer. MoS_2 has a mobility of 200 cm^2 V^{-1} s^{-1} in the monolayer and approximately 517 cm^2 V^{-1} s^{-1} in few layers.[11] Therefore, after optimization of N_D to 10^{18} cm^{-3}, the N_C/N_V was optimized in the range 3×10^{18} to 9×10^{20} cm^{-3} as shown in Fig. 2c and d. It can be seen in Fig. 2d that the V_{OC} increased from 646.1 to 668 mV with an increase of the N_C/N_V

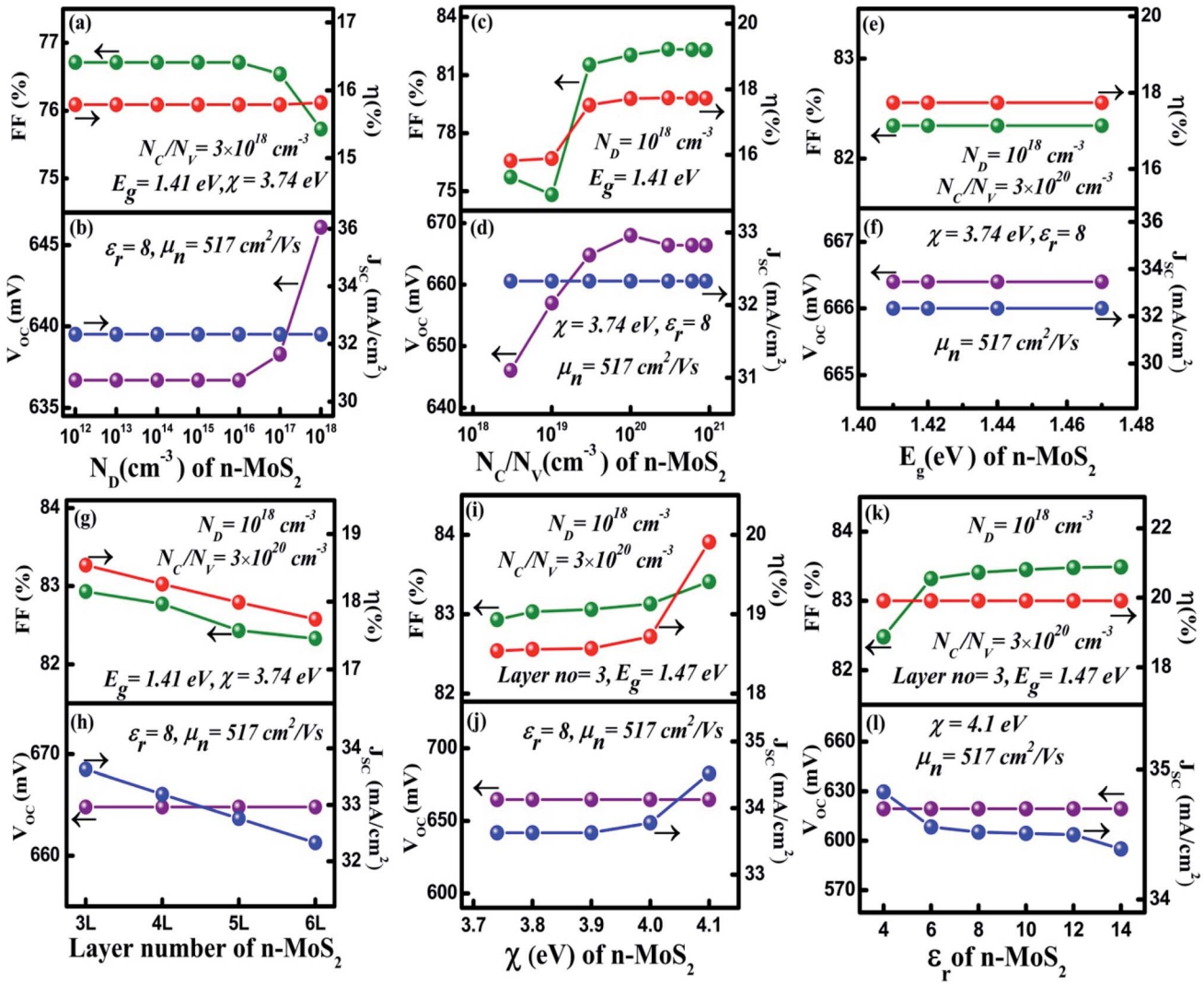


Fig. 2 Optimization of the n-MoS_2 layer: (a) and (b) donor concentration N_D (cm^{-3}), (c) and (d) effective conduction band and valance band density N_C/N_V (cm^{-3}), (e) and (f) the bandgap energy E_g (eV), (g) and (h) the number of layers of MoS_2, (i) and (j) the electron affinity χ (eV), and (k) and (l) the dielectric constant ε_r. The p-cSi parameters were maintained as: thickness = 100 μm, $N_A = 1 \times 10^{16}$ cm^{-3}, $N_C/N_V = 3 \times 10^{19}$ cm^{-3} and $\chi = 4.05$ eV, and the graphene parameters were maintained as: number of layers = 5, reflectance = 0.092 and work function = 4.31 eV. The other parameters are given in Tables 1 and 2.

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

from 3×10^{18} cm^{-3} to 10^{20} cm^{-3}. The increment in V_{OC} can be understood by considering the equation: $V_{OC} = \frac{kT}{q} \ln\left(\frac{np}{{n_i}^2}\right)$ in which n_i is the carrier concentration, n is the intrinsic electron concentration and p represents the intrinsic hole concentration. Here, $n = n_o + \Delta n$ and $p = p_o + \Delta p$, in which n_o and p_o are the concentrations of the electrons and holes in the equilibrium condition and Δn, Δp are the excess photo-generated electron and hole concentrations, respectively. As, ${n_o}{p_o} = {n_i}^2$, $n = N_D$ and $\Delta n = \Delta p$, V_{OC} can therefore be rewritten as $V_{OC} = \frac{kT}{q} \ln\left(1 + \frac{\Delta p}{N_D} + \frac{\Delta p^2}{{n_i}^2} + \frac{\Delta p N_D}{{n_i}^2}\right)$.[57] We can assume that V_{OC} increased with N_C/N_V owing to an increase of Δp and Δn. As shown in Fig. 2d, the V_{OC} decreased a little bit to 666.4 mV as the N_C/N_V increased to 9×10^{20} cm^{-3}, which is due to the decrement of eV_{bi} according to eqn (1). The J_{SC} was found to remain constant at 32.33 mA cm^{-2} with an increase of N_C/N_V from 3×10^{18} to 9×10^{20} cm^{-3}. Although the generation rate (G) of the photo-generated carriers is reported to increase with higher values of N_C/N_V.[58] The constant J_{SC} indicates that the photo-generated carriers either recombine or are trapped close to the metal/semiconductor interface. As with the V_{OC}, the FF is also found to increase and a maximum value of 82.33% was obtained at $N_C/N_V = 3 \times 10^{20}$ cm^{-3}. Therefore, η was found to be enhanced from 15.82% to a maximum value of 17.73%.

After optimization of the N_D and N_C/N_V to values of 10^{18} cm^{-3} and 3×10^{20} cm^{-3}, respectively we varied the bandgap (E_g) in the range 1.41–1.47 eV (ref. 39) and results are shown in Fig. 2e and f. As we can see, all of the cell parameters are constant and a value of $\eta = 17.74\%$ was obtained in this range of E_g.

For the simulation with different numbers of layers of n-MoS_2 as shown in Fig. 2g and h, the V_{OC} was found to remain constant at a value of 664.8 mV, whereas the J_{SC} reduced from 33.63 to 32.33 mA cm^{-2}, when the number of layers was increased from 3 to 6. This happens because the increase in the number of layers reduces the transmittance of photons to reach the junction, which results in a reduction of the electron–hole pair generation rate. The reduction of the transmittance with the increase in the n-MoS_2 layer number leads to a reduction of the FF. The η was found to decrease from 18.54% to 17.74% as the number of layers increased from 3 to 6. Hence, we used 3L as the optimized number of layers of MoS_2 for further simulation.

Fig. 2i and j show the cell parameters for the optimization of χ in the ranges 3.7 to 4.1 eV. Here we have considered $E_g = 1.47$ eV for 3L MoS_2.[42] The V_{OC} remained constant at a value of 664.8 mV when χ (eV) was increased from 3.7 to 4 eV. Upon a further increase in the χ to 4.1 eV, a small increase in the V_{OC} to approximately 691.4 eV was obtained owing to an increase in the eV_{bi}. A constant J_{SC} of 33.63 mA cm^{-2} was found in the range 3.74–3.90 eV of χ and this was enhanced to 34.16 mA cm^{-2} for $\chi = 4.1$ eV. A small increase in the ΔE_C (see Fig. 1b) was observed with the increase in χ, which leads to a slightly higher V_{OC} and J_{SC} at $\chi = 4.1$ eV. The FF was observed to increase from 82.93% (@ $\chi = 3.74$ eV) to 83.41% (@ $\chi = 4.1$ eV) as a result of the reduction in the sheet resistance. Therefore, η was found to increase from 18.54% to 19.19%.

It has been found from electric transportation measurements that the dielectric response of the MoS_2 layers is sensitive to the substrate that is used and the dielectric effect relies on the thickness of the sample.[43] The effect of the dielectric constant (ε_r) on the cell parameters is depicted in Fig. 2k and l. It can be observed in Fig. 2l, that the V_{OC} does not depend on the variation of ε_r and it was found to be a constant at 691.4 eV. However, the J_{SC} shows a light decrement from 34.83 to 34.39 mA cm^{-2} with an increase of ε_r up to 14. In a p–n junction, $eV_{bi} \propto \frac{1}{\varepsilon_r}$, which affect the relative permittivity of the MoS_2 layer.[54,57] The generation of the e–h pair increases in the MoS_2 layer with an increase in the ε_r following Coulombs law as $F \propto \frac{1}{\varepsilon_r}$, in which F is the electrostatic force between the electron and holes, but at the same time the increase in ε_r reduces eV_{bi}. As a consequence of both of these effects, we obtained a constant V_{OC}. In contrast, under short circuit conditions, the photo-generated carriers experienced reduced eV_{bi}, which boosts the recombination of the carriers. Thus, the J_{SC} was found to decrease. The FF showed a minor increase from 82.43% to 83.41% when ε_r was increased from 4 to 6. After this the increase in FF slowed down and only reached 83.49% for $\varepsilon_r = 14$. In Fig. 2k, η was found to largely depend on the V_{OC} and remained constant at 19.91% with ε_r. The cell parameters of the proposed cell were found to be independent of the electron mobility (μ_n) of the n-MoS_2 layer, which has been varied in the range of 517–50 cm^2 V^{-1} s^{-1}.

3.2 Effect of the a-Si:H(i) layer thickness

In the HIT solar cell, an intrinsic layer of a-Si:H is inserted, mainly for passivation of the c-Si surface in the cell.[51] Moreover, the insertion of the thin intrinsic layer also contributes to the depletion region, as well as to the drift current and electric field strength.[59] Fig. 3a and b shows the dependence of the cell parameters on the thickness variation of the a-Si:H(i) layer is in the 3 to 10 nm range. It can be observed in Fig. 3b that along with the increase of a-Si:H(i) thickness from 3 to 4 nm, the V_{OC} was slightly enhanced from 691.4 to 693 mV owing to a minor

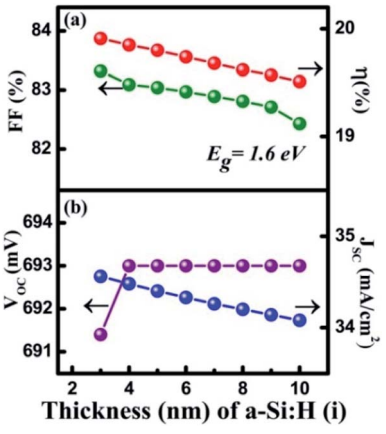


Fig. 3 Optimization of the a-Si:H(i) layer: (a) and (b) thickness (nm) in which E_g was maintained at 1.6 eV.

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

increase in the barrier height at the a-Si:H(i) layer and p-cSi interface.[43] The V_{OC} then became constant at 693 mV upon further increasing the thickness of a-Si:H(i). The J_{SC} was found to decrease linearly from 34.56 to 34.08 mA cm^{-2}. It is reported that a thicker intrinsic layer increases the series resistance in the cell.[50] Therefore, we observed a tiny change in the J_{SC}. Thus, the FF was found to reduce from 83.32% to 82.43%. Consequently, the η obtained reduced from 19.91% to 19.51%. To observe the cell performance for the optimized bandgap energy of a-Si:H(i), a further simulation was carried out by varying the E_g a-Si:H(i) in the range 1.6 to 2 eV. However, the maximum η = 19.19% was only obtained at E_g = 1.6 eV. This might be because the absorption of the photons by the MoS_2 layer only corresponds to this E_g of the a-Si:H(i) layer.

3.3 Optimization of the p-cSi layer

After optimizing the n-MoS_2 and a-Si:H(i) parameters to their optimal values, the simulation was further performed to optimize the parameters of the p-cSi layer, while keeping the parameters of n-MoS_2 and a-Si:H(i) at their optimal values, as discussed in the previous section. The results of the cell performance against variation in the N_A variation of p-cSi are shown in Fig. 4a and b. It is observed in Fig. 4b that when the N_A is increased from 10^{15} to 10^{17} cm^{-3}, the V_{OC} also increased from 691.2 to 699.2 mV owing to the higher eV_{bi} with N_A according to eqn (1). Similarly, J_{SC} was also found to increase from 34.23 to 39.15 mA cm^{-2} as the increased eV_b pushes the photo-generated carriers effectively from the depletion region into the contact region without experiencing a recombination effect or the charge carrier trapped effect in the neutral region. This resulted in an increment in the J_{SC}.[57] The increment in the FF from 81.48% to 82.76% can be attributed to the reduction of the sheet resistance as J_{SC} is also increased. These changes lead to an improvement in η from 19.46% to 21.88%. Hence, $N_A = 10^{17}$ cm^{-3} with a maximum η of 21.88% is considered to be the optimal value for the c-Si layer.

To understand the effect of N_C and N_V on the performance of the HIT solar cell, we varied these parameters within the range of 3×10^{19} to 1×10^{21} cm^{-3} for the p-cSi layer. It can be observed from Fig. 4c and d, with the increase in the N_C/N_V, the V_{OC} decreased linearly from 699.2 to 560.2 mV. According to eqn (1), with the increase in the N_V of p-cSi layer the eV_{bi} decreases and therefore V_{OC} also decreases. In contrast, the value of J_{SC} shows a small increase from 37.81 to 37.91 mA cm^{-2} with N_C/N_V. It has also been reported that N_C/N_V influences the bandgap as well as the absorption of photons.[58] It can be estimated that for $N_C/N_V > 3 \times 10^{19}$ cm^{-3}, the high absorption of photons in the active silicon layer increased the G of electron–

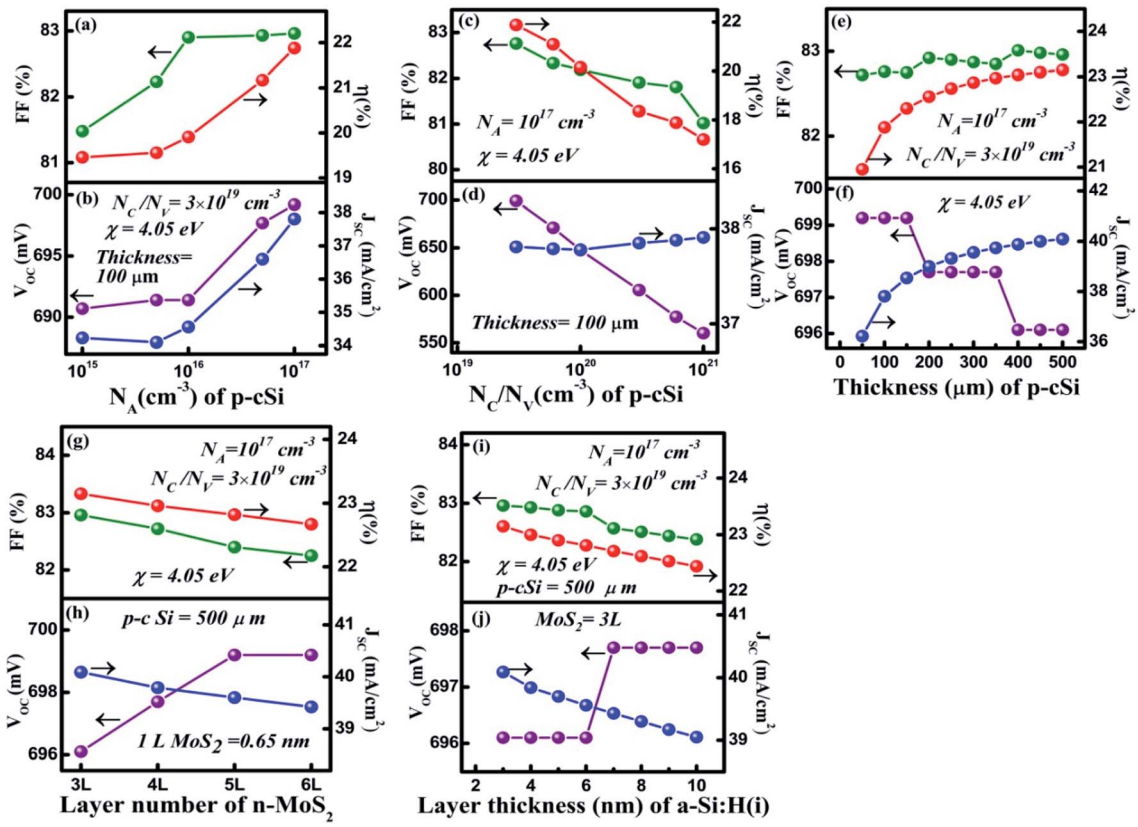


Fig. 4 Optimization of the p-cSi layer: (a) and (b) acceptor concentration N_A (cm^{-3}), (c) and (d) effective conduction band and valance band density N_C/N_V (cm^{-3}), (e) and (f) thickness (μm), (g) and (h) number of layers of MoS_2 and (i) and (j) layer thickness of a-Si:H(i). The optimized parameters of a-Si:H(i) were maintained as: thickness = 3 nm, E_g = 1.6 eV, and the parameters for n-MoS_2 were maintained as: $N_D = 10^{18}$ cm^{-3}, $N_C/N_V = 3 \times 10^{20}$ cm^{-3}, number of layers = 3, E_g = 1.47 eV, χ = 4.1 eV and ε_r = 6.

This journal is © The Royal Society of Chemistry 2020

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

hole pairs. In Fig. 4d, the FF showed a deterioration from 82.76% to 81.02%, which resulted in the η dropping to 17.2% from 21.88%. The maximum value of $\eta = 21.88\%$ was obtained at $N_C/N_V = 3 \times 10^{19}$ cm^{-3}.

In solar cell technology, the thickness of the p-cSi wafer plays an important role in the fabrication of HIT solar cells. Fig. 4e and f represents the variation of cell parameters as a function of the p-cSi wafer thickness from 50 to 500 μm. In Fig. 4f, the V_{OC} remains constant at 699.2 mV. Upon a further increase in the thickness, the V_{OC} reduced sharply to 697.7 mV and again remained constant up to a thickness of 350 μm. Above this thickness, the V_{OC} was found to deteriorate to the value 696.1 mV and becomes independent of the change of thickness. This variation may be due to the variation in the recombination of the photo-generated carriers under open circuit conditions. The observed enhancement in the J_{SC} from 36.22 to 40.09 mA cm^{-2} can be ascribed to the boost in the generation of the charge carriers with the increase in the p-cSi thickness under short circuit conditions. The FF was observed to increase slowly from 82.72% to 82.96% owing to the reduction in the series resistance. These processes help to increase the η from 20.95% to 23.15% at a thickness of 500 μm.

To study the role of the number of layers of n-MoS_2 and the a-Si:H(i) layer thickness with the best optimized p-cSi parameters, the proposed cell was again simulated by varying the number of layers of MoS_2 and the thickness of a-Si:H and the results are depicted in Fig. 4g, h and i, j respectively. From Fig. 4g and h, it can be clearly seen that the V_{OC} increased from 696.1 to 699.2 mV when the number of layers of n-MoS_2 was increased from 3 to 5. Furthermore, the increase in the number of layers to 6V_{OC} remains constant at 699.2 mV. On the other hand, the J_{SC}, FF, and η were observed to decrease with the increase in the number of layers. The J_{SC} decreased from 40.09% to 39.42%, the FF from 82.96% to 82.25 and the η from 23.15% to 22.67%. In Fig. 4i and j, as the layer thickness of a-Si:H(i) was changed to 6 from 3 nm, the V_{OC} gave a constant value of 696.1 mV and increased slightly to 697.7 mV when the thickness was increased to 6 nm. The V_{OC} then became independent of the

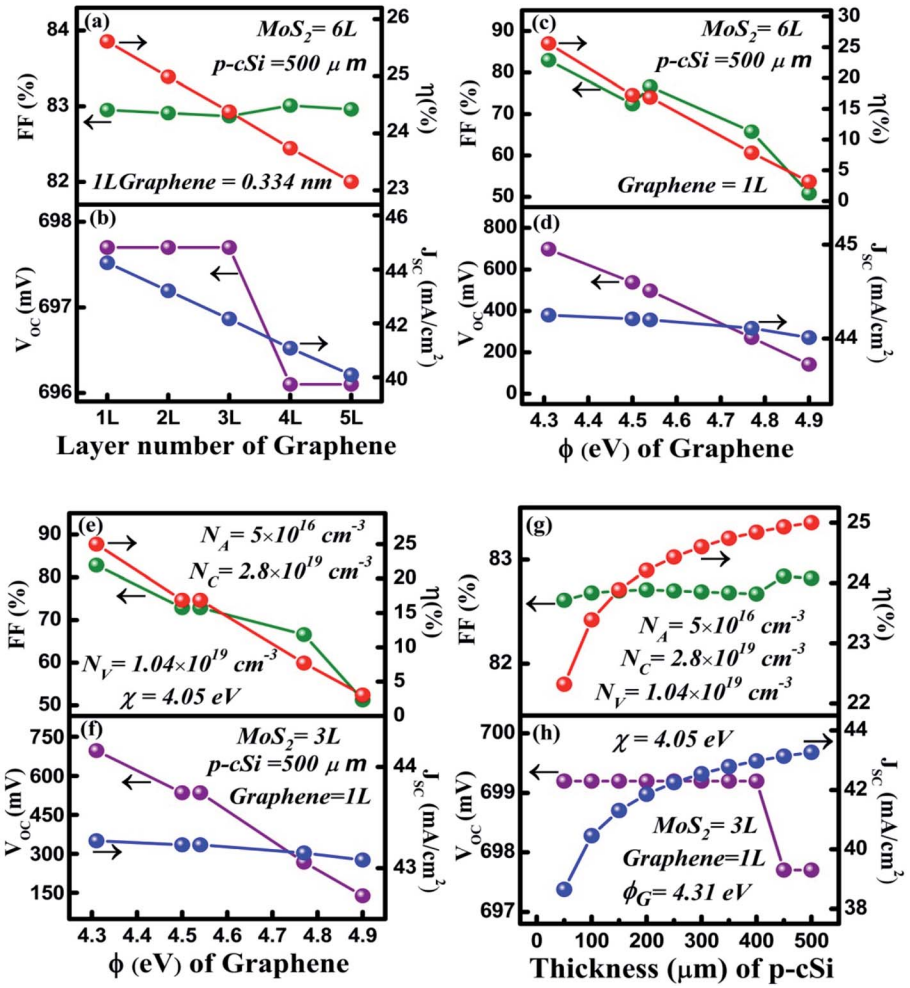


Fig. 5 Optimization of the graphene (front contact) layer: (a) and (b) optimization of the graphene layer number, (c) and (d) optimization of the Φ (eV) of the graphene layer in which the parameters of n-MoS_2 and p-cSi are maintained at their best-optimized values, (e) and (f) optimization of the Φ (eV) of the graphene layer and (g) and (h) optimization of the thickness of a commercially available silicon wafer with practically available p-cSi parameters.

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

thickness even when it was increased to 10 nm. The J_{SC} was found to decline from 40.09 to 39.05 mA cm^{-2} with the thickness of a-Si:H(i). The FF was found to decrease from 82.96% to 82.38% and a minimum η of 22.14% was achieved at 10 nm.

3.4 Simulation and optimization of the front contact (graphene) parameters

In a solar cell, contact plays a transparent conducting electrode role in the solar cell by allowing the light to transmit through and maintain a high electrical conductivity. Fig. 5a and b shows the dependence of the HIT solar cell parameters on the variation of the graphene layer (front contact) number. It is observed in Fig. 5b that with the increase in the number of graphene layers from 1 to 3, the V_{OC} remained constant at 697.7 mV. The reason behind this consistency might be the decrease in the recombination rate which limits the reduction of the charge carrier generation under open-circuit voltage conditions. However, upon further increasing of the graphene layer number to 4, the absorbed light might not be sufficient to produce an efficient electron–hole pair, which causes the V_{OC} to slightly decrease to 696.5 mV and then become constant with an increase in the layer number. On the other hand, the J_{SC} was observed to decline linearly from 44.25 to 40.09 mA cm^{-2}. When the number of layers of graphene was increased the optical transmittance reduced and thus the G of the charge carriers decreased and therefore the J_{SC} reduced too. Under short circuit conditions, the photo-generated carriers could be trapped close to the graphene/n-MoS_2 junction. Therefore, we have observed a linear decrease in the J_{SC}, which also, in turn, reduced the η from 25.61% to 23.15%. An insignificant change in the FF was found from 82.95% to 82.96%.

The presence of a strong covalent bond between the C and Mo atoms or the presence of a dangling bond of sulfur present in the interfaces affects the interfacial barrier height and a variation in the barrier height results in a variation in the contact resistance.[60,61] Usually, the contact resistance is a vital factor that affects the solar cell performance.[62] However, many experimental reports suggest that by controlling the Fermi level (*i.e.* the work function) of graphene, the contact resistance can be minimized in the graphene/MoS_2 interfaces.[63,64] Therefore, in this work, we have focused primarily on the work function of graphene. The depicted results shown in Fig. 5c and d, show the variation of the cell parameters with the variation of the graphene work function (Φ_{Gr}) from 4.31 to 4.9 eV. A large reduction in the V_{OC} was observed in Fig. 5d, in which it decreased from 697.7 to 141.4 mV with an increase of Φ_{Gr}. From eqn (2), it can be observed that with the increase in Φ_{Gr}, the potential barrier energy eV_{bi} of graphene/n-MoS_2 decreases. It has also been reported that when the Φ_{TCO} (contact) of Si HIT solar is low, the eV_{bi} of the TCO (transparent conducting oxide)/n-type emitter has the same direction as that of the eV_{bi} of the n-emitter/p-Si. On the other hand, when the Φ_{TCO} (contact) is high, the eV_{bi} of the TCO/n-type emitter contact will have an inverted direction to that of the eV_{bi} of the n-emitter/p-Si junction. Moreover, with the increase in the Φ of the TCO contact in the silicon HIT solar cell, the depletion region in the emitter region increases and widens and results in the overlap of the contact/n-type emitter region and the n-type emitter/p-type silicon junction region if the emitter is not sufficiently thin enough.[4] Thus, the higher Φ_{Gr} might increase the possibility of overlapping between the graphene/n-MoS_2 depletion region and the n-MoS_2/p-cSi depletion region, which has resulted in a reduction in the V_{OC}. In contrast, owing to the overlap of the graphene/n-MoS_2 depletion region and the n-MoS_2/p-cSi depletion region, the eV_{bi} of n-MoS_2/p-cSi reduces. This results in a reduction of the built-in electric field which is unable to push the photo-generated carriers effectively to their respective sides under short circuit conditions and thus the J_{SC} is observed to decline linearly from 44.25 to 40.09 mA cm^{-2} with an increase in the Φ_{Gr}. The overall reduction in the V_{OC} and J_{SC} reduced η from 25.61% to 23.15%. The maximum η = 25.61% was achieved at Φ_{Gr} = 4.31 eV of pristine graphene.

To validate our simulated results, a comparison has been performed by considering the best-optimized values of n-MoS_2 and a-Si:H(i) with the practically available p-cSi parameters by varying the Φ_{Gr} and p-cSi thickness within the range of 4.31–4.9 eV and the 50–500 μm range respectively, as depicted in Fig. 5e–h and Table 4, using the practically available data for p-cSi.[52,65] It can be seen from Fig. 5e and f, that the maximum η =

Table 4 Summary of the most optimized cell

Cell parameter	Most optimized cell (before optimization of the parameters of the graphene layers)	Graphene layer number optimized cell (after optimization of n-MoS_2 and p-cSi)	n-cSi wafer optimized cell for practically available silicon parameters
V_{OC} (mV)	696.1	697.7	697.7
J_{SC} (mA cm^{-2})	40.09	44.25	43.27
FF (%)	82.96	82.95	82.82
η (%)	23.15 [@ 3L n-MoS_2 at $N_D = 10^{18}$ cm^{-3}, $N_C/N_V = 3 \times 10^{20}$ cm^{-3}, E_g = 1.47 eV, χ = 4.47 eV, ε_r = 6, μ_n = 517 cm^2 V^{-1} s^{-1}, μ_p = 8.5 cm^2 V^{-1} s^{-1}, and @ 6 nm a-Si:H(i) at E_g = 1.6 eV and @ 500 μm p-cSi at $N_A = 1 \times 10^{17}$ cm^{-3}, $N_C/N_V = 3 \times 10^{19}$ cm^{-3}, χ = 4.05 eV and @ 5L graphene at Φ = 4.31 eV]	25.16 [@ 3L n-MoS_2 at $N_D = 10^{18}$ cm^{-3}, $N_C/N_V = 3 \times 10^{20}$ cm^{-3}, E_g = 1.47 eV, χ = 4.47 eV, ε_r = 6, μ_n = 517 cm^2 V^{-1} s^{-1}, μ_p = 8.5 cm^2 V^{-1} s^{-1}, and @ 6 nm a-Si:H(i) at E_g = 1.6 eV and @ 500 μm p-cSi at $N_A = 1 \times 10^{17}$ cm^{-3}, $N_C/N_V = 3 \times 10^{19}$ cm^{-3}, χ = 4.05 eV and @ 1L graphene at Φ = 4.31 eV]	25 [@ 3L n-MoS_2 at $N_D = 10^{18}$ cm^{-3}, $N_C/N_V = 3 \times 10^{20}$ cm^{-3}, E_g = 1.47 eV, χ = 4.47 eV, ε_r = 6, μ_n = 517 cm^2 V^{-1} s^{-1}, μ_p = 8.5 cm^2 V^{-1} s^{-1}, and @ 6 nm a-Si:H(i) at E_g = 1.6 eV and @ 500 μm p-cSi at $N_A = 5 \times 10^{16}$ cm^{-3}, $N_C/N_V = 2.8 \times 10^{19}/1.04 \times 10^{19}$ cm^{-3}, χ = 4.05 eV and @ 1L graphene at Φ = 4.31 eV]

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

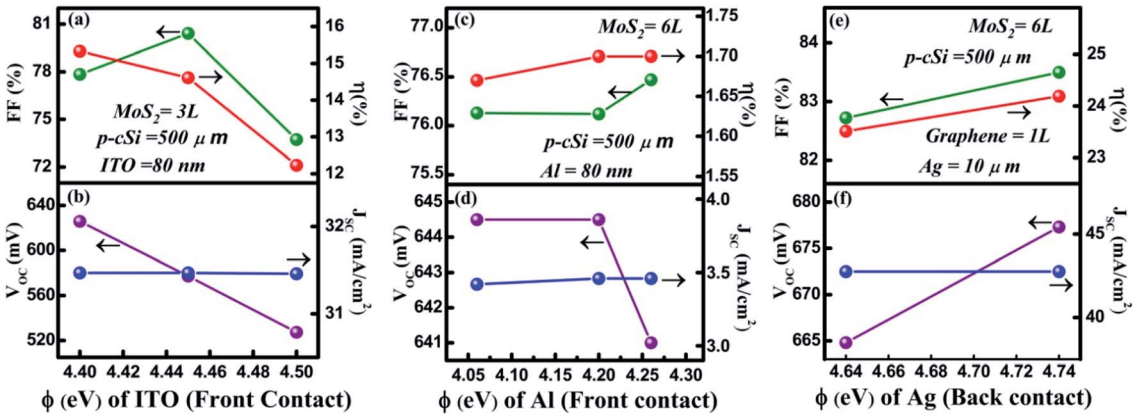


Fig. 6 Changing the front and back contact: (a) and (b) variation in the Φ (eV) of ITO, (c) and (d) variation in the Φ (eV) of Al, and (e) and (f) variation in the Φ (eV) of Ag.

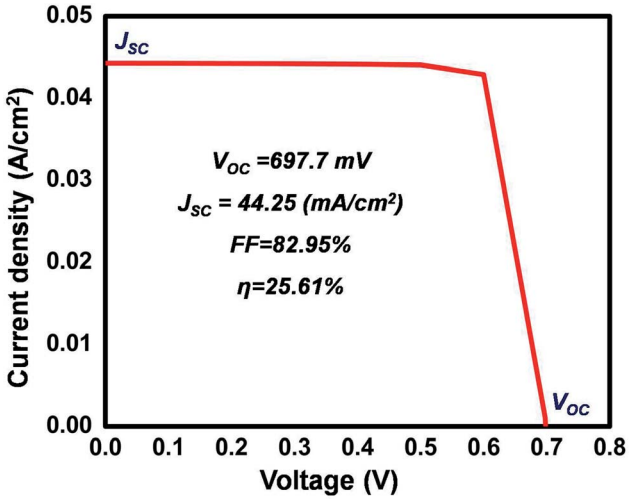


Fig. 7 *IV* characteristics of the most optimized cell.

25% was obtained at $\Phi_{Gr} = 4.31$ eV, which is close to our best-optimized results as shown in Fig. 5c and d. Similarly, in Fig. 5g and h, the same maximum $\eta = 25\%$ was achieved for a 500 μm thick p-cSi wafer.

3.5 Variation and simulation of the front contact and back contact material

In this section, a comparison of the cell parameters has been performed by changing the front contact material graphene to ITO and Al[19] and the back contact to Ag. As seen in Fig. 6a and b the thickness of ITO is considered to be 80 nm and the Φ_{ITO} has been varied from 4.40–4.50 eV.[34] The maximum values for V_{OC}, J_{SC}, FF and η were obtained and found to be 625.8 mV, 31.47 mA cm^{-2}, 80.41% and 15.33% respectively. Similarly, in Fig. 6c and d, when Φ_{Al} was varied, the maximum values for V_{OC}, J_{SC}, FF, and η were found to be 650.8 mV, 4.36 mA cm^{-2}, 82.79% and 2.353%, respectively. In both cases, η was significantly lower than that of η obtained using the graphene layer as a contact.

Fig. 6e and f shows the variation in the cell parameters when the back contact was changed from Au to Ag and the Φ of Ag was maintained from 4.64 to 4.74 eV. It was observed that in this range of Φ for Ag, the V_{OC} increased from 664.8 to 677.3 mV, whereas the J_{SC} remained constant at 42.76 mA cm^{-2}. The FF and η were observed to decrease from 82.72% to 83.5% and 23.51% to 24.19%, respectively. This η was close to our simulated η_{max} value, that is, 25.61% when we used Au as a contact. Here, the back contacts have been considered as the uniform thin film deposited on the backside of p-cSi.

3.6 maximum power and series resistance

The proposed solar cell as shown in Fig. 1a was simulated and characterized under a standard AM1.5G spectrum and forward voltage in the range of 0 to 0.7 V. The *IV* characteristics of the best optimized solar cell generated by the software are shown in Fig. 7. The maximum power delivered by the cell, as well as the

Table 5 Graphene and MoS_2 silicon-based solar cells

Type of study	Solar cell structure	V_{OC} (V)	J_{SC} (mA cm^{-2})	FF (%)	η (%)	Ref.
Experimental	ITO/graphene/MoS_2/n-cSi/(Ti/Ag)	0.59	36.8	73	15.8	66
Experimental	Au/graphene/MoS_2/n-Si/In	0.56	33.4	60	11.1	67
Experimental	Au/graphene/MoS_2/n-Si/Au	0.50	28.1	47	6.56	68
Experimental	Pd/n-MoS_2/i-SiO_2/p-Si/In	0.30	5.5	42	4.5	69
Experimental	Ni/graphene/MoS_2/p-csi/Al	0.51	—	—	2.58	70
Theoretical	TCO/MoS_2/SiGe:H/p-Si/Al	0.652	40.01	83.7	21.85	3
Theoretical	Graphene/n-MoS_2/a-Si:H/p-cSi/Au	0.697	44.25	82.95	25.61	This work

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

series resistance (R_s), was calculated from the *IV* characteristics and found to be 25.4 mW cm^{-2} and 2.5 Ω cm^2 was achieved respectively (see ESI†).

Finally, a comparison between the experimental and theoretical demonstration of graphene, MoS_2, and Si-based solar cells was performed and the results are detailed in Table 5. However, except for our work none of the studies used graphene as a contact material or TCE.

4. Conclusions

In order to develop an efficient HIT solar cell, a graphene/n-MoS_2/a-Si:H(i)/p-cSi/Au structure has been proposed and simulated using AFORS-HET v 2.5 software. The individual effects of the six parameters of the n-MoS_2 layer were analyzed individually. After optimizing the parameters of the n-MoS_2, a-Si:H(i), and p-cSi layers, an optimum efficiency of 23.15% was achieved. As the contact material in a solar cell has an adverse effect on the cell performance, a further simulation was performed to observe the effect of graphene on the proposed HIT solar cell by varying the number of layers and the work function. From this simulation, a maximum efficiency of 25.61% was achieved for single-layer graphene at a work function of 4.1 eV. On the other hand, by changing the front contact graphene to ITO and Al we achieved efficiencies of 15.33% and 2.35%. This confirms that graphene could be an efficient contact material for use in an n-MoS_2/a-Si:H(i)/p-cSi/Au HIT solar cell. Our simulated results establish that the graphene layer could be an effective contact TCE material and n-MoS_2 could be an efficient emitter layer in a single facial n-MoS_2 and p-cSi based HIT solar cell.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors are very thankful to Helmholtz-Zentrum Berlin for providing us with the AFORS-HET simulation software and the Department of Science and Technology (DST), Ministry of India for their financial support.

References

1 N. Dwivedi, S. Kumar, A. Bisht, K. Patel and S. Sudhakar, *Sol. Energy*, 2013, **88**, 31–41.
2 M. Belarbi, M. Beghdad and A. Mekemeche, *Sol. Energy*, 2016, **127**, 206–215.
3 S. Xu, X. Zeng, W. Wang, G. Zhou, Y. Hu and S. Wu, *J. Miner. Mater. Char. Eng.*, 2017, **5**, 323–338.
4 L. Zhao, C. L. Zhou, H. L. Li, H. W. Diao and W. J. Wang, *Sol. Energy Mater. Sol. Cells*, 2008, **92**, 673–681.
5 K. Masuko, M. Shigematsu, T. Hashiguchi, D. Fujishima, M. Kai and N. Yoshimura, *IEEE J. Photovolt.*, 2014, **4**, 1433–1435.
6 I. Khrapach, F. Withes, T. H. Bointon, D. K. Polyushkin, W. L. Barnes, S. Russo and M. F. Craciun, *Adv. Mater.*, 2012, **24**, 2488–2849.
7 P. Lazar, R. Zboril, M. Pumera and M. Otyepka, *Phys. Chem. Chem. Phys.*, 2014, **16**, 14231–14235.
8 J. H. Kim, J. H. Hwang, J. Suh, S. Tongay, S. Kwon, C. C. Hwang, J. Wu and J. Y. Park, *Appl. Phys. Lett.*, 2013, **103**, 171604.
9 A. S. Mayorov, R. V. Gorbachev, S. V. Morozov, L. Britnell, R. Jalil, L. A. Ponomarenko, P. Blake, K. S. Novoselov, K. Watanbe, T. Taniguchi and A. K. Geim, *Nano Lett.*, 2011, **11**, 2396–2399.
10 X. Zhong, W. Zhou, Y. Peng, Y. Zhou, F. Zhou, Y. Yin and D. Tang, *RSC Adv.*, 2015, **5**, 45239.
11 B. Radisavljevic, A. Radenovic, J. Brivio, V. Giacometti and A. Kis, *Nat. Nanotechnol.*, 2011, **6**, 147.
12 B. Radisavljevic and A. Kis, *Nat. Mater.*, 2013, **12**, 815.
13 D. S. Tsai, K. K. Liu, D. H. Lien, M. L. Tsai, C. F. Kang and C. A. Lin, *ACS Nano*, 2013, **7**, 3095–3911.
14 X. Yang, W. Fu, W. Liu, J. Hong, Y. Cai, C. Jin, M. Xu, H. Wang, D. Yang and H. Chen, *J. Mater. Chem. A*, 2014, **2**, 7727.
15 J. M. Yun, Y. J. Noh, J. S. Yeo, Y. J. Go, S. I. Na, H. G. Jeong, J. Kim, S. Lee, S. S. Kim, H. Y. Kim and D. Y. Kim, *J. Mater. Chem. C*, 2013, **1**, 3777.
16 S. Choi, Z. Shaolin and W. Yang, *J. Korean Phys. Soc.*, 2014, **64**, 1550–1555.
17 J. H. Li, D. Lee, D. Qu, X. Liu, J. Ryu and A. Seabaugh, *Nat. Commun.*, 2015, **6**, 6564.
18 Y. Du, H. Liu, A. T. Neal, M. Si and P. D. Ye, *IEEE Electron Device Lett.*, 2013, **34**, 1328–1330.
19 M. L. Tsai, S. H. Su, J. K. Chang, D. S. Tsai, C. H. Chen and C. I. Wu, *ACS Nano*, 2014, **8**, 8317–8322.
20 L. Hao, Y. Liu, W. Gao, Z. Han, Q. Xue and H. Zeng, *J. Appl. Phys.*, 2015, **117**, 114502.
21 D. D. Tune, N. Mallik, H. Fornasier and B. S. Flavel, *Adv. Energy Mater.*, 2020, **10**, 1903261.
22 M. Bat-Erdene, M. Batmunkh, S. A. Tawfik, M. Fronzi, M. J. Ford, C. J. Shearer, L. Yu, M. Dadkhah, J. R. Gascooke, C. T. Gibson and J. G. Shapter, *Adv. Funct. Mater.*, 2017, 1704488.
23 L. Yu, A. S. R. Bati, T. S. L. Grace, M. Batmunkh and J. G. Shapter, *Adv. Energy Mater.*, 2019, 1901063.
24 S. Lin, P. Wang, X. Li, Z. Wu, Z. Xu and S. Zhang, *Appl. Phys. Lett.*, 2015, **107**, 153904.
25 Z. Qiao, J. L. Ji and Y. L. Zhang, *Chin. Phys. B*, 2017, **26**, 068802.
26 L. Fan, F. Wang, J. Liang, X. Yao, J. Fang and D. Zhang, *J. Semicond.*, 2017, **38**, 014003.
27 H. Keum, H. J. Chung and S. Kim, *ACS Appl. Mater. Interfaces*, 2013, **5**, 6061–6065.
28 *Introduction to Quantum Transport in Low-Dimensional Systems*, http://www.hri.res.in/∼nanotr16/notes/KJThomas-1.pdf, accessed July 2018.
29 S. D. Sarma, S. Adam, E. H. Hwang and E. Rossi, *Rev. Mod. Phys.*, 2011, **83**, 407.

 This journal is © The Royal Society of Chemistry 2020

Open Access Article. Published on 23 June 2020. Downloaded on 7/21/2020 10:26:52 AM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

30 J. Suh, T. E. Park, D. Y. Lin, D. Fu, J. Park and H. J. Jung, *Nano Lett.*, 2014, **14**, 6976–6982.
31 Z. Lin, B. R. Carvalho, E. Kahn, R. Lv and R. Rao, *2D Matter.*, 2016, **3**, 022002.
32 A. I. Pantazi and V. Yannopapas, *Opt. Commun.*, 2016, **374**, 24–28.
33 K. Patel and P. K. Tyagi, *Carbon*, 2017, **116**, 744–752.
34 Y. Park, V. Choong, Y. Gao, B. R. Hsieh and C. W. Tang, *Appl. Phys. Lett.*, 1996, **68**, 2699–2701.
35 A. Kahn, Fermi level, work function and vacuum level, *Mater. Horiz.*, 2016, **3**, 7–10.
36 R. M. Eastment and C. H. B. Mee, *J. Phys. F: Met. Phys.*, 1973, **3**, 1738–1745.
37 A. W. Dweydari and C. H. B. Mee, *Phys. Status Solidi A*, 1975, **27**, 223–230.
38 Z. H. Ni, H. M. Wang, J. Kasim, H. M. Fan, T. Yu and Y. H. Wu, *Nano Lett.*, 2007, **7**, 2758–2763.
39 F. J. Nelson, V. K. Kamineni, T. Zhang, E. S. Comfort, J. U. Lee and A. C. Diebold, *Appl. Phys. Lett.*, 2010, **97**, 253110.
40 C. P. Lu, G. Li, J. Mao, L. M. Wang and E. Y. Andrei, *Nano Lett.*, 2014, **14**, 4628.
41 B. Radisavljevic and A. Kis, *Nat. Mater.*, 2013, **12**, 815–820.
42 K. F. Mak, C. Lee, J. Hone, J. Shan and T. Y. Heinz, *Phys. Rev. Lett.*, 2010, **105**, 136805.
43 E. J. G. Santos and E. Kaxiras, *ACS Nano*, 2013, **7**, 10741–10746.
44 S. L. Howell, D. Jariwala, C. C. Wu, K. S. Chen, V. K. Sangwan and J. Kang, *Nano Lett.*, 2015, **15**, 2278–2284.
45 O. Salehzadesh, N. H. Tran, X. Liu, I. Shihand and Z. Mi, *Nano Lett.*, 2014, **14**, 125–4130.
46 H. Zhang, Y. Ma, Y. Wan, X. Rong, Z. Xie and W. Wang, *Sci. Rep.*, 2015, **5**, 8440.
47 W. Zhou, X. Zou, S. Naijmaei, Z. Liu, Y. Shi and J. Kong, *Nano Lett.*, 2013, **13**, 2615–2622.
48 X. Garros, G. Reimbold, J. Cluzel, D. Muñoz and P.-J. Ribeyron, *Microelectron. Eng.*, 2011, **88**, 1247–1250.
49 L. Qin, Y. Xiao-Jun, L. Cheng and C. Ming-bo, *Optoelectron. Lett.*, 2010, **6**, 0108–0111.
50 W. Lisheng and C. Fengxiang, *J. Optoelectron. Adv. Mater.*, 2011, **13**, 81–88.
51 A. Rawat, M. Sharma, D. Chaudhary, S. Sudhakar and S. Kumar, *Sol. Energy*, 2014, **110**, 691–703.
52 L. Zhao, H. L. Li, C. L. Zhou, H. W. Diao and W. J. Wang, *Sol. Energy*, 2009, **83**, 812–816.
53 F. Rana, Semiconductor Heterostructure, in *Semiconductor Optoelectronics*, 2017, ch. 2, https://courses.cit.cornell.edu/ece533/Lectures/handout2.pdf.
54 J. Singh, *Semiconductor devices: Basic Principles*, Wiley India (P) Ltd, New Delhi, 2004.
55 V. K. Thakur and M. K. Thakur, *Chemical Functionalization of Carbon Nanomaterials: Chemistry and Applications*, CRC Press, New York, 2016, p. 918.
56 A. V. Sachenko, V. P. Kostylyov, M. V. Gerasymenko, R. M. Korkishko, M. R. Kulish and M. I. Slipchenko, *Quantum Electronics & Optoelectronics*, 2016, **19**, 67–74.
57 Y. Zhao, C. Liang, M. Sun, Q. Liu, F. Zhang and D. Li, *J. Appl. Phys.*, 2014, **116**, 154506.
58 Y. Zhou and G. Long, *J. Phys. Chem. C*, 2017, **121**, 1455–1462.
59 W. Lisheng, C. Fengxiang and A. Yu, *J. Phys.: Conf. Ser.*, 2011, **276**, 012177.
60 N. Jensen, R. M. Hausne, R. B. Bergmann, J. H. Werner and U. Rau, *Prog. Photovoltaics Res. Appl.*, 2002, **10**, 1–13.
61 X. Liu, J. Gao, G. Zhang and Y. W. Zhang, *Nano Res.*, 2017, **10**, 2944–2953.
62 M. Houssa, K. Iordanidou, A. Dabral, A. Lu, R. Meng, G. Pourtois, V. V. Afanas'ev and A. Stesmans, *Appl. Phys. Lett.*, 2019, **114**, 163101.
63 D. K. Schroder and D. L. Meier, *IEEE Trans. Electron Devices*, 1984, **31**(5), 637–647.
64 Y.-J. Yu, Y. Zhao, S. Ryu, L. E. Brus, K. S. Kim and P. Kim, *Nano Lett.*, 2009, **9**, 3430–3434.
65 H. Wang, L. Yu, Y.-H. Lee, Y. Shi, A. Hsu, M. L. Chin, L.-J. Li, M. Dubey, J. Kong and T. Palacios, *Nano Lett.*, 2012, **12**, 4674–4680.
66 J. Ma, H. Bai, W. Zhao, Y. Yuan and K. Zhang, *Sol. Energy*, 2018, **160**, 76–84.
67 Y. Tsuboi, F. Wang, D. Kozawa, K. Funahashi, S. Mouri, Y. Miyauchi, T. Takenobub and K. Matsuda, *Nanoscale*, 2015, **7**, 14476–14482.
68 K. Jiao, C. Duan, X. Wu, J. Chen, Y. Wang and Y. Chen, *Phys. Chem. Chem. Phys.*, 2015, **17**, 8182.
69 L. Z. Hao, W. Gao, Y. J. Liu, Z. D. Han, Q. Z. Xue, W. Y. Guo, J. Zhu and Y. R. Li, *Nanoscale*, 2015, **7**, 8304–8308.
70 W. Shi and M. Xiying, *Coatings*, 2018, **8**, 2.

24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

TransSketchNet: Attention-based Sketch Recognition using Transformers

Gaurav Jain*, Shivang Chopra*, Suransh Chopra*, and Dr. Anil Singh Parihar

Delhi Technological University, India

Abstract. Sketches have been employed since the ancient era of cave paintings for simple illustrations to represent real-world entities. The abstract nature and varied artistic styling makes automatic recognition of drawings more challenging than other areas of image classification. Moreover, the representation of sketches as a sequence of strokes instead of raster images introduces them at the correct abstract level. However, dealing with images as a sequence of small information makes it challenging. In this paper, we propose a Transformer-based network, dubbed as TransSketchNet, for sketch recognition. This architecture incorporates ordinal information to perform the classification task in real-time through vector images.

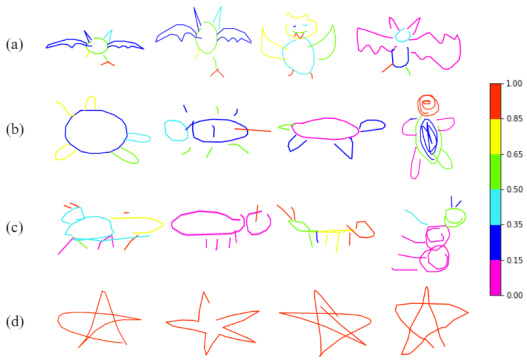


Figure 1. Visualization of few classes in the QuickDraw [2] dataset, (a) Bat, (b) Sea Turtle, (c) Ant, (d) Star. Attention heatmaps depicting the relative importance of strokes while inference.

1 Introduction

The ability to draw and comprehend eclectic notions through sketches reflects the intellectual capabilities of human beings. Drawing has long been a part of human behavior, mutating its utility from rock carvings in the ancient era, to blueprints of drafts in the modern age. In this work, we focus on *sketch recognition*, which aims to identify human drawings and classify them into their respective categories.

Perceiving sketch is a challenging task due to two main reasons, (1) heterogeneous representations and (2) level of abstraction. Representation of the same object is dependent upon the interpretation of that object, which consequently leads to high intraclass variation between samples of the same class. Fig. 1 illustrates this diversity among different samples of the same class. For example, distinct views of a bat in Fig. 1 (a) include different representations which render the task of finding patterns relatively difficult for the models. Alternatively, sketches of one class could be interpreted as belonging to another class. For instance, the second sketch of the sea turtle class in Fig. 1 (b) resembles an ant shown in Fig. 1 (c). Hence, learning such overlapping representations which may exhibit a significantly anomalous, yet acceptable, view proves to be an arduous task.

In order to address the aforementioned issues, a transformer-based approach is proposed for attentive sketch recognition. This is the first approach to the best of our knowledge that employs transformers for sketch recognition. We leverage the attention mechanism of Transformers to identify objects using the vector image format. For this, sketches are interpreted as a sequence of strokes, like humans actually comprehend drawings in real-time.

2 Proposed Methodology

Input Preprocessing: The input data is preprocessed to transform the sketches into the vector-image format. In this, $\mathcal{S}_v$ is a sketch with sequence of strokes s_i, i.e. $\mathcal{S}_v = \{s_1, s_2, ..., s_n\}$, where n is the sequence length. Each stroke, s_i, is defined using a 3-point format:

$$s_i = \{\Delta x_i, \Delta y_i, p_i\}, \forall i \in \{1, 2, ..., n\} \quad (1)$$

where $(\Delta x_i, \Delta y_i)$ is the offset distance in the x and y direction. For each sketch, $(\Delta x_1, \Delta y_1)$ begin with origin as the initial starting point. Pen-state, p_i is a binary variable, with $p_i = 1$ indicating that the pen is in contact with surface, while $p_i = 0$ represents that the pen is lifted off the surface and moved from the previous point in the direction of offsets.

Architecture: Fig. 2 illustrates the proposed architecture, which consists of two modules, (1) *auto-encoder*, and (2) *transformer-encoder*. The *auto-encoder* performs two major functions, (1) extract latent vector, $\mathcal{Z}_v$, and (2) reshape the input dimension to feed a larger context vector to the transformer module effectively. The input $\mathcal{S}_v$ is not directly fed into the auto-encoder. Instead, it is decomposed as $\mathcal{S}_v = \{\mathcal{S}_v^{\Delta}, \mathcal{S}_v^{p}\}$, where $\mathcal{S}_v^{\Delta} = (\Delta x_i, \Delta y_i)$ and $\mathcal{S}_v^{p} = p_i$. We only feed the offsets, $\mathcal{S}_v^{\Delta} \in \mathbb{R}^{2\times n}$, to obtain latent vector, $\mathcal{Z}_v \in \mathbb{R}^{127\times n}$, with a sequence length of n. To preserve the temporal nature of our data, we use *coordinate embedding*, which computes a tuple with (position, time). In the *transformer module*, we use 10 identical transformer blocks connected via skip-connections. Each block consists of multi-head self-attention with $m = 8$ heads. The cus-

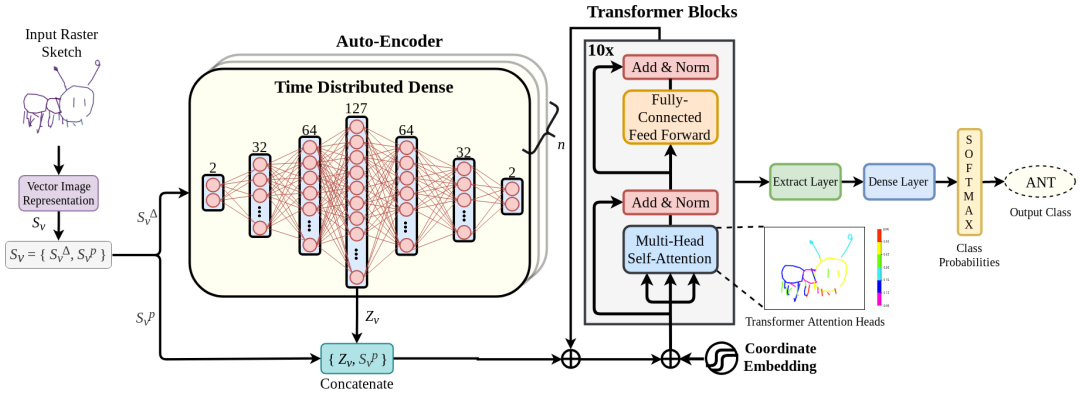


Figure 2. Overview of the proposed Network depicting the pre-processing of raster sketch to form the vector image representation. Projecting the input to a higher latent dimension using the Auto-Encoder and the final classification using the attention maps from the Transformer blocks.

tom *Extract layer* is used to pick the features (ext_{out}) corresponding to the fist timestamp of the transformer output. Further, ext_{out} is passed through a dense layer with a softmax activation function to obtain class probabilities $\Psi = \{\psi_1, \psi_2, ..., \psi_c\}$, where c represents the number of classes. Finally, dense layers with softmax activation facilitate the model to return class probabilities.

Method	5	20	50
HOG-SVM [1]	75.21%	66.79%	63.22%
Fisher-Vectors [5]	79.53%	75.80%	72.90%
AlexNet [4]	77.18%	75.22%	73.06%
Sketch-a-Net v2 [6]	94.78%	88.64%	85.19%
Resnet50-CNN [3]	**96.47%**	90.06%	86.20%
TransSketchNet (Ours)	96.21%	**90.31%**	**88.72%**

Table 1. Comparative evaluation of recognition accuracy on the Quick Draw dataset, with (a) 5 classes, (b) 20 classes, (c) 50 classes. Values in **bold** depict the best accuracy for each subset.

3 Results and Evaluation

The performance of the proposed Transformer-based network is evaluated using the benchmark Quick Draw dataset [2]. Further, attention heatmaps have been plotted to support the use of transformers for sketch recognition. A train-test split of $80-20\%$ is performed with $50,000$ samples per class for training, and $10,000$ samples per class during testing.

Comparative Analysis. We compared the proposed TransSketchNet with three types of approaches, (1) traditional classifiers, (2) CNN-based approaches, and (3) RNN-based approaches. Table 1 reports the recognition accuracy. TransSketchNet outperforms the state-of-the-art approaches for 20 and 50 classes. However, Resnet50-CNN [6] performs marginally better than TransSketchNet with a gain of 0.26% for dataset with 5 classes. For the proposed approach, increasing the number of classes observes a lower drop in accuracy, compared to other approaches. Due to limited resource availability, analysis over the complete dataset could not be performed. To account for this, we randomly selected c classes, where $c \in \{5, 20, 50\}$, and report the averages over 10-fold cross-validation. Quantitative evaluation confirms the effectiveness of representing sketches as vector images. Moreover, this representation facilitates real-time recognition of sketches.

Attention Analysis. The proposed method introduces transformers to exploit the sequence of strokes to draw attention to parts of sketches. Fig. 1 shows few attention heatmaps. For the *bat* class, attention on both the wings are equally focused. This is analogous to how humans recognize bats through their peculiar wing shape. It is interesting to note that for the *star* class, equally high attention is given to the complete structure. Perhaps, structures like stars that are symmetric in nature and constitute a distinct shape are identified based on the overall view of the sketches. Similarly, certain parts of the sketch are more important in recognizing an object; which is supported by the attention heatmaps presented.

4 Conclusion and Future Work

In this work, we proposed TransSketchNet for recognizing sketches using vector images. The autoencoder extracts information from the input data in the form of a latent vector. This renders the input dimension compatibility between the input data and transformer-blocks. Further, transformer-blocks enable the model to focus on characteristic information from each sketch. The proposed approach achieves favorable recognition accuracy when compared with state-of-the-art approaches. In the future, Transformers can be adapted to solve challenging problems in the domain of computer vision.

REFERENCES

[1] Mathias Eitz, James Hays, and Marc Alexa, 'How do humans sketch objects?', *ACM Trans. Graph.*, **31**(4), 44–1, (2012).
[2] David Ha and Douglas Eck, 'A neural representation of sketch drawings', *CoRR*, **abs/1704.03477**, (2017).
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, (2016).
[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'Imagenet classification with deep convolutional neural networks', in *Advances in neural information processing systems*, pp. 1097–1105, (2012).
[5] Rosália G Schneider and Tinne Tuytelaars, 'Sketch classification and classification-driven analysis using fisher vectors', *ACM Transactions on Graphics (TOG)*, **33**(6), 174, (2014).
[6] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, 'Sketch-a-net: A deep neural network that beats humans', *International journal of computer vision*, **122**(3), 411–425, (2017).

Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2020)
IEEE Xplore Part Number:CFP20K74-ART; ISBN: 978-1-7281-4876-2

Understanding and Analyzing Methods of Character Identification in Movies

Prashant Giridhar Shambharkar
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
prashant.shambharkar@dtu.ac.in

Anubhav Gupta
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
anubhavgupta_bt2k16@dtu.ac.in

Anirudh Gupta
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
anirudhgupta_bt2k16@dtu.ac.in

Anshuman
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
anshuman_bt2k16@dtu.ac.in

Dr. M. N. Doja
Department of Computer Engineering, Faculty of Engineering & Technology, Jamila Millia Islamia
Delhi, India
mdoja@jmi.ac.in

***Abstract*— Film and television industry is one of the fastest-growing sectors across the globe. Character identification and its method classification have attracted a lot of attention from researchers because of its applications in video summarization, personalized video retrieval, and semantic video analysis. In this paper, we study and analyze 14 algorithms categorizing them into three categories and compare them based on various parameters and present our findings in a tabular form.**

Keywords—Character identification, graph matching, error-correcting graph matching (ECGM), optical character recognition (OCR), multiple kernel learning (MKL), earthmover distance (EMD)

I. INTRODUCTION

The film and television industry is one of the fastest-growing sectors which leads to the generation of a lot of movie data [1]. The revenue in the global film industry was determined to be 41.7 billion USD as of 2018 [2]. There are more than 6,534,894 titles present on IMDB [3]. Therefore, it is vital to have efficient and useful techniques to manage this data which understands video content and organizes video data. While watching a news broadcast, character names are mentioned, but when you watch a movie, we don't know the names of many characters. Now and then people take third-person references to identify real character names, implying that the audience is more focused on character and its real name while watching a movie. Hence, most important subjects to the index are characters, and analysis of character identification becomes a crucial step [4, 5].

Annotation of characters in a video is called Automatic video annotation. The main objective of automatic video annotation is to identify and annotate characters in the movie, which is known as movie character identification [6]. Character identification is therefore important in many aspects. It labels the character names of people by processing a frame of video one by one. Character identification helps in scene segmentation. A scene is a unit of a story that when combined becomes a film. The scene is segmented as it improves the detection of a character and simplifies understanding. Methods employing character identification helps in movie summarization which helps movie producers to promote the movie. It also helps in getting an overview of the theme of a movie and helps in gaining valuable insight.

Summarization of the movie is based on the analysis of a character and employs a relation between characters to manipulate the structure of a movie [7]. Character identification helps those with hearing impairment by using dynamic captioning. Character identification also helps by providing security in the domain of the banking sector. There is also frequent use of character identification in an MNC by restricting the entry of unauthorized persons. Hence, it is also used in maintaining attendance records in an MNC [6].

Movie character identification has seen broad research efforts and different applications in movies. However, techniques for identifying characters efficiently and effectively are improving day by day. Moreover, every technique has its characteristics and requirements which are applicable in specific scenarios. Therefore, it becomes difficult to study the techniques systematically. This is the problem we seek to address so that it can help individuals to select a technique for their project. We present a comprehensive review and analysis by extensively studying and evaluating some of the most common character identification methods.

In this paper, our focus is on discussing various techniques for Automatic character identification. We have organized the rest of the paper as follows: In section 2, we will give a brief overview of 14 different methods of character identification along with their merits and demerits. In section 3, we analyze the mentioned techniques and present our observations. In section 4, we compare them based on certain criteria. In section 5, we will give the conclusion of our research.

II. CHARACTER IDENTIFICATION METHODS

Existing character identification methods can be divided into three categories:

A. Cast List Based

a) "On affine invariant clustering and automatic cast listing in movies" [8]: A. W. Fitzgibbon and A. Zisserman [8] gave the first unsupervised clustering technique for the visual clustering of characters in movies. This was one of the very first papers to use the clustering technique to detect characters. It was one of the uncomplicated techniques but required quite a lot of time as it involved detecting more than thousands of images with their variances while keeping viewpoint, scale in check [8].

This technique presented a new distance metric which is affine invariant and accurately uses priors. Priors were managed efficiently and thus, showed how the use of priors

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

makes the clustering efficient. "Trust region" is used along with strategies developed by Levenberg-Marquardt in nonlinear optimization [8]. Robustness was achieved by using image-to-image matching, and a band-pass filter was used for changes to observe illumination changes [9]. This technique showed that Clustering invariant is a crucial strategy in further computer vision problems.

Merits:

- It is simple to understand as it used a clustering technique to determine the main characters from a movie automatically.

Demerits:

- Poor tolerance to expression change in characters [8].

*b) "Automatic cast listing in feature-length films with anisotropic manifold space" [9]: The m*ethod mentioned by O. Arandjelovic and R. Cipolla [9] mainly deals with the effect on change in appearance leading to factors such as blur in an image. Its significant contributions are the estimation of appearance manifold from the generic one.

This algorithm majorly utilizes the property of consistency of dissimilarities between appearance manifolds.

Since some people look similar to each other than the rest of them. Reference [9] utilizes this to select and define a structure interpersonal similarity from an unidentified collection of video frames.

The algorithm works by progressively updating generic manifold with generated data so that it would merge to form a data-specific one [9].

The algorithm executes majorly in 3 steps.



Fig. 1. Algorithm steps

Automatic Data Acquisition proceeds by first detecting the transition between shots. Shot transition is detected by using the Edge Change Ratio algorithm [33]. Then the face is tracked from the shots obtained by using the Viola-Jones [34] cascaded algorithm and automatic data acquisition step is completed. Now unsupervised discriminative learning technique such as clustering is used to cluster short sequences by identity. A comparison of two appearance manifolds occurs by frame by frame eliminating extrinsic factors such as illumination. Since many little pose variations may occur, they can be corrected by drawing randomly manifold samples. Then we compare pairwise using the Constraint Mutual Subspace method (CMSM) [10] obtained Normalized Appearance Manifolds. We create a manifold space which is a vector space used to represent an appearance manifold denoted by each point. Gaussian clusters are used to represented classes generated from the given manifold space [9].

Merits:

- Algorithm proposed helps to distinguish when some character looks similar to one another.

Demerits:

- In some cases, the algorithm fails to identify the correct character when a character appears for a brief period.

c) "Leveraging archival video for building face datasets" [11]: The method mentioned by D. Ramanan, S. Baker, and S. Kakade developed a system for finding characters in large actual video collections. Their work is based on two approaches. One is 'Names and Faces' [12], which explains building face datasets automatically by using photos containing captions. The second one is the 'Buffy' system which inherently used automatic face identification in movies [13]. This system also helps in providing one of the most extensive datasets of 611,770 faces. Therefore, it incorporates varying factors such as age, gain, weight, changes in facial features, and other factors that could not have been observed in smaller datasets [11].

The system works in hierarchical stages. First, parts of the face that are in front are combined by building a histogram model constituting hair, face, and overall structure. Dynamic constraints are then added by tracking neighboring frames around the detection of a group. In a scene, the same characters are found to wear the same type of clothes. In the next stage, this fact is used. The tracks obtained earlier can then be grouped using a clustering procedure that takes into the fact that the same body appearance across a scene. In the final stage, a labeled dataset is constructed by manually labeling groups that are small in number. This is done quickly as instead of tracks, they are clusters. After this algorithm becomes the nearest neighbors [11].

Merits:

- Accuracy of 80% with the best half of the dataset.
- Can track, find, and recognize people irrespective of the length of the scene.
- One of the largest database spanning images over 11 years.

d) "Combining generative and discriminative head models" [14]: Since many methods put so far have not been robust in concern with variations in poses and appearance. The method mentioned used considered the issue of identifying a distinct character by forming a classifier from the cast list based on the appearance of a character. 3D model obtained can be obtained in a way giving an estimate from the sample data served. Then the individual is detected by building a tree structure, and the individual's poses are estimated.

The figure below uses a classifier to detect the individual in each step. The pose is evaluated by iterating over a wide range of scales. Refinement of pose occurs after which matching and edge features are used to scale down the effect of lighting and face changes [14].

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

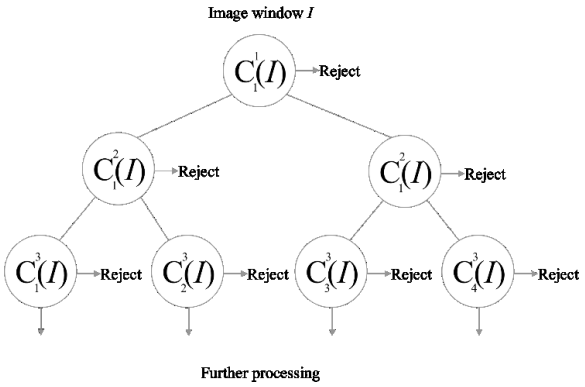


Fig. 2. A framework of pose estimation.

Every node contains a classifier. Each node which is an extension of poses is always a subgroup of their parents. A tree structure is used to divide the degree of 3D rotation into smaller partitions. The tree contains 1024 number of leaf nodes that constitute different poses. Since every node has a classifier and detected images of the head in the poses are mentioned by the training classifier. A classifier is applied to an input image. If the classifier responds to an image, then the node is explored further else, the entire branch is snipped [14].

Merits:

- Pose refinement and recognition was done by exploiting an edge-based measure.

Demerits:

- Other factors to appearance such as clothing are not taken into account which could significantly improve results.

e) "Unsupervised Metric Learning" [15]: The method introduced determines a character by checking whether faces, when applied, reveal the same character or not. The face identification problem was addressed by using the technique of unsupervised metric learning. This system tests a pair of faces with face track and determines if the pair of faces selected are of a person in the face track. If they apply to the same individual, they are concluded as positive examples whereas when they belong to different people, then they are concluded as negative examples. Labels or metrics which are particular to face tracks of a cast can be used with semi-supervised learning [15].

We begin by extracting facial features. Face tracks are built from videos by using a face detector on every single frame, and derived observations are linked together. This type of approach has shown great results in uncontrolled videos. Now face pairs are extracted using metrics in an unsupervised manner. When a metric is recognized and learned, then a distance metric is defined to denote the distance between tracks for identification and recognition. Recognition experiments are conducted using a set of face tracks that have been already tagged to classify tracks which are not classified as labeled tracks in the test set [15].

Merits:

- Cast-specific metric has significantly improved results in character identification.

Demerits:

- The cast-specific metric cannot be computed when subtitles and scripts are used as a measure for recognition.

f) "Cast2Face" [16]: The algorithm explores the idea of automatically recognizing actors in a film, with the actor-character relations given by the film cast. The suggested structure is illustrated by (1) limiting the names to be assigned to the cast within the list of character names, (2) a set of pictures for each actor is obtained by performing queries on search engines, (3) the probe face tracks in a film are then classified by a rigorous multi-task combined sparse representation and classification system as one of the performers [16].

On a face chart, the given actor name is then categorized again to the character's name depending on the cast. In addition to face labeling, the suggested approach is further extended to spotlights a description of a single actor in his / her films. Empirical tests on many feature-length films show our method's satisfying efficiency.

Merits:

- Easy implementation.
- Textual Analysis is not required.

Demerits:

- Inefficient in motion blur video stills and low lighting sequences.

g) "Clustering Faces in Movies" [17]: This attempts to cluster faces in videos using a social network. Attempts are made to generate the social network automatically. A social network represents the social connectivity among peoples and is represented by groups of relationships. For this purpose, a weighted undirected graph is used, where edges represent closeness between two peoples while the vertices represent the people themselves. Most of the time, it is built on a robust face-clustering result [4].

To automate the process of analysis, the following are addressed: (1) construction of a social network of characters from the video, (2) use of social networks for clustering. The results of the work are compared to the then previous face recognition methods.

Consider an undirected weighted graph $G = (V, E, W)$, where $V = \{ v_1, v_2, \dots, v_N \}$ is the set of characters, $E = \{e_{ij} | if\ v_i\ and\ v_i\ have\ a\ relationship\}$, and $w_{ij} \in$ W signifies closeness between v_i and v_j [4]. w_{ij} is directly proportional to the number of scenes in which character *i* and *j* are present [4, 20]. The graph G is a social network.

The time complexity of the construction of social networks is $O(F)$, F is the total faces found. AP clustering is the most expensive task involved [21], used for grouping faces from a proximity matrix, and has a complexity of $O(V^2)$ for each iteration. It is still faster than the k-means technique for higher dimensions inputs.

h) "DeepStar" [22]: A unique framework is used for identifying characters based on robust features. They proposed a threefold system. Shots containing faces are extracted from the video stream. Then robust discriminative features are used to cluster faces. Finally, starring characters

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

are selected based on the generated occurrence matrix.

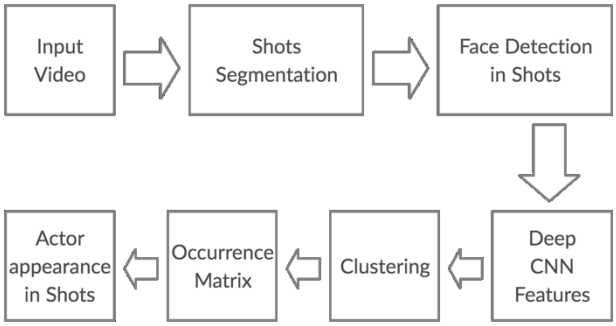


Fig. 3. A framework of DeepStar [22].

The first step is the segmentation of input video into shots employing a histogram-based method. The histogram-based technique is one of the automatic Shot boundary detection (SBD) technique, also known as temporal video segmentation [31]. This first step involving video segmentation is called preprocessing. To solve the problem of face detector capturing almost every face instead of only detecting starring characters, three additional constraints are added to eliminate side pose, blur, and small size [22].

In this method, the deep face model is chosen as it is trained on face data and features such as discriminative facial features are used in this model [32]. The Deep face CNN model is used to cluster faces of a character in a movie to identify performing characters. Starring characters have a greater impact on the audience, and identifying them correctly becomes much more important. Therefore, clusters with maximum points are given preference as a starring character.

Merits:

- Highest accuracy of greater than 83% as compared to previous algorithms.

B. Local Matching Based

a) "Hello! My name is... Buffy" [13]: High precision is achieved more efficiently by using this algorithm to identify the characters in a movie, by involving different dimensions:

- Subtitles - Automatic character annotation generation with time-stamped by aligning subtitles and transcripts
- Scripts - Strengthening surveillance intelligence by determining when characters communicate
- Video - Using complementary face matching indications and clothing matching to propose common face track annotations

The overall algorithm consists of three threads. The subtitles and scripts are used to get the first impression on the characters present in the video. Then video shots are used to identify character based on the facial attributes and outfit or dress attributes. As a final step, the amalgamation of the textual and optical data is used to nominate label to a particular person in the video [13].

Merits:

- The accuracy of the audio detection and the optical-facial detection nominations reached 25% and 90% respectively.

Demerits:

- No error correction.

b) "Who are you?" [23]: The algorithm is constructed by significantly expanding the scope by identifying and understanding characters in profile images. Additionally, the following inputs are made: (1) Seamless surveillance, profile integration, and frontal identification recognition, (2) A character-specific multiple kernel classifier that can learn the features that adequately distinguish characters [23].

The new aspect we introduce here is using a linear classification combination of kernels, and in particular, using Multiple Kernel Learning to evaluate the combination of features used.

It has the advantage of learning an optimum mix of features – others have considered several features, but only through a greedy learning algorithm. Here we discover a character-specific mix of kernels, discriminatively. This means it is possible to learn the features which best differentiate one character from another [23].

Compared to the baseline process, which trains a standard single kernel SVM classifier using patch descriptors extracted only from frontal faces, all proposed methods significantly improve performance.

Merits:

- Improved precision over the baseline by 20–30 %.
- That is because the new methods will mark face tracks that only include profile images, while the baseline approach doesn't.

Demerits:

- The MKL framework does not take into consideration other non-facial cues like hair and clothing.

c) "Talking Pictures" [24]: All previous research in the domain expects some visual information. This algorithm takes 'dialog cues' as the only root of administration into consideration.

Each state in the model represents partitions of the current k face tracks and transitions correspond to consecutive partitions. All operations provide us with a distinguishable learning model and are dynamic. After the said operations, each face track is labeled by a name from the available constraints by the learning classifier [24].

C. Global Matching Based

a) "Naming faces in films using hypergraph matching" [25]: In this method, the relationships between the video and the film scripts are used. The matching takes place between the extracted faces and the detected character names in the video.

Interactions between characters can be detected by the concurrence of character faces in a frame of a scene. The interactions signify a relationship between the characters in the relationship network. When a character speaks, there is a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

very high chance that the character is present in the scene. Thus, a face hypergraph and a name hypergraph network are constructed. The process is performed in three steps: (1) face track clustering, (2) hypergraph construction, (3) matching in hypergraph [25].

In face track clustering, detection and extraction of faces from all video frames are done using a face tracker. A normal-sized video may contain about 1,000,000 detected faces. The face tracker helps reduce duplicate faces, thereby reducing the amount of data to be processed. A parameter for measuring similarity between two faces is defined as:

$$S(T_m, T_n) = \mu * \ max_{i,j}(s(f_{m,i}, f_{n,j})) \quad (1)$$

where T_m and T_n are two tracks, $f_{m,i}$ represents a face in track m, $s(f_{m,i}, T_{n,j})$ represents a similarity between $f_{m,i}$ and $f_{n,j}$, and μ is used for standardization [12]. Face descriptors, such as eyes, forehead, nose, mouth, etc., are extracted. Finally, the grouping of face tracks is done using k-means, where each cluster represents a single character.

Merits:

- The correctness of face-name matching is more (96.3%) compared to existing methods.

Demerits:

- It does not take into account the noise in the face track clustering.

b) "Robust Face-Name Graph Matching" [26]: It aims to improve the face tracking and clustering methods by reducing the noise generated in the face tracking data. They worked on: (1) generating a character relationship which is noise insensitive, (2) graph matching method based on edit operation, (3) handling complex face changes.

Here, two methods are proposed. Both ways are based on global matching. For increasing the validity, an ordinal graph for name and face representation is used along with a group matching method known as error correcting graph matching (ECGM) [27]. In the first proposed method, clustering (k-means) is performed, while in the second method, face tracks are grouped using their native data structure (e.g. mean shift).

Merits:

- It significantly reduces the noise in face tracking data.

Demerits:

- It takes into account a significantly smaller number of character relationships.

c) "Character Identification in Feature-Length Films" [28]: This method aims to perform global matching between the character names and the face tracks in circumstances where there are not enough name hints found. The steps are: (1) building the association between affinity networks of face and name, (2) effectively measuring the distance between face tracks in track clustering, (3) social network analysis is used to mine character relationships.

The co-occurrence of the faces represents interaction and thus, the relationship between characters. Similarly, the speeches of various characters for a given scene represents an interaction. The names at the starting in subtitles can help make the name affinity network [28].

The script is used directly to make the name affinity network. For the face affinity network, Earth Mover's Distance is utilized to calculate the distance between face track clusters [29]. Face track clusters are generated by detecting face tracks from the clip and grouping them according to the characters. Finally, the relationships among the characters are mined.

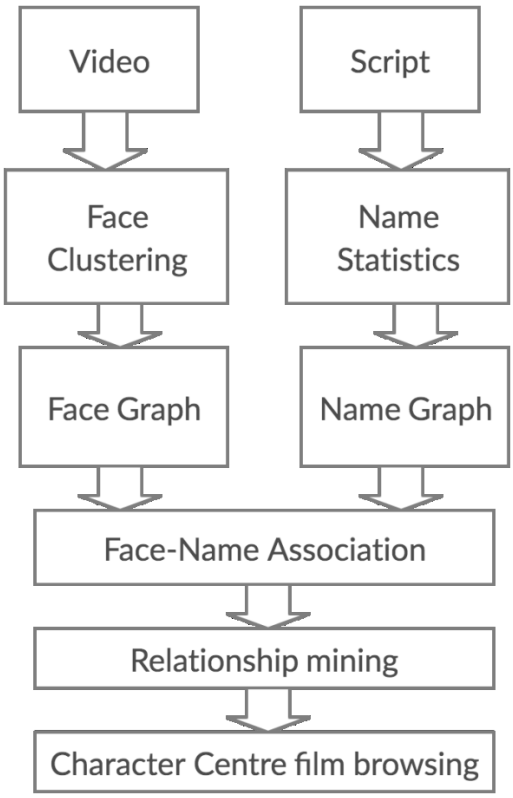


Fig. 4. Steps in character identification [28].

Merits:

- It provides a robust platform for film browsing based on characters.

Demerits:

- Searching and browsing happen only on the level of the scene, not the entire film.

III. OBSERVATIONS AND ANALYSIS

As part of the analyzing process, we made the following observations and inferences as far as character identification methods are concerned.

a) Clustering: It is observed that the way of clustering is dependent on the type of textual resources utilized by the different methods.

Most of the methods which utilize the cast list text resources perform face clustering by the appearance of the characters. When subtitle based techniques are taken into consideration, then faces are clustered as provided by the nearest neighbor classifier [7]. In the methods utilizing the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

film script, face clustering is done from extracted video frames.

b) Error Correction: Most algorithms lack error correction, which is performed to rectify the wrong character mapping. However, the recently developed algorithms consider this factor by using machine learning methods for mapping, like MKL classifiers [30]. It is a crucial part of development in this domain and the achievement of which now seems feasible by the use of AI and ML concepts.

c) Implementation: Different methods have different levels of complexity based on ease of understanding and ease of implementation. Most of the cast list-based methods are found to be easy to understand, as well as easy to implement. The amount of thinking required to understand is comparatively less than other methods not utilizing the cast list. Similarly, they are comparatively easier to implement. This is because the cast list-based methods do not require a time stamp dialog for alignment.

The subtitle based methods utilize OCR for extracting information regarding time stamps. The script-based approach builds a face-name association for clustering. Cast list-based methods require manual labeling for clustering [7]. Thus implementation of cast list-based methods is easier.

d) Noise Sensitivity: Mapping of character based on visual and textual resources is vulnerable to noise. Often in movies stills the video quality, motion blur, inefficient lightning leads to an improper determination of character. This sensitivity to noise has been a major challenge, and so there are a few methods developed to overcome it by using image processing algorithms, using textual cues in dialogues, using voice recognition, etc.

e) Resources: There are two resources used in the character identification: (a) textual, (b) video.

The textual resources include scripts and subtitles. The subtitles and scripts are used to get the first impression on the characters present in the video. The visual resources are used in the form of video stills. This helps in character identification based on facial features. Also, several algorithms use the clothing and dress of the characters to further help in identification.

f) No Universal Best Method: As far as we know, there is no universally best method. All the methods have varying accuracy, resource requirements, and employ different error correction methods. The comparative study between the methods does not produce (or signify) a best-of-all method.

IV. COMPARISION

Table I provides a comparison between various character identification methods analyzed in this paper. This comparison covers various aspects, such as textual resource used, timestamp, clustering method, accuracy, and error correction. These parameters have been selected, keeping in mind the common considerations that arise when choosing a character identification method.

a) Accuracy: It is important to establish the accuracy of the methods if we want to use them in production. The accuracy of film character identification methods is the measure of characters identified correctly. It is measured as the percentage of the characters identified correctly over the total number of characters present in the film.

It is observed that accuracy is not related to any of the parameters undertaken for the study.

b) Time Stamps: Timestamps are the records of the actual time in the film where the actual dialogues have been spoken by the characters.

Cast list and film scripts based methods do not have the time stamps available with them, while the methods utilizing subtitles have the timestamps.

In cases where the timestamps are not available, efforts have been made to extract the subtitles and timestamp, say using OCR [26].

c) Textual Resources: The algorithms in consideration use visual resources and textual resources. Visual resources include movie stills, character-actor image data sets, etc. Textual resources include scripts, subtitles, a cast list of the movie in consideration.

All the algorithms use the same visual resources. However, most of the cast list-based methods use cast lists as textual resources, while most of the global matching methods use scripts as resources and local matching methods use both script and subtitles as resources.

d) Error Correction: Many algorithms are efficient in mapping characters accurately. However, there are many instances where the mapping may be incorrect.

Most of the algorithms take error correction into account and try to rectify it. Others do not use any such error correction technique.

e) Clustering Method: Various methods compared above fall mainly into two types of clustering techniques as Hierarchical and Partitioning methods.

Agglomerative algorithms are very often used in the various character identification methods as it uses hierarchies. Local matching based methods use mainly K - Partition among Medoids as the clustering method to cluster faces into face examplers. Global matching based methods generally use k-means or Spectral clustering but sometimes a combination of both. Cast list-based methods use the agglomerative technique as well as affine invariant clustering, anisotropic clustering, deep adaptive clustering to improve face clustering by the appearance of a character.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

TABLE I. COMPARISON OF VARIOUS CHARACTER IDENTIFICATION METHODS

Research Paper	Textual Resource	Timestamp	Clustering Method	Accuracy	Error Correction
[8]	Cast list	No	Affine invariant clustering	-	Yes
[9]	Cast list	No	Anisotropic clustering	-	No
[11]	Cast list	No	Agglomerative clustering	> 80%	No
[13]	Script, Subtitle	Yes	K - Partition among Medoids	< 80%	No
[14]	Cast list	No	-	> 80%	Yes
[15]	Cast list	No	Agglomerative clustering	-	Yes
[16]	Cast list	No	Clustering based on the Posterior Similarity Matrix	< 92%	No
[17]	Cast list	No	Affinity propagation (AP)	-	No
[22]	Cast list	No	Deep adaptive clustering	> 83%	No
[23]	Script, Subtitle	Yes	K - Partition among Medoids	< 95%	Yes
[24]	Script	Yes	Agglomerative clustering	< 80%	No
[25]	Script	No	K-Means clustering	< 90%	No
[26]	Script	No	K-means, spectral clustering	-	Yes
[28]	Script	No	Spectral clustering	> 84.6%	No

V. CONCLUSION

We reviewed and analyzed different methods for character identification in movies. These methods utilize different types of textual resources: (1) cast list, (2) subtitle, (3) script, etc. We also discussed the merits and demerits of different methods. We also analyzed the mentioned techniques based on observations made. We also compared various methods based on common parameters in a simplified manner for better understanding.

After analyzing and comparison, we have concluded that there is no universal best method that exists for character identification in movies. Therefore, among various techniques, global matching was found to perform better concerning face tracking and clustering and maintaining the least noise.

We hope that with this review paper, our analysis and comparison can be extended to include other character identification methods in the future. This would act as a foundation for further research and development of the character identification methods.

REFERENCES

[1] K. Muhammad, T. Hussain, and S. W. Baik, "Efficient CNN based summarization of surveillance videos for resource-constrained devices," Pattern Recognition Letters, 2018.

[2] F. Schwandt, "statista," 2007.

[3] C. Needham, "Internet Movie Database (IMDb)," 1990.

[4] C.-Y. Weng, W.-T. Chu, and J.-L. Wu, "Rolenet: Movie analysis from the perspective of social networks," IEEE Transactions on Multimedia, vol. 11, pp. 256-271, 2009.

[5] J. Zhang, X. Yao, G. Han, and Y. Gui, "A survey of recent technologies and challenges in big data utilizations," in Information and Communication Technology Convergence (ICTC), 2015 International Conference on, 2015, pp. 497-499.

[6] K. K. Ingale, S. S. Patil, "A Review: Movie Character Identification Based on Graph Matching", International Journal on Recent and Innovation Trends in Computing and Communication, vol. 6, no. 3, pp. 205 - 207, Mar. 2018.

[7] B. S. Salve, S. A. Shinde,'A Novel Method for Movie Character Identification Based on Graph Matching: A Survey ', International Journal on Recent and Innovation Trends in Computing and Communication ,Volume: 2 Issue: 4,pp 873 – 879, April 2014.

[8] A. W. Fitzgibbon and A. Zisserman, "On affine invariant clustering and automatic cast listing in movies,"in Proc. ECCV, 2002, pp. 304-320.

[9] O. Arandjelovic and R. Cipolla,"Automatic cast listing in feature length films with anisotropic manifold space,"in Proc. Comput. Vis. Pattern Recognition, 2006, pp. 1513-1520

[10] K. Fukui and O. Yamaguchi, "Face Recognition Using MultiViewpoint Patterns for Robot Vision," Proc. Int'l Symp. Robotics Research, pp. 192-201, 2003

[11] D. Ramanan, S. Baker, and S. Kakade,"Leveraging archival video for building face datasets," in Proc. Int. Conf. Comput. Vis., 2007, pp. 1-8.

[12] T. L. Berg, A. C. Berg, J. Edwards, M. Maire,R.White,Y.W.Teh,E.G. Learned-Miller, and D. A. Forsyth, "Names and faces in the news,"in Proc. Comput. Vis. Pattern. Recognit., 2004, pp. 848–854.

[13] M. Everingham, J. Sivic, and A. Zissserman, "Hello! My name is Buffy automatic naming of characters in TV video," in Proc. BMVC, 2006, pp. 889-908D. Kornack and P. Rakic, "Cell Proliferation without Neurogenesis in Adult Primate Neocortex," Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.

[14] M. Everingham and A. Zisserman, "Identifying individuals in video by combining 'generative' and discriminative head models," Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Beijing, 2005, pp. 1103-1110 Vol. 2.

[15] R. G. Cinbis, J. Verbeek, and C. Schmid,"Unsupervised metric learning for face identification in TV video,"in Proc. Int. Conf. Comput. Vis., 2011, pp. 1559-1566.

[16] G. Gao, M. Xu, J. Shen, H. Ma and S. Yan, "Cast2Face: Assigning Character Names Onto Faces in Movie With Actor-Character Correspondence," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, no. 12, pp. 2299-2312, Dec. 2016.

[17] M. Yeh and W. Wu, "Clustering Faces in Movies Using an Automatically Constructed Social Network," in IEEE MultiMedia, vol. 21, no. 2, pp. 22-31, Apr.-June 2014.

[18] P.Wu and F. Tang, "Improving Face Clustering Using Social Context," Proc. ACM Int'l Conf. Multimedia, 2010, pp. 907–910.

[19] Q. Yin, X. Tang, and J. Sun, "An Associate-Predict Model for Face Recognition," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR), 2011, pp. 497–504.

[20] P.Wu and D. Tretter, "Close & Closer: Social Cluster and Closeness from Photo Collections," Proc. ACM Int'l Conf. Multimedia, 2009, pp. 709–712.

[21] B.J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," Science, vol. 315, 2007, pp. 972–976.

[22] I. U. Haq, K. Muhammad, A. Ullah and S. W. Baik, "DeepStar: Detecting Starring Characters in Movies," in IEEE Access, vol. 7, pp. 9265-9272, 2019.

[23] J. Sivic, M. Everingham and A. Zisserman, ""Who are you?" - Learning person specific classifiers from video," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1145-1152.

[24] T. Cour, B. Sapp, A. Nagle and B. Taskar, "Talking pictures: Temporal grouping and dialog-supervised person recognition," 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 1014-1021.

[25] Y. Zhang, C. Xu, J. Cheng and H. Lu, "Naming faces in films using hypergraph matching," 2009 IEEE International Conference on Multimedia and Expo, New York, NY, 2009, pp. 278-281.

[26] J. Sang and C. Xu, "Robust Face-Name Graph Matching for Movie Character Identification," in IEEE Transactions on Multimedia, vol. 14, no. 3, pp. 586-596, June 2012.

[27] H. Bunke, "Error correcting graph matching: on the influence of the underlying cost function," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 9, pp. 917-922, Sept. 1999.

[28] Y. Zhang, C. Xu, H. Lu and Y. Huang, "Character Identification in Feature-Length Films Using Global Face-Name Matching," in IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1276-1288, Nov. 2009.

[29] Q. Zhao, Z. Yang and H. Tao, "Differential Earth Mover's Distance with Its Applications to Visual Tracking," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 2, pp. 274-287, Feb. 2010.

[30] Cappelli, R., Maio, D. & Maltoni, D. A Multi-Classifier Approach to Fingerprint Classification. Pattern Anal Appl 5, 136–144 (2002). https://doi.org/10.1007/s100440200012

[31] N. J. Janwe and K. K. Bhoyar, "Video shot boundary detection based on JND color histogram," 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013), Shimla, 2013, pp. 476-480.

[32] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in BMVC, 2015, p. 6.

[33] R. Zabih, J. Miller, and K. Mai. A feature-based algorithm for detecting and classifying scene breaks. ACM Multimedia, 1995.

[34] P. Viola and M. Jones. Robust real-time face detection. IJCV, 57(2), 2004.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on July 21,2020 at 09:33:10 UTC from IEEE Xplore. Restrictions apply.