SCHOLARLY PUBLICATIONS

A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT

@DTU

(105th Edition)

SEPTEMBER 2021

BY: CENTRAL LIBRARY

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

PREFACE

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

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

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

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

Central Library

CONTENTS

1. A Comparative Study of Frequency Ratio, Shannon's Entropy and Analytic Hierarchy Process (AHP) Models for Landslide Susceptibility Assessment, ***6.Sandeep Panchal*** and ***3.Amit K. Shrivastava***, Civil, DTU

2. A Comparative Study of Machine Learning and NLP Techniques for Uses of Stop Words by Patients in Diagnosis of Alzheimer's Disease, ***8.Surabhi Adhikari***, ***8.Surendrabikram Thapa***, Priyanka Singh, Huan Huo, Gnana Bharathy and Mukesh Prasad, CSE, DTU

3. A Language-independent Network to Analyze the Impact of COVID-19 on theWorld via Sentiment Analysis, ***6.ASHIMA YADAV*** AND ***3.DINESH KUMAR VISHWAKARMA***, IT, DTU

4. A Novel Low-Power Nonvolatile 8T1M SRAM Cell, ***6.Damyanti Singh***, Kirti Gupta and ***3.Neeta Pandey***, Electronics, DTU

5. A Parametric Study of Eco-Friendly Refrigeration with Waste Heat Recovery from Gas Power Plants using Loop Heat Pipe, ***7.1.Ankit Dwivedi*** and ***3.R S Mishra***, Mechanical, DTU

6. A regression test case prioritization technique targeting 'hard to detect' faults, Sourav Biswas, ***8.Raghav Rathi***, Arpita Dutta, Pabitra Mitra and Rajib Mall, CSE, DTU

7. A Review on Protein Misfolding Diseases and therapeutics for preventing Diseases, ***8.Yogesh Kumar Singh***, Biotechnology, DTU

8. Analysing and Identifying Crucial Evidences for the prediction of False Information proliferated during COVID-19 Outbreak: A Case Study, ***6.Deepika Varshney*** and ***3.Dinesh Kumar Vishwakarma***, IT, DTU

9. Analysis of Organic Rankine Cycle with Waste Heat Recovery from Gas Power Plant using Loop Heat Pipes, ***7.1.Ankit Dwivedi*** and ***3.R S Mishra***, Mechanical, DTU

10. Analytical model to study the effect of alignment mechanism on the growth characteristics of carbon nanotubes in PECVD chamber, ***3.Suresh C. Sharma*** and ***6.Umang Sharma***, Physics, DTU

11. Artificial intelligence enables advanced financial modeling, ***6.Tanveer Aslam*** and ***7.Abhishek Sarkar***, DTU

12. Automated Threat Objects Detection with Synthetic Data for Real-Time X-ray Baggage Inspection, ***8.Kunal Chaturvedi***, Ali Braytee, ***3.Dinesh Kumar Vishwakarma***, Muhammad Saqib, Domingo Mery and Mukesh Prasad, IT, DTU

13. Behavioural Mechanisms of Microplastic Pollutants in Marine Ecosystem Challenges and Remediation Measurements, ***6.Megha Bansal***, ***3.Deenan Santhiya*** and ***3.Jai Gopal Sharma***, Biotechnology and Chemistry, DTU

14. Big Five Personality Detection Using Deep Convolutional Neural Networks, ***8.Waiel Tinwala*** and ***8.Shristi Rauniyar***, CSE, DTU

15. COMPARATIVE ANALYSIS OF CLASSIFICATION METHODS FOR PREDICTION SOFTWARE FAULT PRONENESS USING PROCESS, METRICS, ***6.Anjali Bansal***, Software Engineering, DTU

16. Comparative Study of Nonlinear Controllers for Standalone PV System, ***6.Manisha***, ***6.MD. Modassir Masoom*** and ***3.Narendra kumar***, Electrical, DTU

17. Data rate over different applications in 5G and beyond Networks, ***8.Manpreet Kaur*** and ***3.Rajesh Kumar Yadav***, CSE, DTU

18. Deep Learning Based Thermal Object Recognition under Different Illumination Conditions, ***8.Rohini Goel***, ***8.Avinash Sharma*** and ***3.Rajiv Kapoor***, CSE & ECE, DTU

19. Design and Analysis of a Novel Y-Type Hydraulic Braking System for Implementation in ATVs, ***8.1.Kunal Mathur***, ***8.1.Lalit Kumar Choudhary*** and ***3.Krovvidi Srinivas***, Mechanical, DTU

20. Design and development of a reconfigurable tiling robot, Rizuwana Parween, ***3.Madan Mohan Rayaguru***, Shi Yuyao, Borusu Charan Satya Chandra Sairam and Mohan Rajesh Elara, Electrical, DTU

21. Design and feasibility analysis of hydrogen based hybrid energy system: A case study, ***8.Srijit Basu***[a], ***8.Alfred John***[a], ***8.Akshay***[a] and ***3.Anil Kumar***[a,b], [a]Mechanical and [b]Environmental, DTU

22. Design and Investigation of Recessed-T-Gate Double Channel HEMT with InGaN Back Barrier for Enhanced Performance, ***6.Megha Sharma*** and ***3.Rishu Chaujar***, Applied Physics, DTU

23. Effect of data packet size on the performance of RIP and OSPF routing protocols in hybrid networks, Neha Jain, Ashish Payal and ***7.1.Aarti Jain***, Electronics, DTU

24. Effect of metal doping on visible light absorption and optical properties of lithium iobate (LiNbO3) crystal: a first-principles calculation,***6.ASHISH RATURI***, ***3.POORNIMA MITTAL*** and SUDHANSHU CHOUDHARY, Electronics, DTU

25. Empirical Validation of cross-version and 10-fold cross-validation for Defect Prediction, ***3.Ruchika Malhotra*** and ***3.Shweta Meena***, CSE, DTU

26. Enhancing the wine tasting experience using greedy clustering wine recommender system, ***3.Rahul Katarya*** and ***8.Rajat Saini***, CSE, DTU

27. Finite element study on the influence of pore size and structure on stress shielding effect of additive manufactured spinal cage, Vijay Kumar Meena, Parveen Kalra and ***3.Ravindra Kumar Sinha***, Physics, DTU

28. Generation and analysis of events for Heavy Ions collisions using the MC generator - therminator , ***7.Anish Kalsi***, Engineering Physics ,DTU

29. Gratitude Questionnaire (GQ-6)—Exploring Psychology ***3.Naval Garg***, Nancy Katiya and Mehak, Management, DTU

30. Integrated Taguchi method-assisted polynomial Metamodelling & Genetic Algorithm based optimisation of a surface inset permanent synchronous motor for performance improvement, ***3.Monika Verma***, ***3.Madhusudan Singh*** and ***3.Mini Sreejeth***, Electrical, DTU

31. Low Light GAN-Based Photo Enhancement, Ashish Chopra, ***8.Akash Maan*** and Anmol Kesharwani, IT, DTU

32. Multilevel profiling of situation and dialogue-based deep networks for movie genre classification using movie trailers, ***3.Dinesh Kumar Vishwakarma***, ***8.Mayank Jindal***, ***8.Ayush Mittal*** and ***8.Aditya Sharma***, IT, DTU

33. Nonlinearity Analysis of Quantum Capacitance and its Effect on Nano-Graphene Field Effect Transistor characteristics, ***6.Munindra*** and ***3.Deva Nand***, ECE, DTU

34. Numerical Study of JAM-GS-GAA FinFET: A Fin Aspect Ratio Optimization for Upgraded Analog and Intermodulation Distortion Performance, ***6.Bhavya Kumar*** and ***3.Rishu Chaujar***, Physics, DTU

35. Personalized and Dynamic top-k Recommendation System using Context Aware Deep Reinforcement Learning, ***8.1.Anubha Kabra*** and ***8.Anu Agarwal***, DTU

36. Power quality investigation of CHB nine-level converter based large-scale solar PV system with different modulation schemes, ***6.Jyoti Kulkarni***, Shivam Kumar Yadav, Bhim Singh and ***3.Narendra Kumar***, Electrical, DTU

37. PV integrated grid synchronization technique usingmodified SOGI-FLL and zero-crossing detector, ***6.Hemant Saxena***, ***3.Alka Singh***, ***3.Jitendra Nath Rai*** and Manoj Badoni, Electrical, DTU

38. Relation between Schrödinger's Equation and Einstein's Equation, ***8.Aman Yadav***, Physics, DTU

39. Restoration circuits for low power reduce swing of 6T and 8T SRAM cell with improved read & write margins***,3.Ram Murti Rawat*** and ***3.Vinod Kumar***, CSE, DTU

40. Review of Dynamic Comparators for ADCs, ***8.Yogesh Sachdeva***, ***8.Nalin Nehra***, ***8.Shikhar Bansal*** and ***3.Garima***, Electrical, DTU

41. Role of Different Peptides for Cancer Immunotherapy, ***8.Ritika Luthra***, ***8.Shreeja Datta*** and ***3.Arpita Roy***, Biotechnology, DTU

42. Soil and Water Assessment Tool (SWAT) Modelling: A Review, ***6.Noopur Awasthi*** and ***3.Munendra Kumar***, Civil, DTU

43. Solution of optimal power flow problem using sine-cosine mutation based modified Jaya algorithm: a case study, ***6.Saket Gupta***, ***3.Narendra Kumar*** and Laxmi Srivastava, Electrical, DTU

44. Status of groundwater resource potential and its quality at Kabul, Afghanistan: a review, ***6.Ali Reza Noori*** and ***3.S. K. Singh***, Environmental, DTU

45. Study of contribution of doubly excited 3d10 configurations in excitation energies and SXR transition data of Fe-like ions, ***3.R Sharma*** and ***6.A Goyal***, Physics, DTU

46. Study on parameter's optimization for the minimizing of the residual stress in powder mixed near-dry electric discharge machining, ***6.1.S. Sundriyal***, ***3.Vipin*** and R.S. Walia, Mechanical, DTU

47. Thermodynamic performances of ejector refrigeration systems (ERS) using low GWP ecofriendly HFO refrigerants, ***3.R. S. Mishra***, Mechanical, DTU

48. Two different Approaches for consistency of Intuitionistic Multiplicative Preference Relation using Directed Graph, ***6.Mamata Sahu*** & ***3.Anjana Gupta***, Mathematics, DTU

49. Waste Cooking Oil as Fuel in CI Engine – A Review, ***7.Adhirath Mandal*** and HaengMuk Cho, Production, DTU

50. When Economics Jitter, Bitcoin Flutters : Evidence from the Impact of Macroecomomic Factors on Bitcoin Returns, ***3.Sonal Thukral***, Ashima Arora and Varun Bhandari, DSM, DTU

51. Zero-Shot Learning with Missing Attributes using Semantic Correlations, Ali Braytee, Mohamad Naji, Ali Anaissi, ***8.Kunal Chaturvedi*** & Mukesh Prasad, IT, DTU

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

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

International Journal of Geo-Information

MDPI

Article

A Comparative Study of Frequency Ratio, Shannon's Entropy and Analytic Hierarchy Process (AHP) Models for Landslide Susceptibility Assessment

Sandeep Panchal * and Amit K. Shrivastava

Department of Civil Engineering, Delhi Technological University, Main Bawana Road, Samaypur Badli, Delhi 110042, India; aksrivastava@dce.ac.in

* Correspondence: 2290sandy@gmail.com; Tel.: +91-87-0863-1157

Abstract: Landslide susceptibility maps are very important tools in the planning and management of landslide prone areas. Qualitative and quantitative methods each have their own advantages and dis-advantages in landslide susceptibility mapping. The aim of this study is to compare three models, i.e., frequency ratio (FR), Shannon's entropy and analytic hierarchy process (AHP) by implementing them for the preparation of landslide susceptibility maps. Shimla, a district in Himachal Pradesh (H.P.), India was chosen for the study. A landslide inventory containing more than 1500 landslide events was prepared using previous literature, available historical data and a field survey. Out of the total number of landslide events, 30% data was used for training and 70% data was used for testing purpose. The frequency ratio, Shannon's entropy and AHP models were implemented and three landslide susceptibility maps were prepared for the study area. The final landslide susceptibility maps were validated using a receiver operating characteristic (ROC) curve. The frequency ratio (FR) model yielded the highest accuracy, with 0.925 fitted ROC area, while the accuracy achieved by Shannon's entropy model was 0.883. Analytic hierarchy process (AHP) yielded the lowest accuracy, with 0.732 fitted ROC area. The results of this study can be used by engineers and planners for better management and mitigation of landslides in the study area.

Keywords: landslide susceptibility mapping; AHP; Shannon's entropy; frequency ratio; landslide inventory

check for updates

Citation: Panchal, S.; Shrivastava, A.K. A Comparative Study of Frequency Ratio, Shannon's Entropy and Analytic Hierarchy Process (AHP) Models for Landslide Susceptibility Assessment. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 603. https://doi.org/10.3390/ijgi10090603

Academic Editors: Wolfgang Kainz and Peter Bliš'an

Received: 3 August 2021
Accepted: 9 September 2021
Published: 12 September 2021

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

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

1. Introduction

Hilly regions are frequently affected by landslide disasters. Landslides are responsible for huge loss of life and damage to infrastructure [1,2]. A landslide is the result of movement of a slope under the effect of gravity, but the movement can be triggered by the factors such as rainfall, earthquakes, erosion, etc. [3,4]. The accurate prediction of landslides can help in the planning and management of landslide hazards, and can ultimately be used for the reduction of risk [5]. Landslide susceptibility maps help in the identification of landslide prone areas. Landslide susceptibility maps can be very efficient tools for planners and risk managers [5]. The occurrence of landslides is a complex phenomenon which depends upon various factors. Geological factors, drainage characteristics, land-use of the region, construction activities, etc., can all be responsible for the occurrence of landslides [6]. Anthropogenic activities and the development of infrastructure in landslide prone areas can disturb natural slopes, resulting in massive disaster [7,8].

The study area selected in this study is the Shimla district in Himachal Pradesh (H.P.), India. The study area is developing at a very rapid rate. Anthropogenic activities have been increasing in the study area, resulting in the instability of natural slopes. Landslide occurrence is very frequent in the study area due to rugged topographic and the typical climatic conditions in the region. Therefore, there is a need for the preparation of a landslide susceptibility map for the region which can be used by planners and engineers.

In different regions of the world, landslide susceptibility mapping has been undertaken by various researchers, using geographic information systems (GIS) and remote sensing [2,4,6,9–14]. Qualitative and quantitative approaches are both widely used. Qualitative techniques consider expert opinion, while quantitative approaches establish a mathematical relationship between causative factors and the occurrence of landslide [2,5,13,15–17]. Landslide inventory-based methods utilize a direct approach in which susceptibility maps are prepared by considering the geomorphological characteristics of the region. Ranking and weightage are introduced to improve direct susceptibility mapping, through analytic hierarchy processes (AHP), analytic network processes (ANP), fuzzy logic, etc. [15–18]. These methods can be categorized as semi-qualitative methods. Quantitative methods such as the bivariate method [19–21], multi-variate method [22–24], weight of evidence [25,26], etc., are widely used with the development of GIS techniques. GIS platforms help in the calculation and visualization of the cumulative effects of causative factors on landslides. Qualitative techniques are based on the subjective judgment of experts, while the output in quantitative techniques depends on the availability of historical landslide data [27].

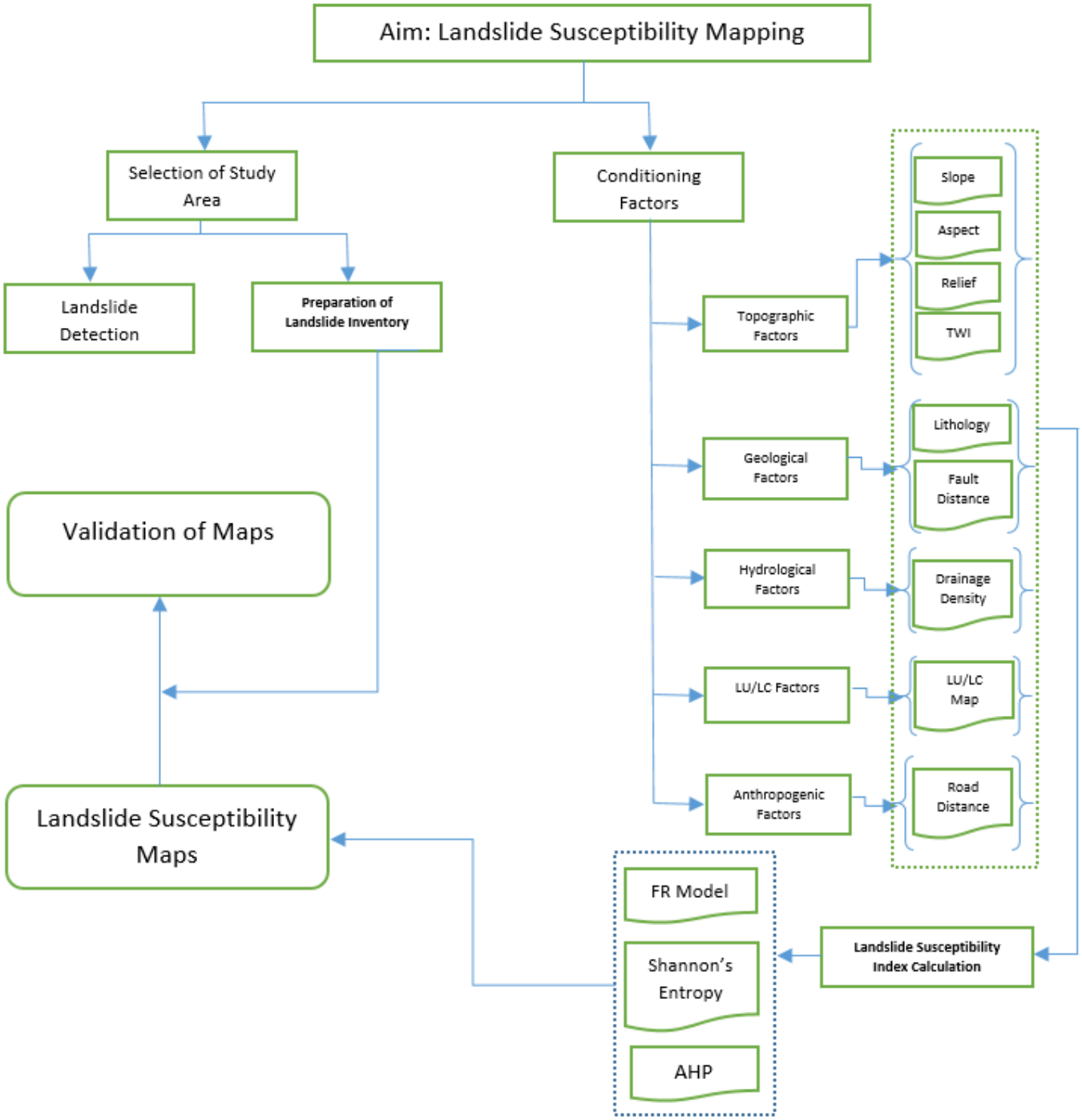
Landslide susceptibility mapping has not been undertaken for the study area until now. There is a need for the preparation of an accurate landslide susceptibility map for the region. In this study, we used three techniques for landslide susceptibility mapping, i.e., frequency ratio (FR), Shannon's entropy model and analytic hierarchy process (AHP). Shannon's entropy model is an improvement over the frequency ratio model and is rarely used in the Indian continent for landslide susceptibility mapping. The results of this study can help in the delineation of landslide prone regions in the study area. Construction planners, environmental engineers and risk managers can use the results of this study for the planning and mitigation of landslide hazards in the study area.

2. Study Area

Figure 1a shows the methodology adopted in this study. In the first step, the study area is selected. Shimla is the capital of Himachal Pradesh (H.P.) province in India. The region lies between longitudes 77°0′ E and 78°19′ and latitudes 30°45′ and 31°44′. The elevation of the region varies from 300 m to 6000 m and shows variation in the vegetation [28,29]. The study area lies in the Kashmir and Western Himalayas seismic region. Shimla district lies in a high damage risk zone in the case of seismic activity (MSK VIII) [28,29]. Major earthquakes that have occurred in Himachal Pradesh (H.P.) are the Kangra earthquake 1905 (M = 8), the Chamba earthquake 1945 (M = 6.5), the Kinnaur earthquake (M = 6.8), etc. The study area is predominantly covered by rock from the Jutogh, Shali, Simla and Rampur groups. The study area consists of lithological units of shale, siltstones, quartzite, dolomites, phyllite, schist, conglomerate, etc. The region is predominantly covered by the Proterozoic age group, Plaeo-proterozoic age group and Neo-proterozoic age group. The study area is mainly forest and cultivated land. Some of the study area in upper Shivaliks consists of hill tops with snow. The temperature varies from 4 °C to 31 °C [28,29]. Therefore, the region is cool in winter and moderately warm in summer. Figure 1b shows the study area and landslide inventory.

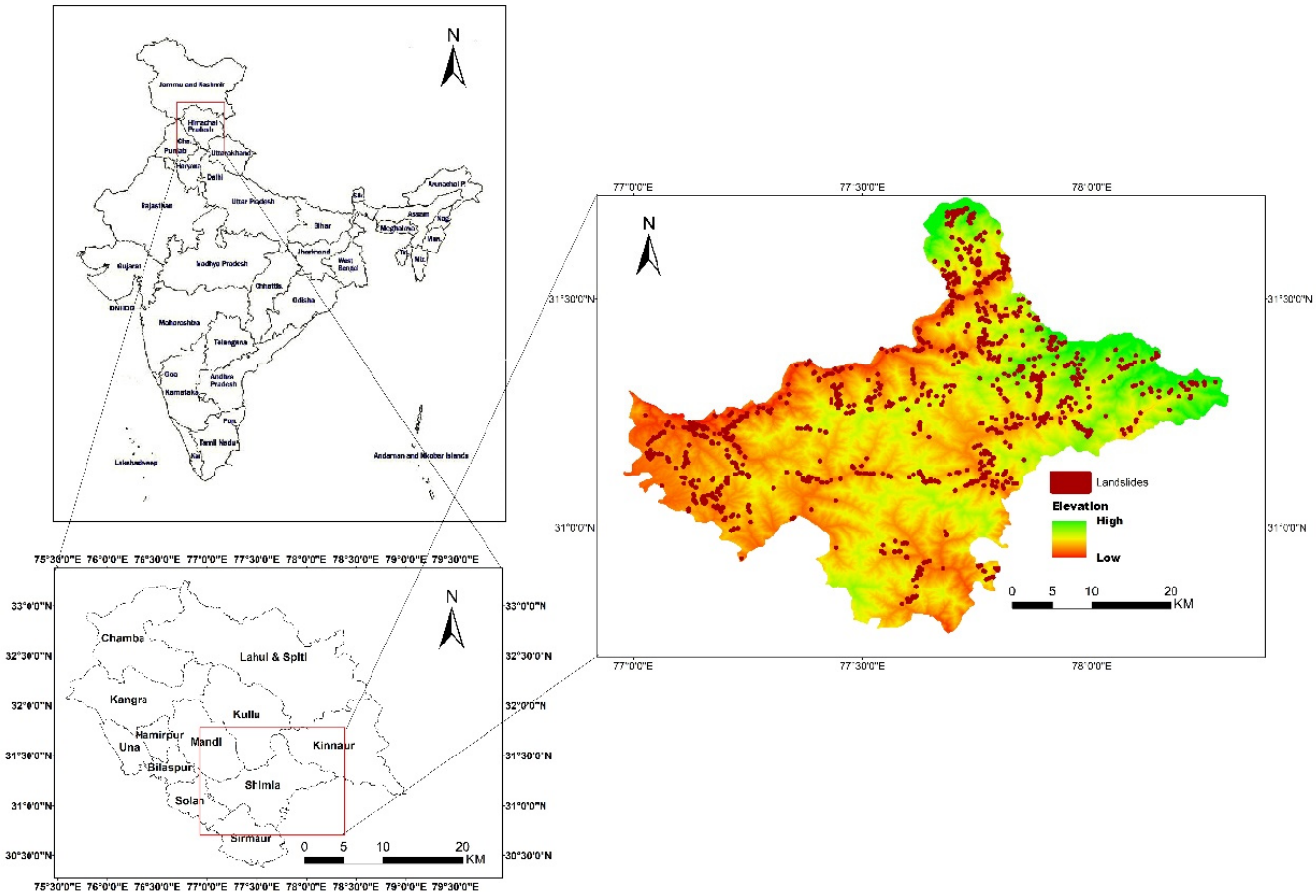
2.1. Landslide Inventory

A landslide inventory plays an important role in the mathematical modelling and validation of the final output [13]. Landslide inventories consist of information regarding the location of landslides, area of landslides, activity and date of landslides, etc. [30]. A landslide inventory can be simple to complex, depending upon the requirement of study. A landslide inventory is prepared from previous literature, Google Earth and field visits. The field visits cannot be undertaken frequently due to restricted movement in the current COVID-19 pandemic. The landslide inventory consists of more than 1500 landslides in the study area. The landslides are represented by polygons. Figure 1 shows the landslide inventory.



(**a**)

Figure 1. *Cont.*



(**b**)

Figure 1. (**a**) Flow chart of methodology, (**b**) Study area and landslide inventory.

2.2. Landslide Conditioning Factors

Understanding the causative factors of landslides for the preparation of landslide susceptibility maps is very important. The data for the preparation of thematic layers of causative factors were obtained from different sources. An ASTER digital elevation model (DEM) was obtained from the USGS website. The resolution of the DEM used in this study is 1 arc second, which is approximately equal to 30 × 30 m. The average route mean square error (RMSE) in the measurement of elevation is ±13.31 m. In terms of absolute positioning, the ASTER DEM models have a tendency for displacement by 5.3 m in our study. Landsat 8 data was used for the preparation of a land-use map for the study area. The resolution for Landsat bands 1–7 is 30 × 30 m. Hydrogeological and topographical maps were used for the digitization of faults, lithology and the road network. The land cover of the study area was obtained from Landsat 8 data.

Selection of condition factors depends on the availability of data, ease of use and the results of a literature survey. In many studies, DEM based factors are considered to be major causative factors, but in recent studies, land use, drainage networks and faults or lineaments are also considered to be responsible for landslide occurrence. The factors considered in this study are explained as follows. The relative importance of the causative factors can be found using Shannon's entropy and the AHP model.

2.2.1. Slope Gradient

Slope gradient may be defined as the rate of change in elevation with respect to horizontal surface. When the shear stresses in a downward direction exceed the shear strength of soil or rock mass, the failure of slopes occur. Therefore, the slope angle directly affects the stability of slope [31]. The slope of the study area was extracted from a digital elevation model (DEM) of the study area. The study area is comprised predominantly of steep slopes greater than 60°. As the slope angle increases, the possibility of slope failure increases. Figure 2a shows slope gradient of the study area.

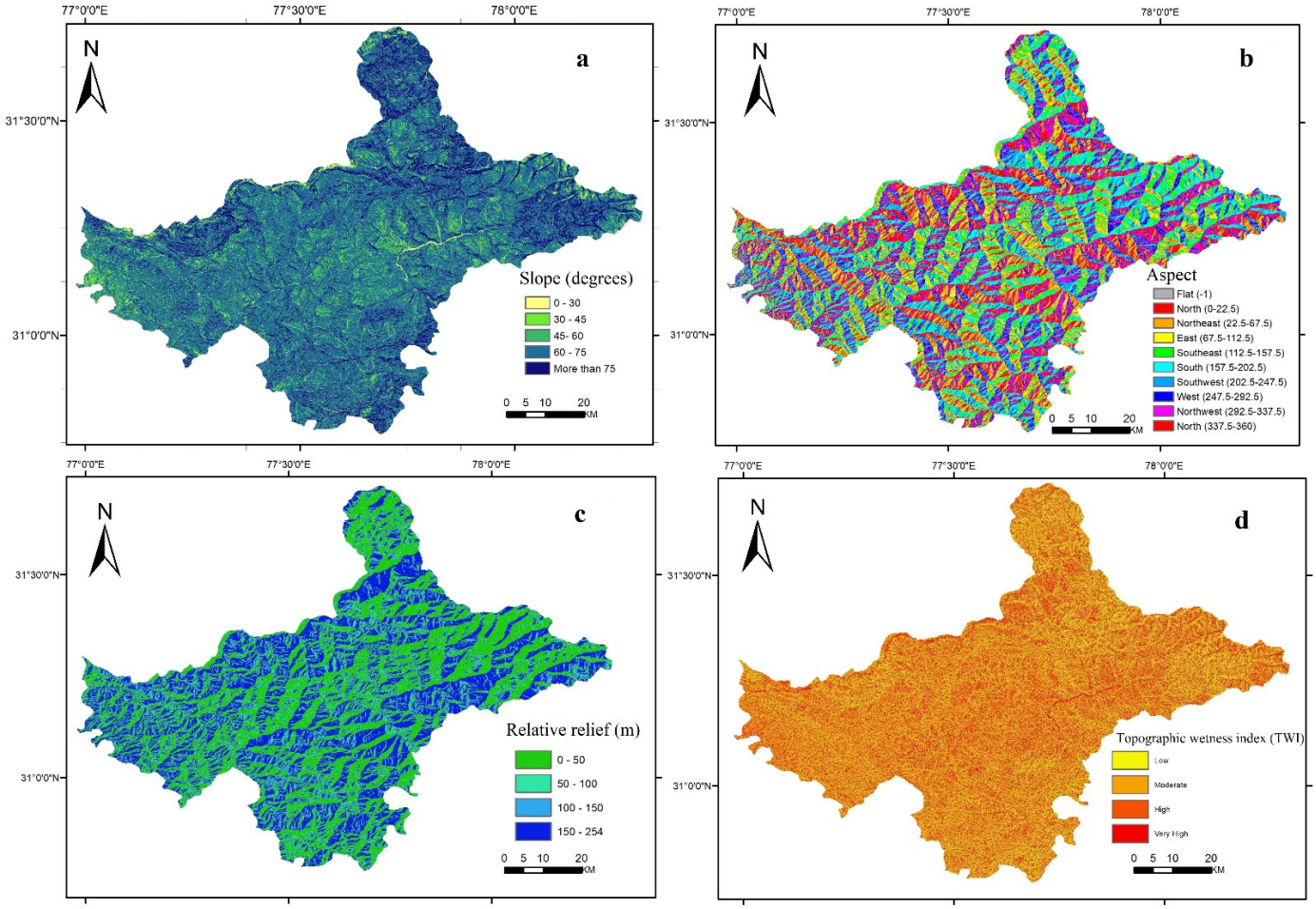


Figure 2. Causative factors: (**a**) Slope gradient (**b**) Slope aspect (**c**) Relative relief (**d**) Topographic wetness index.

2.2.2. Slope Aspect

Slope aspect gives the direction of maximum slope with respect to magnetic north. Slope aspect affects erosion, evapotranspiration, desertification, solar heating and soil moisture [5]. Slope aspect indirectly shows the impact of water on the mass. Figure 2b shows the slope aspect of the study area.

2.2.3. Relative Relief

Relative relief shows the change in elevation. The change in elevation indicates variation in vegetation type and topographical characteristics of the region [13]. Vegetation roots hold the rock and soil mass together and make it more stable. The relative relief varies from 0 to 254 m in the present study. Figure 2c shows the relative relief of the region.

2.2.4. Topographic Wetness Index (TWI)

The topographic wetness index (TWI) is the tendency of terrain to accumulate water. It indicates variation in soil moisture [32]. Figure 2d shows the TWI of the region. If SCA is specific catchment area and φ is the angle of slope, TWI is given by the following formula:

$$TWI = ln\left(\frac{SCA}{\tan\varphi}\right) \quad (1)$$

2.2.5. Lithology

Lithology is an important causative factor of landslides. Variation in rock characteristics and age prominently affect landslide events [33]. Detailed information about the lithological and geological character of the region is given in Table 1. Figure 3a shows the lithology of the region.

Table 1. Lithology of the study area.

Age	Group Name	Geological Formation	Lithology
Mesoproterozoic	Shali	Tatapani	Pink and grey dolomite, phyllite, shale
Neoproterozoic	Shali	Sorgharwari	Pink and grey limestone, sporadic shale
Palaeocene-Eocene	Sirmur Dharmshala Group	Subathu	Green carbonaceous shale, limestone, quartzite
Palaeoproterozoic	Kulu	Khokhan	Schist and quartzite
Palaeozoic	Not available	Not available	Medium to coarse biotite granite
Proterozoic (Undiff)	Jutogh	Manal, chor, pabar	White grey quartzite, schist, carbonaceous dolomite, granite, gneiss

2.2.6. Drainage Density

The drainage characteristics of the region significantly affect the occurrence of landslides. River water erodes the soil at the toe of the slope and weakens the base, which may result in slope failure [33,34]. Drainage density is divided into four categories for the study area. Higher drainage density increases the risk of landslides. Figure 3b shows the drainage density of study area.

2.2.7. Distance from Roads

Road construction disturbs the natural slopes along highway corridors. Road construction activities involve significant cutting and filling activities. The bottom section of a slope is weakened by the soil and rock cutting activities during highway construction [35]. Vehicular vibrations, excessive excavations and disturbed natural slopes all induce instability and increase the risk of landslide occurrence. Slopes existing near roads are more prone to experience landslides [13,35]. A detailed analysis of slope stability during the planning of construction of projects can reduce the risk of landslide occurrence. Distance from roads indirectly represents the effects of human interference on natural slopes. If proper slope protection is provided and the slopes along highways are planned properly, the risk of landslides along highways can be reduced. Figure 3c shows the distance from roads.

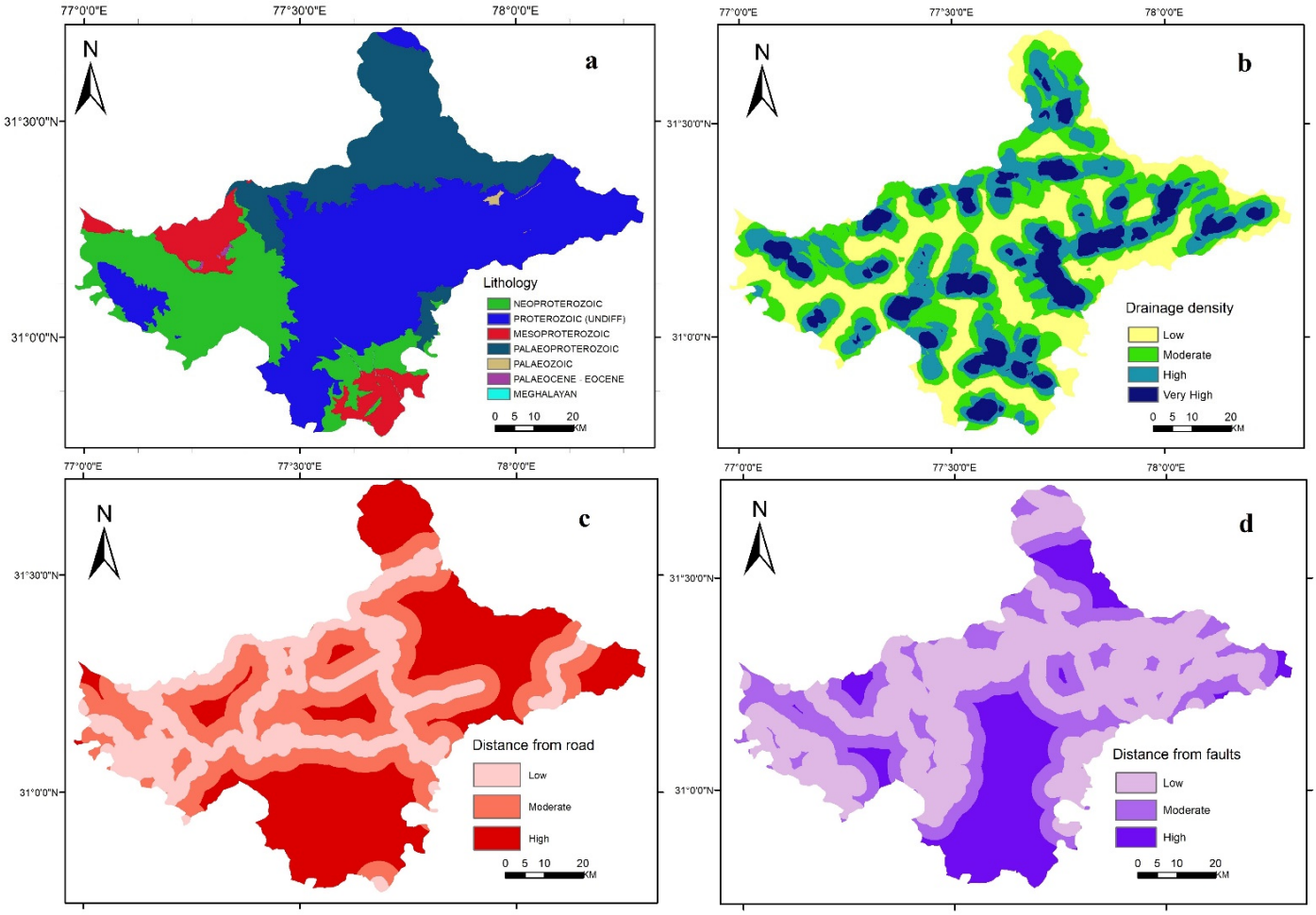


Figure 3. Causative factors: (**a**) Lithology (**b**) Drainage density (**c**) Distance from road (**d**) Distance from faults.

2.2.8. Distance from Faults

Cracks, faults, fissures and lineament can increase pore pressure in the rock and soil mass, and can induce instability in slopes [36,37]. In this study, faults were digitized from NRSA data. Proximity to faults is considered a causative factor in this study, and Euclidean distance from faults was calculated in ArcGIS. Figure 3d shows distance from faults.

2.2.9. Land Cover

Land cover is an important factor that affects the occurrence of landslides. Vegetation area, forest area, barren land, etc. are shown in a land cover map, and may affect the occurrence of landslides due to their diverse characteristics. A land cover map was extracted from Landsat 8 imagery. The land use was classified using a supervised learning technique of image classification in ArcGIS. The study area is predominantly covered with the cropland and forests. In the upper elevations, there are glaciers and snow covered hill tops as well. Figure 4 shows the land cover map of the region.

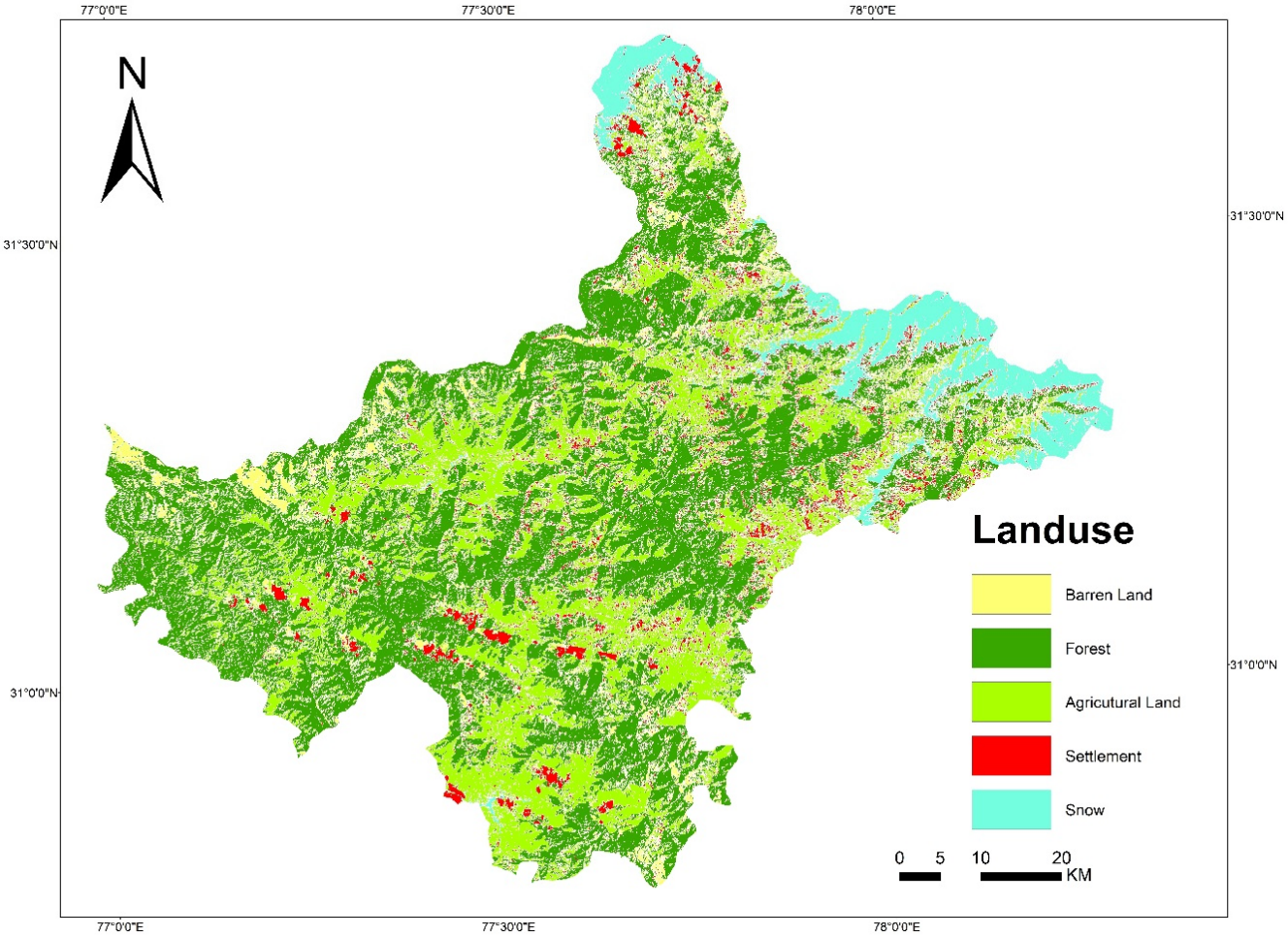


Figure 4. Land-use/land cover map.

3. Modelling Approach

3.1. Frequency Ratio

Frequency ratio models are based upon the relationship between the occurrence of landslide events and the causative factors [38,39]. A relationship between the landslide inventory and causative factors can be established in a GIS environment. A landslide will occur under the same conditions as past landslides [5]. The frequency ratio is the ratio between the percent areas where landslides have occurred in a class to the percent area of the influencing class relative to the whole study area as shown in the following formula [40].

$$Fr = \frac{\left(\frac{Mi}{M}\right)}{\left(\frac{Ni}{N}\right)} \quad (2)$$

where Mi = Number of pixels containing landslides in a class; M = Total number of pixels in a class; Ni = Total number of pixels containing landslides; N = Total number of pixels in the study area.

The landslide susceptibility index is calculated by summing up the Fr for each sub-factors:

$$\begin{aligned} &Landslide\ Susceptibility\ Index\ (LSI) \\ &= Fr_1 + Fr_2 + Fr_3 + Fr_4 + \ldots\ldots\ldots\ldots\ldots\ldots + Fr_n \end{aligned} \quad (3)$$

where Fr is frequency ratio and n is number of causative factors. Table 2 shows the frequency ratio (FR) calculation.

Table 2. Frequency ratio and Shannon's entropy weightage.

Conditioning Factor	Classes	No. of Pixels in Class	Percentage in Class (a)	No. of Landslide Pixel in Class	Percentage of Landslide Pixel (b)	FR (b/a)	Pij	Eij	Hij = 1 − Eij	Wj
Slope										
	<30	107,718.000	1.710	13.000	0.800	0.468	0.115	−0.108		
	30–45	246,772.000	3.900	13.000	0.800	0.205	0.050	−0.065		
	45–60	1,001,205.000	15.910	91.000	5.950	0.374	0.092	−0.095	0.442	0.120
	60–75	3,779,957.000	60.050	803.000	52.510	0.874	0.214	−0.143		
	More than 75	1,158,974.000	18.410	609.000	39.830	2.163	0.530	−0.146		
		6,294,626.000		1529.000		4.085		−0.558		
Aspect										
	Flat	105.000	0.002	0.000	0.000	0.000	0.000	0.000		
	North	916,307.000	14.560	210.000	13.730	0.943	0.116	−0.108		
	Northeast	792,179.000	12.590	156.000	10.200	0.810	0.100	−0.100		
	East	627,294.000	9.970	255.000	16.680	1.673	0.206	−0.141		
	South East	804,073.000	12.770	290.000	18.970	1.486	0.183	−0.135	0.124	0.033
	South	900,237.000	14.310	230.000	15.040	1.051	0.129	−0.115		
	South West	809,177.000	12.860	135.000	8.830	0.687	0.084	−0.091		
	West	643,538.000	10.220	150.000	9.810	0.960	0.118	−0.109		
	North West	801,716.000	12.730	103.000	6.740	0.529	0.065	−0.077		
		6,294,626.000		1529.000	100.000	8.139		−0.876		
Relative relief										
	0−50	2,612,642.000	41.660	804.000	52.580	1.262	0.323	−0.159		
	50−100	691,263.000	11.023	145.000	9.480	0.860	0.220	−0.145		
	100−150	710,948.000	11.337	186.000	12.160	1.073	0.274	−0.154	0.408	0.110
	More than 150	2,256,417.000	35.980	394.000	25.770	0.716	0.183	−0.135		
		6,271,270.000		1529.000		3.911		−0.592		
TWI										
	Low	2,810,725.000	44.650	712.000	46.560	1.043	0.340	−0.159		
	Moderate	2,473,126.000	39.290	596.000	38.980	0.992	0.323	−0.159		
	High	855,581.000	13.590	207.000	13.540	0.996	0.325	−0.159	0.500	0.135
	Very High	155,194.000	24.650	14.000	0.910	0.037	0.012	−0.023		
		6,294,626.000		1529.000		3.068		−0.500		
Lithology										
	Neoproterozoic	1,364,515.000	21.680	415.000	27.150	1.252	0.408	−0.159		
	Proterozoic (Undiff)	3,023,399.000	48.030	415.000	27.140	0.565	0.184	−0.135		
	Mesoproterozoic	502,058.000	7.970	143.000	9.350	1.173	0.382	−0.160		
	Plaeoproterozoic	1,384,339.000	22.000	556.000	36.360	1.653	0.539	−0.145	0.401	0.109
	Palaeozoic	10,282.000	0.160	0.000	0.000	0.000	0.000			
	Palaleocene-eocene	7564.000	0.130	0.000	0.000	0.000	0.000			
	Meghalayan	2469.000	0.030	0.000	0.000	0.000	0.000			
		6,294,626.000		1529.000		4.643		−0.599		
Drainage density										
	0–15	2,309,122.000	36.680	320.000	20.930	0.571	0.186	−0.136		
	15–30	2,194,591.000	34.860	628.000	41.070	1.178	0.384	−0.160		
	30–45	1,441,690.000	22.900	438.000	28.640	1.251	0.408	−0.159	0.403	0.109
	More than 45 (up to 66)	349,223.000	5.550	143.000	9.350	1.685	0.549	−0.143		
		6,294,626.000		1529.000				−0.597		

Table 2. *Cont.*

Conditioning Factor	Classes	No. of Pixels in Class	Percentage in Class (a)	No. of Landslide Pixel in Class	Percentage of Landslide Pixel (b)	FR (b/a)	Pij	Eij	Hij = 1 − Eij	Wj
Distance from road										
	0–1.5 KM	1,743,800.000	27.700	531.000	34.720	1.253	0.409	−0.159		
	1.5–5.5 km	2,114,630.000	33.590	286.000	18.700	0.557	0.181	−0.134	0.547	0.148
	More than 5.5	2,436,196.000	38.700	712.000	46.560	1.203	0.392	−0.159		
		6,294,626.000		1529.000				−0.453		
Distance from faults										
	0–1.5KM	2,790,363.000	44.330	674.000	44.080	0.994	0.324	−0.159		
	1.5 km–3.0 km	1,295,638.000	58.660	298.000	19.490	0.332	0.108	−0.105		
	more than 3 km	2,208,625.000	35.090	557.000	26.420	0.753	0.245	−0.150	0.587	0.159
		6,294,626.000		1529.000				−0.413		
Landuse/landcover										
	Snow	440,046.000	6.990	220.000	14.390	2.059	0.671	−0.116		
	Settlement	604,702.000	9.610	104.000	6.800	0.708	0.231	−0.147		
	Agricultural Land	1,800,554.000	28.600	246.000	16.080	0.562	0.183	−0.135	0.287	0.078
	Forest	3,148,252.000	50.010	855.000	55.910	1.118	0.364	−0.160		
	Barren Land	301,072.000	4.790	104.000	6.800	1.420	0.463	−0.155		
		6,294,626.000		1529.000				−0.713	3.700	

3.1.1. Shannon's Entropy

Shannon's entropy model is an improvement on the frequency ratio model. The frequency ratio model only considers the weightage of sub-factors, and does not consider the weightage of causative factors. Shannon's entropy measures the uncertainty or instability of a system [41]. With landslide susceptibility mapping, it measures the influence of causative factors on the occurrence of landslides [42–45]. Weightage of causative factors is calculated using the following procedure:

$$P_{ij} = FR / \sum_{i=1}^{m} FR \tag{4}$$

$$\sum_{i=1}^{m} E_{ij} = \sum_{i=1}^{m} (P_{ij}) \times (lnP_{ij}) \tag{5}$$

$$H_{ij} = 1 + \sum_{i=1}^{m} E_{ij} \tag{6}$$

$$W_j = H_{ij} / \sum_{i=1}^{n} H_{ij} \tag{7}$$

where Pij is the probability density and FR is the frequency ratio of sub-factors. Wj is the weightage of causative factors obtained from Shannon's entropy technique. These values can then be used for assigning weight to causative factors, while frequency ratio values are used for sub-factors. Table 2 shows Shannon's entropy weightage of causative factors.

3.1.2. Analytic Hierarchy Process (AHP)

Analytic hierarchy process (AHP) is a decision-making tool which helps to solve complex problems with simple criteria. AHP is based upon three principles i.e., decomposition of the problem, comparative judgment, and synthesis of relative importance or rankings [43]. In AHP, the problem is broken down into hierarchical criteria. These criteria are compared to each other. This process of relative comparison is called pair-wise com-

parison. The eigen vector method is used to calculate the rankings and the consistency of the solution is then also checked by calculating the consistency ratio [43]. The consistency of the weights for relative importance assigned during the pairwise comparison can be checked for consistency using the equation given below.

$$Consistency\ Ratio\ (CR) = \frac{CI}{RI} \tag{8}$$

where, CI is consistency index while RI is randomness index.

CI is calculated as follows:

$$Consistency\ Index\ (CI) = \frac{\lambda max - n}{n - 1} \tag{9}$$

where, λmax = Major eigen value and n = order of matrix.

Randomness index values are given by Saaty, which is dependent on the value of n. RI is the result of extensive experimentation on large dataset samples. If the CR values are less than 10%, the pairwise comparison is considered consistent. If the CR value is more than 10%, the solution is considered inconsistent and weights are reassigned in a pairwise comparison matrix. Tables 3 and 4 show the relative comparison matrix of AHP for causative factors and sub-factors.

Table 3. AHP weightage for causative factors.

Causative Factors	1	2	3	4	5	6	7	8	9	Weightage
Slope	1									0.105
Aspect	0.14	1								0.016
Relative relief	0.33	4	1							0.036
TWI	2	5	3	1						0.081
Lithology	0.33	6	3	3	1					0.096
Drainage density	3	7	5	3	3	1				0.205
Distance from road	2	9	7	5	3	3	1			0.28
Distance from faults	3	8	6	4	2	0.33	0.33	1		0.161
Landuse	0.14	2	0.33	0.14	0.14	0.14	0.14	0.17	1	0.02
CR = 0.09										

Table 4. Analytic hierarchy process (AHP) weightage and relative importance of sub-factors.

Conditioning Factor	Classes	1	2	3	4	5	6	7	8	9	Weightage (Wi)
Slope											0.04
	>30	1									0.054
	30–45	2	1								0.102
	45–60	3	3	1							0.209
	60–75	5	5	3	1						0.596
	More than 75	9	8	7	5	1					
					CR = 0.059						
Aspect											
	Flat	1									0.023
	North	2	1								0.047
	Northeast	7	4	1							0.204
	East	4	2	0.33	1						0.082
	South East	9	8	3	4	1					0.362
	South	5	3	0.5	2	0.25	1				0.125
	South West	4	2	0.25	1	0.2	0.5	1			0.078
	West	3	0.5	0.25	0.5	0.14	0.33	0.5	1		0.046
	North West	2	1	0.14	0.33	0.12	0.25	0.33	0.5	1	0.033
					CR = 0.025						

Table 4. *Cont.*

Conditioning Factor	Classes	1	2	3	4	5	6	7	8	9	Weightage (Wi)
Relative relief											
	0–50	1									0.581
	50–100	0.33	1								0.255
	100–150	0.2	0.33	1							0.114
	More than 150	0.11	0.2	0.33	1						0.05
					CR = 0.028						
TWI											
	Low	1									0.565
	Moderate	0.33	1								0.262
	High	0.2	0.33	1							0.118
	Very high	0.14	0.2	0.33	1						0.055
					CR = 0.043						
Lithology											
	Neoproterozoic	1									0.194
	Proterozoic (Undiff.)	0.33	1								0.098
	Mesoproterozoic	0.5	4	1							0.168
	Palaeoproterozoic	5	7	3	1						0.43
	Palaeozoic	0.2	0.33	0.33	0.2	1					0.053
	Paleocene-eocene	0.17	0.2	0.2	0.14	0.5	1				0.033
	Meghalayan	0.2	0.14	0.14	0.11	0.33	0.5	1			0.024
					CR = 0.085						
Drainage density											
	Low	1									0.046
	Moderate	3	1								0.094
	High	5	5	1							0.203
	Very high	9	7	5	1						0.657
					CR = 0.063						
Distance from road											
	Low	1									0.751
	Moderate	0.2	1								0.178
	High	0.11	0.33	1							0.07
					CR = 0.03						
Distance from faults											
	Low	1									0.751
	Moderate	0.2	1								0.178
	High	0.11	0.33	1							0.07
					CR = 0.03						
Landuse/Landcover											
	Snow	1									0.5
	Settlement	0.14	1								0.046
	Agricultural land	0.11	0.5	1							0.034
	Forest	0.2	5	5	1						0.137
	Barren land	0.33	9	7	3	1					0.284
					CR = 0.072						

4. Results and Discussion

4.1. Frequency Ratio

The frequency ratio (FR) of the causative sub-factors is shown in the Table 2. The relative importance of the sub-factors can be evaluated from the weightage calculated in the Table 2. The impact of slopes less than 45° is very low, but as the slope increases above 45°,

the frequency ratio increases sharply. More than 39% of landslides occur in regions where the slope is more than 75°, while such slopes only cover 18% of the study area. Higher slope values trigger the effect of gravity and also increase shear stress. The maximum frequency ratio (FR) value for slopes is 2.163. East, south-east, south, and North Slope aspects affect the occurrence of landslides more compared other directions. The frequency ratio (FR) is highest for the east direction (1.673), followed by the south east direction (1.486). These aspects receive more rainfall and are also subjected to erosion, which makes them more prone to landslides. The areas of low (0–50 m) and high (100–150 m) relative relief were more prone to landslides. The frequency ratio value for relative relief was highest (1.262) for 0–50 m relief, followed by 1.073 for 100–150 m relief. Lower values of topographic wetness index (TWI) have a significant effect on landslide occurrence, with a frequency ratio of 1.043, while very high values of TWI (FR = 0.037) had minimal effect on landslide occurrence. As TWI increases, FR decreases. Neoproterozoic age deposits had the highest FR values (1.252), followed by Mesoproterozoic deposits (FR = 1.653). The frequency ratios follow an increasing trend as the drainage density increases. For very high drainage density, the FR value was 1.685. Smaller distance from roads had a significant effect on the occurrence of landslides. Small distances from roads have the highest FR value i.e., 1.253, and similarly, smaller distances from faults also showed maximum FR (0.994) values in its category. Cutting activities performed during road construction reduced strength at the toe of slopes, making them more prone to landslides. Snow covered land (FR = 2.059) was more prone to landslides, while agricultural land (FR = 0.562) had the least effect on landslide occurrence. After snow cover, barren land due to less vegetative cover was found to be more prone to landslides, with an FR value of 1.420.

The landslide susceptibility index (LSI) for the frequency ratio model varied from 3.173 to 13.297. The landslide susceptibility map is shown in Figure 5. The map is divided using natural breaks in ArcGIS showing very low, low, moderate, high and very high susceptibility zones. It is observed that high and very high landslide susceptibility zones cover around 45% of the study area, but more than 70% landslides occur in these regions. The very high susceptibility zone covers only 16.62% of the total study area but it is subjected to 40.7% of the total landslide area. 11% of the total landslide area occur in very low and low susceptibility zones while these zones cover more than 25% area.

4.2. Shannon's Entropy Model

The weightage of sub-factors in Shannon's entropy model was based on frequency ratio (FR) values. The weightage of causative factors was evaluated from FR values of sub-factors. It is found that distance from faults and distance from roads are the major causative factors which have a very high impact on the occurrence of landslides. The weightage of distance from faults is 0.159 and weightage of vicinity to roads is found to be 0.148. The anthropogenic activities of road cutting and development have significantly increased landslide activities in the study area. The topographic wetness index (TWI) is also an important factor, with a weightage of 0.135 as per Shannon's entropy model. Slope has a weightage of 0.120. Lithology, drainage density, and relative relief have almost equal weightage.

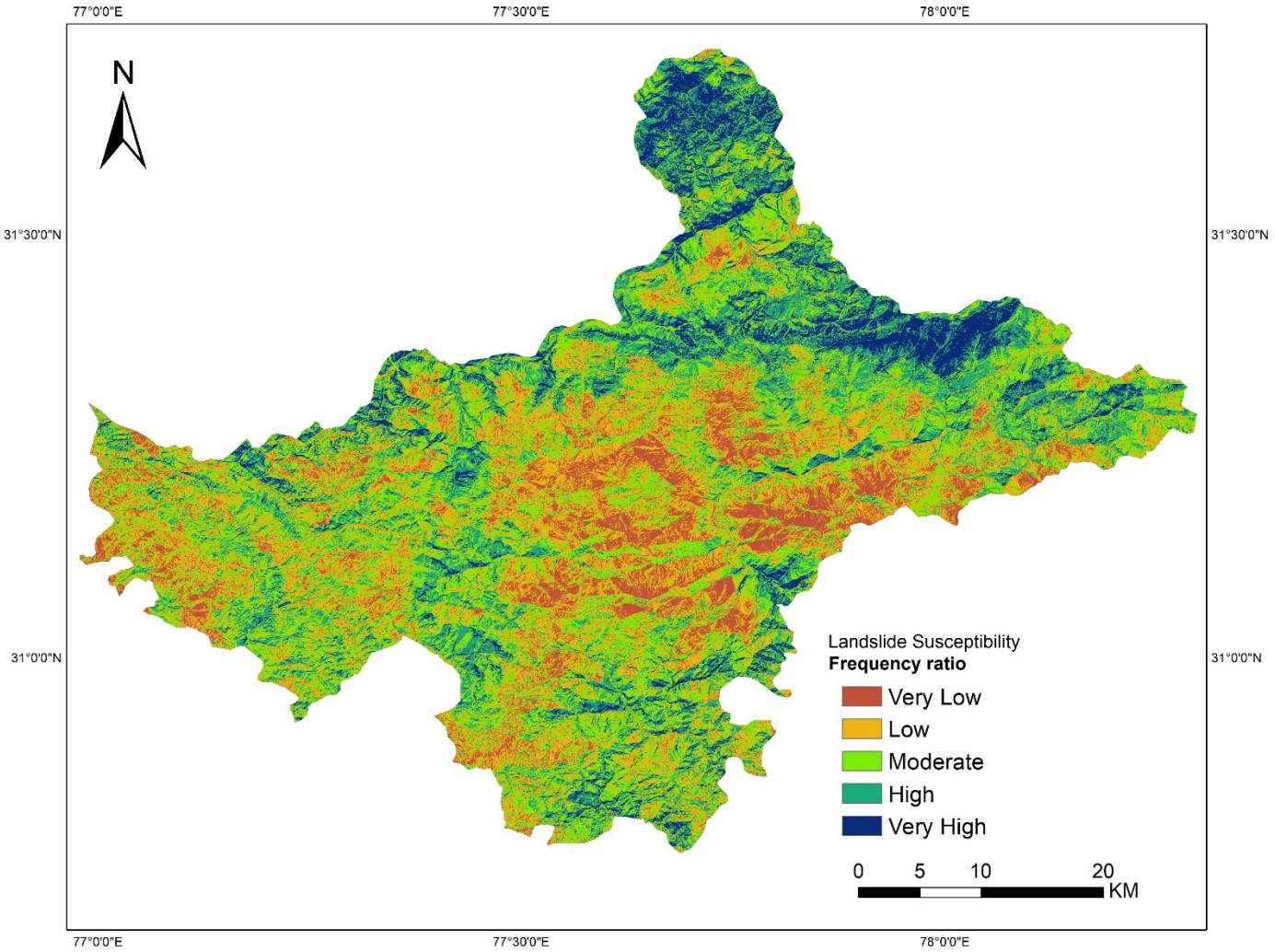


Figure 5. Landslide Susceptibility Map using Frequency Ratio.

The results of the FR model have been changed a little following the implementation of Shannon's entropy weightage to major causative factors. The simplest landslide susceptibility equation for this model is given as follows:

$$\begin{aligned} &Landslide\ Susceptibility\ Index\ (LSI) \\ &\quad = 0.120 \times Slope + 0.033 \times Aspect \\ &\quad + 0.110 \times Relative\ Relief + 0.135 \times TWI \\ &\quad + 0.109 \times Lithology + 0.109 \times Draiage\ Density \\ &\quad + 0.148 \times Distance\ from\ Roads \\ &\quad + 0.159 \times Distance\ from\ Faults + 0.078 \times Land\ use \end{aligned} \tag{10}$$

LSI varies from 0.402 to 1.412 for Shannon's entropy model. The final landslide susceptibility map is divided into five categories, i.e., very low, low, moderate, high and very high, using natural breaks. The landslide susceptibility map by Shannon's entropy is shown in the Figure 6. Shannon's entropy model shows that 12.06% of the area has very high susceptibility, which contains 31.93% landslide area of the region. The model shows that 23.66% of the area has high susceptibility, which contains 34.45% landslide area. Very low and low landslide susceptibility zones cover 9.84% and 24.85% of study area, respectively.

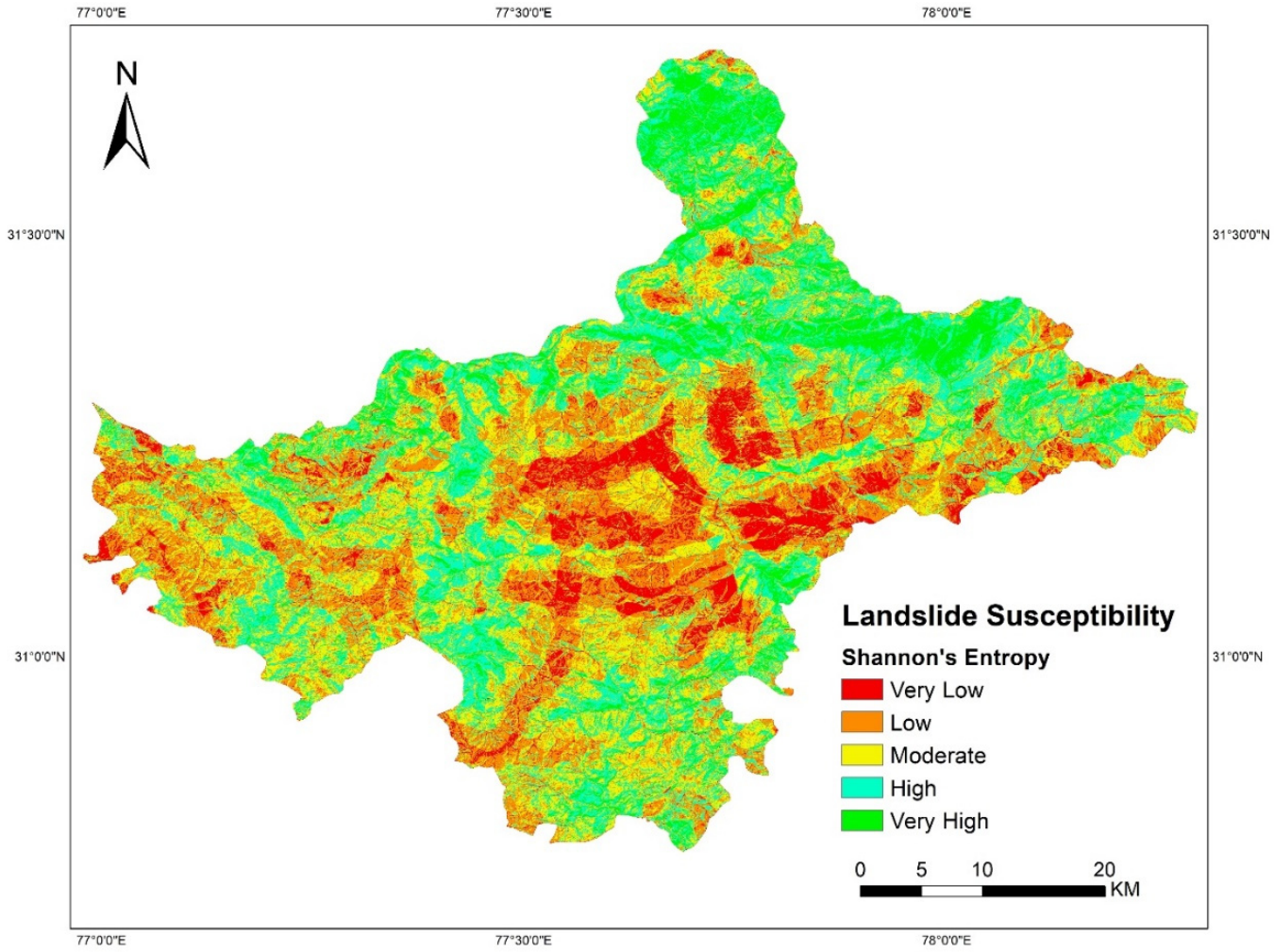


Figure 6. Landslide susceptibility map using Shannon's entropy.

4.3. Analytic Hierarchy Process (AHP) Approach

According to the AHP model, distance from roads and drainage density are the major factors which have high weightage, i.e., 0.280 and 0.205 respectively. Higher weightage of distance from roads shows the significance of anthropogenic activities towards occurrence of landslides. Lithology has a weightage of 0.096 in the AHP model. Neoproterozoic and mesoproterozoic deposits have been more responsible for occurrence of landslides compared to other lithological units, as per the AHP model. Distance from faults has a weightage of 0.161, which states the high significance of this factor in the occurrence of landslides, while aspect has the lowest weightage, i.e., 0.016. Slope (weightage = 0.105) and (weightage = 0.081) are important DEM-based factors which affect the landslide phenomenon. Lower TWI values are more responsible for the occurrence of landslides compared to higher values. Tables 3 and 4 show the AHP weightage for all the factors and sub-factors, respectively.

The landslide susceptibility index (LSI) for the AHP approach is given by following equation:

$$\begin{aligned} &Landslide\ Susceptibility\ Index\ (LSI) \\ &\quad = 0.105 \times Slope + 0.016 \times Aspect \\ &\quad + 0.036 \times Relative\ Relief + 0.081 \times TWI \\ &\quad + 0.096 \times Lithology + 0.205 \times Draiage\ Density \\ &\quad + 0.280 \times Distance\ from\ Roads \\ &\quad + 0.168 \times Distance\ from\ Faults + 0.020 \times Land\ use \end{aligned} \tag{11}$$

The landslide susceptibility index (LSI) for the landslide susceptibility map obtained from the AHP approach varies from 0.061 to 0.648. The landslide susceptibility map is divided by natural breaks based on the landslide susceptibility index shown in Figure 7.

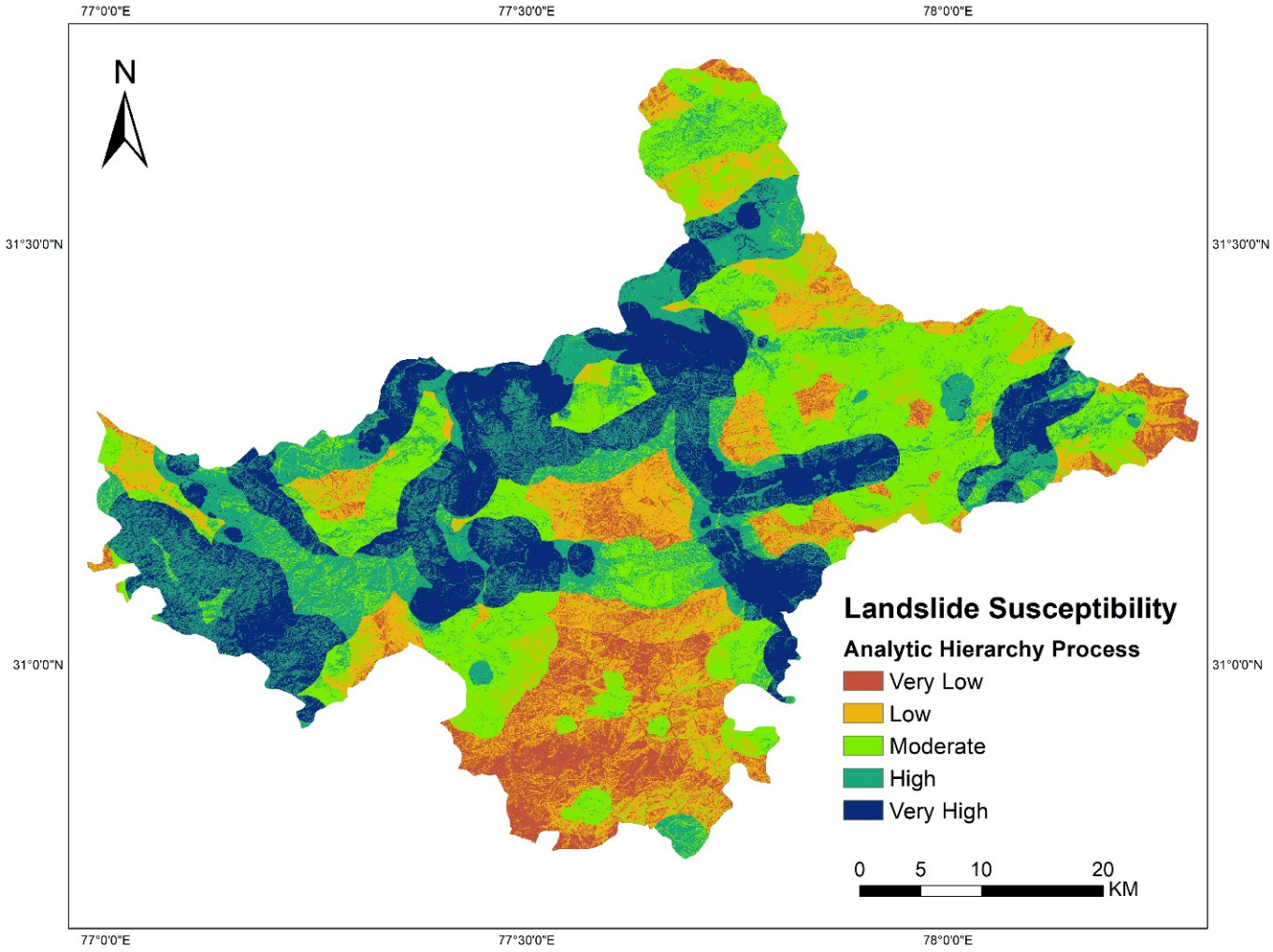


Figure 7. Landslide susceptibility map using analytic hierarchy process.

It is found that 23.05% area lies under very low landslide susceptibility according to the AHP approach, and contains around 10% landslide area. Around 24% area lies under high susceptibility and contains around 29% landslide area, while around 8% area lies under very high susceptibility containing 9.24% of the landslides. It seems evident that the AHP model remains less efficient in the classification of landslide susceptibility as compared to the previous two models. The relative importance of the causative factors has been selected based on the subjective approach of experts. There is no mathematical relationship for the comparison of the relative importance of causative factors and sub-factors. In qualitative techniques such as AHP, the input can be varied for better output. The lower accuracy in the AHP may be due to underestimation and overestimation of the impacts of some causative factors and sub-factors.

4.4. Validation of Results

The results were validated using a receiver operating curve (ROC). Figure 8 shows the ROCs for the landslide susceptibility maps prepared with different techniques. The ROC curves show a relationship between sensitivity and (1-specificity). The area under curve (AUC) value for the frequency ratio model is 0.925, which shows very high accuracy. The frequency ratio model shows very encouraging results in the study area. Shannon's

entropy model shows an accuracy of 0.883. The analytic hierarchy process (AHP) based landslide susceptibility map shows an AUC value of 0.732, which is the lowest among all the models.

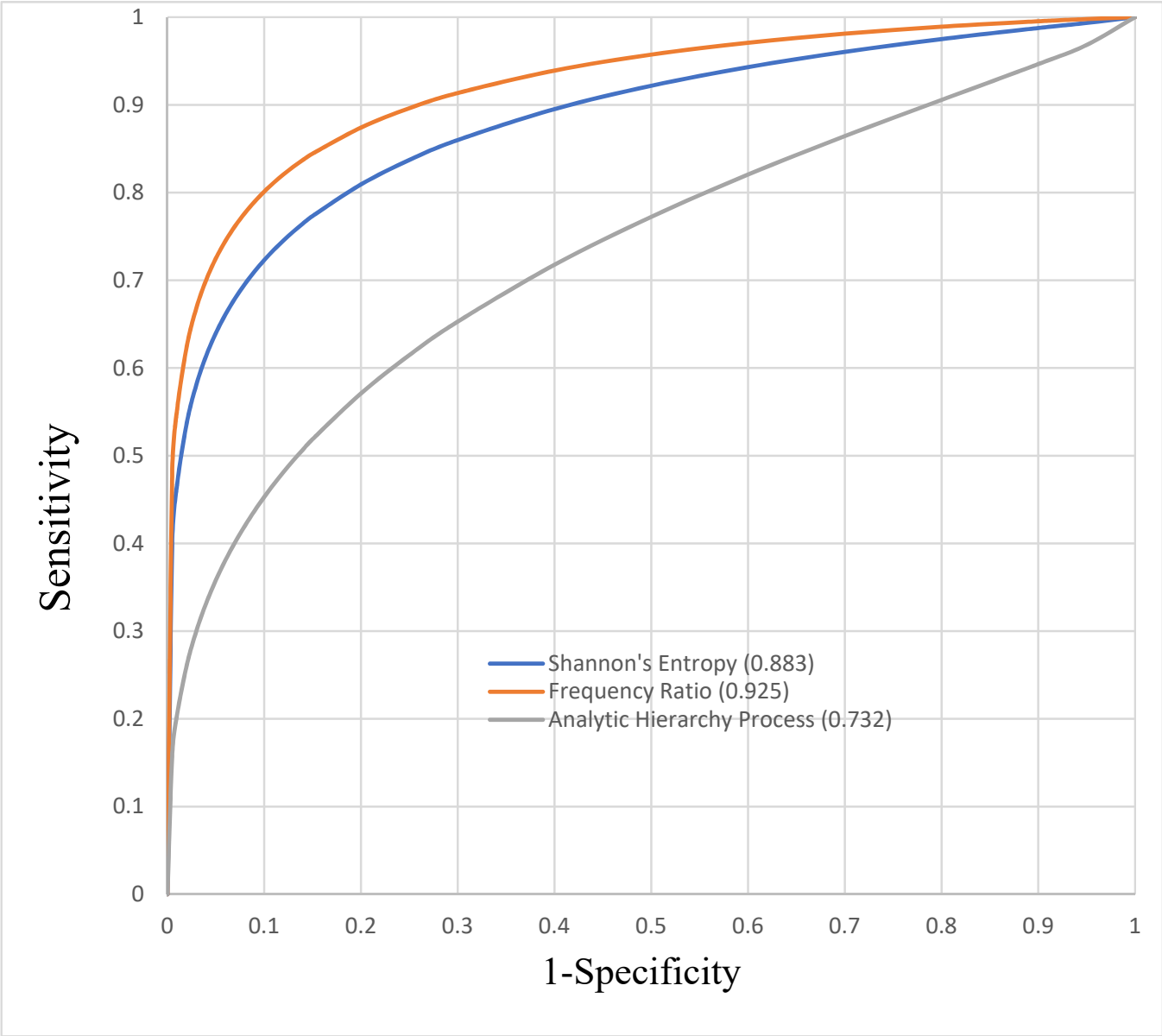


Figure 8. Receiver operating characteristic curve for validation.

5. Conclusions and Future Scope

Landslides are one of the most disastrous phenomena in the Himalayan region. In this study, statistical approaches and expert based approaches are compared for landslide susceptibility mapping in a GIS environment. The frequency ratio (FR) model shows better accuracy i.e., 0.925 compared to Shannon's entropy model and the AHP. Shannon's entropy model had an accuracy of 0.883, and AHP model had an accuracy of 0.732 using the ROC technique. The FR model and Shannon's entropy model have inputs based on landslide inventories which cannot be varied, while in the AHP model, the input can be varied to improve accuracy. The relative importance of each factor and sub-factor can be observed from these models. It can be concluded that frequency ratio is the most efficient method, and is the simplest to implement in quantitative methods. An accurate and extensive

landslide inventory is required for the efficient implementation of frequency ratio models. The results from qualitative methods like AHP depend upon the selection of weights, so need additional research. AHP can be used in the regions for which historical landslide data is not available. There is evidence of rock fall, debris flows and landslides in high and very high susceptibility zones. Landslide susceptibility maps using these models have not previously been prepared in this study area. Therefore, the results of this study can provide a very useful input for the planner and risk managers.

The study could be extended to predict landslide hazard and risk assessment of the study area. The seasonal variation in causative factors such as rainfall, land-use variation and vegetation cover variation can be studied and their impact on the occurrence of landslides can be observed.

Author Contributions: Conceptualization: Sandeep Panchal, Amit K. Shrivastava, Formal Analysis: Sandeep Panchal, Investigation: Sandeep Panchal; Resources: Sandeep Panchal, Amit K. Shrivastava; Data Curation; Sandeep Panchal, Amit K. Shrivastava; Writing—Original Draft Preparation: Sandeep Panchal; Writing—Review and Editing: Sandeep Panchal, Amit K. Shrivastava; Figure editing: Sandeep Panchal; Supervision; Amit K. Shrivastava. Both authors have read and agreed to the published version of the manuscript.

Funding: No funding is obtained for this work.

Data Availability Statement: Not applicable.

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

References

1. Aleotti, P.; Chowdhury, R. Landslide hazard assessment: Summary review and new perspectives. *Bull. Int. Assoc. Eng. Geol.* **1999**, *58*, 21–44. [CrossRef]
2. Lee, S. Application and verification of fuzzy algebraic operators to landslide susceptibility mapping. *Environ. Earth Sci.* **2006**, *52*, 615–623. [CrossRef]
3. Xu, C.; Xu, X.; Yu, G. Landslides triggered by slipping-fault-generated earthquake on a plateau: An example of the 14 April 2010, Ms 7.1, Yushu, China earthquake. *Landslides* **2012**, *10*, 421–431. [CrossRef]
4. Mersha, T.; Meten, M. GIS-based landslide susceptibility mapping and assessment using bivariate statistical methods in Simada area, northwestern Ethiopia. *Geoenviron. Disasters* **2020**, *7*, 20. [CrossRef]
5. Khan, H.; Shafique, M.; Khan, M.A.; Bacha, M.A.; Shah, S.U.; Calligaris, C. Landslide susceptibility assessment using Frequency Ratio, a case study of northern Pakistan. *Egypt. J. Remote Sens. Space Sci.* **2018**, *22*, 11–24. [CrossRef]
6. Meten, M.; Bhandary, N.P.; Yatabe, R. GIS-based frequency ratio and logistic regression modelling for landslide susceptibility mapping of Debre Sina area in central Ethiopia. *J. Mt. Sci.* **2015**, *12*, 1355–1372. [CrossRef]
7. Van Westen, C.J.; Rengers, N.; Soeters, R. Use of Geomorphological Information in Indirect Landslide Susceptibility Assessment. *Nat. Hazards* **2003**, *30*, 399–419. [CrossRef]
8. Prasad, A.S.; Pandey, B.W.; Leimgruber, W.; Kunwar, R.M. Mountain hazard susceptibility and livelihood security in the upper catchment area of the river Beas, Kullu Valley, Himachal Pradesh, India. *Geoenviron. Disasters* **2016**, *3*, 1. [CrossRef]
9. Gokceoglu, C.; Aksoy, H. Landslide susceptibility mapping of the slopes in the residual soils of the Mengen region (Turkey) by deterministic stability analyses and image processing techniques. *Eng. Geol.* **1996**, *44*, 147–161. [CrossRef]
10. Lee, S.; Ryu, J.-H.; Won, J.-S.; Park, H.-J. Determination and application of the weights for landslide susceptibility mapping using an artificial neural network. *Eng. Geol.* **2003**, *71*, 289–302. [CrossRef]
11. Yeon, Y.-K.; Han, J.-G.; Ryu, K.H. Landslide susceptibility mapping in Injae, Korea, using a decision tree. *Eng. Geol.* **2010**, *116*, 274–283. [CrossRef]
12. Pourghasemi, H.R.; Mohammady, M.; Pradhan, B. Landslide susceptibility mapping using index of entropy and conditional probability models in GIS: Safarood Basin, Iran. *Catena* **2012**, *97*, 71–84. [CrossRef]
13. Singh, K.; Kumar, V. Hazard assessment of landslide disaster using information value method and analytical hierarchy process in highly tectonic Chamba region in bosom of Himalaya. *J. Mt. Sci.* **2018**, *15*, 808–824. [CrossRef]
14. Gautam, P.; Kubota, T.; Sapkota, L.M.; Shinohara, Y. Landslide susceptibility mapping with GIS in high mountain area of Nepal: A comparison of four methods. *Environ. Earth Sci.* **2021**, *80*, 359. [CrossRef]
15. Hasekioğulları, G.D.; Ercanoglu, M. A new approach to use AHP in landslide susceptibility mapping: A case study at Yenice (Karabuk, NW Turkey). *Nat. Hazards* **2012**, *63*, 1157–1179. [CrossRef]

16. Kayastha, P.; Dhital, M.; De Smedt, F. Application of the analytical hierarchy process (AHP) for landslide susceptibility mapping: A case study from the Tinau watershed, west Nepal. *Comput. Geosci.* **2013**, *52*, 398–408. [CrossRef]
17. El Jazouli, A.; Barakat, A.; Khellouk, R. GIS-multicriteria evaluation using AHP for landslide susceptibility mapping in Oum Er Rbia high basin (Morocco). *Geoenviron. Disasters* **2019**, *6*, 3. [CrossRef]
18. Sujatha, E.R.; Sridhar, V. Mapping debris flow susceptibility using analytical network process in Kodaikkanal Hills, Tamil Nadu (India). *J. Earth Syst. Sci.* **2017**, *126*, 116. [CrossRef]
19. Shahabi, H.; Ahmad, B.B.; Khezri, S. Evaluation and comparison of bivariate and multivariate statistical methods for landslide susceptibility mapping (case study: Zab basin). *Arab. J. Geosci.* **2012**, *6*, 3885–3907. [CrossRef]
20. Kavzoglu, T.; Sahin, E.K.; Colkesen, I. An assessment of multivariate and bivariate approaches in landslide susceptibility mapping: A case study of Duzkoy district. *Nat. Hazards* **2014**, *76*, 471–496. [CrossRef]
21. Ram, P.; Gupta, V.; Devi, M.; Vishwakarma, N. Landslide susceptibility mapping using bivariate statistical method for the hilly township of Mussoorie and its surrounding areas, Uttarakhand Himalaya. *J. Earth Syst. Sci.* **2020**, *129*, 167. [CrossRef]
22. Pradhan, B. Landslide susceptibility mapping of a catchment area using frequency ratio, fuzzy logic and multivariate logistic regression approaches. *J. Indian Soc. Remote Sens.* **2010**, *38*, 301–320. [CrossRef]
23. Sarkar, S.; Roy, A.; Martha, T.R. Landslide susceptibility assessment using Information Value Method in parts of the Darjeeling Himalayas. *J. Geol. Soc. India* **2013**, *82*, 351–362. [CrossRef]
24. Akgun, A.; Erkan, O. Landslide susceptibility mapping by geographical information system-based multivariate statistical and deterministic models: In an artificial reservoir area at Northern Turkey. *Arab. J. Geosci.* **2016**, *9*, 165. [CrossRef]
25. Pradhan, B.; Oh, H.-J.; Buchroithner, M. Weights-of-evidence model applied to landslide susceptibility mapping in a tropical hilly area. *Geomat. Nat. Hazards Risk* **2010**, *1*, 199–223. [CrossRef]
26. Polykretis, C.; Chalkias, C. Comparison and evaluation of landslide susceptibility maps obtained from weight of evidence, logistic regression, and artificial neural network models. *Nat. Hazards* **2018**, *93*, 249–274. [CrossRef]
27. Shano, L.; Raghuvanshi, T.K.; Meten, M. Landslide susceptibility evaluation and hazard zonation techniques—A review. *Geoenviron. Disasters* **2020**, *7*, 1–19. [CrossRef]
28. Mathews, R.P. Ground Information Water Booklet Shimla District, Himachal Pradesh, Ministry of Water Resources, Central Ground Water Board, Government of India. 2013. Available online: http://cgwb.gov.in/District_Profile/HP/Shimla.pdf (accessed on 1 April 2021).
29. District Survey Document Shimla. Available online: https://emerginghimachal.hp.gov.in/miningstone/survay_docs/shimla.pdf (accessed on 3 May 2021).
30. Guzzetti, F.; Mondini, A.C.; Cardinali, M.; Fiorucci, F.; Santangelo, M.; Chang, K.-T. Landslide inventory maps: New tools for an old problem. *Earth Sci. Rev.* **2012**, *112*, 42–66. [CrossRef]
31. Regmi, A.D.; Yoshida, K.; Pourghasemi, H.R.; DhitaL, M.R.; Pradhan, B. Landslide susceptibility mapping along Bhalubang—Shiwapur area of mid-Western Nepal using frequency ratio and conditional probability models. *J. Mt. Sci.* **2014**, *11*, 1266–1285. [CrossRef]
32. Saleem, N.; Huq, E.; Twumasi, N.Y.D.; Javed, A.; Sajjad, A. Parameters Derived from and/or Used with Digital Elevation Models (DEMs) for Landslide Susceptibility Mapping and Landslide Risk Assessment: A Review. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 545. [CrossRef]
33. Bahrami, S.; Rahimzadeh, B.; Khaleghi, S. Analyzing the effects of tectonic and lithology on the occurrence of landslide along Zagros ophiolitic suture: A case study of Sarv-Abad, Kurdistan, Iran. *Bull. Int. Assoc. Eng. Geol.* **2019**, *79*, 1619–1637. [CrossRef]
34. Nebeokike, U.C.; Igwe, O.; Egbueri, J.C.; Ifediegwu, S.I. Erodibility characteristics and slope stability analysis of geological units prone to erosion in Udi area, southeast Nigeria. *Model. Earth Syst. Environ.* **2020**, *6*, 1061–1074. [CrossRef]
35. Nepal, N.; Chen, J.; Chen, H.; Wang, X.; Sharma, T.P.P. Assessment of landslide susceptibility along the Araniko Highway in Poiqu/Bhote Koshi/Sun Koshi Watershed, Nepal Himalaya. *Prog. Disaster Sci.* **2019**, *3*, 100037. [CrossRef]
36. Sen, S.; Mitra, S.; Debbarma, C.; De, S.K. Impact of faults on landslide in the Atharamura Hill (along the NH 44), Tripura. *Environ. Earth Sci.* **2014**, *73*, 5289–5298. [CrossRef]
37. Yang, L.; Liu, E. Numerical Analysis of the Effects of Crack Characteristics on the Stress and Deformation of Unsaturated Soil Slopes. *Water* **2020**, *12*, 194. [CrossRef]
38. Demir, G.; Aytekin, M.; Akgun, A. Landslide susceptibility mapping by frequency ratio and logistic regression methods: An example from Niksar–Resadiye (Tokat, Turkey). *Arab. J. Geosci.* **2014**, *8*, 1801–1812. [CrossRef]
39. Park, S.; Choi, C.; Kim, B.; Kim, J. Landslide susceptibility mapping using frequency ratio, analytic hierarchy process, logistic regression, and artificial neural network methods at the Inje area, Korea. *Environ. Earth Sci.* **2012**, *68*, 1443–1464. [CrossRef]
40. Silalahi, F.E.S.; Arifianti, Y.; Hidayat, F. Landslide susceptibility assessment using frequency ratio model in Bogor, West Java, Indonesia. *Geosci. Lett.* **2019**, *6*, 10. [CrossRef]
41. Roodposhti, M.S.; Aryal, J.; Shahabi, H.; Safarrad, T. Fuzzy Shannon Entropy: A Hybrid GIS-Based Landslide Susceptibility Mapping Method. *Entropy* **2016**, *18*, 343. [CrossRef]
42. Sharma, L.P.; Patel, N.; Ghose, M.K.; Debnath, P. Development and application of Shannon's entropy integrated information value model for landslide susceptibility assessment and zonation in Sikkim Himalayas in India. *Nat. Hazards* **2014**, *75*, 1555–1576. [CrossRef]

43. Saaty, R.W. The analytic hierarchy process—what it is and how it is used. *Math. Model.* **1987**, *9*, 161–176. [CrossRef]
44. Zhu, L.; Huang, J.-F. GIS-based logistic regression method for landslide susceptibility mapping in regional scale. *J. Zhejiang Univ. A* **2006**, *7*, 2007–2017. [CrossRef]
45. Map of India. Available online: https://d-maps.com/ (accessed on 2 January 2021).

A Comparative Study of Machine Learning and NLP Techniques for Uses of Stop Words by Patients in Diagnosis of Alzheimer's Disease

Surabhi Adhikari[1], Surendrabikram Thapa[1], Priyanka Singh[2], Huan Huo[2], Gnana Bharathy[3], Mukesh Prasad[2]

[1] Department of Computer Science and Engineering, Delhi Technological University, India
[2] School of Computer Science, FEIT, University of Technology Sydney, Australia
[3] School of Information, Systems and Modeling, FEIT, University of Technology Sydney, Australia

***Abstract*—**Alzheimer's Disease (AD) is one of the most common forms of neuropsychological disorder in elderly people. It is a slow progressive disease affecting the brain cells. This affects the cognitive abilities of people and their daily activities. During the course of the disease, memory gets brutally affected too. Working as well as long-term declarative memory deteriorates in AD patients. Due to this deterioration of the memory, AD patients tend to show a decline in their communicative skills as well. This decline is reflected in their speech. AD patients usually have poor grammar along with very low coherent ideas. Also, they tend to repeat the words very often and hence become unclear on the message they are trying to convey. As the disease progresses, the speech is completely impaired, and the patients are left to sing or utter words that are totally out of context. Stopwords are the words that are most commonly used in language and it is often hypothesized that AD patients use them much often as compared to Control Normal (CN) subjects. It is seen that due to the degeneration of brain cells in AD patients, they have a tendency to use a lot of stopwords to fill their perplexities in their statements. In this paper, the usefulness of the stopwords in capturing the linguistic information of the patients suffering from AD are discussed. Learning algorithms are evaluated by including stopwords and dropping stopwords at preprocessing to draw comparisons.

Keywords*—Alzheimer's Disease, Natural Language Processing, Data Preprocessing, Feature Extraction*

I. INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease that is characterized by a progressive decline in cognitive domains, such as language, personality, memory, and behavior, which hampers the ability of a patient to perform daily activities [1]. There isn't any clinical cure to delay the progression or completely reverse AD, and hence it is very severe considering the population growing with elderly people. There are different sorts of symptoms AD patients show. This is due to the changes in cortical anatomy [2]. The major effect of AD in a patient is the formation of atrophies in various regions of the brain that degenerates brain cells and hence leads to cognitive impairment. Atrophies in the left anterior temporal lobe, for example, lead to difficulties in naming tasks like picture description tasks [3]. One of the major areas of the brain that gets affected due to AD is the hippocampus, which is responsible for forming memories. The formation of the memories is linked to semantic understanding, allowing individuals to finish sentences, connect words to items, and name objects. Difficulty in communication has been described as one of the most frequent and hardest problems in most of the AD patients [4, 5]. Due to the damage in brain cells, there is a severe condition known as aphasia that is associated with the impairment of language and speech. It affects the production of speech and hence AD patients are unable to comprehend regular activities such as reading or writing. As AD progresses, aphasia worsens. They are not able to process thoughts instantly. It eventually leads to AD patients forgetting the names of their family members, friends, and are even left confused about family relationships [6]. During the early stage, although AD patients speak fluently, they face difficulties finding the right word to use. They usually speak empty speech (For e.g. "the stuff is over there") [7]. Although what they speak is syntactically and morphologically correct, they face semantic and lexical language difficulties. Naming objects can be considered as an example. In this stage, AD patients face loss of verbal fluency, lexical retrieval problems, and inability to comprehend written and spoken languages due to language impairment [7]. Similarly, in the mild stage, AD patients are not able to remember their thoughts and hence cannot express them into words [8]. Due to reduced reading, speaking, writing, or mathematical skills, they tend to utter words that are not in line with the ongoing conversation. A person can replace the word with the wrong one or not find a substitute at all [6]. They find difficulties in forming even simple sentences. In severe stages, AD patients tend to stutter a lot, babble or even sing in an ongoing conversation. When speech occurs, it is occasionally incoherent or illogical. In very severe stages of AD, speech is mostly restricted to verbal stereotypy and echolalia [9]. Language impairment may be an early sign of AD mainly because of the rupture caused by frontotemporal dementia (FTD).

Diagnosis of AD and other forms of dementia manually has become a very big challenge. Currently, certain clinical diagnostic measures and neuropsychological test scores are used as means to diagnose AD. The process is very long and costly, and hence the early diagnosis of the disease is still a big hassle. Hence, using Machine learning algorithms for building diagnostic models with low-level features that are collected from verbal utterances can automate the entire process [10-12]. This can help in early and probable diagnosis of AD. Since last decade, image modalities have been used for diagnosis of AD using machine learning methods. Advanced neuroimaging techniques, such as magnetic resonance imaging (MRI), computed tomography (CT) scans, and positron emission tomography (PET), have been in practice to enable the identification of AD-related structural and molecular biomarkers [13]. Due to high dimensional multimodal neuroimaging data, various machine learning and deep learning algorithms have been used excessively to automate and ease the process. Similarly, feature extraction, dimensionality reduction, feature-based classification, and feature selection techniques are applied for better accuracy in classification of AD patients [14]. Over the years, machine learning algorithms get better accuracy by using hand-crafted

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9534449

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

and refined features. For this reason, deep learning techniques attract significant attention in the field of high-dimensional medical imaging analysis [15].

Other than the image, speech modality is also an important factor for the early diagnosis of AD. Speech is indeed the most easily accessible data for analyzing the cognitive status of a person since other medical approaches can be uncomfortable to some extent [16]. The variation in language can be an easy and significant way to delineate AD patients from the cognitively normal group. In this regard, Natural Language Processing (NLP) techniques prove to be a very significant tool in diagnosing AD patients through their transcripts. AD patients use a very reduced vocabulary. They also tend to use verbal filters such as Uhm, uh, etc. repetitively that depict their confused state of mind. Similarly, AD patients are also seen to use words that aren't significant in the context quite often, such as 'and', 'or', 'it'. NLP can hence be leveraged here to distinguish the linguistic cues of AD patients and make it easier for clinicians to diagnose. This study compares the speech transcripts of AD patients with control normal (CN) subjects with respect to the amount of stopwords people from each category use. The stopwords were taken from the NLTK library [17].

This paper mainly focuses on backing the hypothesis that AD patients use stopwords more frequently than CN subjects. The rest of the paper organised as follows: Section II discusses the methodology, Section III discusses the experimental results and finally, Section IV colcludes the findings of the paper along with the future works.

II. Methodology

The process of this study begins with the collection of data, relevant pre-processing steps, feature extraction, training of classifiers, and comparing the results obtained. The performance of the classifiers are analyzed after the training and 10-fold cross-validation on the dataset. Fig. 1 shows the flow diagram for the methodology that is used in this experiment. The experiment is not done with just a single learning model. Since a detailed study of how stopwords affect the classification is being done in the experiment, ten learning models are used in the study as the conclusion drawn from the results that just a single learning model would not strongly compare the effect of stop words.

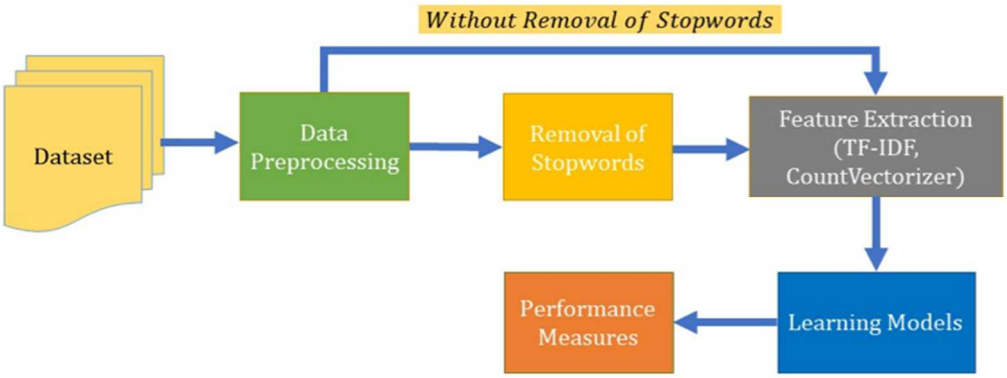


Fig. 1. Setup for the experimental analysis

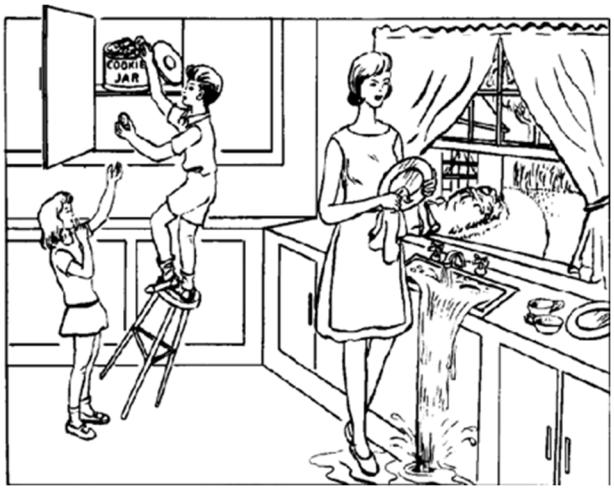


Fig. 2. Boston Cookie theft description task

Table I. Demographics of Participants under study

Attributes	CN Subjects	AD Patients
No. of participants	168	98
Gender	31M / 67F	55M / 113F
Age	64.7 (7.6)	71.2 (8.4)
Education	14.0 (2.3)	12.2 (2.6)

A. Data Collection

The data used for this experiment is collected through the DementiaBank's Pitt Corpus [18]. It consists of audio files and text transcripts of the participants that participated in the research conducted by Becker et al. of the University of Pittsburgh School of Medicine. This study specifically uses the transcripts of audio recordings for the Cookie Theft picture

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

description task as shown in fig. 2. The cookie theft description task was first used by the Boston Diagnostic Aphasia Examination protocol and hence mentioned as Boston Cookie Theft picture description task in literature [31]. There were 292 participants in this study. Among 292 participants, 194 participants had dementia. 309 transcripts in the corpus belong to the dementia category. Among 194 demented patients, there were 168 subjects who had AD whereas 26 participants had other forms of dementia.

The demographic information of the subjects under study are shown in table I. Also, fig. 3 and 4 show the distribution of age and education of the participants under study. The average age of the CN subjects and AD patients are 64.7 and 71.2 years respectively. Similarly, the average education of CN participants and AD subjects are 14.0 years and 12.2 years respectively. This study compares transcripts from both patients with AD as well as control normal (CN) subjects. From the 168 AD patients, 255 CHAT transcripts were used. Similarly, 244 CHAT transcripts from the 98 CN participants of the study were used. Table II shows a sample of the transcript of CN and AD patients.

B. Data Preprocessing

For data preprocessing, noise and unwanted words were removed from the transcripts. All the letters were changed to lowercase, and punctuation marks had been removed too. The words from CHAT transcripts that just showed actions of the interviewer or the subjects, and had no meaning in the real transcripts, had been removed too. For this comparative study, first the models were trained without removing the stopwords. Later, stopwords, from the NLTK library [17] for the English language, were removed from the transcripts. Table II shows an example of the transcripts of CN and AD subjects without removing stopwords. Table III shows an example of preprocessed transcript along with the original transcript.

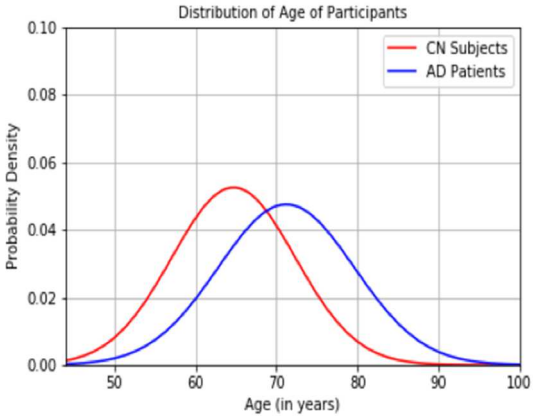


Fig. 3. Distribution of the age of participants

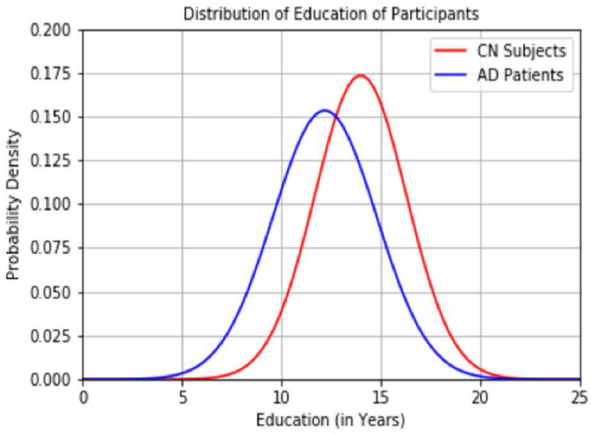


Fig. 4. Distribution of education of the participants

Table II. Example of Transcript of CN and AD patients

Control Normal (CN) Subject		Alzheimer's Disease (AD) Patient	
CHAT Transcript ID	**Transcript**	**CHAT Transcript ID**	**Transcript**
013-2.cha	A boy is getting cookies out of the cookie jar. He's standing on a stool that's gonna fall. The girl is reaching for a cookie. The mother's drying dishes. The faucet's running water. It's dripping out of the sink. Spilling onto the floor. dishes are on the counter. Window is open. Must be summertime. The girl is laughing. Looks like she's laughing. That's about it.	178-1.cha	Taking some cookies. And falling over. And mother's drying dishes. And things falling ... sink's running over. And two cups and saucer on the sink. and ... and the girl's ... curtains on the windows. and the spigot's on the sink. and ...well that's about it.

Table III. Example of Preprocessed Text of ad patient

CHAT Transcript ID	Original Transcript	Preprocessed Transcript
178-1.cha	Taking some cookies. And falling over. And mother's drying dishes. And things falling ... sink's running over. And two cups and saucer on the sink. and ... and the girl's ... curtains on the windows. and the spigot's on the sink. and ...well that's about it.	Taking cookies falling mothers drying dishes things falling sinks running two cups saucer sink girls curtains windows spigots sink well thats

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

C. Feature Extraction

Machine learning models do not understand the exact words and hence cannot be fed into them directly. Therefore, to represent words in the form of vectors, vectorization techniques have been used. Vectorization is used to map words and phrases to vectors to give it as an input to a learning algorithm. In this paper, two vectorization methods viz. CountVectorizer [19] and Term Frequency-Inverse Document Frequency (TF-IDF) [20] were used. CountVectorizer is a method that is used to convert a collection of text documents to a vector of term/token counts. The count vectorizing method gives the accurate count of tokens/ word contents of documents whereas TF-IDF represents the text in a way that gives importance to certain words. It is based on the idea that the words occurring naturally many times should be given less weightage and the importance of words is given on the basis of their presence in the documents.

D. Learning Models

In this experiment, machine learning algorithms as single/ individual leaners (some times simply referred to as machine learning algorithms) and ensemble learners were used for the classification task of AD and CN subjects. Models were trained two times. First, the models were trained without the removal of stopwords, and later, the models were trained after the removal of the stopwords. The single machine learning algorithms used in this study are Naive Bayes (NB) [21], Support Vector Machines (SVM) [22], Decision Tree (DT) [23], K-Nearest Neighbor (KNN) [24], and Logistic Regression (LR) [25]. Similarly, ensemble learners like AdaBoost (ADB) [26], XGBoost (XGB) [27], Random Forest (RF) [28], Light Gradient Boosting Method (LGBM) [29] and Categorical Boosting (CatBoost) [30] were used. 10-fold cross-validation (CV) is done for validating each learning model.

E. Performance Measures

The performance of the above-mentioned models was evaluated using accuracy, precision, recall, and F1-score. 10-fold cross-validation has been used to validate the results. The expressions of the evaluation metrics are given below:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Apart from accuracy, precision, recall, and f1-score, the area under the curve (AUC) has also been calculated. AUC is the performance measure that depicts the ability of a classification algorithm to distinguish between the given classes and is used as a summary of the ROC curve. The higher value of the area under the curve (AUC) signifies the better performance of the model at the classification of the positive and negative classes. The performance of the models was compared using the given evaluation metrics.

III. Experimental Results

From the transcripts, a statistical visualization of the number of stopwords used by CN and AD subjects has been shown in the figures below. The simulations were done in an Intel Core i5 machine with 8 GB RAM equipped with 2 GB of AMD graphics. The codes were run in python 3.6.7 on Ubuntu 16.04.6 Operating System. Fig. 5 shows a normal distribution curve of the number of stop words used by AD patients. Similarly, fig. 6 shows a normal distribution curve of the number of stop words used by CN subjects. Fig. 7 on the other hand, gives a comparison of the distribution of the stop words used by CN and AD subjects. Table IV shows the probability and mean of stopwords used by AD patients in every 100 words. Similarly, Table V shows the probability of the number of stopwords used by CN subjects per 100 words.

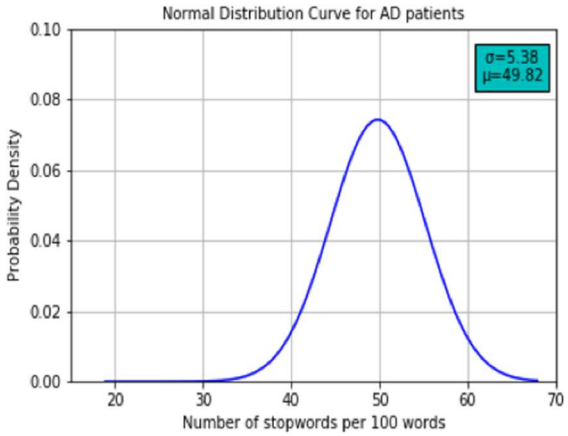


Fig. 5. Normal Distribution Curve for the number of stopwords per 100 words used by AD patients

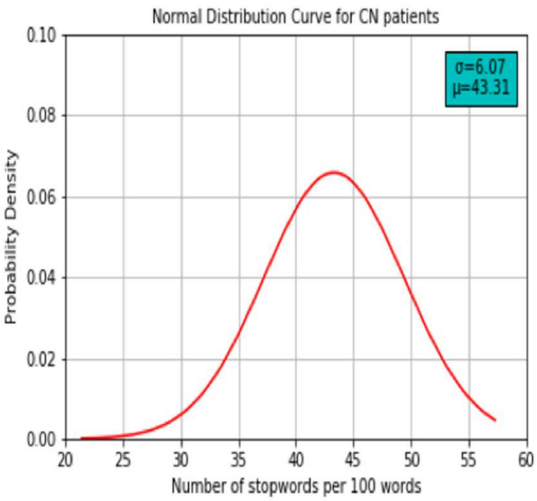


Fig. 6. Normal Distribution Curve for the number of stopwords per 100 words used by CN subjects

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

Table IV. Probability and the mean of number of stopwords used by AD patients per 100 words

Probability	No. of stopwords per 100 words
68%	49.82 ± 5.38
95%	49.82 ± 10.76
99.7%	49.82 ± 16.14

Table V. Probability and the mean of number of stopwords used by CN subjects per 100 words

Probability	No. of stopwords per 100 words
68%	43.31 ± 6.07
95%	43.31 ± 12.14
99.7%	43.31 ± 18.21

From the given tables IV and V, it is seen that the AD patients use stopwords more frequently than CN subjects on average. AD patients on average use 49.82 stopwords per 100 words. Similarly, CN subjects use 43.31 stopwords per 100 words on average. Further statistical analysis of the comparative number of stopwords used by Control Normal subjects and patients with Alzheimer's disease can be found in table IV and V. The use of stopwords in greater number by AD patients than that of CN subjects is further backed by the simulation results obtained from different machine learning algorithms as discussed.

From the normal distribution curve in fig. 5, it can be deciphered that 68% of the AD patients used stopwords in the range of 44.44 per 100 words to 55.2 per 100 words. Similarly, for CN subjects, 68% of the CN subjects used stopwords in the range of 37.27 per 100 words to 49.38 per 100 words which can be shown in fig. 6. The statistics show that stopwords is used more often by the patients with Alzheimer's disease. It is thus hypothesized that the stopwords preserve linguistic information of the AD patients and help in better classification.

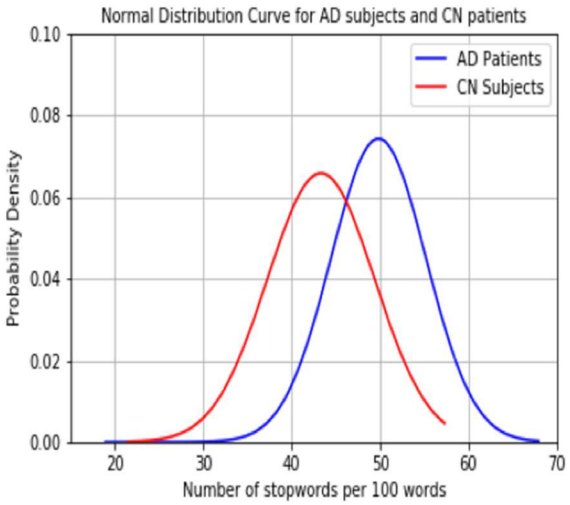


Fig. 7. A comparison of the normal distribution curves of the number of stopwords used by CN and AD subjects.

The results of various models with and without removing stopwords and using the two feature extraction methods viz. TF-IDF vectorizer and Count vectorizer can be found in Table VI and VII respectively. Table VI shows results with TF-IDF Vectorizer with stopwords and without stopwords. The best performing model without removing stopwords from the transcripts is found to be Support Vector Machines (SVM) with the highest F1-score of 0.840. Similarly, with TF-IDF and without stopwords, the SVM model performed the best with an F1-score of 0.820.

Table VI. Performance of ML Algorithms with Tf-Idf Vectorizer with stopwords and without stopwords

Learning Algorithms		With 'Stopwords'				Without 'Stopwords'			
		Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-score
Machine Learning Algorithms as Single/ Individual Learners	Multinomial NB	0.819	0.820	0.820	0.820	0.799	0.820	0.800	0.800
	SVM	0.843	0.840	0.840	0.840	0.824	0.820	0.820	0.820
	Decision Trees	0.705	0.710	0.710	0.710	0.658	0.660	0.660	0.660
	KNN	0.745	0.760	0.750	0.740	0.734	0.760	0.750	0.749
	Logistic Regression	0.832	0.830	0.830	0.830	0.832	0.830	0.830	0.830
Ensemble Learners	XGBoost	0.808	0.810	0.810	0.810	0.774	0.770	0.770	0.770
	AdaBoost	0.792	0.790	0.790	0.790	0.769	0.770	0.770	0.770
	Random Forest	0.745	0.750	0.750	0.750	0.742	0.760	0.740	0.749
	LGBM	0.798	0.800	0.800	0.800	0.794	0.790	0.790	0.790
	CatBoost	0.824	0.820	0.820	0.820	0.816	0.820	0.820	0.820

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

Table VII. Performance of ML Algorithms with CountVectorizer with stopwords and without stopwords

Learning Algorithms		With 'Stopwords'				Without 'Stopwords'			
		Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-score
Machine Learning Algorithms as Single/ Individual Learners	Multinomial NB	0.824	0.830	0.820	0.820	0.814	0.820	0.810	0.810
	SVM	0.792	0.790	0.790	0.790	0.732	0.760	0.730	0.720
	Decision Trees	0.685	0.690	0.690	0.690	0.694	0.690	0.690	0.690
	KNN	0.756	0.760	0.760	0.760	0.754	0.750	0.750	0.750
	Logistic Regression	0.822	0.820	0.820	0.820	0.818	0.820	0.820	0.820
Ensemble Learners	XGBoost	0.812	0.810	0.810	0.810	0.796	0.800	0.800	0.800
	AdaBoost	0.776	0.780	0.780	0.780	0.774	0.770	0.770	0.770
	Random Forest	0.764	0.770	0.760	0.760	0.763	0.770	0.760	0.760
	LGBM	0.774	0.770	0.760	0.70	0.756	0.760	0.750	0.750
	CatBoost	0.815	0.820	0.820	0.820	0.794	0.790	0.790	0.790

Table VIII. Accuracy of various models with stopwords and without stopwords

Learning Algorithms		TF-IDF Vectorizer		CountVectorizer	
		With 'Stop Words'	Without 'Stop Words'	With 'Stop Words'	Without 'Stop Words'
Machine Learning Algorithms as Single/ Individual Learners	Multinomial Naïve Bayes	0.822	0.802	0.824	0.814
	Support Vector Machines	0.838	0.824	0.792	0.728
	Decision Tree	0.706	0.658	0.686	0.693
	KNN	0.748	0.731	0.757	0.753
	Logistic Regression	0.832	0.832	0.822	0.818
Ensemble Learners	XGBoost	0.808	0.773	0.812	0.796
	AdaBoost	0.792	0.769	0.778	0.774
	Random Forest	0.747	0.742	0.765	0.765
	LGBM	0.798	0.794	0.766	0.753
	CatBoost	0.824	0.816	0.816	0.794

Similarly, from table VII, the best performing model with CountVectorizer with stopwords is seen to be Multinomial NB that achieved an F1-score of 0.820 and the highest accuracy of 0.824. Similarly, without stopwords, Logistic Regression performed the best with an F1-score of 0.820 with CountVectorizer.

Table VIII shows the comparison of AUC values with and without 'Stopwords' for different algorithms using both TF-IDF Vectorizer and CountVectorizer. In all the cases, the AUC is higher for the classification that included stopwords.

From the results, it can be concluded that better accuracy is achieved when stopwords are not removed from the transcripts. It can further be deciphered that since there is degeneration of speech in AD patients as the disease progresses, they use more stopwords than the CN group of people. Hence, the models described above were able to classify the AD and CN transcripts with better accuracy when the stopwords were not removed from the transcripts.

Fig. 8 shows the AUC values with TF-IDF for different machine learning algorithms. Similarly, fig. 9 shows the AUC values with CountVectorizer with removal and without removal of stopwords.

IV. Conclusion

It is important to be able to make the entire diagnosis process of AD smooth and effective. Also, as the disease progresses, the symptoms worsen and patients have severe difficulties due to their declining cognition. Hence, it is really important to be able to make an early diagnosis of such neurodegenerative disorder. It is also seen that speech is the most easily available data from AD patients. Hence, using the speech of AD patients models can be built that help in the early diagnosis of the disease. The comparative analysis in this study helps to picturize the differences in the speech of AD patients and CN subjects. The hypothesis that stopwords can be phenomenal for the speech analysis of AD patients is strongly supported by the results and hence including them in analysis can make models more accurate. It is important to find out the linguistic cues in people's speech since speech disorder seems to be one of the earliest symptoms of AD. This study can further help researchers in the field to understand some specific language differences that can be used in the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

predictive analysis of AD. The work can further be extended by using acoustic features such as the length of pause a person takes, the number of times a person repeats a word and including them in building the models. Also, a speech synthesis system can be built in order to help the clinicians in the delineation of AD patients.

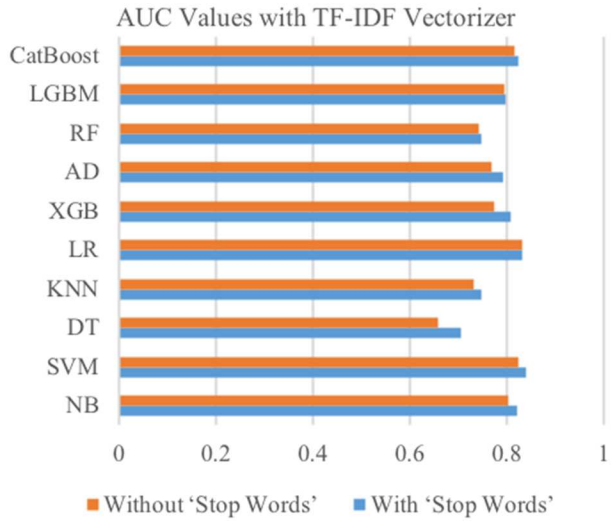


Fig. 8. Comparison of AUC values with and without stopwords with TF-IDF vectorizer

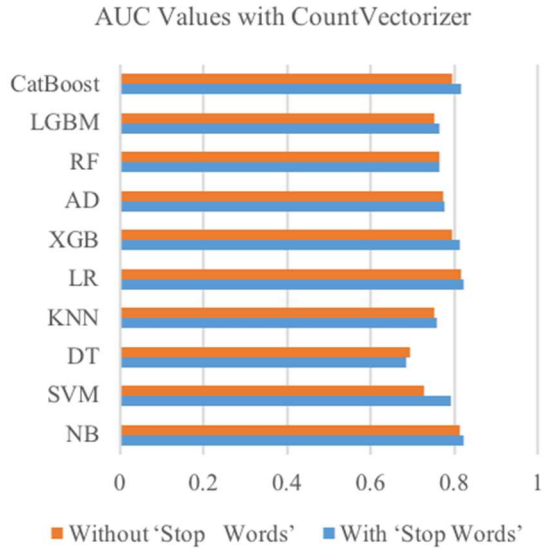


Fig. 9. Comparison of AUC values with and without stopwords with CountVectorizer

REFERENCES

[1] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease," in 2020 International Joint Conference on Neural Networks (IJCNN), 2020: IEEE, pp. 1-8.

[2] B. C. Dickerson et al., "The cortical signature of Alzheimer's disease: regionally specific cortical thinning relates to symptom severity in very mild to mild AD dementia and is detectable in asymptomatic amyloid-positive individuals," Cerebral cortex, vol. 19, no. 3, pp. 497-510, 2009.

[3] K. Domoto-Reilly, D. Sapolsky, M. Brickhouse, B. C. Dickerson, and A. D. N. Initiative, "Naming impairment in Alzheimer's disease is associated with left anterior temporal lobe atrophy," Neuroimage, vol. 63, no. 1, pp. 348-355, 2012.

[4] M. Egan, D. Bérubé, G. Racine, C. Leonard, and E. Rochon, "Methods to enhance verbal communication between individuals with Alzheimer's disease and their formal and informal caregivers: a systematic review," International Journal of Alzheimer's Disease, vol. 2010, 2010.

[5] D. J. Braun and L. J. Van Eldik, "In vivo brainstem imaging in Alzheimer's disease: potential for biomarker development," Frontiers in aging neuroscience, vol. 10, p. 266, 2018.

[6] L. Miller, "Effective communication with older people," Nursing Standard, vol. 17, no. 9, p. 45, 2002.

[7] S. H. Ferris and M. Farlow, "Language impairment in Alzheimer's disease and benefits of acetylcholinesterase inhibitors," Clinical interventions in aging, vol. 8, p. 1007, 2013.

[8] B. Klimova, P. Maresova, M. Valis, J. Hort, and K. Kuca, "Alzheimer's disease and language impairments: social intervention and medical treatment," Clinical interventions in aging, vol. 10, p. 1401, 2015.

[9] D. F. Tang-Wai and N. L. Graham, "Assessment of Language Function in Dementia," Geriatrics, vol. 11, no. 2, pp. 103-110, 2008.

[10] E. Moradi, A. Pepe, C. Gaser, H. Huttunen, J. Tohka, and A. D. N. Initiative, "Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects," Neuroimage, vol. 104, pp. 398-412, 2015.

[11] E. Pellegrini et al., "Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review," Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, vol. 10, pp. 519-535, 2018.

[12] M. Tanveer, B. Richhariya, R. U. Khan, A. H. Rashid, P. Khanna, M. Prasad, C. T. Lin, "Machine learning techniques for the diagnosis of Alzheimer's disease: A review," ACM Transactions on Multimedia Computing, Communications, and Applications, 2019.

[13] D. P. Veitch et al., "Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's disease neuroimaging initiative," Alzheimer's & Dementia, vol. 15, no. 1, pp. 106-152, 2019.

[14] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in Alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data," Frontiers in aging neuroscience, vol. 11, p. 220, 2019.

[15] S. Thapa, S. Adhikari, U. Naseem, P. Singh, G. Bharathy, and M. Prasad, "Detecting Alzheimer's Disease by Exploiting Linguistic Information from Nepali Transcript," in International Conference on Neural Information Processing, 2020: Springer, pp. 176-184.

[16] T. G. Beach, S. E. Monsell, L. E. Phillips, and W. Kukull, "Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010," Journal of neuropathology and experimental neurology, vol. 71, no. 4, pp. 266-273, 2012.

[17] E. Loper and S. Bird, "NLTK: the natural language toolkit," arXiv preprint cs/0205028, 2002.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis," Archives of Neurology, vol. 51, no. 6, pp. 585-594, 1994.

[19] N. Buhagiar, B. Zahir, and A. Abhari, "Using deep learning to recommend discussion threads to users in an online forum," in 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1-8: IEEE.

[20] T. Peng, L. Liu, and W. Zuo, "PU text classification enhanced by term frequency‐inverse document frequency - improved weighting," Concurrency and computation: practice and experience, vol. 26, no. 3, pp. 728-741, 2014.

[21] S. Rajora, D. L. Li, C. Jha, N. Bharill, O. P. Patel, S. Joshi, D. Puthal, M. Prasad, "A comparative study of machine learning techniques for credit card fraud detection based on time variance," in 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018, pp. 1958-1963: IEEE.

[22] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

[23] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3, pp. 660-674, 1991.

[24] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, D. W. Ping, C. T. Lin, "A review of clustering techniques and developments," Neurocomputing, vol. 267, pp. 664-681, 2017.

[25] D. Pregibon, "Logistic regression diagnostics," The Annals of Statistics, vol. 9, no. 4, pp. 705-724, 1981.

[26] R. E. Schapire, "Explaining adaboost," in Empirical inference: Springer, 2013, pp. 37-52.

[27] R. P. Sheridan, W. M. Wang, A. Liaw, J. Ma, and E. M. Gifford, "Extreme gradient boosting as a method for quantitative structure–activity relationships," Journal of chemical information and modeling, vol. 56, no. 12, pp. 2353-2360, 2016.

[28] M. Pal, "Random forest classifier for remote sensing classification," International journal of remote sensing, vol. 26, no. 1, pp. 217-222, 2005.

[29] P. Chun, S. Izumi, and T. Yamane, "Automatic detection method of cracks from concrete surface imagery using two - step light gradient boosting machine," Computer - Aided Civil and Infrastructure Engineering

[30] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in neural information processing systems, 2018, pp. 6638-6648

[31] E. Giles, K. Patterson, and J. R. Hodges, "Performance on the Boston Cookie Theft picture description task in patients with early dementia of the Alzheimer's type: missing information," Aphasiology, vol. 10, no. 4, pp. 395-408, 1996.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:41:45 UTC from IEEE Xplore. Restrictions apply.

A Language-independent Network to Analyze the Impact of COVID-19 on the World via Sentiment Analysis

ASHIMA YADAV and DINESH KUMAR VISHWAKARMA, Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi, India

Towards the end of 2019, Wuhan experienced an outbreak of novel coronavirus, which soon spread worldwide, resulting in a deadly pandemic that infected millions of people around the globe. The public health agencies followed many strategies to counter the fatal virus. However, the virus severely affected the lives of the people. In this paper, we study the sentiments of people from the top five worst affected countries by the virus, namely the USA, Brazil, India, Russia, and South Africa. We propose a deep language-independent **Multilevel Attention-based Conv-BiGRU network (MACBiG-Net)**, which includes embedding layer, word-level encoded attention, and sentence-level encoded attention mechanisms to extract the positive, negative, and neutral sentiments. The network captures the subtle cues in a document by focusing on the local characteristics of text along with the past and future context information for the sentiment classification. We further develop a COVID-19 Sentiment Dataset by crawling the tweets from Twitter and applying topic modeling to extract the hidden thematic structure of the document. The classification results demonstrate that the proposed model achieves an accuracy of 85%, which is higher than other well-known algorithms for sentiment classification. The findings show that the topics which evoked positive sentiments were related to frontline workers, entertainment, motivation, and spending quality time with family. The negative sentiments were related to socio-economic factors like racial injustice, unemployment rates, fake news, and deaths. Finally, this study provides feedback to the government and health professionals to handle future outbreaks and highlight future research directions for scientists and researchers.

CCS Concepts: • **Information systems** → **Information retrieval**; **Retrieval tasks and goals**; **Sentiment analysis**; • **Computing methodologies** → **Machine learning**; **Machine learning approaches**; **Neural networks**;

Additional Key Words and Phrases: Attention, coronavirus, deep learning, pandemic, sentiment analysis, social media

ACM Reference format:
Ashima Yadav and Dinesh Kumar Vishwakarma. 2021. A Language-independent Network to Analyze the Impact of COVID-19 on the World via Sentiment Analysis. *ACM Trans. Internet Technol.* 22, 1, Article 28 (September 2021), 30 pages.
https://doi.org/10.1145/3475867

Authors' address: A. Yadav and D. K. Vishwakarma, Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi, India; emails: {ashimayadavdtu, dvishwakarma}@gmail.com.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
1533-5399/2021/09-ART28 $15.00
https://doi.org/10.1145/3475867

1 INTRODUCTION

The outbreak of COVID-19 pandemic[1] (popularly referred to as coronavirus), caused by severe acute respiratory syndrome coronavirus 2, has badly hit the entire world. According to the WHO,[2] this pandemic has affected 7,941,791 people with a death toll of 434,796 worldwide. Thus, the WHO declared a Public Health Emergency of International Concern on January 30, 2020. Nearly every country of the world has reported active cases, following which the WHO has advised them to follow strict COVID-19 mitigation guidelines like social distancing, contact tracing, non-essential business closures, quarantines, bans on travel, and bans on mass gatherings, etc. [1]. This creates a strong need to assess the attitudes and behavior of the people related to the pandemic.

Social networks have become a crucial part of disseminating public health information to the people. Twitter is increasingly used to extract real-time health data and understand public reactions towards any crisis or outbreak. Earlier, tweets have been analyzed for pandemic like Zika [2], Ebola [3], and natural disasters like earthquakes and Hurricane Irma [4]. The results show that the real-time analysis of public sentiments contributed to developing effective medical decisions to control the outbreak. Thus, analyzing the sentiments generated during pandemics like the most recent COVID-19 becomes critical.

Some of the works have utilized tweets to identify different aspects of the COVID-19 pandemic. Samuel et al. [5] presented a study that explored the sentiments of people about reopening the US economy amidst the COVID-19 crisis. A binary logit model was developed to analyze the crucial factors that generated the COVID-19 reopening sentiment. Medford et al. [6] evaluated public sentiment and emotions in tweets for a period of 15 days (January 14 – January 28, 2020). The dominant emotions were fear and surprise, and the maximum number of tweets were directed towards the economic and political impact of the virus. Garcia et al. [7] performed topic detection and sentiment analysis on Twitter to extract the sentiments emerging from Brazil and the USA for four months. However, their approach was unable to understand the context in tweets.

Different from these works, we aim to classify the opinions of the people from the countries worst affected by the virus. Our network can model both the word-level and sentence-level attention differently by assigning larger weights to relevant contexts for the document classification. In this way, the network can understand the context in tweets and can easily detect the sentiments. Further, the network is language-independent, thus not restricting itself to any particular language. We make an in-depth analysis to visualize the fluctuation patterns of sentiments by applying topic modeling to explain the story behind these fluctuations. Hence our main objectives are:

(1) To obtain the sentiment polarity of the tweets posted by citizens in different languages effectively.
(2) To study the country-wise fluctuation pattern of sentiments over a period of time.
(3) To identify the essential topics which received significant attention from the public to uncover the issues or factors related to the pandemic.
(4) To analyze the reason behind the popularity of the extracted topics.

The main contribution of our research can be summarized as:

- We propose a deep **Multilevel Attention-based Conv-BiGRU Network (MACBiG-Net)** to capture the subtle cues in a document for classifying the opinions of the people from the

[1] https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020.
[2] https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200815-covid-19-sitrep-208.pdf?sfvrsn=9dc4e959_2 (As on 15 August, 2020).

top five worst affected countries by the coronavirus, namely the USA, Brazil, India, Russia, and South Africa.
- Moreover, the proposed network is language independent as it can classify the tweets belonging to multiple languages like Hindi, Japanese, Arabic, Spanish, Urdu, etc., effectively by encoding the words and sentences and assigning larger weights to relevant contexts for the document classification. To the best of our knowledge, this is the first work that performs sentiment classification on multiple languages about the COVID-19 pandemic.
- We use topic modeling to uncover the major topics (or issues) related to the pandemic on Twitter for different time frames corresponding to each country.
- We develop a COVID-19 Sentiment dataset by crawling and downloading the tweets from the famous microblogging site Twitter and analyzing this dataset to study the fluctuation patterns of sentiments and topics for five months.

Our findings give policymakers crucial insights for formulating policies and strategies to accommodate the public needs more precisely. The study contributes to cyberpsychology, which monitors the behavior of humans during an outbreak with the help of social media. Further, the framework can be utilized by researchers and practitioners to study a diverse range of topics and sentiments. The proposed COVID-19 Sentiment Dataset can easily be fetched using Twitter APIs and library. Researchers can use this dataset for designing and training new algorithms for other NLP-related tasks like aspect-based sentiment classification. The remainder of the paper is organized as follows: Section 2 explores the previous literature related to sentiment analysis, multilingual sentiment analysis, and its applications in the healthcare domain and outbreak management. Section 3 focuses on the proposed methodology by describing the proposed MACBiG-Net. Section 4 presents the experimental results, including the dataset collection process, classification results, baseline comparison, sentiment trend analysis, and thematic analysis. Section 5 explains the theoretical and practical research implications. Section 6 discusses the limitations and future research directions. Finally, Section 7 concludes our paper.

2 LITERATURE REVIEW

Since our work focuses on multilingual sentiment analysis, outbreak management, and sentiment analysis in the healthcare domain, we summarize the previous literature in these areas. We also discuss the latest work related to sentiment analysis on COVID-19, their limitations, and how our proposed work is different from them.

2.1 Sentiment Analysis

The basic tasks for sentiment analysis and affective computing include emotion recognition and polarity detection [8]. The former focuses on extracting emotions [9] like Anger, Happy, Sad, etc. The latter is used in binary or multi-class classification problems with output labels such as positive, negative, highly-positive, highly-negative, etc. The general categorization for sentiment analysis and affective computing [8] includes knowledge-based techniques, which contains lexicons like WordNet-Affect [10] and SenticNet [11] to compute the polarity of the words for sentiment classification, statistical techniques which has machine learning-based methods, and finally the hybrid-based techniques, which applies both the knowledge-based and statistical-based approaches. Cambria et al. [12] argued that sentiment analysis is not only a simple categorization problem, but a suitcase of 15 problems related to NLP, which are grouped into three layers: syntactic layer, which involves preprocessing and normalizing the words; semantics layer, which deconstructs the text into concepts that can be single word or multi-words; and pragmatics layer to extract the opinionated words and compute their polarity. One of the excellent surveys in this

field is [13], where the authors analyzed the development achieved so far in sentiment analysis by exploring the trends, applications, achievements, and tasks in sentiment analysis. The authors also identified interesting future directions like multimodal sentiment analysis [14, 15, 16], sentiment reasoning [17], sarcasm analysis [18], etc.

2.2 Multilingual Sentiment Analysis

The majority of work in sentiment analysis has focused on a single language, usually English. However, with the advent of the Internet, a large amount of multilingual data is being generated online. Earlier, sentiment classification primarily focused on subjectivity analysis and polarity detection. To perform these tasks on multilingual data, efforts were being made to develop resources for other languages, like building a corpus, lexicon, or applying a translator for converting the data into a resource-rich language like English. SentiWordNet [19] is a polarity lexicon inspired by WordNet [20] for the English language. The nouns, verbs, adjectives, etc., are grouped to form synsets (synonym sets) having polarity scores in the range [0,1]. OpinionFinder [21] is a subjectivity lexicon that automatically finds the subjective sentences and their various aspects in English sentences. For a detailed analysis of resources and techniques used for multilingual sentiment analysis, we encourage the readers to refer to [22] and [23].

Apart from these, concept-level sentiment analysis incorporating common-sense reasoning is now getting popular. BabelSenticNet [24] is a multilingual concept-level resource that extends the SenticNet version to 40 different languages. However, these lexicons were unable to extract the granularity in text. One of the major drawbacks of these approaches is that human efforts are needed to create lexicons or corpus for other languages, requiring some funding support like crowdsourcing to get good quality annotation. Moreover, the scope of multilingual sentiment classification is not restricted to the above-mentioned tasks. They include information retrieval from the web, where researchers apply different machine translation tools like Google Translator [25], Bing Translator [24], etc., or unlabelled parallel corpus [26], which can be used for multiple different languages without requiring huge time, effort, and cost. However, translation always involves some noise as some part of the text is not translated correctly in the target language [27]. Finally, the popularity of deep learning models and pre-trained word embeddings inspired researchers to apply them for multilingual sentiment analysis [28, 29].

2.3 Word Vector Embeddings During Outbreaks

During the outbreak of an epidemic, the administration is responsible for making quick decisions and taking necessary actions. Since social media is increasingly being used to communicate information quickly and effectively to the public, there exists a need to transform and classify unstructured textual data into a structured form. Hence, researchers apply word embeddings to extract meaningful information from the unstructured data during any crisis-related event like natural disasters or pandemics. In recent years, deep learning-based methods like word2vec [30], doc2vec [31], Glove [32], and FastText [33] have shown promising results in analyzing the textual data. Since these prediction methods are trained on a vast corpus, they can efficiently capture the syntactic and semantic meaning of a word.

Ghosh et al. [34] designed disease-specific word embeddings from health-related news corpus for finding disease characterizations. However, constructing such a word vector model needs a tremendous amount of domain-rich text for training the model, which is difficult, especially during the early stage of the crisis. Hence, Khatua et al. [35] constructed relatively small disease-specific word embeddings from scholarly abstracts published on PubMed and Twitter tweets related to the 2014 Ebola and 2016 Zika outbreaks. The authors claimed that the relatively small domain-specific embeddings could model the contextual relationship among the data compared to the pre-trained

word2vec or Glove model. However, it becomes difficult to build an embedding model for every crisis. Moreover, this domain-specific vocabulary might not be rich in resources as compared to the pre-trained models. Sciandra [36] analyzed the Italian tweets generated during the COVID-19 pandemic by comparing the word embeddings with the term-frequencies approach. The lexicon-based approach was compared with supervised learning methods for classifying the tweets into different sentiments. Experimental results demonstrated that Glove and word2vec embeddings capture the semantic and contextual similarities in the dataset, whereas the frequency-based techniques failed to improve the classification accuracy.

The pre-trained embeddings are also used to initialize the embedding layer in the deep learning models. Wang et al. [37] applied 200-dimensional Glove embeddings with bi-directional RNNs for classifying the eight different types of diseases or symptoms from the NTCIR13-MedWeb dataset and around 2,000 tweets. The experimental results were compared with machine learning classifiers like SVM, Decision trees, Naïve Bayes, logistic regression. Lampos et al. [38] combined neural word embeddings with feature selection methods from Google search queries to study the influenza-like illness rates. The similarity between queries and the set of positive and negative words were computed using embeddings from Wikipedia articles and Twitter. The hybrid feature selection technique outperformed the baseline approaches by 12%. However, these queries did not give the best predictions. Since GloVe works better in recognizing the sub-linear relationships in the word vectors by assigning less weight to more frequent words so that the training process is not dominated by them [39], we apply them on our Twitter data to extract meaningful information.

2.4 Sentiment Analysis in Healthcare

The facets of sentiments in medicine are broadly categorized into [40]: capturing patient's health status (good/bad/no change), diagnosing the certainty or severity of a disease, and the effectiveness of the treatment or drug [41]. The early work in this area focused on in-built lexicons or machine learning-based approaches for extracting the sentiments. Yang et al. [42] proposed a framework based on the weighted scheme LDA to analyze breast cancer health-related information. The framework extracts the medical topics related to symptoms, side effects, etc., and clutters them together using weighted scheme LDA. The sentiment intensities along with physiological and psychological sentiments were extracted using the AFINN lexicon. The authors extracted health-related information from social media sites like MedHelp. The framework can benefit other users in seeking advice or taking necessary precautions if they experience similar symptoms. Sabra et al. [43] proposed a framework based on semantic extraction and sentiment assessment of the risk factors to predict Venous Thromboembolism disorder diagnosis. The clinical narratives were parsed to extract the risk factors with the help of MetaMap. According to the severity, these risk factors were stored in a dictionary, and semantic-rich rules were derived from weighting the risk factors. Finally, the polarity of the adjectives and adverbs were extracted with the help of the SentiWordNet lexicon. The framework achieved 54.5% precision and 85.7% recall to identify the patient's likelihood of developing Venous Thromboembolism early and starting the treatment as soon as possible.

Moh et al. [44] developed two approaches to study the side effects of five crucial drugs by capturing tweets. In the first approach, tweets were extracted in real-time, followed by rule-based classification. MetaMap was applied to extract the adverse drug effects, and subjectivity classification was performed for getting the opinionated sentences. This approach gave 91% accuracy on drug-related data. In the second approach, tweet similarity calculation was done, followed by drug classification only, which helped in identifying a large number of adverse drug effects. However, this approach gives lower drug-related predictions. The first approach was very complicated and could extract only a limited number of adverse drug effects. The sentiment scores were extracted

using several lexicons like SentiWordNet, AFINN, etc. Similarly, Jiménez-Zafra et al. [45] applied machine learning classifiers and lexicon-based approaches to compute the sentiment polarities for the physicians and the drugs. The reviews related to physicians contain informal language, while discussions about drugs contain the combination of colloquial language and specific terms like adverse effects or name of the drug. Korkontzelos et al. [46] identified the adverse drug reactions on social media platforms and forum posts. The correlation between the adverse drug reaction occurrences and sentiments was evaluated by considering different features like context features, parts of speech, negation, and cluster-based features. Sentiment-aware features like token n-grams, character n-grams, negation indication, token normalization, etc., were added into the conditional random field classifier. Experimental results show that the proposed method could locate the adverse drug mentions using sentiment analysis features and improved the F-score from 72.14% to 73.22%.

Recently, deep-based approaches were being explored to extract the healthcare sentiments expressed by the people. Limsopatham et al. [47] proposed deep neural networks CNN and RNN for medical concept normalization. This helps in mapping a long social media message to the medical concepts. CNN and RNN model the sentence representation separately to capture the semantic information. Different word embeddings were used to convert the inputs into word vectors. Experimental results show that both the methods were able to outperform the baselines. Grisstte et al. [48] normalized the adverse drug reactions found on social media into formal medical concepts. The undefined subwords were mapped into new concepts based on a similarity measure. BiLSTM network captured the context and understood the sentiment-related information for each word vector. Hence, the model can efficiently identify the target medical terms from unlabeled data. However, the model could not detect complex or unsuspected adverse drug reactions. Talpada et al. [49] studied the impact of Telemedicine on heart attacks and epilepsy. They compared the lexical and semantic methods with the deep-based methods. The lexical and semantic methods were based on VADER and TextBlob, whereas LSTM was employed for deep-based classification. Experimental results signify that sentiments can be used to obtain the demographic distribution of Telemedicine for heart attacks and epilepsy. Moreover, the lexical and semantic methods outperform deep learning methods for evenly distributed datasets. However, these methods can only work when enriched with domain-specific words, whereas deep learning networks can learn the domain-specific information by training the relevant dataset [50]. Thus, sentiment in the context of medicine shows the impact of a disease or drug on a person's life over a period of time.

2.5 Sentiment Analysis for COVID-19

Recently, some works have focused on analyzing different concerns about COVID-19 using sentiment analysis. We summarize the popular methods as follows:

Rahman et al. [51] explored the socio-economic factors like education, income, population density, number of COVID-19 cases, and emotional factors like mental health for 50 different US states to analyze the sentiments of people related to the reopening of the USA economy. Twitter data was collected with the keyword "reopen" to understand the positive and negative sentiments of the US people. A binary logit model was applied to identify which factors contributed towards positive or negative sentiment. The results show that poor-income people, families with low education levels, and people paying house rent were more inclined to reopen the USA economy. Imran et al. [52] compared variants of RNN based on different embeddings for understanding the cultural behavior during the COVID-19 pandemic. For this, tweets from February 2020 to April 2020 were downloaded using trending hashtags focusing on different dimensions like demography, culture, polarity, emotions, timeline, etc. The results demonstrate that highly correlated polarity tweets were shown by the USA and Canada, followed by India and Pakistan. Further, it can be concluded

that NLP-based techniques can help in identifying the emotional trend between different countries. However, the study does not explore CNN and its variants to focus on local textual features. Moreover, the proposed method cannot handle certain aspects of language like Roman Hindi. Alamodi et al. [53] presented a survey that focused on how sentiment analysis can fight infectious diseases. The authors have majorly concentrated on lexicon-based, machine learning-based, and hybrid techniques for sentiment analysis related to different outbreaks, epidemics, pandemics, and other infectious diseases. The study shows that social media plays a crucial role in spreading information about the pandemic.

Some researchers also applied topic modeling to analyze various issues related to the pandemic. Jelodar et al. [54] applied information retrieval methods like Gibbs sampling, topic modeling, and topic ranking to understand the sentiments of people related to COVID-19-based comments on Reddit. The authors utilized LSTM networks for classification. However, their study was limited to the English language only, and the authors considered the comments for two months only. Garcia et al. [7] extracted Portuguese and English tweets from April 17, 2020 to August 08, 2020. For analyzing English tweets, a combination of tweets, lexicons, n-grams, word embeddings, and word counts was applied using an SVM classifier. A combination of unigrams, bigrams, and sentence embeddings was selected with a logistic regression classifier for Portuguese tweets. The results show that mostly the topics were similar for both languages. The political actions mainly affected the flow of sentiments during the four months. One of the major challenges was to train the algorithms on Portuguese data since very few labeled datasets are available for this language.

Different from the above approaches, we apply attention at word-level and sentence-level to understand the meaning of a document. The algorithm focuses on the most informative word or sentence to understand a document better. Since the meaning of a word changes continuously in different contexts, modeling the context becomes essential to improve the classification process. Further, earlier works have majorly focused on some popular languages like English, Portuguese, Chinese, etc., and cannot handle multilingual data. We have proposed a language-independent framework that can process multiple different languages and eliminates the need for training the model on diverse language-specific datasets. We have also shown the sentiment trend of different countries for five months and extracts of crucial topics that influenced the sentiments of citizens.

3 PROPOSED METHODOLOGY

In this section, we explain the proposed framework for sentiment classification. It describes the data preprocessing process where we follow a two-step data cleaning approach, followed by the proposed MACBiG-Net to learn the semantic information of a sentence by encoding the essential words and sentences in a sequence.

3.1 Data Preprocessing

We scrapped the data from one of the most popular microblogging sites, Twitter, for all five countries based on their location. The complete data scraping, cleaning, and labeling process are discussed in Section 4. The final dataset comprises people's views belonging to three output classes: positive, negative, and neutral. Since data processing is one of the most crucial tasks for handling the unstructured textual data, hence we meticulously designed a two-step data cleaning approach. The first approach is followed before translating the data into the English language so that the translation process is not affected by special characters or non-ASCII codes in the sentence. This includes removing hyperlinks, special characters/symbols, retweet headers, Twitter usernames, non-ASCII codes, NaN (Not a Number) characters, and whitespaces. Finally, all the uppercase words are converted into lowercase by using the NumPy library functions.

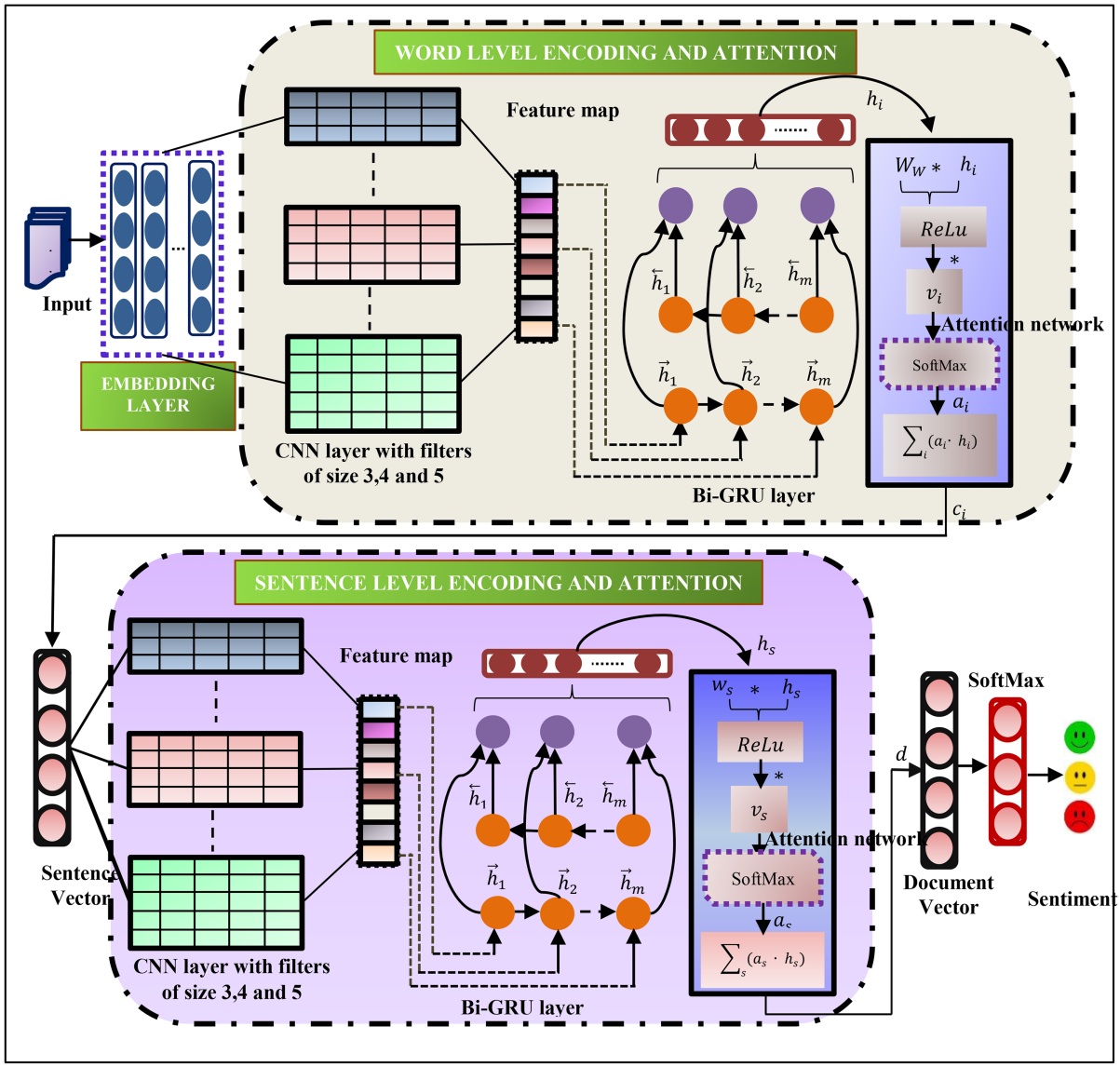


Fig. 1. Block diagram of the proposed network.

In the second approach, the labeled dataset goes through stemming and lemmatization to reduce the word to its root form. Porter Stemmer and WordNet Lemmatizer are used for stemming and lemmatization, respectively. We also removed the stop words that frequently occur in the text and convey no meaningful information, followed by hashtags and punctuation removal. The cleaned data is then passed to the MACBiG-Net, as discussed in the subsequent section for the final classification.

3.2 MACBiG-Net

The section describes the proposed MACBiG-Net, as shown in Figure 1. The network includes three main steps: Embedding Layer, which encodes the input words into the low-dimensional vectors; Word-Level Encoding, in which the words of a sentence are encoded by 1D Conv-BiGRU based representation to obtain the sentence representation, followed by Word-Level Attention, which computes the essential weights for the sentence vector, Sentence-Level Encoding, where the sentences of a document are again encoded by 1D Conv-BiGRU based representation to obtain

the document representation, and finally by Sentence-Level Attention to get the document vector for the classification.

3.2.1 Embedding Layer. Each word in the sequence of the sentence is converted into a real-valued vector. Formally, let $\{r_1, r_2, \ldots.., r_n\}$ be the sequence of n words in a sentence. We first embed each word w_i into a real-valued vector through a word embedding matrix $E \in R^{d_r * |V|}$, where d_r is the word embedding size, and V is the vocabulary size. We use the Glove[3] embedding matrix to get the final sequence of vector $\{e_1, e_2, \ldots.., e_n\}$, which serves as an input for the next layer.

3.2.2 Word-level Encoding and Attention. Our proposed architecture is inspired by [55], which intends to capture the structure of words from the sentence and the structure of sentences from a document. The word encoding is performed by 1D convolution, followed by the Bi-GRU mechanism. We first describe the convolution process as follows:

The local features are extracted by using a 1D CNN network with different kernel sizes to generate a feature map. The different convolution kernels help in learning the various local characteristics of the text. In general, for e_i^{th} embedding corresponding to r_i^{th} word, the concatenation operation $\oplus$ is expressed as follows:

$$e_{1:n} = e_1 \oplus e_2 \oplus \cdots\cdots\cdots \oplus e_n \tag{1}$$

The convolution operation is applied on the window of j words to generate a new feature value m_i as shown below:

$$m_i = f\ (W \cdot e_{i:i+j-1} + b) \tag{2}$$

Where, b = bias and f = Rectified Linear unit activation function (ReLu). This filter is applied to a different possible window of words in a sentence which generates the following feature map:

$$M = [m_1, m_2, \ldots\ldots., m_{n-j+1}] \tag{3}$$

We apply the filter of sizes 3, 4, and 5 to obtain multiple features, followed by a max-pooling layer to capture the maximum feature value. The concatenated output of all three filter sizes is passed into the BiGRU based layer, which contains the hidden states to store crucial information. The bidirectional nature of GRU is based on the fact that words in a sentence are not only related to the previous words but also to the following words. We have chosen GRU over LSTM due to its efficiency in training, which makes it computationally cheaper than LSTM. The forward GRU with $\vec{h}_i = \overrightarrow{GRU}\ (e_i, \vec{h}_{i-1})$is concatenated with backward GRU, $\overleftarrow{h}_i = \overleftarrow{GRU}\ (e_i, \overleftarrow{h}_{i+1})$ to obtain the output of BiGRU at step w as $h_i = [\vec{h}_i\ , \overleftarrow{h}_i]$. Thus, the BiGRU layer fetches the contextual information of every sentence centered around the word w_i. This makes sense as the orientation of sentiment needs to consider the past and future context information in a sentence. Thus, the word is encoded by the Conv-BiGRU block by storing the local and contextual features.

The attention mechanism is based on the idea that when users read a document word by word in each sentence, then they would pay attention to the most informative word or sentence. Hence, we model both the word level and sentence level attention differently. In word-level attention, the word annotation h_i is fed into the network to get its vector representation v_i. We apply the *ReLu* activation function over *tanh* as specified in [55], as *ReLu* converges quickly and results in cheaper computation. This is described in Equation (4):

$$v_i = ReLu\ \ (W_w \cdot\ h_i + b_w) \tag{4}$$

[3] https://nlp.stanford.edu/projects/glove/.

Table 1. Country-wise Details of COVID -19 Sentiment Dataset

Country	Sentiment			Total
	Positive	Negative	Neutral	
USA	383	321	431	1,135
Brazil	267	318	444	1,029
India	302	164	299	765
Russia	147	168	249	564
South Africa	200	193	232	625
Total	**1,299**	**1,164**	**1,655**	**4,118**

The resulting annotation v_i is multiplied (dot product) by the word-level context vector v_{wt} which is learned during the training. The v_{wt} gives a high-level representation to the question "*Which is the most important word*" [56]. It is initialized randomly and learned during the training process. The softmax function is applied to obtain attention weights a_i as shown in Equation (5) below:

$$a_i = \frac{\exp(v_{wt} \cdot v_i)}{\sum_i \exp(v_{wt} \cdot v_i)} \tag{5}$$

Finally, the attention weights a_i is concatenated with the word annotations to obtain the sentence vector c_i as described in Equation (6) below:

$$c_i = \sum_i (a_i \cdot h_i) \tag{6}$$

The layer details of the word-level encoding and attention step are shown in the Appendix (Table 1: *Layer description of Step 2: Word-level Encoding and Attention*).

3.2.3 Sentence-level Encoding and Attention. As discussed, selecting important sentences in the document becomes another crucial task for the classification. Hence, we follow the same procedure for generating the document representation of the sentences and, as discussed in Section 3.2.2, by encoding the sentences with relevant context and computing the crucial weights of these contexts for the document classification.

The sentence vector c_i obtained in Section 3.2.2 is passed into the 1D CNN layer and Bi-GRU layer to encode the sentence and obtain information h_s of the neighboring sentences centered around sentence s. The sentence-level context vector v_{st} is used to compute the sentence level attention a_s, which further yields the document vector d that summarizes all the information of the sentences in a document. The v_{st} give a high-level representation to the question "*Which is the most important sentence*" [56]. It is initialized randomly and learned during the training process. Mathematically, this is represented as below in Equation (7)–(8):

$$v_s = ReLu\ (W_s \cdot h_s + b_s) \tag{7}$$

$$a_s = \frac{\exp(v_{st} \cdot v_s)}{\sum_s \exp(v_{st} \cdot v_s)} \tag{8}$$

$$d = \sum_s (a_s \cdot h_s) \tag{9}$$

The layer details of the Sentence level encoding and attention step are shown in the Appendix (Table 2: *Layer description of Step 3: Sentence-level Encoding and Attention*). The learned vector d of [1* 100] dimension is passed to the dropout layer with a 0.5 rate to handle the overfitting problem

Table 2. Hyperparameters

Parameter	Value
Number of filters in 1D CNN	128
Filter sizes	3,4,5
Number of BiGRU units	100
Glove Embedding dimension	100
Maximum length of sequences	200
Dropout	0.5
Optimizer	Adam
Decay rate for first moment estimate (β_1)	0.9
Decay rate for second moment estimate (β_2)	0.999
Batch size	64
Epochs	100
Learning Rate	0.0001
L2 regularization parameter	0.001

in deep neural network architectures. This is followed by a dense layer and softmax activation function. The entire network is trained end to end for obtaining the best weights by monitoring the validation accuracy of the model. We use categorical cross-entropy loss, which is defined in Equation (10) as follows:

$$Loss = -\sum_{j=1}^{3}(y_j \cdot \log \hat{y}_j) \tag{10}$$

Where, y_j is the target value corresponding to the model output $\hat{y}_j$ for the j^{th} sample. We have used the L2-regularization parameter, which penalizes the larger weights in the network for avoiding the high variance problem. This makes the objective function of the network as follows:

$$Cost\ function = Loss + \left(\frac{\lambda}{2m} * \sum ||w||^2\right) \tag{11}$$

Here, λ is the regularization parameter and one of the hyperparameters of the network, which is tuned according to the validation set. Our model achieves optimized results with $\lambda = 0.001$.

4 EXPERIMENTAL ANALYSIS

This section discusses the data collection and labeling process in Section 4.1, implementation details in Section 4.2, classification results in Section 4.3, baseline comparison in Section 4.4, word-level attention visualization in Section 4.5, sentiment trend analysis in Section 4.6, and the topic extraction and thematic analysis in Section 4.7.

4.1 Dataset Collection and Labeling

In order to analyze the impact of the novel coronavirus, we developed Python scripts to download the Tweets from January 1, 2020 to June 7, 2020 of top five most affected countries, namely, USA (5,203,206 cases), Brazil (3,224,876 cases), India (2,526,192 cases), Russia (917,884 cases), and South Africa (579,140 cases) as per data available from WHO[4]. We used GetOldTweets3[5] library to extract

[4]https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200815-covid-19-sitrep-208.pdf?sfvrsn=9dc4e959_2 (As on 15 August, 2020).
[5]https://pypi.org/project/GetOldTweets3/.

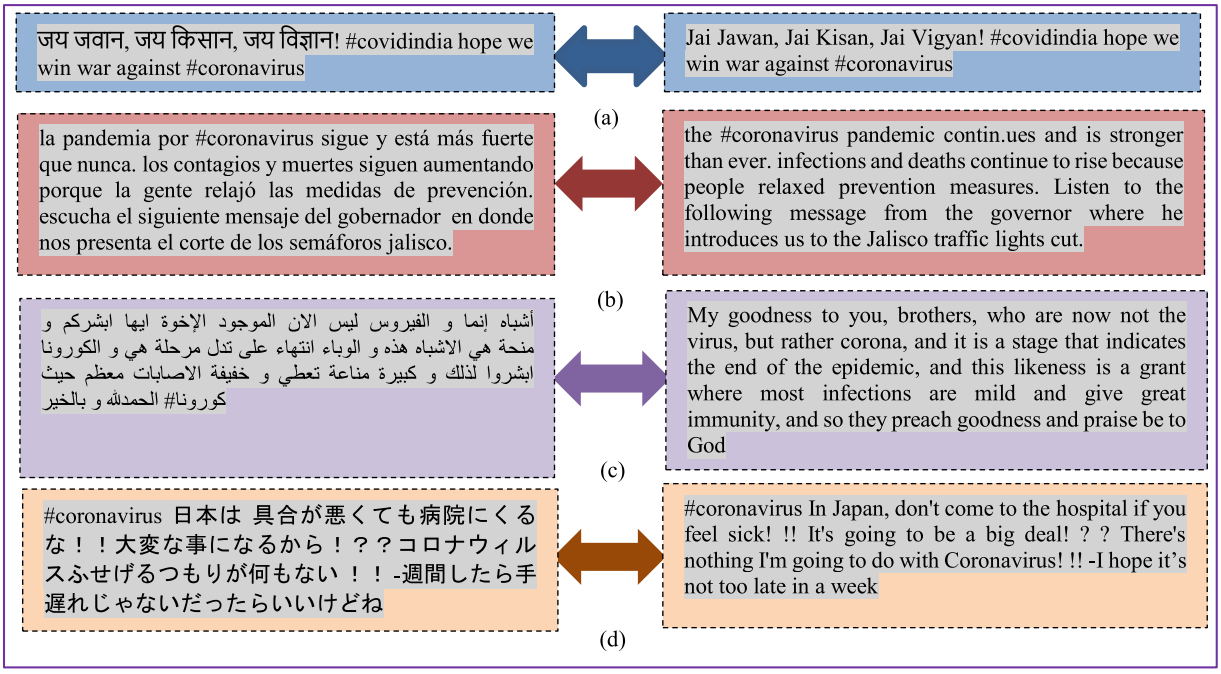


Fig. 2. Sample tweets in different languages from COVID-19 Sentiment Dataset (a) Hindi to English (b) Spanish to English (c) Arabic to English (d) Japanese to English.

the tweets using some popular hashtags like #coronavirus, #covid19, and #COVID_19. The query object was fed with the location and duration of tweets to get the following attributes of the data, which we stored into a CSV file:

- Tweet_id: The unique ID corresponding to every tweet.
- Tweet_data: The UTC time when the tweet was created.
- Tweet_retweets: The number of times tweet was retweeted by the users.
- Tweet_favorites: The number of times the tweet has been liked.
- Tweet_geo: The location specified by the user at the time of tweet.
- Tweet_text: The actual tweet.

After saving the data, we performed tweet filtering and cleaning steps, which resulted in a major reduction in the number of tweets. The filtering process involved removing those tweets which were duplicate or retweets, incomplete and blank. For example, tweets that did not contain geographical information were removed. This way, we could get the origin of each tweet. After this, we followed a two-step cleaning approach (one before the translation of the tweet and the other after translating the tweet) as described in Section 3.1. We applied Google translator for translating the tweet. Finally, a total of 4118 tweets were extracted after the filtering and cleaning process. Some sample tweets in a different language, and their English translation is shown in Figure 2.

We used the Textblob[6] library to automatically get the sentiment score of the processed tweets in the range of $[-1, 1]$. The tweets with $+1$ sentiment scores were labeled as positive, tweets with -1 score were labeled as negative, and finally, tweets with 0 sentiment scores were labeled as neutral. However, we observe that in many cases, the Textblob library fails to classify the data into correct sentiment labels. For example, consider the following tweet, "*My dad tested positive for the corona*"

[6]https://textblob.readthedocs.io/en/dev/.

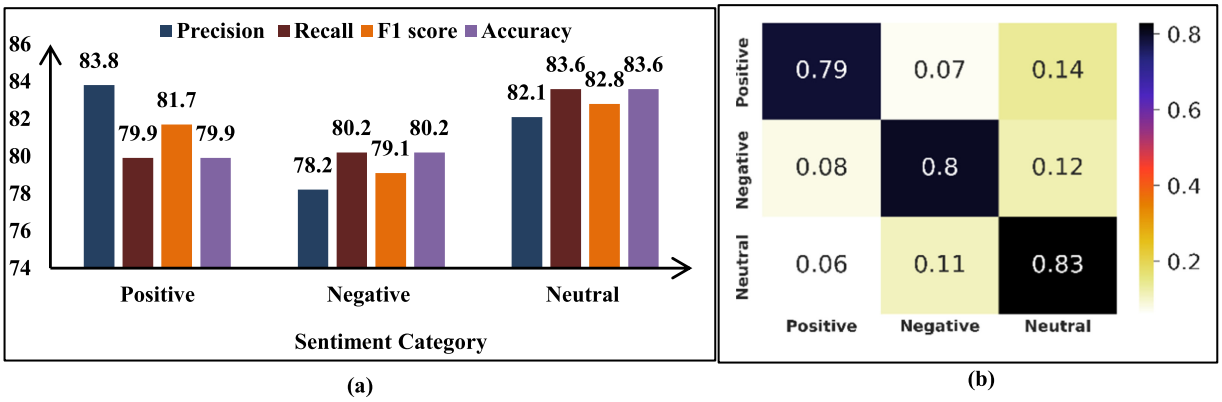


Fig. 3. (a) Sentiment classification results on COVID-19 Sentiment Dataset (%) (b) Confusion Matrix.

was labeled as positive sentiment, though it belongs to the negative sentiment category. Hence, the shortlisted tweets were again manually annotated by three humans with positive (+1), negative (−1), and neutral (0) sentiment category. Finally, the tweets were annotated with an average value of Cohen Kappa inter-annotator agreement as 0.85. The country-wise dataset details corresponding to each sentiment category are shown in Table 1.

4.2 Implementation Details

The proposed MACBiG-Net is implemented on Python 3 using the popular Keras framework. The experiments were performed on Windows 10 machine with 128GB RAM using NVIDIA Titan RTX GPUs, and Punkt sentence Tokenizer is used for dividing the text into a list of sentences. We have used a ten-fold cross-validation testing strategy with 80% data for training, 5% data for validation, and 15% data for testing. The validation accuracy is monitored during training the network. The model achieving the highest validation accuracy is selected for the testing phase. The hyperparameter details are shown in Table 2 that researchers can use for reproducing the results.

4.3 Classification Results and Observations

We compare the sentiment prediction performance of the MACBiG-Net in terms of precision, recall, F1 score, and accuracy corresponding to different sentiment categories, as shown in Figure 3 (a). Since we used a ten-fold cross-validation strategy, the final results are reported by averaging the results across each of the test folds. The accuracy of the positive, negative, and neutral sentiment categories is 79.9%, 80.2%, and 83.6%, respectively. The average accuracy of the network is 81.5%. We also compare the F1 scores of the individual sentiment category. This shows that our model can detect individual sentiments very well. Further, the confusion matrix in Figure 3 (b) gives a detailed analysis of the classification results. It can be seen that misclassification occurred when positive and negative samples were incorrectly classified as neutral, and the neutral sentiment was misclassified as negative. This means the network gets confusion majorly between negative and neutral sentiment classes. Yet, we can see that the network can discriminate well across each of the sentiment categories. Since we translated the tweets of different languages into English and then perform sentiment classification, we can see the proposed network can classify the translated tweets and capture the true sentiments invoked by them. These results show that MACBiG-Net can process multiple different languages and eliminate the need to train the model on diverse language-specific datasets [57].

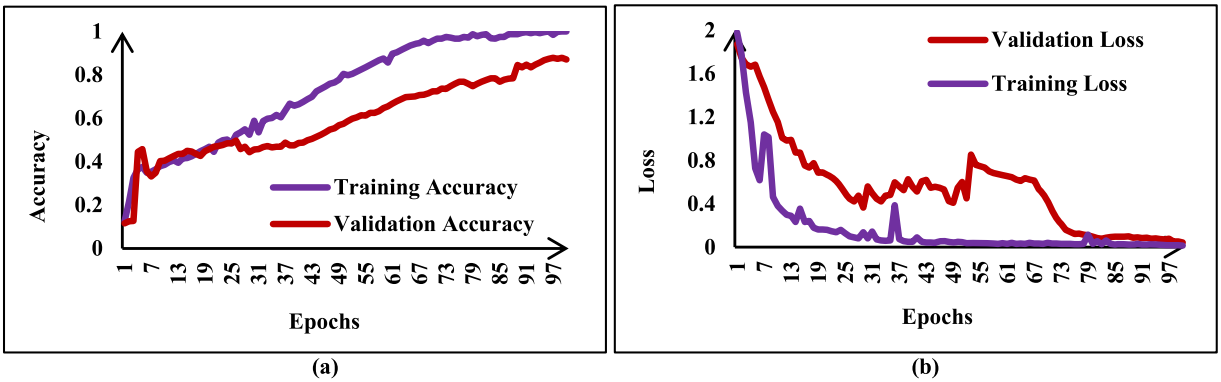


Fig. 4. Tracing internal performance of MACBiG-Net with (a) Accuracy curve for train and validation set (b) Loss curves for train and validation set.

We also provide accuracy and loss curves for training and validation set in Figure 4 (a) and (b), respectively. As we can see, the training accuracy is 100%, which means the model is trained completely, and the validation accuracy shows how effectively the model can perform on the unseen samples. Similarly, decreasing loss curves with each epoch confirm the adequate learning of the model and validate its internal classification performance. These graphs indicate that the proposed algorithm is strong enough to learn the training data structure and validate its performance on the validation set. This also shows that our proposed dataset contained adequate samples, which helped to train the model and perform well on unseen data.

4.4 Baseline Comparison

We use several popular baseline methods for comparison with the proposed MACBiG-Net. The details are discussed as follows:

- **LSTM** [58]**:** The classical long-short memory network is applied. Word vectors are initialized by Glove. The hidden layer size is 300 units. Adam optimizer with learning rate of 0.001 is used.
- **Bi-LSTM** [59]**:** This represents standard bi-directional LSTM. We follow the same settings as described above for LSTM.
- **CNN** [60]**:** Convolutional neural network with multiple filters of size [3, 4, 5], each having 100 filters is applied. Adam optimizer is used with a learning rate of 0.001.
- **CNN-RNN** [61]**:** We successively stack the CNN layer with max-pooling layer followed by RNN units. For RNN, we experimented with LSTM and GRU variants. The GRU based integration shows the highest performance.
- **HAN** [55]**:** The hierarchical attention network uses two LSTM layers followed by an attention mechanism at word-level and sentence-level for composing the final text representation. The HAN is the simplest form of our proposed approach as it does not encounter localized features.
- **Self-attention-based LSTM** [62]**:** The self-attention mechanism is used with Bi-LSTM, which learns the contribution of each hidden state by providing a set of summarized weight vectors for each hidden state. A learning rate of 0.06 is used.
- **RNN-Capsule** [63]**:** We apply capsule networks where the input instance representation is taken from the LSTM unit. The dimension of hidden vectors is set to 256. The attention mechanism is used to construct the capsule representation inside another capsule. All word vectors are initialized with Glove. Adam optimizer is used with a learning rate of 0.001.

Table 3. Comparative Results of Different Methods on COVID-19 Sentiment Dataset (%)

Method	Accuracy	Precision	Recall	F1 score	AUC
LSTM [58]	72.6	73.3	73.6	73.4	86.1
Bi-LSTM [59]	73.2	74.9	72.0	73.4	90.2
CNN [60]	75.8	73.1	75.2	74.1	93.5
CNN-RNN [61]	77.7	78.6	74.2	76.3	95.4
HAN [55]	79.4	77.4	78.9	78.1	96.2
Self-attention-based LSTM [62]	70.9	68.2	66.7	67.4	78.2
RNN-Capsule [63]	71.5	68.5	70.8	69.6	77.6
BERT [64]	76.6	75.2	74.8	74.9	92.2
MACBiG-Net (Ours)	**81.5**	**81.3**	**81.2**	**81.2**	**97.8**

- **BERT** [64]**:** We use a pre-trained BERT model with the help of an online released TensorFlow[7] library. The transformer blocks are set to 12, hidden size to 768, and self-attention head to 12.

As evident from Table 3, the proposed network has outperformed the baseline methods and recent works in terms of sentiment classification by achieving an average accuracy of 81.5%. We observe that the HAN [55] and CNN-RNN [61] also demonstrates better classification performance. This serves as the primary motivation of our proposed MACBiG-Net, which inherits the basic idea of both methods and achieves nearly 2% and 4% higher accuracy compared to [55] and [61], respectively. Thus, the 1D CNN and Bi-GRU-based encoding are better than stacking the layers of CNN-RNN [61]. It can capture the local characteristics of text and consider the past and future context information for the classification. Moreover, the algorithm pays attention to the most informative word or sentence to better understand a document, just what a reader does while reading a document. This improves the performance of the sentiment classification. We have also compared our network's performance with previous methods using ROC and AUC measures, as shown in Figure 5. ROC curves are a graphical plot between True Positive Rate and False Positive Rate at different possible thresholds. The AUC values help in comparing the ROC curves in a better way. The higher the AUC value, the better is the classification performance of the network. Thus, AUC gives a more comprehensive classification performance. From the AUC values, we can see that RNN-based capsule networks [63] show the worst behavior, with an AUC value of 0.77. On the other hand, HAN [55] gives an AUC score of 0.96. This shows the importance of attention networks in handling the data. The irrelevant information is not considered, and the network focuses on sentiment-rich words or sentences to get the best results. Thus, our network has consistently outperformed all the previous sentiment classification methods with an AUC value of 0.97.

4.5 Visualizing Word-level Attention

In this section, qualitative analysis is performed to visualize the word-level attention and extract the positive and negative sentiment topics that were generated during the pandemic. In this section, qualitative analysis is performed on MACBiG-Net to investigate the role of word-level attention for the sentiment classification. We randomly picked four tweets from our dataset, two of them were positive, and two were negative. Thus, in Figure 6, (a)–(b) belongs to the positive sentiment class (shown in Red color), and (c)–(d) belongs to the negative class (shown in Blue color). The

[7] https://github.com/tensorflow/tensor2tensor.

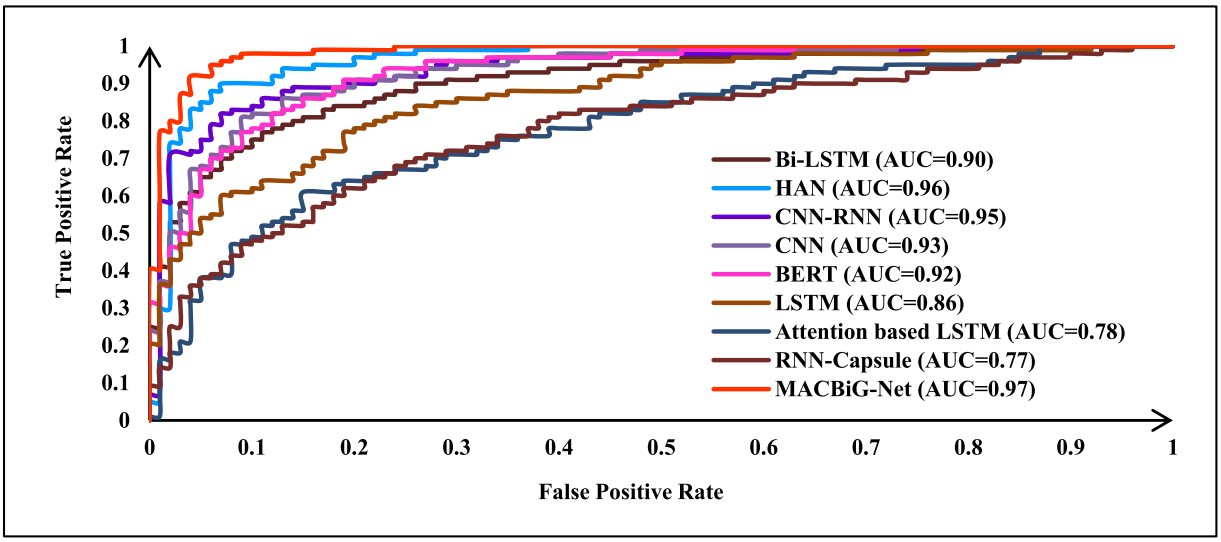


Fig. 5. Comparison of MACBiG-Net with baseline methods in terms of ROC curve and AUC.

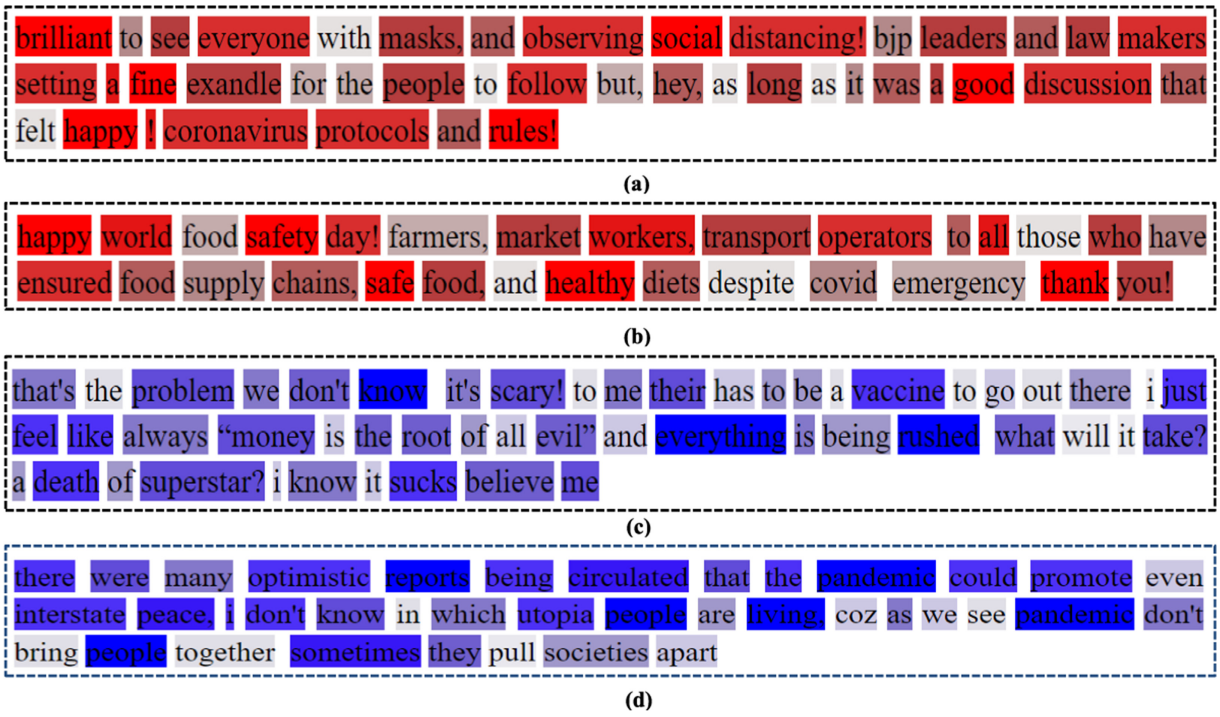


Fig. 6. Visualization the word-level attention weights from proposed dataset: (a)–(b) Positive Sentiment (c)–(d) Negative Sentiment.

color signifies the attention weights of each word. The darker the color in the respective category, the higher amount of attention is given to the word.

We can see that attention focuses on the prominent part of the sentence by giving darker color to the most important word. For example, in review sentences (a)–(b), positive sentiment is conveyed by crucial words like *'brilliant'*, *'happy'*, *'healthy'*, *'safety'*, and *'good'*, which have received more score by the attention mechanism for better sentiment prediction. Similarly, in reviews (c)–(d), negative sentiment is reflected by words like *'scary'*, *'pandemic'*, and *'death'*, which again have received more attention score. The words which convey no informative meaning (like stop words),

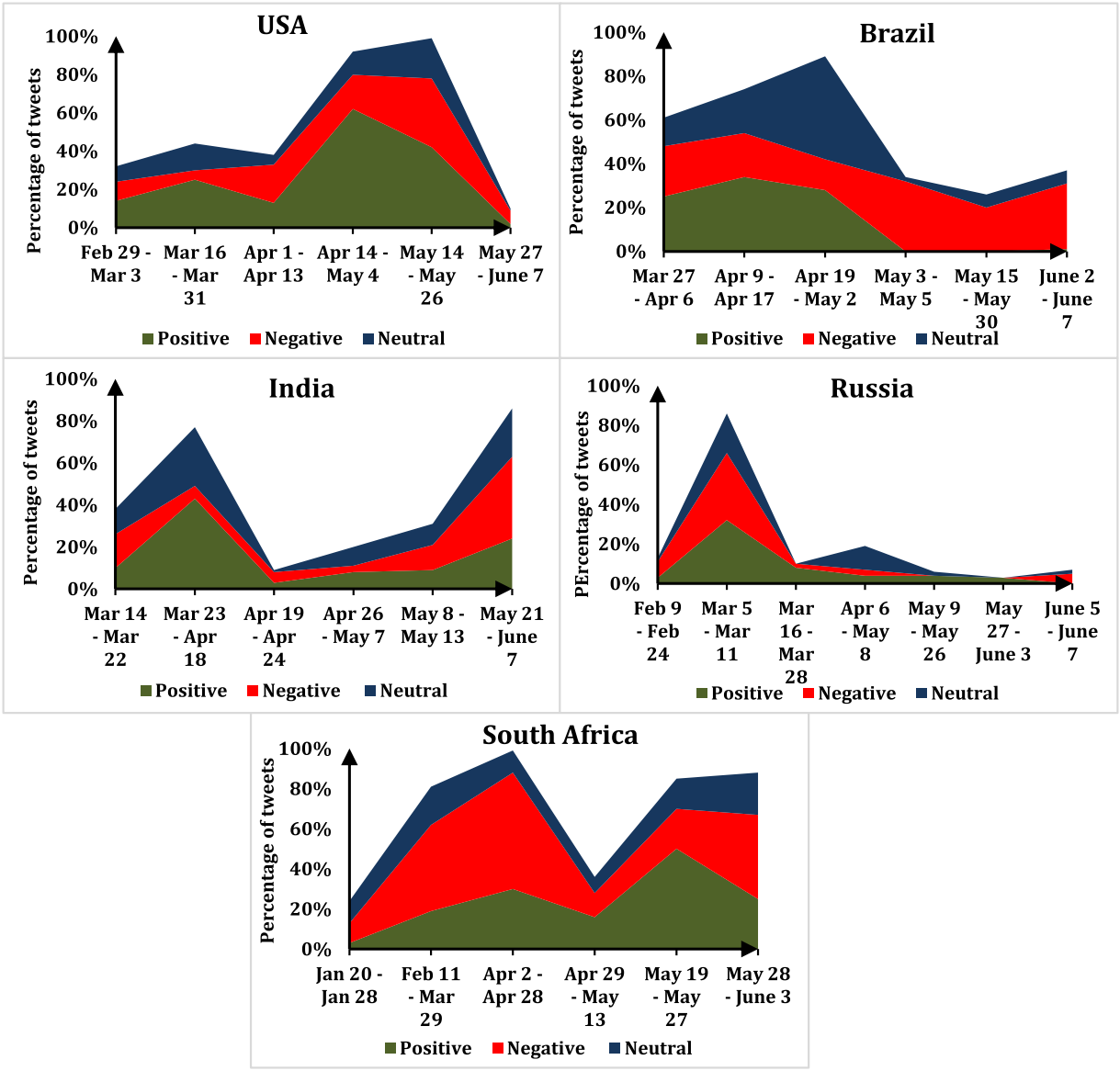


Fig. 7. Visualizing the sentiments of people over time in top five worst affected countries by COVID-19 pandemic.

were given very low weights by highlighting them with light colors. This shows that the attention weights focus on those words that relate to the output sentiment category. Since RNN can capture the sequential structure of a text, we need a mechanism to give higher weights to essential words in a sentence which are aggregated together to form a sentence vector. Thus, a score is assigned to each word relative to its importance in predicting the output sentiment class. This visualization confirms that the attention mechanism in our algorithm is concentrating more on informative words by assigning them more weights to classify the positive and negative sentiments effectively.

4.6 Sentiment Trend Analysis

The foremost aim of our work is to analyze the sentiments of people expressed globally on the COVID-19 pandemic. This can be visualized in Figure 7, where we plot the fluctuating patterns of positive, negative, and neutral sentiments of people over time located in the top five countries worst affected by the pandemic. As discussed, we retrieved the tweets from January 1, 2020 to June

7, 2020. However, in many countries, sentiments were generated in the month of late February or March. This may be because the first confirmed case in that country started in late February (e.g. *The virus was first confirmed in Brazil on February 25, 2020* [8]), or no specific sentiments were generated during January and February. We summarize our observations as follows:

- In the USA, significant sentiments emerged from March onwards. This may be because Trump declared a national emergency[9] on March 13, and by the end of March, the cases were reported in all 50 US States.[10] In early March, positive sentiment was recorded. However, by the end of March till mid-April, negative sentiment was prevalent. Similarly, by the end of May, we observed mostly negative sentiments. One of the primary reasons was the breaking news of George Floyd, which created a state of panic and terror in the USA from May 25 onwards. People started tweeting that COVID-19 will continue to spread as citizens were protesting and moving in groups without following any preventive measures. Overall, it was observed that 43.6% of sentiments were neutral, 37.9% were positive, and 18.3% were negative.
- Brazil has shown positive sentiments in the early days of April. However, mostly negative sentiments were conveyed in May and June. This may be because the number of active cases and deaths was doubled in May. Further, this number only kept on increasing, creating a condition of fear among people. Overall, 45.4% of negative sentiment, 39.3% of positive sentiment, and 15.1% of neutral sentiments were conveyed by the citizens of Brazil.
- In India, positive sentiments were observed, majorly in March and April. This may be because people were hopeful in the initial stages to overcome the virus. Further, PM Narender Singh Modi came up with some ideas to boost all the frontline worker's support. On March 22, the PM requested every Indian to clap their hands or ring some bells from their house to boost the morale of frontline workers. Similarly, on April 5, the PM urged everyone to light candles or *diyas* for nine minutes as a mark to fight against the deadly virus. Several people followed these tasks and happily enjoyed them. Towards the end of May, we see the people expressed negative sentiments. Since, by the end of May, a large number of active cases were being reported each day. Overall, 50.6% of positive sentiments, 28.3% of neutral sentiments, and 20.9% of negative sentiments were observed in India.
- Russia reported its first positive case on January 31, 2020. People have majorly shown neutral and negative sentiments during the mid of February. However, positive sentiments were expressed in April. Overall, 39.2% of neutral sentiment, 33.9% of positive sentiment, and 26.7% of negative sentiment were observed in Russia.
- In South Africa, majorly negative sentiments were being observed from late February till mid of March. People finally realized that the deadly virus had hit their country as the number of cases started rising sharply. However, in April, we can see some positive sentiments as citizens started motivating everyone by asking them to wear masks, read books, be involved in daily exercise, and post positive ideas by their former President Nelson Mandela. Overall, South Africa expressed 39.6% of negative sentiments, 32.6% of positive sentiments, and 27.7% of neutral sentiments.
- Finally, we conclude that overall, 38.3% of positive sentiments, 32.6% of neutral sentiments, and 28.3% of negative sentiments were expressed by these top five affected countries.

[8] https://www.gov.br/saude/pt-br.
[9] https://edition.cnn.com/2020/03/13/politics/donald-trump-emergency/index.html.
[10] https://archive.vn/20200428073138/https://banningca.gov/DocumentCenter/View/7139/CDC_COVID19-Weekly-Key_Messages_03292020_FINAL.

Based on Figure 7, we can see the oscillating pattern of various sentiments evoked from different countries. However, we tried to further explore the reason behind these patterns by building topic models for each country in Section 4.7.2.

4.7 Topic Extraction and Thematic Analysis

This section performs thematic analysis by applying topic modeling to get (i) the overall picture of positive and negative sentiment topics for the entire dataset and (ii) in-depth country-wise analysis of positive and negative sentiment topics. We also discuss the reason behind the fluctuations of these sentiments to explore the behavior of the people.

4.7.1 Visualizing Overall Sentiment Topics. We apply LDA-based topic modeling to determine the set of words (topics) that contribute to generate the positive and negative sentiments. This will help in finding the hidden thematic structure of the document and drawing essential conclusions from them. We use LDAvis [65] tool for visualization, which further helps in finding the prevalent topics in a document, relation between the topics, and the conditional distribution of a given term over the set of topics. The optimal number of topics was determined by building many LDA models with a different number of topics (k) and selecting the one with the highest coherence score [66]. This gives us $k = 10$. The layout of LDAvis is shown in Figure 8 (a)–(b) for positive and negative sentiment labels, respectively.

If we consider (a), the entire layout is divided into two panels: the left panel shows the ten topics in the form of circles where the similar topics are plotted close to each other, the right panel shows the bar graphs which represent the terms or words that are most useful for understanding the currently selected topic (from the left panel). The length of the gray bars depicts the entire document-wise frequencies of each term, and the red bars depict the topic-wise frequencies of each term. Thus, for a given topic, we can compare the length of the red and gray bars to understand the relevance of the topic. Further, selecting a term on the right panel tells the conditional distribution of the selected term over the topics (on the left panel). In this way, topic modeling helps in explaining the underlying information in a document. Hence, we summarize our observations as follows:

- The prevalent positive words related to COVID-19 on Twitter are: 'good', 'stay safe', 'happy', 'best', 'thank you', 'support', 'doctors', 'mask', 'together', 'proud'; and the popular negative words linked to COVID-19 on Twitter are: 'virus', 'deaths', 'bad', 'trump', 'help', 'lost', 'fake', 'died', 'outbreak', 'infected', 'hate', 'worst', 'crisis', 'fight'.
- Many positive tweets were directed towards the frontline workers where users were motivating and boosting the morale of the health care workers by appreciating them, donating them masks, dedicating some songs, providing free accommodation to them, and donating money to restaurants for distributing meals to essential workers.
- In some countries like Brazil, health care workers celebrated the discharge of the first COVID-19 patient from the hospital. In India, many citizens were delighted to see everyone unified with candles or flashlights in their hands from their balconies.

  Finally, positive sentiments were directed towards maintaining a healthy diet, reuniting with families after travel restrictions were lifted, boosting the immune system, asking everyone to stay safe, binge-watching television series, encouraging everyone to take precautions, wishing and motivating the users on special occasions by posting positive messages on occasions like Birthdays, Mother's Day, Earth Day, World Bicycle Day, Memorial Day (celebrated in the US), World Labor Day, International Nurses Day, and staying together to beat the deadly virus.

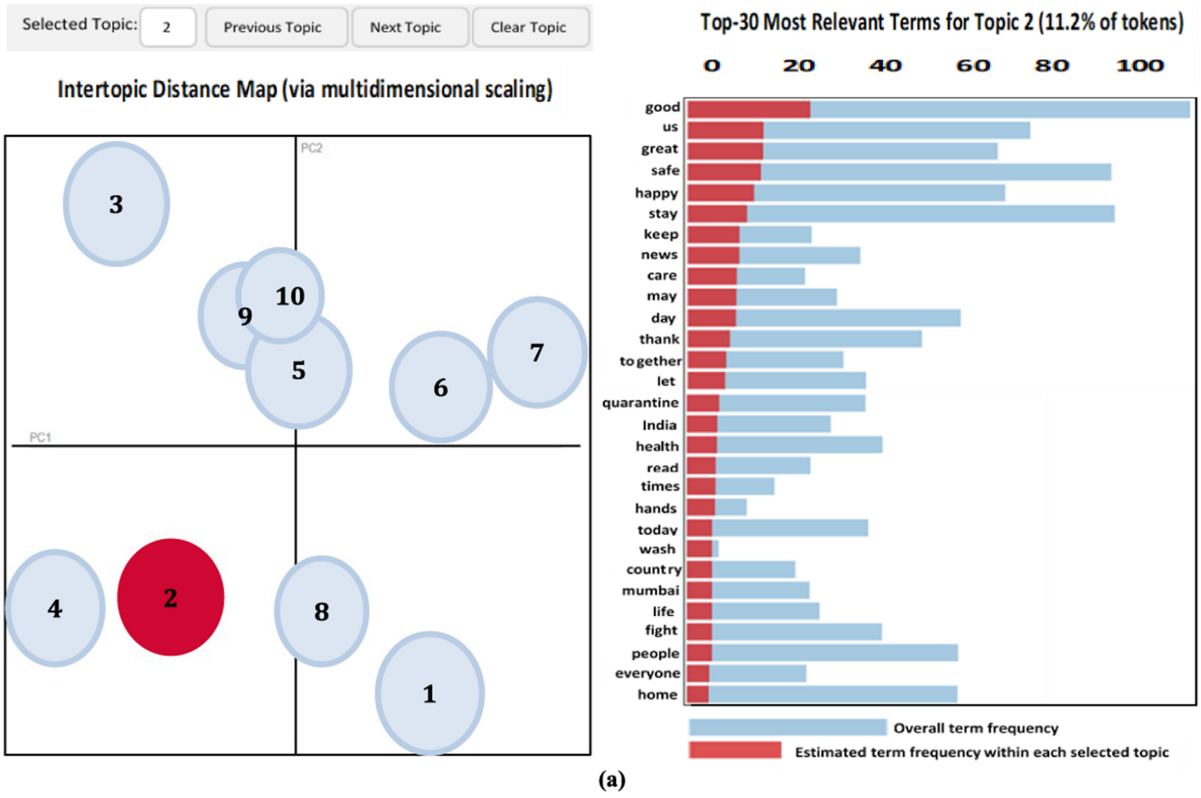


Fig. 8. Identifying the hidden information in the document through LDAvis visualization for (a) Positive and (b) Negative sentiment category.

Table 4. Sentiment Topic Analysis for USA

Duration	Sentiment	Topic terms
February 29 – March 3	Positive	eat, great, beer, believe, difference, sing, weekend, biking, new, radio
March 16 – March 31	Positive	us, stay, protect, protection, god, home, grateful, hospital, security, advice
April 1 – April 13	Negative	lost, vaccine, income, trump, prison, job, actions, left, health, unemployment, death
April 14 – May 4	Positive	indoors, us, food, working, quarantine, thank, helping, save, lives, appreciate, healthcare
May 14 – May 26	Positive	mask, quarantine, happy, Sunday, peace, truth, fighting, vaccine, environment
May 27 – June 7	Negative	protest, health, trump, safe, floyd, death, black, help, hate, racists, criminals, trump

- During the pandemic, several social-economic factors also contributed to generate negative sentiments in people all around the world. In the US, the negative tweets were mainly focused on the racial injustice related to the death of George Floyd, which resulted in massive protests and violence even during the time of the pandemic. People feared that these mass demonstrations would increase the count of COVID-19 cases.
- In India, Mumbai (one of the worst-hit cities by COVID-19), was severely affected by cyclone winds. In West Bengal, Cyclone Amphan caused extensive damage. The poor, jobless migrants went traveling back to their homes by walking thousands of kilometers due to lockdown restrictions in India.
- In general, many people also lost their jobs. The unemployment rate was increased during the pandemic. People were bored and anxious during home quarantine. The upsurge in infected cases also created fear and stress among people. Further, lots of fake news[11] was generated on social media, creating panic about the disease.
- The popular positive hashtags are: #salutetocoronafighters, #inthistogether, #lovemycountry, #withme, #thankyouindia, #indiafightscorona, #flattenthecurve, #fightthevirus, and #frontlineheroes and negative hashtags are: #DictatorTrump, #BlackLivesMatter, #GeorgeFloyd, #AmericaInCrisis, #Riot2020, #migrantlivesmatter, #pandemia, #chinavirus, #wuhanvirus, and #fuck_coronavirus.

4.7.2 Exploring Country-wise Sentiment Topics. In this section, we build topic models by considering tweets for different time frames for three countries, namely the USA, India, and Russia, to see the variation of topic terms and sentiments during this period and tell the story behind each country's fluctuation patterns. We also found that in some cases, the extracted topics were representative of the news published online at that specific time. The detailed country-wise analysis is as follows:

1) USA: Table 4 represents the topic terms for the USA. From this table, we infer the following:

- From February 29 – March 3, positive sentiments emerged in the USA. People were not believing in COVID-19 as some states reported no cases. We observed entertainment-related terms like watch, radio, streaming, eat, beer, sing.

[11] https://timesofindia.indiatimes.com/blogs/toi-editorials/social-media-menace-beware-of-fake-news-going-viral-faster-than-covid-19-it-will-cost-lives-literally/.

Table 5. Sentiment Topic Analysis for India

Duration	Sentiment	Topic terms
March 14 – March 22	Negative	jantacurfew, widespread, virus, china, care, safe, guidelines, awareness, people, ipl
March 23 – April 18	Majorly Positive	fighting, good, happy, efforts, healthcare, speaks, indian, stage, statement, leadership, love, modi, win
April 19 – April 24	Negative	lord, lockdown, disagreements, staff, pandemic, lakh, violators, divide, addressing, jamaat, fake
April 26 – May 7	Positive	best, happy, day, vaccine, liquor, love, bollywood, home, appreciate, wishes, union, leader, salute, fashion, nurses
May 8 – May 13	Majorly Negative	mumbai, increased, vaccines, virus, maharashtra, please, spread, help, need, depression
May 21 – June 7	Negative	China, hospital, worsthit, sonusood, mask, crisis, testing, reports, overtake, worst, migrant

- Although Trump declared a national emergency on March 13, positive sentiments still appeared from March 16 – March 31. This may be because the citizens were ready and determined to fight the disease. We noticed that citizens were advising everyone to stay home, encouraged each other to follow the guidelines, and take preventive measures.
- However, negative sentiments were detected from April 1 – April 13. Tweets related to unemployment, losing income, people losing lives due to disease were posted. Some criticized Trump, his administration, speech, and actions as they believed that he fumbled the pandemic.
- From April 14 – May 3, positive sentiments were noticed as people posted messages appreciating the healthcare workers, meals were given to frontline workers, blessing and hope messages were circulated to motivate and thank all the frontline heroes.
- Then, some neutral tweets were posted related to case reports, statistics, and preventive measures by various news channels.
- During May 14 – May 26, positive sentiments emerged again, where people posted messages related to corona fighters, enjoying quarantine with family, and loving and appreciating nature, and they distributed care packages to the needy ones.
- Finally, negative sentiments emerged from May 27 onwards. We observed that one of the primary reasons was the breaking news of George Floyd, which filled the Internet with anti-racism tweets. People were leaving home, joining protests, and were in deep anger and pain.

2) India: Table 5 represents the topic terms for India. From the table, we infer the following:

- In India, negative sentiment emerged from March 14 – March 22. People tweeted information related to the virus to spread awareness. The government released rules and guidelines related to the pandemic. Jantacurfew [67] was imposed by the Prime Minister, which restricted people from going out of their homes. Further, the popular IPL (Indian Premier League) was postponed [68], and many fans were sad, blamed the pandemic for ruining it.
- However, some positive sentiments were observed from March 23 – April 18. This was majorly because, on March 22, the Prime Minister requested every Indian to clap their hands to applaud the frontline workers. Similarly, on April 5, he asked everyone to light the candles

Table 6. Sentiment Trend Analysis for Russia

Duration	Sentiment	Topic terms
February 9- February 24	Majorly Negative	Outbreak, china, Chinese, virus, wuhan, news, wuhanvirus, country, novel, stop, suspected
March 5 – March 11	Majorly Negative	Die, flu, mayor, now, services, patient, death, measures, tackle, advice, fever, number
March 16 – March 28	Majorly Positive	Quarantine, best, times, working, home, good, mental, disinfectants, safety, sports, ill, exposed
April 6 – May 8	Positive	Advertisement, safe, stayed, prevention, believe, rules, home, morning, congratulations, wishes, meme, friends
May 9– May 26	Negative	Affected, crisis, angry, black, show, kill, risk, cases, Floyd, outbreak
May 27 – June 3	Positive	Best, help, thank, patients, feeling, family, us, sanitizer
June 5 – June 7	Negative	Death, million, infected, patient, heath, virus, government, drug, community, vaccine, city, time

or switch on the flashlights in the night to show unity in the fight against the pandemic. Indians gladly participated in these activities and expressed themselves on social media.

- From April 19 – April 24, the containment zones were increased in India. The Tablighi Jamaat event [69] became the super spreader source of COVID-19 in India, and people expressed their dissatisfaction and anger regarding it. Thus, negative sentiments emerged during this time.
- During April 26 – May 7, positive tweets were generated towards PM Narender Singh Modi's progressive actions to combat the pandemic. People supported the lockdown extension and appreciated frontline workers. Some entertainment-related terms like Netflix, fashion, movies, lifestyle, and YouTube were observed.
- From May 8 – May 19, negative sentiments emerged majorly for Mumbai state, which remained the worst COVID-affected state in India [70]. This was followed by other factors like the economy going down, people losing jobs, poor hospital conditions, unavailability of beds in the hospital, and psychological factors like depression, anxiety, fear, etc.
- The negative sentiments continued from May 21 – June 7. By this time, India became the seventh worst-hit country by pandemic [71]. However, some positive sentiments emerged about Bollywood actor Sonu Sood [72], who started trending on social media. People applauded him for his continuous efforts towards all the migrant workers and students in this pandemic.

3) Russia: Table 6 represents the topic terms for Russia. From this table, we infer the following:

- During February 9 – February 24, majorly negative sentiments were observed, followed by some neutral sentiments. The first case in Russia was reported when a man returned from China [73]. Thus, people were majorly tweeting about the coronavirus, the situation of COVID-19 in China, and restricting the entry of people from China.
- The negative sentiments continued from March 5 – March 11 again. People started giving advice and recommendations during this time. They even tweeted the number of active cases, deaths related to the pandemic, and the pandemic symptoms.

- From March 16 – March 28, positive sentiments focused on entertainment, sports, vacation, enjoying time with family, long morning walks, etc., were observed.
- However, from April 6 – May 8 a major number of positive sentiments were recorded. These sentiments were related to topics like posting birthday or anniversary wishes on social media platforms, defeating viruses, and celebrating life. Since the government imposed stay-at-home orders [74], people started tweeting about the same and encouraged each other to stay at home safely.
- From May 9 – May 26, negative sentiments were noticed regarding the economic situation of people. Moreover, anti-racism and justice tweets were posted related to the George Floyd case in the USA.
- Positive sentiments were noticed for a short period of time, from May 27 – June 3. These tweets focused on subjects like charity, support, distributing food, donating clothes to the needy, washing hands, and standing together in this pandemic.
- Finally, from June 5 – June 7, negative sentiment emerged. Several sentiments were directed towards Sputnik V, the Russian first covid vaccine [75]. People were anxious about the effectiveness and distribution of the vaccine. Moreover, some mixed reactions were observed regarding topics like partial re-opening of borders, schools, universities, etc.

Based on the experimental results, we can conclude that the proposed algorithm has several practical advantages. It can capture the subtle cues in a document by extracting the local characteristics of the text and incorporates the past and future context information for the classification. The algorithm pays attention to the most informative word or sentence to understand a document better, just like what a reader does while reading a document. This improves the performance of the sentiment classification. The algorithm can process multiple different languages and eliminates the need for training the model on diverse language-specific datasets. Finally, the hidden thematic structure of the document is extracted by applying LDA-based topic modeling, which gives the conditional distribution of a given word over a set of topics. This helps in getting an in-depth analysis of the sentiment trend during the COVID-19 pandemic.

5 RESEARCH IMPLICATIONS

This section discusses the theoretical and practical implications of our research.

5.1 Theoretical Implication

This study introduces many theoretical implications. Firstly, applying sentiment analysis during the COVID-19 era helps in analyzing the attitudes and perceptions of people from all over the world. This helps in identifying the psychological risks in the population. Accordingly, the health professionals and federal government can take timely actions to improve the well-being of the citizens during any crisis. Secondly, the study empirically detected the positive and negative sentiments of tweets and investigated the factors that contributed to generating the sentiments. The relevant contextual words are identified, which help in revealing the hidden patterns in data. The positive sentiments were directed towards the frontline workers, spending quality time with family, subjects like motivation, and entertainment. The negative sentiments were mainly focused on racial injustice, deaths, active cases related to the pandemic, and unemployment rates.

Moreover, the study can help the researchers to direct their efforts towards several strategies like work from home, stay at home order, social distancing, etc., to get more valuable insights about the people's perception towards these strategies. Researchers can use the results of this study for comparative analysis and case studies.

5.2 Practical Implication

This study can guide the government and decision-making body to propose timely responses and take necessary actions to handle any outbreak in the future. The fluctuating sentiment patterns in different countries reflect the impact of COVID-19 on the world. Several views and opinions were generated towards vaccine development and medical treatment given to the patients during the pandemic. This may provide crucial instructions and guidelines to health professionals and policymakers. Besides this, it becomes very important to protect the mental health of citizens and ensure their well-being. The cultural differences and behavioral changes of the citizens were identified by studying the timeline of sentiments for different countries.

6 LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Although this study tries to provide an in-depth analysis of the sentiment trend during the pandemic, yet it is impossible to cover all the aspects regarding the same. Hence, in this section, we discuss the limitations of our work.

- The current research has focused only on Twitter data, which doesn't reflect the entire population of any country. Most of the people are not using Twitter and may or may not be on other social media platforms. Thus, our study doesn't provide a comprehensive idea about the entire country or nation.
- The keywords (or hashtags) used to fetch the Twitter data might not give all the relevant tweets conveying crucial sentiments. Moreover, the quality of tweets may also be affected by spam or fake tweets.
- Social media is filled with multimedia content. Our study has only focused on textual data. People might have expressed themselves using other modalities like images [76], speech [29], videos [77], or text embedded in an image. We have focused on extracting the positive, negative, and neutral sentiments from the text. We have not analyzed emotions like fear, happy, sad, etc., and fine-grained sentiments like most positive, least negative.
- We have used Google translator to translate our multilingual data into English. During translation, it is possible that some tweets might have lost their true meanings.

Based on the above limitations, future research can be directed towards the following:

- Our study has focused only on positive, negative, and neutral sentiments. Researchers can apply hybrid fuzzy-deep learning-based techniques to get fine-grained sentiment scores of the citizens, which can provide rich information about the sentiments of a population.
- Aspect-based sentiment analysis [78] is one of the most popular sub-fields in sentiment analysis, which focuses on identifying implicit and explicit aspects to achieve more granularity in text sentiment classification. Researchers can apply aspect-based sentiment analysis with topic modeling to identify the nuanced sentiments towards the crucial aspect terms.
- The latest trend in multilingual sentiment analysis, like BabelSenticNet [24], a multilingual concept level resource for 40 different languages, can be applied to improve the quality of the translated versions.
- As discussed, the study is focused on textual modality. It can be expanded to include multimodal data to understand the sentiment of citizens fully.
- In the future, the sentiment trend of different countries can be monitored over time using social media, and specific measures could be developed whenever a negative sentiment is observed for one particular area over a longer period.

7 CONCLUSION

This paper is aimed to analyze the impact of the novel coronavirus on people by extracting and analyzing the sentiments of the top five countries most affected with the virus. We propose MACBiG-Net, which aims to extract the positive, negative, and neutral sentiment using word-level encoded attention and sentence-level encoded attention. This is achieved by developing and labeling the COVID-19 Sentiment Dataset, which contains 4,118 labeled tweets crawled from Twitter. Further, we visualize the sentiments of the people from January 31, 2020 to June 7, 2020 and summarize essential observations. We conclude that lots of sentiments were being generated during the pandemic, and people were posting their views or opinions on several topics like frontline workers, upsurge in active cases, travel restrictions, and about the virus itself. Much of the fake news was generated during this time, which was misleading the people. However, despite being stressed, people appreciated all frontline workers' efforts and motivated each other to follow all the precautions.

APPENDIX

Table A1. Layer Description of Step 2: Word-level Encoding and Attention

Layer Name	Input Shape	Output Shape	# Parameters	Connected to
Input Layer	(None, 200)	(None, 200)	0	-
Embedding	(None, 200)	(None, 200, 100)	1835200	Input Layer
Conv1D_1	(None, 200, 100)	(None, 198, 128)	38528	Embedding
Conv1D_2	(None, 200, 100)	(None, 197, 128)	51328	Embedding
Conv1D_3	(None, 200, 100)	(None, 196, 128)	64128	Embedding
MaxPooling1D_1	(None, 198, 128)	(None, 66, 128)	0	Conv1D_1
MaxPooling1D_2	(None, 197, 128)	(None, 65, 128)	0	Conv1D_2
MaxPooling1D_3	(None, 196, 128)	(None, 65, 128)	0	Conv1D_3
Concatenate	(None, 66, 128), (None, 65, 128), (None, 65, 128)	(None, 196, 128)	0	MaxPooling1D_1, MaxPooling1D_2, MaxPooling1D_3
MaxPooling1D_4	(None, 196, 128)	(None, 65, 128)	0	Concatenate
Bidirectional_GRU	(None, 65, 128)	(None, 65, 200)	183200	MaxPooling1D_4
TimeDistributed (Dense)	(None, 65, 200)	(None, 65, 100)	20100	Bidirectional_GRU
Hierarchical_attention_Network	(None, 65, 100)	(None, 100)	10200	TimeDistributed (Dense)

Table A2. Layer Description of Step 3: Sentence-level Encoding and Attention

Layer Name	Input Shape	Output Shape	# Parameters	Connected to
Input Layer	(None, 15, 200)	(None, 15, 200)	0	-
TimeDistributed (Model)	(None, 15, 200)	(None, 15, 100)	2202684	Input Layer
Conv1D_1	(None, 15, 100)	(None, 13, 128)	38528	TimeDistributed (Model)
Conv1D_2	(None, 15, 100)	(None, 12, 128)	51328	TimeDistributed (Model)
Conv1D_3	(None, 15, 100)	(None, 11, 128)	64128	TimeDistributed (Model)
MaxPooling1D_1	(None, 13, 128)	(None, 4, 128)	0	Conv1D_1
MaxPooling1D_2	(None, 12, 128)	(None, 4, 128)	0	Conv1D_2
MaxPooling1D_3	(None, 11, 128)	(None, 3, 128)	0	Conv1D_3
Concatenate	(None, 4, 128), (None, 4, 128), (None, 3, 128)	(None, 11, 128)	0	MaxPooling1D_1, MaxPooling1D_2, MaxPooling1D_3
MaxPooling1D_4	(None, 11, 128)	(None, 3, 128)	0	Concatenate
Bidirectional_GRU	(None, 3, 128)	(None, 3, 200)	183200	MaxPooling1D_4
TimeDistributed (Dense)	(None, 3, 200)	(None, 3, 100)	20100	Bidirectional_GRU
Hierarchical_attention _Network	(None, 3, 100)	(None, 100)	10200	TimeDistributed (Dense)
Dropout	(None, 100)	(None, 100)	0	Hierarchical_attention _Network
Dense	(None, 100)	(None, 3)	303	Dropout

REFERENCES

[1] R. M. Anderson, H. Heesterbeek, D. Klinkenberg, and T. D. Hollingsworth. 2020. How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet* 395, 10228 (2020), 931–934.

[2] K.-W. Fu, H. Liang, N. Saroha, Z. T. Ho, P. Ip, and I. C.-H. Fung. 2016. How people react to Zika virus outbreaks on Twitter? A computational content analysis. *American Journal of Infection Control* 44, 12 (2016), 1700–1702.

[3] A. S. Imran, S. M. Daudpota, Z. Kastrati, and R. Batra. 2019. How did Ebola information spread on Twitter: Broadcasting or viral spreading? *BMC Public Health* 19, 1 (2019), 1–11.

[4] K. A. Lachlan, Z. Xu, E. E. Hutter, R. Adam, and P. R. Spence. 2019. A little goes a long way: Serial transmission of Twitter content associated with Hurricane Irma and implications for crisis communication. *Journal of Strategic Innovation and Sustainability* 14, 1 (2019), 16–26.

[5] J. Samuel, M. Rahman, G. N. Ali, Y. Samuel, A. Pelaez, P. H. J. Chong, and M. Yakubov. 2020. Feeling positive about reopening? New normal scenarios from COVID-19 US reopen sentiment analytics. *IEEE Access* 8, 1 (2020), 42173–142190.

[6] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann. 2020. An "Infodemic": Leveraging high-volume Twitter data to understand early public sentiment for the Coronavirus disease 2019 outbreak. *Open Forum Infectious Diseases* 7, 7 (2020), 1–7.

[7] K. Garcia and L. Berton. 2021. Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA. *Applied Soft Computing* 101 (2021), 1–15.

[8] E. Cambria. 2016. Affective computing and sentiment analysis. *IEEE Intelligent Systems* 31, 2, 1541–1672.
[9] Y. Susanto, B. C. Ng, A. G. Livingstone, and E. Cambria. 2020. The hourglass model revisited. *IEEE Intelligent Systems* 35, 5 (2020), 96–102.
[10] C. Strapparava and A. Valitutti. 2004. WordNet-affect: An affective extension of Wordnet. *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. 1083–1086.
[11] E. Cambria, Y. Li, F. Z. Xing, S. Poria, and K. Kwok. 2020. SenticNet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020), 105–114.
[12] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32, 6 (2017), 74–80.
[13] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea. 2020. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing* 1, 1 (2020), 1–29.
[14] A. Agarwal, A. Yadav, and D. K. Vishwakarma. 2019. Multimodal sentiment analysis via RNN variants. *2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD)*. 19–23.
[15] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang. 2020. WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia* 22, 5 (2020), 1358–1371.
[16] A. Yadav, A. Agarwal, and D. K. Vishwakarma. 2019. XRA-net framework for visual sentiments analysis. *IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. 219–224.
[17] E. Cambria, A. Hussain, and A. Vinciarelli. 2017. Affective reasoning for big social data analysis. *IEEE Transactions on Affective Computing* 8, 4 (2017), 426–427.
[18] D. I. H. Farías, V. Patti, and P. Rosso. 2016. Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology (TOIT)* 16, 3 (2016), 1–24.
[19] S. Baccianella, A. Esuli, and F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh Conference on International Language Resources* 10, 10 (2010), 2200–2204.
[20] G. A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM* 38, 11 (1995), 39–41.
[21] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, and J. Wiebe. 2005. OpinionFinder: a System for subjectivity analysis. *Proceedings of Empirical Methods in Natural Language Processing*. 34–35.
[22] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth. 2017. Multilingual sentiment analysis: From formal to informal and scarce resource languages. *Artificial Intelligence Review* 48, 4 (2017), 499–527.
[23] K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Hawalah, A. Gelbukh, and Q. Zhou. 2016. Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Cognitive Computation* 8, 4 (2016), 757–771.
[24] D. Vilares, H. Peng, R. Satapathy, and E. Cambria. 2018. BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis. *IEEE Symposium Series on Computational Intelligence*. 1292–1298.
[25] W. X. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 553–561, 2008.
[26] X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. 2012. Cross-lingual mixture model for sentiment classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* 1 (2012) 572–581.
[27] M. Araújo, A. Pereira, and F. Benevenuto. 2020. A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences* 512 (2020), 1078–1102.
[28] H. Nankani, H. Dutta, H. Shrivastava, P. V. N. S. Rama Krishna, D. Mahata, and R. R. Shah. 2020. Multilingual sentiment analysis. *Deep Learning-Based Approaches for Sentiment Analysis*. 193–236.
[29] A. Yadav and D. Vishwakarma. 2020. A multilingual framework of CNN and Bi-LSTM for emotion classification. *11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (2020). 1–6.
[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. 3111–3119.
[31] Q. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. *International Conference on Machine Learning*. 1188–1196, 2014.
[32] J. Pennington, R. Socher, and C. D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543, 2014.
[33] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.
[34] S. Ghosh, P. Chakraborty, E. Cohn, J. S. Brownstein, and N. Ramakrishnan. 2016. Characterizing diseases from unstructured text: A vocabulary driven Word2vec approach. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. 1129–1138, 2016.
[35] A. Khatua, A. Khatua, and E. Cambria. 2019. A tale of two epidemics: Contextual Word2Vec for classifying Twitter streams during outbreaks. *Information Processing and Management* 56, 1 (2019), 247–257.

[36] A. Sciandra. 2020. COVID-19 outbreak through Tweeters' words: Monitoring Italian social media communication about COVID-19 with text mining and word embeddings. *IEEE Symposium on Computers and Communications (ISCC)*. 1–6.
[37] C.-K. Wang, O. Singh, Z.-L. Tang, and H.-J. Dai. 2017. Using a recurrent neural network model for classification of tweets conveyed influenza-related information. *Proceedings of the International Workshop on Digital Disease Detection using Social Media*. 33–38.
[38] V. Lampos, B. Zou, and I. J. Cox. 2017. Enhancing feature selection using word embeddings: The case of flu surveillance. *Proceedings of the 26th International Conference on World Wide Web*. 695–704.
[39] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown. 2019. Text classification algorithms: A survey. *Information* 10, 4 (2019).
[40] K. Denecke and Y. Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine* 64 (2015), 17–27.
[41] A. Yadav and D. K. Vishwakarma. 2020. A weighted text representation framework for sentiment analysis of medical drug reviews. *IEEE Sixth International Conference on Multimedia Big Data (BigMM)*. 326–332.
[42] F.-C. Yang, A. J. Lee, and S.-C. Kuo. 2016. Mining health social media with sentiment analysis. *Journal of Medical Systems* 40, 11 (2016), 1–8.
[43] S. Sabra, K. M. Malik, and M. Alobaidi. 2018. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. *Computers in Biology and Medicine* 94 (2018), 1–10.
[44] M. Moh, T. Moh, Y. Peng, and L. Wu. 2017. On adverse drug event extractions using Twitter sentiment analysis. *Network Modeling Analysis in Health Informatics and Bioinformatics* 6, 1 (2017), 1–12.
[45] S. M. Jiménez-Zafra, M. T. Martín-Valdivia, M. D. Molina-González, and L. A. Ureña-López. 2019. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. *Artificial Intelligence in Medicine* 93 (2019), 50–57.
[46] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez. 2016. Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics* 62 (2016), 148–158.
[47] N. Limsopatham and N. Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1014–1023.
[48] H. Grisstte and E. Nfaoui. 2019. Daily life patients sentiment analysis model based on well-encoded embedding vocabulary for related-medication text. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 921–928.
[49] H. Talpada, M. N. Halgamuge, and N. T. Q. Vinh. 2019. An analysis on use of deep learning and lexical-semantic based sentiment analysis method on Twitter data to understand the demographic trend of telemedicine. *11th International Conference on Knowledge and Systems Engineering (KSE), Vietnam*. 1–9.
[50] A. Yadav and D. K. Vishwakarma. 2019. Sentiment analysis using deep learning architectures: A review. *Artificial Intelligence Review* 53, 6 (2019), 4335–4385.
[51] M. Mokhlesur Rahman, G. N. Ali, X. J. Li, J. Samuel, K. C. Paul, P. H. Chong, and M. Yakubov. 2021. Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data. *Heliyon* 7, 2 (2021), 1–11.
[52] S. A. Imran, S. Daudpota, Z. Kastrati, and R. Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. *IEEE Access* 8 (2020), 181074–181090.
[53] A. Alamoodi, B. Zaidan, A. Zaidan, O. Albahri, K. Mohammad, R. Malik, E. Almahdi, M. Chyad, and Z. Tareq. 2021. Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert Systems With Applications*, 167, 1–13.
[54] H. Jelodar, Y. Wang, R. Orji, and S. Huang. 2020. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics* 24, 10 (2020), 2733–2742.
[55] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego*. 1480–1489.
[56] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. 2015. End-to-end memory networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems* 2 (2015) 2440–2448.
[57] A. Yadav and D. K. Vishwakarma. 2020. A comparative study on bio-inspired algorithms for sentiment analysis. *Cluster Computing* 23, 4 (2020), 2969–2989.
[58] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[59] A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18, 5–6 (2005), 602–610.

[60] Y. Kim. 2014. Convolutional neural networks for sentence classification. *Empirical Methods in Natural Language Processing*. 1746–1751.
[61] S. Lai, L. Xu, K. Liu, and J. Zhao. 2015. Recurrent convolutional neural networks for text classification. *Twenty-ninth AAAI Conference on Artificial Intelligence*. 2267–2273.
[62] Z. Lin, M. Feng, C. Nogueira, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. 2017. A structured self-attentive sentence embedding. *International Conference on Learning Representations*. 2017.
[63] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu. 2018. Sentiment analysis by capsules. *Proceedings of the 2018 World Wide Web Conference*. 1165–1174.
[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.
[65] C. Sievert and K. S. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. 63–70.
[66] K. Stevens, P. Kegelmeyer, D. Andrzejewski, and D. Buttler. 2012. Exploring topic coherence over many models and many topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 952–961.
[67] "India to observe 'Janata curfew' on Sunday amid spurt in Coronavirus cases," [Online]. Available: https://economictimes.indiatimes.com/news/politics-and-nation/india-to-observe-janata-curfew-on-sunday-amid-spurt-in-coronavirus-cases/articleshow/74750784.cms?from=mdr.
[68] "IPL postponed, India ODIs called off: How coronavirus has rocked cricketing world," [Online]. Available: https://www.indiatoday.in/sports/story/ipl-2020-postponed-india-vs-south-africa-called-off-coronavirus-impact-on-cricket-1655346-2020-03-14.
[69] "30 Per Cent Of Coronavirus Cases Linked To Delhi Mosque Event: Government," [Online]. Available: https://www.ndtv.com/india-news/coronavirus-tablighi-jamaat-30-per-cent-of-coronavirus-cases-linked-to-delhi-mosque-event-government-2206163.
[70] "Mumbai accounts for over 62% cases in Maharashtra: How can worst-hit Indian city stop spread of coronavirus," [Online]. Available: https://www.indiatoday.in/india/story/mumbai-cases-maharashtra-worst-hit-indian-stop-spread-coronavirus-bmc-commissioner-1682328-2020-05-27.
[71] "India Climbs To 7th From 9th Spot Among 10 Nations Worst-Hit By COVID-19," [Online]. Available: https://www.ndtv.com/india-news/india-climbs-to-8th-from-9th-spot-among-10-nations-worst-hit-by-coronavirus-with-over-1-85-lakh-cases-2238310.
[72] "Sonu Sood, the 'messiah' of migrant workers, bags UN honour," [Online]. Available: https://www.newindianexpress.com/entertainment/hindi/2020/sep/29/sonu-sood-the-messiah-of-migrant-workers-bags-un-honour-2203485.html#:~:text=Sood%20helped%20more%20than%2015%2C000,people%20struck%20in%20the%20lockdown.
[73] "First two cases of coronavirus infection detected in Russia," [Online]. Available: https://tass.ru/obschestvo/7656549.
[74] "Moscow mayor orders all residents to stay at home," [Online]. Available: https://www.business-standard.com/article/pti-stories/moscow-mayor-orders-all-residents-to-stay-at-home-120033000008_1.html.
[75] "Exclusive: Russia to roll out its first approved COVID-19 drug next week," [Online]. Available: https://www.reuters.com/article/us-health-coronavirus-russia-exclusive/exclusive-russia-after-approving-japanese-covid-19-drug-to-roll-out-game-changer-next-week-idUSKBN2381FR.
[76] A. Yadav and D. K. Vishwakarma. 2020. A deep learning architecture of RADLNet for visual sentiment analysis. *Multimedia Systems* 26 (2020), 431–451.
[77] A. Yadav and D. K. Vishwakarma. 2020. A unified framework of deep networks for genre classification using movie trailer. *Applied Soft Computing* 96, 106624.
[78] C. Huang, W. Jiang, J. Wu, and G. Wang. 2020. Personalized review recommendation based on users' aspect sentiment. *ACM Transactions on Internet Technology (TOIT)* 20, 4 (2020), 1–26.

Received October 2020; revised June 2021; accepted July 2021

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-06035-2

RESEARCH ARTICLE-ELECTRICAL ENGINEERING

A Novel Low-Power Nonvolatile 8T1M SRAM Cell

Damyanti Singh[1] · Kirti Gupta[2] · Neeta Pandey[1]

Received: 18 February 2021 / Accepted: 28 July 2021
© King Fahd University of Petroleum & Minerals 2021

Abstract
In Static Random Access Memory (SRAM) that is most ubiquitous of portable devices, the power consumption is a major concern. The emerging nonvolatile device-based SRAM designs have shown reduced power consumption by enabling power down mode without the loss of data. This paper presents a novel nonvolatile SRAM cell that employs eight transistors and a single TiO_2 memristor as a nonvolatile device. The proposed 8T1M has low write power consumption, low store/restore energy and does not require reset phase. The performance of the proposed cell is compared with existing 8 T nvSRAM cells to demonstrate its versatility over others. The proposed cell shows 99.7% and 47% reductions in write '0' power consumption and store/restore energy, respectively, with respect to existing 8 T nvSRAM counterparts, while the corresponding improvement in write margins and average store/restore delay is 69% and 70%. The results at different power supply and technology nodes are also captured to validate the impeccable performance of the proposed cell.

Keywords Nonvolatile memory · Memristor · HRS · LRS · SRAM · nvSRAM

1 Introduction

The continuous downscaling of process technology has resulted in high density integrated circuits with improved device performance at the cost of increased power consumption. This has raised concerns for devices which rely on extended battery life, reliable power on/off operations. A static random access memory (SRAM) is an essential component in mobile systems and occupies a significant portion of the chip area [1]. A memory with low power consumption, high capacity, good noise margins is desirable. The reduction in supply voltage and dynamic voltage scaling is popular choices for lowering the power consumption [2–5]. Alternatively, the use of nonvolatile memory (NVM) devices in mobile chips enables to power-down mode which further saves the power without the loss of data [6]. Thus, a two-macro scheme wherein a high-speed SRAM is used in conjunction with NVM device to provide a balance between the power and delay performance is followed. However, this scheme is vulnerable to the loss of data due to long store/restore time in case of sudden power failure. This makes it unattractive, and therefore, a SRAM cell with integrated NVM device is preferred solution to design low power and fast devices.

A variety of nvSRAM cells depending upon the type of NVM device employed are existing in the literature [7–15]. Flash memory is the conventional one, while there are other emerging NVMs which provide better speed and low voltage operation in its comparison. The FeRAM, MRAM, PCM use ferroelectric capacitance, magnetic tunnel junctions, GST material as storage element, respectively [7–12]. However, they possess number of challenges related to data signal degradation, excessive write power consumption and heating requirement. A memristor (also known as resistive memory (RRAM)) is an amenable candidate for next generation memory due to its simple metal-oxide structure, fast-switching, low voltage and potential for device scaling characteristics [13–15]. A variety of nvSRAM cells [16–30] that employs one or two memristors are suggested in the literature. The cells are able to restore the data after power down, and the efforts are directed to further improve their nonvolatile performance such as reduce store/restore energy and delay. However, the power consumption of these nvSRAM cells is still a drawback which restricts their usage in

✉ Kirti Gupta
kirtigupta22@gmail.com

[1] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India

[2] Department of Electronics and Communication Engineering, Bharati Vidyapeeth's College Of Engineering, Delhi 110063, India

Published online: 01 September 2021

low-power applications. The existing nvSRAM cells except [24, 29] have two data bitlines which gets either charged or discharged in every write access cycle due to the complementary nature which consumes power. In this paper, therefore the efforts are made toward the reduction in power consumption. The objective of the paper is to present a novel 8T1M nvSRAM cell with the following features:

a. The proposed 8T1M nvSRAM cell lowers the power consumption by modifying the write mechanism of the cell. In existing nvSRAM cells, either of the two bitlines BL/BLB is charged/discharged for every write cycle but in the proposed cell the charging/discharging of bitlines is avoided during write '0' operation.
b. The proposed cell suggests the use of only one memristor and transistor to provide a non-volatile features.
c. The proposed cell structure avoids multiple cell power supply voltage levels, phases in store/restore mode, additional RESET phase as well as complex sequencing of control signals for proper operation in comparison with existing nvSRAM cells
d. The proposed cell offers the advantages of lower, store/restore energy along with higher write margin and fast restoration of data in comparison with its existing counterparts.

The organization of the paper is as follows: A background work on memristor and a discussion on state-of-the-art nvSRAM cells is provided in Sect. 2. A new 8T1M nvSRAM cell is presented in Sect. 3, and its operation is elucidated. The functional verification of the proposed cell, its design and performance comparison with the conventional 6 T SRAM cell, existing 8 T nvSRAM cells is provided in Sect. 4. The effect of voltage scaling and technology is also discussed. In the last, Sect. 5 concludes the paper.

2 Basic Concepts

This section presents an overview of memristor structure and its characteristics which is followed by its application in SRAM designs.

2.1 Memristor

Memristor is a two terminal passive element which store the data in terms of resistance. It was postulated by Leon Chua in 1971 as a fourth circuit element Memristor, besides the other three elements namely, resistor, inductor and capacitor [13]. Initially, it was a hypothetical component, but later on in 2008, William and his co-worker from HP labs introduce memristor physically [14]. A memristor has a Titanium dioxide (TiO_2) layer sandwiched between platinum (Pt) electrode as shown in Fig. 1a. The doped region of TiO_2 layer is oxygen deficient, i.e., TiO_{2-x} and act as conductor, while the undoped region of TiO_2 layer is purely an insulator. The symbolic representation of memristor is shown in Fig. 1b where the shaded portion represents the undoped region, while the other end corresponds to doped region.

When a positive voltage is applied across the doped side of the memristor, the oxygen deficiencies of doped TiO_{2-x} layer will repel toward the undoped TiO_2 layer. It increases width of TiO_{2-x} layer causing lowering of memristor resistance. In this condition, the memristor is defined to be in Low Resistance State (LRS). Conversely, if a negative voltage is applied across the memristor, it will attract the oxygen deficiencies, resulting into the increase in the width of undoped TiO_2 layer. This state of memristor is known as the High Resistance State (HRS). If no voltage is applied across the memristor, the oxygen deficiencies will remain at their position, this is how memristor remains in its previous state.

Mathematically, if voltage V(t) is applied across memristor then

$$V(t) = M(t)i(t) \tag{1}$$

where $i(t)$ is the current flowing through memristor. The $M(t)$ corresponds to memristance, the resistance of memristor and is represented as:

$$M(t) = R_{\mathrm{ON}}\frac{w(t)}{\mathrm{D}} + R_{\mathrm{OFF}}\left(1 - \frac{w(t)}{D}\right) \tag{2}$$

Here, D is the thickness of the semiconductor film sandwiched between two Pt electrodes and μ_v is the mobility of the oxygen ions. The state variable $w(t)$ represents the width of the doped region TiO_{2-x}. When the doped region extends to the full length (i.e., $w/D = 1$), then the resistance of the memristor is denoted by R_{ON}. Similarly, if the undoped region extends to the full length (i.e., $w/D = 0$), then the total resistance is denoted as R_{OFF}.

2.2 Existing nvSRAM Cells

The conventional 6 T SRAM cell with two back-to-back connected inverters (Q1–Q4) and two access transistors (Q5, Q6) is shown in Fig. 2a. It has two data bitlines (BL/BLB) to perform read and write operations. The cell is volatile in

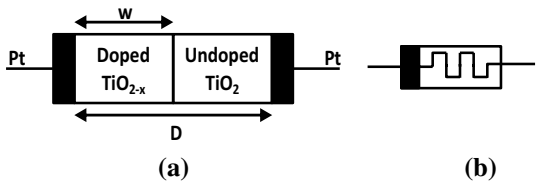


Fig. 1 Memristor representation **a** Physical **b** Symbolic

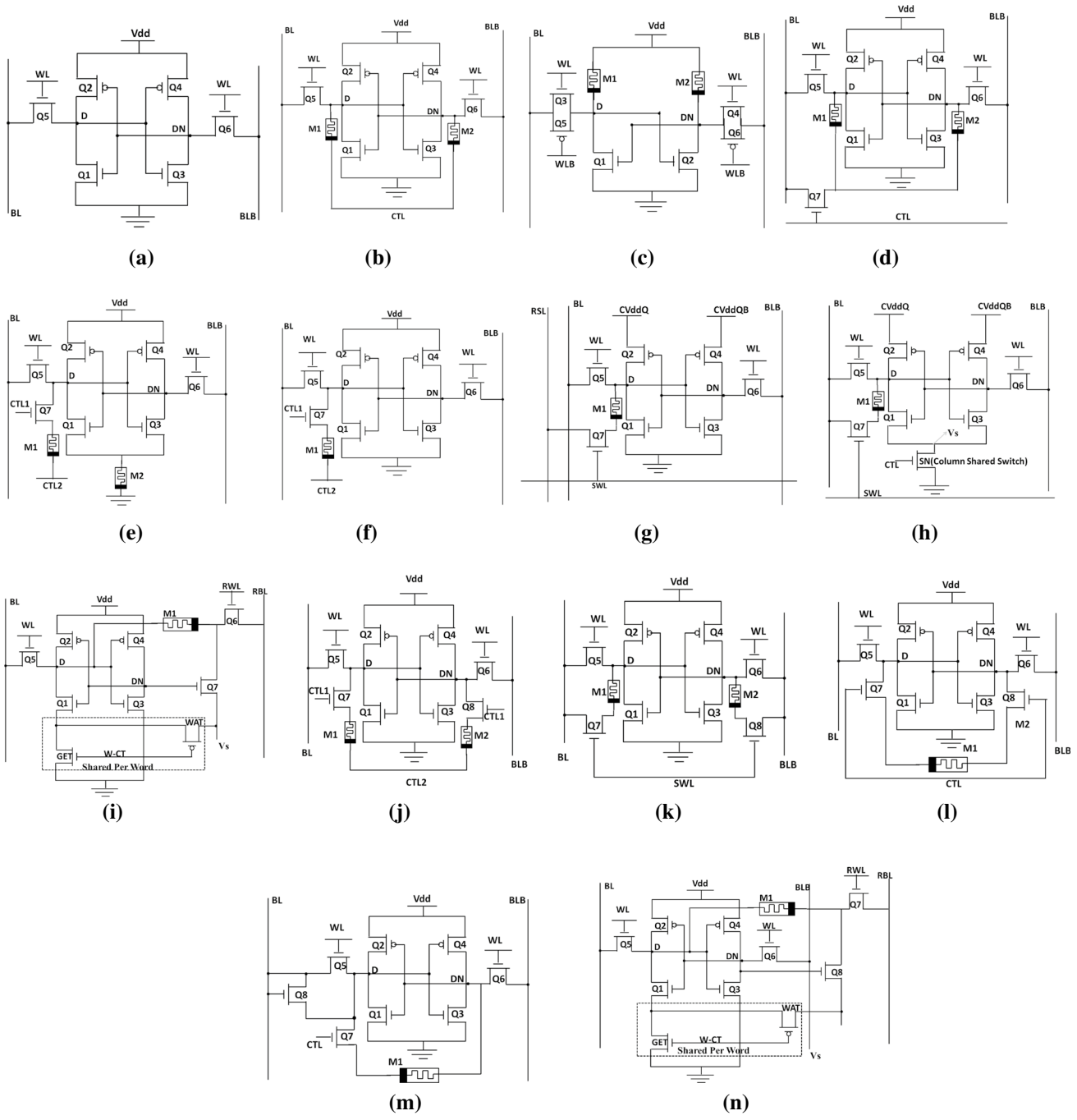


Fig. 2 **a** Conventional 6 T SRAM; Existing nvSRAM cells, **b** 6T2R [16], **c** 6T2R [17], **d** 7T2R [19], **e** 7T2M [20], **f** 7T1R [21], **g** IOW 7T1R [22], **h** Avg. 7T1R [23], **i** MS 7T1R [24], **j** 8T2R [25], **k** Rnv8T2R [27], **l** 8T1R[28], **m** 8T1R[17], **n** RD 8T1R [29]

nature as the stored data are lost once the supply voltage to the cell is turned OFF. The data are retained in SRAM structures by adding a NVM device such as memristor. In the event of power down, the data on the internal nodes of the cell are stored in memristor and is restored back into the cell after power up. The nvSRAM cells thus have an initial, store, power down, restore and reset phases as a representative of nonvolatile operation.

Several nvSRAM cells having one or two NVM devices are presented in the literature. The nvSRAM cells with six-to ten transistors are drawn in Fig. 2b–n. The 6T2R nvSRAM cell [16] shown in Fig. 2b is derived from conventional 6 T SRAM cell by connecting two NVM devices directly to the

internal storage nodes (D and DN). The use of NVM device allows the selected blocks to be powered off to save significant stand-by power. The direct connection of NVM devices to internal nodes results in DC-short circuit current as well as degraded cell stability. The 6 T nvSRAM cell as shown in Fig. 2c presented in [17] follows a different design strategy wherein the PMOS load transistors in CMOS inverters are replaced by two NVM devices and transmission gates are used as access transistors for speed enhancement. However, the use of transmission gates can easily flip the bitcell content during a read operation in the cell [18].

Thus, the other nvSRAM cells avoid the direct connection to the internal nodes by introducing controlling transistors. A variety of nvSRAM cells are presented in the literature [19–29] that differ in the way controlling transistor is connected. The devices in 7T2R [19] shown in Fig. 2d are connected to each of the two internal nodes and are further connected to either of the two data bitlines through a single controlling transistor. This arrangement improves the write margin; however, the leakage current issue is not resolved. The 7T2R nvSRAM cell structure suggested in [20] shown in Fig. 2e uses two NVM devices wherein M1 enable to achieve nonvolatile characteristic, while M2 improves the write performance. However, the nvSRAM cell structure with one resistive device is preferred for energy efficient designs. The 7T1R [21] shown in Fig. 2f uses a single resistive device and is connected to the internal node D through a controlling transistor involving two separate control signals. It provides "Instant ON" operation and achieves significant energy reduction. A similar 7T1R nvSRAM cell [22] shown in Fig. 2g has two phases namely the pulse overwrite (POW) phase and differentially supplied initialization phase to suppress the DC-short circuit current with improved restore yield than 7T1R [21]. The other 7T1R [23] shown in Fig. 2h uses an additional column-shared switch with reference to 7T1R [22] to further improve read–write performance. The cell, however, suffers from reduced stability of column-selected cells and limited read margins. The recent 7TIR [24] shown in Fig. 2i addresses the stability issue by incorporating a pair of shared write ability enhanced transistors and decoupled read port. The cell involves sequencing of various control signals and clock signals which increases the overall complexity.

There also exist eight transistors nvSRAM cells. The 8T2R cell [25, 26] in Fig. 2j addresses the DC short ci rcuit current issue of 6T2R [16] by adding two controlling transistors between the internal nodes and the devices. The cell offers significant power reduction in comparison with volatile SRAM cell; however, its application is limited by high energy consumption. The other 8T2R nvSRAM cell [27], as depicted in . 2k, attaches two NVM devices between the storage nodes and their respective data bitlines through controlling transistors. This arrangement enhances the write-ability while maintaining the nonvolatile characteristics. The 8T1R cell presented in [28] employs two transistors to connect the resistive device between the two storage nodes as shown Fig. 2l. The cell achieves compatible delay and noise margin as conventional 6 T SRAM cell along with low store/restore energy values. The circuit uses an additional power supply restore circuit for the delayed activation of the inverter in cell during restore operation. A recent alternate structure [17] using two transistors and a resistive device is shown in Fig. 2m. These transistors are controlled by bitline potential and an additional complementary word line signal to enable store and write execution in different cycles. The cell, however, requires maintaining different potential at power supply, bitlines for proper execution of different phases in memory operation. A read-decoupled 8T1R nvSRAM cell as drawn in Fig. 2n is suggested in [29]. It can be programmed to operate in stable or high-speed mode. The circuit adds two additional transistors GET and WAT shared per word to improve the stability margin and restore performance.

Further, it can be observed that the nvSRAM cells [16, 17, 22, 23, 28] require changes in the cell power supply usually to high potential for the proper execution of the nonvolatile phases. In the cells [21, 22, 24, 27, 29], the store operation requires the execution of two sub-operations due to the presence of single resistive device, whereas the cells [16, 17, 19, 21, 25–27] have additional RESET phase to program the devices in the HRS state after restore operation. Few cells [17, 27] connect bitlines with the NVM devices which adds extra capacitances to the bitlines thereby degrading the cell volatile performance. Though the cells in [20, 24, 29] show improved performance, the use of large number of control signals and their complex sequencing along with the requirement of maintaining different voltage level makes its unattractive. Further, most of the nvSRAM cells except [24, 29] exhibit differential write operation. In these cells, either of the two bitlines charge/discharge results in significant write power consumption. Therefore, in this work, a novel nvSRAM cell addressing the above concern is proposed. The efforts are made to present a cell structure which avoids the use of multiple cell power supply, voltage levels, two phases in STORE/RESTORE mode, additional RESET phase as well as complex sequencing of control signals for proper operation.

3 Proposed 8T1M nvSRAM Cell

The proposed 8T1M nvSRAM cell with two data bitlines (BL/BLB) is shown in Fig. 3. It consists of two cross-coupled inverters (Q1–Q4), transistors (Q5–Q7) to control read/write operations and transistor Q8, memristor M1 for nonvolatile feature. The new data on BLB line are written in

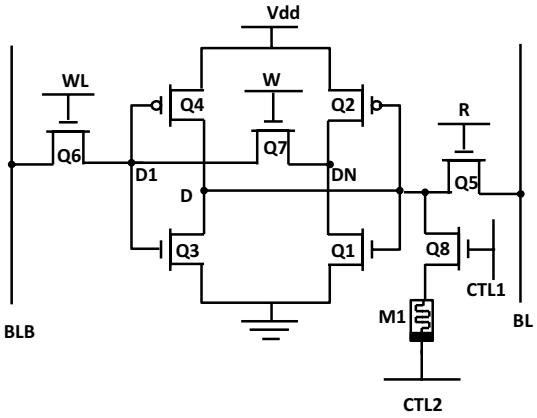


Fig. 3 Proposed 8T1M nvSRAM Cell

the cell through the access transistor Q6 controlled by the word line (WL) signal. The data are read through the BL/BLB line via the access transistors Q5, Q6, Q7 driven by read *R* signal, WL and W signal, respectively. During write operation, it is turned OFF such that the feedback connection of the inverters is disconnected and they exhibit sequential write operation. Alternatively, it is turned ON during the read operation to retain the advantage of fast differential read as exhibited in conventional SRAM structures. The control signals CTL1 and CTL2 monitor the operation of 1T1M nonvolatile structure.

In a write operation, the two bitlines BL and BLB are precharged to Vdd prior to writing new data values in the cell. The control signal WL is asserted high, while the signals *W*, *R* are kept at ground potential. The resulting structure represents cascade of two inverters as shown in Fig. 4. The new data loaded on the BLB line get stored at node D1 through Q6, which is then updated sequentially at nodes D and DN through the two inverters (Q1–Q2, Q3–Q4). During a write '0' operation, BLB remains at Vdd (logic 1) and the cascade of two inverters stores logic '0' and logic '1' at nodes D and DN sequentially. Thus, it can be noted here that the proposed cell eliminates the need of discharging of BL bitline in contrast to conventional nvSRAM cells. This reduces the activity factor of the bitlines and offers a method to reduce power of the cell in a write '0' operation. Alternatively, in write '1' event, the BLB line discharges, this in such cases the proposed cell consumes the same power as the conventional structures. The new value stored at D is subsequently stored in memristor M1 by setting CTL1 and CTL2 at high and ground potential, respectively, during store operation.

The proposed cell exhibits differential read operation. During read operation, the control signals WL, *R* and *W* are asserted high, while CTL1, CTL2 are maintained at ground potential. Depending on the value stored at either of the two bitlines, a discharge path is formed as illustrated in Fig. 5. For a read '0', BL discharges through Q3 and Q5 as depicted in Fig. 5a. Alternatively, in a read '1' event, BLB discharges through a path consisting of three transistors Q6, Q7 and Q1 as represented in Fig. 5b. The difference in number of transistors in discharging requires optimum sizing of transistors. During the hold mode, W is asserted high to enable the feedback connection between the two inverters so as to retain

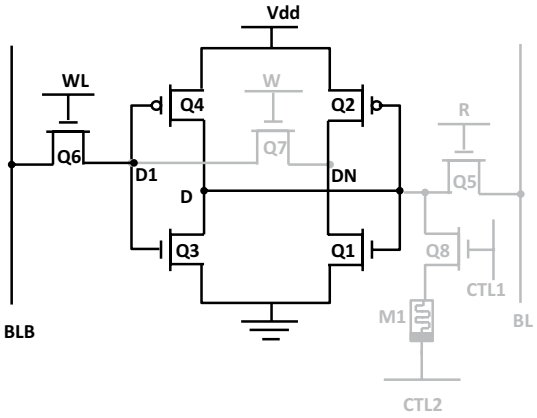


Fig. 4 Proposed 8T1M nvSRAM cell during write operation

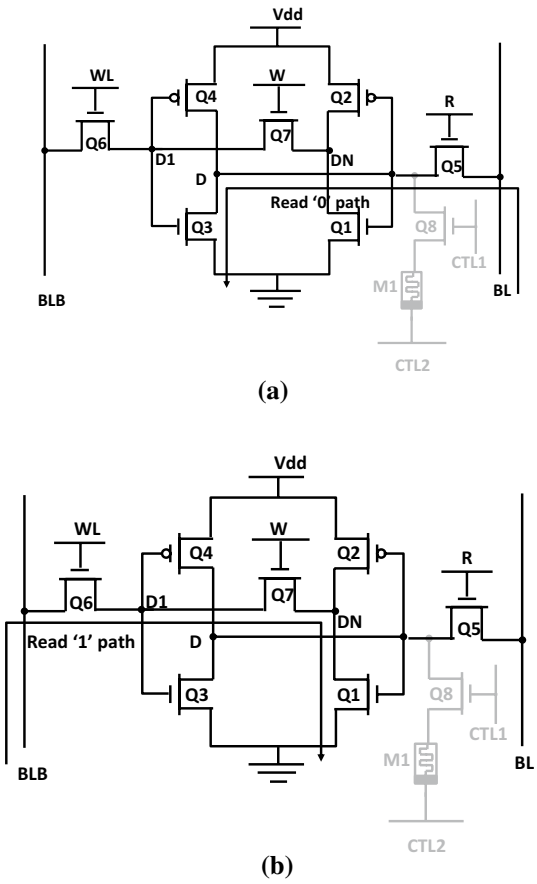


Fig. 5 Proposed 8T1M nvSRAM cell during **a** read '0' operation, **b** read '1' operation

the data values, whereas WL and *R* are asserted low. The rest of the control signals are maintained at ground potential.

The proposed cell exhibits non-volatile operation in four phases namely initial, store, power down and restore mode. In the initial mode, the control signal CTL1 and CTL2 are maintained at ground potential, while the memristor M1 is initialized to HRS state. During the store phase, the data at node D are stored in the memristor. The control signal CTL1 is asserted high, while CTL2 is set to ground potential. If logic '1' is stored at D, then the potential drop across M1 is such that it changes its state from HRS to LRS. Alternatively, for logic '0' stored at D, M1 will not change its state and continue to remain in HRS. At the time of power down, the supply voltage as well as the control signals are turned OFF. Both the storage nodes D and DN discharges and M1 remain in its last state. After the power down, the supply voltage is turned ON, while the control signals WL and R remain at ground potential. To restore the data, the signals W, CTL1 and CTL2 are asserted high. If M1 is in LRS, there will be large flow of current through Q8, the storage node D will be restored to logic high level and M1 will change its state from LRS to HRS, which is also its initial state. If M1 is in HRS, node DN will attain high level due to the cross-coupled inverters and D will settle at low potential. It can be observed that memristor comes into the initial HRS state after the store, power down and restore mode. Thus, the proposed scheme eliminates the requirement of reset state which is necessary in the existing 8T2R structures [25].

4 Simulation Results

In this section, the functionality of proposed 8T1M nvSRAM cell is verified through simulations using 32 nm PTM model and memristor model suggested in [32]. The memristor used in the nvSRAM cells is simulated to determine the cycle time of various operations performed in the memory. Following this, the aspect ratios of various transistors are obtained so that the proposed cell exhibits the same read margin as conventional 6 T SRAM cell. Thereafter, the memory operation in different phases of the proposed cell is verified. The performance of the proposed cell is then compared existing eight transistor nVSRAM cells with same design conditions for the sake of fair comparison. Another design option is also presented wherein the proposed cell having comparable read delay to conventional 6 T SRAM cell is designed and its performance is analyzed.

4.1 Memristor Simulation

The following simulation parameters of the memristor are considered: $R_{OFF} = 16$ KΩ (HRS), $R_{ON} = 100$ Ω (LRS), mobility $\mu_v = 350 \times 10^{-14}$ m^2 V^{-1} S^{-1}, $D = 10$ nm. The change in memristance and conducting width (w) for different values of voltage across the memristor is shown in Fig. 6a, b, respectively. From Fig. 6a, it can be observed that resistance variation from HRS to LRS is completed in 250 ps, 300 ps and 370 ps for the voltage difference of 1.0 V, 0.8 V and 0.6 V, respectively. The resistance switching time is necessary to decide the cycle time required to exhibit the nonvolatility in nvSRAM cell [21]. It can be noted that the resistance does not switch from HRS to LRS at voltage of 0.4 V, a phenomenon that was observed for voltages higher than 0.6 V, 0.8 V and 1.0 V. This behavior is also supported by the conducting width variation plotted in Fig. 6b which also shows a fast change in conducting width for high voltage in contrast to the low voltage.

4.2 Design of Proposed Cell Based on Read Margin

As the number of transistors traced for read '0' and '1' in the proposed cell is not same, therefore the butterfly curve is asymmetric culminating into smaller read margin (RM). In this subsection, various transistors are sized so that the proposed cell has RM comparable to conventional 6 T SRAM cell. Extensive simulations are performed to examine the impact of width of transistors on butterfly curve of the proposed cell, and the observations are plotted in Fig. 7a. It may be noted that there are two curves in Fig. 7a. The horizontal

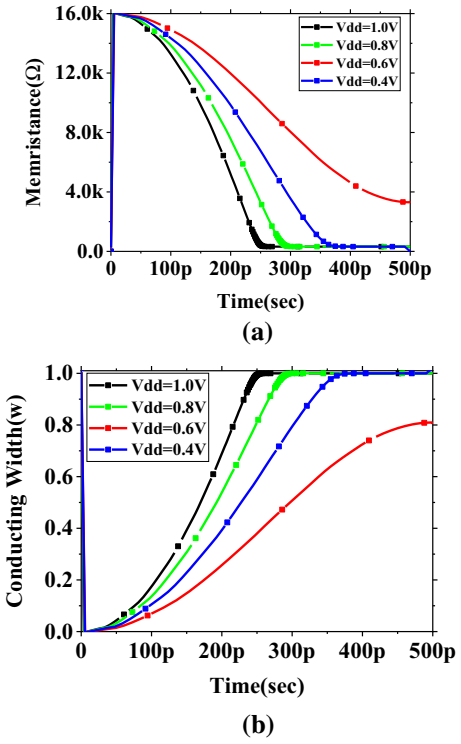


Fig. 6 Memristor switching **a** Memresistance and **b** Conducting width

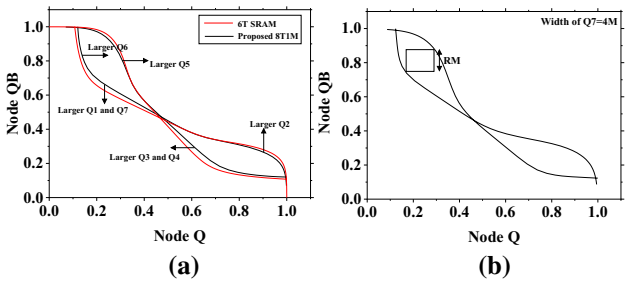


Fig. 7 **a** Butterfly curves for proposed cell **b** RM for proposed cell with width of Q7 = 4 M

curve depends on the width of transistors Q1, Q2, Q6 and Q7, while the vertical curve is influenced by widths of transistors Q3, Q4 and Q5. If the width of transistors Q1 and Q7 is increased, the horizontal curve shifts downwards resulting in improved RM. The increase in the width of transistor Q6 shifts the horizontal curve toward the right side leading to deteriorated RM. Similarly, the width of transistor Q2 will shift the horizontal curve in the upper direction, while the width of transistor Q3 will shift the vertical curve to the left side. Both these variation increases the butterfly curve width. The increase in the width of Q4 will shift the lower vertical curve to the left side, while width of Q5 shifts the upper vertical curve to the right side leads to increase in the RM of the proposed cell.

The RM for the conventional 6 T SRAM cell is 336 mV. To design, the widths of all transistors of proposed cell initially are set to the ones used conventional 6 T SRAM cell and the width of Q7 is varied in multiples (M) of 64 nm. The RM for Q7 width = 1 M, 2 M, 3 M, 4 M and 5 M is obtained as 255 mV, 272 mV, 285 mV, 292 mV and 296 mV, respectively. It may be noted that there is improvement in RM with increase in the width of Q7; however, the gain in RM beyond Q7 width = 4 M becomes insignificant. Therefore, the width of Q7 may be taken as 4 M. Figure 7b shows butterfly curve of the proposed cell with Q7 width = 4 M and other transistor width same as those of convention 6 T SRAM cell. The nature of the butterfly curve for proposed cell is asymmetric, and RM is observed as 292 mV. As discussed in preceding para, the RM improves when.

- widths of Q1, Q3, Q5 and Q6 are reduced from those used in conventional 6 T SRAM cell.
- width of Q2 and Q4 is increased from the one used in conventional 6 T SRAM cell.

The simulations are carried out by varying widths of these transistors, and it is found that RM = 332 mV for the aspect ratios of various transistors listed in Table 1. This value of RM is close to the conventional 6 T SRAM cell. Further, as the transistor Q8 is used to perform the nonvolatile operation, and it remains OFF during read/write operation, the width of transistor Q8 is set to '1 M'. Table 1 also enlists the aspect ratio of various transistors used in existing 8 T nVSRAM cell for comparable read margin.

4.3 Nonvolatile Operation

The proposed nvSRAM cell is simulated to verify its nonvolatile feature. This evaluation includes verifying the complete execution of the cell in store, power down and restore phases. The cycle time which is twice or greater than the time required to change the memristor resistance state at particular Vdd is taken [21].

The timing waveform for the complete execution of store '1' operation in the proposed cell is shown in Fig. 8a. The following observation can be made:

a. In the first cycle, the new data are written on the memory cell. For a write '1' operation, BLB is set to low logic level, the signal WL is asserted high, while W, CTL1 and CTL2 are changed to ground potential. As shown in Fig. 8a, the internal nodes D and DN are set to high and low logic levels, respectively.
b. During the second cycle, the store operation takes place. The control signals CTL1, CTL2 are set to high and

Table 1 Aspect ratio of SRAM cells

Transistor	Cell						
	6 T SRAM	8T1R_C [17]	8T2R [25]	Rnv8T2R [27]	8T1R [28]	RD 8T1R [29]	Proposed 8T1M
Q1	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	211.2 nm/32 nm
Q2	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	128 nm/32 nm
Q3	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	256 nm/32 nm	166.4 nm/32 nm
Q4	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm
Q5	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	64 nm/32 nm
Q6	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	128 nm/32 nm	83.2 nm/32 nm
Q7	–	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	128 nm/32 nm	128 nm/32 nm	256 nm/32 nm
Q8	–	64 nm/32 nm	64 nm/32 nm	64 nm/32 nm	128 nm/32 nm	128 nm/32 nm	64 nm/32 nm

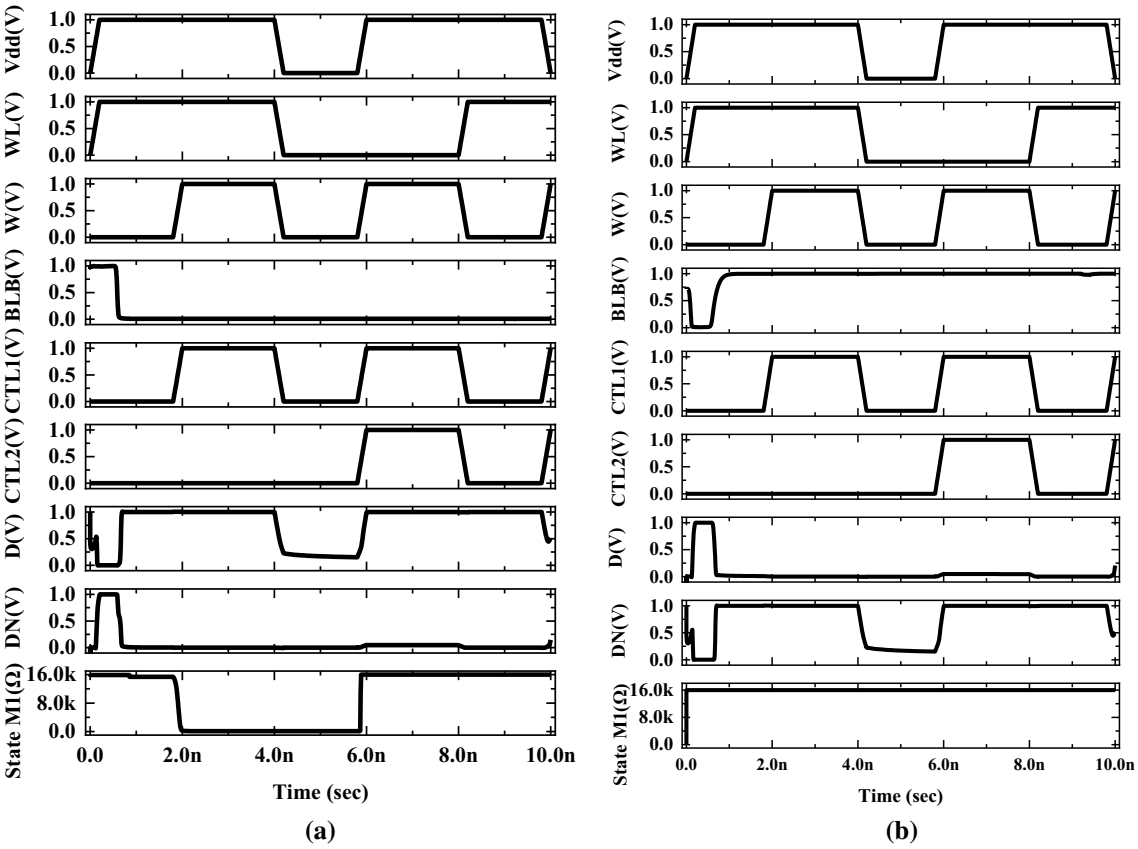


Fig. 8 Nonvolatile operation of the proposed 8T1M nvSRAM cell for **a** Logic '1' **b** Logic '0'

low logic levels, respectively, such that the transistor Q8 turns ON. Since node D is at high logic level and CTL2 is ground potential, the memristor M1 which changes its state from HRS to LRS.

c. The next cycle demonstrates the operation in the power down mode wherein the power supply is turned OFF. It can be observed that the storage node D discharges to ground potential, but the memristor M1 retains its LRS state.
d. Finally, in the restore operation, Vdd is turned ON and CTL1 and CTL2 are made high. The internal node D is charged back to high logic level due to the large flow of current through Q8. After restoring, the memristor M1 changes its state back to HRS which is also its initial state.

Similarly, Fig. 8b shows the complete execution timing waveform for store '0' operation. The control signals CTL1, CTL2 are set high and low, respectively, the transistor Q8 turns ON. As node D is at low logic level, memristor M1 continues to remain in the HRS state. In the power down mode, the internal nodes get discharged. Subsequently, in the restore phase, node D does not get charged as memristor is in HRS state, but node DN attains high logic level due to the cross-coupled inverters in the proposed cell. It is worth noting that after the restore operation, the memristor is in HRS; thus, it eliminates the need of a reset state which is an additional state required in existing 8T2R [25], Rnv8T2R [27] nvSRAM cell.

4.4 Performance Evaluation

In this subsection, the performance of the proposed 8T1M nvSRAM cell is compared with existing eight transistors nvSRAM cells namely 8T2R [25], Rnv8T2R [27], 8T1R [17], 8T1R [28] and the conventional 6 T SRAM cells by designing for same read margins. Various performance parameters such as write power consumption, read power, leakage current, read/write delay, read/write margin, store/restore delay and store/restore energy are evaluated and compared. The effect of variation in power supply and technology is also studied for completeness.

4.4.1 Write Power Consumption

The write power of the proposed 8T1M, existing eight transistors nvSRAM cells and the conventional 6 T SRAM cell for different write data patterns is measured at

Table 2 Write Power Consumption (μW) for write '0' and write '1' patterns in different SRAM cells

Write '0'						
SRAM Cell	0–0 Write Pattern			1–0 Write Pattern		
	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V
6T SRAM	85.75	92.54	98.43	91.25	96.58	100.97
8T1R_C[17]	85.61	92.51	98.85	92.36	97.56	102.01
8T2R [25]	85.91	91.95	98.83	92.65	97.49	101.11
Rnv8T2R [27]	95.36	100.56	104.81	101.57	106.36	109.78
8T1R [28]	85.56	92.05	98.51	92.25	97.05	101.35
RD 8T1R[29]	87.95	95.56	100.05	93.89	98.66	102.35
Proposed 8T1M	0.29	0.32	0.48	2.96	3.85	4.35
Write '1'						
SRAM Cell	0–1 Write Pattern			1–1 Write Pattern		
	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V
6T SRAM	90.78	95.43	100.37	86.93	95.01	98.63
8T1R_C[17]	92.25	96.05	101.56	88.06	95.95	99.15
8T2R [25]	91.78	96.21	100.98	87.03	95.89	98.85
Rnv8T2R [27]	100.85	105.86	110.88	96.62	101.98	104.45
8T1R [28]	92.01	95.99	100.65	87.15	95.52	98.86
RD 8T1R[29]	95.65	98.25	102.95	92.45	97.85	100.12
Proposed 8T1M	93.85	96.56	102.31	90.28	96.75	99.73

different power supply values and summarized in Table 2. It is observed that at 1 V, a write '0' operation in the proposed cell offers 99.5% and 95.7% power saving with respect to conventional 6 T SRAM cell when the original cell data were '0' and '1', respectively. In reference to the existing eight transistors nvSRAM cells, it can be observed that the proposed cell shows a maximum power reduction of 99.7% for the write '0'. This is attributed to the fact that a write '0' operation in the proposed cell does not require BLB to discharge as exhibited in its counterparts. Conversely, in a write '1' operation, the power consumption is increased marginally (1.9%) in comparison with 6 T SRAM cell. However, it can be noted that the increase is not as much percentage of power saving achieved in the proposed cell. Also, it can be noted that with the decrease in power supply, the proposed cell shows similar reduction percentage for write '0' operation, while slight increase of 3.8% is observed in write '1' power consumption value. The effect of technology variation on write power consumption plotted in Fig. 9 reveals that the proposed cell power consumption decreases on reducing the transistor feature size from 90 to 22 nm which is in tune with its counterparts.

4.4.2 Read Power Consumption

The read power of the cells is measured at different power supply values, and the observations are summarized in Table 3. It is evident that all the nvSRAM cells show increase in read power with reference conventional 6 T SRAM cell. The cells 8T2R [25], 8T1R [28] show marginal

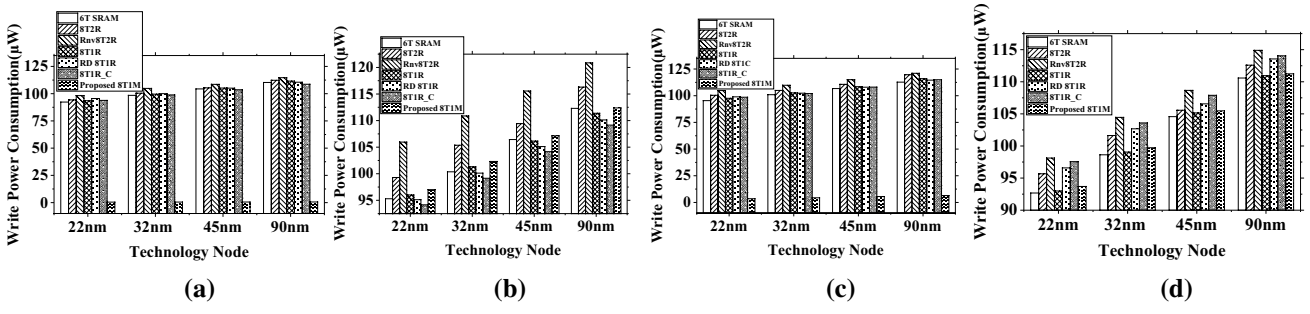


Fig. 9 Variation in write power consumption of the SRAM cells with technology scaling for different write pattern **a** 0–0 **b** 0–1 **c** 1–0 **d** 1–1

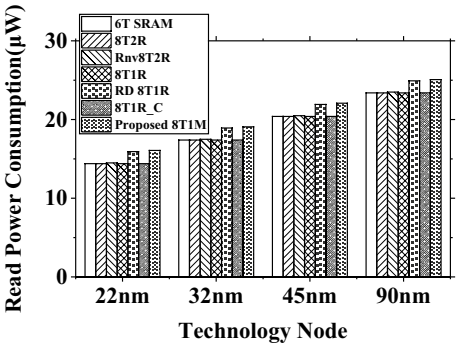


Fig. 10 Read power consumption with respect to technology node

increase due to the addition of controlling transistor. In Rnv8T2R [27] and RD 8T1R [29], the memristor is connected to the storage node which adds to the bitline capacitance leading to a slight increase in power. The higher read power in RD8T1R [29] is an outcome of high value of the read current in the isolated read port which can be traded-off with the read delay. The proposed cell, however, shows an increase of 8.6% in the read power which may be attributed to its larger cell size and activation of the control signals for read operation. Further, it can be noted that the proposed cell and its eight transistor nvSRAM counterparts show similar decreasing trend in read power with lowering of supply voltage. The effect of technology scaling on read power plotted in Fig. 10 reveals that the proposed cell power consumption decreases as is observed for other existing cells on reducing the transistor feature size from 90 to 22 nm.

4.4.3 Leakage Current

The leakage current for the considered cells is noted for different supply values and is summarized in Table 4. It can be observed that all the existing eight transistors nvSRAM cells show increased leakage current values in comparison with conventional 6 T SRAM cell as there exist additional paths in all nvSRAM cells to connect NVM device through a control transistors. However, in proposed cell, the leakage current increases by 12.53%. This happens as one of the PMOS load transistor in the inverter remains in weak conduction since Q7 reduces the high logic level voltage by one threshold voltage drop. The leakage current value can be lowered through the introduction of multiple threshold voltage transistors in the cell design. A low threshold voltage Q7 and/or high threshold voltage PMOS load can be used for leakage current reduction. To examine the effect of such a change on performance parameters, three design cases are considered: Case-1 with low V_T Q7, Case-2 with high V_T Q2/Q4, Case-3 with low V_T Q7 and high V_T PMOS Q2/Q4, and proposed cell is considered as Case-4. The performance parameters are obtained through simulations for all design

Table 3 Read power consumption (μW)

Cell	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V
6 T SRAM	4.15	9.76	17.38
8T1R_C [17]	4.15	9.78	17.38
8T2R [25]	4.16	9.78	17.40
Rnv8T2R [27]	4.18	9.82	17.49
8T1R [28]	4.16	9.78	17.40
RD 8T1R [29]	4.31	10.42	18.71
Proposed 8T1M	4.76	10.82	19.02

Table 4 Leakage Current (pA) variation with supply voltage Vdd

Cell	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V
6 T SRAM	576	1456	2436
8T1R_C [17]	685	1692	2555
8T2R [25]	656	1685	2533
Rnv8T2R [27]	695	1705	2589
8T1R [28]	675	1698	2575
RD 8T1R [29]	701	1756	2605
Proposed 8T1M	845	1855	2785

cases and are enlisted in Table 5. From the three solutions, it can be deduced that use of high V_T PMOS Q2/Q4 shows significant improvement in leakage reduction as well as overall performance improvement. The effect of technology variation on leakage current plotted in Fig. 11 reveals the same increasing trend as observed for other existing 8 T nvSRAM cells.

4.4.4 Read/Write Margin

The read margin (RM) is measured by fitting a smallest square inside the butterfly curve wherein the length of the square is considered as RM [33]. The butterfly curves of the proposed cell shown in Fig. 12a show a decrease in RM with respect to power supply. The word line sweep method suggested by N. Gierczynski et al. [34] is used for write margin (WM) calculations. The difference between Vdd and the point where store data are flipped determines WM, and the corresponding values for the proposed cell are plotted in Fig. 12b for different values of Vdd. Table 6 shows WM values of the existing and the proposed SRAM cells with respect to power supply and technology node variations. It can be observed that the WM of the 8T2R [25] and 8T1R [17], 8T1R [28] is same as the conventional 6 T SRAM cell while for Rnv8T 2R [27], RD 8T1R[29] and proposed cell WM is more than the conventional 6 T cell. In Rnv8T2R [27], the transistors Q7 and Q8 provide an extra write path and decrease the pull up to access transistor ratio, i.e., pull up ratio to provide the improved WM. In RD8T1R [29], cell read–write conflict of 6 T cell is resolved by providing the separate read path and hence, result in the more stable write operation. The proposed cell alters the writing mechanism in the cell by connecting them in cascade and thus eliminating the dependence on pull up to access transistor ratio. From Table 6, the maximum percentage of improvement reduces from 69 to 21% with the lowering of the supply voltage values, while a decreasing trend in WM values with technology scaling is observed in the proposed cell.

4.4.5 Read/Write Delay

The read delay is measured as a time interval between the instance when WL reaches its 50% value to the time, the sense amplifier attains 90% of its full swing value [31]. The write delay is defined as the time interval between the instant when WL reaches its 50% value to the time, DN has reached 90% of its full swing [31]. The read and write delays of the cells are summarized in Table 7. The cells 8T2R [25], 8T1R [28], 8T1R_C [17], Rnv8T2R [27] have comparable delay values to the conventional 6 T cell, while the cell RD8T1R [29] showed reduced delay values due to high read current values in the read port for the considered design. The read delay of the proposed 8T1M nvSRAM cell is greater than the existing counterparts by maximum of 9.45 ps at 1 V. This increase in the delay parameter is due to the stack of transistor formed during the read '1'operation as illustrated in Fig. 5b. In terms of write delay, the proposed cell has 9% increase as the new data are written on the internal nodes of the proposed cell through a cascade connection of inverters in contrast to the cross-coupled connection in the existing cells. An increase in delay values of the proposed cell is observed with respect to decrease in power supply values which is similar to trend exhibited by the existing counterparts. Figure 13 shows the read/write delay for different cells at different technology node. A decreasing trend in the delay values is observed with technology scaling.

4.4.6 Store/Restore Performance Delay

The performance of the nvSRAM cells during store and restore operation is evaluated in terms of delay and energy parameters dependent on the data value. The average delay and energy values for store and restore operation of logic 0/logic 1 are listed in Tables 8 and 9, respectively. It is observed that the proposed cell is the fastest among all in performing the store and restore operations. Quantitatively, there is a maximum reduction of 54.94% and 67.9% is observed in store and restore delays in comparison with the existing eight transistors nvSRAM cells. In the existing eight transistors nvSRAM cells, the programming of the NVM takes more time due to switching of power supply voltage to higher levels or two phases in store/restore operations. This leads to the increase in the store/restore delay of the existing cells. Such a programming is not required in the memristor

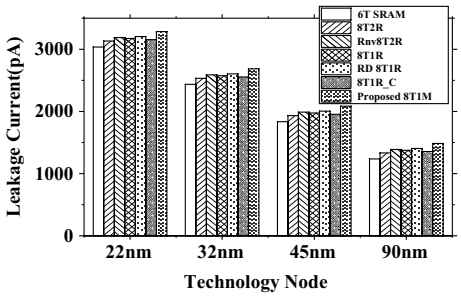


Fig. 11 Leakage Current variation with respect to technology node

Table 5 Variation in performance parameters of the proposed nvSRAM cell for different design cases

Performance parameter	Case			
	Case-1	Case-2	Case-3	Case-4
Leakage (pA)	2659	2485	2566	2785
Read power '0'(μW)	18.36	14.15	18.45	19.02
Read delay (ps)	62.86	64.59	61.56	65
Write delay (ps)	57.76	58.85	56.58	59
Read margin (mV)	325	310	305	336
Write margin (mV)	595	605	582	629
Write power (0–0)	0.39	0.45	0.42	0.48
Write power (0–1)	99.65	101.95	100.85	102
Write power (1–0)	4.25	4.31	4.28	4.35
Write power (1–1)	97.56	99.05	98.51	99.73
Store delay (ps)	39.15	40.68	38.65	41
Restore delay(ps)	25.15	25.55	24.56	26
Store energy (fJ)	181	231	196	263
Restore energy(fJ)	255	295	269	313

Table 6 WM (mV) for different SRAM cells

SRAM cell	WM at different power supply			WM at different technology node			
	Vdd = 0.6 V	Vdd = 0.8 V	Vdd = 1.0 V	22 nm	32 nm	45 nm	90 nm
6 T SRAM	271	335	371	278	371	390	410
8T1R_C [17]	271	335	371	278	371	390	410
8T2R [25]	271	335	371	278	371	390	410
Rnv8T2R [27]	310	430	605	550	605	650	690
8T1R [28]	271	335	371	278	371	390	410
RD 8T1R [29]	340	482	635	585	635	682	712
Proposed 8T1M	326	466	629	570	629	675	705

Table 7 Read/write delay of different cells

Delay	Read (ps)			Write (ps)		
Cell	Vdd					
	0.6 V	0.8 V	1.0 V	0.6 V	0.8 V	1.0 V
6 T SRAM	38	48	56	43	51	54
8T1R_C [17]	39.6	48.9	57.09	44.05	52.56	54.1
8T2R [25]	39	50	57	44.56	52.09	55.15
Rnv8T2R [27]	52	55	59	40	49	51
8T1R [28]	40	49	57	45.05	52.85	55.92
RD 8T1R [29]	29.95	36.12	41.77	41.81	49.58	52.5
Proposed 8T1M	50	59	65	52	56	59

Table 8 Average store/restore delay for nvSRAM cells

Delay	Avg. Store Delay (ps)			Avg. Restore Delay (ps)		
	Vdd					
Cell	0.6 V	0.8 V	1.0 V	0.6 V	0.8 V	1.0 V
8T1R_C[17]	45.5	54.6	57.5	36.8	40.2	43.65
8T2R [25]	59	61	63	40	43	45
Rnv8T2R [27]	85	89	91	75	78	81
8T1R [28]	40	51	56	30	35	40
RD 8T1R[29]	38.89	44.16	53.55	28.64	32.82	36.51
Proposed 8T1M	36	39	41	23	25	26

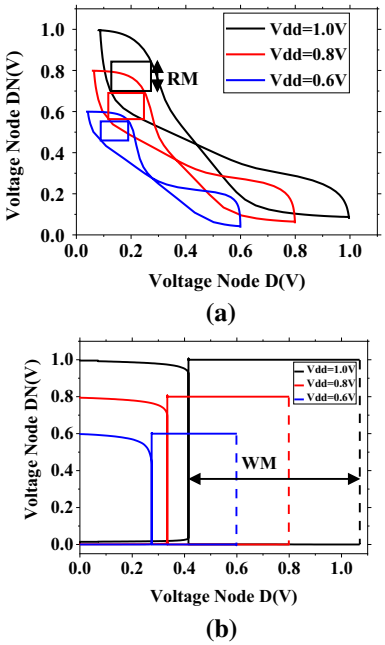


Fig. 12 Margins of the proposed 8T1M nvSRAM cell **a** RM **b** WM

used in the proposed nvSRAM cell. Furthermore, there is maximum of 39.95% and 37.27% reduction in store and restore energy, respectively, in comparison with the existing

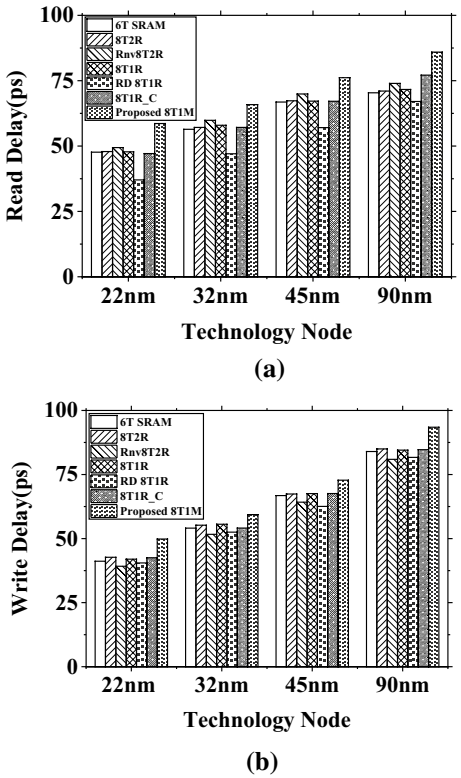


Fig. 13 **a** Read delay **b** Write delay of SRAM cells at different technology node

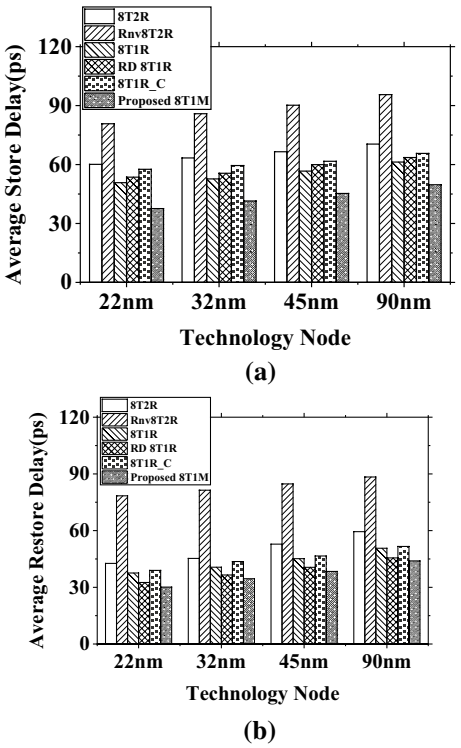


Fig. 14 **a** Store delay **b** restore delay of the proposed 8T1M nvSRAM cell at different technology node

eight transistors nvSRAM cell. The proposed cell avoids the requirement of raising the power supply to higher voltage levels as is performed in the existing nvSRAM cells. Thus, the proposed cells exhibit better nonvolatile performance than the existing eight transistors nvSRAM cells. The delay and energy variation with respect to technology is shown in Figs. 14 and 15, respectively. A decreasing trend in delay and energy values is observed with technology scaling as exhibited by other cells.

4.5 Design Based on Timing parameters

In this subsection, another design option is investigated wherein the aspect ratios of transistors for proposed cell are obtained for comparable read delay with conventional 6 T SRAM cell. The performance parameters are also examined and compared with the proposed design given in Sect. 4.2 and conventional 6 T SRAM cell.

In the proposed cell, the transistors Q7 is turned ON during read operation, while it remains OFF during write operation. The path for read '1' operation comprises of three transistors (Q6, Q7 and Q1) connected in cascade, while read '0' operation follows path comprises of transistors Q5 and Q3. The cascade connection of three transistors during read '1' operation makes it critical read path; hence, it necessitates appropriate sizing of transistors Q1, Q6 and Q7 so that read delay is comparable to the conventional 6 T SRAM cell. As there should be minimum variation across process corners, due consideration is given while selecting aspect ratios. In the following discussion, the transistor width is given in multiples of '*M*' where $M = 64$ nm and length is kept minimum, i.e., 32 nm.

The sizing of various transistors is carried out by taking timing parameters of conventional 6 T SRAM cell as reference in context of this subsection. Initially, widths of all the transistors of proposed cell except Q1, Q6 and Q7 are kept same as those used in reference cell to achieve the comparable read delay. To start with, width of the transistor Q1 is taken as 4 M. The variation in read delay with the widths of transistors Q6 and Q7 is observed and findings are plotted in Fig. 16a It is noted that the read delay majorly depends on the width of transistor Q6 and is its value is nearly same as reference cell for Q6 = 1 M. The read delay is approximately same as the reference cell for transistor widths of Q7 = 3 M, 4 M and 5 M, which is 56 ps. To investigate the impact of process variation and appropriate dimensioning, read delay is observed at different process corners and best to worst read delay ratio plotted in Fig. 16b. It is observed that for Q1 = 4 M, Q6 = 1 M and Q7 = 5 M, the best to worst ratio for

Table 9 Average store/restore energy for nvSRAM cells

Energy	Avg. Store Energy (fJ)			Avg. Restore Energy (fJ)		
Cell	Vdd					
	0.6 V	0.8 V	1.0 V	0.6 V	0.8 V	1.0 V
8T1R_C[17]	260	299	316	325	420	462
8T2R [25]	295	319	342	455	471	483
Rnv8T2R [27]	391	410	438	460	480	499
8T1R [28]	250	290	310	305	400	450
RD 8T1R[29]	275	305	320	350	430	475
Proposed 8T1M	205	230	263	279	300	313

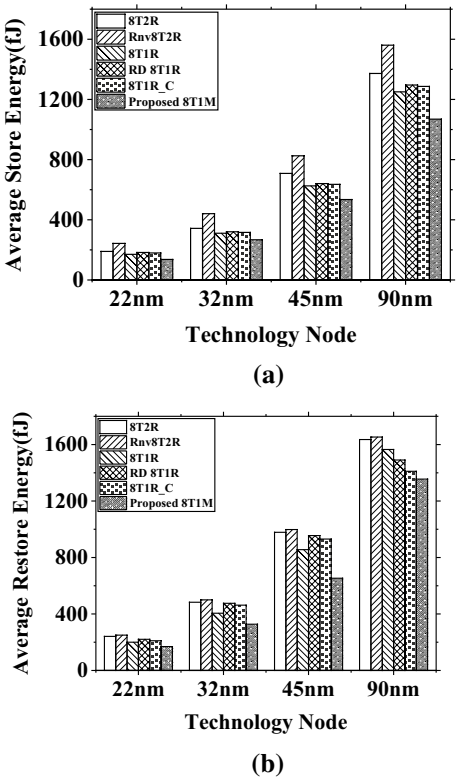


Fig. 15 **a** Average Store Energy **b** Average Restore Energy of the proposed 8T1M nvSRAM cell at different technology node

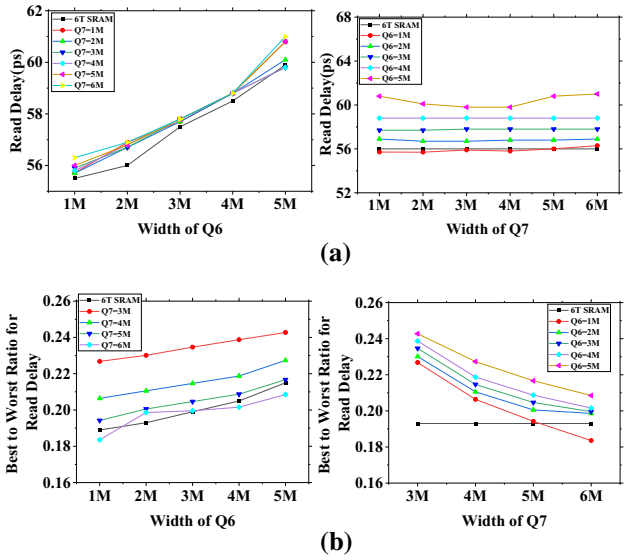


Fig. 16 Effect of width of Transistor Q6/Q7 on **a** Read Delay **b** Ratio of best to worst Read delay at different process corners

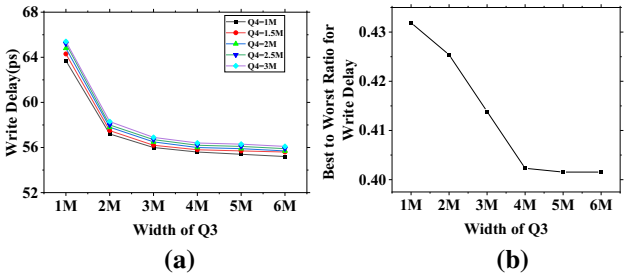


Fig. 17 **a** Write Delay with Width of Q3 **b** Ratio of best to worst Write delay at different process corners

read delay is minimum; therefore, for further investigation these dimensions are considered.

Similar analysis is carried out for write delay which is cumulative delay of inverter1(Q1-Q2) and inverter2 (Q3-Q4). The write '1' is performed by discharging BLB to GND, whereas BLB remains at Vdd for write '0' operation. As the access transistor does not pass a strong '1', the transistor Q3 must be properly sized so that a strong '0' is obtained at the output of inverter2. The write delay is observed for different widths of transistors Q3 and Q4, and observations are plotted in Fig. 17a. It is noted that for transistor Q3 widths = 4 M, 5 M, 6 M and Q4 = 1 M, the write delay is minimum. The impact of process variation on write delay is investigated through corner analysis. The width of Q4 is kept at 1 M, while width of Q3 is varied from 4 to 6 M. The best to worst write delay ratio is plotted in Fig. 17b. It is observed that the best to worst ratio for write delay is minimum for Q3 = 4 M and for Q3 = 5 M and 6 M. Hence, Q3 = 4 M and Q4 = 1 M is fixed for the minimum write delay.

The widths of remaining transistors are finalized for RM. As the proposed cell has two different read paths, its butterfly curve has two asymmetric lobes. The width of transistor Q2 will affect the lower lobe, while the width of transistors Q5 will affect the upper lobe of the butterfly curve. In Fig. 18, the variation in RM with the dimension of transistors Q2 and Q5 is plotted. It is observed that though RM is maximum for Q2 = 2 M and Q5 = 2 M, its value is nearly same for widths of transistor Q2. Therefore, Q2 = 1 M and Q5 = 2 M is chosen.

The widths of various transistors for the proposed cell for comparable read delay are: Q1 = 4 M, Q2 = 1 M, Q3 = 4 M, Q4 = 1 M, Q5 = 2 M, Q6 = 1 M, Q7 = 5 M and Q8 = 1 M. Various performance parameters are measured and are placed in Table 10 under column Design1, and the corresponding parameters with design given in Sect. 4.2 are placed under column Design2. Conventional 6 T SRAM cell performance parameters are also placed in Table for quick reference and comparison.

Lastly, the effect of transistors Q8 width on store and restore delay is examined and findings are plotted in Fig. 19. It may be noted that the proposed cell requires more store/restore time as width of transistor Q8 increases. Hence, Q8 = 1 M is enough for successful store/restore operation

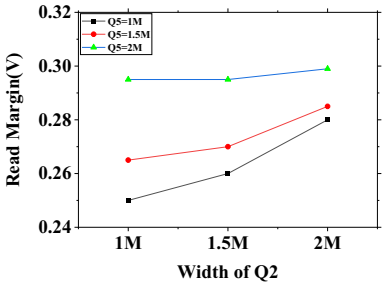


Fig. 18 Read Margin variation with width of Q2

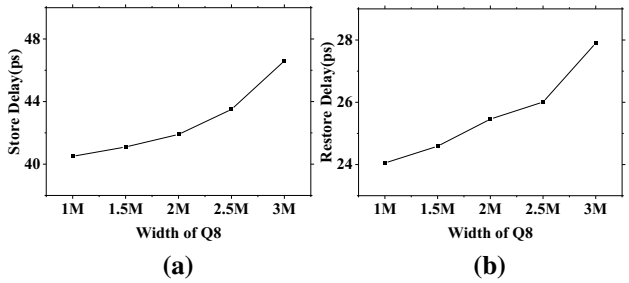


Fig. 19 Variation in **a** Store Delay **b** Restore Delay with width of Transistor Q8

Table 10 Summary of performance parameters of 6 T SRAM and proposed designs

Parameters	6 T SRAM	Design1	Design2
Write power (0–0) (μW)	98.43	0.41	0.48
Write power (0–1) (μW)	100.37	99.56	102.31
Write power (1–0) (μW)	100.97	3.52	4.35
Write power (1–1) (μW)	98.63	95.86	99.73
Read power (μW)	17.38	21.56	19.07
Leakage current (pA)	2436	3156	2785
Write margin (mV)	371	585	629
Read margin (mV)	336	295	336
Read delay (ps)	56	56	65
Write delay (ps)	54	55.6	59
Store delay (ps)	–	40.5	41
Restore delay (ps)	–	24.05	26
Store energy (fJ)	–	256	263
Restore energy (fJ)	–	305	313

with minimum delay. A close observation of Table 10 reveals that.

- both designs have significant saving in power for writing '0' in the cell in comparison with conventional 6 T SRAM cell. The write power consumption of the proposed cell is lower than with Design1 as the width of transistor Q6 is smaller in Design1 as compared to the Design2.
- the read power in Design2 is smaller than the Design1 as the width of transistors in read path for Design2 is smaller than those in Design1.
- the leakage power in Design1 is higher than Design2 which may be attributed to larger aspect ratios in former one.
- Due to adopted designing process in Design1 and Design2—the read margins of Design2 are better than Design1; and the read and write delay of Design1 are comparable to conventional 6 T SRAM cell, whereas these parameters are larger for Design2.
- the nonvolatile performance parameters of the Design2 are superior to Design1.

Thus, a trade-off exists in the design and appropriate design option may be adopted as per constraints.

5 Conclusion

In this paper, a low-power 8T1M nonvolatile SRAM cell is proposed that employs a single TiO2 memristor as a nonvolatile/NVM device. The proposed cell reduces the power consumption by avoiding the discharge of bitlines during write '0' operation. Additionally, the proposed cell avoids the use of multiple cell power supply voltage levels, two phases in store/restore mode, additional RESET phase as well as complex sequencing of control signals in comparison with the existing nvSRAM cells. Two designs of the proposed cell based on the conditions of maintaining same read margins as well as read delay are also put forward. The functionality of the cell is verified through simulations, and a performance comparison with the existing cells is performed. The results indicate that the proposed 8T1M nvSRAM cell reduces the write power consumption, low store/restore energy by 99.7% and 47%, respectively, in comparison existing 8 T nvSRAM counterparts. Also, the corresponding improvement of 69% and 70% is observed in write margins and average store/restore delay of the proposed cell with respect to the existing counterparts.

References

1. Morifuji, E.; Yoshida, T.; Kanda, M.; Matsuda, S.; Yamada, S.; Matsuoka, F.: Supply and threshold-voltage trends for scaled logic and SRAM MOSFETs. IEEE Trans. Electron Dev. **53**(6), 1427–1432 (2006)
2. Kozma, R.; Pino, R. E.; Pazienza, G. E.: Advances in Neuromorphic Memristor Science and Applications. Springer Series in Cognitive and Neural Systems. (2012).

3. Mohammad, B.; Homouz, D.; Elgabra, H.: Robust hybrid memristor-CMOS Memory Modeling and design. IEEE Trans Very Large Scale Integr. Syst. **21**, 2069–2079 (2013)
4. Kvatinsky, S.; Friedman, E.G.; Kolodny, A.; Weiser, U.C.: (2013) The desired memristor for circuit designers. IEEE Circ. Syst. Mag. **13**, 17–22 (2013)
5. Talukdar, A.; Radwan, A.G.; Salama, K.N.: Non linear dynamics of memristor based 3rd order oscillatory syst. Microelectron. J. **43**, 169–175 (2012)
6. Adamatzky, C. L. : Memristor Networks. Springer International Publishing, Switzerland (2014)
7. Takashima, D.; Nagadomi, Y.; Hatsuda, K.; Watanabe, Y.: A 128 Mb Chain FeRAM and system design for HDD application and enhanced HDD performance. IEEE J. Solid State Circ . **46**, 530–536 (2011)
8. Qazi, M.; Clinton, M.; Bartling, S.; Chandrakasan, A.P.: A Low-Voltage 1 Mb FRAM in 0.13µm CMOS featuring time-to-digital sensing for expanded operating margin. IEEE Int. Solid State Circ. Conf. (ISSCC) **47**, 141–150 (2012)
9. Nebashi, R. et al.: A 90nm 12ns 32Mb 2T1MTJ MRAM. IEEE Int. Solid State Circ. Conf. (ISSCC). 462–463 (2009).
10. Lee, K.; Cho, B.; Kang, S.; Choi, B., et al.: A 90 nm 1.8V 512 Mb Diode-Switch PRAM with 266Mb/s read throughput. IEEE J. Solid State Circ. **43**, 150–162 (2008)
11. Hanzawal, S. et al.: A 512kB embedded phase change memory with 416kB/s Write throughput at 100A cell write current. IEEE Int. Solid State Circ. Conf. (ISSCC) 474–475 (2007).
12. Kawahara, T. et al.: 2 Mb spin transfer torque RAM(SPRAM) with bit-by-bit bidirectional current write and parallelizing-direction current read. IEEE Int. Solid State Circ. Conf. (ISSCC) **18**, 280–281 (2007)
13. Chua, L.: Memristor-The missing circuit element. IEEE Trans. Circ. Theory. **18**, 507–519 (1971)
14 Williams, R.S.: How we found the missing memristor. IEEE Spectr. **45**, 29–35 (2008)
15. Strukov, D.B.; Snider, G.S.; Stewart, D.R.; Williams, R.S.: The missing memristor found. Nature **453**, 80–83 (2008)
16. Wang, W.; Gibby, A.; Wang, Z.; Chen, T.W.; Fujita, S.; Griffin, P.; Nishi, Y.; Wong, S.: Nonvolatile SRAM cell. Tech. Dig. Int. Elect. Dev. Meet. IEDM. 1–4 (2006).
17. Bazzi, H.; Harb, A.; Aziza, H.; Moreau, M.; Kassem, A.: RRAM-based non-volatile SRAM cell architectures for ultra-low-power applications. Analog Integr. Circ. Sign. Process. **106**, 351–361 (2021)
18. Singh, J.; Mohanty, S.P.; Pradhan, D.K.: Robust SRAM designs and analysis. Springer, New York (2013)
19. Sheu, S.S.; Kuo; et.al: A ReRAM integrated 7T2R non-volatile SRAM for normally-off computing application. Proc. 2013 IEEE Asian Solid-State Circuits Conf. A-SSCC 2013. 245–248 (2013).
20. Singh, J.; Raj, B.: Design and investigation of 7T2M-NVSRAM with enhanced stability and temperature impact on store/restore energy. IEEE Trans Very Large Scale Integr. Syst. **27**, 1322–1328 (2019)
21. Wei, W.; Namba, K.; Han, J.; Lombardi, F.: Design of a non-volatile 7T1R SRAM cell for instant-on operation. IEEE Trans. Nanotechnol. **13**, 905–916 (2014)
22. Lee A.; Chang, M.;, Lin, C.; Chen, C.; Ho, M.; Kuo, C.; Tseng, P.; Ku, S.S.D.Q.G.T.: RRAM-based 7T1R Nonvolatile SRAM with 2x Reduction in Store Energy and 94x Reduction in Restore Energy for Frequent-Off Instant-On Applications. in Proc. IEEE Symp. VLSI Circuits, Kyoto, Japan. C76–C77 (2015).
23. Peng, C.; Xiao, S.; Lu, W.; Zhang, J.; Wu, X.; Chen, J.; Lin, Z.: Average 7T1R Nonvolatile SRAM with R/W Margin Enhanced for Low-Power Application. IEEE Trans. Very Large Scale Integr. Syst. **26**, 584–588 (2017)
24. Lin, Z.; Wang, Y.; Peng, C.; Wu, X.; Li, X.; Chen, J.: Multiple Sharing 7T1R Nonvolatile SRAM with an Improved Read/Write Margin and Reliable Restore Yield. IEEE Trans Very Large Scale Integr Syst **28**, 607–619 (2020)
25. Yamamoto, S.; Shuto Y.; Sugahara, S.: Nonvolatile SRAM (NV-SRAM) using functional MOSFET merged with resistive switching devices. Proc. Cust. Integr. Circuits Conf. 531–534 (2009).
26. Turkyilmaz, O.; Onkaraiah, S.; Reyboz, M.; Clermidy, F.; Hraziia, A.C.; Portal, J.M.; Bocquet, M.: RRAM-based FPGA for "normally Off, Instantly On" applications. J. Parallel Distrib. Comput. **74**, 2441–2451 (2014)
27. Chiu, P.F.; Chang, M.F.; Wu, C.W.; Chuang, C.H.; Sheu, S.S.; Chen, Y.S.; Tsai, M.J.: Low store energy, low vddmin, 8t2r non-volatile latch and sram with vertical-stacked resistive memory (Memristor) devices for low power mobile applications. IEEE J. Solid State Circ. **47**, 1483–1496 (2012)
28. Tosson, A.M.S.: 8T1R : A Novel Low-power High-speed RRAM-based Non-volatile SRAM Design. 2016 Int. Gt. Lakes Symp. VLSI. 239–244 (2016).
29. Lin, Z.; Wang, Y.; Peng, C.; Lu, W.; Li, X.; Wu, X.; Chen, J.: Read-decoupled 8T1R non-volatile SRAM with dual-mode option and high restore yield. Electron. Lett. **55**, 519–521 (2019)
30. Singh, S.; Mishra, V.: Enhanced static noise margin and increased stability sram cell with emerging device memristor at 45-nm technology. Radioelectron. Commun. Syst. **61**, 267–274 (2018)
31. Aly, R.E.; Bayoumi, M.A.: Low-Power Cache Design Using 7T SRAM Cell. IEEE Trans. Circ. Syst. II Express Briefs **54**, 318–322 (2007)
32. Biolek, D.; Di Ventra, M.; Pershin, Y.V.: Reliable SPICE simulations of memristors, memcapacitors and meminductors. Radioengineering. **22**, 945–968 (2013)
33. Takeda, K.; Hagihara, Y.; Aimoto, Y.; Nomura, M.; Nakazawa, Y.; Ishii, T.; Kobatake, H.: A read-static-noise-margin-free SRAM cell for low-Vdd and high-speed application. IEEE J. Solid-State Circ. **41**, 113–121 (2006)
34. Gierczynski, N., Borot, B., Planes, N., Brut, H.: A new combined methodology for write-margin extraction of advanced SRAM. IEEE Int. Conf. Microelectron. Test Struct. https://doi.org/10.1109/ICMTS.2007.374463 (2007)

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354313208

A Parametric Study of Eco-Friendly Refrigeration with Waste Heat Recovery from Gas Power Plants using Loop Heat Pipe

Article *in* Journal of Mechanical Engineering Research and Developments · September 2021

CITATIONS
0

READS
26

2 authors:



Ankit Dwivedi
Delhi Technological University
12 PUBLICATIONS **7** CITATIONS
SEE PROFILE



Radhey Shyam Mishra
Delhi Technological University
53 PUBLICATIONS **175** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Doctoral Research View project

Performance Measurement View project

All content following this page was uploaded by Ankit Dwivedi on 02 September 2021.

The user has requested enhancement of the downloaded file.

Journal of Mechanical Engineering Research and Developments
ISSN: 1024-1752
CODEN: JERDFO
Vol. 44, No. 9, pp. 286-310
Published Year 2021

A Parametric Study of Eco-Friendly Refrigeration with Waste Heat Recovery from Gas Power Plants using Loop Heat Pipe

Ankit Dwivedi* and R S Mishra

Research Scholar, Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India.

*Corresponding Author Email: ankitdwivedi2008@gmail.com, rsmishra@dtu.ac.in

ABSTRACT

In numerous Industrial setups such as Power Plants, Steel Plants, Food Processing, etc., a considerable amount of waste heat is available which goes into the environment to contribute to Global Warming and Thermal Pollution. In this research investigation Heat Recovery from Gas Power Cycle (GPC) exhaust has been objectified to operate Ejector Refrigeration Cycle (ERC) through the superconductor of heat i.e., Loop Heat Pipes (LHP). In LHP fluids like Acetone, Ethanol, Methanol & Water, and in ERC new eco-friendly refrigerants such as R236ea, R1224yd (Z), R1233zd (E), R245fa, R365mfc along with R718 have been selected for study as the working fluids. Results obtained from the software-based mathematical modeling have been presented and an assessment of this novel system for industrial viability has been done. It has been observed that the boiler temperature of ERC can be maintained near or above the critical temperature of the Eco-friendly refrigerants working fluids. COP & refrigeration capacity of the combined system has been obtained in the range of 0.25-0.28 and 10.35kW-315kW respectively for various combinations of fluids. Based on the eco-friendliness, compactness & industrial viability the system with Water (LHP)-R1224yd (Z) (ERC) has been recommended as the mass flow required for working fluids is the least & utilization of Heat Input Available to ERC is maximum for the operations.

KEYWORDS

Heat Pipes, Ejector Refrigeration System, Brayton Cycle, Modern Refrigerants, Heat Exchangers, Combined Cycle

INTRODUCTION

The world has seen numerous methods of refrigeration across the millenniums. From the use of Ice & low temperatures at night to the modern huge HVAC plants, there has been extraordinary development in the field of refrigeration technology. The most common and most efficient refrigeration system is the Vapour Compression Refrigeration System for it works on High-Grade Energy & has the highest Coefficient of Performance (COP) of all. However, to generate this high-grade energy extremely high costs are incurred in the power plants along with wastage of Energy, Exergy, Fuel, and the Generation of all sorts of pollution. It can be simply stated that everything comes at a cost. One of such power plants is the Gas Power Plant (GPP) working on Brayton Cycle which we all are well acquainted with. Even after the regenerator of the GPP, a considerable amount of energy remains in form of the exhaust gases. Hence, a Highly Flexible & Highly Efficient Superconductor Loop Heat Pipe Heat Exchanger (LHP HEx) can be incorporated to harness the waste heat and utilize it in other systems which can satisfactorily work on Low-Grade Energy reducing thermal pollution and global warming.

From Figure 1, the LHP can be explained as a device that can be used to transfer heat from one point to another point over a long distance with the help of an insulated vapor line and liquid line. It works on the phenomenon of Evaporation-Condensation. It can be simplified to be similar to Power Cycle less the Turbine & Pump; the remaining components of the power cycles such as the evaporator & the condenser are connected through thermally insulated lines called vapor lines & liquid lines and the evaporator is in porous wicked construction. A

working fluid is filled under pressure within the LHP system, which after receiving heat evaporate and moves out of the evaporator through the porous structure, it enters the transfer lines (Vapour Line & Liquid Line) to move to the condenser part, where the fluid from evaporator condenses and returns to the evaporator owing to the buoyancy, hence, completing the cycle. Hence no input work is required here to transfer heat from one point to another distant point. More than one LHP Ex. May also be suggested for one system to take the complete heat load.

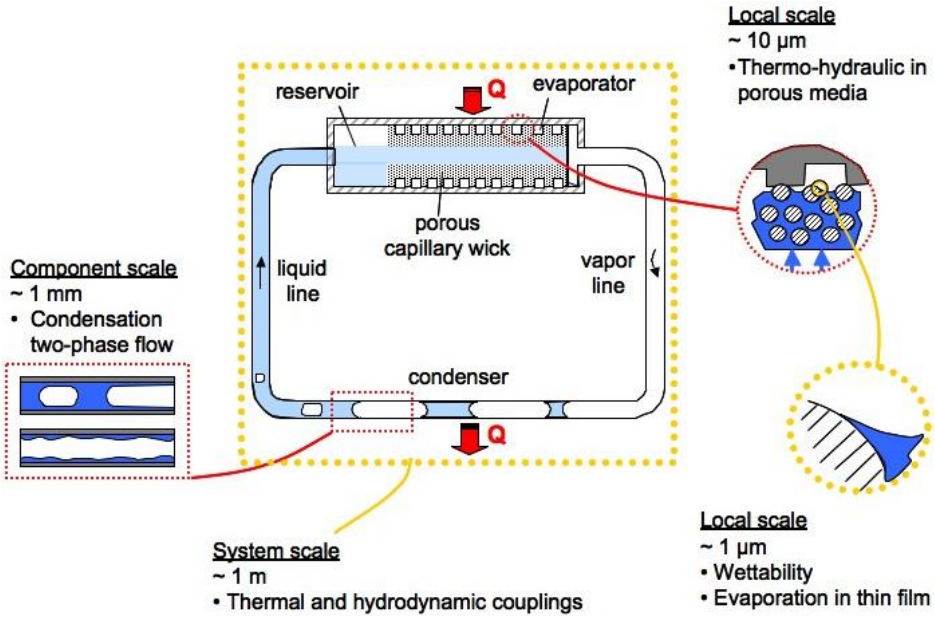


Figure 1. Porous Wick Structure in the LHP [33]

In search of such a system that can work on waste heat and generates desired effects, we came across the Ejector Refrigeration System (ERS) working on new eco-friendly refrigerants which have a relatively low boiling point, hence, the waste energy at relatively low temperature can be harnessed. It has a complicated operation in which converging-diverging nozzles, flash evaporation and refrigeration can be observed to take place. Flash evaporation of ERS & flash evaporation in VCRS both occur due to sudden pressure drop, however, the causes of pressure drops are different for the two.

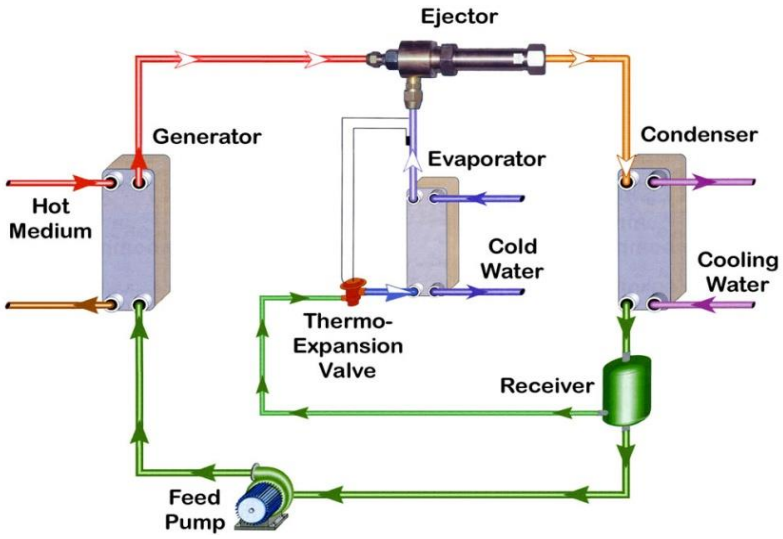


Figure 2. Ejector Refrigeration System [34]

Owing to the requirement of heat at low temperatures only, ERS has been attached with many systems including solar power sources. Through the literature review followings, advancements have been studied on ERS & LHP: Untea et. al. [1] performed energy and exergy investigation of an ejector refrigeration system working on

different working fluids. For 4 working fluids water, methanol, ammonia, and R134a, the best performances were achieved for water. The optimum value was reached for Boiler Temperature 140°C, Condenser Temperature 30° C, Evaporator Temperature 5°C which resulted in a COP of 0.48 and ηex of 0.085. Ebadollahi et. al. [2] used multi-parallel ejector and low-temperature heat sources in the Ejector Refrigeration Cycle (ERC) to evaluate the first and second law analysis keeping the system running at optimum conditions under various working conditions. The maximum and minimum coefficient of performance (COP) was recorded at 0.344 for R152a and 0.285 for R236fa, respectively, whereas the maximum and minimum exergy efficiencies were 33.98% for R152a and 27.63% for isobutene, respectively. Besagni et. al. [3] performed an overall review of the ERS system working on low GWP Refrigerants.

Akkurt et. al. [4] analyzed a solar-assisted ejector cooling system having varying ejector area ratios such as Ar = 6.56, Ar = 7.17, and Ar = 7.86. Exergy destruction proportions found were 42.9%, 44.7%, and 45.2% in the ejector, 10.9%, 9.1%, and 10.1% for generator, 7.6%, 7.7%, and 8.7% for condenser, 5.9%, 5.6%, and 5.8% for evaporator of total cooling subsystem for different ejector area ratios. Besagni et. al. [5] performed another comprehensive study and review of the ERS. An overview of ejector technology, relationships among the working fluids, and the ejector performances, with a focus on past, present, and future trends, were presented. Liu [6] studied a performance improvement potential analysis for a booster-assisted ejector refrigeration system. Outcomes showed that 61.6% of the system exergy destruction occurs due to the components. Also, 55.5% of the overall exergy destruction may be avoided by improving component efficiencies. Memet et. al. [7] investigated theoretically an Ejector Refrigeration System, started with a typical ejector design. For the chosen input conditions, it was revealed that the Coefficient of Performance increased with the increase of the boiler temperature with the best COP being 0.178. Taleghani et. al. [8] analyzed the exergy of a CO2 (R744) two-phase ejector using a 1D model for single and double choking conditions.

It was observed that the transiting exergy flow had an important effect on the exergy analysis of the system. It was also observed that the Grassmann exergy efficiency was not the suitable standard for evaluating the performance of a trans-critical CO2 ejector. Al-Sayyab et. al. [9] analyzed a compound PV/T waste heat driven ejector-heat pump system for simultaneous data center cooling and waste heat recovery for district heating. The compressor contributed the largest exergy destruction source at 26%, the boiler was the lowest source at 2%. Moreover, advanced exergy analysis indicated that 59.4% of the whole system may be avoided. The condenser has the highest &the ejector has the lowest exergoeconomic factors. Sharma et. al. [10] applied exergy analysis to each component of the system. The refrigerant R1234yf had been used and the results exhibited that maximum exergy destruction in the generator followed by ejector and other components. Kumar et. al. [11] investigated the performance of the ejector refrigeration system working on R-134a.It was observed that an optimum area ratio of the ejector was required for better performance.

Moreover, a higher value of area ratio gives higher COP for a lower value of critical condenser pressure. The critical condenser pressure for the area ratio 10.08 was 778.9 kPa, whereas, it was 916.72 kPa for the ejector of area ratio 6.451 while boiler and evaporator temperature were kept at 80 °C and 15 °C respectively. Ebadollahi et. al. [12] conducted a theoretical analysis of the triple-evaporator ejector refrigeration cycle (TEERC) for triples applications of cooling, freezing, and ventilation, with Nine appropriate working fluids i.e., R717, R152a, R134a, R290, cis-2-butene, butane, isobutene, isobutane, R236fa. The maximum & minimum COP were observed for R717 and R236fa at 0.333 and 0.268, respectively. The maximum & minimum exergy efficiencies were considered for R717 and isobutene at 21.43% and 12/51 %, respectively. Ventilation, cooling, and freezing capacities recorded were 11.68 kW, 3.86 kW & 1.904 kW, respectively. Reddy [13] presented a numerical analysis of ejectors optimizing operating evaporator temperature, condenser temperature, and generator temperature with R245fa as the working fluid. Li et. al. [14] chose refrigerants R1234yf and R1234ze(E). R1234yf had a greater entrainment ratio than R134a and R1234ze(E).

Besana et. al. [15] proposed screening of refrigerants based on an integrated Computational Fluid Dynamic (CFD) - Lumped Parameter Model (LPM) approach. Performances for different refrigerants were obtained through a validated CFD approach, whereas, the cycle was modeled by a Lumped Parameter Model. For Different refrigerants, the energy performances were calculated and the effects of the “component-scale” on the

"system-scale" were investigated. Mishra [16] used HFO refrigerants in the ejector refrigeration system and comparisons with other HFC and HCFC refrigerants were presented. Dwivedi et. al. [17] performed investigation with eco-friendly refrigerants R-404A, R-410A, R-407C, R-423A, R-500, R-502 and R-507C. The maximum first Law Efficiency (η_I) was in the range of 1.5-1.8 with fluid R-404A and at condenser temperature 328K. At 253K evaporator temperature, the Second Law Efficiency (η_{II}) was 37% for R-404A.The highest COP was associated with R-404A with subject temperature variations. He et. al. [18] designed and experimented with neon Cryogenic LHP with infrared point-to-point heat transfer elements in future space applications. The results from the experiments exhibited that the supercritical startup was realized effectively at cases of 1.5 W secondary evaporator power, unsuccessful when 0.5 and 1W heat load was subjected to the second evaporator.

The maximum heat transport capacity of the primary evaporator was in the range of 4.5and 5 W with the proper auxiliary heat load. Hossain et. al. [19] performed an experimental investigation on the loop performance at different heat loads. Thermocouples were installed on LHP to gauge the temperature record at various locations. The minimum thermal resistance of LHP recorded was 0.78 °C/W for a heat load of 100 W, whereas, the maximum was 3.1 °C/W for a heat load of 20 W. The maximum heat transfer coefficient in the evaporator was recorded as 14114 W/m2 for a heat load of 100 W. Setyawan et. al. [20] aimed to manufacture a loop heat pipe uses capillary wick copper sintered with centrifugal casting. The filling ratio was also varied for the investigation FR: 50%, 60%, and 80%, with 60% giving the best performance. Martvoňová et. al. [21] worked on increasing the efficiency of fireplace inserts by preheating the combustion air with the heat from the flue gases through an LHP. Anikivi et. al. [22] executed an experimental investigation and to evaluate the performance of Flat loop heat pipe (LHP) with stainless-steel mesh wick. Different fill ratios such as 40%, 50%, 60% were used subject to different heat loads of 20 W, 40 W, 60 W, 80 W, and 100 W, using De-Ionized water as fluid. Results exhibited that with forced water-cooling conditions, the LHP can transfer a maximum heat load of 100 W.

The evaporator temperature was recorded at 92˚C, with minimum thermal resistance at 0.5332 ˚C/W. Gai et. al. [23] established that the temperature hysteresis of the LHP was connected to the gas-liquid distribution in the compensation chamber (CC). It relied on the interaction between the heat leak of the evaporator and the reflux liquid from the condenser. The temperature of the LHP evaporator increased as the vapor phase in the compensation chamber began to increase. Shukla [24] presented a thermo-fluid dynamic model for the transient operations and analyzed the thermo-fluid dynamic feature of a Stainless Steel/Ammonia. Buz et. al. [25] prepared a mathematical model and performed computational analysis on the processes in the loop heat pipe to avoid auto- oscillation and increasing of the efficiency and temperature. Shioga et. al. [26] conducted an experimental study measuring the temperature of each division of the thin LHP aligned with the subjected heat input evaluating the heat transfer capability. The Thermal resistance between the evaporator and the condenser was found to be 0.11 K/W for horizontal orientation. For a bottom heat orientation, it was0.03 K/W and 0.28 K/W for a top heat orientation at 20W. Nagano et. al. [27] performed a comprehensive test program with start-up, power step-up power cycle, and low power. Korn [28] reported the major working principles and the most important potential to calculate heat. Dwivedi et al. [29] studied the viability of LHP to be used for intra-cycle heat exchange for the refrigeration cycles. Hirasawa et. al. [30] investigated the thermal performance of a loop heat pipe having two evaporators and two condensers applying a lumped network model. The Results exhibited that the vapor and liquid flow rates and the thermal conductance of the heat pipe varied considerably based on the distribution ratio of the heating rate of the evaporators.

GAP IN THE LITERATURE

The review of the previous investigations suggests that there is a huge scope of using both LHP & ERC for utilizing the waste heat from power plants or steel plants. Hence, an industrially viable eco-friendly refrigeration unit can be objectified which benefits the environment and well as the industry in cost savings on multiple fronts. In this research analysis, the combining of GPC to ERC is done through LHP HEx. The limits of the cycle have been fixed along with the selection of working fluids. The objective of this investigation is not to study the GPC, LHP, or ERS for sufficient search that has already been published on the First Law & Second Law analysis of these systems. The sole objective of this research work is to study the industrial feasibility of

such an eco-friendly novel combination of cycles through a software-based equation solver. Moreover, mass flow ratio in the ERS, Pressure of Condenser of ERS, and others have been chosen based on the available literature and standard texts. Peak Temperature of GPC, Pressure of the Boiler of ERS, the mass ratio of ERS have been made input parameters and other performance parameters have been recorded through the equation solver and have been presented.

MATERIALS AND METHODS

As summarized in the previous section, in this research investigation a Gas Power Cycle (GPC) has been combined with an Ejector Refrigeration Cycle (ERC) through a Loop Heat Pipe Heat Ex. In this section, a basic overview of the operation of this cycle has been presented regarding Fig 3. The Topping cycle is the GPC working of the standard working fluid (Air) which has a Reheater &Regenerator in addition to other compulsory parts such as Compressor, Generator & Gas Turbine. The exhaust gas from the regenerator has been routed through the Evaporator of the LHP Ex. back to the Compressor. The heat of the exhaust is absorbed by the LHP fluid while evaporating and the vapors through the wicked structure move to the condenser, from where the liquid generated after vapor releasing heat to the boiler of the ERC returns to the Evaporator. The working fluids for the LHP HEx have been chosen from the easily available fluids such as Water, Ethanol, Methanol, and Acetone. It has been studied how these working fluids affect the performance of the components as well as the combined cycle.

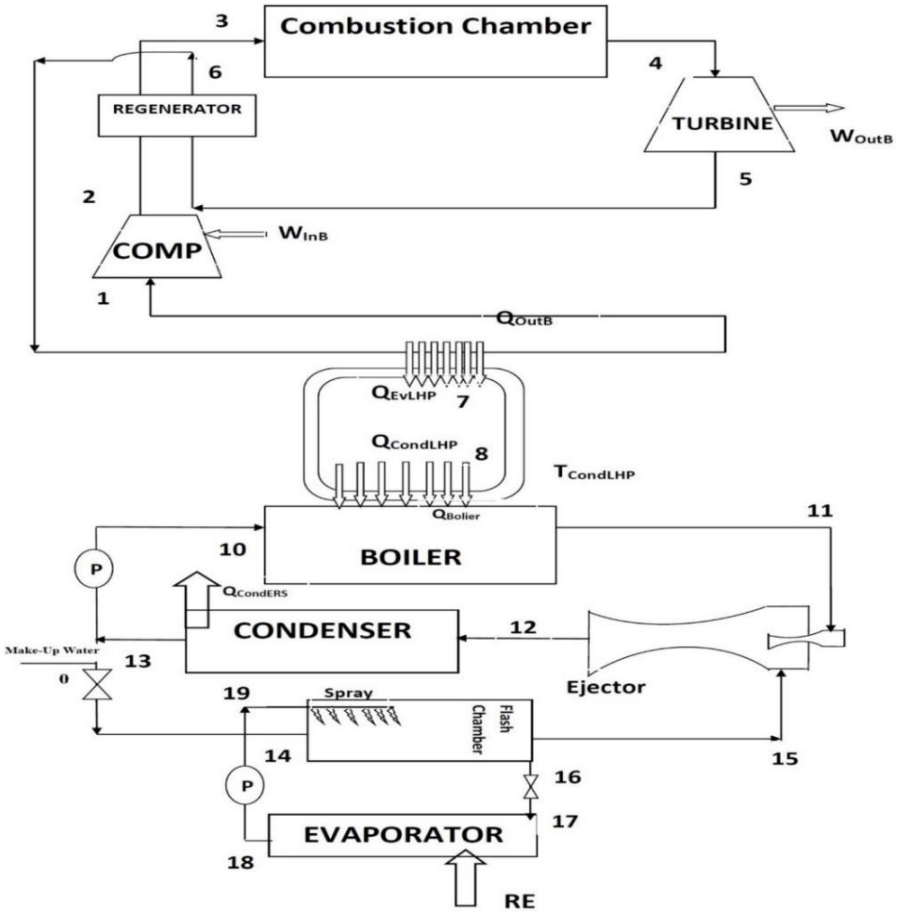


Figure 3. Schematic of Ejector Refrigeration System Combined with Organic Vapour Power Cycle through Loop Heat Pipes

From Figure 3 is can be seen that the vapor generated in the boiler is expanded in the Ejector Nozzle (Fig 3a) at 11, causing the sudden drop in of the pressure at the converging section which causes the liquid inside the flash chamber to evaporate and move into the ejector nozzle at15 resulting into the cooling of the remaining liquid in the flash chamber. This cool liquid at 16 is throttled down to the Evaporator at 17 which maintains the desired low temperature. The fluid coming out of the evaporator at 18 is pumped back to the flash chamber in form of

spray at 19 to cool down the liquid further. The fluid coming out of the mail ejector nozzle at 12 is pumped back to the boiler at 10 through the condenser at 13. Hence, 2 parallel cycles operate in the ERC.05 New Eco-friendly refrigerants namelyR236ea, R1224yd (Z), R1233zd (E), R245fa, R365mfc along with 1 established refrigerant Water (R718) for standardization purpose have been chosen for the investigation so that the objective of environmental protection and conservation has been strengthened. The corresponding T-s plot can be seen in Fig 4.

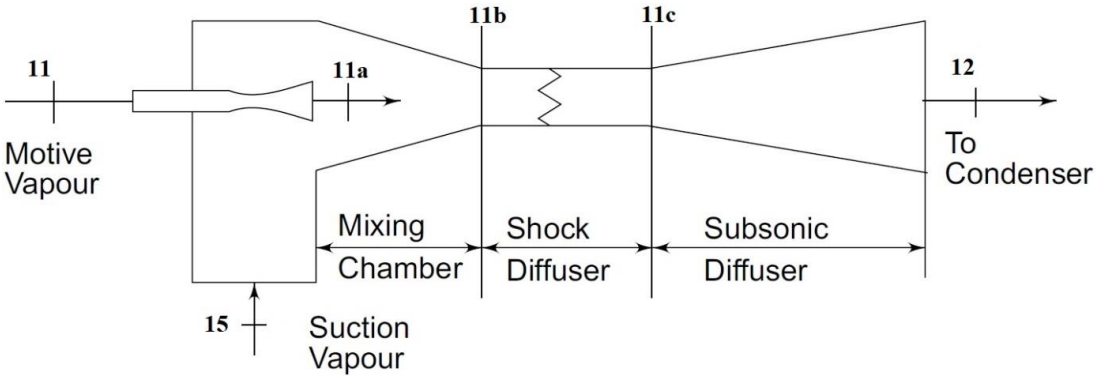


Figure 3a. The ejector of the ERS [32]

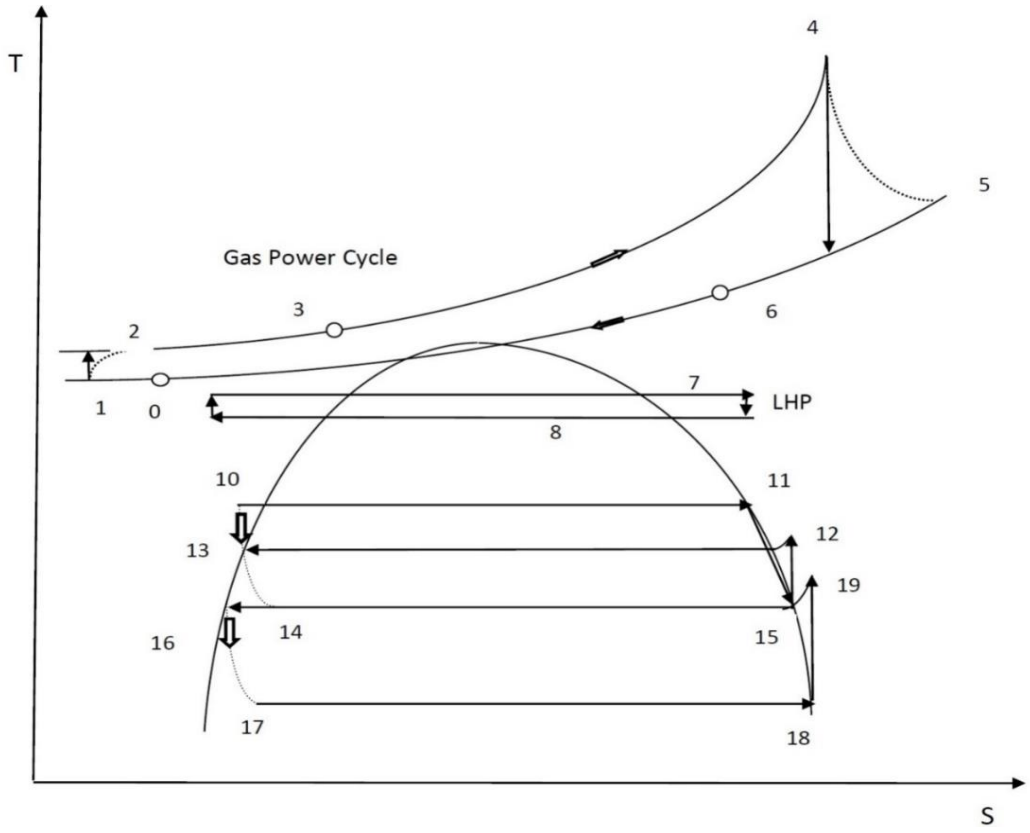


Figure 4. T-s Plot for Gas Power Cycle Combined with Organic Vapour Power Cycle through Loop Heat Pipes.

Tables 1 & 2 contain the 10-working fluids being used in the combined system. Table 1 has all the refrigerants being considered for the ERS whereas table 2 has the working fluids being considered for the LHP Hex. It needs to be noticed that the GWP & GDP for R1224yd (Z) is 0, hence, the most attractive from the list. Moreover, it can be seen that the critical temperatures for the refrigerants fall in the same line whereas for water it is quite high. Hence, it needs to be pointed out that ERS working on these refrigerants can work on even low-temperature heat sources. Hence, the peak temperature of the GPC may be kept even in the range of 500K. The chiller temperature has not been altered keeping because of the operating temperature of the Water.

Table 1. Combination of working fluids for ORC

ORC Working Fluid	R236ea	R1224yd(Z)	R1233zd(E)	R245fa	R365mfc	R718
Critical Temperature (K)	412.4	428.7	438.8	427	460	674.14
Ozone Depletion Potential (ODP)	0	0	0	0		0
Global Warming Potential (GWP)	1200	<1	4.5	1030	794	0

Table 2. Combination of working fluids for LHP-Ex

LHP Working Fluid	Acetone	Ethanol	Methanol	Water
Critical Temperature (K)	508K	516.25	512.6K	674.14

METHODOLOGY AND EQUATIONS USED

The GPC Peak temperature has been varied suitably from 600 K to 1000K. The pressure ratio has been kept at 5 with the assumption of perfect inter-cooling & reheating. The values of the Efficiencies & Effectiveness for the different components of the complete combined cycle such as Compressors, Turbines, Nozzles, Diffusers, etc. have been suitably assumed with the help of referred previous research works and textbooks. The condenser temperature has been kept at 323K; the Evaporator Temperature has been kept at 278K. The mass ratio has been varied from 2 to 7. The Boiler temperature will be dependent on the Condenser temperature of the LHP Ex. The state of Fluid out of the boiler has been assumed as Saturated Vapor. The equations available from texts and literature available have been mathematically modeled into the Engineering Equation Solver to run the desired iterations to obtain the results. The results have been displayed in form of comparative graphs based on the eco-friendly working fluid combinations and input pressure & temperature to the cycle to ascertain the industrial feasibility.

Gas Power Cycle [35]

Isentropic compression the work required for the compressor is given by

$$T_2/T_1 = r^{(\gamma-1/\gamma)} = T_5/T_6 \quad (1)$$

$$T_3 = T_1 + (T_2 - T_1)/\eta_C \& T_7 = T_5 - (T_5-T_6)\,\eta_{T,} \quad (2)$$

Where η_T can be taken as 0.92 &η_C can be taken as 0.86

$$W_{inB} = H_3 - H_1 = \dot{m}_a c_{pa} (T_3-T_1) \quad (3)$$

Isobaric heat addition the net heat added is given by

$$Q_{add} = H_5 - H_4 = \dot{m}_a c_{pa} (T_5-T_4) = \dot{m}_F CV \text{ (Calorific Value of Fuel= 43 MJ)} \quad (4)$$

Isentropic expansion the work done by turbine is given by

$$W_{outB} = H_7 - H_5 = \dot{m}_g c_{pg} (T_7-T_5) \quad (5)$$

$$W_{NetB} = W_{outB} - W_{inB} \quad (6)$$

$$\eta_B = W_{NetB} / Q_{add} \quad (7)$$

Heat Exchanged in the Regenerator

$$Q_R = H_4 - H_3 = \dot{m}_a c_{pa} (T_4-T_3) \quad (8)$$

$$\text{Where, } T_4 = T_3 + \varepsilon_R(T_7-T_3) \quad (9)$$

Isobaric heat rejection (in a heat exchanger), the net heat rejected is given by

$$Q_{Rej} = H_8 - H_1 = \dot{m}_g c_{pg} (T_8-T_1) \quad (10)$$

$$\text{Where } T_8 = T_7-\varepsilon_R(T_7-T_3) \text{ \& } \varepsilon_R \text{ can be taken as 0.75} \quad (11)$$

Table 3. List of parameters & respective abbreviations for GPC

Parameters	**Abbreviations/ Symbols**
Gas Power Cycle	
Inlet Temperature (K) & Pressure (kPa) to Compressor of Gas Power Cycle (GPC)	$\mathbf{T_1, P_1}$
Outlet Temperature (K) & Pressure (kPa) to Compressor of GPC	$\mathbf{T_3, P_2}$
Inlet Temperature (K) & Pressure (kPa) to Combustion Chamber of GPC	$\mathbf{T_4, P_2}$
Inlet Temperature (K) & Pressure (kPa) to Turbine of GPC	$\mathbf{T_5, P_2}$
Inlet Temperature (K) & Pressure (kPa) to Regenerative Heat Exchanger of GPC	$\mathbf{T_7, P_1}$
Inlet Temperature (K) & Pressure (kPa) to Loop Heat Pipe Connecting the Power Cycles	$\mathbf{T_8, P_1}$
Work Input in the GPC (kW)	$\mathbf{W_{inB}}$
Heat Input in the GPC (kW)	$\mathbf{Q_{add}}$
Work Output in the GPC (kW)	$\mathbf{W_{outB}}$
Net-Work Output in the GPC (kW)	$\mathbf{W_{netB}}$
Heat Exchange between GPC & LHP in kW	$\mathbf{Q_{GP\text{-}LHP}}$
Flow Rate of Air in GPC (kg/s)	$\mathbf{\dot{m}_a}$
Flow Rate of Gas in GPC (kg/s)	$\mathbf{\dot{m}_g}$
Pressure Ratio	**r**
Heat Exchange in Regenerator in kW	$\mathbf{Q_R}$
Turbine & Compressor Efficiency	$\mathbf{\eta_T \& \eta_B}$

Loop Heat Pipes [31]

$$Q_{GP\text{-}LHP} = m_g c_{Pg}(T_8 - T_1) \quad (12)$$

$$\dot{E}_{GP\text{-}LHP} = m_g (h_8 - h_1) - T_o\, m_g (s_8 - s_1) \quad (13)$$

$$\text{Where, } s_8 - s_1 = c_{Pg} (T_8 - T_1) - R_g \ln(P_8/P_1) \quad (14)$$

Effectiveness of LHP Evap-GPC Heat Exchanger $\boldsymbol{\varepsilon_{LHPEv\text{-}GPC}} = 0.75$

Effectiveness of LHP Cond-ORC Heat Exchanger $\boldsymbol{\varepsilon_{LHPCond\text{-}ORC}} = 0.85$

Figure of Merit for comparing different working fluids for LHP:

$$MF_h = \rho_l \sigma h_{fg} / \mu_l \quad (15)$$

The axial heat flow rate due to the sonic limitation was calculated from the following equation

$$Q_S = \pi\, r_p^{\,2} \rho_v h_{fg} \sqrt{\frac{\gamma RT}{2(\gamma+1)}} \quad (16)$$

The maximum heat transfer due to the entrainment limit was determined using the equation

$$Q_{Ent} = \pi\, r_p^{\,2} h_{fg} \sqrt{(2\, \pi\, \rho_v \sigma \cos\phi / \lambda)} \quad (17)$$

The maximum heat transfer due to the capillary limit was determined using the equation

$$Q_{Cap} = (MF_h)(A_w k_{hp}/L_{eff}) \{(2/r_p) - (\rho_l g L_{eff} \sin\phi / \sigma)\} \quad (18)$$

The degree of superheat to cause nucleation is given by:

$$\Delta T = 3.06\, \sigma_l T_{vLHP} / \rho_v h_{fg} \delta \quad (19)$$

With reference to mass flow rate of warm and cool air, the heat transfer rates to the evaporator and condenser sections are calculated as follow:

$$q_e = mc_p(T_{e,in} - T_{e,out}) = \rho\, u\, Ac_p\, (T_{e,in} - T_{e,out.}) \quad (20)$$

$q_c = mc_p(T_{c,in} - T_{c,out}) = \rho \, u \, Ac_p \, (T_{c,in} - T_{c,out.})$ (21)

$HTF = 2MF_h Ak_{hp}/ r_p$ (22)

Table 4. List of parameters & respective abbreviations for LHP-Ex

Parameters	**Abbreviations/ Symbols**
Loop Heat Pipe HEx.	
Evaporator Temperature LHP Connecting the Power Cycles (K)	$\mathbf{T_{EvLHP}}$
Condenser Temperature LHP Connecting the Power Cycles (K)	$\mathbf{T_{CondLHP}}$
Heat Absorbed in the Evaporator of the LHP Connecting the Power Cycles (kW)	$\mathbf{Q_{EvLHP}}$
Heat Transported to the Condenser of the LHP Connecting the Power Cycles (kW)	$\mathbf{Q_{CondLHP}}$
Heat Input to the Boiler of the ORC Through the LHP Connecting the Power Cycles (kW)	$\mathbf{Q_{inLR}}$
The surface tension N/m	$\boldsymbol{\sigma}$
The latent heat of vaporization at heat pipe temperature in kJ/kg	$\mathbf{h_{fg}}$
The viscosity of the liquid in Pa-s.	$\boldsymbol{\mu_l}$
The Figure of Merit in kW/m^2	$\mathbf{MF_h}$
The X-sectional area of wick in m^2	$\mathbf{A_w}$
The Permeability of the wick material in m^2	$\mathbf{k_{hp}}$
The effective length of the heat pipe in m	$\mathbf{L_{eff}}$
The pore radius in m	$\mathbf{r_p}$
The acceleration due to gravity in m/s^2	$\mathbf{g}$
The inclination angle in degrees	$\boldsymbol{\Phi}$
The density of vapor in heat pipe in kg/m^3	$\boldsymbol{\rho_l}$
Mass Flow Rate in the Evaporator & Condenser of LHP in kg/s	$\dot{\mathbf{m}}_{\mathbf{LHP}}$
Heat Exchange in the Evaporator& Condenser of LHP Respectively in kW	$\mathbf{Q_{EvLHP}}$ $\mathbf{\&Q_{CondLHP}}$
Enthalpy inlet & outlet of the Evaporator of LHP kJ/kg	$\mathbf{h_{EviLHP}\&}$ $\mathbf{h_{EvoLHP}}$
Enthalpy inlet & outlet of the Condenser of LHP kJ/kg	$\mathbf{h_{CondiLHP}}$ $\mathbf{\&h_{CondoLHP}}$
Mean Temperature Diff. in the Evaporator & Condenser of LHP Respectively in K	$\mathbf{T_{H\text{-}HP}}$ $\mathbf{\&\ T_{L\text{-}HP}}$
The temperature of the vapor in LHP in K	$\mathbf{T_{vLHP}}$
The characteristic dimension of the liquid/vapor interface (m)	$\boldsymbol{\lambda}$
Thickness thermal layer (m)	$\boldsymbol{\delta}$
Heat Transfer in LHP as per Sonic Limitation in kW	$\mathbf{Q_S}$
Heat Transfer in LHP as per Entrainment Limit Limitation in kW	$\mathbf{Q_{Ent}}$
Heat Transfer in LHP as per CapillaryLimit Limitation in kW	$\mathbf{Q_{Cap}}$
X-sectional area of the duct of LHP (m^2)	$\mathbf{A}$
Permeability of the Wick Material	$\mathbf{k_{hp}}$

Ejector Refrigeration Cycle [32]

The Jet Velocity at the exit of the Nozzle,

$C_{11a} = \sqrt{[2\eta_N(h_{11}-h_{12})]}$, where η_N is nozzle efficiency 0.85. (23)

Continuity Equation at the Mixing Section $m_{11} + m_{15} = m_{11bc}$, (24)

Where; point 11_{bc} shows the throat right before the shock.

Momentum Equation at the Mixing Section $(m_{11} + m_{15})\, C_{11b} = m_{11}\, C_{11a} - m_{15}\, C_{15}$ (25)

Velocity& Mass Flow at the Suction of vapor from flash chamber m_{15}&C_{15} are considered to be negligible.

$m_{11bc} C_{11b} = m_{11}\, C_{11a}$ (26)

Energy Equation of Mixing Process

$m_{11} h_{11} = m_{11bc}\, (h_{11b} + c_{11a}^2/2)$ (27)

Continuity Equation for Shock Diffuser Section

$m_{11bc} = C_{11b}\, A/v_{11b} = C_{11c} A/v_{11c}$, where v is the specific volume. (28)

Momentum equation at shock diffuser section:

$(p_{11c} - p_{11b})\, A = (C_{11b} - C_{11c})\, m_{11bc}$ (29)

Energy equation at shock diffuser section:

$h_{11b} + C_{11b}^2 / 2 = h_{11c} + C_{11c}^2 / 2$ (30)

Energy equation at shock diffuser section:

$h_{12} - h_{11c} = C_{11c}^2 / 2 = h_{12s} - h_{11c} / \eta_D$, where $\eta_D = 0.8$ is the diffuser efficiency (31)

Neglecting the shock waves,

$\eta_E m_{11}\, C_{12}^2 / 2 = m_{11bc}\, C_{11bc}^2 / 2$, where $\eta_E = 0.65$ is the entrainment efficiency (32)

Mass Ratio, $m_R = m_{11}/m_{15} = h_{12s} - h_{11bc} / [\eta_N \eta_D \eta_E (h_{11} - h_{11bc}) - (h_{12s} - h_{11bc})]$ (33)

$\eta_N \eta_D \eta_E$= Overall Efficiency of Ejector, η_O (34)

Refrigeration Effect, $RE = h_{18} - h_{17}$ (35)

Heat Supplied, $Q_{Sup} = h_{11} - h_{10}$ (36)

$COP = RE/Q_{Sup}$ (37)

Table 5. List of parameters & respective abbreviations for ORC

Parameters	**Abbreviations/ Symbols**
Ejector Refrigeration Cycle	
Temperature (K) & Pressure (kPa) at Different States	**T, P**
Enthalpy in kW & kJ/kg respectively	**H & h**
The temperature of the Environment in K	$\mathbf{T_o}$
Nozzle, Diffuser, Entrainment & Overall Efficiency	$\boldsymbol{\eta_N, \eta_D, \eta_E \& \eta_o}$
Coefficient of Performance	**COP**
Refrigerating Effect kW	**RE**
Heat Supplied in the Boiler kW	$\mathbf{Q_{Sup}}$
Ejector Area Ratios	$\mathbf{A_R}$

Mass at Different States kg/s	**m**
Velocity at different State's m/s	**C**
Dryness Factor at Ejector Exit	$\mathbf{x_{12}}$
Ratio of Mass Flow in Boiler to Evaporator	$\mathbf{m_{Ratio}}$

Table 6 shows the recommended and non-recommended wick material for the LHP-Evaporator with the corresponding working fluids. In this research work Copper, Silica or Stainless Steel can be invariably used based on thermal effectiveness & cost-effectiveness.

Table 6. Heat Pipe working fluids & material compatibility [31]

Working Fluid	**Recommended Material**	**Not Recommended Material**
Acetone	Copper, Silica, Aluminium, Stainless Steel	
Methanol	Copper, Silica, Stainless Steel	Aluminium
Water	Copper, Silica, Aluminium, Stainless Steel, Nickel, Carbon Steel	Silica, Aluminium, Stainless Steel, Nickel, Carbon Steel, Inconel
Ethanol	NA	NA

Similar to the table above, Table 7 presents the suitable operating temperature range, based on which the working fluid can be decided for the LHP. In this research work, the operating range is about 400-450 K. Hence, the corresponding working fluids i.e., Ethanol, Acetone, Methanol &Water have been chosen. It can be observed that the LHP can be used to extract heat from relatively low temperatures as well as to run the bottoming cycle such as chosen in this investigation.

Table 7. Heat Pipe working fluids useful operating temperatures [31]

Working Fluid	**Melting Point (ºC)**	**Boiling Point at ATP (ºC)**	**Useful Range (ºC)**
Acetone	-95	57	0-120
Methanol	-98	64	10-130
Ethanol	-112	78	0-130
Water	0	100	30-200

The next section discusses the results obtained from the simulations through various plots and data obtained.

RESULTS AND DISCUSSIONS

The results obtained from the have been presented in this section. As input parameters the Boiler Pressure (P_{11}) of ERS has been varied from 500 kPa to 800 kPa, Peak Temperature of the GPC (T_5) from 600 K to 1000 K, 6 Eco-friendly Working Fluids for ERS, and 4 Working Fluids of the LHP Ex. have been used for the investigation. This section has been separated into 3 sub-sections which present the results based on T_5 (subsection-3.1), P_{11} (subsection-3.2) &Working Fluids in the LHP (subsection-3.3). These results have been presented for the different working fluids being used in the ERS. In sub-section 3.3, results based on 3 input variables namely ERC working fluid-LHP working fluid-P_{11} and ERC working fluid-LHP working fluid-T_5 respectively have been presented. The results have been emphasized for R718 for being the widely available and origin of Steam Jet Refrigeration System as well as all other ancient refrigeration systems and for R1224yd (Z) as it has 0 ODP & 0 GWP and is the most eco-friendly refrigerant of all the chosen refrigerants. By the end of this section, we will be able to assess the feasibility and best operating parameters for such a combined system. The dotted – discontinuous lines/curves may be neglected for they are mere projections.

The Peak Temperature Of Gpc (T_5)

As the T_5 has been varied from 600K to 1000K, the effect of temperature available for heat transfer has been studied. In fig 5 it can be observed that the Temperature available in the boiler of ERC varies from 380 K to 455 K. It can be reiterated that the critical temperatures of the Eco-friendly refrigerants chosen for the study are in the range of 380K – 450K as well. Hence, the boiler can be operated at high temperatures near critical points low heat consumption for the cycle to work.

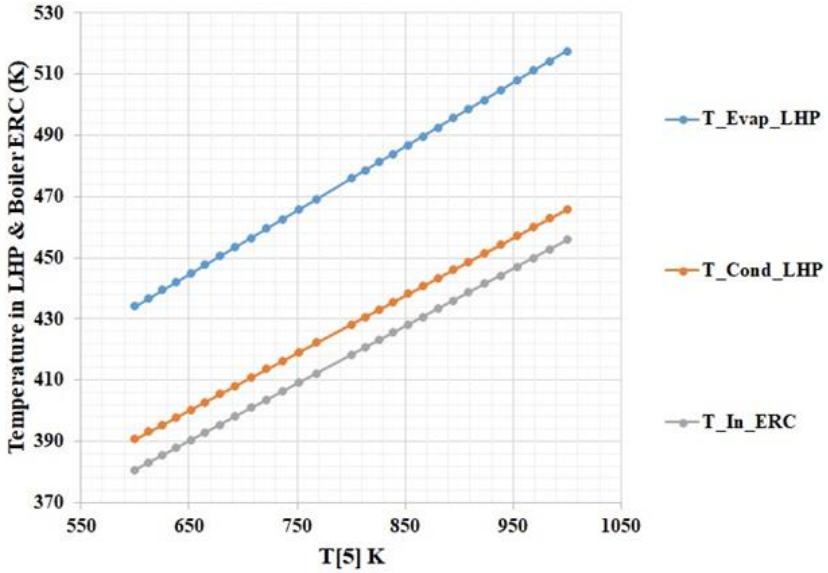


Figure 5. Temperature Profiles of the Components of Combined Cycle with T_5

In Fig 6 refrigeration capacity obtained for the temperature range has been presented. It can be seen that owing to the critical temperature limits, all the ERC working fluids except R718 have been studied for the range of 600K to 800K and 800K -1000K for R718. It can be seen R365mfc (i.e., 8.601kW to 10.35kW) &R1233zd (E) (i.e., 7.657kW to 11.05kW) provide consistent high values of the refrigeration capacities. R245fa &R1224yd (Z) (7.87kW-9.903kW) have coinciding values. With the temperature rise, it can be observed that the Refrigeration capacity also increases. For R718 the rise is the steepest from 6.835kW to 31.545kW.

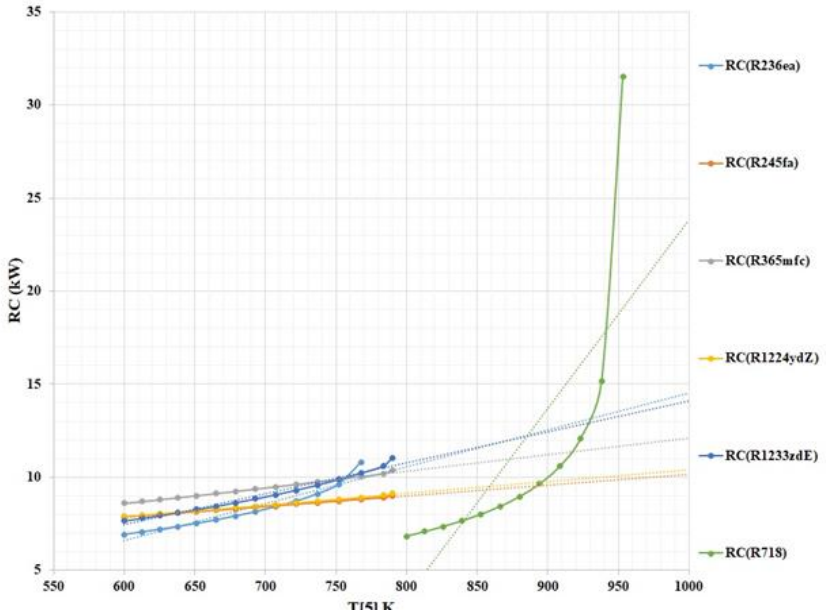


Figure 6. Refrigeration Capacity of the ERS with T_5

The COP can be seen to be decreasing with the increase in temperature Fig 7. For R365mfc the value of COP has been noted to be the maximum from 0.2856 to 0.249. Whereas, for R718 the COP is more consistent and is in the range of 0.285 to 0.282. The main drawback of R718 being not able to run below 278 K. For R1224yd (Z) the value of COP is close and parallel with R245fa in the range of 0.241 to 0.225.

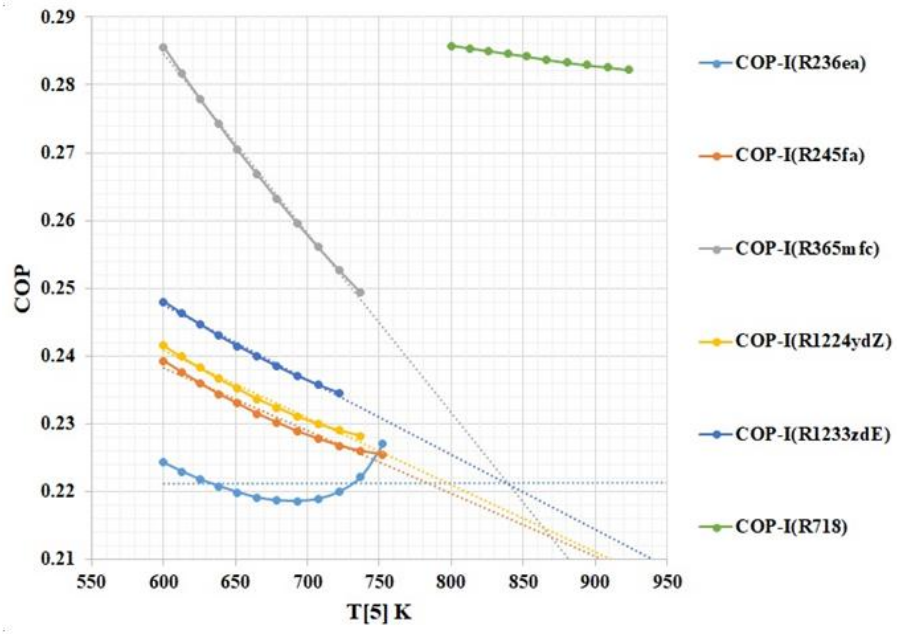


Figure 7. COP of the ERS with T_5

The Boiler Pressure of ERS (P_{11})

Refrigeration capacity and COP have also been presented with the varying Boiler Pressure in Fig 8 & 9 respectively. Unlike the previous sub-section, all the refrigerants have been studied for the same range of pressure 500kPa to 800kPa. A continuous and gradual increase can be seen in the refrigeration capacity with the increase in pressure.

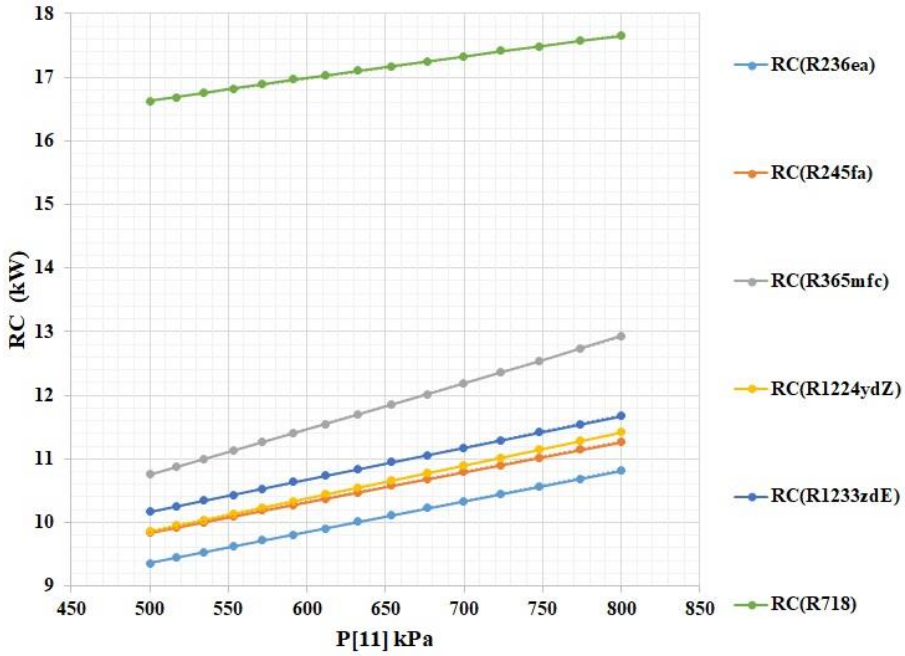


Figure 8. Refrigeration Capacity of the ERS with P_{11}

It can be seen in Figure 8 that for R718 the refrigeration capacity has been recorded as a maximum of 16.62 to 17.65 kW, whereas, for the refrigerants, R365mfc provides a cooling effect of 10.75 to 12.93 kW under the subject conditions. It must be mentioned that the R1224yd (Z) which has 0 GWP & 0 ODP in a system has recorded RC of 9.86 to 11.41 kW.

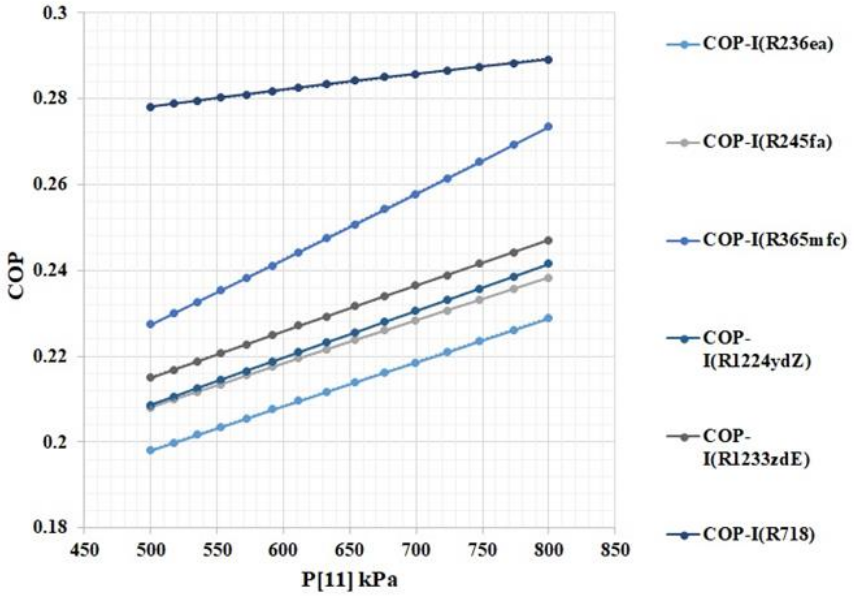


Figure 9. COP of the ERS with P_{11}

Fig 9 presents the data recorded for COP under the variable Boiler pressure. The COP can be seen gradually increasing with an increase in boiler pressure. R718, R365mfc & R1224yd (Z) have maximum COP of 0.289, 0.275& 0.241 respectively.

Working Fluids In The Lhp

Four easily available and eco-friendly working fluids have been selected for the LHP Ex. As mentioned in Section 2, the working fluids have defined working temperatures as well as suitable materials based on the corresponding thermo-physical properties. This section has focussed on the mass flow rate of the working fluids, availability of Heat at the inlet of LHP & ERC for all the working fluids subject to variable Peak Temperature of GPC, Peak Pressure of ERC, and different working fluids. Smaller the mass flow handling, Compacter the whole combined system. It can be seen that mass flow rate and Heat Input for LHP for all the refrigerants except R718 and R718 it's different. Moreover, as mentioned above the primary focus of this section is on R718 & R1224yd (Z), which has the potential of providing the most compatible eco-friendly system subject to the variable Temperature & Pressure

Acetone

In Figure 10 to Figure 13 the results based on Acetone as LHP working fluid have been presented based on variable Pressure & Temperature. The mass flow rate of refrigerants in the ERC is maximum for R1224yd (Z) (1.041kg/s) and least for R718 (0.097kg/s). The mass flow rate in the LHP for the R718 system is 0.639kg/s, whereas, for other refrigerants, it was recorded as 0.4635kg/s.

In Figure 11 heat inputs can be seen to be constant with the varying Pressure. Input to the LHP can be seen to be carried away for transfer with R718 system is at 197.1 kW & for other refrigerants 155.9 kW. Whereas, the heat input to the ERC for R718 is 160.1kW & for other eco-friendly refrigerants is 142kW.

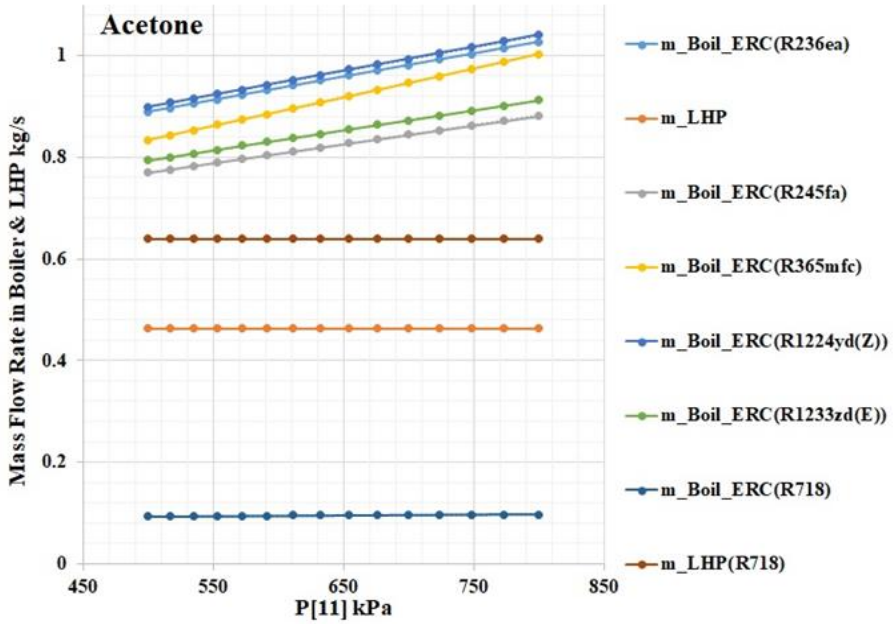


Figure 10. Mass Flow Rate of ERS & LHP with P_{11} for Acetone

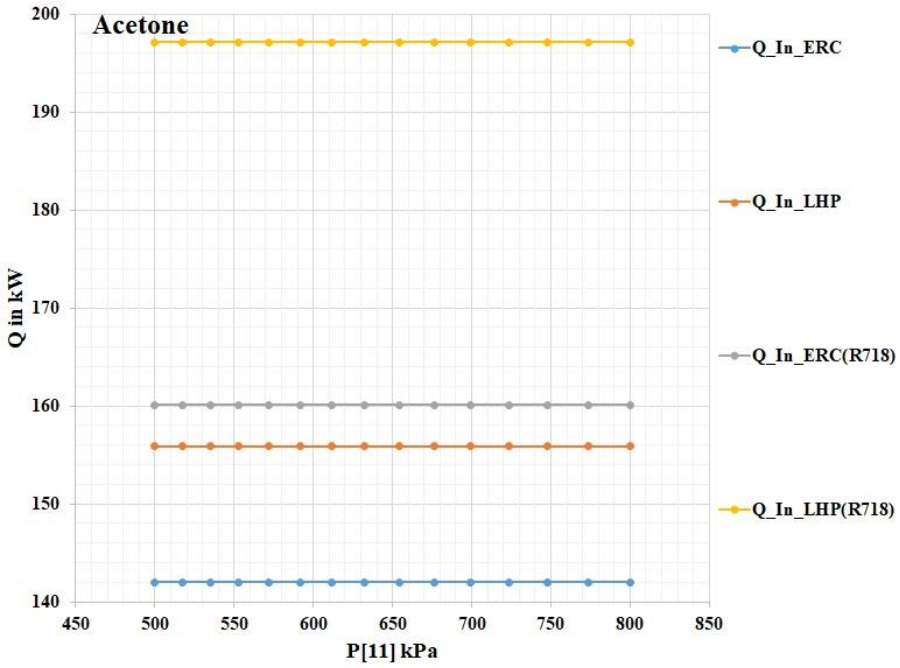


Figure 11. Heat Available as Input to ERS & LHP with P_{11} for Acetone

The results for the analysis of Acetone based on Peak Temperature of GPC Fig 12 & 13. From Fig 12 the maximum range of mass flow rate in ERC has been recorded as 0.821-1.283 kg/s for R236ea and for R1224yz (D) it has been recorded as 0.85-1.245 kg/s. For R718 the range of mass flow rate in ERC is 0.096-0.17 kg/s. The maximum mass flow rate range inside the LHP is recorded for R365mfc at 0.342-0.439 kg/s and for R1224yz (D) at 0.3401-0.498 kg/s. It can also be seen that for R 718 the LHP range of mass flow rate is 0.6396 - 1.321 kg/s.

The heat input range at ERC for R718 can be seen from Fig 13 as 160.1- 182.6 kW & for LHP it has been recorded as 197.1- 353.1 kW. Whereas, for other refrigerants, the range has been recorded as 123.8- 154.3 kW &127.2- 181.5 kW respectively for ERC & LHP.

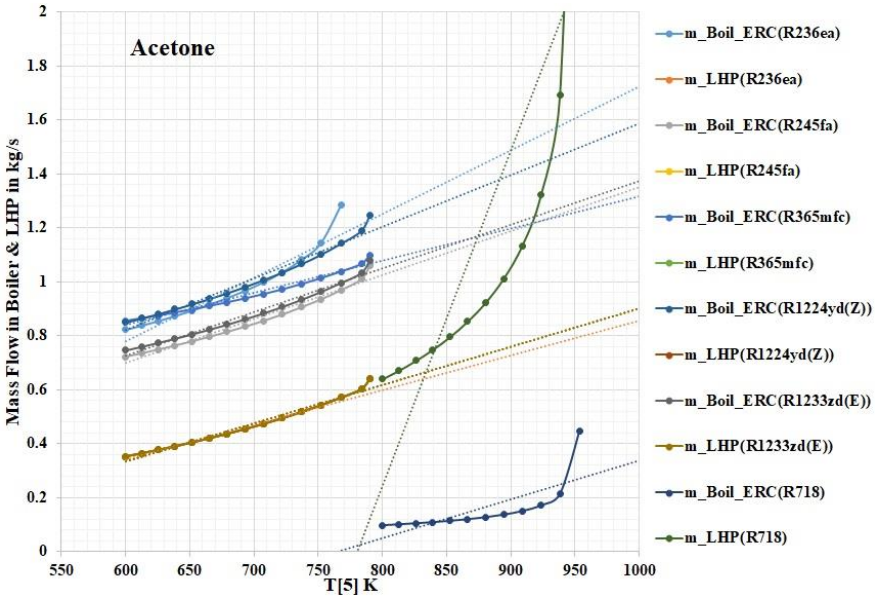


Figure 12. Mass Flow Rate of ERS & LHP with T_5 for Acetone

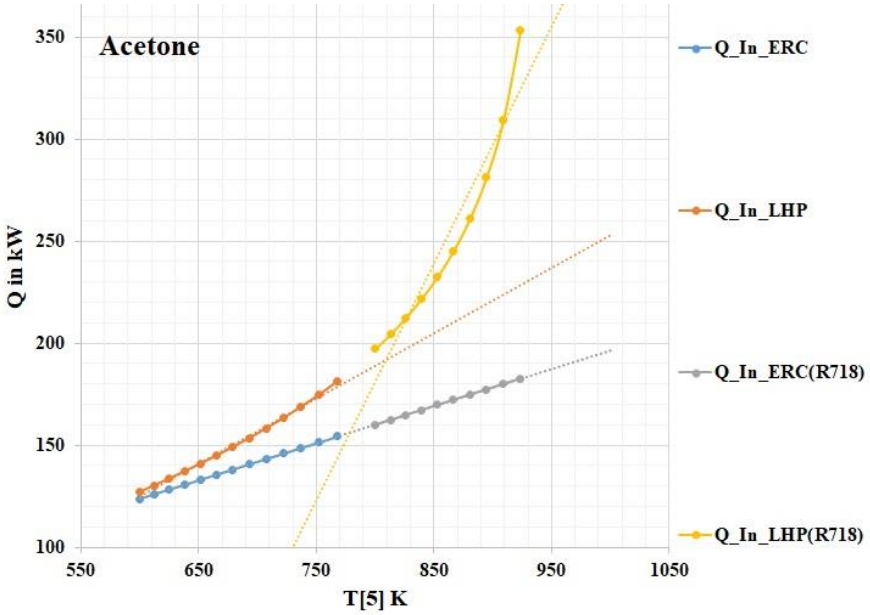


Figure 13. Heat Available as Input to ERS & LHP with T_5 for Acetone

ETHANOL

In Figures 14 & 15 the results of the analysis of Ethanol with variable pressure have been presented. It can be observed from Fig 14 that the maximum mass flow in ERC is for R1224yz(D) 1.01 kg/s and the minimum is for R718 at 0.096 kg/s. Maximum mass flow in LHP for refrigerants is0.244 kg/s and for R718 is at 0.342 kg/s.

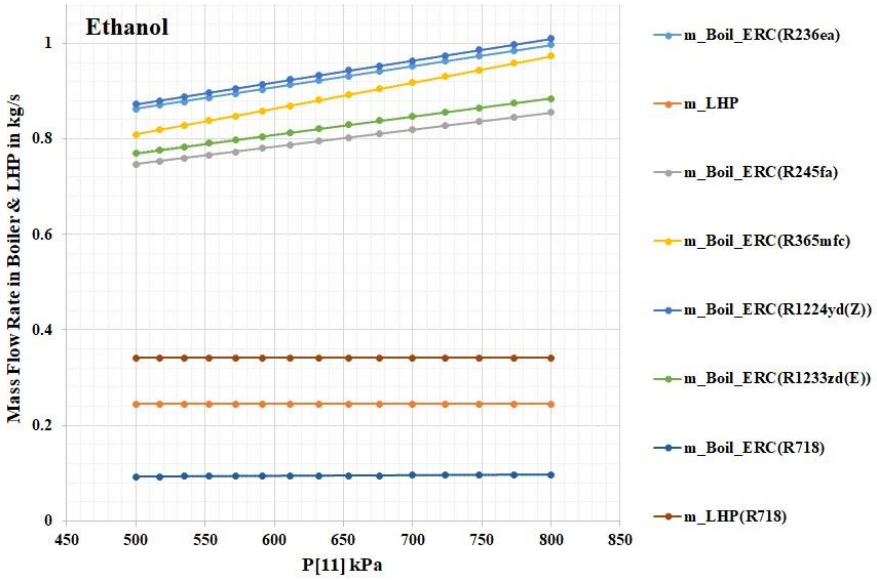


Figure 14. Mass Flow Rate of ERS & LHP with P_{11} for Ethanol

From Figure 15 it can be observed that heat input available at ERC & LHP for refrigerants are 142 kW &155.9 kW. For R718 heat input available at ERC & LHP are 160.1kW & 194.4 kW.

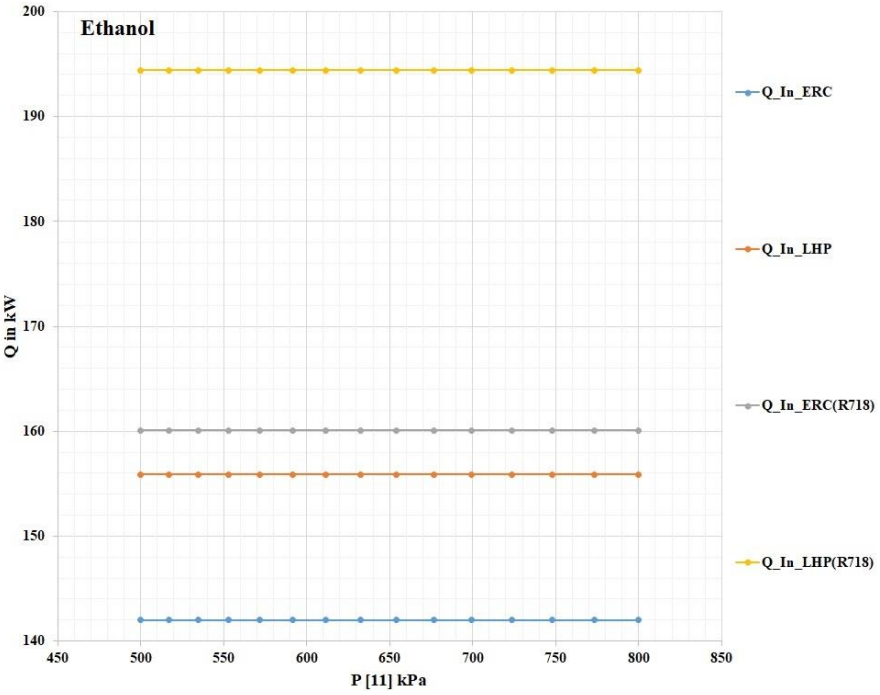


Figure 15: Heat Available as Input to ERS & LHP with P_{11} for Ethanol

Similarly, the results of analysis based on the Peak Temperature have been reported in Figure 16 & 17. The range of Maximum mass flow rate in ERC from Fig 16 is for R1224yz(D) at 0.8321- 1.178 kg/s. The range of mass flow rate for LHP for R1224yz (D) is 0.1893 - 0.3261 kg/s. The range of mass flow rate in ERC & LHP for R718 can be observed as 0.09514 - 0.1877 kg/s &0.3419 - 0.7967 kg/s respectively.

Furthermore, from Figure 17 it has been studies that the Heat input range available for ERC & LHP for refrigerants is 123.8 - 154.3 kW &124.5 - 173.7 kW. Moreover, for R718 it has been observed for ERC & LHP is 160.1 - 185.4 kW &194.4 - 388.9 kW

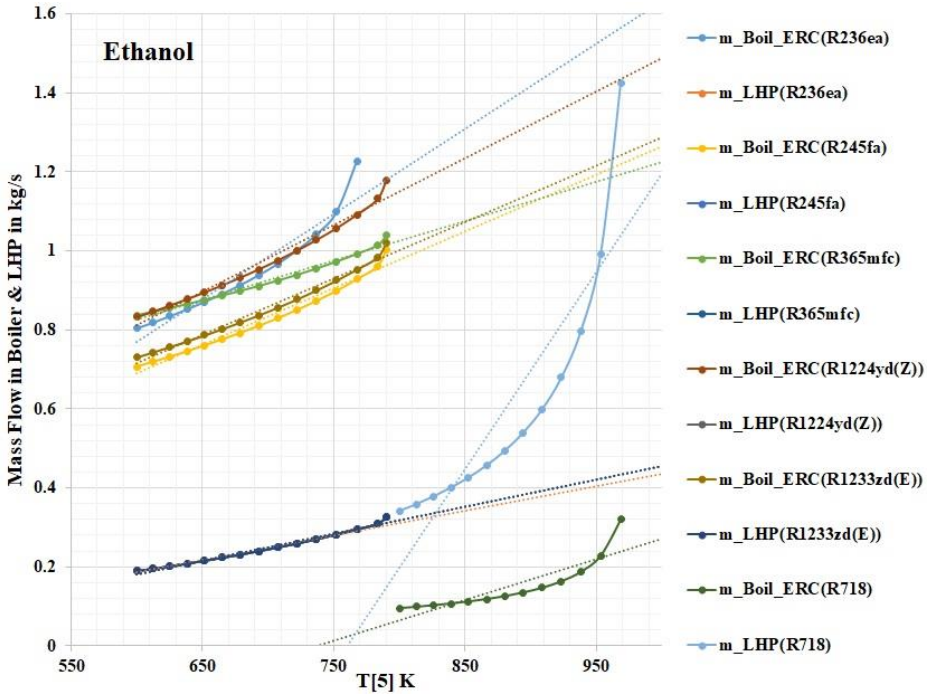


Figure 16. Mass Flow Rate of ERS & LHP with T_5 for Ethanol

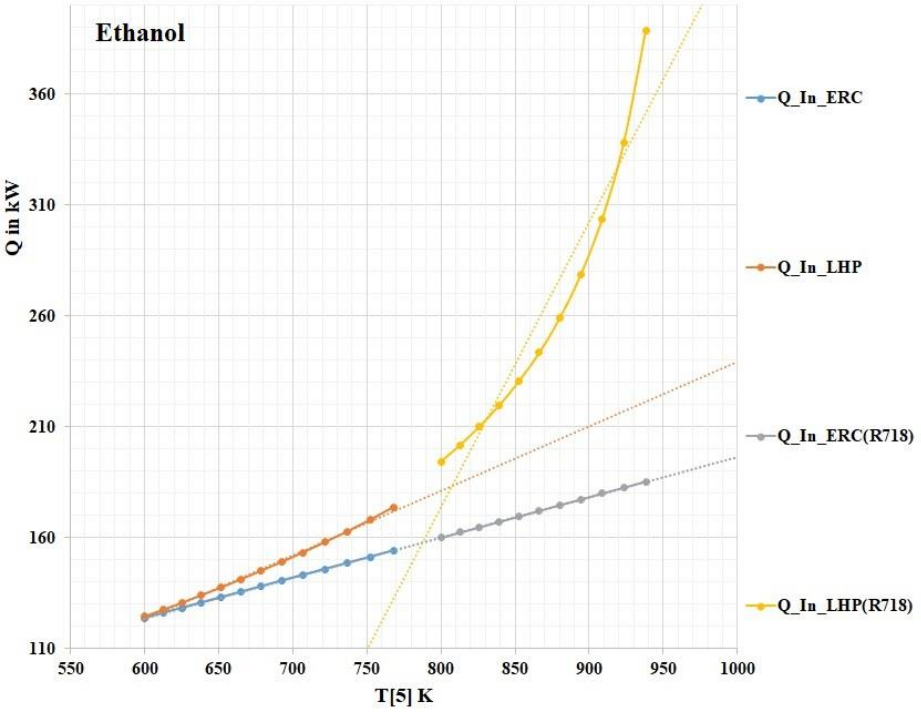


Figure 17. Heat Available as Input to ERS & LHP with T_5 for Ethanol

METHANOL

For methanol as well Maximum Mass flow for variable boiler pressure can be observed for R1224yz(D) as 1.023 Kg/s from Figure 18. For R718 the maximum mass flow rate is 0.09247 kg/s. The maximum mass flow rate for all the refrigerants has been recorded as 0.1968 kg/s and 0.2618 kg/s for R718.

From Figure 19 it has been studied that the Heat Input available for refrigerants at ERC & LHP is 142 kW & 153.2 kW respectively. Similarly, for R718 Heat Input available is 160.1 kW & 186.7 kW.

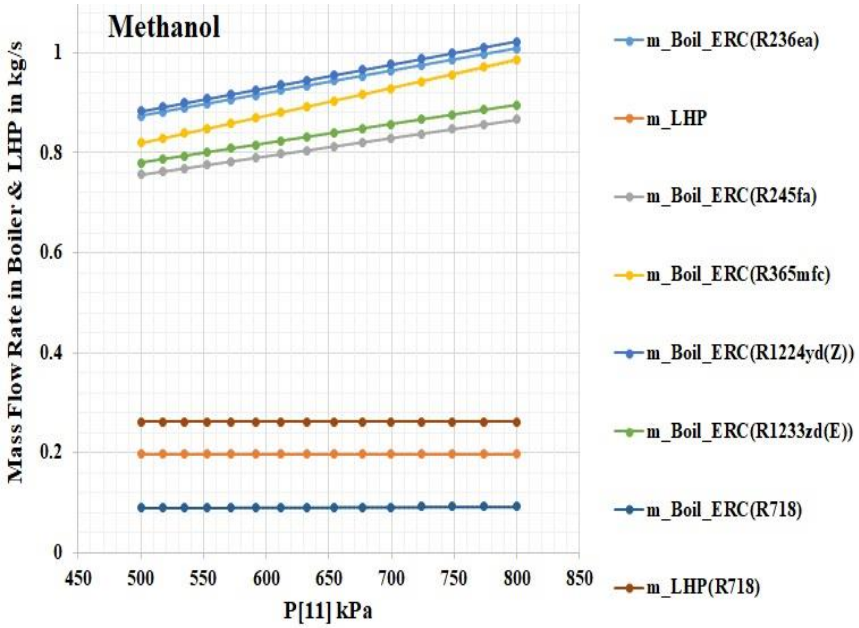


Figure 18. Mass Flow Rate of ERS & LHP with P_{11} for Methanol

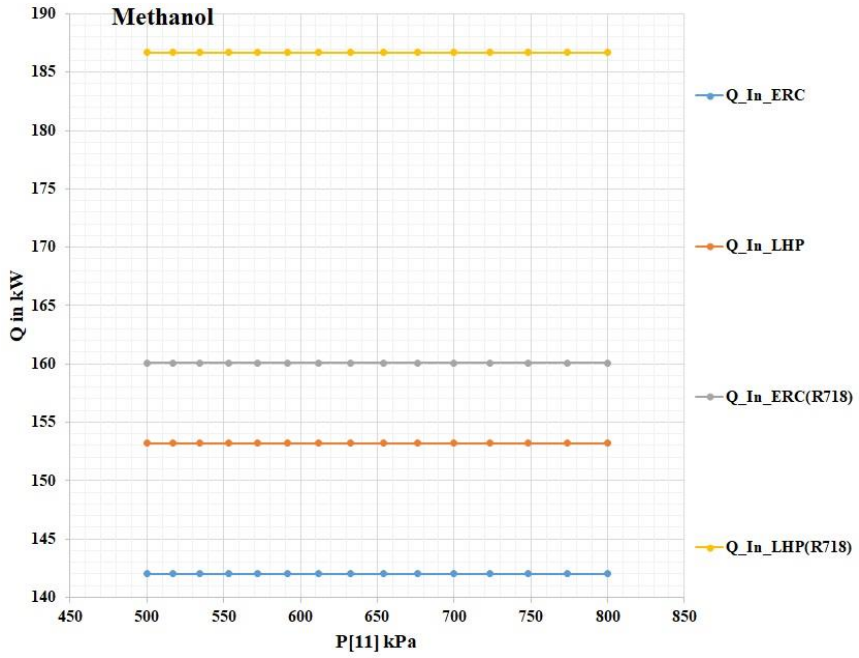


Figure 19. Heat Available as Input to ERS & LHP with P_{11} for Methanol

For the study based on the Peak GPC Temperature Fig 20 & 21 have been presented. It has been observed from Fig 20 that the range for maximum mass flow rate in ERC is for R236ea as 0.8109 - 1.239 kg/s. For R1224yz (D) the range has been observed to be 0.8389 - 1.186 kg/s. Moreover, mass flow rate range for LHP for R236ea & R1224yz (D) is 0.1512 - 0.2375 kg/s &0.1512 - 0.2618 kg/s respectively. For R718 mass flow rate range for ERC & LHP are 0.09139- 0.1621 kg/s & 0.2618 - 0.5451 kg/s respectively.

The ranges of Input Heat available for ERC & LHP for Refrigerants are 123.8 - 154.3 kW &125.5 - 175.3 kW respectively. Moreover, for the R718 range of Heat Input Available for ERC & LHP are 160.1- 185.4 kW &186.7 – 336 kW respectively.

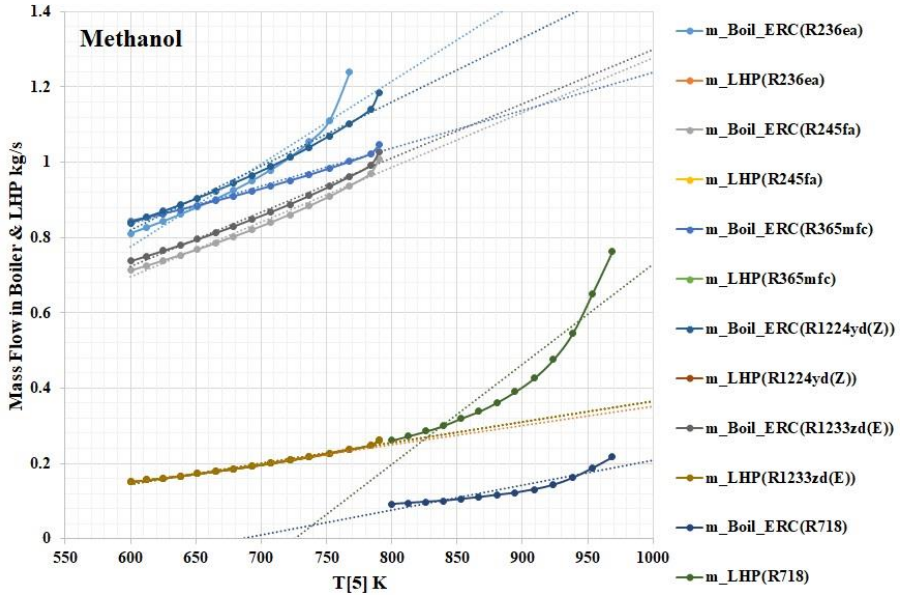


Figure 20. Mass Flow Rate of ERS & LHP with T_5 for Methanol

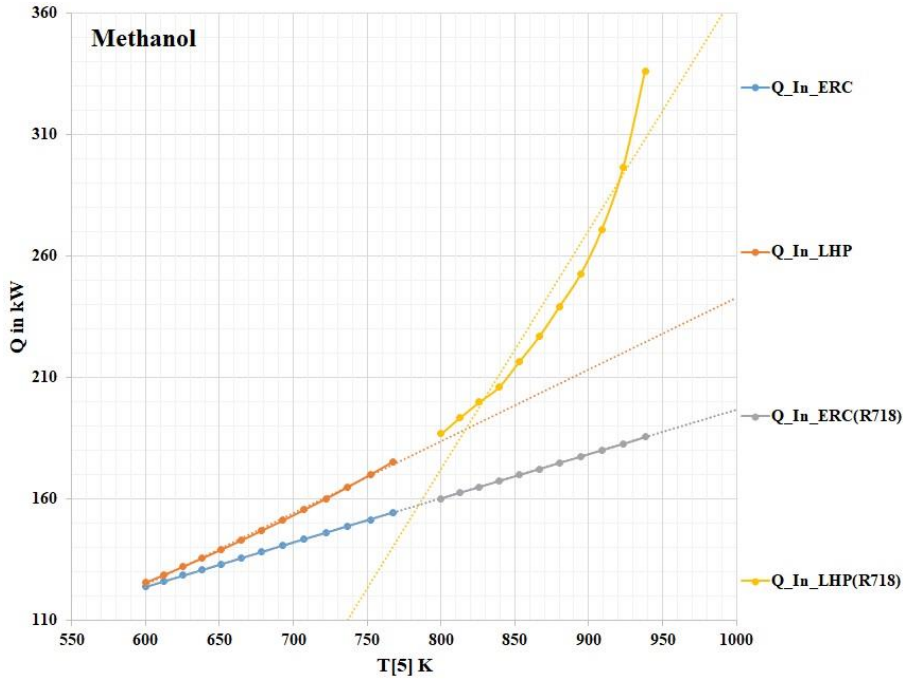


Figure 21. Heat Available as Input to ERS & LHP with T_5 for Methanol

WATER

In Figure 22 & 23 results for water in LHP have been presented for variable Boiler Pressure. In Fig 22, the Maximum mass flow rate in ERC has been required for R1224yz (D) as 0.8651 kg/s whereas, for R718 the value of the same rounds up to 0.07329 kg/s. Furthermore, the mass flow rate for the LHP for R1224yz (D) &R718 is 0.07071 kg/s &0.08299 kg/s.

Heat Input available at ERC & LHP for refrigerants is 129.5 kW &142 kW, while, the heat available at ERC % LHP for R718 is 148 kW &160.1 kW.

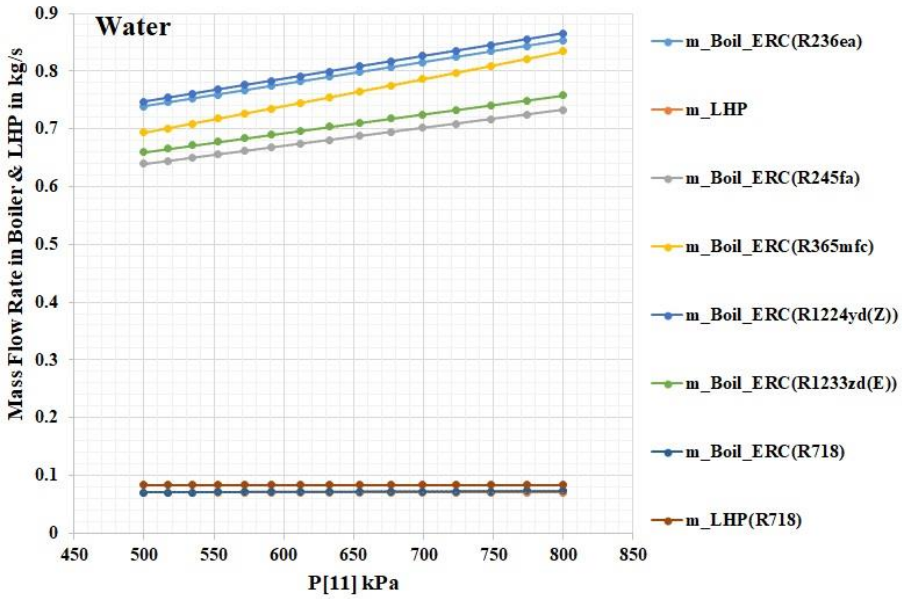


Figure 22. Mass Flow Rate of ERS & LHP with P_{11} for Water

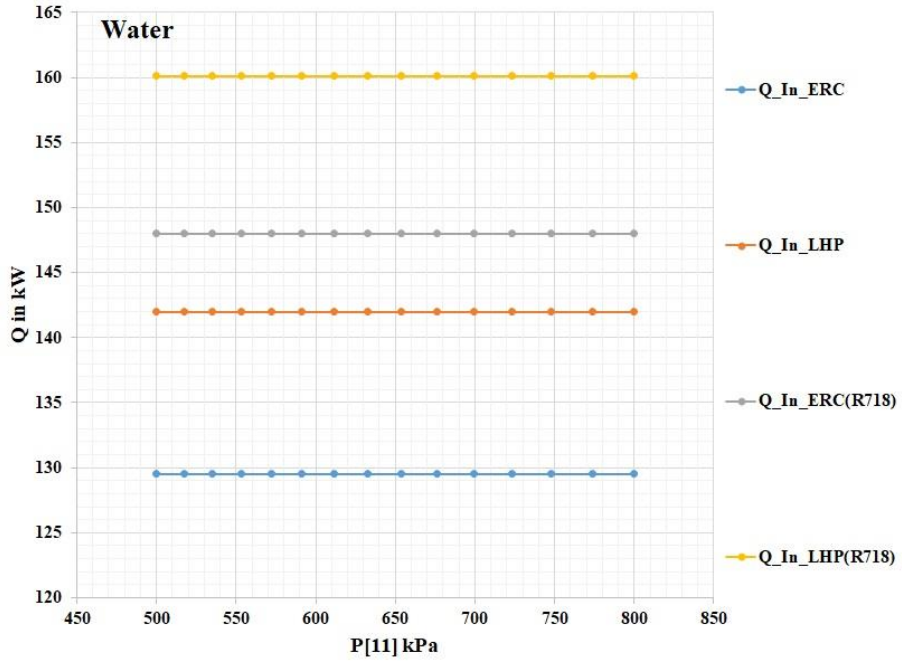


Figure 23. Heat Available as Input to ERS & LHP with P_{11} for Water

Figure 24 & 25 present the study with variable temperature. In the Fig 24, the Range of mass flow rate in ERC & LHP for R1224yz(D) is0.7475 - 0.9348 kg/s &0.05957 - 0.08301kg/s respectively, whereas, for R718 the range of mass flow rate 0.07244 - 0.0906 kg/s &0.08299 - 0.1128 kg/s respectively.

From Figure 23, the heat input range available at the ERC & LHP for the refrigerants is 111.9 - 142 kW &123.8 - 154.3 kW, whereas, for R718 it is 148 - 188.7 kW &160.1 - 196.7 kW respectively.

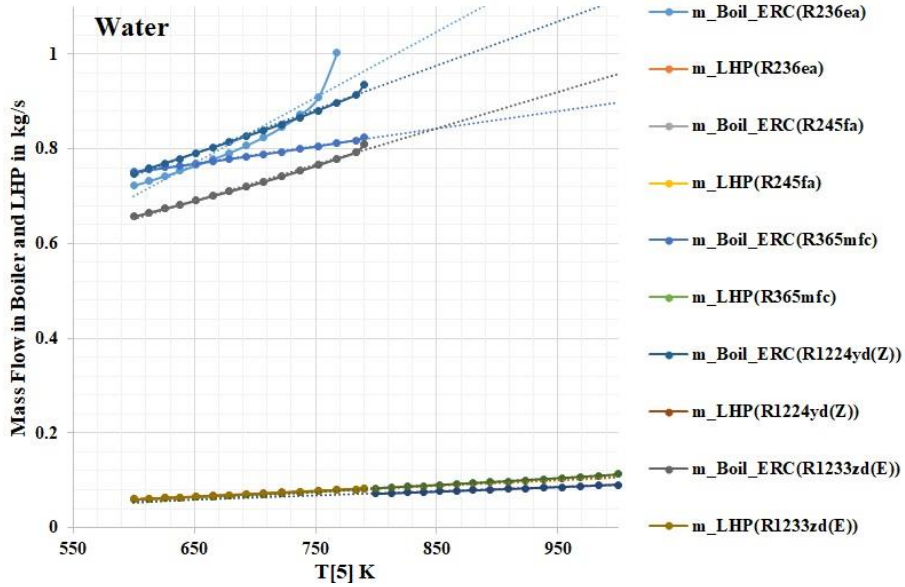


Figure 24. Mass Flow Rate of ERS & LHP with T_5 for Water

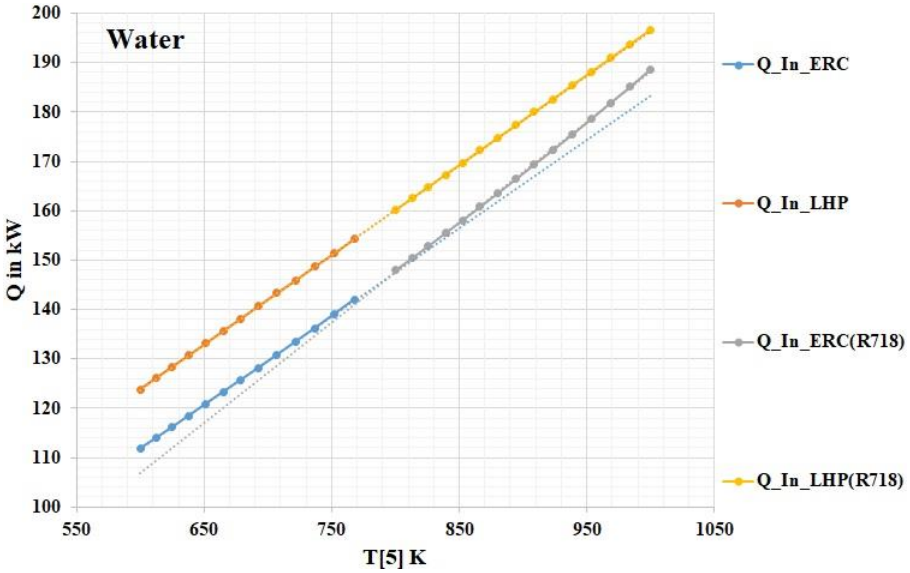


Figure 25. Heat Available as Input to ERS & LHP with T_5 for Water

CONCLUSIONS

The results obtained based on various Input parameters have been discussed, especially for R718 &R1224yz (D) (i.e., 0 ODP & 0 GDP). It has also been observed that the Temperature that can be made available for heat transfer with varying the Peak Temperature of GPC can be near or beyond the critical temperature of the eco-friendly refrigerants hence reducing the overall requirement of heat input. Hence, making the combined cycle highly attractive for industrial & eco-friendly usage. Moreover, the followings are the quantitative conclusions to the research investigation.

i. The temperature available at the ERC Boiler Input is in the range of 380 K to 455K which is very close to and above the Critical Temperature of the New Eco-Friendly Refrigerants.

ii. The Maximum Refrigeration capacity for variable Peak GPC Temperature conditions for R365mfc, R1224yd (Z) & R718 has been recorded as 10.35kW, 9.903kW & 31.545kW respectively, increasing with the temperature.

iii. Furthermore, the Maximum COP for variable Peak GPC Temperature Condition for R365mfc, R1224yd (Z) & R718 has been observed to be as 0.2856, 0.241 & 0.285 respectively, showing a downwards trend with increasing temperature.

iv. Whereas, for Peak Boiler Pressure Maximum Refrigeration Effect has been calculated for R365mfc, R1224yd (Z) & R718 at 12.93, 11.41 & 17.65 increasing with the pressure.

v. The COP under the variable pressure condition for R365mfc, R1224yd (Z) & R718 is 0.275, 0.241& 0.289.

vi. For Water as working fluids of the LHP, the least Mass Flow Rate in ERC & LHP for R1224yd (Z) & R718 has been 0.8651 kg/s &0.07329 kg/s and 0.07071 kg/s & 0.08299 kg/s respectively.

vii. Moreover, for Acetone as working fluids of the LHP, Heat Input Available has been maximum at ERC & LHP for R1224yd (Z) & R718 is 142 kW & 160.1kW and 155.9 kW & 197.1 kW respectively, for variable pressure conditions.

viii. For Variable Peak GPC temperature condition for Water as working fluids of the LHP, the least Mass Flow Rate in ERC & LHP for R1224yd (Z) & R718 has been 0.747 kg/s &0.072 kg/s and 0.0595 kg/s & 0.08299 kg/s respectively.

ix. Moreover, for Acetone as working fluids of the LHP, Heat Input Available has been maximum at ERC & LHP for R1224yd (Z) & R718 is 154.3 kW & 182.6kW and 181.5 kW & 363.1 kW respectively, for variable pressure conditions.

It can be observed that for the combination of 1224yd (Z) (ERC)-Water (LHP) as compact, eco-friendly & industrially viable combined system can be presented.

RECOMMENDATIONS FOR FUTURE RESEARCH WORKS

It can be recommended that further experimental work be initiated on this novel system to obtain a low-cost, eco-friendly Industrial refrigeration system that can be used for the storage of different sensitive items as well as for comfort usages. Studies on customization of the LHP must also be taken up for heavy industrial applications.

REFERENCES

[1] G. A. Untea, A. Dobrovicescu, L. Grosu, and E. C. Mladin “Energy and Exergy Analysis of an Ejector Refrigeration System” ISSN 1454-2358 U.P.B. Sci. Bull., Series D, Vol. 75, Iss. 4, 2013

[2] M. Ebadollahi, H. Rostamzadeh, P. S. Matin, H. Ghaebi, and M. Amidpour “Performance Evaluation of Ejector Refrigeration Cycle (ERC) Using Multi-parallel Ejectors” The 25th Annual International Conference on Mechanical Engineering ISME2017, TarbiatModares University, Tehran, Iran, 2-4 May 2017.

[3] G. Besagni, L. Croci and R. Nesa “A Screening of Low-GWP Refrigerant for Ejector Refrigeration” ISSN 2283-9216 Chemical Engineering Transactions vol. 70, 2018.

[4] F. Akkurt and A. Kahraman “An Exergy Analysis of Solar-Assisted Ejector Cooling System for Different Area Ratios at Their Maximum Coefficients of Performance Values” Thermal Science: Vol. 23, No. 2A, pp. 717-726, the Year 2019.

[5] G. Besagni, R. Mereu and F. Inzoli “Ejector refrigeration: A comprehensive review” Renewable and Sustainable Energy Reviews 53, 373–407,2016.

[6]J. Liu “Performance Improvement Potential Analysis of a Booster-Assisted Ejector Refrigeration System” 2169-3536 IEEE Translations, 2019.

[7] F. Memet and A. Preda “An Analysis of the Performance of an Ejector Refrigeration Cycle Working with R134a” Modern Technologies in Industrial Engineering (ModTech2015), Materials Science and Engineering 95, 012035,2015.

[8] S. T. Taleghani, M. Sorin, and S. Poncet “Analysis and Optimization of Exergy Flows inside a Transcritical CO2 Ejector for Refrigeration, Air Conditioning and Heat Pump Cycles” Energies, 12, 1686, 2019.

[9] A. K. S. Al-Sayyab, J. Navarro-Esbrí, V. Manuel, S. Francés and A. Mota-Babiloni "Conventional and Advanced Exergoeconomic Analysis of a Compound Ejector-Heat Pump for Simultaneous Cooling and Heating" Energies 2021, 14, 3511.

[10] B. Sharma, G. Sachdeva and V Kumar "Exergy analysis of heat assisted ejector refrigeration system using R1234yf" IOP Conf. Series: Materials Science and Engineering 1104, 012035,2021.

[11] V. Kumar and G. Sachdeva "Experimental investigation of an ejector refrigeration system using R-134a" IOP Conf. Series: Journal of Physics: Conf. Series, 1240, 012168,2019.

[12] M. Ebadollahi, H. Rostamzadeh, H. Ghaebi, M. Amidpour, F. Rostamian and H. Abioghli "Energy analysis and performance evaluation of a novel multi-evaporator ejector refrigeration cycle (ERC)" Journal of Energy Management and Technology (JEMT) Vol. 1, Issue 2, 2017.

[13] L. R. Reddy "Numerical Investigation of Ejectors for Ejector Refrigeration System" ISSN No: -2456-2165, International Journal of Innovative Science and Research Technology, Volume 3, Issue 5, May–2018.

[14] S. Li, Y. Liu, Y. Liu and J. Zhang "Performance comparison of ejectors in ejector-based refrigeration cycles with R1234yf, R1234ze(E) and R134a" Environmental Science and Pollution Research,2021.

[15] G. Besagni, L. Croci, N. Cristiani1, G. R. Guédon, and F. Inzoli "Refrigerant selection for ejector refrigeration systems: a multiscale evaluation" E3S Web of Conferences 197,10011,2020.

[16] R. S. Mishra "Thermodynamic performances of ejector refrigeration systems (ERS) using low GWP ecofriendly HFO refrigerants" ISSN 2456-6934, International Journal of Research in Engineering and Innovation Vol-4, Issue-5, 236-240, 2020.

[17] R S Mishra, A. Dwivedi and Shadab Ahmad "A thermodynamic analysis of ejector type vapour refrigeration system using eco-friendly refrigerants" ISSN 2456-6934, International Journal of Research in Engineering and Innovation, Vol-1, Issue-2, 40-48,2017.

[18] J. He, Y. Guo. H. Zhang, J. Miao, L. Wang and G. Lin "Design and experimental investigation of a neon cryogenic loop heat pipe" Heat Mass Transfer, 53:3229–3239,2017.

[19] M. M. Hossain and J. Kizito "Experimental Study of Thermal Performance of A Flat Loop Heat Pipe" ISSN: 2155-658X, Frontiers in Heat Pipes (FHP), 6, 5, 2015.

[20] I. Setyawan, I. Ibnu Hakim and N. Putra "Experimental Study on A Hybrid Loop Heat Pipe" MATEC Web of Conferences,101, 03011,2017.

[21] L. Martvoňová, M. Malcho and J. Jandačka "Increase the Efficiency of the Fireplace Insert with Loop Heat Pipe" MATEC Web of Conferences 328, 04002, 2020.

[22] A. Anikivi, R. Naik and K R. Narasihma "Investigation of Copper-Water Loop Heat Pipe for Different Filling Ratios", IOP Conf. Series: Materials Science and Engineering, 872,012070, 2020.

[23] D. Gai, Y. Yin and C. Chen "Investigation of Instability on Loop Heat Pipe with Flat Evaporator", IOP Conf. Series: Earth and Environmental Science 467, 012024, 2020.
[24] K.N. Shukla "Thermo-fluid Dynamics of Loop Heat Pipe Operation" International Communications in Heat and Mass Transfer 35,916–920, 2008.

[25] V. Buz and K. Goncharov "Loop Heat Pipe Dynamics Modeling and Analysis" IOP Conf. Series: Materials Science and Engineering, 1139, 012011, 2021.

[26] T. Shioga, Y. Mizuno and H. Nagano "Operating Characteristics of a New Ultra-Thin Loop Heat Pipe" International Journal of Heat and Mass Transfer 151, 119436, 2020.

[27] H. Nagano, H. Nagai, F. Fukuyoshi and H. Ogawa" Study on Thermal Performances of a Small Loop Heat Pipe" Journal of Thermal Science and Technology, June 2008.

[28] F. korn, "Heat Pipes and its Applications" Project Report 2008 MVK160 Heat and Mass Transport May 07, Lund, Sweden,2008.

[29] A. Dwivedi and R. S. Mishra," Thermodynamic Analysis of Heat Pipe Using Ammonia, Water and Ethanol with a View to Being Used in Refrigeration" ISSN 2347 - 3258 International Journal of Advance Research and Innovation, Volume 3, Issue 3, 498-502,2015.

[30] S. Hirasawa, T. Kawanami and K. Shirai, "Heat Transfer and Flow Analysis in Loop Heat Pipe with Multiple Evaporators Using Network Model", Journal of Mechanics Engineering and Automation, 6, 319-325,2016.

[31] P.D. Dunn, D.A. Reay, "Heat Pipes", Pergamon Press, Oxford, 1993.

[32] C.P. Arora, "Refrigeration and Air Conditioning", Tata McGraw-Hill Publishing Company Limited, Delhi, Third Edition, 2009.

[33] http://iusti.univ-provence.fr/document.php?pagendx=11288&project=iusti

[34] http://www.ertc.od.ua/en/about_ert_en.html

[35] P. K. Nag, "Basic & Applied Thermodynamics", Tata McGraw-Hill Publishing Company Limited, Delhi, Second Edition, 2017.

View publication stats

Int J Syst Assur Eng Manag
https://doi.org/10.1007/s13198-021-01385-4

ORIGINAL ARTICLE

A regression test case prioritization technique targeting 'hard to detect' faults

Sourav Biswas[1] · **Raghav Rathi**[2] · **Arpita Dutta**[1] · **Pabitra Mitra**[1] · **Rajib Mall**[1]

Received: 4 April 2021 / Revised: 14 May 2021 / Accepted: 8 September 2021
© The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2021

Abstract We propose a novel regression test case prioritization technique targeting to detect 'hard to detect' faults. A 'hard to detect fault' is a fault that is detected by only one test case. In our technique, we first seed a large number of bugs into a program to create mutants. Each mutant is executed with the test suite and their execution results are recorded in a fault matrix. Using the fault matrix, we first prioritize the test cases based on their 'hard to detect' values. Our technique assigns higher priority to those test cases which reveal bugs in a program-component that are hard to expose. Subsequently, the remaining test cases are prioritized based on their fault detection capability. Our experimental results show that on an average our proposed TCP technique performs 43.82% better than existing TCP techniques.

Keywords Regression testing · Test case prioritization · Mutation testing

✉ Sourav Biswas
bsws.sourav@gmail.com

Raghav Rathi
rathi.raghav00@gmail.com

Arpita Dutta
arpitad10j@gmail.com

Pabitra Mitra
pabitra@cse.iitkgp.ac.in

Rajib Mall
rajib@cse.iitkgp.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur, India

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

1 Introduction

Regression testing provides confidence that a set of changes made to a program have not degraded the behavior of the unchanged functionalities of the program Mall (2018). This can simplistically be ensured by rerunning all the available test cases after each change. However, rerunning all test cases after every change to a program is time-consuming and expensive. For this reason, suitable regression test selection (RTS) techniques (Rothermel 1994; Harrold et al. 2001; Gregg et al. 1999) are deployed. RTS techniques select a subset of the test cases to run after a change. Test case prioritization (TCP) techniques (Elbaum et al. 2002; Rothermel et al. 2001; Elbaum et al. 2004; Malishevsky et al. 2006) have been proposed to prioritize the regression test cases. TCP techniques help when the tester is unable to run the full regression test suite due to time or cost considerations. Also, by executing the regression test suite in order of their priority assigned by a TCP technique, the rate of fault detection improves, thereby providing more time to the developers for bug correction.

TCP techniques rank the test cases of a test suite in a priority order. For ranking the test cases, TCP techniques optimize specific criteria such as the maximization of the rate of bug[1] detection, and maximization of the rate of increase in code coverage. In other words, a TCP technique assigns a priority score to each test case based on the number of yet undetected faults detected by a test case or the extent of incremental code coverage achieved. Based on the assigned score, the test cases are ordered in a

[1] In this paper, we have used the terms 'bug' and 'fault' interchangeably.

Published online: 18 September 2021

prioritized list. A tester can choose the top most test cases from the prioritized list on time and cost considerations.

In the literature, several TCP techniques have been proposed by different researchers. Most of the techniques are based on coverage of code Rothermel et al. (2001); Elbaum et al. (2002), requirements Srikanth et al. (2005); Krishnamoorthi and Sahaaya Arul Mary (2009), or various types of model elements Korel et al. (2005); Panigrahi and Mall (2010, 2013). However, test case prioritization techniques based on the capability of test cases to expose 'hard to detect' bugs has not been reported.

In this context, we proposed a novel test case prioritization technique based on the potential of the test cases to expose 'hard to detect' bugs. Our technique prioritizes the test cases in three different groups. First group contains the list of prioritized test cases that detect bugs not detected by other test cases, thereby these test cases are indispensable. The second group contains a priority list of test cases which detect bugs that are also detected by other test cases. Third group contains the prioritized order of the redundant test cases. A higher priority is accorded to a test case that detects bugs detected by a very few test cases.

Rest of the paper is organized as follows. In Sect. 2, we discuss related work. Our proposed TCP technique is presented in Sect. 3. In Sect. 4, we discuss the experimental results. Finally, Sect. 5 concludes this paper.

2 Related work

A large number of TCP techniques have been proposed by researchers. These TCP techniques can broadly be classified into: Coverage-based Rothermel et al. (2001); Elbaum et al. (2002), requirement-based Srikanth et al. (2005); Krishnamoorthi and Sahaaya Arul Mary (2009), model-based Korel et al. (2005); Panigrahi and Mall (2010, 2013) techniques. We briefly review these techniques in the following.

Rothermel et al. (2001, 1999) have proposed a number of TCP techniques. They proposed a statement coverage-based approach (Panda et al. 2019; Lou et al. 2015) that orders test cases according to the number of statement exercised. Their branch coverage-based technique (Rothermel et al. 2001, 1999) rank the test cases on the basis of total number of branches exercised by each test case. They also proposed TCP techniques based on fault-exposing-potential (FEP) (Rothermel et al. 2001, 1999) of the test cases. To prioritize the test cases based FEP, they calculated mutation-score for each test case. Mutation score is a ratio of mutants detected by a test case t to the total number of mutants. They used APFD (Rothermel et al. 2001) metric to determine the performance of their proposed techniques. Later, Elbaum et al. (2002) extended the work of Rothermel et al. (2001) by proposing a function coverage approach. In their technique, test cases are prioritized according to the number of functions exercised by each test case. Their experimental studies showed that their proposed technique helps in early fault detection. Jones and Harrold (2003) proposed effective test suite reduction and prioritization technique using a modified condition/decision coverage (MC/DC) based approach (Dutta et al. 2019).

Srikanth et al. (2005) proposed a system-level requirement-based TCP technique. They used factors such as the number of customer requirements, number of changes to requirements, difficulty to implement requirements, and fault susceptibility of requirements. Krishnamoorthi and Sahaaya Arul Mary (2009) proposed a software requirement specification (SRS) document-based TCP prioritization technique. Their technique prioritized test cases using six factors: customer requirements, difficulty in implementing requirements, changes in requirements, faults in requirements, completeness of requirements, and traceability of requirements.

Korel et al. (2005) presented a state model-based TCP technique. They constructed two different Extended Finite State Machine (EFSM) models: one for the original software, and another for the modified software. These two models were compared to identify any differences in transitions. Based on the differences in transitions, test cases were prioritized. Panigrahi and Mall (2010, 2013, 2013) proposed a model-based TCP technique for object-oriented programs. An extended system dependency graph (ESDG) Larsen and Harrold (1996) was constructed to capture the changes to a given program. They computed a slice on the constructed model to identify the affected parts. Test cases were prioritized based on the extent of their execution of the affected nodes. Their experimental study showed a significant improvement in APFD value over code-based techniques (Rothermel et al. 2001; Elbaum et al. 2002).

Panda et al. (2019) proposed a UML diagram based test scenario prioritization for object-oriented software. They have constructed a state machine graph (SMG) from a source code. Then they identified all the nodes which are got affected due to any changes in the source code. They have stored the information of affected node after each modification of the software. Next time when any changes made in the software their technique order the test cases based on the analysis of previously stored information.

Lou et al. (2015) proposed two different test case prioritization techniques with mutation testing. The test cases were prioritized based on their fault detection capabilities. Their first proposed method is statistics-based and the second method is based on a probabilistic approach. Both the approaches follow the following three steps for TCP.

First, they identify the modified part in the original program. Subsequently, they generates the mutants using the mutation operators involved in the modified part. Finally, the test cases are scheduled based on their fault detection capabilities. They have evaluated the performance of both the proposed TCP techniques on fourteen different versions of three open-source Java projects. They compared their techniques with two approaches namely addition and total. These two techniques are widely used as the control approaches in test case prioritization. Also, they applied their methods at two different levels of system design viz., test-method level and test-class level. Based on their experimental results, they reported that the statistics based TCP approach outperforms the probability based TCP as well as the other TCP techniques.

Dhareula and Ganpati (2020) proposed a flower pollination algorithm (FPA) based regression test case prioritization technique. FPA is a nature inspired metaheuristic optimization algorithm that tries to emulate the pollination process of flowering plants. Authors used only the code coverage information of test cases as input to the flower pollination algorithm. FPA does not contain the prior knowledge of the fault detection abilities of test cases. The test cases are prioritized based on the maximum coverage of faults based in the minimum execution time. They have validated their approach using APFD metric on three Java applications. Based on their experimental results, it was observed that FPA based TCP is more effective than reverse order, random order and reverse-random order of test cases.

Vescan et al. proposed an optimized test case prioritization technique using ant colony optimization algorithm. They have named their approach as Test Case Prioritization-ANT (TCP-ANT). The criteria used by the optimization algorithm is the number of faults that are not covered by the test case and the sum of the severity of the uncovered faults. Along with this, the execution time of the test cases are used to compute the pheromone deposited on the edges of the graph. They validated their proposed approach using Grep SIR (2001) data set and found it to be encouraging.

Mirarab and Tahvildari (2008) used Bayesian networks (BNs) to prioritize regression test cases. They used several code coverage metrics, fault proneness of the software components and source code quality metrics to construct a unified model using BNs. They have used APFD metric to compute the effectiveness of their approach. Later, Zhao et al. Zhao et al. (2015) extended their approach by integrating the clustering method based on code coverage and BNs. Their approach helps to break the ties between the test cases. Busjaeger et al. Busjaeger and Xie (2016) introduced the usage of SVMmap Tan et al. (2016) for TCP. They used code coverage, test-failure history, similarity between tests and changes, test case age, and fault history information for test prioritization. Different neural network models have also been used for TCP Gökçe et al. (2006); Spieker et al..

3 Proposed technique: HDFDC

In this section, we present our technique for test case prioritization. We have named our technique HDFDC, being an abbreviation of test cases prioritization using '**H**ard to **D**etect' bugs and **F**ault **D**etection **C**apability. We first discuss a few important terminologies that we later use to describe our HDFDC technique.

Fault Matrix (FM) A fault matrix (FM) is a matrix of order $n \times m$, where n represents the number of test cases and m represents the total number of faults present in a program. Let TC_i denote the i^{th} test case and F_j represent the j^{th} fault. The bug detection information of a test case is recorded in FM as given by Eq. 1.

$$FM_{ij} = \begin{cases} 1, & \text{if bug}j\text{is detected by test case}i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Table 1 represents an example of a Fault Matrix in which the rows represent the test cases and the columns represent the faults.

Fault Vector (FV): Fault vector FV=$(FV_0, FV_1, FV_2, ..., FV_{m-1})$ is a vector of length m, in which a vector component FV_j represents the number of test cases that detect the j^{th} fault. Variable m represents the total number of faults present. Initially, for each fault j, FV_j is *0*. Later, on the execution of each test case, FV_j gets incremented if the j^{th} fault is detected by that test case.

For example, fault vector $FV = (2, 1, 0, ..., 1)$. Here, $FV_0 = 2$ which shows *Fault0* is detected by two test cases. Similarly, $FV_1 = 1$, and $FV_2 = 0$ show that the *Fault1* and *Fault2* are detected by test cases 1 and 0 respectively.

Fault Detection Capability (FDC) of a test case : FDC measures the fault detection capability of test cases. A test case gets a higher priority based on the obtained FDC score. A higher FDC score is accorded to a test case that detects bugs which are detected by a very few test cases. FDC metric for each test case i is evaluated using Eq. 2:

$$FDC_i = \sum_{j=0}^{n-1} \frac{1}{FV_j + 1}, \quad where \quad FM_{ij} = 1 \quad (2)$$

where, j indicates the j^{th} fault, n indicates the total number of faults, FM shows the Fault Matrix and FV represents the fault vector.

We illustrate computation of the FDC metric using an example provided in Table 1. Table 1 represents an example of fault matrix where bug detection information of

Table 1 Illustration of Fault Matrix (Here, F stands for Fault)

	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
TC_0	1	1	1	1	1	0	1	0	0	0	0	0	0	0
TC_1	0	1	0	1	1	0	0	0	0	0	0	0	0	0
TC_2	0	0	0	0	0	1	0	0	0	1	1	0	0	1
TC_3	0	0	0	1	1	0	0	1	1	0	0	1	1	0
TC_4	0	1	0	1	1	0	0	1	1	0	0	0	0	0
TC_5	0	0	0	1	0	0	0	1	0	0	0	0	0	0
TC_6	0	1	0	1	0	0	0	0	1	0	0	0	0	0
TC_7	0	0	0	1	1	0	0	0	1	0	0	1	1	0

eight test cases (TC_0 to TC_7) shown for fourteen faults (*Fault*0 to *Fault*13). Initially the Fault Vector will be assigned with 0 for all bugs. Hence, $FV = (0,0,0,0,0,0,0,0,0,0,0,0,0,0)$. After execution of TC_0, the updated FV is (1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0). It can be observed from Table 1 that TC_1 detects *Fault*1, *Fault*3, and *Fault*4. Based on the fault detection of TC_1, we can calculate the *FDC* metric for TC_1 using the updated *FV* by TC_0 as follows:

$$\begin{aligned} FDC_1 &= \frac{1}{FV_1+1} + \frac{1}{FV_3+1} + \frac{1}{FV_4+1} \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \\ &= 1.5 \end{aligned}$$

3.1 An overview of our HDFDC approach

Figure 1 depicts a schematic representation of our proposed technique. For a given Java class, we first generate a large number of mutants. Each mutant contains a single fault. In the Figure, M_1, M_2, ..., M_m are the mutants generated by a Mutator. In the next step, each mutant is executed with the original test suite and the output is stored for each test case in a separate file. In Fig. 1, it can be seen that mutant M_1 is executed using the test cases TC_1, TC_2, ..., TC_n, and the corresponding output files are $O_{1,1}$, $O_{1,2}$, ..., $O_{1,n-1}$, $O_{1,n}$. Similarly, the original program is also executed using the test cases TC_1, TC_2, ..., TC_n and generated output files are labelled O_1, O_2, ..., O_{n-1}, O_n respectively. Subsequently, we compare the output files corresponding to each mutated version with the output files generated for the original program and generate a Fault Matrix (FM). Using the FM, we prioritize the test cases in three separate groups: *Group 1*, *Group 2*, and *Group 3*. These three groups are defined as follows.

Group 1 This group contains all the test cases that detect 'hard to detect' bugs. All the test cases of this group are selected compulsory. If we ignore any test case of this group, there can be undetected faults. This group is prioritized based on the total number of 'hard to detect' bugs.

Group 2 This group contains all the test cases which detect faults that are also detected by other test cases. It contains the test cases in priority order based on the capability of exposing faults which are detected by more than one test cases. To measure the fault exposing potential of a test case, we used our proposed Fault Detection Capability (FDC) metric.

Group 3 This group contains all the test cases which detect the bugs which are already detected by the test cases of *Group 2*. Therefore this group contains all the redundant

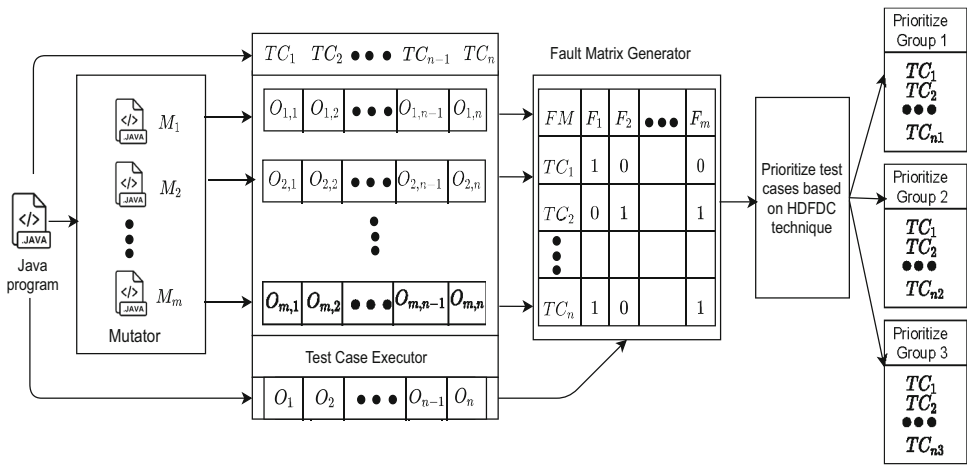


Fig. 1 A Schematic representation of HDFDC technique

test cases. Even though Group 3 are redundant test cases, however this have slightly inferior potential of detecting bugs on a change. So, if the test budget permits, at least the more promising of these should be executed. Therefore, the redundant test cases are also prioritized based on the FDC metric.

Figure 2 represents the activity diagram model for HDFDC. Various activities that are carried out shown in Fig. 2 are described below:

- *Bugs Seeding* In this step, we select a program and seed a large number of bugs into it and create different mutants for that program. Each mutant contain a single bug.
- *Execute Mutated Program using the Test Suite* In this step, we select a mutant and execute it with all the available test cases and outputs are stored in a separate file. This step is repeated for each mutant.
- *Execute the Original Program using the Test Suite* In this step, the selected original program is executed with all the existing test cases, and the output is written into a file. For each test case, a separate file is created, which contains the result of the original output. For example, if there are n test cases available for a program then n output files are created.
- *Compare the Results of the Original and the Mutated Programs*: In this step, for each test case, we compare the original output of a program with the output of each mutated version of the same program. If the original output and the output of a mutated version are not the same for a test case, then we say that the test case is able to detect the fault introduced in the mutated version. For example, for a test case TC_i and a mutated version M_j, the original output is O_i, and the output of the mutated version is $O_{j,i}$. If O_i and $O_{j,i}$ are not the same, then test case TC_i is failing for the mutant M_j. In other words, the test case TC_i is able to detect the bug which is present in the version M_j.
- *Record Mutant Detected* In this step, the mutant detected by a test case is recorded in the 'Fault Matrix'.
- *Prioritize the Test Cases*: In this step, the test cases are prioritized in three different groups: *Group 1*, *Group 2*, and *Group 3*. Test cases which detect at least one 'hard to detect' bugs are prioritized in *Group 1*, and the rest of the test cases are prioritized in *Group 2*, and *Group 3* based on the FDC metric.

3.2 Algorithmic formulation of HDFDC

In this subsection, we present the algorithmic description of our proposed HDFDC technique for TCP.

Algorithm 1 HDFDC

Require: Fault Matrix FM and Fault Vector array FV
1: Initialize: FV[j] = 0, for all bug j
2: Select all the test cases that detect at least one 'hard to detect' bug
3: Add the selected test cases into *Group 1*
4: Prioritize the test cases of *Group 1* based on maximum number of 'hard to detect' bugs are detected.
5: // Update FV array by *Prioritize Group 1* test cases:
6: **for** each test case i ∈ *Prioritize Group 1* **do**
7: **for** each bug j **do**
8: **if** (FM[i][j] ==1) **then**
9: FV[j]=FV[j]+1;
10: **end if**
11: **end for**
12: **end for**
13: Add the remaining test cases into Group 2 (test case $TC_i \in$ *Group 2*, if $TC_i \notin$ *Group 1*)
14: // Compute FDC score:
15: **for** each test case i ∈ *Group 2* **do**
16: FDC[i]=0
17: **for** each bug j **do**
18: **if** (FM[i][j] ==1) **then**
19: FDC[i]=FDC[i]+$\frac{1}{FV[j]+1}$
20: FV[j]=FV[j]+1;
21: **end if**
22: **end for**
23: **end for**
24: Remove the test cases which detected same bugs from *Group 2* and add the test cases into *Group 3*
25: Prioritize the test cases of *Group 2* and *Group 3* based on the obtained FDC score
26: return *Prioritize Group 1*, *Prioritize Group 2*, and *Prioritize Group 3*

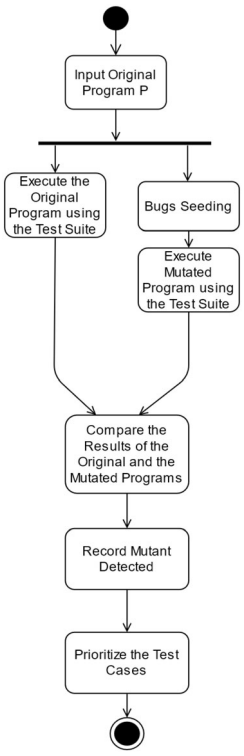


Fig. 2 Activity diagram of our proposed HDFDC technique

HDFDC algorithm takes Fault Matrix and Fault Vector as input and generates a prioritized list of test cases as its output. Steps 2 and 3 generate *Group 1* test cases. Step 4 prioritizes the test cases present in *Group 1*. Steps 6 to 12 update the FV array after the prioritization of *Group 1* test cases. Steps 13 to 24 generate prioritized test cases of *Group 2* based on the FDC score. In Steps 15 to 23, the FDC score for each test cases of the Group 2 is computed and updated the FV array. In step 24, redundant test cases are selected from *Group 2* and the test cases are added into *Group 3* . Step 25, HDFDC algorithm prioritized the test cases of *Group 2 Group3* based on the *FDC* score and return the *Prioritized Group 1*, *Prioritized Group 2*, and *Prioritized Group 3* in Step 26.

3.3 An illustrative example

We illustrate the working of our HDFDC approach for TCP using the Fault Matrix in Table 2. For better visual representation, we have used '×' instead of 1 and have used 'blank; instead of 0 in the Table 2. '×' denotes that the bug is detected by the corresponding test case, and blank indicates that a particular test case does not detect that particular bug.

HDFDC first selects all the test cases which detect 'hard to detect' faults. From Table 2, it can be observed that *Fault*0, *Fault*2, *Fault*5, *Fault*6, *Fault*9, *Fault*10, and *Fault*13 are 'hard to detect' faults and these are detected by TC_0, and TC_2. So, the selected test cases are stored in *Group 1*= $\{TC_0, TC_2\}$. HDFDC technique orders the selected test cases according to maximum number of 'hard to detect' bugs detection. Among our selected test cases TC_0 detects three 'hard to detect' faults and TC_2 detects four 'hard to detect' faults. Therefore, TC_2 will get a higher priority than TC_0. The *Prioritized Group 1* is $\{TC_2, TC_0\}$. Table 3, highlighted the faults that are detected by TC_2 and TC_0.

After ordering the test cases of *Group 1*, our algorithm updates the Fault Vector FV for each test case of *Prioritized Group 1*. To start with, all the entries of the FV array are initialized to 0. Table 4, shows the initial FV array. Table 5, shows the FV array after the test cases of Group 1 have been prioritized. The remaining test cases would be added in the *Group 2*.

Table 2 Illustration of fault matrix

	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
TC_0	×	×	×	×	×		×							
TC_1		×		×	×									
TC_2						×				×	×			×
TC_3				×	×			×	×			×	×	
TC_4		×		×	×			×	×					
TC_5				×				×						
TC_6		×		×					×					
TC_7				×	×				×			×	×	
TC_8				×	×				×			×	×	
TC_9		×		×					×					

Table 3 Fault Matrix highlighted bug detection by Prioritize Group 1 test cases

	F0	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13
TC_0	×	×	×	×	×		×							
TC_1		×		×	×									
TC_2						×				×	×			×
TC_3				×	×			×	×			×	×	
TC_4		×		×	×			×	×					
TC_5				×				×						
TC_6		×		×					×					
TC_7				×	×				×			×	×	
TC_8				×	×				×			×	×	
TC_9		×		×					×					

Table 4 Initial *FV* array

0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 5 Fault Vector array *FV* after the *Prioritize Group 1* test cases get prioritized

1	1	1	1	1	1	1	0	0	1	1	0	0	1
0	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 6 Updated Fault Vector array *FV* after the computation of *FDC* of TC_1

1	2	1	2	2	1	1	0	0	1	1	0	0	1
0	1	2	3	4	5	6	7	8	9	10	11	12	13

Table 7 *FDC* score obtained by the test cases of *Group 2*

Test Case	FDC Score
TC1	1.5
TC3	4.66
TC4	1.83
TC5	0.53
TC6	0.75
TC7	1.59
TC8	1.17
TC9	0.47

$$Group\ 2 = \{\{TS\} - \{Group\ 1\}\}$$

where, TS represents the entire test suite

Table 8 Prioritized order of the test cases of *Prioritize Group 2*

Test Case	FDC Score
TC3	4.66
TC4	1.83
TC7	1.59
TC1	1.5
TC6	0.75
TC5	0.53

$$\begin{aligned} Group\ 2 = \{\{&TC_0, TC_1, TC_2, TC_3, TC_4, TC_5, \\ &TC_6, TC_7, TC_8, TC_9\} \\ &- \{TC_2, TC_0\}\} \\ = \{&TC_1, TC_3, TC_4, TC_5, TC_6, TC_7, TC_8, TC_9\} \end{aligned}$$

HDFDC algorithm computes the FDC score of each test

Table 9 Prioritized order of the test cases of *Prioritize Group 3*

Test Case	FDC Score
TC8	1.17
TC9	0.47

case present in *Group 2*. The computation of FDC score is carried out in Steps 15 to 23 of our Algorithm 1 HDFDC. The FDC metric for *TC*1 is computed using the Fault Matrix shown in Table 2 and Fault Vector shown in Table 5:

$$FDC_1 = \frac{1}{FV[1]+1} + \frac{1}{FV[3]+1} + \frac{1}{FV[4]+1} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1.5$$

where, FM[1][1] =1, FM[1][3]=1, and FM[1][4]=1.

Along with the computation of the FDC score for each test case, FV array is also updated. As can be observed, the test case *TC*1 detects the faults *Fault*1, *Fault*3, and *Fault*4. So, FV[1], FV[3], and FV[4] value will be incremented by 1 (FV[j] = FV[j]+1). The updated FV array after the FDC computation of TC1 is shown in Table 6. Similarly, for the rest of the test cases of *Group 2*, HDFDC algorithm will compute the FDC score and simultaneously update the FV array.

The final FDC score for all the test cases of Group 2 is shown in Table 7. From Table 7, it can be observed that *TC*3 obtained highest FDC score. From Table 2, it can be observed that *TC*3 detects the faults *Fault*7, *Fault*8, *Fault*11, and *Fault*12 which are not detected by any other test cases before the execution of *TC*3. From the Fault Matrix in Table 2, it can be observed that *TC*8 and *TC*9 are detecting same bugs which are already detected by *TC*7 and *TC*6 respectively. So, *TC*8 and *TC*9 are added in the *Group 3* as per the algorithmic formulation of HDFDC technique. Based on the computed FDC scores, Group 2 and Group 3 are prioritized next. Prioritized Group 2 is shown in Table 8 8, and prioritized Group 3 is shown in Table 9.

4 Experimental studies

In this section, we first discuss the experimental setup and subject programs that we have used. We then present an evaluation metric to compare the effectiveness of different techniques. Subsequently, we analyze the obtained results and discuss the threats to the validity of the results.

Table 10 Project characteristics

S. No.	Project Name	LOC	No. of Classes	No. of Test cases	No. of Mutants
1	Book	80	1	15	176
2	Elevator	580	1	142	800
3	JFreeChart	1481	1	654	1250
4	Jtcas	169	1	160	180
5	Tritype	106	1	65	50
6	Area and Perimeter	916	1	110	50
7	Cli	4499	55	38	39
8	Closure	149516	1626	412	174
9	Codec	5834	30	43	18
10	Collections	64908	898	506	4
11	Compress	12816	92	81	47
12	Csv	1674	11	15	16
13	Gson	11560	529	261	18
14	JacksonCore	27908	190	135	26
15	JacksonDatabind	75069	2448	590	112
16	JacksonXml	7500	106	98	6
17	Jsoup	4512	171	41	93
18	JxPath	28657	397	201	22
19	Lang	61289	1298	236	64
20	Math	186609	1540	1310	106
21	Mockito	22654	1406	319	38
22	Time	68712	546	317	26

4.1 Experimental setup

All experiments have been conducted on a 64-bit Ubuntu machine, 2.5 GHz Intel Core i5 processor, with 8 GB RAM. All the programs which are used to evaluate the performance of our technique are Java classes. The Java programs are compiled on JDK version 8. We have used the mutation tool muJava muJava (2016) to create the mutants of considered Java programs. All the modules were coded using Python.

4.2 Subject programs

To evaluate the effectiveness of our HDFDC technique, we have used twenty two publicly available Java projects. Table 10 shows the important characteristics of these Java projects. We have considered the programs *Book*, *Elevator*, *JFreeChart*, *Jtcas*, *Tritype* etc. *JFreeChart* and *Elevator* are obtained from an open source library that has been used in test coverage of Java programs Gligoric et al. (2015). The code for *Book* and *Area* and *Perimeter* classes are taken from a student assignment program. Rest of the programs are available at the open-source GitHub repository Defects4j Just et al. (2014) and SIR Rothermel et al..

4.3 Mutant generation

We have used muJava muJava (2016) tool for mutant generation. The muJava tool supports injection of different type of bugs that are frequently committed by the programmers. We have used class-level and method-level mutation operators. In method-level mutation operators we have used the following:

- *Arithmetic operators* One of the example of this category of bugs is replacing $++$ by $--$
- *Relational operators* One of the example of this category of bugs is replacing $>$ by $<$
- *Conditional operators* One of the example of this category of bugs is replacing&& by ||
- *Logical operators* One of the example of this category of bugs is replacing & by |
- *Assignment operators* One of the example of this category of bugs is replacing $-=$ by $+=$

In class-level category we have used muJava operators for Java-specific features, such as JTI, JTD, JSI, JSD, etc.

- *JTI* The operator inserts the keyword *this*.
- *JTD* The operator deletes uses of the keyword *this*.
- JSI : The operator adds the *static* modifier to change instance variables to class variables.
- *JSD* The operator removes the *static* modifier to change class variables to instance variables.
- *JID* The operator removes the initialization of member variables in the variable declaration so that member variables are initialized to appropriate default values of Java.

In the table 10, for the projects S. No. 1 to S. No 6, we have used muJava tool for mutant generation. Rest of the projects (S. No. 7 to S. No. 22) are taken from Defects4j repository and these projects are already seeded with real faults. The number of mutants for each of the projects are shown in the table 10.

Table 11 Illustration of APFD value calculation

TC\Fault	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
TC_1						×	×			
TC_2		×			×					×
TC_3										
TC_4				×					×	
TC_5										
TC_6	×				×					
TC_7			×					×		×
TC_8		×					×			
TC_9	×			×				×		
TC_{10}			×						×	

4.4 TCP evaluation metric

To evaluate the performance of various considered TCP techniques, we have used the average percentage of faults detected (APFD) metric. Lou et al. (2015) proposed APFD metric. It captures how fast a prioritized test suite is capable of detecting faults. A higher APFD value denotes a higher rate of fault detection. For a test suite T, APFD metric is evaluated using Eq. 3:

$$APFD = 1 - \frac{\sum_{i=1}^{m} TF_i}{n * m} + \frac{1}{2 * n} \tag{3}$$

where,

- n is the total number of test cases presents in test suite T
- m represents total number of faults to be detected
- TF_i represents the position of the first test case in T which detects fault i.

Suppose, there are ten faults $(m = 10)$ and ten test cases $(n = 10)$ for a given program. The obtained fault matrix is shown in Table 11. APFD value for this program is calculated as follows:

Prioritized test case order: $TC_4, TC_2, TC_1, TC_7, TC_6, TC_9, TC_{10}$,

TC_5, TC_8, TC_3. APFD applied w.r.t prioritized Test Suite:

$$APFD = 1 - \frac{5+2+4+1+2+3+3+4+1+2}{10*10} + \frac{1}{2*10} = 1 - \frac{27}{100} + \frac{1}{20} = 1 - 0.27 + 0.05 = 0.78$$

The APFD value for non-prioritized test suite:

$$APFD = 1 - \frac{6+2+7+4+2+1+1+7+4+2}{10*10} + \frac{1}{2*10} = 1 - \frac{36}{100} + \frac{1}{20} = 1 - 0.36 + 0.05 = 0.69$$

From the above example, it can be observed that APFD value (0.78) obtained for prioritized test suite is higher than the APFD value (0.69) for non-prioritized test suite. Therefore, we can say that a prioritized order is necessary to achieve higher fault detection rate.

Another metric we have used to evaluate the performance of various TCP techniques is $APFD_C$ Elbaum et al. (2001) which is the 'cost-cognizant' version of the $APFD$ metric. $APFD_C$ consider both the test case cost and fault severity to measure the fault detection rate of a test case. For a test suite T, $APFD_C$ metric is evaluated as:

$$APFD_C = \frac{\sum_{i=1}^{m}(f_i * (\sum_{j=TF_i}^{n} t_j - \frac{1}{2} * t_{TF_i}))}{\sum_{i=1}^{n} t_i * \sum_{i=1}^{m} f_i} \quad (4)$$

where,

- n is the total number of test cases presents in test suite T
- m represents total number of faults revealed by T
- TF_i represents the position of the first test case in T which detects fault i.
- t_j is the cost of the j^{th} test case.
- f_i is the severity of the i^{th} fault.

4.5 Experimental results

We compare the results of HDFDC with six other techniques: No-Prioritization (NP), random prioritization, Statement Coverage (SC), Branch Coverage (BC), Flower Pollination Algorithm (FPA) based regression test case prioritization technique, and ant colony optimization algorithm based test case prioritization technique (TCP-ANT). No-prioritization technique considers the test cases

Table 12 Obtained APFD scores

S. No.	Programs	NP	Random	SC	BC	FPA	TCP-ANT	HDFDC
1	Book	0.538	0.643	0.761	0.834	0.849	0.842	0.93
2	Elevator	0.461	0.486	0.691	0.7	0.743	0.865	0.927
3	JFreeChart	0.44	0.529	0.639	0.739	0.786	0.852	0.986
4	Jtcas	0.46	0.51	0.82	0.80	0.83	0.92	0.939
5	Tritype	0.49	0.542	0.564	0.581	0.688	0.795	0.814
6	Area and Perimeter	0.398	0.421	0.437	0.437	0.468	0.69	0.7
7	Cli	0.526	0.592	0.685	0.685	0.848	0.832	0.904
8	Closure	0.573	0.551	0.530	0.530	0.791	0.858	0.973
9	Codec	0.516	0.674	0.549	0.618	0.915	0.919	0.953
10	Collections	0.529	0.461	0.602	0.602	0.995	0.995	0.996
11	Compress	0.543	0.605	0.638	0.638	0.787	0.872	0.952
12	CSV	0.469	0.325	0.512	0.528	0.9	0.9	0.9
13	Gson	0.581	0.635	0.617	0.617	0.883	0.958	0.991
14	JacksonCore	0.547	0.788	0.602	0.602	0.945	0.935	0.983
15	JacksonDatabind	0.565	0.729	0.595	0.612	0.886	0.855	0.984
16	JacksonXml	0.698	0.908	0.705	0.705	0.961	0.998	0.943
17	Jsoup	0.453	0.764	0.706	0.706	0.794	0.879	0.936
18	JxPath	0.553	0.723	0.598	0.897	0.952	0.962	0.986
19	Lang	0.571	0.573	0.626	0.937	0.880	0.836	0.963
20	Math	0.566	0.415	0.541	0.795	0.913	0.777	0.985
21	Mockito	0.566	0.583	0.734	0.734	0.845	0.914	0.981
22	Time	0.490	0.591	0.593	0.625	0.860	0.926	0.978

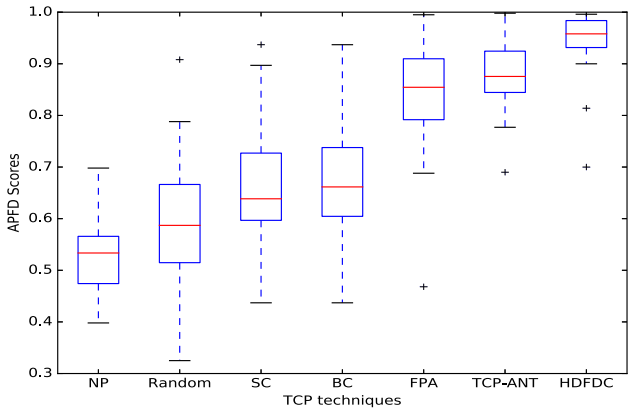


Fig. 3 Box plot representation of APFD achieved by different TCP techniques

without any prioritization. Random prioritization technique randomizes the test cases. Rothermel et al. (2001, 1999) proposed statement coverage (SC), and branch coverage (BC) based TCP approaches. These are two prominent TCP techniques available in the literature. Dhareula and Ganpati (2020) proposed flower pollination algorithm (FPA) based-Test Prioritization technique. It is a metaheuristic-based approach which prioritize the test cases very effectively. Vescan et al. proposed TCP-ANT technique. TCP-ANT prioritized the test cases based on the severity of the uncovered faults. This technique showed very encouraging results on all the experimented programs. FPA and TCP-ANT are the recent techniques in the literature and considered *APFD* metric to measure the effectiveness of the techniques.

Table 12 shows the APFD values obtained using seven different techniques. We can observe from Table 12 that our HDFDC based TCP technique obtained the maximum APFD value (0.996) for *Collections* project and the lowest APFD value (0.7) for *Area* and *Perimeter* class. It can be observed from project characteristics Table 10, *Collections* are seeded with lowest number of mutants (4) that may lead HDFDC to achieved highest APFD for *Collections*. BC obtained a similar or a better APFD score than SC except for *Jtcas*, where SC scored (0.82) and BC scored (0.80). Similarly, Random prioritization obtained better APFD than NP for most of the projects except for *Collections*, *CSV*, and *Math*, where NP scored better APFD value than Random prioritization. It can be also observed that FPA and TCP-ANT performs better than NP, Random, SC and BC for all the programs. But, our HDFDC technique performs more effectively than FPA and TCP-ANT techniques for all the projects except *CSV* and *JacksonXml* projects. FPA, TCP-ANT, and HDFDC obtained exactly same APFD value (0.9) for *CSV* and for *JacksonXml* HDFDC obtained lower APFD (0.943) than FPA (0.961) and TCP-ANT (0.998). It can be also observed from the Table 12 that among all the techniques, HDFDC showed highest

Table 13 Relative improvement achieved on APFD

S.No.	Projects	NP (%)	Random (%)	SC (%)	BC (%)	FPA (%)	TCP-ANT (%)
1	Book	72.86	44.63	22.2	11.51	9.54	10.45
2	Elevator	101.08	90.74	34.15	32.42	24.76	7.16
3	JFreeChart	124.09	86.38	54.30	33.41	25.44	15.72
4	Jtcas	104.13	84.11	14.51	17.37	13.13	2.06
5	Tritype	66.12	50.48	44.32	40.10	18.31	2.58
6	Area and Perimeter	75.87	66.27	60.18	60.18	49.57	1.44
7	Cli	71.86	52.70	31.97	31.97	6.60	8.65
8	Closure	69.8	76.5	83.5	83.5	23	13.40
9	Codec	84.68	41.39	73.5	54.2	4.15	3.69
10	Collections	88.27	116.05	65.44	65.44	0.10	0.10
11	Compress	75.32	57.35	49.21	49.21	20.96	9.17
12	CSV	91.89	176.92	75.78	70.45	0	0
13	Gson	70.56	56.06	60.61	60.61	12.23	3.44
14	JacksonCore	79.7	24.74	63.28	63.28	4.08	5.13
15	JacksonDatabind	74.15	34.97	65.37	60.78	11.06	15.08
16	JacksonXml	35.1	3.854	33.569	33.569	−1.87	−5.51
17	Jsoup	106.6	22.51	32.5	32.5	17.88	6.4
18	JxPath	81.23	36.37	64.88	9.92	3.57	2.49
19	Lang	68.65	68.06	53.83	2.77	9.43	15.1
20	Math	74.02	137.34	82.07	23.89	7.88	26.66
21	Mockito	73.32	68.26	33.65	33.65	16.09	7.33
22	Time	99.59	65.48	64.92	56.48	13.72	5.61

Table 14 Total average relative improvement on APFD achieved for each of the project

S.No.	Projects	Total average relative improvement (%)
1	Book	28.53
2	Elevator	48.53
3	JFreeChart	56.55
4	Jtcas	39.21
5	Tritype	36.95
6	Area and Perimeter	52.25
7	Cli	33.95
8	Closure	58.28
9	Codec	43.60
10	Collections	55.9
11	Compress	43.53
12	CSV	69.17
13	Gson	43.91
14	JacksonCore	40.02
15	JacksonDatabind	43.56
16	JacksonXml	16.44
17	Jsoup	30.98
18	JxPath	33.07
19	Lang	36.30
20	Math	58.64
21	Mockito	38.71
22	Time	50.96

average APFD (0.941) and NP showed lowest Average APFD score (0.524). SC and BC prioritization techniques achieved very similar average APFD scores (0.665 & 0.678 respectively). If we arrange them in decreasing order of average APFD scores, the arrangement will be: HDFDC (0.941) > TCP-ANT (0.88) > FPA (0.841) > BC (0.678) > SC (0.665) > Random (0.593) > NP (0.524).

Figure 3 represents the box plot diagram showing the APFD scores achieved on twenty-two java projects by seven different prioritization techniques. As can be seen from Fig. 3, both the outlying and the maximum values of NP and Random are much smaller than corresponding values of other five techniques. TCP-ANT achieved highest maximum value compared to others techniques. Among all the techniques, HDFDC achieved highest median. This implies in most cases HDFDC performs better than other six techniques under consideration. The performance of SC and BC are much similar and they outperform NP and Random with larger median and maximum values. TCP-ANT has shorter interquartile range (IQR) and slightly larger median than FPA that means its performance is more effective and consistent than FPA. HDFDC has the shortest IQR among all the seven techniques. This implies HDFDC showed most consistent performance for all the projects under experimentation.

Table 13 shows relative improvement achieved in APFD score by the HDFDC technique over all the other

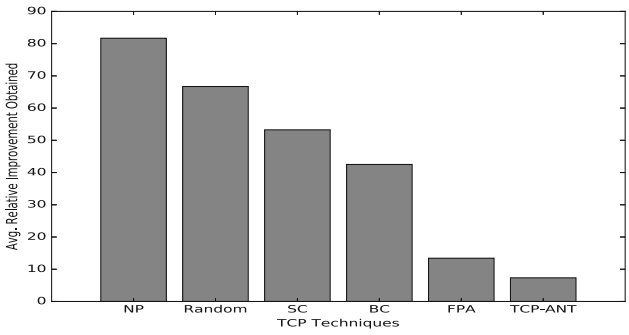


Fig. 4 Average relative improvement achieved on APFD over existing methods using our HDFDC technique

TCP techniques. We can observe that HDFDC shows a significant relative improvement over NP and Random prioritization for all the projects under test. It shows highest relative improvement (176.92%) over Random prioritization for *CSV* project among all other techniques considering all the projects. HDFDC performs more effectively than SC and BC for all twenty-two Java projects. Among SC and BC, HDFDC shows lowest relative improvement (2.77%) over BC for *Lang* project and it achieves same highest relative improvement (83.5%) over both SC and BC for *Closure* project. Whereas, as compared with FPA and TCP-ANT, HDFDC showed positive improvement in relative APFD value for most of the

Table 15 Obtained $APFD_C$ scores

S. No.	Programs	NP	Random	SC	BC	FPA	TCP-ANT	HDFDC
1	Book	0.435	0.564	0.489	0.546	0.722	0.798	0.811
2	Elevator	0.297	0.321	0.398	0.42	0.618	0.696	0.784
3	JFreeChart	0.3	0.296	0.487	0.287	0.656	0.793	0.963
4	Jtcas	0.226	0.274	0.43	0.43	0.805	0.876	0.898
5	Tritype	0.44	0.378	0.522	0.541	0.727	0.794	0.851
6	Area and Perimeter	0.387	0.482	0.428	0.428	0.433	0.62	0.657
7	Cli	0.412	0.559	0.655	0.655	0.828	0.815	0.891
8	Closure	0.426	0.301	0.369	0.269	0.704	0.741	0.950
9	Codec	0.415	0.489	0.344	0.452	0.848	0.856	0.925
10	Collections	0.348	0.289	0.459	0.359	0.986	0.986	0.984
11	Compress	0.395	0.451	0.481	0.481	0.742	0.819	0.936
12	CSV	0.262	0.244	0.408	0.426	0.909	0.909	0.909
13	Gson	0.272	0.406	0.396	0.344	0.857	0.917	0.975
14	JacksonCore	0.459	0.639	0.459	0.359	0.802	0.827	0.949
15	JacksonDatabind	0.456	0.588	0.436	0.436	0.657	0.718	0.961
16	JacksonXml	0.556	0.821	0.541	0.541	0.931	0.987	0.943
17	Jsoup	0.468	0.760	0.473	0.554	0.76	0.878	0.938
18	JxPath	0.434	0.554	0.389	0.584	0.806	0.899	0.971
19	Lang	0.346	0.298	0.489	0.586	0.704	0.646	0.908
20	Math	0.388	0.310	0.352	0.447	0.647	0.611	0.957
21	Mockito	0.332	0.265	0.506	0.506	0.773	0.807	0.941
22	Time	0.31	0.329	0.426	0.522	0.829	0.862	0.945

projects except for *CSV* and *JacksonXml*. HDFDC showed no improvement (0%) over FPA and TCP-ANT for *CSV*. On the other hand, for *JacksonXml* project, as compared with FPA and TCP-ANT, HDFDC showed deterioration in relative APFD value by −1.87% and −5.51% respectively. HDFDC showed less than 1% relative improvement over FPA and TCP-ANT for *Collections*, *CSV*, and *JacksonXml* projects. One of the reasons for this low effectiveness of HDFDC for these specific projects is that it adds most of the test cases in *Group 2*. One of the observations can be drawn from here is that HDFDC performs more effectively when there are more test cases in the *Group 1*. In other words, when there are more numbers of 'hard to detect' faults in the program, HDFDC shows a significant improvement in early fault detection.

Table 14 shows total average relative improvement by the HDFDC technique over all other techniques for each of the project. It can be observed that HDFDC showed the highest total average relative improvement (69.17%) for *CSV* project and the lowest total average relative improvement (16.44%) for *JacksonXML*. It can be observed from Table 13, for *CSV* project, HDFDC showed significant relative improvement over NP, Random, SC, and BC that leads HDFDC to achieved highest total average relative improvement for *CSV* project. On the other hand, for *JacksonXML*, HDFDC showed deterioration in relative improvement over FPA and TCP-ANT. That is the main reason for HDFDC to achieved lowest total average relative improvement for *JacksonXML* project. Another observation can be made from Table 14, for *Math* and *Closure* HDFDC achieved 58.64% and 58.28% total average relative improvement respectively which are the second and third highest total average relative improvement among all the projects. Now, it can be observed from Table 10, *Math* and *Closure* has highest number of lines of code (LOC) compared to other projects. So, one of the observations can be drawn from this is that our HDFDC technique performs better for large projects.

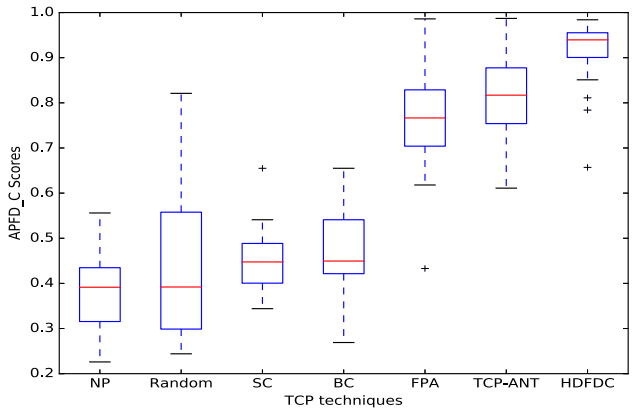


Fig. 5 Box plot representation of $APFD_C$ achieved by different TCP techniques

Figure 4 represents the average relative improvement achieved by the HDFDC technique over NP, Random, SC, BC, FPA, and TCP-ANT techniques. Among this six techniques, HDFDC showed the highest average relative improvement over NP and the lowest average relative improvement over TCP-ANT. On an average, it shows relative improvement in APFD value by 73.85% than NP and Random. On the other hand, HDFDC shows an average relative increment in APFD value by 28.81% over SC, BC, FPA, and TCP-ANT. On an average HDFDC shows 43.82% average relative improvement over the existing TCP techniques. This improvement may be because our technique considers different types of bugs based on their difficulty of detection. So, it may lead to a situation where the test cases that detect 'hard to detect' bugs have a higher probability of revealing more faults than the test cases that detect normal bugs.

We have also conducted experimental evaluation based on $APFD_C$ achieved by different TCP techniques of our consideration. $APFD_C$ achieved by different techniques shown in Table 15. From the Table 15, we can observed that No Prioritization (NP) showed lowest $APFD_C$ score (0.226) for *Jtcas* followed by Random prioritization which showed second lowest $APFD_C$ score (0.244) for *CSV* Project. NP and Random showed poor performance compared to other techniques in terms of $APFD_C$. Like our previous *APFD* scores analysis, SC and BC perform very similar on $APFD_C$ also. TCP-ANT outperformed FPA for more most of the projects and $TCP-ANT$ achieved highest $APFD_C$ (0.987) for *JacksonXml* project among all the techniques. Our technique *HDFDC* showed lowest $APFD_C$ (0.657) for *Area* and *Perimeter* and highest $APFD_C$ (0.984) for *Collections* project. In our previous analysis on APFD, we found that there also HDFDC achieved lowest and highest *APFD* scores for the same projects *Area* and *Perimeter* and *Collection* respectively. For most of the considered projects, HDFDC obtained higher $APFD_C$ than other techniques. This established that HDFDC helps detect more severe faults earlier in a regression testing cycle.

Figure 5 shows the box plot representation of $APFD_C$ scores achieved by seven different techniques. As can be seen from Fig. 5, Random technique has largest IQR than other techniques that means Random showed more variance in $APFD_C$ scores than all other techniques under consideration. SC showed shorter IQR than BC than means in terms of $APFD_C$ scores, SC showed more consistency than BC. FPA and TCP-ANT has very similar IQR but TCP-ANT showed slightly larger median than FPA. Like box plot representation of APFD scores (Fig. 3), here in $APFD_c$ box plot also HDFDC has shortest IQR and largest median among all the seven techniques which implies HDFDC showed most consistent performance and more effective than other techniques in terms of cost and severe faults detection.

4.6 Threats to the validity

- Our experiments have been carried out using a limited set of data. However, to mitigate this risk, we have considered several programs with widely varying LOCs, complexity, and specifications.
- We have injected only basic programming bugs using the mutation tool. We have not injected higher-order bugs such as algorithmic errors. However, it has been reported that basic programming bugs can lead to higher order mutants Nguyen and Madeyski (2014).
- In our experimentation, every subject program contains exactly one seeded fault. On the other hand, in practice programs can have more than one bug. However, since debugging is carried out by considering one test failure at a time, we argue that our approach does not compromise the prioritization of the test cases.
- We have considered the test cases which are available in Defects4j Just et al. (2014) and SIR Rothermel et al. repository for any particular project. Results might differ for other test cases. However, the two considered repositories contain programs that are realistic and have widely differing characteristics.

5 Conclusion

We have proposed a TCP technique based on the capability of test cases to detect 'hard to detect' bugs. Our TCP algorithm prioritized the test cases in three different groups based on 'hard to detect' bugs and overlapped bugs. For this, we have proposed a metric called Fault Detection Capability (FDC) to measure the bug detection capability of test cases. We have compared the effectiveness of our proposed HDFDC technique with existing prioritization techniques. Our experimental results show that on an average, our HDFDC technique achieves 73.85% relative improvement over No Prioritization (NP) and Random Prioritization, 52.88% relative improvement of APFD over statement coverage (SC), 42.13% than branch coverage (BC) based techniques, and on an average 10.11% relative improvement over Flower Pollination Algorithm (FPA) based TCP technique, and Test Case Prioritization technique based on ANT colony optimization algorithm (TCP-ANT). Our HDFDC technique also shows a significant improvement on $APFD_C$ over other existing TCP techniques. That implies HDFDC is effective in terms of cost and severe faults detection. Another advantage of our

technique over existing techniques is that our works not required prepossessing steps like code instrumentation, coverage analysis etc.

We have considered class-level and method-level mutation operators for seeding bugs into a program. In future, we plan to extend our approach by injecting other types of bugs such as inter-class mutants. At present, we have considered only single bug per program. But, in practice, a program may contain multiple faults. We are planning to handle programs with multiple faults as our future work.

Funding No funding used.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Busjaeger B, Xie T (2016) Learning for test prioritization: an industrial case study, in: Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering, pp. 975–980

Dhareula P, Ganpati A (2020) Flower pollination algorithm for test case prioritization in regression testing. In: ICT Analysis and Applications, Springer 155–167

Dutta A, Kumar S, Godboley S (2019) Enhancing test cases generated by concolic testing. In: Proceedings of the 12th Innovations on Software Engineering Conference (formerly known as India Software Engineering Conference), pp. 1–11

Elbaum S, Malishevsky AG, Rothermel G (2002) Test case prioritization: a family of empirical studies. IEEE Transact Softw Eng 28(2):159–182. https://doi.org/10.1109/32.988497

Elbaum S, Rothermel G, Kanduri S, Malishevsky AG (2004) Selecting a cost-effective test case prioritization technique. Softw Qual J 12(3):185–210. https://doi.org/10.1023/B:SQJO.0000034708.84524.22

Elbaum S, Malishevsky A, Rothermel G (2001) Incorporating varying test costs and fault severities into test case prioritization. In: Proceedings of the 23rd International Conference on Software Engineering. ICSE 2001, pp. 329–338. https://doi.org/10.1109/ICSE.2001.919106

Gligoric M, Groce A, Zhang C, Sharma R, Alipour MA, Marinov D (2015) Guidelines for coverage-based comparisons of non-adequate test suites. ACM Transact Softw Eng Methodol (TOSEM) 24(4):1–33

Gökçe N, Eminov M, Belli F (2006) Coverage-based, prioritized testing using neural network clustering. In: International Symposium on Computer and Information Sciences, Springer, pp. 1060–1071

Gregg R, Mary Jean H, Dedhia, (1999) Regression test selection for c++ software. Tech rep, USA

Harrold MJ, Jones JA, Li T, Liang D, Orso A, Pennings M, Sinha S, Spoon SA, Gujarathi A (2001) Regression test selection for java software. SIGPLAN Not 36(11):312–326. https://doi.org/10.1145/504311.504305

Jones JA, Harrold MJ (2003) Test-suite reduction and prioritization for modified condition/decision coverage. IEEE Trans Software Eng 29(3):195–209

Just R, Jalali D, Ernst MD (2014) Defects4J: A Database of existing faults to enable controlled testing studies for Java programs. In: ISSTA 2014, Proceedings of the 2014 International Symposium on Software Testing and Analysis, San Jose, CA, USA, pp. 437–440, tool demo

Korel B, Tahat LH, Harman M (2005) Test prioritization using system models. In: 21st IEEE International Conference on Software Maintenance (ICSM'05), 2005, pp. 559–568. https://doi.org/10.1109/ICSM.2005.87

Krishnamoorthi R, Sahaaya Arul Mary SA (2009) Factor oriented requirement coverage based system test case prioritization of new and regression test cases. Inf Softw Technol 51(4):799–808. https://doi.org/10.1016/j.infsof.2008.08.007

Larsen L, Harrold MJ (1996) Slicing object-oriented software. In: Proceedings of the 18th International Conference on Software Engineering, ICSE '96, IEEE Computer Society, Washington, DC, USA, pp. 495–505. http://dl.acm.org/citation.cfm?id=227726.227837

Lou Y, Hao D, Zhang L (2015) Mutation-based test-case prioritization in software evolution. In: IEEE 26th International Symposium on Software Reliability Engineering (ISSRE). IEEE 2015:46–57

Malishevsky AG, Ruthruff JR, Rothermel G, Elbaum S (2006) Cost-cognizant test case prioritization. Tech rep

Mall R (2018) Fundamentals of software engineering. PHI Learning Pvt. Ltd., Delhi

Mirarab S, Tahvildari L (2008) An empirical study on bayesian network-based approach for test case prioritization. In: 2008 1st International Conference on Software Testing, Verification, and Validation, IEEE, pp. 278–287

muJava (2016) https://cs.gmu.edu/offutt/mujava/

Nguyen Q-V, Madeyski L (2014) Problems of mutation testing and higher order mutation testing. Adv Intell Syst Comput 282:157–172. https://doi.org/10.1007/978-3-319-06569-4_12

Panda N, Acharya AA, Mohapatra D (2019) Test scenario prioritization for object-oriented systems using uml diagram. Int J Syst Assur Eng Manag. https://doi.org/10.1007/s13198-019-00759-z

Panigrahi CR, Mall R (2010) Model-based regression test case prioritization. In: Prasad SK, Vin HM, Sahni S, Jaiswal MP, Thipakorn B (eds) Information systems. Technology and Management, Springer, Berlin Heidelberg, Berlin, Heidelberg, pp 380–385

Panigrahi CR, Mall R (2013) An approach to prioritize the regression test cases of object-oriented programs. CSI Transact ICT 1(2):159–173. https://doi.org/10.1007/s40012-013-0011-7

Panigrahi CR, Mall R (2013) A heuristic-based regression test case prioritization approach for object-oriented programs. Innovations Syst Softw Eng 10:155–163

Rothermel Harrold (1994) Selecting regression tests for object-oriented software. In: Proceedings 1994 International Conference on Software Maintenance, pp. 14–25. https://doi.org/10.1109/ICSM.1994.336793

Rothermel G, Elbaum S, Kinneer A, Do H, Software-artifact infrastructure repository, http://sir.unl.edu/portal

Rothermel G, Untch RH, Chengyun Chu, Harrold MJ (1999) Test case prioritization: an empirical study. In: Proceedings IEEE International Conference on Software Maintenance - 1999 (ICSM'99). Software Maintenance for Business Change (Cat. No.99CB36360), pp. 179–188

Rothermel G, Untch RH, C Chengyun, Harrold MJ (2001) Prioritizing test cases for regression testing. IEEE Transact Softw Eng 27(10):929–948. https://doi.org/10.1109/32.962562

SIR (2001)https://sir.csc.ncsu.edu/portal/index.php

Spieker H, Gotlieb A, Marijan D, Mossige M, Reinforcement learning for automatic test case prioritization and selection in continuous integration, arXiv preprint arXiv:1811.04122

Srikanth H, Williams L, Osborne J (2005) System test case prioritization of new and regression test cases. In: 2005 International Symposium on Empirical Software Engineering, 2005., p. 10. https://doi.org/10.1109/ISESE.2005.1541815

Tan P-N, Steinbach M, Kumar V (2016) Introduction to data mining. Pearson Education, India

Vescan A, Pintea C.-M, Pop PC, Test case prioritization–ant algorithm with faults severity, Logic Journal of the IGPL

Zhao X, Wang Z, Fan X, Wang Z (2015) A clustering-bayesian network based approach for test case prioritization. In: 2015 IEEE 39th Annual Computer Software and Applications Conference, Vol. 3, IEEE, pp. 542–547

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

Available online on 15.09.2021 at http://jddtonline.info

Journal of Drug Delivery and Therapeutics

Open Access to Pharmaceutical and Medical Research

Copyright © 2021 The Author(s): This is an open-access article distributed under the terms of the CC BY-NC 4.0 which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited

Open Access Full Text Article

Review Article

A Review on Protein Misfolding Diseases and therapeutics for preventing Diseases

Yogesh Kumar Singh*

Dept. of Biotechnology, Delhi Technological University (DTU) New Delhi 110042 India

Article Info:

Article History:

Received 11 July 2021
Reviewed 27 August 2021
Accepted 09 Sep 2021
Published 15 Sep 2021

Cite this article as:

Singh YK, A Review on Protein Misfolding Diseases and therapeutics for preventing Diseases , Journal of Drug Delivery and Therapeutics. 2021; 11(5):221-223

DOI: http://dx.doi.org/10.22270/jddt.v11i5.5047

*Address for Correspondence:

Yogesh Kumar Singh, Dept. **o**f Biotechnology, Delhi Technological University (DTU) New Delhi 110042 India

Abstract

Background: Protein misfolding sicknesses are the gathering of irresistible lethal neuro and non-neurodegenerative infections and in ebb and flow researchers and specialists accepted that unusual folding of protein is the essential or main key of such illnesses are Alzheimer's infections, Parkinson's diseases, Huntington's sickness, Creutzfeldt-Jakob infection, cystic fibrosis, Gaucher's infection and numerous other degenerative and neurodegenerative problems. The motive of this review article is to gave a detailed of the existing structural information for prion and prion protein and also we will trying to find out their preventing root causes with respect to structural information of prions within the context of what is known about the protein misfolding diseases.

Objective: This article presents a brief overview of research on the use of these therapeutics for the treatment or improvement in prion diseases or protein misfolding.

Material and Methods: This article begins with the brief introduction about protein misfolding diseases or infections and the therapeutic materials which are used in researches or explain this article (pentosan polysulfate, Quinacrine, Doxycycline, Chaperone based therapy, Resveratrol and curcumin) etc.

Results and Conclusions: In this present context of protein misfolding/prion diseases diagonsis.Therapeutic approaches predicts that person infected with prion diseases prolongs the survival time of the patient and improvement in the conditions of the prion diseased infected patient which provides good result for future medicine development.

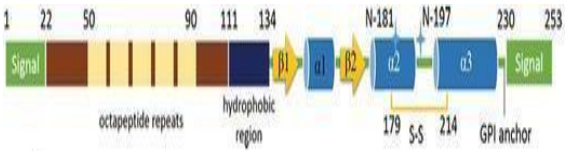
Keywords: Amyloid, Beta-sheet, neurodegenerative, prion, protein misfolding, therapeutics etc

Introduction

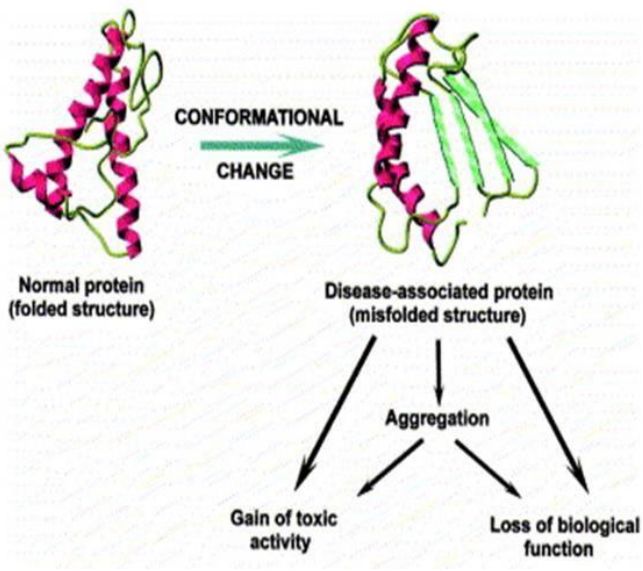
Protein misfolding infections otherwise called contagious spongiform encephalopathies (TSEs or prion sicknesses) or can says amyloid illnesses, are a gathering of lethal sicknesses that influence the cerebrum and sensory system of human and animals. unusual folding(misfolding) of proteins currently stay mutually to form round "oligomers", whose subunits consist of the unusual folding(misfolding) prions. The gathering of such protein oligomers in nerve cells is the basic key of neurodegenerative side effects and eventually, demise, in patients with PR(prion) illnesses. An abundance of natural and biophysical proof currently proposes that the atomic reason for prion illnesses might be encoded by protein conformation. Prions are proteinaceous irresistible particle which is created generally, if not altogether, of a strange isotype of the PrP(prion protein) assigned, on account of scrapie, prion protein scarpie [1]. Prion grounds four neuro-degenerative infections of people and six of creatures, together with scrapie of sheep and ox-like *spongiform encephalopathy* [1]. The human prions sicknesses show as irresistible, ancestral, and inconsistent problems represented a puzzler until it was found that changes in the PrP quality are hereditarily connected to advancement of neurodegeneration.

Prion and Prion Protein

Prion protein is present in two forms first is normal form and other one is folded protein which is a type of abnormal fold. Prion sicknesses can influence the two people and animals and are at times spread to people by contaminated meat items. The most common form of prion disease that affects humans is Creutzfeldt- Jakob disease (CJD).Typical (PrP^C) cellular prion protein is an N-allied glycoprotein positioned on the external membranes by means of a glycophosphatidyl-inositol affix [**show in fig-1]**, broadly spoken in a variety of tissues and enrich in the central nervous system [2]



(Fig-1) [1]



(Fig-2) [3]

Conformational Change in Prion Protein (PRP^{C} TO PRP^{SC} Conversion)

PrP^{c}(cellular prion protein) is a dispersible monomeric protein that is eagerly degraded by proteinase K, in comparison, prion protein scarpie (PrP^{Sc}) form unsolvable aggregate that predicts property of resistance to proteinase K absorption. Conformational change of α-helix loaded prion protein scarpie to unusual folded(misfolded) β-sheet-loaded prion protein scarpie, account for the growth of prion infections [2]. Prion protein scarpie is synthesize from the typical cellular isotype cellular prion protein by a post-translational method that most likely occur in endosomes [1]. Endeavors to recognize a post-translational compound change that highlights in the transformation of cellular prion protein into prion protein scarpie have been ineffective.[1]All discoveries raise the likelihood that prion protein scarpie contrasts from cellular prion protein just concerning adaptation [1]. As per Fourier-change infrared spectroscopy (FTIR) showed that cellular prion protein has a elevated α-helix content 42% and no β-sheet (3%only), discoveries that were affirmed by round about dichroism estimations. In difference, the β-sheet content of prion protein scarpie was 43% and the α-helix 30% as calculated by fourier-change infrared spectroscopy [1].

Amyloid Fibrills

Amyloid refers to the irregular fibrous, extracellular, proteinaceous deposit originates in organs and tissues. Amyloid is unsolvable and is fundamentally overwhelmed by β-sheet form. Dissimilar to further fibrous proteins it doesn't usually have an underlying, strong or motility job however is related with the pathology found in a scope of infections called as the *amyloidosis*. Many researchers finding and recommended that amyloid is fibrillous in[4] structure this means it consist of linear biopolymer, and are characterize by rod-like form with high length-to-diameter ratios. Oftentimes, they suddenly position into helical forms. Amyloid fibrills are occurred by usually dissolvable proteins, which bring together to form unsolvable fibres that are opposed to proteinase K degrade [4]. Their arrangement can go with illness and every sickness is described by a particular protein or peptide that totals (prion protein) [4]. Notable instances of amyloid sicknesses incorporate alzheimer's infection, diabetes type 2 and the spongiform encephalopathies (e.g., frantic cow illness) [4]. His amyloid fibrills are kept extracellularly in the tissues and are attention to have [4] an infectious impact. The fibrillar congregations are characteristically steady and underlying examinations have uncovered that they are made dominatingly out of β-sheet from in a trade name cross-beta [4] conformity. As of late, various instances of practical amyloid has been recognized as well as a composition of melanosomes, curli and hydrophobins[4].

Approaches for treatment of prion or *prion disease*

Pentosan Polysulphate(PPS) is a type of polyglycoside [5] chemical compound having low heparin-like [5] work. Pentosan polysulphate acts as a co-receptor for prion type of protein(PrP) on cell area [5] when compared to interior heparin sulfate proteoglycans and this PPS predicts the better capacity to stop the making of newly maked PrP^{sc} in neuroblastoma cells [5] or cancerous cell(tumor cell). In an experiment ,the mice is taken and injects with PPS in the cerebral ventricles in prion diseased mouse, PPS shows activity of [5] increment or increase the lag phase of diseased mice. After treatment with PPS their is no significance or no reversal of foregoing neurological deficit are experiential [5].

Quinacrine is a derivative of acridine that shows the activity of inhibition in PrP^{sc} occurance in scarpie-infected neuroblastoma cells. Many studies predicts that use of quinacrine results in increase the incubation time and survival time of the infected mice, which are infected with different scarpie strains, but, there is no specific reason for this, the particular mechanism is controversial. Quinacrine stabilizes the PrP^{c} that prevent PrP^{sc} polymerization and then reduces the conversion to PrP^{c} to PrP^{sc}. Practically in a experiment with a patient having CJD, the 300mg quinacrine is administered in patient; after 300mg dose of quinacrine the symptoms of disease slowly deteriorated. Then, again some quantity of quinacrine is given to patient continuously for 3 weeks. However, subsequent testing of quinacrine and observational data has predict changes in the survival of patient with CJD.

In CJD brains PrP^{sc} is extracted then, **Doxycycline** is treated with CJD, which shows the activity of U-turn of protease resistance of PrP^{sc} which then, prolongs the survival time when experimentally done in animals which are infected with prions. Doxycycline shows the property of positive kinetics, good ability to eliminate the blood brain barriers and low deadly property [5], when administered. Many researches describe that patients with CJD received the doxycycline, which results that, exhibit longer survival time, when compared to other patient which are not receive doxycycline. But, during doxycycline is orally administer in patients (100mg) does not predicts or profit the continued existence of patient with creutzfeldt-jakob disease [5].

Iron Tetrapyrole Derivative Fe(III) – TMPyP interact with the PrP^{c} and Fe(III0- TMPyP works just like therapeutic chaperone for prion protein cellular,that reduces the ground state of the occupant conformation and [5] shows the activity of inhibition in prion- induced unusual folding in vitro. Whereas, iron tetrapyrole derivative Fe(III) and its strongly same type of porphyrins predicts the increase in survival activity in prion-diseased mice [5]. Fe(III)- TMPyP combines to PrP^{c} produces the two type of effects [5].(Blocks the prion duplication and also stops the activity of PrP intervened toxicity) [5].

Another approach for stops or reduces the fibrils formation and aggregation in prion diseases or neurodegenerative diseases are developing chaperone based therapeutics. In an

experimental setup, this is found that the **Yeast Disaggregation HsP104** contains [6] a activity of fibrils formed dissolve type of activity from a different type of neurodegenerative disease proteins, and it also predicts the same activity in other type of fibrils or aggregation compounds; such as tau, polyglutamine, A-beta42, alpha-synuclein and prion protein, in another case high level or high concentration of HsP104 is required to liquefy these proteins; some change in HsP104 series [6] shows high level of disaggregation activity and slows the toxicity level. Other chaperone which are used as a good approach for this type of neurodegerative diseases (i.e A-beta42,tau). This chaperone approach include yeast chaperone Tric, which shows protectiveactivity against Htt toxicity and another one chaperone is metazoan chaperone which has the property of disaggregation(e.g HsP110.HsP70 and HsP40).

Some of the naturally occurring plant based compounds shows promising effect to reduce/eliminate the prion diseases. **Resveratrol** a type of polyphenol compound that found naturally in foods/fruits, such as blueberries, peanuts, grapes and [5] red wine. Resveratrol shows many potential activities such as antioxidant, anti-allergic and neuro-protective activities, and in some researches it is predict/discover that resveratrol reduces prion duplication and prion disease in vivo [5]. Also it shows that they clears/eliminate PrP^{sc} by reducing endogenous sirtuin 1levels in brain. In different researches it is showed that resveratrol act together with prion and disruption of prion fibrils formation, it too steady the prion protein peptide during pi-pi stacking interaction amid resveratrol and Tyr128. And the bonding linkage of hydrogen of resveratrol and prion protein peptide additionally reduces the suppleness of the peptide and its vulnerability to the aggregates [5].

Curcumin yellowish coloured present or draw out from turmeric rhizomes, this plant belongs to the family of ginger [5]. More studies set up that curcumin stop the anti-inflammatory work, antioxidant and antitumor properties, also curcumin shows the neuro safety property by altering themitochondrial assimilation and ease unusual free-radical stress [5]. Curcumin in less quantity of doses effectively enhance the free-radicle stress reaction in cells, it also exert antiprion property by stopping of amyloid occurred by prion [5]. Curcumin increment the quantity of oligomer and holdup the occurrence time of the birth of fibers, also affects the change of prion protein cellular to prion protein scarpie by reduces the birth rate also yields of aggregation in prions [5].

A hormone a type Neuroendocrine hormone [5] called as **Melatonin** which is excreted by pineal gland. When melatonin directly combine with free radicals it shows the antioxidant effect and also this blocks the sequence of chain [5] reaction of superoxide oxidiation and also by enhancing the property of antioxidant macromolecules [5]. Melatonin shows effective property for patients with fatal familiar insomnia, in different words it is beneficial in case of insomnia.

Gallic Acid (polyphenol) in green tea extracts gallocathechin gallate, it shows the clear activity of reducing or clears the PrP^{sc} from ScNza cells. But, in another case it hinders with simple PrP^{c} utterance in undiseased cells [5].

Flupirtine Maleate acts centrally, monopioid analgesic that exhibits cytoprotective action in opposition to apoptotic assistant in brain neuron in vivo or in artificial environment [5]. It is occured by PrP106-126 type of different piece like structure which guide toxicity in neurons, increasingly slows the neurotoxicity in the availability of this fragment, in double blind placebo prohibited study of CJD, this chemical molecule shows well organized against cognitive degradation. In another words it does not show any other significant effecton other neurological symptoms, or survival.

Ubiquitin-Specific Protease 14 recent research work predicts that UPS are concerned in quality assurance of [7] PrP^{c}. we expose the implication of PrP ubiqitination, so firstly paying attention on ubiquitin-specific protease 14 (USP14), a type of de-ubiqutinating enzyme [7]which catalyze adornment of polyubiquitin sequence and play as a function in direction of proteasomal progress. Overexpressed of the leading negative deviant forms of USP14 reduces PrP^{sc}, whereas in case of wild type USP14 increases PrP^{sc} in prion-diseased cells. All datas propose that USP14 stops the detoriation of simple and abnormal forms of Prion protein. jointly a well-organized considerate about the directive of PrP^{sc} clearance occured by USP14 may gave highly to the expansion of therapeutic approach for prion's type of infections [7].

Conclusion

In this review we are suggested some types of chemical compounds and enzyme which degrade or reduces the prion protein scarpie formation which causes amyloid fibril formation in neurons. But, no promising treatment provides for prion diseases. some chemical compounds or enzymes shows some type of prevention property for prion diseases, they can reduce or decrease the level of PrP^{sc} formation at some levels. so, we can suggest that by use of these compounds which we are explain in this review shows promising results to prevent many prion disease or misfolding of proteins which leads to amyloid formation.

References

1. Zhou, Z., & Xiao, G. Conformational conversion of prion protein in prion diseases. Acta Biochim Biophys Sin, 2013; 45(6):465-476. https://doi.org/10.1093/abbs/gmt027
2. Pan, K. M., Baldwin, M., Nguyen, J., Gasset, M., Serban, A. N. A., Groth, D., ... & Cohen, F. E. Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. Proceedings of the National Academy of Sciences, 1993; 90(23):10962-10966. https://doi.org/10.1073/pnas.90.23.10962
3. Scheinost, J. C., Boldt, G. E., & Wentworth, P. Protein Misfolding and Disease. Chemical Biology: Approaches to Drug Discovery and Development to Targeting Disease, 2012; 379-400. https://doi.org/10.1002/9781118435762.ch19
4. Rambaran, R. N., & Serpell, L. C. Amyloid fibrils: abnormal protein assembly. Prion, 2008; 2(3):112-117. https://doi.org/10.4161/pri.2.3.7488
5. Chen, C., & Dong, X. Therapeutic implications of prion diseases. Biosafety and Health. 2020.
6. Sweeney, P., Park, H., Baumann, M., Dunlop, J., Frydman, J., Kopito, R., Hodgson, R. Protein misfolding in neurodegenerative diseases: implications and strategies. Translational neurodegeneration, 2017; 6(1):1-13. https://doi.org/10.1186/s40035-017-0077-5
7. Nargeh, H., Aliabadi, F., Ajami, M., & Pazoki-Toroudi, H. Role of Polyphenols on Gut Microbiota and the Ubiquitin-Proteasome System in Neurodegenerative Diseases. Journal of Agricultural and Food Chemistry. 2021. https://doi.org/10.1021/acs.jafc.1c00923

2021 8th International Conference on Smart Computing and Communications (ICSCC)

Analysing and Identifying Crucial Evidences for the prediction of False Information proliferated during COVID-19 Outbreak: A Case Study

Deepika Varshney[1] and Dinesh Kumar Vishwakarma[1]

[1]Department of Information Technology, Delhi Technological University, New Delhi, India

{deepikavarshney_phdit2k18,dinesh}@dtu.ac.in

2021 8th International Conference on Smart Computing and Communications (ICSCC) | 978-1-7281-9687-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSCC51209.2021.9528205

***Abstract*—In the current scenario social media platforms are one the efficient way to share opinions and thoughts of an individual. User can freely share their thoughts on an event/ situation. This can be a curse for the society if social media platform is utilized with some bad intention to spread false information and create chaos/ confusion among public which greatly degrades user experience. In the current pandemic many people have their eye on any news article related to corona cure. Malicious users take this as an opportunity to spread fake news in order to create confusion among public or some monetary benefits, the detection of which is of paramount importance. The proposed technique is leverages to learn crucial evidences based on Context Knowledge, Distance Metric and Word Resemblance with respect to news article headline and its content concerning top 10 google search results related to the claim, where considering COVID-19 as one of the special case studies from the application perspective. This paper proposed a novel scheme for the prediction of false information and generated a covid-fakenews dataset that further be utilized for the analysis and evaluation of our model. The results reveals that the proposed intelligent strategy gives promising experimental results and quite effective in predicting False information.**

***Index Terms*—COVID-19, Information Pollution, Fake News Detection.**

I. Introduction

In the recent scenario, a new coronavirus disease spread around the world. The disease emerges as a respiratory infection with indicative concern for global public health hazards. Initially, it is suspected that the disease is transmitted from animal to individual, later the paradigm is shifted that the infection is transmitted towards individual to individual via droplets, close contacts creating huge panic with approximate 6,359,182 confirmed cases and 380,663 deaths have been encountered till now and growth rate is still high which has alarmed the global authorities including world health organization (WHO) [1]. The pandemic affects worldwide badly, however some of the users take this crisis as an opportunity for doing some malicious activities. During this crisis lot of health related misleading information related to fake cures are suggested for COVID-19 has been posted to create chaos and confusion/misconceptions about the disease. During this pandemic, people have their eye on any new announcement from the government official or some news that can help to get rid of COVID-19. As the disease is deadly, the people are also desperate to know some cure and in rush to find a treatment for the new coronavirus disease. Some of the fake cures posted on social media are harmful too and give bad health advice. The Fig.1 shows an image with the false claim reported that the silver solution can kill coronavirus within 12 hours, however, there is no significant evidence has been found concerning to this claim.The circulation of such misleading information has led to dire need for a system that can differentiate fake from real. Most of the previous studies reported approaches to detect misleading information based on multimedia data (Text, Images and Videos) [2],[3]. The methods based on text are mainly use text patterns and match them with already existing patterns of fake news. Like text,

4. US authorities have been cracking down on fake cures for COVID-19, the disease caused by the novel coronavirus — including colloidal silver, vitamins, teas, and essential oils. None of these are approved treatments or preventative measures against the disease.



Fig. 1: Examples of Misinformation.

many researchers are keen towards extracting useful evidences from images and videos for detecting misleading content on social media [2], [3], [4], [5]. In our work we have employed three set of novel evidences based on Context Knowledge, Distance Metric and Word Resemblance with respect to news article headline and its content for the detection of misleading information. Here, we have provided a case study on detecting misleading information during covid-19 pandemic as well as

978-1-7281-9687-9/21/$31.00 © 2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:10:23 UTC from IEEE Xplore. Restrictions apply.

built a self-generated dataset related to fake news spreaded during this period. One of the major lack in this case that there is no dataset concerning to corona fake news that is publically available to test our proposed algorithm.One of the novelty and contribution of the work is we have built a dataset with 335 claim samples having 168 fake and 167 real news related to fake cure spreaded during COVID-19 outbreak, we are also working to expand the given dataset.

The remaining of the paper is organized as follows. In Section II, we are going to discuss the states-of-the-art techniques to detect misleading information, whereas the detailed description of data collection process and the method adopted for the prediction of misleading information is given in Section III, which followed by a discussion of experimental results in Section IV and lastly, the paper is concluded with some suggested future work aspects.

II. Related Work

In the current era due to widespread use of social media, spreading of false information is one of the crucial problem nowadays, where discriminating fake news from the real is one of the challenging task for the users. Most of the earlier studies considered misleading information detection as a task of classification intending to associate label as true or false and from survey analysis, it has been observed that the classification approaches are turn segregated into methods based on machine and deep learning. From various studies, it has been proven that machine learning approach are extremely useful in countering numerous problem in engineering field. In particular, various different types of machine learning methods based on supervised learning strategy are applied to detect misleading/fake information, where SVM (support vector machine) is one of the prominently used approach for classification. The authors of [6], employed a graph-kernel-based SVM classifier for the detection of rumors and provided an accuracy of 0.91 on sina-weibo dataset. Whereas, in [7], the author reported a crucial feature set to distinguish news from fake, real and satire. The SVMs are also employed for clickbait detection in [8]. Like SVM, the random forest has also been exploited in numerous works for fake news and rumor detection. Most of the studies have reported random forests as a strong performer among other machine learning algorithms [9], [10], [11], [12], [13]. The random forest has also been used for stance detection in [9],[14]. The authors of [15], incorporated and analyse tweets related to 5g conspiracy and COVID-19 theories to detect sources of spreading misleading information. Whereas, in [16], the authors proposed a two stage pipeline for COVID-19 misinformation prediction. The authors of [17], incorporated natural language processing(NLP) and ontology reasoning to indicate false information related to Covid-19. Whereas, the authors of [18], proposed three set of features based on news-article content, sentiment and mixtures of both as hybrid features for the detection of false information, where from the analysis it has been found that random forest classifier gives the best result with an accuracy of 95%. Whereas, in [19], authors incorporated three set of features based on user-profiling, video-content and human consensus to retrieve significant evidences for detection of clickbaitâs videos. In [20], authors have proposed a novel RNN architectures, namely tanh-RNN, LSTM, and Gated Recurrent Unit(GRU) for the detection of rumors. Whereas, in [21], the author has proposed a multi-task learning approach and designed a multi-task learning framework with an LSTM layer shared among all tasks to counter the problem of rumor classification. From the study, it has been observed that most of the recent work has explored the combination of RNN and CNN in their model [22],[23],[24].The authors of [25], provided a study on fake news prediction during COVID-19 using the uses and gratification architecture, which further be enhanced and extended by an altruism motivation and it has been reported that it is one of the efficient and significant factor for the prediction of fake news sharing during COVID-19. Whereas, the author of [26], build a COVID-19 fake news data set and propose a semi- supervised end-to-end neural attention model(Cross-SEAN) that largely rely on unlabeled data. The propose technique outperform other baselines with an F1-SCORE of 0.95. In the previous studies, it has been observed that very few works have been reported the important clues concerning news article headlines and news article content based on Context, Distance-Metric, and Word resemblance (paragraph based) for the prediction of misleading information that are found to be efficient and gives novelty to our proposed work.

III. Methodology to Detect Misleading Information

This paper considered a binary class classification problem. We assume that the posted source claim c=c1,c2,c3âŠck can be divided into two classes Class=M,R: 1) Real (R), namely the posted claim faithfully represents the event that it refers to, 2) Misleading (M), namely the posted claim does not truly shows the event that it refers to. The claims that have been included here are related to coronavirus COVID-19. Take an example of the post concerning to coronavirus. The post that âchlorine and alcohol products cannot kill viruses within the bodyâ is true information and belongs to the real class since many authoritative and authentic news sources have noted and reported significant news and has also been acknowledged by WHO [1]. While one of the posts says that âcoronavirus is caused by 5G technologyâ is absolutely a misleading information spreaded on web, it not only lacks factual support but also deviates from scientific principles. So, the goal is to learn a classifier from the labeled feature set, that is f: $X^k \rightarrow Y^k$, where Y^k takes one of the two fine-grained classes: R, M. Given the input feature set X^k, the classifier f can output the classification result for the posted claim C^k.

A. Dataset Creation

In the coronavirus outbreak, lots of misleading information was posted by users, and most of the misleading information

[1]https://www.indiatoday.in/world/story/drinking-alcohol-will-not-protect-you-from-covid-19-says-who-1653555-2020-03-08

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:10:23 UTC from IEEE Xplore. Restrictions apply.

related to fake cures, lockdown, and others have gone viral during the period. The posts have gone viral on different social media platforms like Twitter, Facebook, and others.We have collected prominent claims related to corona posted during the period from different web sources (News Media, WHO, Buzzfednews) in the form of images and text [2]. Some of the sources of misleading news are Buzz feed news and the web sources that are reporting myths regarding coronavirus. These set of source provides the latest running list of hoaxes/false information spreading related to coronavirus.Whereas, most of the real claims are retrieved from the WHO website for a more reliable collection. In all total of 335 queries are collected with having 168 misleading and 167 real claims (Covidfakenews 2019).

B. Clues identification from News Article Headlines/Titles

The effective clues are retrieved from news article title/headlines based on three proposed measures. A detailed description of each of these measures have shown below:

1) Word Resemblance: : The first set of evidential clues are collected from word resemblance based evidence, where word similarity ratio is calculated to identify how useful or relevant title is concerning the input query, Word Similarity Ratio (WSR) is the first measure considered for retrieving some important clues from the news headline. This is the ratio of the number of keywords k_i that are found to be similar in news title and the query to the total number of words in the query t_i.

2) Distance-Metric: : This is the second set of evidential clues that have been incorporated to find how apart the two sequences of words. Levenshtein distance ratio has calculated to find the distance ratio concerning (query, title/headline) and (query, article content).

3) Context Knowledge: : The third set of evidential clues is to retrieve context knowledge from the news article titles. In this work, we have incorporated N-Grams (uni-gram, bi-gram, and tri-gram) for context analysis. The set of clues helps in identifying a sequence of phrases that further be used in determining the false context that exists in the news article title.We empirically analyzed titles and find prominent keywords used in false query search responses and build a false phrases dictionary with 20 keywords concerning to unigrams, bigrams, and trigrams. The word cloud representation of the respective keyword is shown in Fig.2.

C. Clues identification from News Article Content

As we have identified the clues in the previous section related to the top 10 news headlines/ titles respective to each query. In the same way, the analysis has also been applied to new article content.New article content is one of the important parts that reveals what the article trying to convey and whether the given content reporting the same thing as it claiming. Analyzing news article content is needful

[2]https://www.buzzfeednews.com/article/janelytvynenko/coronavirus-fake-news-disinformation-rumors-hoaxes



Fig. 2: Word Cloud Representing Key Phrases used for False Information Prediction.

to extract possible clues to predict whether the query is misleading or not. The first 8 paragraph ($< p >$) tags of each article are extracted.Each paragraph is inspected concerning two clue measures Distance Metric and Context-Knowledge as discussed previously in section B. Both measures are applied in the same way but the context is changed, instead of applying it into news headline/title, news articles paragraphs are considered for fetching the possible clues. The Fig.3. shows the detailed flow diagram of the proposed architecture.

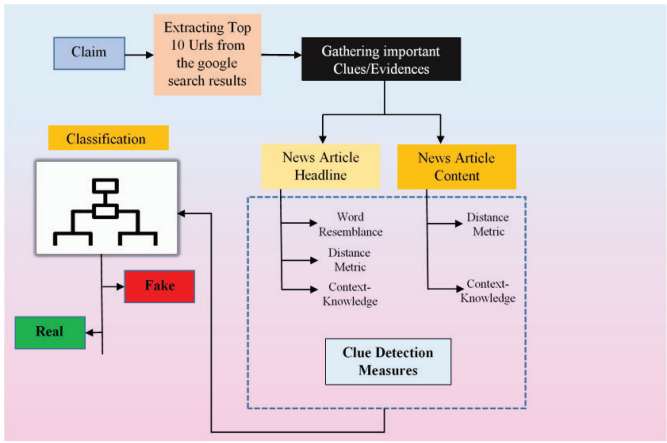


Fig. 3: The figure shows the flow diagram of the proposed method.

IV. Experimental Results

This section describes the experiments conducted to respond to the questions we draw up in the introduction part (Section I). As currently, no fake news dataset is available for COVID-19, the experiment is conducted on the proposed dataset, and the performance of the model is evaluated on the same. The metric used in this paper is the accuracy(acc), precision(Pre), recall(Rec), and f1-score(F1). The experiment is conducted to analyze the performance of the proposed model. The evaluation results analysis on the self-generated dataset is discussed in Section A.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:10:23 UTC from IEEE Xplore. Restrictions apply.

A. Self-Generated Dataset (Covid-fakenews2019)

To evaluate the proposed algorithm on the Self-Generated Dataset, the percentage split technique has been employed with 70% samples are used to train the model and 30% to test the data. The performance has been analyzed based on three different ways. Firstly, the effectiveness of the proposed model is evaluated with respect to both news article title/headline and its content using cross validation scheme 10 fold and 20 fold and percentage split technique (70:30) as shown in Table I and Table II respectively. From the given tables it has been observed that Random-forest(RF) and SVM(support vector machine) outperform all others on both the performance testing schemes with an f1-score of 0.99.

TABLE I: Effectiveness of Proposed Model on the Covid-fakenews2019 dataset with respect to both New Article Headline and Content:Cross validation

Classifier	**10 Fold**				**20 Fold**			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Logistic Regression	99.98	0.99	0.99	0.99	99.98	0.99	0.99	0.99
NaÃ¯ve Bayes	98.41	0.98	0.98	0.98	98.41	0.98	0.98	0.98
Random-Forest(RF)	99.99	0.99	0.99	0.99	99.99	0.99	0.99	0.99
Adaboost	70.42	0.71	0.70	0.69	69.98	0.70	0.70	0.69
Linear SVM	99.99	0.99	0.99	0.99	99.99	0.99	0.99	0.99

TABLE II: Effectiveness of Proposed Model on the Covid-fakenews2019 dataset with respect to both New Article Headline and Content:Percentage-Split

Classifier	**Percentage-split(70:30)**			
	Acc	Pre	Rec	F1
Logistic Regression	99.98	0.99	0.99	0.99
Naive Bayes	98.33	0.98	0.98	0.98
Random-Forest(RF)	99.99	0.99	0.99	0.99
Adaboost	69.98	0.70	0.70	0.69
Linear SVM	99.99	0.99	0.99	0.99

V. Conclusion

In this paper, we developed an intelligent generalized strategy for identifying possible clues to predict misleading information, where fake news proliferated during the COVID-19 outbreak is considered as a special case study and detailed analysis has been discussed. To extract important clues for the verification of news content, the scheme leverages three sets of novel evidences proposed in this paper that are based on Word Resemblance, Distance-Metric, and Context-Knowledge. These features are extracted from the news article headlines/titles and their content. The derived feature-set is given as an input to the machine learning model and performance has been analyzed by considering different variants of feature category by employing a different set of classifier applied on self-generated dataset "Covid-fakenews 2019". It has been observed that most of the classifiers are performing well on the dataset we have employed, where SVM and Random-Forest outperforms all other classifiers with an f1-score of 0.99. In the future, we are planning to build the real-time application of the proposed work and also extend the proposed dataset by incorporating a large number of samples with different categories of news that proliferated during the COVID-19 outbreak, like in this work we are more focusing on the fake news propagated with respect to cures/treatment of the disease. In the future, we are also including other types of news like false information regarding lockdowns in the cities, fake claims that are pretended to be posted by some government officials, etc.

References

[1] K. K. Sahu, A. K. Mishra, and A. Lal, âComprehensive update on current outbreak of novel coronavirus infection (2019-nCoV),â Ann. Transl. Med., vol. 8, no. 6, 2020.

[2] D. K. Vishwakarma, D. Varshney, and A. Yadav, âDetection and veracity analysis of fake news via scrapping and authenticating the web search,â Cogn. Syst. Res., vol. 58, pp. 217â229, 2019.

[3] D. Varshney and D. K. Vishwakarma, âA Review on Rumour Prediction and Veracity Assessment in Online Social Network,â Expert Syst. Appl., p. 114208, 2020.

[4] W. Zhou, A. Wang, F. Xia, Y. Xiao, and S. Tang, âEffects of media reporting on mitigating spread of COVID-19 in the early phase of the outbreak,â 2020.

[5] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, âNovel Visual and Statistical Image Features for Microblogs News Verification,â Trans. Multi., vol. 19, no. 3, pp. 598â608, Mar. 2017.

[6] K. Wu, S. Yang, and K. Q. Zhu, âFalse rumors detection on Sina Weibo by propagation structures,â 2015 IEEE 31st Int. Conf. Data Eng., pp. 651â662, 2015.

[7] B. D. Horne and S. Adali, âThis just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news,â in Eleventh International AAAI Conference on Web and Social Media, 2017.

[8] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, âStop clickbait: Detecting and preventing clickbaits in online news media,â in 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 9â16.

[9] A. Aker, L. Derczynski, and K. Bontcheva, âSimple open stance classification for rumour analysis,â arXiv Prepr. arXiv1708.05286, 2017.

[10] E. J. Briscoe, D. S. Appling, and H. Hayes, âCues to deception in social media communications,â in 2014 47th Hawaii international conference on system sciences, 2014, pp. 1435â1443.

[11] G. Giasemidis et al., âDetermining the Veracity of Rumours on Twitter,â CoRR, vol. abs/1611.0, 2016.

[12] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, âProminent Features of Rumor Propagation in Online Social Media,â 2013 IEEE 13th Int. Conf. Data Min., pp. 1103â1108, 2013.

[13] L. Zeng, K. Starbird, and E. S. Spiro, â unconfirmed: Classifying rumor stance in crisis-related social media messages,â in Tenth International AAAI Conference on Web and Social Media, 2016.

[14] L. Derczynski, K. Bontcheva, M. Liakata, R. Procter, G. W. S. Hoi, and A. Zubiaga, âSemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours,â CoRR, vol. abs/1704.0, 2017.

[15] A. Hamid et al., âFake News Detection in Social Media using Graph Neural Networks and NLP Techniques: A COVID-19 Use-case,â arXiv Prepr. arXiv2012.07517, p. 114208, 2020.

[16] R. Vijjali, P. Potluri, S. Kumar, and S. Teki, âTwo stage transformer model for covid-19 fake news detection and fact checking,â arXiv Prepr. arXiv2011.13253, 2020.

[17] A. Groza, âDetecting fake news for the new coronavirus by reasoning on the Covid-19 ontology,â arXiv Prepr. arXiv2004.12330, 2020.

[18] D. Varshney and D. K. Vishwakarma, âHoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles,â J. Ambient Intell. Humaniz. Comput., pp. 1â14, 2020.

[19] D. Varshney and D. K. Vishwakarma, âA unified approach for detection of Clickbait videos on YouTube using cognitive evidences,â Appl. Intell., pp. 1â22.

[20] F. Yu et al., âDetecting Rumors from Microblogs with Recurrent Neural Networks,â CoRR, vol. 8, no. 1, pp. 1435â1443, Aug. 2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:10:23 UTC from IEEE Xplore. Restrictions apply.

[21] E. Kochkina, M. Liakata, and A. Zubiaga, âAll-in-one: Multi-task Learning for Rumour Verification,â in Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3402â3413.
[22] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun, âCED: Credible early detection of social media rumors,â IEEE Trans. Knowl. Data Eng., 2019.
[23] O. Ajao, D. Bhowmik, and S. Zargari, âFake news identification on twitter with hybrid cnn and rnn models,â in Proceedings of the 9th International Conference on Social Media and Society, 2018, pp. 226â230.
[24] W. Wang, ââLiar, Liar Pants on Fireâ: A New Benchmark Dataset for Fake News Detection,â 2017.
[25] Apuke, Oberiri Destiny, and Bahiyah Omar. "Fake news and COVID-19: modelling the predictors of fake news sharing among social media users." Telematics and Informatics 56 (2021): 101475.
[26] Paka, William Scott, et al. "Cross-SEAN: A cross-stitch semi-supervised neural attention model for COVID-19 fake news detection." Applied Soft Computing (2021): 107393.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:10:23 UTC from IEEE Xplore. Restrictions apply.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354313426

Analysis of Organic Rankine Cycle with Waste Heat Recovery from Gas Power Plant using Loop Heat Pipes

Article *in* Journal of Mechanical Engineering Research and Developments · September 2021

CITATIONS
0

READS
15

2 authors:



Ankit Dwivedi
Delhi Technological University
12 PUBLICATIONS **7** CITATIONS
SEE PROFILE



Radhey Shyam Mishra
Delhi Technological University
53 PUBLICATIONS **175** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Doctoral Research View project

Performance Measurement View project

All content following this page was uploaded by Ankit Dwivedi on 02 September 2021.

The user has requested enhancement of the downloaded file.

Journal of Mechanical Engineering Research and Developments
ISSN: 1024-1752
CODEN: JERDFO
Vol. 44, No. 9, pp. 265-285
Published Year 2021

Analysis of Organic Rankine Cycle with Waste Heat Recovery from Gas Power Plant using Loop Heat Pipes

Ankit Dwivedi*, R S Mishra

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

*Corresponding Author Email: ankitdwivedi2008@gmail.com, rsmishra@dtu.ac.in

ABSTRACT

In this research work Waste Heat Recovery from a Practical Gas Power Cycle (GPC) combined with an Organic Rankine Cycle (ORC) through a Loop Heat Pipe Heat Exchanger (LHP Hex.) has been presented for Industrial. Six Eco-friendly refrigerants namely R600, R134a, R290, R152a, R125 & R124 have been used as working fluids for ORC along with 4 easily available working fluids such as Water, Ethanol, Methanol & Acetone for LHP. Peak Temperature of GPC, Working Fluids of ORC & LHP Hex. and Condenser Pressure of ORC have been kept as the input variables for the combined cycle, whereas, the net-work output of the ORC, First Law & Second Law Efficiency, Irreversibility Associated with the major components, required flow rates of the different working fluids and Most Suitable working fluids are the Output Performance Parameters. The combination of Water (LHP H Ex.) & R290 (ORC) is the most suited working fluids among all the 24 Combinations studied. Moreover, the First & Second Law Efficiencies of the System at 1500K Peak GPC Temperature & 65 kPa ORC Condenser Pressure have been recorded as 32% & 80 % respectively with 170 kW net-work output. The irreversibility related to each different component of the lower system has been around 80 kW. The ratio of the mass flow rate of R290 to Water has been around 1.42.

KEYWORDS

Heat Pipes Heat Exchangers, Organic Power Cycle, Brayton Cycle, Combined Cycle, Eco-friendly Refrigerants

INTRODUCTION

Energy has been the source of life for ages, while, humankind has been looking for different highly efficient energy sources. From small-scale hydro-power plants to large-scale hydel plants, from low power-producing steam & gas engines to large capacity steam power plants, gas power plants & IC engines, the world has seen a huge technological enhancement. It has been observed in the high energy input cycles (Rankine/Brayton), there's a large energy source left unutilized, which results in unutilized monetary expenditure, loss of exergy & Global Warming. Hence, the study of combining two different cycles has been persistent for a very long time i.e., power cycles with power cycles & power cycles with refrigeration cycles. In such combined systems, waste heat recovery has been the topmost priority. In this research work, two power cycles namely Gas Power Cycle (GPC) and Organic Vapor Power Cycle (Organic Rankine Cycle (ORC)) are connected through a Loop Heat Pipe (LHP) as a Loop Heat Pipe Heat Exchanger (LHP HEx.) on computer-based simulations. The waste heat of the Brayton Cycle is utilized through LHP Hex. for input to the Boiler of the ORC. The working fluids in the ORC are eco-friendly refrigerants namely R124, R125, R152a, R600, R290&R134a, while the working fluids for the LHP have been chosen to be Acetone, Ethanol, Methanol & Ammonia. In this research work, 24 combinations have been studied with different input variables. The benefit of these refrigerants is that they can be converted into vapors at relatively low temperatures, hence making possible the use of low-temperature sources quite effectively. The LHP acts as a superconductor of Heat, providing almost constant effectiveness around 90% (Evaporator & Condenser of LHP), moreover, for the interfaces between GPC, ORC & LHP average effectiveness may be taken as 0.65-0.7 based on the literature. The previous research works are summarized in the following Para;

Angelino et. al. [1] studied the geothermal Organic Rankine Cycle (ORC) with combinations of n-butane & n-hexane. Where the results showed that the ORC with such combinations delivered 6.8% more force than with

sole n-pentane. Papadopoulos et. al. [2] projected a thorough method for fluid selection through a simulation-aided system following process optimization, physical, chemical, eco-friendly, safety evaluating the economic properties of pure ORC fluids. Demuth [3] performed a case study on a geothermal power plant working on two-component mixtures of natural hydro-carbons for specified compositions at 137℃ and 182℃ temperature of geothermal water. The efficiency increases up to 14% for the sub-critical cycles compared to the most efficient pure working fluid as propane. Hung et. al. [4] analyzed an ORC employing diverse working fluids among wet, dry & isentropic fluids. Dry and isentropic fluids exhibited better thermal efficiencies, moreover, did not get condensed during the expansion process, hence, resulted in smaller damage. Tchanche et. al. [5] analyzed thermo-physical characteristics along with the results of 20 fluids in a low-temperature solar ORC. R134a was finally recommended. Heberle et al. [6] performed the exergy analysis of zeotropic mixtures as the working fluids in which the results exhibited that the efficiency had increased up to 15% in comparison with that of pure fluid for heat source temperature below 120°C. Liu et. al. [7] studied the effect of temperature glide during the zeotropic condensing process on the thermal efficiency, exergy efficiency, and work of the ORC system. Yang et. al. [8] deliberated the performance of eight zeotropic mixtures working as fluids for a waste heat recovery system of the vehicle engine. Wang et. al. [9] performed a comparative study on three different compositions (0.9/0.1, 0.65/0.35, and 0.45/0.55) of R245fa/R152a to pure R245fa in a low-temperature solar ORC. The second law efficiency of subcritical cycles was studied. Dwivedi et al. [10] analyzed the viability of LHP Ex for intra-cycle heat exchange for the VARS. Fabian Korn [11] accomplished numerous experiments on heat pipes to establish it to be the most effective method to transport thermal energy.

GAP IN THE LITERATURE

It has been observed through the literature review that the Brayton Cycle and the ORC can be connected through the LHP to obtain better utilization of waste heat and highly effective performance. The outlet of the Brayton Cycle has been routed through the LHP for waste heat utilization. This in turn results in a novel system on which investigation has been made in this research work. Hence, the research investigation aims at the feasibility study of such a system, and the results have been discussed and concluded. The objective of this research article is not to analyze and report on the GPC, LHP, or ORC as sufficient work has been published on the different parameters of such systems. It can be observed that the flexible nature of LHP makes its use very attractive in an industrial set-up in which heat can be efficiently transferred from the source to the sink over large distances.

MATERIALS AND METHODS

In this newly envisaged combined system, ORC has been connected with the Gas Power Cycle (GPC) and works on the heat rejected from the Power Cycle (Fig. 1). The exhaust of the GPC is passed through the evaporator of the Loop Heat Pipe (LHP), where it vaporizes the working fluid in the LHP; further, these vapors are then used to supply input heat to the boiler of the ORC through the Condenser of LHP. The condenser of the LHP is integrated with the evaporator of the boiler to achieve better flexibility in the operations. Moreover, LHP performs as a superconductor of heat owing to the extremely extensive heat transfer coefficient observed in the Boiling-Condensation phenomenon. The software-based mathematical modeling analysis has been performed keeping the variation in peak temperature (T_5 or T_{PEAK}) and the pressure ratio (r or P_3/P_1) of the GPC in consideration. Parametric analysis of the rest of the components is recorded with the help of the simulations based on the equations for GPC, LHP & ORC in published literature and textbooks.

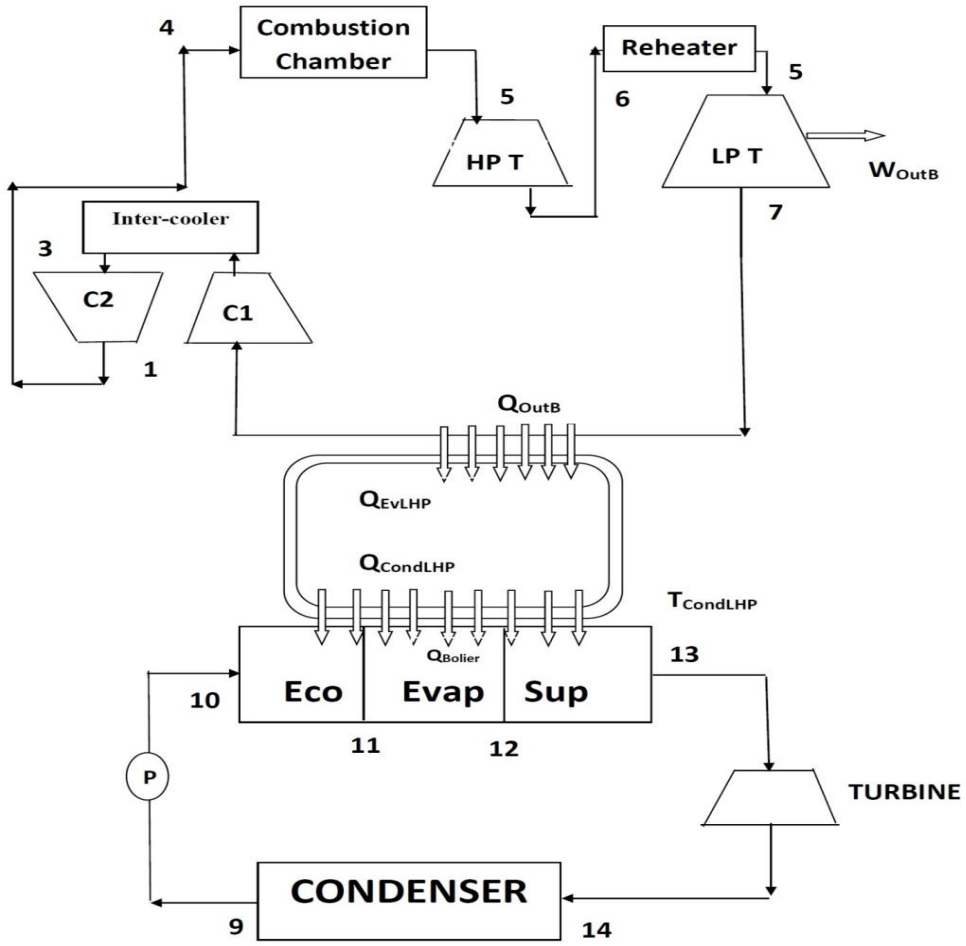


Figure 1. Gas Power Cycle Combined with Organic Vapour Power Cycle through a Loop Heat Pipe

Fig 2 has been drawn based on the available literature and texts to present the T-s plot of the complete cycle. The bottom cycle is the ORC & Top cycle is the GPC. In the middle of both the LHP Heat Exchangers are used to connect the both instead of a regular heat exchanger. The Evaporator of the LHP has been equipped with baffles, fins assembly & multiple passes to maximize the effectiveness between the heat transfer between the High Energy Gas Mixture & the working fluid of LHP in the evaporator of LHP, whereas, in the condenser of the LHP boiling & condensation heat transfer will take place which has the maximum effectiveness of all the arrangements. To even facilitate the drop-wise condensation & nucleate boiling phases to prevail necessary precautions have been made. The symbols have their usual meaning and the cycles consist of the usual components such as compressor, turbine, boiler, condenser combustion chamber, and others as consisted of a simple cycle for the software-based simulations.

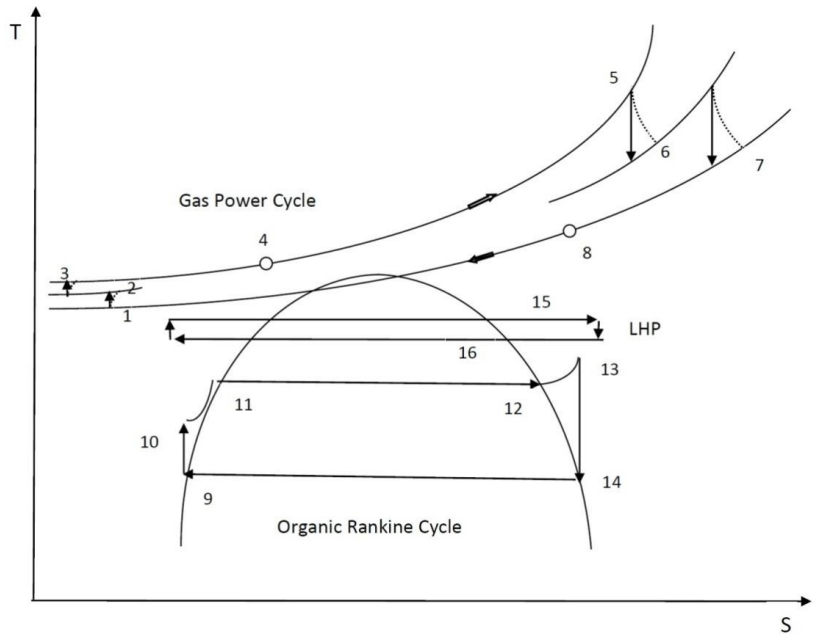


Figure 2. T-s Plot for Gas Power Cycle Combined with Organic Vapour Power Cycle through a Loop Heat Pipes

Figure 3 explains the structure of the LHP as well as the methodology of working. It shows the different compartments of the LHP-Evaporator. Moreover, the flow of the fluid within the LHP can be explained in Figure 3. A fluid near-saturated liquid state enters the evaporator where it evaporates to become saturated/superheated vapor. Further, the subcooled liquid goes out of the condenser tube to the compensation chamber of the LHP hence completing the loop.

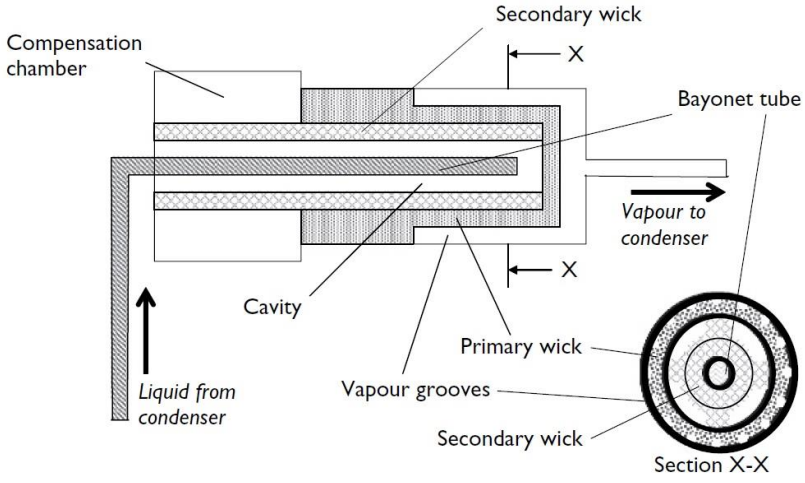


Figure 3. The Sectional view of Loop Heat Pipe Evaporator [10].

Table 1A & 2B contains the 10-working fluids being considered in the analysis. The working fluids of the LHP and that of the ORC have been presented. The equation solver software draws the value of properties corresponding to the state and generates results through several simulations.

Table 1A. Combination of working fluids for ORC

ORC Working Fluid	R124	R125	R134a	R152a	R290	R600

Table 1B. Combination of working fluids for LHP-Ex

LHP Working Fluid	Acetone	Ethanol	Methanol	Water

METHODOLOGY & EQUATIONS USED

The novel system works on a wide range of working parameters as input in the computer-based modeling software. Working fluids combinations, Peak Temperature of GPC, Condenser Pressure of ORC, the mass flow rate in GPC, etc. have been taken as the basic input parameter. The pressure ratio of the GPC has been taken as 5. Similarly, for example, the boiler temperature & pressure of the ORC has been connected to the Condenser Temperature of the LHP which in turn is dependent on the pressure ratio and peak temperature of the GPC, an attribute of the combined system. First Law & Second Law Analysis have been done for the combined system & components. To assess the feasibility of such a system for industrial applications, the applicable equations from standard texts and literature available have been modeled into the equation solving software to run iterations and to obtain the results based on which conclusions could be presented from this investigation.

Following Input, variables have been used to obtain a comprehensive analysis of the combined system & components. The limits of both the cycles have been fixed i.e. Peak Temperature (T_5) & Sink Pressure & Temperature (P_9 & T_9). All

Table 2. Input parameters & Abbreviations/Symbols

Input Parameters for the Analysis	**Abbreviation/Symbols**
Working Fluids of LHP	
Working Fluid of ORC	
Inlet Temperature (K) to Turbine of GPC	$\mathbf{T_5}$
Condenser Pressure of ORC (kPa)	$\mathbf{P_9}$

Gas Power Cycle [14]

Table 3A, 3B & 3C show the list of parameters & abbreviations/symbols used during the analysis, whereas, Table 2 shows the list of input parameters for the analysis. Effect of mass flow rate in GPC has been kept as input similar to the Peak Temperature so assess the total heat exchange through the LHP Ex.

Table 3A. List of parameters & respective abbreviations for GPC

Parameters	**Abbreviations/ Symbols**
Gas Power Cycle	
Inlet Temperature (K) & Pressure (kPa) to Compressor of Gas Power Cycle (GPC)	$\mathbf{T_1, P_1}$
Outlet Temperature (K) & Pressure (kPa) to Compressor of GPC	$\mathbf{T_3, P_2}$
Inlet Temperature (K) & Pressure (kPa) to Combustion Chamber of GPC	$\mathbf{T_4, P_2}$
Inlet Temperature (K) & Pressure (kPa) to Turbine of GPC	$\mathbf{T_5, P_2}$
Inlet Temperature (K) & Pressure (kPa) to Regenerative Heat Exchanger of GPC	$\mathbf{T_7, P_1}$
Inlet Temperature (K) & Pressure (kPa) to Loop Heat Pipe Connecting the Power Cycles	$\mathbf{T_8, P_1}$
Work Input in the GPC (kW)	$\mathbf{W_{inB}}$
Heat Input in the GPC (kW)	$\mathbf{Q_{add}}$
Work Output in the GPC (kW)	$\mathbf{W_{outB}}$
Net-Work Output in the GPC (kW)	$\mathbf{W_{netB}}$
Heat Exchange between GPC & LHP in kW	$\mathbf{Q_{GP\text{-}LHP}}$
Flow Rate of Air in GPC (kg/s)	$\mathbf{\dot{m}_a}$
Flow Rate of Gas in GPC (kg/s)	$\mathbf{\dot{m}_g}$
Pressure Ratio	**r**
Heat Exchange in Regenerator in kW	$\mathbf{Q_R}$

Isentropic compression the work required for the compressor is given by:

$T_2/T_1 = r^{(\gamma-1/\gamma)} = T_5/T_6$ (1)

$T_3= T_1+ (T_2 – T_1)/\eta_C \& T_7 = T_5 – (T_5-T_6)\ \eta_{T,}$ (2)

Where η_T can be taken as 0.92 &η_C can be taken as 0.86

$W_{inB}= H_3 – H_1=\dot{m}_a c_{pa} (T_3-T_1)\ldots$ (3)

Isobaric heat addition the net heat added is given by

$Q_{add} = H_5 – H_4 =\dot{m}_a c_{pa} (T_5-T_4) = \dot{m}_F CV$ (Calorific Value of Fuel= 43 MJ) (4)

Isentropic expansion the work done by turbine is given by

$W_{outB}= H_7 – H_5 =\dot{m}_g c_{pg} (T_7-T_5)$ (5)

$W_{NetB}= W_{outB} - W_{inB}$ (6)

$\eta_B = W_{NetB} / Q_{add}$ (7)

Heat Exchanged in the Regenerator

$Q_R = H_4 – H_3 =\dot{m}_a c_{pa} (T_4-T_3)$ (8)

Where, $T_4 = T_3 + \varepsilon_R(T_7-T_3)$ (9)

Isobaric heat rejection (in a heat exchanger), the net heat rejected is given by

$Q_{Rej} = H_8– H_1= \dot{m}_g c_{pg} (T_8-T_1)$ (10)

Where $T_8 = T_7-\varepsilon_R(T_7-T_3) \& \varepsilon_R$ can be taken as 0.75 (11)

Exergy Analysis

Compressor

Exergy & Irreversibility for Compressor

$X_{Comp}= m_a (h_1-h_3) – T_o\ m_a (s_3-s_1)$ (12)

$I_{Comp}= T_o\ m_a (s_3-s_1)$ (13)

Where$s_3-s_1= c_{Pa} (T_3-T_1) – R_a \ln(P_3/P_1)$ (14)

Regenerator

$\Delta s_{RH}= s_{Ho} – s_{Hi}$ (15)

$\Delta s_{RC}= s_{Co} – s_{Ci}$ (16)

$x_C = (h_{Co}-h_{Ci}) – T_o (s_{Co}-s_{Ci})$ (17)

$x_H = (h_{Hi}-h_{Ho}) – T_o (s_{Hi}-s_{Ho})$ (18)

$I_R = T_o\Delta\dot{S}_o= T_o[\dot{m}_g(s_7-s_8) - \dot{m}_a(s_4-s_3)]$ (19)

Where, $s_7-s_8= c_{pg} \ln(T_7/T_8) \& s_4-s_3= c_{Pa} \ln(T_4/T_3)$ (20)

Combustion Chamber

$X_{CC} = Q_{add} – T_o S_{gen}$ (21)

$I_{CC}= T_o S_{gen}$ (22)

Where, $S_{gen}= \dot{m}_g c_{Pg} \ln(T_5/T_4) - \dot{Q}_{add}/ T_{av}$ (23)

Turbine

$X_T = m_g (h_5-h_7) - T_o m_g (s_5-s_7)$ (24)

$I_T = T_o m_g (s_5-s_7)$ (25)

Where, $s_5-s_7 = c_{Pg} (T_5-T_7) - R_g \ln(P_5/P_7)$ (26)

$I_{TotalORC} = I_{Comp} + I_{CC} + I_R + I_T$ (27)

Exergetic efficiency = $\frac{Wnet}{Wnet + \dot{I}total}$ (28)

Loop Heat Pipes [12]:

Table 3B. List of parameters & respective abbreviations for LHP-Ex

Parameters	**Abbreviations/ Symbols**
Loop Heat Pipe Hex.	
Evaporator Temperature LHP Connecting the Power Cycles (K)	$\mathbf{T_{EvLHP}}$
Condenser Temperature LHP Connecting the Power Cycles (K)	$\mathbf{T_{CondLHP}}$
Heat Absorbed in the Evaporator of the LHP Connecting the Power Cycles (kW)	$\mathbf{Q_{EvLHP}}$
Heat Transported to the Condenser of the LHP Connecting the Power Cycles (kW)	$\mathbf{Q_{CondLHP}}$
Heat Input to the Boiler of the ORC Through the LHP Connecting the Power Cycles (kW)	$\mathbf{Q_{inLR}}$
The surface tension N/m	$\boldsymbol{\sigma}$
The latent heat of vaporization at heat pipe temperature in kJ/kg	$\mathbf{h_{fg}}$
The viscosity of the liquid in Pa-s.	$\boldsymbol{\mu_l}$
The Figure of Merit in kW/m^2	$\mathbf{MF_h}$
The X-sectional area of wick in m^2	$\mathbf{A_w}$
The Permeability of the wick material in m^2	$\mathbf{k_{hp}}$
The effective length of the heat pipe in m	$\mathbf{L_{eff}}$
The pore radius in m	$\mathbf{r_p}$
The acceleration due to gravity in m/s^2	$\mathbf{g}$
The inclination angle in degrees	$\boldsymbol{\Phi}$
The density of vapor in heat pipe in kg/m^3	$\boldsymbol{\rho_l}$
Mass Flow Rate in the Evaporator & Condenser of LHP in kg/s	$\dot{\mathbf{m}}_{\mathbf{LHP}}$
Heat Exchange in the Evaporator& Condenser of LHP Respectively in kW	$\mathbf{Q_{EvLHP}}$ $\mathbf{\&Q_{CondLHP}}$
Entropy inlet & outlet of the Evaporator of LHP in kJ/kgK	$\mathbf{s_{EviLHP}\&}$ $\mathbf{s_{EvoLHP}}$
Entropy inlet & outlet of the Condenser of LHP in kJ/kgK	$\mathbf{s_{CondiLHP}}$ $\mathbf{\&s_{CondoLHP}}$
Enthalpy inlet & outlet of the Evaporator of LHP kJ/kg	$\mathbf{h_{EviLHP}\&}$ $\mathbf{h_{EvoLHP}}$
Enthalpy inlet & outlet of the Condenser of LHP kJ/kg	$\mathbf{h_{CondiLHP}}$ $\mathbf{\&h_{CondoLHP}}$
Entropy Generation in the Evaporator & Condenser of LHP Respectively in kW/K	$\dot{\mathbf{S}}_{\mathbf{genEvLHP}}$ $\mathbf{\&}\dot{\mathbf{S}}_{\mathbf{genCondLHP}}$
Mean Temperature Diff. in the Evaporator & Condenser of LHP Respectively in K	$\mathbf{T_{H\text{-}HP}}$ $\mathbf{\&\ T_{L\text{-}HP}}$
The temperature of the vapor in LHP in K	$\mathbf{T_{vLHP}}$
The characteristic dimension of the liquid/vapor interface (m)	$\boldsymbol{\lambda}$

Thickness thermal layer (m)	$\boldsymbol{\delta}$
Heat Transfer in LHP as per Sonic Limitation in kW	$\mathbf{Q_S}$
Heat Transfer in LHP as per Entrainment Limit Limitation in kW	$\mathbf{Q_{Ent}}$
Heat Transfer in LHP as per Capillary Limit Limitation in kW	$\mathbf{Q_{Cap}}$
X-sectional area of the duct of LHP (m^2)	**A**
Irreversibility in different components in kW	**I**
Permeability of the Wick Material	$\mathbf{k_{hp}}$

$$Q_{GP\text{-}LHP} = m_g c_{Pg}(T_8\text{-}T_1) \quad (29)$$

$$\dot{E}_{GP\text{-}LHP} = m_g (h_8\text{-}h_1) – T_o\, m_g (s_8\text{-}s_1) \quad (30)$$

Where, $s_8\text{-}s_1 = c_{Pg} (T_8\text{-}T_1) – R_g \ln(P_8/P_1)$ (31)

Effectiveness of LHP Evap-GPC Heat Exchanger $\boldsymbol{\varepsilon_{LHPEv\text{-}GPC}}$= 0.75

Effectiveness of LHP Cond-ORC Heat Exchanger $\boldsymbol{\varepsilon_{LHPCond\text{-}ORC}}$= 0.85

Figure of Merit for comparing different working fluids for LHP:

$$MF_h = \rho_l\, \sigma h_{fg} / \mu_l \quad (32)$$

The axial heat flow rate due to the sonic limitation was calculated from the following equation

$$Q_S = \pi\, r_p^2 \rho_v h_{fg} \sqrt{\frac{\gamma RT}{2(\gamma+1)}} \quad (33)$$

The maximum heat transfer due to the entrainment limit was determined using the equation

$$Q_{Ent} = \pi\, r_p^2 h_{fg} \sqrt{(2\, \pi\, \rho_v \sigma \cos\phi / \lambda)} \quad (34)$$

The maximum heat transfer due to the capillary limit was determined using the equation

$$Q_{Cap} = (MF_h)(A_w k_{hp}/L_{eff})\, \{(2/r_p)\text{-}(\rho_l g L_{eff} \sin\phi / \sigma)\} \quad (35)$$

The degree of superheat to cause nucleation is given by:

$$\Delta T = 3.06\, \sigma_l T_{vLHP} / \rho_v h_{fg} \delta \quad (36)$$

With reference to mass flow rate of warm and cool air, the heat transfer rates to the evaporator and condenser sections are calculated as follow:

$$q_e = mc_p(T_{e,in} - T_{e,out}) = \rho\, u\, Ac_p\, (T_{e,in} - T_{e,out.}) \quad (37)$$

$$q_c = mc_p(T_{c,in} – T_{c,out}) = \rho\, u\, Ac_p\, (T_{c,in} – T_{c,out.}) \quad (38)$$

$$HTF = 2MF_h A k_{hp} / r_p \quad (39)$$

Effectiveness

$$\epsilon_{LHP} = q_{act} / q_{max} = T_{e,in} - T_{e,out} / T_{e,in} – T_{c,in)} \quad (40)$$

Exergy Analysis of LHP- Evaporator:

$$Q_{EvLHP} = \dot{m}_g c_p\, (T_8 – T_1) \quad (41)$$

$$\dot{m}_{EvLHP} = Q_{evLHP} / (h_{EvoLHP} - h_{EviLHP}) \quad (42)$$

$$\dot{S}_{genEvLHP} = \dot{m}_{EvLHP}(s_{EvoLHP} - s_{EviLHP}) - Q_{evLHP} / T_{H\text{-}HP} \quad (43)$$

$$T_{H\text{-}HP} = (T_8\text{-}T_1) / \ln (T_8/T_1) \quad (44)$$

$$I_{genEvLHP} = T_o\, (\dot{S}_{genEvLHP}) \quad (45)$$

$$X_{EvLHP} = \dot{m}_g c_p\, (T_8 – T_1) - I_{genEvLHP} \quad (46)$$

Exergy Analysis of LHP-Condenser:

$$Q_{CondLHP} = \dot{m}_{ORC} h_{fgevORC} = Q_{inLR} \quad (47)$$

$$\dot{m}_{CondLHP} = \dot{m}_{EvLHP} = \dot{m}_{LHP} \quad (48)$$

$$\dot{S}_{genCondLHP} = \dot{m}_{CondLHP}(s_{CondoLHP} - s_{CondiLHP}) - Q_{CondLHP} / T_{L\text{-}HP} \quad (49)$$

$$T_{L\text{-}LHP} = (T_{CondLHP} + T_{EvapORC}) / 2 \quad (50)$$

$$I_{genCondLHP} = T_o(\dot{S}_{genCondLHP}) \quad (51)$$

$$X_{CondLHP} = Q_{CondLHP} - I_{genCondLHP} \quad (52)$$

$$I_{TotalLHP} = I_{genEvLHP} + I_{genCondLHP} \quad (53)$$

Organic Rankine Cycle [14]

Table 3C. List of parameters & respective abbreviations for ORC

Parameters	**Abbreviations/ Symbols**
Organic Rankine Cycle	
Inlet Temperature (K) & Pressure (kPa) to the Pump of Organic Vapour Power Cycle (ORC)	$\mathbf{T_9, P_9}$
Inlet Temperature (K) & Pressure (kPa) to the Boiler of ORC	$\mathbf{T_{10}, P_{10}}$
Inlet Temperature (K) & Pressure (kPa) to the Turbine of ORC	$\mathbf{T_{13}, P_{13}}$
Inlet Temperature (K) & Pressure (kPa) to the Condenser of ORC	$\mathbf{T_{14}, P_{14}}$
Work Output from the Turbine of ORC (kW)	$\mathbf{W_{outR}}$
Work Input to the Pump of ORC (kW)	$\mathbf{W_{inP}}$
Net-Work Output from the ORC (kW)	$\mathbf{W_{netR}}$
Exergy in kW & kJ/kg respectively	**X&x**
Enthalpy in kW & kJ/kg respectively	**H & h**
The temperature of the Environment in K	$\mathbf{T_o}$
Effectiveness of Different Components	$\boldsymbol{\varepsilon}$
The efficiency of Different Components	$\boldsymbol{\eta}$
Specific Heats Ratio	$\boldsymbol{\gamma}$

The power absorbed by the pump is estimated by equation

$$W_{inR} = (h_{10} - h_9) \quad (54)$$

$$\eta_p = (h_{10s} - h_9)/(h_{10} - h_9) \quad (55)$$

$$Q_{EvpR} = (h_{13} - h_{10}) \quad (56)$$

$$(h_{13} - h_{10}) = c_{pR}(T_{in} - T_{pp}) \quad (57)$$

The useful work output can be estimated by equation.

$$W_{outR} = (h_{13} - h_{14}) \quad (58)$$

$$Q_{ConR} = (h_{14} - h_9) \quad (59)$$

Where Q_{ConR}stands for the heat rejected heat in the condenser (kW)

Thermal efficiency is defined as the ratio of the network output to heat supplied to the evaporator.

$$\eta_{Th} = (W_{outR} - W_{inP})/Q_{EvpR} \quad (60)$$

Exergy destruction in pump

The exergy destruction rate in the pump is given by equation

$I_{pump} = T_o(s_9 - s_{10})$ (61)

$T_{H\text{-}ORC} = (T_{in} + T_{pp})/2$, the energy destruction in evaporator can be estimated by equation.

$I_{evp} = T_o[(s_{13} - s_{10}) - (h_{13} - h_{10}/T_{H\text{-}ORC})]$ (62)

Where;

$I_{evp}$is the exergy destruction rate in the evaporator (kW), s_{13} is specific entropy at boiler outlet kJ/kgK)

Exergy destruction in turbine

The equation gives the exergy destruction rate in the expander.

$I_{Turb} = T_o (s_{14} - s_{13})$ (63)

$I_{Turb}$is the exergy destruction rate in the turbine (kW), s_{14} is specific entropy at turbine outlet (kJ/kgK)

Exergy destruction in the condenser

$T_{L\text{-}ORC} = (T_{ci} + T_{co})/2$ can be used to estimate the exergy destruction in the condenser. The equation gives the exergy destruction in the condenser.

$I_{con} = \dot{m}_R \times T_o[(s_9 - s_{14}) - (h_9 - h_{14}/T_L)]$ (64)

I_{con} refers to the exergy destruction rate in condenser (kW)

The system's total energy destruction can be calculated by combining equations

$I_{TotalORC} = I_{Pump} + I_{Evap} + I_{Turb} + I_{Con}$ (65)

This gives:

$I_{Total} = T_o \times \left[\frac{h13-h10}{TH-ORC} \frac{h9-h14}{TL-ORC}\right]$ (66)

Exergetic efficiency = $\frac{Wnet}{Wnet + \dot{I}total}$ (67)

Table 4 contains the details of the recommended and non-recommended wick material for the LHP-Evaporator with the corresponding working fluids. While constructing the LHP, these details need to be kept in mind. In this research work Copper, Silica or Stainless Steel can be invariably used based on thermal effectiveness & cost-effectiveness.

Table 4. Heat Pipe working fluids & material compatibility [12]

Working Fluid	**Recommended Material**	**Not Recommended Material**
Acetone	Copper, Silica, Aluminium, Stainless Steel	
Methanol	Copper, Silica, Stainless Steel	Aluminium
Water	Copper, Silica, Aluminium, Stainless Steel, Nickel, Carbon Steel	Silica, Aluminium, Stainless Steel, Nickel, Carbon Steel,Inconel
Ethanol	NA	NA

Similar to the table above, Table 5 contains the details of the suitable operating temperature range, based on which the working fluid can be decided for the LHP. In this research work, the operating range is about 400-450 K. Hence, the corresponding working fluids i.e., Ethanol, Acetone, Methanol & Water have been chosen. It can be seen that the LHP can be used to extract heat from relatively low temperature sources, making it extremely useful for such a system.

Table 5. Heat Pipe working fluids useful operating temperatures [12]

Working Fluid	**Melting Point (ºC)**	**Boiling Point at ATP (ºC)**	**Useful Range (ºC)**
Acetone	-95	57	0-120

Methanol	-98	64	10-130
Ethanol	-112	78	0-130
Water	0	100	30-200

The next section discusses the results obtained from the simulations through various plots and data obtained.

RESULTS AND DISCUSSIONS

Software-based simulations have been prepared on engineering equation solver based on the modeled equations of GPC, LHP & ORC. The Peak Temperature of the GPC (T_5), Condenser Pressure of ORC (P_9), and working fluids for LHP &ORC have been kept as the input variables for this research work. The input heat source that has been obtained for the Boiler of ORC is in the range of 400K-450K (127 ˚C-177 ˚C) as a constant source owing to the condensation in the LHP. All 4 working fluids selected for this range are compatible. For the construction of such LHP, Silica, Copper, and Stainless Steel can be selected. Different performance parameters such as Net-Work Output of ORC (W_{netR}), Exergy (X) &Irreversibility (I) of different components, Corresponding Efficiencies & Effectiveness of the Components (η &ε), and Mass Flow rates of the working fluids have been closely recorded and studied through different simulations and the results have been presented and explained in the following sections.

Peak Temperature Of Gpc (T_5) As Input Variable

The Peak Temperature (T_5) has been varied from 1400 K to 1800K. In Fig 4 and Fig 5 comparative study of Exergy and Irreversibility has been recorded for Condenser and Boiler of the ORC for 6 selected refrigerants. For Condenser of ORC, it can be seen that the Irreversibility range from 10kW to 50 kW with R125 having the highest value, whereas, the top most Exergy results with R600 with a range from 40 kW to130 kW.

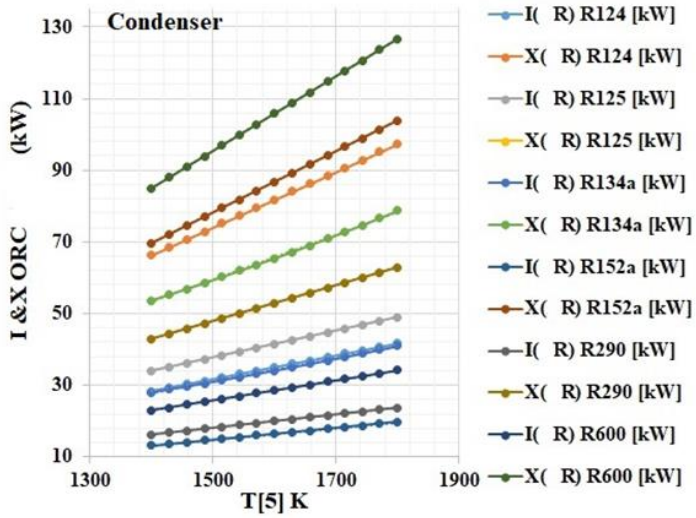


Figure 4. Irreversibility & Exergy associated with the Condenser of the ORC

The comparative study of Exergy and Irreversibility have been recorded for Condenser and Boiler of the ORC for 6 selected refrigerants. It can be assessed that the Irreversibility for Boiler for all the refrigerants lies within the range of 10kW to 30 kW. Moreover, R125 shows the maximum Irreversibility in the condenser whereas R600 has the lowest. Exergy of the condenser varies from a range of 40 kW to 190 kW. The highest Exergy has been observed for R125 whereas the lowest has been observed for R600.

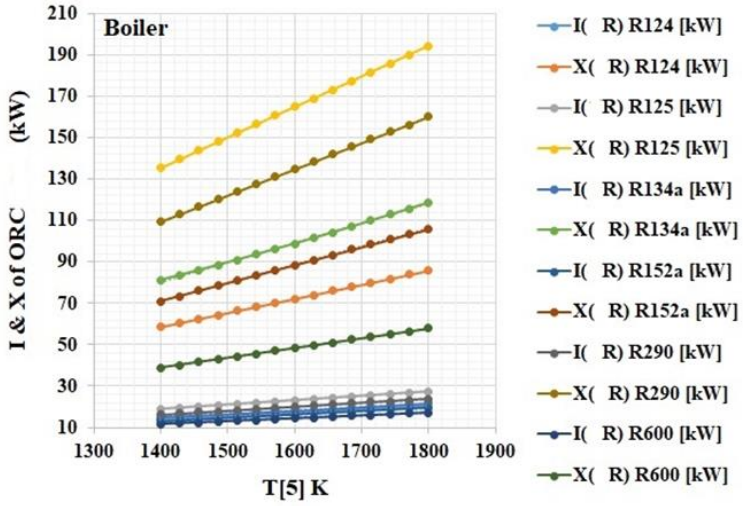


Figure 5. Irreversibility & Exergy associated with the Boiler of the ORC

The following Fig 6 & 7 represent the results of Exergy & Irreversibility associated with the LHP Condenser & Evaporator. It can be seen that the exergy is being transferred from the evaporator to the condenser with high effectiveness. Moreover, sufficient exergy is available at the condenser to transfer to the Boiler of the ORC to complete the connections. It can again be seen that R600 has generated the best possible results for the subject input variables. In the Evaporator the range for Exergy is from 700 kW to 1150 (kW), whereas, the Irreversibility ranges from 150 kW to 350 kW. On the other hand, for the Condenser, the Exergy available varies from 550 kW to 900 kW and Irreversibility from 60 kW to 250 (kW).

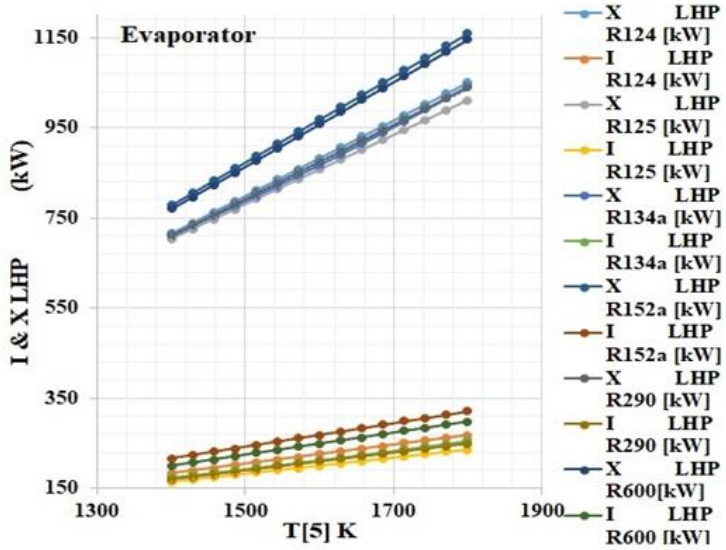


Figure 6. Irreversibility & Exergy associated with the Evaporator of the LHP

It can be seen that the LHP effectively with high efficiency interconnects the two cycles and high work output can be achieved with maximum waste heat utilization. Refrigerants being used for ORC can also be seen affecting the overall performance of the heat exchange taking place within the Condenser of the LHP. For R125, an exceptional ratio of transfer of Exergy from Evaporator to Condenser can be seen to have taken place.

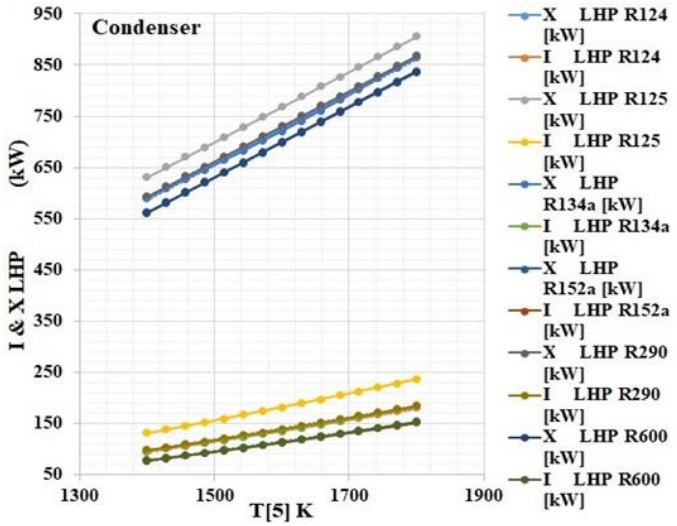


Figure 7. Irreversibility & Exergy associated with the Condenser of the LHP

The First Law & Second Law Efficiency for the ORC has been presented in Fig 8. The range for First Law Efficiency is from 25% to 32% with the maximum value for R125 whereas the Second Law Efficiency has the range of 73 % to 80% with R152a having maximum results.

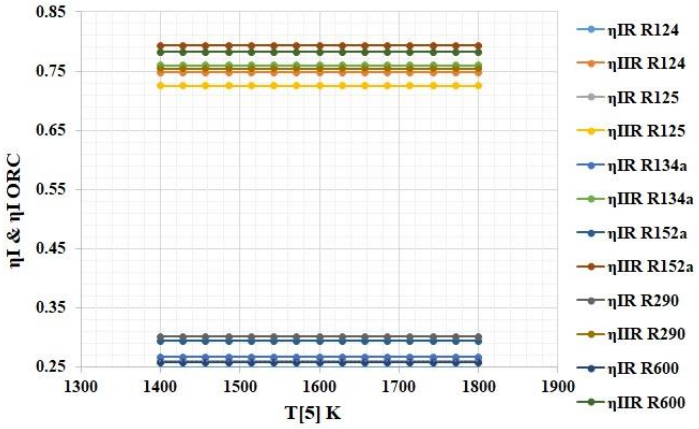


Figure 8. First & Second Law Efficiencies of the ORC

Subsequently, in Fig 9 Net-Work Output of the ORC can be observed. R600 & R124 have coinciding and least results. For R290, the maximum net-work output has been obtained.

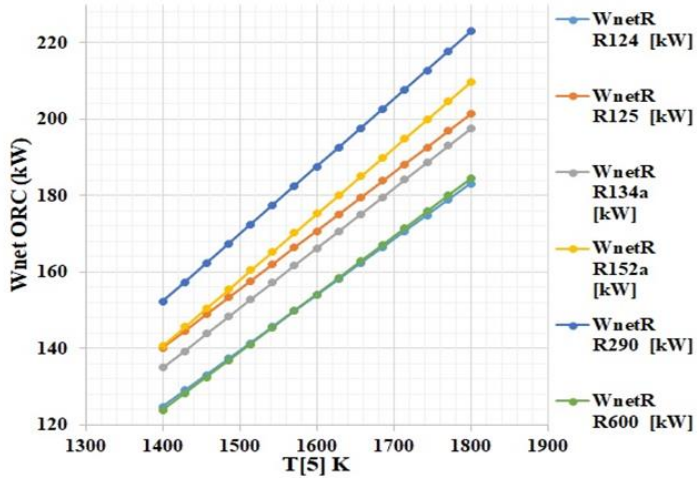


Figure 9. Net-Work Output of the ORC

In Fig 10 the Mass Flow rate inside the LHP and the ORC has been presented. It can be seen that requirement of the amount of working fluid in the LHP is quite less i.e., in the range of 0.3 kg/s to 0.5 kg/s when compared to the mass flow rate required for the ORC in which the range varies from 0.8 kg/s to 3.4kg/s for the subject input parameters.

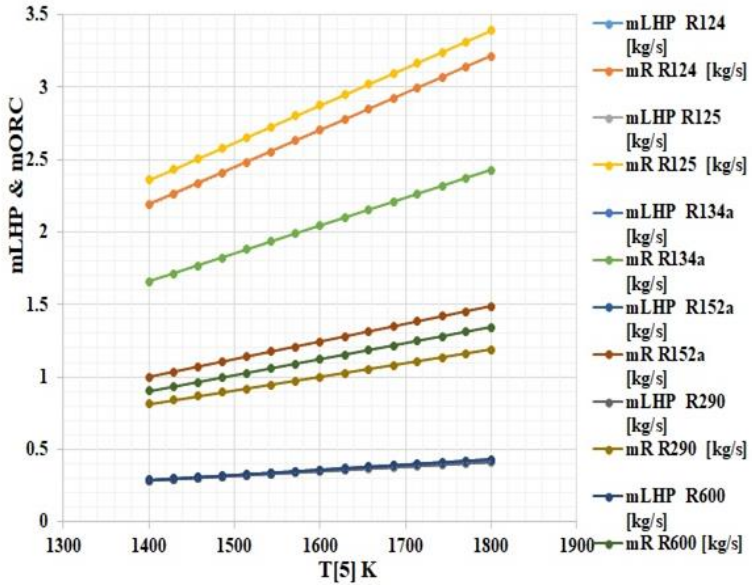


Figure 10. Mass Flow Rates in LHP & ORC

It can be seen that for the given input condition of Peak Temperature of GPC, Maximum Network Output in the ORC has been obtained for R290 and the Max First Law & Second Law Efficiency for R290 & R152a respectively. For the different components such as LHP, Boiler & Condenser maximum Exergy achieved is with R125, R152a, and R600& R125 respectively. It can also be seen that the Mass Flow of LHP is relatively lower than the ORC for the given condition. The overall Exergy transfer within the LHP is highly effective and efficient.

Condenser Pressure (P_9) Of Orc As Input Variable

The performance parameters have been studied for LHP & ORC by varying the Condenser Pressure of the ORC (P_9). It has been varied from 10 kPa to 100 kPa. Fig 11 shows the Irreversibility & Exergy in the condenser of the ORC. R125 isn't suitable for the pressure range as the value of the Exergy is extremely low. R600 has the maximum Exergy ranging from 65kW to 90kW. It can also be seen that the trend for other refrigerants is the same and parallel except R124 which has a zigzag trend ranging from 45 kW to 67.5 kW. Irreversibility also decreases with the range of pressure for all the refrigerants. For R152a, the irreversibility can be seen to be the lowest and constant at 25 kW. Moreover, for R124 the irreversibility is maximum. Fig 12 shows the Exergy of the boiler. Trends for all the refrigerants are similar and decreasing with maximum Exergy for R125 & ranging from 200 to 110 kW. The lowest Exergy value is associated with R600 ranging from 100 to 40 kW. Irreversibility data for all the refrigerants are coinciding and almost constant near 10 kW. Only R124 shows the zigzag trend.

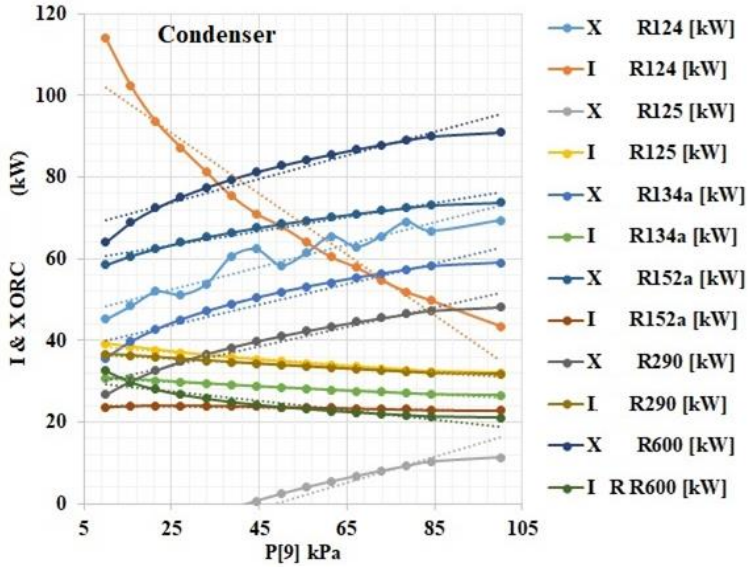


Figure 11. Irreversibility & Exergy associated with the Condenser of the ORC

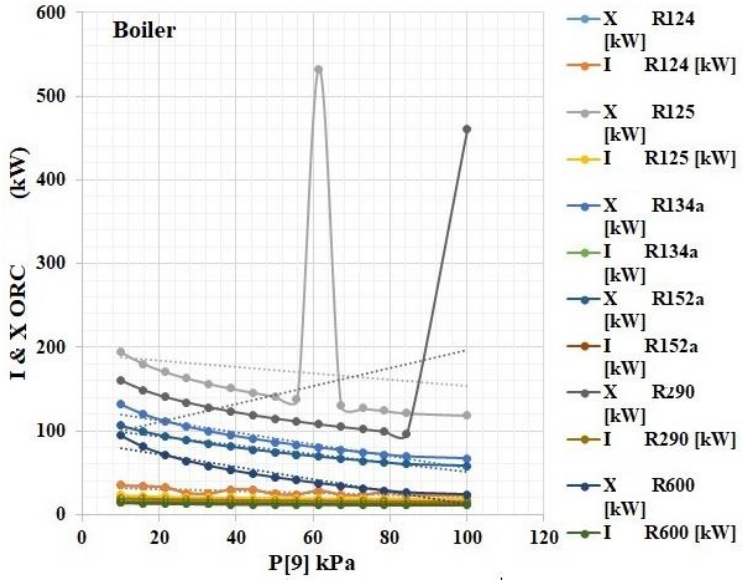


Figure 12. Irreversibility & Exergy associated with the Boiler of the ORC

Exergy & Irreversibility in the LHP has been studied in Fig 13& 14. It can be seen from the Figures that for both Evaporator and Condenser of the LHP, the values are quite distinct and have the same trends for Exergy and Irreversibility. It can also be seen that the values of the Exergy in the evaporator are constant except for R134a. Similarly, Irreversibility is constant except for R152a. The Evaporator Exergy has the range of 1000kW for R152a -700kW for R125 and Irreversibility has values varying from 380kW for R152a to 280 kW for R125.

The Condenser Exergy has values from 640 kW for R125 to 540 kW for R 152a and Irreversibility varies from 140 kW for R125 to 60 kW for 152a. The performance of the Evaporator & Condenser has been observed to be high and low for the same refrigerants.

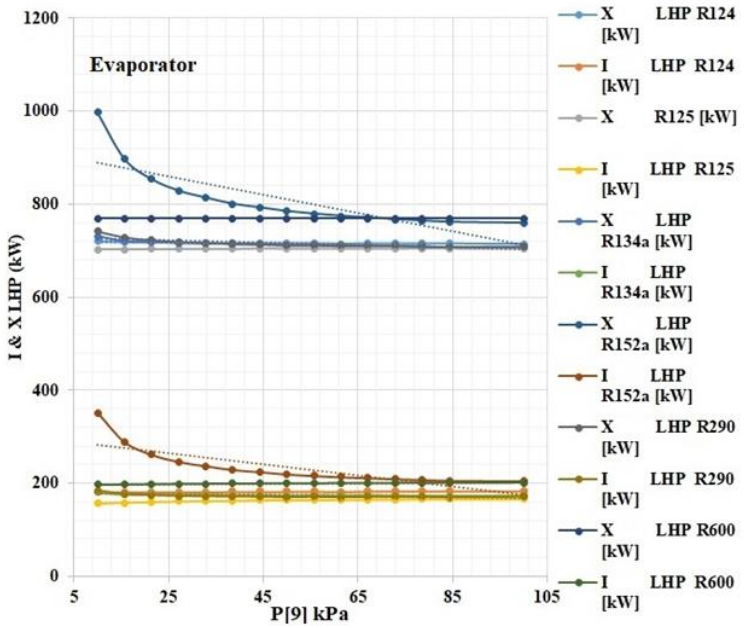


Figure 13. Irreversibility & Exergy associated with the Evaporator of the LHP

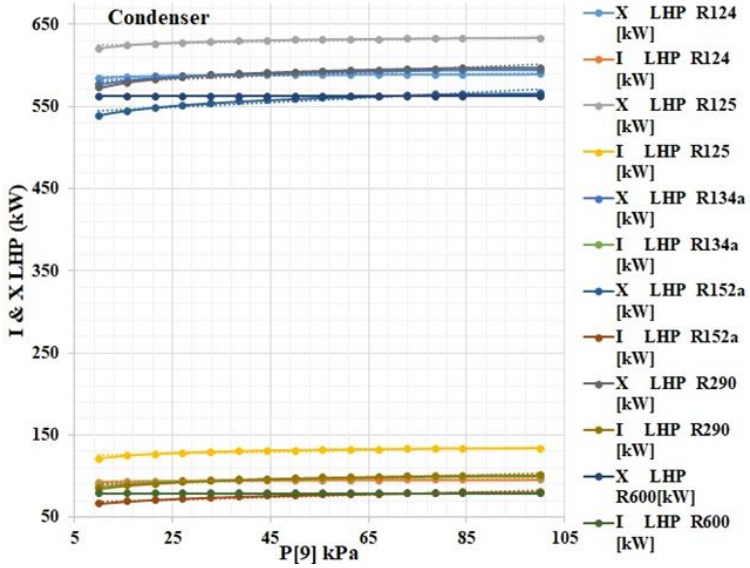


Figure 14. Irreversibility & Exergy associated with the Condenser of the LHP

In Figure 15 the First Law & Second Law Efficiencies are presented. It can be observed that the First Law Efficiency varies from 38% for R290 to 22% for R124. The trend for the first law efficiencies is as expected.

The Second Law Efficiency for the ORC ranges from 81% for 152a to 74% for R125. The results vary with every refrigerant and input variable. Brief disruptions can be observed forR125 & R290 hence it can't perform for the complete range.

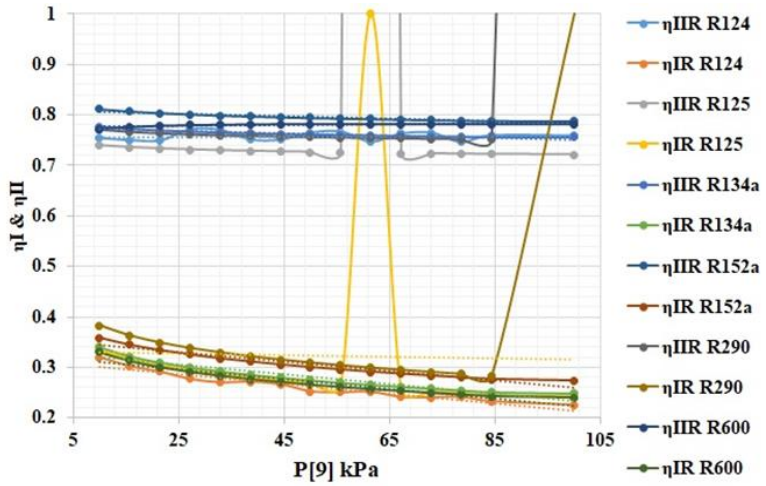


Figure 15. Irreversibility & Exergy associated with the Condenser of the LHP

Work out from Figure 16 can have been studied to be 190 kW for R290 to 110 kW for R600. R125 & R290 can't perform for the complete range. Brief disruptions can be observed.

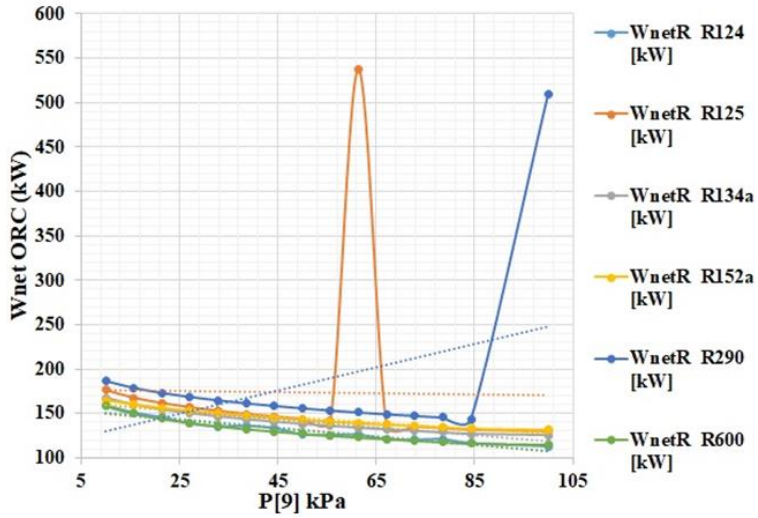


Figure 16: Irreversibility & Exergy associated with the Evaporator of the LHP

Working Fluid Of Lhp As Input Variable

In this segment results on the Mass Flow Rate of LHP & ORC have been presented varying the working fluid of the LHP, the Condenser Pressure of the ORC. Following Fig 17presents the results with Acetone as the working fluid in the LHP. It can be seen that for the 3 working fluids in the ORC namely R290, R600 & R152a the mass flow trend is parallel, increasing & in the same range 0.5 kg/s to 1.1 kg/s. For R134 the increasing trend from 1.3 to 2.5 kg/s is similar to R 125 from all other refrigerants. For R125 we get a continuously increasing trend from 1.9kg/s to 2.5 kg/s and for R124 the discontinuously increasing trend of in the range of 1.8 kg/s to 2.3 kg/s. The same trend in the ORC can be seen independent of the working fluid of the LHP.

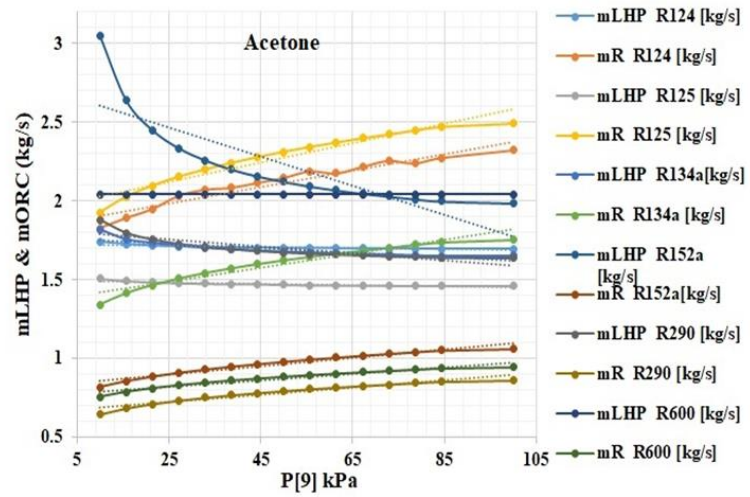


Figure 17. Mass Flow Rates in LHP & ORC with Acetone as LHP Fluid

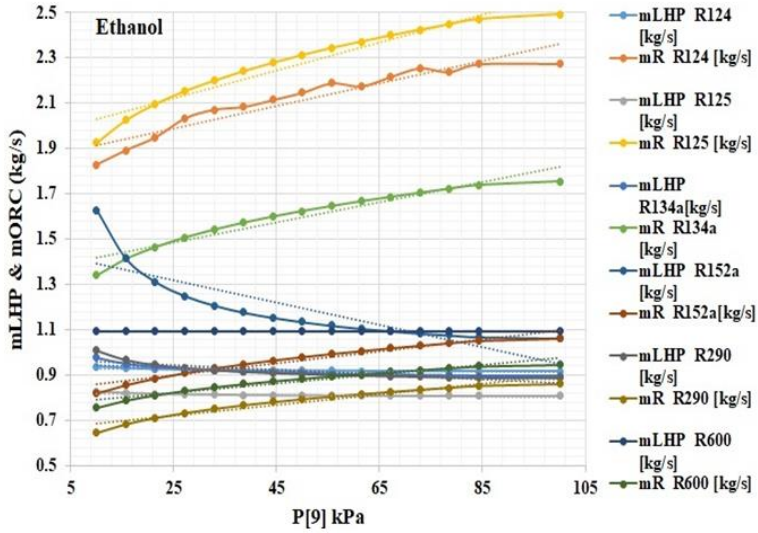


Figure 18. Mass Flow Rates in LHP & ORC with Ethanol as LHP Fluid

In Figure 17 the mass flow rate for Acetone in the LHP varies from 1.5 kg/s with R125 in the bottom cycle to 3 kg/s with R152a as the fluid in the ORC. Moreover, we can see that with pressure increase in the boiler the required mass flow rate has a decreasing trend. Whereas, in Fig 18, the results have been discussed for Ethanol in the LHP. The mass flow rate varies from 0.85kg/s forR125 and 1.65 kg/s for R152a. The trends are similar to the Acetone. In Fig 19 mass flow of Methanol can be seen to vary from 0.65 kg/s for R125 to 1.2 kg/s for R152a.

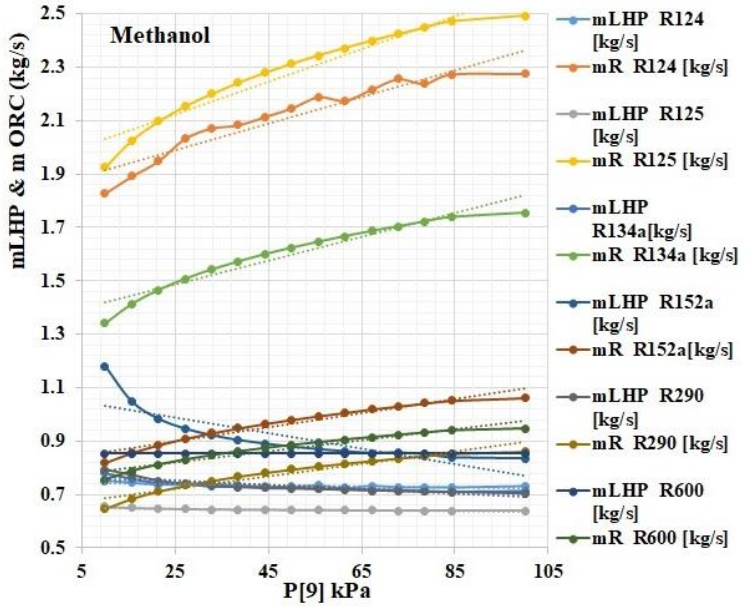


Figure 19. Mass Flow Rates in LHP & ORC with Methanol as LHP Fluid

The results can be seen to be constant and coinciding for all the refrigerants in the case of water in LHP in Fig 20. The mass flow rate can be seen to be 0.3kg/s.

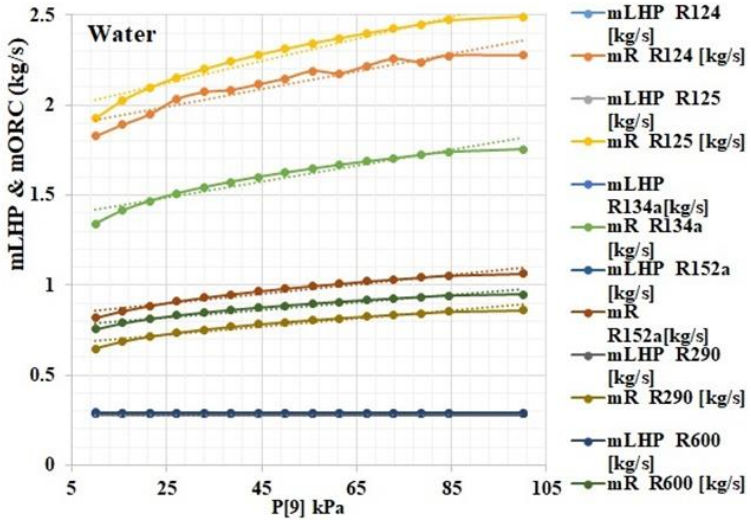


Figure 20. Mass Flow Rates in LHP & ORC with Water as LHP Fluid

CONCLUSIONS

It can be observed from the temperature profiles observed during the operations of the combined system that the temperatures available at the Condenser side of the LHP, the ORC also be run super-critically, in turn, increasing its efficiency. The superconductor nature of the LHP Ex can be very effectively utilized in running a highly efficient system. Furthermore, the following numerical conclusions can be presented on the Investigation done in this research work:

i. The temperature source available for the ORC boiler is at 400-450K for which all the working fluids of the LHP seem to be compatible and show continuous performances.

ii. Silica, Stainless Steel & Copper can be chosen as materials for the LHP being compatible for all.

iii. Work output for the ORC at Peak Operating Temperature 1500K can be observed to be 170kW for R290 and 165 kW for R290 at 65 kPa Condenser Pressure.

iv. The Maximum First Law Efficiency for the ORC has been recorded at 30% for R290 across the Peak Temperature and 32% at 65 kPa Condenser Pressure.

v. The Maximum Second Law Efficiency for the ORC has been presented to be75% for R290& 82 % for R152a across the Peak Temperature and 80% for R132a & 78% for R290 at 65 kPa Condenser Pressure.

vi. Mass flow rate for working fluids in ORC can be observed to be in the range of 0.5kg/s & 0.35kg/s in LHP for R290-Water. All other Combinations result in more mass handling per unit work output. Hence, multiple LHP Ex Pipes will be used for the operation of the Combined Cycle.

vii. Irreversibility associated with the components studied is in the range of 80kW.

For the Steel plants and Power Plants where low-grade energy is abundant, such a system can either be used to run either a centralized plant or separate systems as the total heat rejection from the Plants can be transferred through multiple LHPs depending upon the exhaust temperature of the GPC. As mentioned earlier the Waste heat which is a great cause for thermal pollution and Global warming can be utilized at higher temperatures and with substantial flexibility in such a Novel Combined Cycle. Hence, the research works offer an Eco-friendly system that works on a low heat source.

RECOMMENDATIONS FOR FUTURE RESEARCH WORKS

It can be recommended after this research article that experimental analysis is also performed for this system to have it finally implemented on an industrial basis. The industrial application of this system pre-requisites the successful design of high heat flux LHP H Ex which has not been covered in detail. The LHP has very little Pressure & Temperature drops within its system providing the combined system the advantage of a constant heat source at relatively high temperatures.

REFERENCES

[1] G. Angelino and P. Paliano, "Multi-component working fluids for organic Rankine cycles (ORCs)." Energy, Vol. 23, Pp. 449–63, 1998.

[2] A.I. Papadopoulos, M. Stijepovic and P. Linke, "On the systematic design and selection of optimal working fluids for Organic Rankine Cycles. Appl. Therm. Eng.," 30:760-9, 2010.

[3] O. Demuth, "Analyses of mixed hydrocarbon binary thermodynamic cycles for moderate temperature geothermal resources. Intersociety energy conversion engineering conference" Atlanta (USA); 1981.

[4] T.C. Hung, S.K. Wang, C.H. Kuo, B.S. Pei and K.F. Tsai, "A study of organic working fluids on system efficiency of an ORC using low-grade energy sources" Energy, Vol. 35, Pp. 1403–1411, 2010.

[5] B.F. Tchanche, G. Papadakis, G. Lambrinos and A. Frangoudakis, "Fluid Selection for a low temperature solar organic Rankine cycle" Appl. Therm. Eng., Vol. 29, Pp. 2468–2476, 2009.

[6] F. Heberle, M. Preibinger and D. Brüggemann, "Zeotropic mixtures as working fluids in organic Rankine cycles for low-enthalpy geothermal resources", Renew. Energy, Vol. 37, Pp. 364–370, 2012.

[7] Q. Liu, Y. Duan and Z. Yang," Effect of condensation temperature glides on the performance of organic Rankine cycles with zeotropic mixture working fluids." Appl Energy, Vol. 115, Pp. 394–404, 2014.

[8] K. Yang, H. Zhang, Z. Wang, J. Zhang, F. Yang and E. Wang," Study of zeotropic mixtures of ORC (organic Rankine cycle) under engine various operating conditions." Energy, Vol. 58, Pp. 494–510, 2013.

[9] X. D. Wang and L. Zhao," Analysis of zeotropic mixtures used in low-temperature solar Rankine cycles for power generation." Sol Energy, Vol. 83, Pp. 605–13, 2009.

[10] A. Dwivedi and R. S. Mishra," Performance Enhancement of Simple Vapour Absorption System Using Loop Heat Pipes" *ISSN: 2321-9653* International Journal for Research in Applied Science & Engineering Technology (IJRASET**),** Volume 6 Issue I, Pp. 865-873, 2018.

[11] F. Korn, "Heat Pipes and its Applications" Project Report 2008 MVK160 *Heat and Mass Transport* May 07, 2008.

[12] P. D. Dunn, D.A. Reay, Heat Pipes, Pergamon Press, Oxford, 1993.

[13] C.P. Arora, Refrigeration and Air Conditioning, Tata McGraw-Hill Publishing Company Limited, Delhi, Third Edition, 2009.

[14] P. K. Nag, Basic & Applied Thermodynamics, Tata McGraw-Hill Publishing Company Limited, Delhi, Second Edition, 2017.

View publication stats

Analytical model to study the effect of alignment mechanism on the growth characteristics of carbon nanotubes in PECVD chamber

Suresh C. Sharma[1] and Umang Sharma[1]
[1] Delhi Technological University, Delhi, India
e-mail (speaker): prof_sureshsharma@dtu.ac.in

The authors have developed an analytical model to study the effect of alignment mechanism of carbon nanotubes (CNTs) in the plasma enhanced chemical vapor deposition (PECVD) system. The alignment mechanism and growth of CNTs in a methane hydrogen plasma are modelled in the current work. The significance of plasma parameters and bias potential of the substrate on the electric field of the plasma sheath is studied. The role of the electric field in the plasma sheath to provide the necessary electrostatic force on the CNTs alignment in the plasma are studied by taking into consideration the steady state fluid equations, kinetics of the plasma sheath, charged species dynamics and neutral atoms in a reactive plasma, kinetics of the catalyst particle, and creation of atomic species of carbon and hydrogen over the surface of catalyst nanoparticle and incorporating various processes vital for the growth of CNTs in plasma. Taking into account the initial conditions and the glow discharge parameters the numerical solutions for the equations are obtained. The outcomes of the studies revealed that the alignment of the nanotube depends on the force exerted at the tip of CNTs. This force is responsible for the plasma sheath electric field which is dependent on the plasma parameters and the bias voltage at the substrate. This electrostatic force is also dependent upon the dimensions of the nanotube growing in the plasma environment and modifies itself continuously conferring to the instantaneous nanotube measurement and alignment during the growth. The present study can be apprehended to manufacture CNTs aligned vertically in plasma for better applications in the areas of field emission devices. The results of the present study have been compared with the existing experimental observations.

References

[1] U. Sharma and S. C. Sharma, Phys. Plasmas **25**, 103509 (2018).
[2] Ravi Gupta, Neha Gupta and Suresh C. Sharma, Physics of Plasmas **25**, 043504 (2018)
[3] C. Bower, W. Zhu, S. Jin, and O. Zhou, Appl. Phys. Lett. **77**, 830 (2000)

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354603954

Artificial intelligence enables advanced financial modeling

Article · August 2021

CITATIONS
0

2 authors, including:



Tanveer Aslam

3 PUBLICATIONS **0** CITATIONS

SEE PROFILE

All content following this page was uploaded by Tanveer Aslam on 15 September 2021.

The user has requested enhancement of the downloaded file.

Artificial intelligence enables advanced financial modeling

Tanveer Aslam, Delhi Technological University, Delhi, India
Abhishek Sarkar, Delhi Technological University, Delhi, India

August 2021

Abstract

The rapid growth of AI in FinTech demonstrates AI's potential and significant developments for smart FinTech, economy finance, and society. AI-powered smart FinTech has emerged in AI, data sciences, finance, and other related research areas. This trend is based on the long history of AI-based finance. The new-generation AI/data science and machine learning are fundamentally transforming the vision and objectives, paradigms. AI makes it possible to create more personalized and sophisticated, safer, and more innovative mainstream and alternative financial mechanisms, products and models, services, systems, and applications. This review reviews the current research on AI finance. It focuses on creating an economic and financial problem-driven research landscape on the roles, research direction, and opportunities of AI for new-generation FinTech, finance.

Keywords: Artificial Intelligence, Financial Modeling, Machine Learning, Financial systems, financial applications, Machine Learning

Introduction

Financial modeling involves a lot of technology. Financial models, at this moment, also use the latest technology to ensure that simulations produce the most accurate results. Technology is rapidly improving over time. Machine learning, robotic process automation, artificial intelligence, and machine learning will all impact financial modeling in the future.

Artificial Intelligence is a science that integrates theory, technology, and method. Artificial Intelligence can extend and expand the human mind and create a similar thinking style to humans. Its purpose is to create an intelligent machine capable of simulating and exploring the behavior of humans. AI theory and technology are maturing over the years. This is evident in the rapidly expanding field of AI-based applications. AI is now more than a simulation for human behavior. It is a multi-functional composite discipline, which can be applied to multiple purposes. It is now known as the "hottest technology in early 21st Century", together with virtual reality and 3D printing technology (Buchanan, B. 2019).

AI has made significant progress in imitating human reasoning over the past 40 years. However, AI's tools have been limited to sequential processing, basic representations of knowledge, and logic. Modern AI refers to creating a computer with the same structure and processing abilities as the human brain. These results can then be used to build knowledge representations. They also allow for pattern recognition that is based purely on experience. Artificial Neural Networks (ANN) or neural computing is another name for this AI technology. This paper outlines the artificial neural system and its wide use in the financial market. We also briefly discuss the various services for artificial neural systems in financial model simulation and the successful implementation of artificial neural nets in various financial institutions.AI technology offers many advantages, but AI also has its downsides like all technologies with double-edged swords (Buchanan, B. 2019). Therefore, the risks inherent in technology must be considered when we promote AI technology development in the financial sector.

Data will be collected by automation

Right now, many details must be manually collected at the point of sale. The process can be tedious, time-consuming, and vulnerable to error. The process is time-consuming, tedious, and

vulnerable to errors. It is still necessary to collect data manually, as processors and sensors are still not widely available. But the internet has enabled sensors and processors to almost be integrated into every device. Companies now have large volumes of data that can be used to describe their processes. At the moment, sensors-equipped machines are costly. However, the technology will be much cheaper in the future. This will have significant implications for financial modeling. For example, companies are currently making vague estimates about their electricity costs (Warin et al., 2019). Companies can now monitor the consumption of electricity by using sensors fitted to their devices. Companies will thus be able to predict their electricity costs better and, if necessary, control them.

Financial Modelling with Big Data will use big data

Financial models only deal with limited amounts of data at the moment. Therefore, complexity is limited. This is expected to change in the future. As I mentioned above, financial models will be interconnected with many devices in the future. As a result, financial models will have to deal with an excessive amount of data. The next step is to find meaningful patterns in the data, which will enable decision-making. Big data is a great tool for this. At the moment, big data is primarily used to find customer and supplier data. As financial data grows in quantity, this technology will find new applications in financial modeling (Greene et al., 2017).

Models can be built using automation

Financial modeling is limited in the current moment by the skill of those creating it. Financial modeling has many challenges, as there are multiple variables to consider and complex interactions. Human minds are not equipped to deal with these complex relationships and only express them in mathematical form.

Computers can do complex calculations, however. Computers should therefore be able to create better models, as they can perform millions of calculations in record time. Many templates are available that companies can quickly use to create financial models. These templates, however, are not very advanced and can still be modified by humans (Greene et al., 2017). As a result, future financial models that can be fully automated without manual intervention will be available right out of the box.

Models Will be Updated Using Automation

Humans currently interpret the data from models. They interpret data and make improvements to improve model performance. Shortly, artificial intelligence will be part of financial models. This will enable computers to assess the success of their models and make predictions about future ones. The system will compare actual results with the predicted ones.

Conclusion

AI is one of today's most innovative and forward-looking investments in international financial institutions. AI has been introduced to financial asset trading. Wealth and asset management. Insurance and banking. Customer service. An enterprise is a complex system. An enterprise has so many interactions that it cannot all be understood fully without a detailed model. The future will help us make complex decisions and understand this complex environment using advanced technologies. The computational base for modeling has already been established. Now, it's time to create an environment where data is automatically fed into, and the system learns. Today's financial modelers are facing a major challenge: how to integrate artificial intelligence into financial modeling.

References

Kunnathuvalappil Hariharan, N. (2018). "Artificial Intelligence and human collaboration in financial planning", *International Journal of Emerging Technologies and Innovative Research*, Vol.5, no. 7, Jul. 2018, pp 1348-1355.

Buchanan, B. (2019). Artificial intelligence in finance.

Greene, D., & Cross, J. P. (2017). Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, *25*(1), 77-94.

Kunnathuvalappil Hariharan, N. (2017). "Predictive model building for driver- based budgeting using machine learning", *International Journal of Emerging Technologies and Innovative Research*, Vol.4, no. 6, Jun. 2017, pp.567-575.

Lu, Y. (2019). Artificial intelligence: a survey on evolution, models, applications and future trends. *Journal of Management Analytics*, *6*(1), 1-29.

Warin, T., & Stojkov, A. (2021). Machine Learning in Finance: A Metadata-Based Systematic Review of the Literature. *Journal of Risk and Financial Management*, *14*(7), 302.

Kunnathuvalappil Hariharan, N. (2017). “Driver-based model- trends factors affecting its adoption", *International Journal of Emerging Technologies and Innovative Research*, Vol.4, no. 10, Oct. 2017, pp.865-869.

View publication stats

Automated Threat Objects Detection with Synthetic Data for Real-Time X-ray Baggage Inspection

Kunal Chaturvedi[1], Ali Braytee[2], Dinesh Kumar Vishwakarma[1], Muhammad Saqib[2], Domingo Mery[3], Mukesh Prasad[2]

[1] Department of Information Technology, Delhi Technological University, India
[2] School of Computer Science, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[3] Department of Computer Science, Pontificia Universidad Catolica de Chile, Chile

***Abstract*— With the recent surge in threats to public safety, the security focus of several organizations has been moved towards enhanced intelligent screening systems. Conventional X-ray screening, which relies on the human operator is the best use of this technology, allowing for the more accurate identification of potential threats. This paper explores X-ray security imagery by introducing a novel approach that generates realistic synthesized data, which opens up the possibility of using different settings to simulate occlusion, radiopacity, varying textures, and distractors to generate cluttered scenes. The generated synthetic data is effective in the training of deep networks. It allows better generalization on training data to deal with domain adaptation in the real world. The extensive set of experiments in this paper provides evidence for the efficacy of synthetic datasets over human-annotated datasets for automated X-ray security screening. The proposed approach outperforms the state-of-the-art approach for a diverse threat object dataset on mean Average Precision (mAP) of region-based detectors and classification/regression-based detectors.**

Keywords*— *X-ray, Threat, Baggage, Synthetic data, Screening, Object detection.

I. INTRODUCTION

Security at public places has been synonymous with the safety of an individual. Massive systems are deployed to safeguard the aviation industry, goods and passenger transportation, and high-risk zones such as defense areas, etc. X-ray security screening is one such system that enables inspection through the scanning of baggage, cargo, vehicle, and full-body scanning [1]. This paper focuses on the X-ray screening for baggage inspection. For accurate identification of suspicious items, security screening requires continuous vigilance from security personnel. However, the limited attention span of humans and the presence of occluded substances in cluttered X-ray scans adversely affect the detection rate of suspicious items. Due to the multitude of suspicious items, there is often confusion in the decision making to identify such items in each baggage, thus making it a time-consuming and tedious operation. Every year millions of baggage go through security screening at checkpoints. To prevent delays during rush hour and consider the number of items to be screened, the baggage screening must be highly effective and accurate. This could be achieved using a robust automated baggage inspection mechanism to assist human inspectors.

Object detection plays a significant role in identifying particular objects, such as distinguishing between threat and non-threat objects. It is a challenging process as it incorporates the two most important tasks: image classification and localization of objects with bounding box coordinates. The advent of deep learning techniques using Convolutional Neural Networks (CNN) replaced conventional computer vision techniques for X-ray baggage screening [2]. However, because of the unavailability of data, the CNN models may suffer from overfitting problems. Unfortunately, the dataset is not readily accessible in the X-ray security imagery domain due to privacy concerns. The available datasets are imbalanced, containing fewer instances of threat objects. This issue can be resolved using synthetic data generation, which is better than augmentation techniques. Synthesized data allows introducing new classes of prohibited items, variation in poses, scales, occlusion, and training on such dataset contributes to highly efficient detectors. This paper performs threat detection using the techniques that can be categorized into the one-step framework and two-steps framework. The one-step framework is based on the traditional object detection pipeline, which at first generates region proposals, and then classifies them into separate objects, and the two-steps framework adopts a unified framework to address object detection as a regression or classification problem called single-shot detectors. Taking into consideration that each technique has its drawback, there must be a trade-off between the two approaches to make screening fast and accurate.

The main contributions of the work are as follows:

- It generates a Synthesized annotated dataset through a mechanism that requires minimal effort and provides good visual coverage appearance of the prohibited items.
- It presents a comparative study of region proposal-based and regression/classification-based object detection pipelines for X-ray security.
- It conducts extensive experiments on the benchmark dataset, namely GDXray [3].

The organization of the remaining paper is as follows. Section II discusses the related work. The methodology is described in detail in Section III, followed by experiment results in Section IV. Finally, Section V concludes the paper with future work.

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9533928

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

II. Related Works

This section briefly addresses the previous methodologies for X-ray security imagery and approaches for generating synthetic data. Security screening has become a major concern for organizations across the world due to a massive increase in the influx and efflux of passengers, goods, and cargo in the past decades. Due to this recent surge of interest in the field of X-ray security imagery [4], conventional machine learning approaches for classification [5] [6] [7], segmentation [8], [9], and detection [10] have been extensively studied to automate the process. A majority of previous classification work was carried out using bag of visual words (BoVW) [5] approach with SVM classifier, and sparse representations [6]. A transfer-learning based approach [11] was proposed for threat object classification. Hu et al. [12] proposed Security X-ray Multi-label Classification Network (SXMNet) to deal with overlapping in X-ray images classification. However, the object detection is more beneficial over the classification of objects for the detailed analysis of the contents in the X-ray imagery of baggage. Automated localization with bounding box dimensions and identification of threat object assists security officers to avert bizarre human-made disasters. Franzel et al. [13] conducted experiments on multiple-view X-ray imagery and compares it to single-view detection that demonstrated superior detection efficiency over single-view detection for handguns. With the advent of deep learning in object detection, CNN based frameworks replaced the approaches based on hand-crafted features. Samet et al. [14] introduced the region-based object detection methodologies, Faster RCNN [15] for this purpose, but the analysis does not take into account the speed of the detector, which is important for rapid threat detection. Dhiraj et al. [16] conducted experiments on the X-ray scans of the GRIMA X-ray Database (GDXray) that proposed by Mery et al. [3]. The dataset consists of instances of guns, shuriken, razor blades, and knives using different object detection frameworks.

However, the training of deep learning models becomes challenging due to the restricted availability of private security screening datasets and the skewed ratio of non-threatening objects to threatening objects. Different strategies have been adopted to compensate for the data for X-ray security imagery by generating synthetic dataset using General Adversarial Networks (GAN) [17] [18] [19], Threat Image Projection [10], and Logarithmic X-ray Imaging Model [20]. Yang et al. [17] and Zhu et al. [18] used GANS that are limited to generating synthetic instances of foreground objects including handgun, knife, etc. ignoring complex baggage information. These approaches do not take into consideration of cluttered background situations, including noise, and the use of non-threatening objects. Most of the recent literature on GANs [21][22] revolves around learning a mapping from a single image source, i.e. image translation that restricts the ability of full control over the rotation, scaling, radiopacity, occlusion, etc. This paper generates simulated X-ray images using real background and object instances with operations that allow full user control over the rotation, scaling, radiopacity, occlusion, etc. on the foreground object to preserve both realism and domain adaptation.

III. Methodology

Due to the lack of available data, the proposed approach (AXSD) aims to generate accurate synthesized images from an existing dataset GDXray. The generated images are evaluated using different object detection methods.

A. AXSD: Automated X-ray Synthetic Data generation

This section presents an automated X-ray synthetic data generation (AXSD) to minimize the empirical limits with the availability of enormous training data for deep learning based object detection. AXSD consists of two components: data collection and object superposition.

i. Data Collection

In the first component, the sampled data are randomly gathered of X-ray images of empty baggage and instances of threat/non-threat objects to perform the experiments in this paper. These instances of objects are extracted using the methodology described as shown in Fig. 1.

Edge Detection: As the first step, for foreground/background segmentation, we use a cellular nonlinear network-based [23] edge detector to predict fine edge-map from the foreground X-ray image. The network takes up three arguments: input image, number of iterations, and a combination of 19 parameters, defined as cloning template. Here, these parameters under optimization are iteratively adjusted so that its output converges to the output of the ideal edge detector. After the template learning, the optimised detector slides through the pixels, computing operations to evaluate edges. The predicted edge-map is the union of edge sets detected along the x and y directions respectively.

Noise Removal: Pre-processing technique such as median blur is applied on the predicted edge-map to eliminate noise. This step is necessary before contour detection, which generates a hierarchy of contours.

Foreground Extraction: Further, the graph-based GrabCut algorithm [24] is applied to the contour with the largest area to extract the foreground mask, which in our case, belongs to the instances of the threat item / non-threat item.

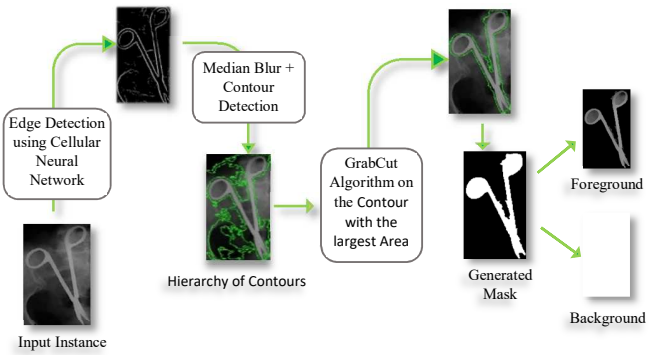


Fig. 1 Illustration of the proposed foreground/background methodology for the extraction of the instances of prohibited items.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

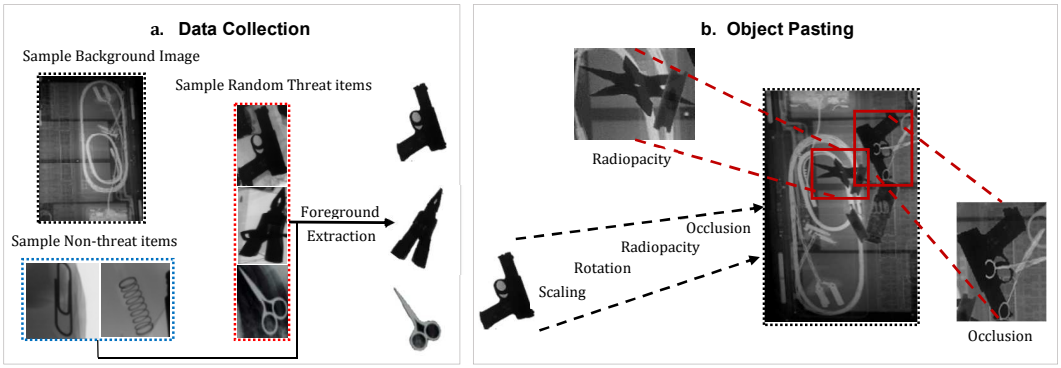


Fig. 2. A detailed explanation of our approach for a) Data collection and b) Object superposition

ii. Object Superposition

The extracted foreground items are superimposed randomly within a small section of the luggage in the background X-ray images. Contour detection is used to determine the boundary regions of the luggage. Due to the limited data, data augmentation techniques are applied to the foreground items to generate synthetic images. The main steps of the proposed method are shown in Fig. 2b and described as follows:

Scaling: A random scale with values ranging from 0.5 to 1 is chosen based on the size of the foreground object and the background luggage image.

Rotation2D: Different 2D viewpoints from the same object with an angle of rotation from -180 to 180 degrees are possible inside the baggage as the position of the object does not matter.

Non-Threat Items: Multiple distractors are superimposed onto the baggage X-ray image to simulate real-world scenarios to prevent bias in the training algorithm.

Radiopacity: The relative inability of electromagnetic radiation to pass through a material varies by thickness and the form of the material. These conditions are simulated by adjusting the alpha channel as shown in Fig. 3.

Occlusion: Partial overlapping with a maximum Intersection over Union (IOU) of 0.75 between different threat items and non-threat items is used when superimposing it on the background image to generate a genuine simulated image. The use of occlusion with the radiopacity principle on both threat objects and distractors increases the X-ray image's scene realism while maintaining global consistency with the existing dataset. A summary of the proposed framework is shown in Fig. 4.



Fig. 3. The radiopacity parameter is adjusted according to the intermediate values between 0 and 1. The higher the value, the less radiopaque the object will be.

B. Evaluate the synthetic data using object detection methods

We evaluated different object detection strategies. These detectors are based on the categorization of the conventional object detection system, which evaluates the regional proposals and then classifies each proposal as a threat or non-threat element. Another one uses a regression or classification problem-based unified framework to evaluate bounding-box coordinates and classification probabilities.

i. Region Proposal based Architectures

Faster RCNN proposed by Ren et al. [15] substitute the selective search algorithm with the introduction of the Region Proposal Network (RPN). It assigns a score to each anchor box that is of varying size and scale generated by sliding a fixed-sized window

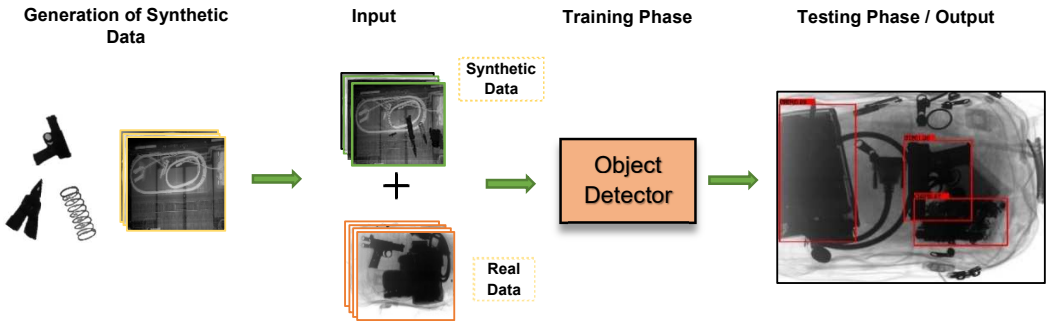


Fig. 4. Overview of our proposed approach. We first use the object superimposition to generate realistic data samples. Then, we use a combination of real and generated synthetic data for the object detection pipeline.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

over the activation maps generated by the base network's last convolutional layer. After applying an ROI pooling layer, the proposals generated are passed onto the RCNN. It uses two different fully connected layers that output the bounding box and classifies whether a suspicious entity is present or not. Any redundant entity (overlapping bounding box) is further eliminated using non-maximum suppression (NMS).

Cascade RCNN [25] developed by Cai et al., is a multi-staged framework based on the popular RCNN. At each stage, cascaded bounding box regression combined with a sequence of detectors operates by changing the bounding boxes to be more selective against the close false positives used in the next stage of training. This reduces the risk of overfitting during training and overcomes a quality mismatch at inference, ensuring high-quality detection performance over Faster RCNN when the items are placed next to one another in the luggage.

ii. Classification/Regression based Architectures

Single Shot MultiBox Detector (SSD) [26] by Liu et al. uses a unified framework for both localizing objects and classification, eliminating the need for a proposal stage. It also resolves the issue of low-resolution object detection within the YOLO [27] framework by incorporating the idea of making faster predictions for bounding boxes and their labels at each feature map than Faster RCNN [15]. Feature maps of varying scales allow the SSD to detect objects of different scales and sizes.

RetinaNet [28] addresses the issue of the class imbalance (foreground/background) problem faced by unified object detection frameworks with a reconfiguration of cross-entropy loss. Focal loss decreases the share of high-probability detections that can dominate the model's performance. It incorporates the feature pyramid network, a regression subnetwork that evaluates bounding box coordinates using a regression head and classification subnetwork to allocate classes to these boxes.

IV. Experiment Results and Discussions

In this section, a comprehensive study is carried out on different combinations of synthesized data and human-annotated data using region-based and classification/regression-based CNN frameworks. Average precision (AP) is used as the evaluation criteria for the different frameworks and precision, and recall is determined by thresholding the Intersection-over-Union (IoU) for detection. The dataset is split into a ratio of 70:30 for training and testing, respectively. The number of epochs is fixed to 50 and the batch size to 4. The PyTorch framework is used for all the experiments. We first describe the characteristics of the dataset. We then report the results of our method using different proportions of synthetic images mixed

Table 1. Comparison of $GDXray$ dataset and $GDXray + Samples\ of\ Synthetic_{\alpha}$ dataset based on the average precisions (AP) metrics using architectures based on the two-stage framework and the one-stage framework.

Model	Backbone Network	Dataset	Average Precision							mAP
			Mobile	Chip	Shuriken	Revolver	Gun	Knife	Blade	
Faster RCNN	ResNet50	GDXray	0.907	0.907	0.917	0.996	0.909	0.905	0.909	0.921
		10% syn	0.910	0.924	0.925	0.979	0.899	0.847	0.913	0.913
		20% syn	0.915	0.925	0.931	0.979	0.902	0.873	0.920	0.920
		30% syn	0.915	0.919	0.932	0.981	0.909	0.898	0.921	0.925
		40% syn	0.903	0.908	0.931	0.999	0.952	0.948	0.921	**0.937**
		50% syn	0.917	0.915	0.929	0.968	0.909	0.901	0.922	0.923
Cascade RCNN	ResNet50	GDXray	0.908	0.904	0.909	0.995	0.906	0.907	0.913	0.920
		10% syn	0.891	0.891	0.902	0.984	0.960	0.899	0.901	0.918
		20% syn	0.899	0.902	0.903	0.989	0.969	0.903	0.901	0.922
		30% syn	0.907	0.901	0.914	0.991	0.959	0.937	0.911	0.931
		40% syn	0.904	0.907	0.929	0.999	0.961	0.954	0.923	**0.939**
		50% syn	0.901	0.906	0.911	0.999	0.961	0.939	0.913	0.932
SSD512	VGG16	GDXray	0.899	0.888	0.897	0.941	0.928	0.834	0.907	0.889
		10% syn	0.894	0.857	0.897	0.901	0.914	0.817	0.878	0.879
		20% syn	0.903	0.833	0.901	0.898	0.925	0.851	0.879	0.884
		30% syn	0.912	0.878	0.915	0.906	0.918	0.863	0.901	0.899
		40% syn	0.903	0.903	0.907	0.964	0.907	0.859	0.909	**0.907**
		50% syn	0.904	0.903	0.897	0.901	0.906	0.858	0.909	0.896
RetinaNet	ResNet50	GDXray	0.905	0.893	0.909	0.953	0.933	0.842	0.909	0.906
		10% syn	0.908	0.875	0.906	0.907	0.904	0.864	0.874	0.891
		20% syn	0.912	0.884	0.913	0.903	0.931	0.844	0.918	0.900
		30% syn	0.925	0.915	0.912	0.922	0.926	0.856	0.924	0.911
		40% syn	0.911	0.907	0.929	0.970	0.907	0.877	0.919	**0.917**
		50% syn	0.915	0.916	0.915	0.927	0.926	0.871	0.927	0.913

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

Table 2. Comparison results of the architectures trained on $Synthetic_{\alpha}$ dataset with 10% and 20% of GDXray dataset used as test samples.

Model	Backbone Network	Test Real Dataset	Average Precision							
			Mobile	Chip	Shuriken	Revolver	Gun	Knife	Blade	
Faster RCNN	ResNet50	10% syn	0.908	0.903	0.967	0.936	0.914	0.848	0.856	0.904
		20% syn	0.899	0.839	0.929	0.909	0.903	0.801	0.848	0.875
Cascade RCNN	ResNet50	10% syn	0.877	0.890	0.969	0.949	0.935	0.885	0.842	0.906
		20% syn	0.885	0.861	0.913	0.916	0.882	0.795	0.831	0.869
SSD512	VGG16	10% syn	0.853	0.850	0.947	0.931	0.901	0.806	0.813	0.871
		20% syn	0.888	0.856	0.887	0.901	0.901	0.834	0.762	0.861
RetinaNet	ResNet50	10% syn	0.893	0.873	0.946	0.933	0.901	0.808	0.832	0.883
		20% syn	0.874	0.849	0.907	0.891	0.915	0.820	0.799	0.865

Table 3. Comparison of the average precisions (AP) on $GDXray + Samples\ of\ Synthetic_{\beta}$ dataset, consisting of additional classes of prohibited items.

Model	Backbone Network	Dataset	Average Precision										mAP
			Mobile	Chip	Shuriken	Revolver	Gun	Knife	Blade	Pliers	Scissor	Hammer	
Faster RCNN	ResNet50	30% syn	0.901	0.892	0.896	0.938	0.875	0.858	0.887	0.894	0.847	0.968	0.895
		40% syn	0.909	0.909	0.908	0.997	0.905	0.907	0.909	0.909	0.896	1.000	**0.924**
		50% syn	0.905	0.879	0.936	0.999	0.901	0.877	0.901	0.910	0.833	0.971	0.911
Cascade RCNN	ResNet50	30% syn	0.878	0.865	0.899	0.931	0.802	0.886	0.901	0.928	0.889	0.979	0.895
		40% syn	0.908	0.907	0.911	0.997	0.899	0.908	0.909	0.989	0.905	0.998	**0.933**
		50% syn	0.917	0.898	0.908	0.989	0.901	0.902	0.899	0.967	0.911	0.989	0.928
SSD512	VGG16	30% syn	0.758	0.846	0.846	0.891	0.799	0.812	0.863	0.872	0.829	0.939	0.845
		40% syn	0.779	0.867	0.898	0.906	0.860	0.837	0.899	0.921	0.896	0.994	0.885
		50% syn	0.791	0.878	0.902	0.899	0.889	0.857	0.891	0.934	0.903	0.999	**0.894**
RetinaNet	ResNet50	30% syn	0.821	0.861	0.881	0.959	0.896	0.857	0.865	0.889	0.891	0.939	0.885
		40% syn	0.814	0.902	0.906	0.992	0.910	0.878	0.902	0.906	0.884	0.996	0.909
		50% syn	0.832	0.879	0.912	0.981	0.917	0.891	0.889	0.910	0.901	0.999	**0.911**

with real images. Finally, we compare our proposed method with the state-of-the-art.

A. Datasets

The Grima X-ray dataset (GDXray) was collected by Mery et al. [3]. The dataset consists of five groups of X-ray scans with more than 21,100 images: castings, welds, baggage, nature, and settings. However, for this research work, we used the baggage group consisting of 1371 X-ray images of prohibited objects. The prohibited items include five classes i.e., shuriken, knives, blades, firearms (guns and revolvers), and electronic items (chips and mobiles). Chips and mobiles are taken into consideration as they are restricted in high-risk environments such as defense areas.

In this work, an exhaustive set of experiments are conducted to choose the best performing synthetic dataset. For this purpose, we evaluate the effectiveness of combining the fixed ratio of the synthetic dataset with the real dataset. We synthesized and created two types of datasets of the same size as the GDXray dataset, $Synthetic_{\alpha}$ and $Synthetic_{\beta}$ using the same dataset generation procedure discussed earlier and randomly took samples in the ratio of 50%, 40%, 30%, 20%, and 10% from $Synthetic_{\alpha}$ and 30%, 40%, and 50% from $Synthetic_{\beta}$. $Synthetic_{\alpha}$ is composed of synthetically generated instances of the classes only from the GDXray dataset and $Synthetic_{\beta}$ consists of three additional classes, namely pliers, scissors, and hammers. The $GDXray + Synthetic_{\beta}$ dataset contains the largest number of classes of prohibited items available.

B. Results

As shown in Table 1 and Table 3, multiclass detection is performed, and the performance is compared in terms of average precision (AP) for each class of prohibited item using different architectures i.e., Faster RCNN, Cascade RCNN, RetinaNet with ResNet50 as configuration and SSD512 with VGG16 [29] as configuration. Table 1 investigates and compares the $GDXray$ dataset and the $GDXray + Synthetic_{\alpha}$ dataset and Table 3 evaluates the effectiveness of the $GDXray + Synthetic_{\beta}$ a dataset comprising ten classes of prohibited items including three additional classes. Table 1 shows that the networks trained on $GDXray +$ samples of $Synthetic_{\alpha}$ dataset achieve a performance gain up to ~2% over the $GDXray$ dataset. As shown in Table 1, combination of $GDXray$ and 40% of $Synthetic_{\alpha}$ resulted in a better and more robust performance than a network trained on $GDXray$ alone for all the architectures.

Moreover, the performance of the models is significant as shown in Table 2 when using a synthetic dataset as the training set and the real dataset as the test set. We used the $Synthetic_{\alpha}$ dataset and randomly chose 10% and 20% of GDXray as the test set. One of the crucial factors observed was the domain

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

adaptation when using the synthetic dataset. The results show that the generalization learned through the model with the use of only synthetic data overcomes the issue of domain adaptation on the test set.

Table 3 shows that the two-stage frameworks, namely Faster RCNN and Cascade RCNN reported higher mAP with 40% of $Synthetic_{\beta}$ Dataset + GDXray while the one-stage frameworks achieved better mAP on 50% of $Synthetic_{\beta}$ Dataset + GDXray. This result is attributed to the increased amount of data for the classes pliers, scissors, and hammers which resulted in inefficiently trained networks. Also, the source background image and target foreground image for the additional classes of $Synthetic_{\beta}$ are obtained using a different X-ray film scanner. However, the networks resulted in superior performance over the $Synthetic_{\beta}$ dataset + GDXray. The Cascade RCNN architecture with ResNet50 achieved the best performance over the $GDXray + 40\%\ of\ Synthetic_{\beta}$ dataset with a mAP of 0.933, while SSD512 performed the worst with a mAP of 0.885.

Table 4. Comparison of average precision (AP) by omitting different settings on $GDXray + 40\%\ Synthetic_{\alpha}$ dataset

Model	No Radiopacity	No Augmentation	No Non-Threat Object	All Operations
Faster RCNN	0.931	0.899	0.926	0.937
Cascade RCNN	0.928	0.902	0.924	0.939
SSD512	0.899	0.879	0.896	0.907
Retinanet50	0.911	0.891	0.910	0.917

A maximal AP of 0.910 for the class gun was observed on the RetinaNet. The visual results are shown in Fig. 6, using different object strategies on the GDXray and Synthesized dataset.

Table 4 shows the results obtained by omitting different settings of our synthetic data generation process. A minor decrease in the mAP for all the architectures was observed with the exclusion of the radiopacity principle in the images. However, the decrease in mAP was significant when augmentation techniques such as rotation and scaling, etc. are not applied. The results also emphasize the importance of the use of distractors which include non-threat objects to force the network to learn to ignore the nearby patterns.

Further, we evaluate and compare different object detectors on the $GDXray + 40\%\ of\ Synthetic_{\beta}$ dataset using the precision-recall curve. For a comparison of the region-based detector and the classification/regression-based detectors, both Faster RCNN and Cascade RCNN outperformed SSD and RetinaNet using the metrics of mAP. We also evaluate and report the inference speed of different models using Nvidia RTX 2080Ti on the GDXray dataset in Table 5. SSD512 based on VGG16 outperformed Faster RCNN, Cascade RCNN, and RetinaNet in speed, however Faster RCNN with 21.8 fps surpassed RetinaNet in both accuracy and speed, which is crucial for security screening. Although Cascade RCNN is comparable to Faster RCNN in terms of mAP, it is not feasible to use it due to its low inference speed. Also, it was observed that the training on the synthesized dataset is challenging due to the diverse nature of the prohibited objects with a diverse cluttered background in the dataset which enabled the network trained on this dataset to be better generalized.

Table 5. Comparison of different models in terms of inference speed on GDXRAY dataset using Nvidia RTX 2080 Ti with batch size 4.

Model	Backbone	Inf(fps)
Faster RCNN	ResNet50	21.8
Cascade RCNN	ResNet50	13.7
SSD512	VGG16	26.1
RetinaNet	ResNet50	16.5

C. Comparison with generative models

The proposed method AXSD is compared with the state-of-the-art generative adversarial models [21][22]. The officially released code is used to produce the results of the GAN models.

BigGAN [22] is a class-conditional GAN model for high fidelity natural image synthesis. The model emphasizes the benefits of scaling up both the batch size and the model size for high-quality image generation. However, there is a lack of

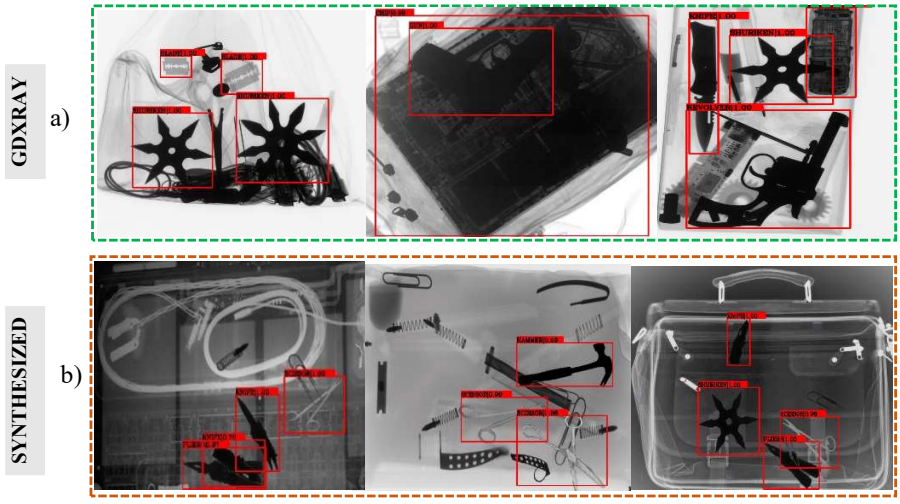


Fig. 6. Performance of the different threat detection pipelines on exemplar images from a) GDXRAY dataset and b) Synthesized dataset. It was observed that the frameworks achieved superior results for cluttered scenes, occluded objects, and illumination changes.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

diversity in the generated images which is attributed to the fact that the model truncates large values of the noise vector to improve the image quality [30].

SinGAN [21] is an unconditional image generation technique that uses single image for serial training. The model consists of a pyramid of fully convolutional GANs, that is trained using a multi-resolution and multi-stage approach. It uses internal patch distribution within the single image to generate diversified samples. However, this method does not consider the relationship between the two images, thus ignoring distribution variations between the images.

Here, we use the Fréchet Inception Distance (FID) [31] as a measure of image quality. It compares the statistics of generated samples to the ground truth samples. A lower FID score indicates better model performance that generates images similar to the real images. As shown in Table 6, the proposed method outperforms the state-of-the-art GAN based methods with the lowest FID score. The GAN-based techniques are resource-intensive and less automated for the controlled synthesis process to deal with complex data.

Table 6. Comparison with the state-of-the-art GAN-based methods. Lower FID indicates better image quality.

Method	FID
BigGAN [22]	334.3
SinGAN [21]	80.6
Ours	**57.9**

V. Conclusions

This paper proposes a novel approach for generating synthetic datasets using different settings to address the scarcity of data in the X-ray imaging domain. The proposed approach generates highly realistic diversified synthetic X-ray scans with less cost and time that assists human-annotated datasets for training robust neural networks. The performance of the combinations of the GDXray and synthesized dataset is validated through an extensive set of experiments on two different detection pipelines. The results obtained are superior compared to the models trained on the real dataset. Also, this paper demonstrates that the performance of models trained purely on the synthetic dataset is comparable to the performance of models trained on the mixture of the real and synthesized dataset. In future, the proposed approach can be extended to other existing classification and detection problems for further improvements in the efficiency and robustness of the trained networks. Also, It can aim to generate realistic multi-view radiographs in the medical domain and implement few-shot learning techniques to reduce dependency on huge amounts of data.

References

[1] S. Akcay and T. Breckon, "Towards Automatic Threat Detection: A Survey of Advances of Deep Learning within X-ray Security Imaging," Jan. 2020, doi: 2001.01293.

[2] S. Akcay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sep. 2016, pp. 1057–1061, doi: 10.1109/ICIP.2016.7532519.

[3] D. Mery, V. Riffo, U. Zscherpel, G. Mondragón, I. Lillo, I. Zuccar, et al., "GDXray: The Database of X-ray Images for Nondestructive Testing," *J. Nondestruct. Eval.*, vol. 34, no. 4, pp. 1–12, Dec. 2015, doi: 10.1007/s10921-015-0315-7.

[4] D. Mery, D. Saavedra, and M. Prasad, "X-Ray Baggage Inspection With Computer Vision: A Survey," *IEEE Access*, vol. 8, pp. 145620–145633, 2020, doi: 10.1109/ACCESS.2020.3015014.

[5] M. Baştan, "Multi-view object detection in dual-energy X-ray images," *Mach. Vis. Appl.*, vol. 26, no. 7–8, pp. 1045–1060, Nov. 2015, doi: 10.1007/s00138-015-0706-x.

[6] D. Mery, E. Svec, and M. Arias, "Object Recognition in Baggage Inspection Using Adaptive Sparse Representations of X-ray Images," 2016, pp. 709–720.

[7] D. Mery, V. Riffo, I. Zuccar, and C. Pieringer, "Object recognition in X-ray testing using an efficient search algorithm in multiple views," *Insight - Non-Destructive Test. Cond. Monit.*, vol. 59, no. 2, pp. 85–92, Feb. 2017, doi: 10.1784/insi.2017.59.2.85.

[8] M. Singh and S. Singh, "Image segmentation optimisation for x-ray images of airline luggage," in *Proceedings of the 2004 IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety, 2004. CIHSPS 2004.*, pp. 10–17, doi: 10.1109/CIHSPS.2004.1360198.

[9] G. Heitz and G. Chechik, "Object separation in x-ray image sets," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2010, pp. 2093–2100, doi: 10.1109/CVPR.2010.5539887.

[10] N. Bhowmik, Q. Wang, Y. F. A. Gaus, M. Szarek, and T. P. Breckon, "The Good, the Bad and the Ugly: Evaluating Convolutional Neural Networks for Prohibited Item Detection Using Real and Synthetically Composited X-ray Imagery," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.11508.

[11] D. Mery, E. Svec, M. Arias, V. Riffo, J. M. Saavedra, and S. Banerjee, "Modern Computer Vision Techniques for X-Ray Testing in Baggage Inspection," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 47, no. 4, pp. 682–692, Apr. 2017, doi: 10.1109/TSMC.2016.2628381.

[12] B. Hu, C. Zhang, L. Wang, and Q. Zhang, "Multi-label X-ray Imagery Classification via Bottom-up Attention and Meta Fusion."

[13] T. Franzel, U. Schmidt, and S. Roth, "Object Detection in Multi-view X-Ray Images," 2012, pp. 144–154.

[14] S. Akcay and T. P. Breckon, "An evaluation of region based object detection strategies within X-ray baggage security imagery," in *2017 IEEE International Conference on Image Processing (ICIP)*, Sep. 2017, pp. 1337–1341, doi: 10.1109/ICIP.2017.8296499.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," Jun. 2015, [Online]. Available: http://arxiv.org/abs/1506.01497.

[16] Dhiraj and D. K. Jain, "An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery," *Pattern Recognit. Lett.*, vol. 120,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

pp. 112–119, Apr. 2019, doi: 10.1016/j.patrec.2019.01.014.

[17] J. Yang, Z. Zhao, H. Zhang, and Y. Shi, "Data Augmentation for X-Ray Prohibited Item Images Using Generative Adversarial Networks," *IEEE Access*, vol. 7, pp. 28894–28902, 2019, doi: 10.1109/ACCESS.2019.2902121.

[18] Y. Zhu, H. Zhang, J. An, and J. Yang, "GAN-based data augmentation of prohibited item X-ray images in security inspection," *Optoelectron. Lett.*, vol. 16, no. 3, pp. 225–229, May 2020, doi: 10.1007/s11801-020-9116-z.

[19] D. Saavedra, S. Banerjee, and D. Mery, "Detection of threat objects in baggage inspection with X-ray images using deep learning," *Neural Comput. Appl.*, 2020, doi: 10.1007/s00521-020-05521-2.

[20] D. Mery and A. K. Katsaggelos, "A Logarithmic X-Ray Imaging Model for Baggage Inspection: Simulation and Object Detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jul. 2017, pp. 251–259, doi: 10.1109/CVPRW.2017.37.

[21] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a Generative Model From a Single Natural Image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 4569–4579, doi: 10.1109/ICCV.2019.00467.

[22] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1809.11096.

[23] L. O. Chua and L. Yang, "Cellular neural networks: theory," *IEEE Trans. Circuits Syst.*, vol. 35, no. 10, pp. 1257–1272, Oct. 1988, doi: 10.1109/31.7600.

[24] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut,'" *ACM Trans. Graph.*, vol. 23, no. 3, p. 309, Aug. 2004, doi: 10.1145/1015706.1015720.

[25] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.09756.

[26] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," Dec. 2015, doi: 10.1007/978-3-319-46448-0_2.

[27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Jun. 2015, doi: 1506.02640.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," Aug. 2017, [Online]. Available: http://arxiv.org/abs/1708.02002.

[29] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556.

[30] A. Razavi, A. van den Oord, and O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," Jun. 2019, [Online]. Available: http://arxiv.org/abs/1906.00446.

[31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," Jun. 2017, [Online]. Available: http://arxiv.org/abs/1706.08500.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:36:29 UTC from IEEE Xplore. Restrictions apply.

Water Air Soil Pollut (2021) 232:372
https://doi.org/10.1007/s11270-021-05301-1

Behavioural Mechanisms of Microplastic Pollutants in Marine Ecosystem: Challenges and Remediation Measurements

Megha Bansal · Deenan Santhiya · Jai Gopal Sharma

Received: 19 January 2021 / Accepted: 9 August 2021
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2021

Abstract Plastic pollution is the biggest threat to marine ecosystem owing to its high rates of disposal and low recovery from the environment. Due to inefficiency in degradation, most of plastic is fragmented into microplastics that are reported as ubiquitous toxicants in marine environment. The abundance of toxic microplastics in marine ecosystem causes adverse impacts on aquatic flora and fauna including oceans, lakes, rivers, coastal areas, and seas. This aggravates its toxicity and induces genomic instability, oxidative stress and disruption of marine organisms. Hence, it is necessary to understand the potential sources, types and behaviour of microplastic in marine environment. In this review, considering the pollution of aquatic ecosystem, major contributors of microplastics in marine environment along with their classification are brought out. Also, behaviour mechanisms of microplastics including physical, chemical and biological behaviours together with their ecological and toxicological impacts on marine ecosystem are illustrated. Finally, the remediation measures to combat against toxic microplastic pollution in aquatic ecosystem are highlighted to bring out an instant remedy for the environment.

Keywords Microplastics · Marine ecosystem · Toxicants · Behaviour · Remediation

Highlights
- Microplastics alter the characteristics of marine environment.
- The types and sources of microplastics in marine ecosystem are highlighted.
- Behavioural mechanisms of microplastics on marine ecosystem are illustrated.
- Potential remediation measures are proposed owing to impact of microplastics.

M. Bansal · J. G. Sharma
Department of Biotechnology, Delhi Technological University, Delhi, India
e-mail: megha_mt2k19@dtu.ac.in

J. G. Sharma
e-mail: sharmajaigopal@dce.ac.in

D. Santhiya (✉)
Department of Applied Chemistry, Delhi Technological University, Main Bawana Road Delhi, Shahbad Daulatpur 110042, India
e-mail: deenan.santhiya@dce.ac.in

1 Introduction

Plastic pollution is considered as major threat to marine environment affecting large number of aquatic organisms. The marine ecosystem has ubiquitous presence of plastic contaminants, prevailing on surface of oceans, on seabed and in water columns (Ritchie & Roser, 2018). Globally, the production of plastic has reached around 360 million tonnes of which a significant proportion is directly thrown into oceans without proper waste management (Jambeck et al., 2015). The major contributors to plastic waste in oceans originate from land-based sources including

Published online: 31 August 2021

landfill operations, agricultural activities, construction and waste released from industries (Geyer et al., 2017). Moreover, various bio solids leached from wastewater treatment plants also contribute substantial proportion of plastic fragments. The microbeads in cosmetic products and fibres also cause pollution of plastics in aquatic ecosystem (Mason et al., 2016). The pollution due to plastic accumulation in marine ecosystem is identified as the most prominent contaminant owing to its properties of flexibility, durability, low cost, corrosion resistance and easy handling (Botterell et al., 2019). In spite of wide range of applications of plastic worldwide, the issues related to its contamination and adverse impact on marine environment cannot be overlooked. Among the major plastic particles present in aquatic ecosystem, smaller fragments have garnered greater concern because of their ability to get ingested by marine organisms. The large plastic particles in marine environment are broken down into smaller pieces by ultraviolet degradation, physical abrasion and wave action, eventually forming microplastics (Dolatabadi & Ahmadzadeh, 2020). Approximately, 90% of plastic waste in oceans is microplastics owing to their very small size of less than 5 mm (Auta et al., 2017). The abundance of microplastics in marine environment poses a potential threat to aquatic flora and fauna with significant adverse impacts on oceans, lakes, rivers, coastal areas and seas.

1.1 Importance to Explore the Behaviour of Microplastics in Marine Environment

Anthropogenic activities and waste from rivers, winds, sewage sludge and water runoff cause accumulation of microplastics in marine environment. Also, tourism and recreational activities at shoreline create large amounts of waste litter that paves its pathway into aquatic water bodies (Avio et al., 2017). Military and research fleets, offshore installations and commercial vessels also cause detrimental effects on seashore (Galgani et al., 2015). Not only they cause damage to water, but they also get dispersed in soil and air causing threat to biodiversity (Wang et al., 2021). Plastic waste reach coasts by influence of water currents and meteorological conditions that cause spatial distribution of litter at coastline which further enter in oceans (Ourmieres et al., 2018). On entering the water column, microplastics having higher density like polyvinyl chloride tend to sink, whereas lower density microplastics like polyethylene and polypropylene tend to float in water. Change in density of particles along water column occurs due to colonization and biofouling of organisms on microplastic surface causing them to collapse at sediments (Avio et al., 2017). Water quality indices including pH, biological oxygen demand, dissolved oxygen and total nitrogen content change when exposed to microplastics (Kataoka et al., 2019). These characteristics of water significantly alter the biological matrix and cause destruction in environment.

Microplastic pollution in marine ecosystem causes potential threats and disorders in aquatic species because of their ingestion and entrapment (Bellasi et al., 2020; Galloway et al., 2017). Microplastics offer more surface area-to-volume ratio for accumulation of various contaminants including toxic metals and polychlorinated biphenyls (PCB) (Ozcan et al., 2013). These chemicals can bio accumulate in biological tissue and cause adverse effects in aquatic food chains. Moreover, added chemicals and additives in plastic manufacturing and organic pollutants are real threats to marine organisms (Hong et al., 2018). Microplastics bioaccumulation in marine environment increases with decreasing size and serve as a surface for proliferation of bacterial pathogens (Michielssen et al., 2016). Hence, in order to minimize the risks linked with ingestion of microplastic particles, it becomes essential to identify potential response and behaviour of microplastics in aquatic ecosystem.

In this review, major sources for various types of microplastic fragments in marine ecosystem are highlighted. Behavioural characteristics of microplastics focusing on physical, chemical and biological mechanisms are discussed to identify their fate in marine environment. In addition, potential impacts of microplastics on marine environment along with plausible remediation measures to overcome microplastic pollution have been provided.

2 Sources of Microplastics in Marine Ecosystem

Prevalence and abundance of microplastic particles in marine environment can be attributed to various sources. The major sources of microplastic can be attributed as primary microplastic that results from

direct release of pellets or powders and microbeads from cosmetic formulations, household products and raw materials from industries, whereas secondary microplastics arise from fragmentation of large plastic particles by ultraviolet radiation, oxidative stress and microbial degradation (Thompson, 2015). Also, evidences suggest the littering of microplastics in marine environments from plastic product manufacturing and waste management industries, from household activities and other commercial sectors. Thus, different source sectors provide substantial amount of microplastics in aquatic ecosystems. A brief summary of all the sources of microplastics from different usage sectors polluting the marine environment is depicted (Fig. 1).

3 Types of Microplastics in Marine Environment

Microplastics are classified into various types because of their different size and shapes (Guzzetti et al., 2018). Additionally, the toxicity and adverse effects of microplastics in marine ecosystem make it necessary to identify them. Hence, microplastics are categorized on the basis of origin, their source and on basis of properties.

3.1 Primary and Secondary Microplastics

This is the most common method to classify microplastics depending on their initial sizes when they enter the aquatic ecosystem. Microplastics fragments

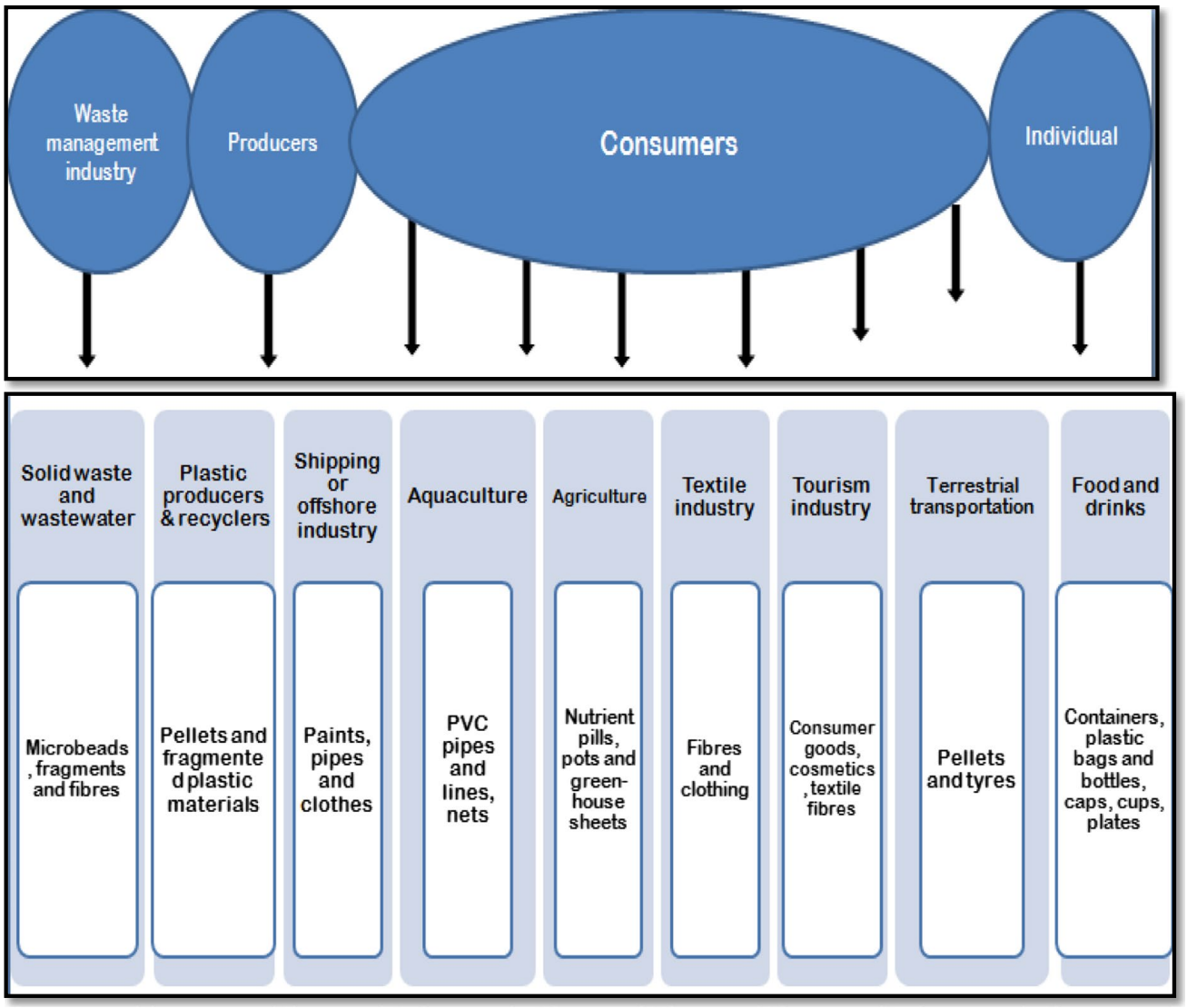


Fig. 1 Classification of microplastics in rivers and coastline

are categorized as primary or secondary microplastics on the basis of their origin. Primary microplastics include the plastic fragments synthesized by small size and used in various commercial applications. For example, microplastics are present in microbeads used for personal care products, abrasives in toothpaste, tiny beads used in exfoliation and pellets used for grinding and polishing (Fu & Wang, 2019).

Secondary microplastics are formed by fragmentation and degradation of bigger microplastic fragments under various atmospheric conditions (Yuan et al., 2020). Different processes including solar radiation, thermal oxidation and hydraulic power in rivers and oceans cause cracking and breakdown of large plastic particles (Liu et al., 2020). Therefore, small fragments with size less than 5 μm are formed commonly called as secondary microplastics. Different sources of secondary microplastics are industrial and plastic goods including plastic bottles, packaged bags, boxes, agricultural plastic films, marine paints, synthetic turfs, clothing, instruments and production wastes (Ammala et al., 2011).

3.2 Industry and Domestic Microplastics

Different industries generate primary or secondary microplastic in marine environment and human activities also add substantial load of microplastic contaminants in water (Kelly et al., 2019). Textile industry generates fibres including natural, regenerated and synthetic fibres that contain large amounts of microplastics released into water columns (Prata, 2018). Various researches have reported presence of microfibres in textile sewage containing microplastic particles approximately accounting for 35% of pollution in marine environment. The domestic washing also contributes to around 700,000 microfibres from wash load of 6 kg of acrylic fabric resulting in huge pollution (Napper & Thompson, 2016). Plastic microbeads occur in various products including shampoos, soaps and lotions used widely for domestic purposes thereby causing contamination when released from domestic sewage and wastewater treatment plants (Cheung & Fok, 2017). Different types of monomer microplastics are produced from automotive tyres. Most emission of microplastics is from car tyres with the road runoff (Kole et al., 2017).

Agricultural soils contain various types of microplastics because of utilization of agricultural plastic films and compost in soils. Farmers utilize various compost and plastic mulches for sustaining crops and ensuring food security (Ding et al., 2020). This indirectly paves the path for proliferation of microplastics that are released from soils to aquatic systems. Different construction sites and landfill operations along with domestic sewage from households also generate microplastics that are considered to be primary in origin (Galafassi et al., 2019).

3.3 Classification on Molecular Basis

Compared to the above two methods, the classification of microplastics at molecular level is achieved based on qualitative analysis of polymeric plastics. This classification helps to study the degradation potential of microplastics as well as recognize their chemical composition at time of degradation mechanisms (Yuan et al., 2020). The classification of microplastics on molecular basis and their properties are summarized in Table 1.

4 Behaviour of Microplastics in Marine Environment

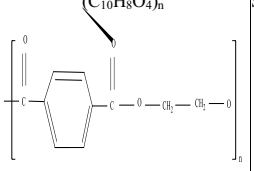
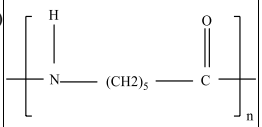
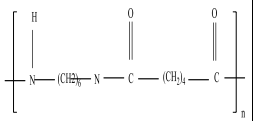
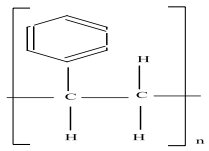
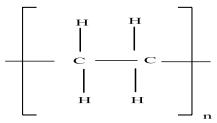
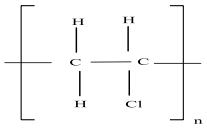
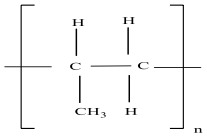
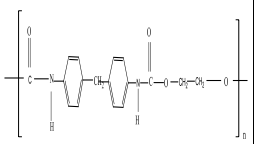
Microplastic bioaccumulation and biomagnifications can cause potential toxicity in humans and marine organisms. Thus, it is essential to determine behaviour of microplastics fragments and their mechanisms to understand harmful impacts in marine environment. The behaviour of microplastics in marine ecosystem can be classified as physical behaviour, chemical behaviour and biological behaviours (Wang et al., 2016). The various mechanisms underlying these behaviours are discussed further to analyse their impacts on aquatic system.

4.1 Physical Behaviour

4.1.1 Sedimentation

The residence of floating plastic debris in aquatic ecosystem causes an increase in the density of plastic debris (Barboza et al., 2019). The different microplastics interact with biotic and abiotic environment to reach high densities in sediment water (Rodrigues et al., 2018). Microplastics that reside on surface of water

Table 1 Microplastics polymers and their properties

Microplastic	Chemical Property	Physical Property			
	Chemical & Structural Formula	Crystallinity	Size (µm)	Glass Transition Temperature (Tg)	Melting Point
Polyethylene terephthalate (PET) (Zhao et al. 2018)	$(C_{10}H_8O_4)_n$ 	Semi-crystalline	100-150	80 ºC	260 ºC
Polyamide 6 (PA6) (Parodi et al. 2017)	$(C_6H_{11}NO)_n$ 	Crystalline	100-150	60 ºC	223 ºC
Polyamide 66 (PA66) (Pellini et al. 2018)	$(C_{12}H_{22}N_2O_2)_n$ 	Crystalline	100-150	58 ºC	260 ºC
Polystyrene (PS) (Sundbæk et al. 2018)	$(C_8H_8)_n$ 	Amorphous	100	100 ºC	240 ºC
Polyethylene (PE) (Peez et al. 2019)	$(C_2H_4)_n$ 	Semi-crystalline	100-150	-125 ºC	130 ºC
Polyvinyl Chloride (PVC) (Wu et al. 2019)	$(C_2H_3Cl)_n$ 	Amorphous	100-150	87 ºC	100-200 ºC
Polypropylene (PP) (Khoironi et al. 2020)	$(C_3H_6)_n$ 	Amorphous and semi-crystalline	100-150	0 ºC	160 ºC
Polyurethane (PU) (Zhang et al. 2018b)	$C_{17}H_{16}N_2O_4$ 	Amorphous	100-200	-20 ºC	90 C

can be drawn down into deep water column and deposited in sediments. The vertical and horizontal forces allow sediment particles to act as carriers of microplastics within rivers (Horton & Dixon, 2018). Sinking or floating behaviour of different sediment particles influence the occurrence of microplastics within the rivers and tend to have similar dynamics as clay sediment particles (Chubarenko et al., 2020). Also, abiotic factors weaken the molecular structure of plastic fragments and enhance their degradation into small microplastics that reach sediment zone due to change in sediment velocities and settling methods (van der Hal et al., 2017). Subsequently, adsorption and interaction of microplastics with sediment particles allows their sedimentation into large water systems.

4.1.2 Migration

Plastic pollution can quickly transfer from one site to another within the seawater by drifter buoys and physical oceanographic models (Law et al., 2010). Various types of microplastic monomers including polyethylene and polypropylene are buoyant and easily transferred within seawater (Ritchie & Roser, 2018). Underlying oceanic currents help in transfer of microplastics that are denser than seawater, for example, polyvinyl chloride (PVC) and polyethylene terephthalate (PET). These microplastics potentially sink in water due to high density and float on the surface only when they have entrapped air (Engler, 2012). Ocean currents are commonly referred as migration of water by different forces acting upon it including wind, temperature and salinity (Mahanty et al., 2016). Moreover, plastic particles could also get transferred by wind, and tsunami occurring in oceans (Lebreton et al., 2019). The strong wind can help to accelerate vortexing and redistribute plastic fragments at surface layers of water (Collignon et al., 2012). Additionally, tides and tsunamis could also offer transfer of plastic debris from one site to another within the seawater (Sadri & Thompson, 2014).

4.1.3 Accumulation

The migration of microplastics from one place to another by littering, landfill, dumping, wastewater treatment plants and accidental inputs causes their accumulation in coastal areas and aquatic system (Phelan et al., 2020). Since microplastics are non-degradable in nature, this allows them to accumulate within sediments and seabed and reside for prolonged period of time (Mason et al., 2016). The major cause of microplastics accumulation in ocean surface is due to their less density as compared with seawater (Gago et al., 2018). Seabed is also likely a sink for microplastics as large amounts of microfibres are also present in deep sea sediments (Sanchez-Vidal et al., 2018) that cause their accumulation. Fragmentation of plastic particles using different processes including physical abrasion, UV radiation and photo degradation could also result in accumulation of microplastics within the seawater (Zettler et al., 2013). The accumulation of plastic debris within the marine environment also results in formation of plastisphere (Zettler et al., 2013). This indirectly helps various microbial species to grow and proliferate on plastic surface and contribute to contamination (Lobelle & Cunliffe, 2011). Plastisphere acts as a habitat for microorganisms to survive and contains various microplastic fragments that are obtained from various sources. Accumulation of plastic debris can create new environment for organisms which indirectly increases their abundance and diffusivity (Diepens & Koelmans, 2018).

4.2 Chemical Behaviour

4.2.1 Degradation

A change in various properties of polymer including tensile strength, colour and shape of polymer under influence of several environmental factors could result in degradation (Bazli et al., 2020). The process of degradation could be categorized as thermal degradation, catalytic degradation, mechano-chemi-

cal degradation, photo-oxidative degradation, ozone-induced degradation and biodegradation (Singh & Sharma, 2008). Different types of microplastics such as polyethylene (PE), polypropylene (PP) and polystyrene (PS) would undergo degradation after exposure to solar radiation or physical abrasion thereby polluting marine environment (Andrady, 2011). Degradation plays an important role in recycling and reusing of microplastics fragments to reduce environmental pollution (Karbalaei et al., 2018). However, full conversion of microplastic particles in to carbon dioxide, water and inorganic compounds is extremely slow especially in marine environment where primary source of degradation is solar-UV radiation (Lithner et al., 2011). Various studies have reported slow degradation of microplastics by other processes as compared to solar radiation (Booth & Sørensen, 2020). Besides this, presence of additives in plastic particles can enhance the tendency for photo-oxidative degradation in marine environment (Zou et al., 2020). Also, oxygen concentrations and temperature changes would lead to degradation of various microplastic monomers (Andrady, 2011; Lithner et al., 2011).

Researchers are focused on determining the chemical reactions that result in degradation of polymers along with potential hazards from chemicals released by degradation. Degradation of plastic polymers usually follows abiotic or biotic pathways. Generally abiotic degradation involves the chemical breakdown process followed by biotic degradation that is initiated by biodegradation pathways (Fazey & Ryan, 2016). The various examples of polymers undergoing chemical degradation pathways are explained further in detail (Fig. 2).

Polyethylene (PE) degradation is initiated by exposing the polymer to ultraviolet radiation for breaking the main polymer chain to produce a free radical (Costa et al., 2018). After the initiation step, propagation phase occurs in which auto-oxidation process helps to form complex oxygenated low molecular weight compounds by reacting with free radicals (Gravouil et al., 2017). These low molecule weight fragments include aliphatic carboxylic acids, aldehydes, alcohols and ketones (Lapointe et al., 2020). In the auto-oxidation step, oxygen is incorporated in polymers thus its presence is necessary to initiate the whole process (Matthews, 2018). After the propagation phase, termination occurs that involves random cross-linking of molecules to generate degradation products including ethane, ethene, propane, propene, butane and hexane (Vasile, 2018). The polymer is degraded and becomes brittle with low molecular compounds formed as degradation products (Bäckström et al., 2017).

Polyvinyl chloride (PVC) is the most sensitive polymer towards UV radiation (Arnaud et al., 2017) and hence photo degradation is necessary (Feldman, 2016). When PVC is exposed to sunlight, it is dechlorinated with formation of conjugated double bonds in polymer and hydrochloric acid along with certain traces of other products (Prociak et al., 2018). The degradation of PVC is enhanced by photo-induced dechlorination in aerobic conditions, presence of HCl, mechanical stress, humidity, presence of other chemicals and high temperatures (Prociak et al., 2018; Seleem et al., 2017). The resulting unsaturated double bonds formed are less stable towards photo-degradation and are prone towards further degradation to smaller fragments (Hatakeyama-Sato et al., 2019).

Polyethylene terephthalate (PET) degradation in marine environment occurs through UV radiation, photo-oxidation as well as hydrolytic degradation (Fotopoulou & Karapanagioti, 2019). As discussed earlier in case of PE, UV radiation results in formation of free radicals that produce carboxylic acid end group in case of PVC (Morgan et al., 2017). The photo-degradation is followed by photo-oxidation process resulting in formation of hydrogen peroxide by various radical reactions (Jia et al., 2020). PET is also susceptible to hydrolytic degradation in marine environment. The low temperature degradation of PET under hydrolysis is the most important procedure and considered as reverse reaction of esterification of PET (Qin et al., 2020; Silva et al., 2018).

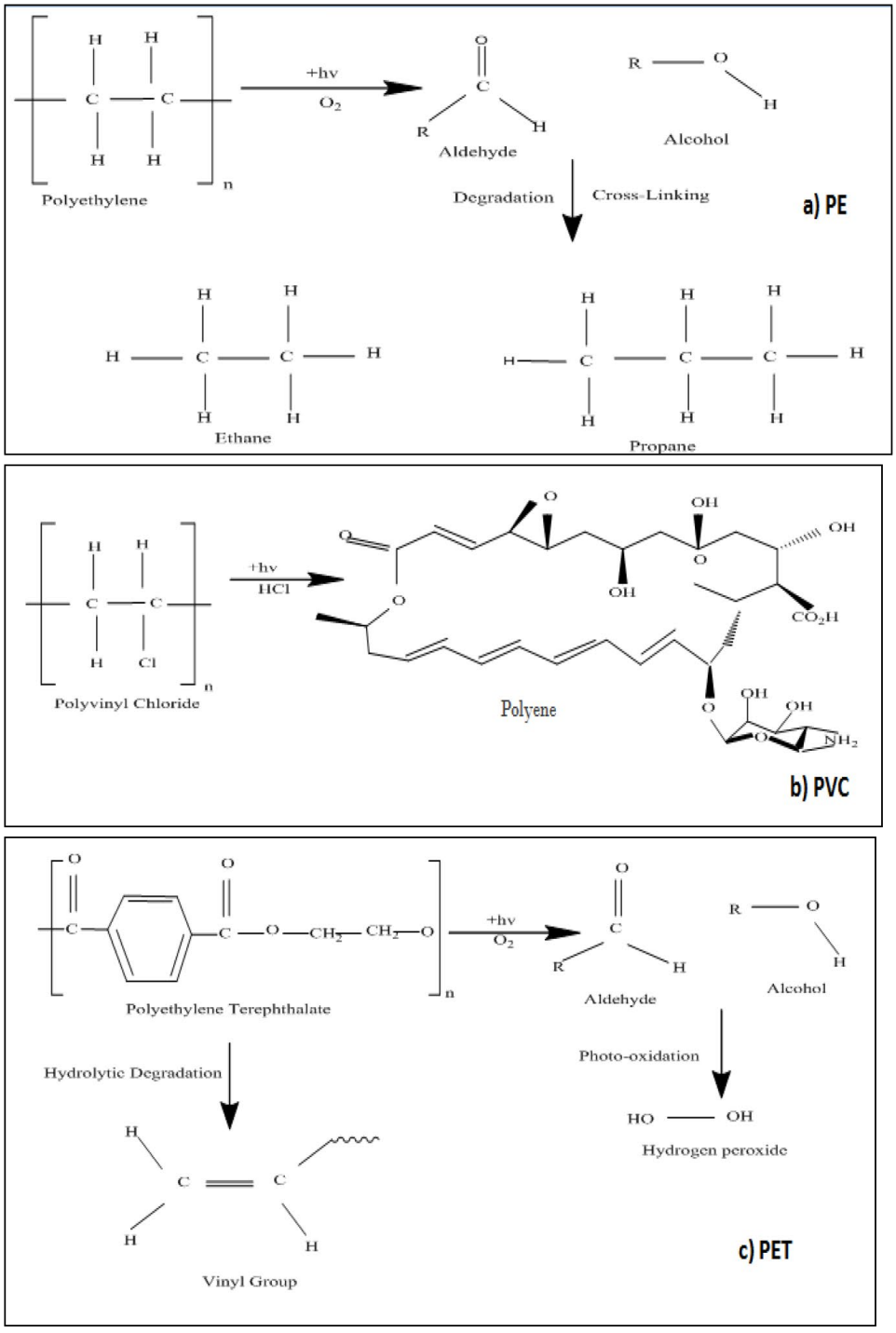


Fig. 2 Chemical reactions of different polymers

4.2.2 *Adsorption*

Adsorption was considered as a physical behaviour but it also shows significant chemical effects. The adsorption by physical means was relying on surface area as well as forces that increased adsorption of pollutants (Zhang et al., 2018a). On a contrary, chemical adsorption allows higher affinity for various pollutants on hydrophobic microplastic surface, and their sorption is influenced by crystallinity, diffusivity and surface area (Karapanagioti & Klontza, 2008). The environment is endowed with various organic polymers that constitute crystalline and amorphous regions (Sun et al., 2019). Crystalline region allows for lattice arrangement of molecules whereas molecules are randomly arranged

in amorphous region (Balzano et al., 2019). Thus, amorphous region exhibits a loose structure that allows greater sorption of organic pollutants on microplastic surfaces (Tien et al., 2009).

Also, properties of sorbates, weathering and residence time conditions, have an influence on adsorption in marine environment (Ogata et al., 2009). Moreover, polymer weathering could effectively increase surface area thereby enhancing the diffusivity (Losaria & Yim, 2020). Even photo-oxidation process could alter the plastic surface by developing cross-links, chain scissions, groups in polymers such as carbon moieties, resulting in cracks on the surface (Grause et al., 2020). Besides, weathering of plastic debris is considered to have long residence time in marine environments thereby resulting in greater sorption of organic pollutants (Lohmann, 2012). Different foulants attached to plastic debris also allow for increased sorption of contaminants including organic pollutants and metals (Hu et al., 2020). Persistent organic pollutants including polychlorinated biphenyls (PCBs), polycyclic aromatic hydrocarbons (PAHs) and dichlorodiphenyltrichloroethane (DDTs) are major pollutants adsorbed on microplastics (Hirai et al., 2011). Charged or neutral regions of plastic surfaces could also direct the adsorption of cations or complexes of metals (Holmes et al., 2012). Study by Turner & Holmes, 2015 reported greater adsorption of various metals including Ag, Cd, Pb, Zn, Ni and Co to plastics surfaces as pH of the water environment was increased. Studies have reported that microplastics make significant contributions to exposure of various chemical contaminants (Avio et al., 2017; Murray & Cowie, 2011). These substances are directly released from gut of marine species or may leach out in marine environment on weathering (Ribeiro et al., 2019).

4.3 Biological Behaviour

4.3.1 Ingestion

Plastic debris has polluted the whole marine environment with significant harmful effects on marine ecosystem. Uptake and behaviour of microplastics has been demonstrated in above sections that shows various microplastics might accumulate in marine species, from zooplankton to crabs, including whales, fishes, mussels, sea reptiles and sea birds (Gallo et al., 2018). Plastic debris is known to be present in gut content of fishes globally including estuaries (Dantas et al., 2012), and demersal habitats (Lusher et al., 2013). Also, wild seafood bivalves have been a source for ingestion of plastic debris (Van Cauwenberghe & Janssen, 2014). Microplastics also allow attachment of microalgae on their surfaces acting as food source for filter feeders (Kershaw & Rochman, 2015).

The most common microplastic debris prevailing in marine environment is polystyrene (PS) (Blettler et al., 2017). The ingestion of microplastic by fish majorly occurred through gills and intestine. Gills provide high surface area for accumulation of most microplastic debris and ingestion could also occur through digestive tract (Ribeiro et al., 2019). These microplastic contaminants could remain in tissue and organs of fish thereby inducing toxic effects (Booth & Sørensen, 2020). The foraging time indicating the activity of searching food also reduced after the accumulation of polymer (Roch et al., 2020). Polystyrene microplastics exposure also reduced the swimming speed of fishes suggesting a weak constitution (Foulon et al., 2016).

Different chemicals disrupt the marine ecosystem including endocrine disrupting chemicals (EDCs) that alter the synthesis of endocrine hormones (Munn & Goumenou, 2019). Various consequences of disrupted endocrine systems are low birth rates, thyroid functioning, metabolism and increased hormone-sensitive cancers (Gore et al., 2015). Chemicals including tetrabromobisphenol A present in *Danio rerio* cause increased oxidation in fish (Yu et al., 2020) whereas silver in *Lemna minor* and *Daphnia magna* causes eco-toxicity in digestive system of these species (Kalčíková et al., 2020). Some precursors employed in textile and paper industry including perfluorooctanesulfonic acid (PFOS) and perfluorooctanoic acid (PFOA) can also cause microplastic ingestion in marine environment (Schaider et al., 2017).

The chronic exposure of these harmful chemicals on marine species has been directly linked to reduced growth and survival of offspring thereby having negative effect on marine biodiversity (Desforges et al., 2015).

4.3.2 *Translocation*

After ingestion of various microplastics by marine organisms, they may be retained in digestive tract, exit in form of feces, absorbed into epithelial lining of gut or translocated to other tissues within an organism (Hale et al., 2020). Various studies have reported the presence of microplastics in gills and digestive tract of mussels using polarized light microscopy (von Moos et al., 2012). Also, microplastics have been found in the ovary, gills and hepato pancreas of crabs (Farrell & Nelson, 2013). These microplastics can be easily translocated from gut to circulatory system (Jovanović, 2017). Depending on the organs of digestive system and difference in dimensional characteristics of ingested microplastics, it is easier to translocate them to other organs (Brennecke et al., 2015). Microplastics may be broken down to smaller fragments and enter the lipid-rich organs of fishes (Anderson et al., 2016). Oxidative stress and transient changes in regulatory activities could also occur in aquatic organisms (Cassia et al., 2018). Microplastics could also be transported from lower trophic level organisms to their predators, for example, from *mytilus edulis* to *carcinus maenas* (Farrell & Nelson, 2013) and from zooplankton to mysid shrimp (Setälä et al., 2014).

4.3.3 *Biodegradation*

Different properties of microplastics including their hydrophobicity and lack of metabolic activity to polymerize the plastics make it difficult to undergo biodegradation (Chowdhary et al., 2020). However, biodegradation is possible by formation of microbial biofilms on surface of microplastic fragments (Rummel et al., 2017). These biofilms allow for growth of bacteria and other organisms that could potentially help in degradation of plastic (Lobelle & Cunliffe, 2011). Also, the weight of various plastic polymers can be reduced by incubating different microbial strains (Harshvardhan & Jha, 2013). Additionally, surface of microplastics containing pits could also be an indicator for bacterial species to degrade the polymers (Zettler et al., 2013). Enzymes play an essential role in degradation of plastic polymers after they are subjected to chemical degradation (Adrio & Demain, 2014). The plastic fragments formed after the chemical (abiotic) degradation are buried deep in marine environment and take years for degradation (Fotopoulou & Karapanagioti, 2019). The microbial biofilms attached to the surface of polymer allow the formation of various enzymes that induce breakdown of plastic by hydrolysis (Ho et al., 2018). Various microorganisms have potential to produce enzymes that result in degradation of polymers, for example, *Thermobifida fusca* produces an enzyme, hydrolase, capable of degrading PET (Barth et al., 2016; Jabloune et al., 2020).

Strains of *Bacillus cereus* and *Bacillus sphericus* produce peroxidase that helps to degrade PE (Yuan et al., 2020) (Figure 3). The primary procedure for biodegradation of plastic is initiated by sticking of microbes on polymer surface and their proliferation (Kawai et al., 2019). These microbes help in excretion of extracellular enzymes that result in breakdown of plastics (Alshehrei, 2017). The enzymatic hydrolysis occurs in two ways: first is oligomers, dimers and monomers release degradation products that are converted to carbon-dioxide and water when enzyme attaches to polymer and hydrolytic division occurs (Roohi et al., 2017). Secondly, polymers are degraded by microbes in absence of air and new enzymes are needed to degrade the plastic in anaerobic conditions (Pathak and Navneet, 2017). Thus, biodegradation of polymers results in production of microbial biomass, carbon-dioxide and water that can be used by aquatic flora and fauna.

The enzymatic degradation of PET using hydrolases, esterases, proteases and cutinases has shown to hydrolyze PET surfaces. Modification in enzymes can improve the specificity and effi-

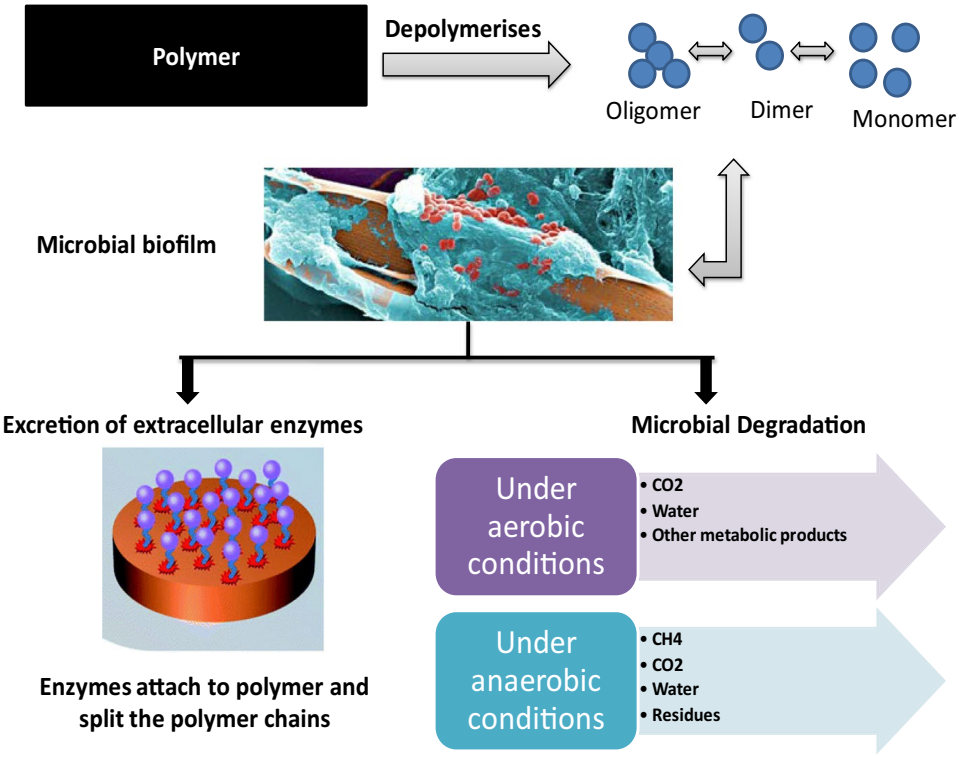


Fig. 3 Polymer degradation mechanism

ciency of PET degradation. For example, recent studies have shown that microbial species, *Ideonella sakaiensis*, has the capability to degrade PET by action of enzyme PETase. A hydrolytic reaction of PET under the enzyme, PETase, has efficiency to produce ethylene glycol and terephthalate that are required for microbial growth (Vandermaesen et al., 2016). Similarly, other enzymes are employed for degradation of microplastics PU, and 6-aminohexanoate oligomers (PA) (Wei & Zimmermann, 2017) (Fig. 4).

5 Challenges and Future Prospects

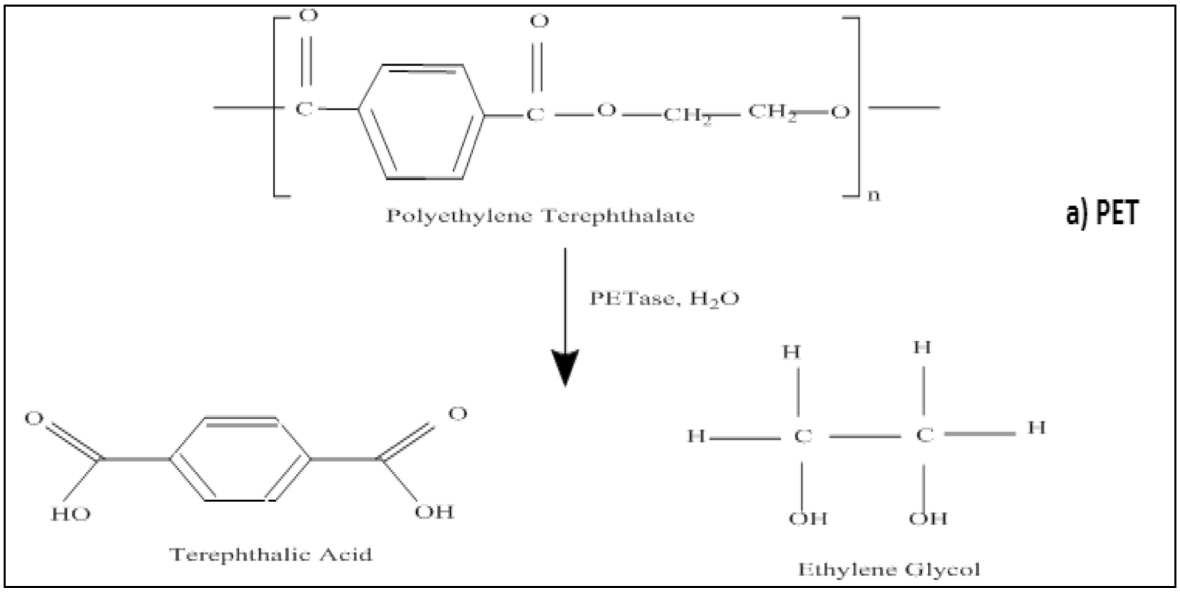
Extensive research studies have focused on the presence and effects of microplastics in marine environment. However, from the above discussions, knowledge gaps and major research questions in understanding the dynamics of microplastics in marine environment need utmost attention. Some broad issues with possible measures to safeguard aquatic environment are briefly summarized below.

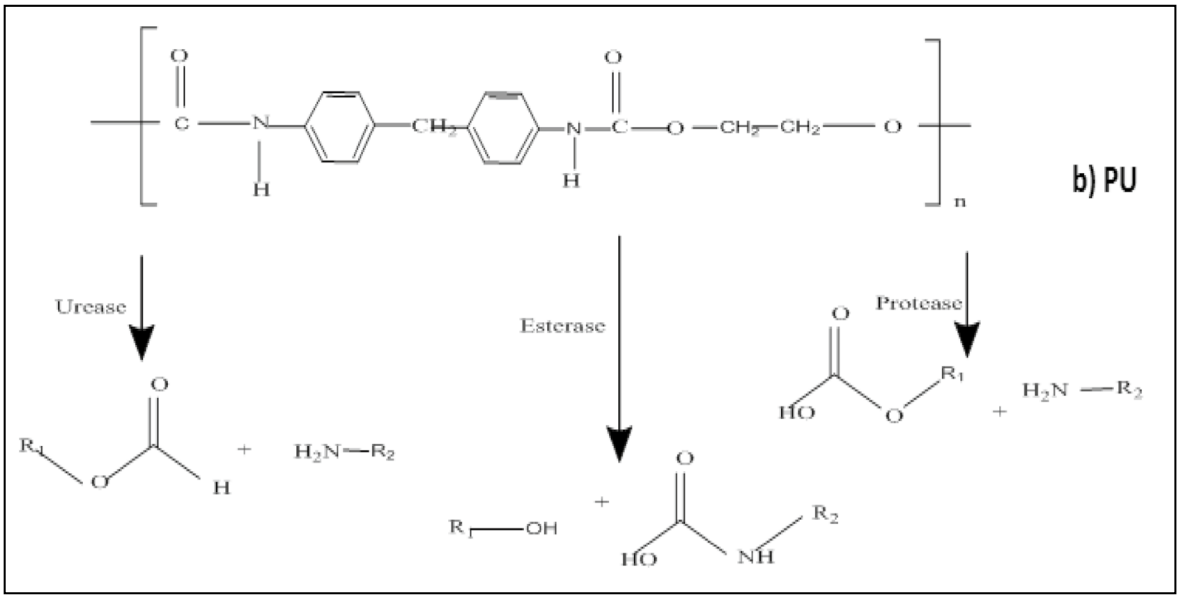
Strategies to control different sources of microplastic pollutants in marine environment

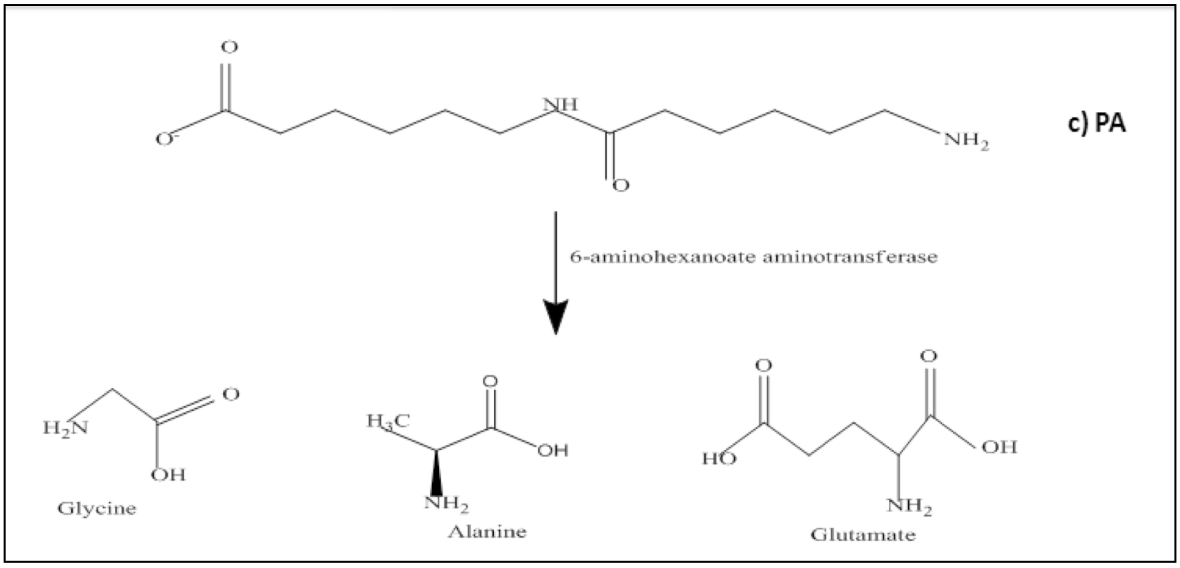
- Source control of microplastic pollution in marine environment is commonly achieved by awareness programs and proper legislation (Picó & Barceló, 2019). Decreasing microplastic release can be achieved most appropriately by banning utilization of microplastic beads in cosmetic products. Some countries like the USA, Australia and Canada are on the forefront of implementing measures to combat microplastic in cosmetics. Also, secondary microplastics obtained from single-use plastics should be restricted for sale and consumption (Steensgaard et al., 2017). The combinatorial approach of restricting plastic usage and levy may attempt to educate public of environmental hazards from microplastics (Picó & Barceló, 2019).

Quantify microplastic pollutants in marine ecosystem and assess water quality parameters at field scale post microplastic contamination.

- Remediation of aquatic environment to eliminate organic matter is performed since a long time but quantifying microplastics is essential since


Polyethylene Terephthalate
a) PET
PETase, H_2O
Terephthalic Acid
Ethylene Glycol


b) PU
Urease
Esterase
Protease


c) PA
6-aminohexanoate aminotransferase
Glycine
Alanine
Glutamate

◂ **Fig. 4** Enzymatic degradation of PET, PU and PA

they are mistaken for various anthropogenic particles (Picó & Barceló, 2019). Visual analysis of microplastics is not adequate that makes physical detection a more reliable approach to quantify microplastics (Sol et al., 2020). FT-IR, Raman spectroscopy, pyrolysis–gas chromatography/mass spectrometry and thermogravimetry coupled to differential scanning calorimetry can help to speculate new insights into microplastics (Gong & Xie, 2020).

- To conquer the problem of water quality analysis following microplastic contamination, new instrumental tools that cater to real-time analysis should be utilized for effective monitoring of pollutants (Varotsos et al., 2020). Application of optical device-making system (ODMS) allows for efficient monitoring of water quality based on detection of emissions from photometry and ellipsometry (Varotsos et al., 2019).

Possible approaches to reduce the physical–chemical behavioural mechanisms of microplastics in marine environment

- The characteristic behaviour of microplastics in marine environment is a major problem which raises questions on the interactions between microplastics and chemicals that cause implications in organisms. Therefore, understanding the surface properties and ecocorona of microplastics is necessitated to address their environmental impact (Galloway et al., 2017). The application of atomic force microscopy, pyrolysis gas-chromatography mass-spectrometry and optical nanosensors can facilitate the characterisation of microplastic particles and explore composition of ecocorona to provide greater insights into surface methodology of microplastics (Burrows et al., 2020). Other techniques like scanning electron microscopy and optical profilometry can help to determine the roughness of microplastic surfaces and enable the measurement of interactions between chemicals that have harmful impact on environment (Connors et al., 2017). Also, action towards chemical management and disposal of polymers from chemical industries could provide beneficial effect in reducing microplastic load in aquatic ecosystem (Hamidian et al., 2021). This can be achieved by adopting green synthetic methodologies like use of green solvents in membrane fabrication process (Xie et al., 2021) and bio-based reagents for manufacturing polymers (Sternberg et al., 2021). Also, bioplastic as alternative to synthetic polymers could provide positive environmental benefits and sustainability to the economy (Pellis et al., 2021).

Control the shortcomings arising from biological behaviour of microplastics in marine environment

- Microbial remediation is necessary to reduce the potential biological impact of microplastics in ecosystem (Tiwari et al., 2020). As discussed above, biodegradation of microplastics is facilitated by microorganisms like *Rhodococcus* sp., *Bacillus* sp., *Clostridium* sp., *Staphylococcus* sp. and *Cladosporium* sp. (Shen et al., 2019). Also, addition of biodegradable additives like starch and stimulants (Satti et al., 2018) to microplastic polymer makes the polymer hydrophilic allowing greater efficiency of enzyme to act on microplastic (Zadjelovic et al., 2020). The formation of biofilms on microplastic surface allows for growth and proliferation of various microorganisms which can facilitate degradation of polymer within aquatic system (Tu et al., 2020).
- Sea animals including annelids (sandworms) and echinoderms (sea cucumber) could be efficient source of microplastic bioremediation (Masiá et al., 2020). Sea grasses like *Halophila ovalis*, *Halophila beccarii* and *Zostera japonica* can help in trapping microplastics and mangrove rhizospheres including *Avicennia marina* and *Aegiceras corniculatum* could act as a sink for microplastic particles (Huang et al., 2021).
- Different algae like *Anabaena spiroides*, *Cocconis, Navicula*, *Chlorella fusca* and *Chlamydomonas reinhardtii* can colonize on microplastic surfaces and result in their biodegradation (Chia et al., 2020). Not only does algae help in biodegradation of microplastics, but it also acts as a potential source for bioplastic production (Rahman & Miller, 2017). Microalgae biomass, biorefinery processing, genetic engineering tools like CRISPR technology and intermixing with other blended materials could help in the production

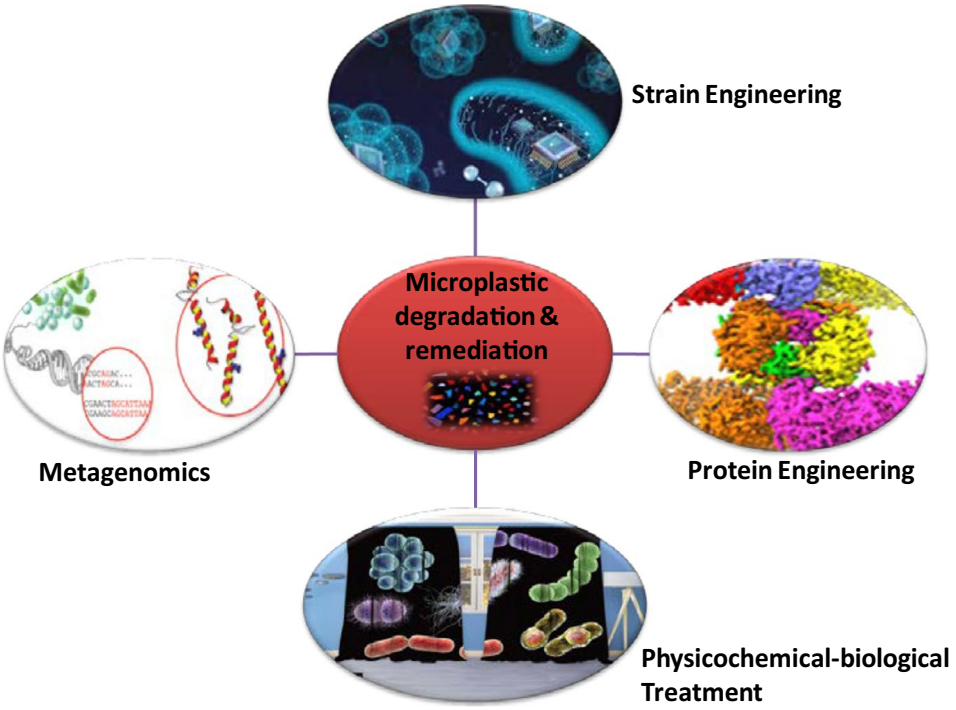


Fig. 5 Methods to remediate and degrade microplastics in marine environment

of algal-based bioplastic that promise non-toxic effect on environment (Karan et al., 2019).

Futuristic strategies to control microplastic pollution in marine environment

- To mitigate the effects of microplastic pollution, futuristic objectives aiming at different bioengineering-based solutions need to be considered (Coyle et al., 2020). Hence, the following tools can provide plausible solutions to microplastic contamination (Fig. 5):
- Strain engineering: This technique relies on using genetic engineering and recombinant DNA technology to manipulate microbial cells for increasing their efficiency of microplastic degradation (Ferreira et al., 2018; Jaiswal et al., 2020). For example, *Bacillus subtilis* strain could efficiently produce PETase enzyme capable of degrading PET microplastic (Huang et al., 2018). Also, microalgae, *Phaeodactylum tricornutum*, generated a microbial cell factory that produced PETase from *Ideonella sakaiensis* for biological decomposition of PET in seawater (Moog and Blank 2019).
- Protein engineering: This approach helps to enhance the activity and efficiency of enzymes for microplastic degradation (Wei et al., 2016). The protein engineered enzymes help to degrade the building blocks of microplastic. For example, the engineered hydrolase were able to degrade and hydrolyze PET films faster as compared to normal enzymes (Han et al., 2017; Ma et al., 2018). Another study by Islam et al., 2019 demonstrated that cutinases derived from *Thermomonospora curvata* could establish faster degradation of polyurethane.
- Metagenomics: This novel approach provides detailed understanding on classification of microorganisms and characterises enzymes capable of degrading microplastic polymers (Bhatt et al., 2021; Ufarté et al., 2015). The microbial biofilms formed on microplastic surfaces can be analysed using targeted or shotgun metagenomic technique to reveal distinct microbial profile (Tiwari et al., 2021) which can be found in the National Centre for Biotechnology Information (NCBI) portal (Bryant et al., 2016). Enzymes responsible for microplastic degradation can be found in biofilms through shotgun metagenomic sequencing process (Pinnell & Turner, 2019). Microplastic bioremediation could also be accelerated using in silico genome mining integrated with metagenomic datasets (Rai et al., 2021; Ziemert et al., 2016).

- Combined physicochemical-biological treatment: Recent studies have reported the use of chemical and biological methods simultaneously to improve degradation of microplastics and provide possible remediation measure in marine environment (Sánchez, 2020). For example, combination of chemical and biological methods was assessed by Tsiota et al., 2018 to degrade secondary microplastics in marine environment. In this study, polyethylene was exposed to ultraviolet radiation resulting in microplastics fragments which were incubated with pelagic microbiomes for increasing their degradation efficiency (Tsiota et al., 2018). Similarly, the microbial strain of *Penicillium variabile* could greatly enhance mineralization of polystyrene under ozone pre-treatment (Sharma et al., 2021).

6 Conclusion

Microplastic particles are emerging pollutants in aquatic ecosystem and also draw attention to other emerging contaminants. Due to the size and feasibility of entering the marine environment, microplastics bioaccumulation is high. The concentration of microplastics from different sources endangers the food security, biodiversity and marine ecosystem. Plastic polymers have different additives and stabilizers that adsorb various toxic pollutants and contaminants from surrounding environment. Their toxic effects are closely related with physical, chemical and biological properties that alter the functioning and characteristics of aquatic flora and fauna. The behavioural mechanisms not only demonstrate their fate but also the response of microplastics to certain environmental conditions. Based on this review, better understanding on types and behaviour of microplastics in marine environment along with their potential impacts could be identified. Development of biodegradable plastic products could help in combating against plastic pollution in marine ecosystem. The utilization of various microorganisms for bioremediation of plastic contaminants can provide sustainable solutions to plastic free marine environment. Further research on engineered enzymes and microbial strains could provide significant insights on degradation and remediation of microplastics in aquatic ecosystem.

Acknowledgements The authors are thankful to Delhi Technological University for providing infrastructure and facility.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication The authors approve for consent for the publication of the article.

Conflict of Interest The authors declare no competing interests.

References

Adrio, J. L., & Demain, A. L. (2014). Microbial enzymes: Tools for biotechnological processes. *Biomolecules, 4*, 117–139. https://doi.org/10.3390/biom4010117

Alshehrei, F. (2017). Biodegradation of synthetic and natural plastic by microorganisms. Environ Microbiol 12

Ammala, A., Bateman, S., Dean, K., et al. (2011). An overview of degradable and biodegradable polyolefins. *Progress in Polymer Science, 36*, 1015–1049. https://doi.org/10.1016/j.progpolymsci.2010.12.002

Anderson, J. C., Park, B. J., & Palace, V. P. (2016). Microplastics in aquatic environments: Implications for Canadian ecosystems. *Environmental Pollution, 218*, 269–280. https://doi.org/10.1016/j.envpol.2016.06.074

Andrady, A. L. (2011). Microplastics in the marine environment. *Marine Pollution Bulletin, 62*, 1596–1605. https://doi.org/10.1016/j.marpolbul.2011.05.030

Arnaud, S. P., Wu, L., Chang, M.-A.W., et al. (2017). New biobased monomers: Tuneable polyester properties using branched diols from biomass. *Faraday Discussions, 202*, 61–77. https://doi.org/10.1039/C7FD00057J

Auta, H. S., Emenike, C. U., & Fauziah, S. H. (2017). Distribution and importance of microplastics in the marine environment: A review of the sources, fate, effects, and potential solutions. *Environment International, 102*, 165–176. https://doi.org/10.1016/j.envint.2017.02.013

Avio, C. G., Gorbi, S., & Regoli, F. (2017). Plastics and microplastics in the oceans: From emerging pollutants to emerged threat. *Marine Environment Research, 128*, 2–11. https://doi.org/10.1016/j.marenvres.2016.05.012

Bäckström, E., Odelius, K., & Hakkarainen, M. (2017). Trash to treasure: Microwave-assisted conversion of polyethylene to functional chemicals. *Industrial and Engineering Chemistry Research, 56*, 14814–14821. https://doi.org/10.1021/acs.iecr.7b04091

Balzano, L., Coussens, B., Engels, T., et al. (2019). Multiscale structure and microscopic deformation mechanisms of gel-spun ultrahigh-molecular-weight polyethylene fibers. *Macromolecules, 52*, 5207–5216. https://doi.org/10.1021/acs.macromol.9b00247

Barboza, L. G. A., Cózar, A., Gimenez, B. C. G., et al. (2019). Chapter 17 — macroplastics pollution in the marine environment. In: Sheppard C (ed) *World Seas: an Environmental Evaluation* (Second Edition). Academic Press, pp 305–328

Barth, M., Honak, A., Oeser, T., et al. (2016). A dual enzyme system composed of a polyester hydrolase and a carboxylesterase enhances the biocatalytic degradation of polyethylene terephthalate films. *Biotechnology Journal, 11*, 1082–1087. https://doi.org/10.1002/biot.201600008

Bazli, M., Zhao, X.-L., Jafari, A., et al. (2020). Mechanical properties of pultruded GFRP profiles under seawater sea sand concrete environment coupled with UV radiation and moisture. *Construction and Building Materials, 258*, 120369. https://doi.org/10.1016/j.conbuildmat.2020.120369

Bellasi, A., Binda, G., Pozzi, A., et al. (2020). Microplastic contamination in freshwater environments: A review, focusing on interactions with sediments and benthic organisms. *Environments, 7*, 30. https://doi.org/10.3390/environments7040030

Bhatt, P., Pathak, V. M., Bagheri, A. R., & Bilal, M. (2021). Microplastic contaminants in the aqueous environment, fate, toxicity consequences, and remediation strategies. *Environmental Research, 200*, 111762. https://doi.org/10.1016/j.envres.2021.111762

Blettler, M. C. M., Ulla, M. A., Rabuffetti, A. P., & Garello, N. (2017). Plastic pollution in freshwater ecosystems: Macro-, meso-, and microplastic debris in a floodplain lake. *Environmental Monitoring and Assessment, 189*, 581. https://doi.org/10.1007/s10661-017-6305-8

Booth, A. M., & Sørensen, L. (2020). Microplastic fate and impacts in the environment. In T. Rocha-Santos, M. Costa, & C. Mouneyrac (Eds.), *Handbook of Microplastics in the Environment* (pp. 1–24). Springer International Publishing.

Botterell, Z. L. R., Beaumont, N., Dorrington, T., et al. (2019). Bioavailability and effects of microplastics on marine zooplankton: A review. *Environmental Pollution, 245*, 98–110. https://doi.org/10.1016/j.envpol.2018.10.065

Brennecke, D., Ferreira, E. C., Costa, T. M. M., et al. (2015). Ingested microplastics (>100μm) are translocated to organs of the tropical fiddler crab Uca rapax. *Mar Pollut Bull* 96:491–495. https://doi.org/10.1016/j.marpolbul.2015.05.001

Bryant, J. A., Clemente, T. M., Viviani, D. A., et al. (2016). Diversity and activity of communities inhabiting plastic debris in the North Pacific Gyre. *mSystems* 1:. https://doi.org/10.1128/mSystems.00024-16

Burrows, S. D., Frustaci, S., Thomas, K. V., & Galloway, T. (2020). Expanding exploration of dynamic microplastic surface characteristics and interactions. *TrAC, Trends in Analytical Chemistry, 130*, 115993. https://doi.org/10.1016/j.trac.2020.115993

Cassia, R., Nocioni, M., Correa-Aragunde, N., & Lamattina, L. (2018) Climate change and the impact of greenhouse gasses: CO2 and NO, friends and foes of plant oxidative stress. *Front Plant Sci* 9.https://doi.org/10.3389/fpls.2018.00273

Cheung, P. K., & Fok, L. (2017). Characterisation of plastic microbeads in facial scrubs and their estimated emissions in Mainland China. *Water Research, 122*, 53–61. https://doi.org/10.1016/j.watres.2017.05.053

Chia, W. Y., Ying Tang, D. Y., Khoo, K. S., et al. (2020). Nature's fight against plastic pollution: Algae for plastic biodegradation and bioplastics production. *Environ Sci Ecotechnology, 4*, 100065. https://doi.org/10.1016/j.ese.2020.100065

Chowdhary, P., Raj, A., Verma, D., & Akhter, Y. (2020). *Microorganisms for sustainable environment and health*. Elsevier.

Chubarenko, I., Efimova, I., Bagaeva, M., et al. (2020). On mechanical fragmentation of single-use plastics in the sea swash zone with different types of bottom sediments: Insights from laboratory experiments. *Marine Pollution Bulletin, 150*, 110726. https://doi.org/10.1016/j.marpolbul.2019.110726

Collignon, A., Hecq, J.-H., Glagani, F., et al. (2012). Neustonic microplastic and zooplankton in the North Western Mediterranean Sea. *Marine Pollution Bulletin, 64*, 861–864. https://doi.org/10.1016/j.marpolbul.2012.01.011

Connors, K. A., Dyer, S. D., & Belanger, S. E. (2017). Advancing the quality of environmental microplastic research. *Environmental Toxicology and Chemistry, 36*, 1697–1703. https://doi.org/10.1002/etc.3829

Costa, J. P. D., Nunes, A. R., Santos, P. S. M., et al. (2018). Degradation of polyethylene microplastics in seawater: Insights into the environmental degradation of polymers. *Journal of Environmental Science and Health, Part A Environmental Science, 53*, 866–875. https://doi.org/10.1080/10934529.2018.1455381

Coyle, R., Hardiman, G., & Driscoll, K. O. (2020). Microplastics in the marine environment: a review of their sources, distribution processes and uptake into ecosystems. *Case Stud Chem Environ Eng* 100010.https://doi.org/10.1016/j.cscee.2020.100010

Dantas, D. V., Barletta, M., & da Costa, M. F. (2012). The seasonal and spatial patterns of ingestion of polyfilament nylon fragments by estuarine drums (Sciaenidae). *Environmental Science and Pollution Research, 19*, 600–606. https://doi.org/10.1007/s11356-011-0579-0

de Silva, J., AA e, Rodrigues JKG, Carvalho MW de, , et al. (2018). Mechanical performance of asphalt mixtures using polymer-micronized PET-modified binder. *Road Mater Pavement Des, 19*, 1001–1009. https://doi.org/10.1080/14680629.2017.1283353

Desforges, J.-P.W., Galbraith, M., & Ross, P. S. (2015). Ingestion of microplastics by zooplankton in the Northeast Pacific Ocean. *Archives of Environmental Contamination and Toxicology, 69*, 320–330. https://doi.org/10.1007/s00244-015-0172-5

Diepens, N. J., & Koelmans, A. A. (2018). Accumulation of plastic debris and associated contaminants in aquatic food webs. *Environmental Science and Technology, 52*, 8510–8520. https://doi.org/10.1021/acs.est.8b02515

Ding, L., Zhang, S., Wang, X., et al. (2020). The occurrence and distribution characteristics of microplastics in the agricultural soils of Shaanxi Province, in north-western China. *Science of the Total Environment, 720*, 137525. https://doi.org/10.1016/j.scitotenv.2020.137525

Dolatabadi M, Ahmadzadeh S (2020) Microplastics pollution in the aquatic environment: problems and challenges. *J Environ Health Sustain Dev*. https://doi.org/10.18502/jehsd.v5i2.3383

Engler, R. E. (2012). The complex interaction between marine debris and toxic chemicals in the ocean. *Environmental Science and Technology, 46*, 12302–12315. https://doi.org/10.1021/es3027105

Farrell, P., & Nelson, K. (2013). Trophic level transfer of microplastic: Mytilus edulis (L.) to Carcinus maenas (L.). *Environmental Pollution, 177*, 1–3. https://doi.org/10.1016/j.envpol.2013.01.046

Fazey, F. M. C., & Ryan, P. G. (2016). Biofouling on buoyant marine plastics: An experimental study into the effect of size on surface longevity. *Environmental Pollution, 210*, 354–360. https://doi.org/10.1016/j.envpol.2016.01.026

Feldman, D. (2016). Polyolefin, olefin copolymers and polyolefin polyblend nanocomposites. *J Macromol Sci Part A, 53*, 651–658. https://doi.org/10.1080/10601325.2016.1212313

Ferreira. R. da G., Azzoni, A. R., Freitas, S. (2018). Techno-economic analysis of the industrial production of a low-cost enzyme using E. coli: the case of recombinant β-glucosidase. *Biotechnol Biofuels 11*:81. https://doi.org/10.1186/s13068-018-1077-0

Fotopoulou, K. N., & Karapanagioti, H. K. (2019). Degradation of various plastics in the environment. In H. Takada & H. K. Karapanagioti (Eds.), *Hazardous Chemicals Associated with Plastics in the Marine Environment* (pp. 71–92). Springer International Publishing.

Foulon, V., Le Roux, F., Lambert, C., et al. (2016). Colonization of polystyrene microparticles by vibrio crassostreae: Light and electron microscopic investigation. *Environmental Science and Technology, 50*, 10988–10996. https://doi.org/10.1021/acs.est.6b02720

Fu, Z., & Wang, J. (2019). Current practices and future perspectives of microplastic pollution in freshwater ecosystems in China. *Science of the Total Environment, 691*, 697–712. https://doi.org/10.1016/j.scitotenv.2019.07.167

Gago, J., Carretero, O., Filgueiras, A. V., & Viñas, L. (2018). Synthetic microfibers in the marine environment: A review on their occurrence in seawater and sediments. *Marine Pollution Bulletin, 127*, 365–376. https://doi.org/10.1016/j.marpolbul.2017.11.070

Galafassi, S., Nizzetto, L., & Volta, P. (2019). Plastic sources: A survey across scientific and grey literature for their inventory and relative contribution to microplastics pollution in natural environments, with an emphasis on surface water. *Science of the Total Environment, 693*, 133499. https://doi.org/10.1016/j.scitotenv.2019.07.305

Galgani, F., Hanke, G., & Maes, T. (2015). Global distribution, composition and abundance of marine litter. In M. Bergmann, L. Gutow, & M. Klages (Eds.), *Marine Anthropogenic Litter* (pp. 29–56). Springer International Publishing.

Gallo, F., Fossi, C., Weber, R., et al. (2018). Marine litter plastics and microplastics and their toxic chemicals components: The need for urgent preventive measures. *Environmental Sciences Europe, 30*, 13. https://doi.org/10.1186/s12302-018-0139-z

Galloway, T. S., Cole, M., & Lewis, C. (2017). Interactions of microplastic debris throughout the marine ecosystem. *Nat Ecol Evol, 1*, 1–8. https://doi.org/10.1038/s41559-017-0116

Geyer, R., Jambeck, J. R., & Law, K. L. (2017). Production, use, and fate of all plastics ever made. *Science Advances, 3*, e1700782. https://doi.org/10.1126/sciadv.1700782

Gong, J., & Xie, P. (2020). Research progress in sources, analytical methods, eco-environmental effects, and control measures of microplastics. *Chemosphere, 254*, 126790. https://doi.org/10.1016/j.chemosphere.2020.126790

Gore, A. C., Chappell, V. A., Fenton, S. E., et al. (2015). Executive Summary to EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. *Endocrine Reviews, 36*, 593–602. https://doi.org/10.1210/er.2015-1093

Grause, G., Chien, M.-F., & Inoue, C. (2020). Changes during the weathering of polyolefins. *Polymer Degradation and Stability, 181*, 109364. https://doi.org/10.1016/j.polymdegradstab.2020.109364

Gravouil, K., Ferru-Clément, R., Colas, S., et al. (2017). Transcriptomics and lipidomics of the environmental strain Rhodococcus ruber point out consumption pathways and potential metabolic bottlenecks for polyethylene degradation. *Environmental Science and Technology, 51*, 5172–5181. https://doi.org/10.1021/acs.est.7b00846

Guzzetti, E., Sureda, A., Tejada, S., & Faggio, C. (2018). Microplastic in marine organism: Environmental and toxicological effects. *Environmental Toxicology and Pharmacology, 64*, 164–171. https://doi.org/10.1016/j.etap.2018.10.009

Hale, R. C., Seeley, M. E., Guardia, M. J. L., et al. (2020). A global perspective on microplastics. *J Geophys Res Oceans* 125:e2018JC014719. https://doi.org/10.1029/2018JC014719

Hamidian, A. H., Ozumchelouei, E. J., Feizi, F., et al. (2021). A review on the characteristics of microplastics in wastewater treatment plants: A source for toxic chemicals. *Journal of Cleaner Production, 295*, 126480. https://doi.org/10.1016/j.jclepro.2021.126480

Han, X., Liu, W., Huang, J.-W., et al. (2017). Structural insight into catalytic mechanism of PET hydrolase. *Nature Communications, 8*, 2106. https://doi.org/10.1038/s41467-017-02255-z

Harshvardhan, K., & Jha, B. (2013). Biodegradation of low-density polyethylene by marine bacteria from pelagic waters, Arabian Sea, India. *Marine Pollution Bulletin, 77*, 100–106. https://doi.org/10.1016/j.marpolbul.2013.10.025

Hatakeyama-Sato, K., Masui, T., Serikawa, T., et al. (2019). Nonconjugated redox-active polymer mediators for rapid electrocatalytic charging of lithium metal oxides. *ACS Appl Energy Mater, 2*, 6375–6382. https://doi.org/10.1021/acsaem.9b01007

Hirai, H., Takada, H., Ogata, Y., et al. (2011). Organic micropollutants in marine plastics debris from the open ocean

and remote and urban beaches. *Marine Pollution Bulletin, 62*, 1683–1692. https://doi.org/10.1016/j.marpolbul.2011.06.004

Ho, B. T., Roberts, T. K., & Lucas, S. (2018). An overview on biodegradation of polystyrene and modified polystyrene: The microbial approach. *Critical Reviews in Biotechnology, 38*, 308–320. https://doi.org/10.1080/07388551.2017.1355293

Holmes, L. A., Turner, A., & Thompson, R. C. (2012). Adsorption of trace metals to plastic resin pellets in the marine environment. *Environmental Pollution, 160*, 42–48. https://doi.org/10.1016/j.envpol.2011.08.052

Hong, S. H., Shim, W. J., Jang, M. (2018). Chapter 9 - Chemicals associated with marine plastic debris and microplastics: analyses and contaminant levels. In: Zeng EY (ed) Microplastic Contamination in Aquatic Environments. Elsevier, pp 271–315

Horton, A. A., & Dixon, S. J. (2018). Microplastics: an introduction to environmental transport processes. *WIREs Water 5*.https://doi.org/10.1002/wat2.1268

Hu, B., Ai, Y., Jin, J., et al. (2020). Efficient elimination of organic and inorganic pollutants by biochar and biochar-based materials. *Biochar, 2*, 47–64. https://doi.org/10.1007/s42773-020-00044-4

Huang, X., Cao, L., Qin, Z., et al. (2018). Tat-independent secretion of polyethylene terephthalate hydrolase PETase in Bacillus subtilis 168 mediated by its native signal peptide. *Journal of Agriculture and Food Chemistry, 66*, 13217–13227. https://doi.org/10.1021/acs.jafc.8b05038

Huang, Y., Xiao, X., Effiong, K., et al. (2021). New insights into the microplastic enrichment in the blue carbon ecosystem: Evidence from seagrass meadows and mangrove forests in Coastal South China Sea. *Environmental Science and Technology, 55*, 4804–4812. https://doi.org/10.1021/acs.est.0c07289

Islam, S., Apitius, L., Jakob, F., & Schwaneberg, U. (2019). Targeting microplastic particles in the void of diluted suspensions. *Environment International, 123*, 428–435. https://doi.org/10.1016/j.envint.2018.12.029

Jabloune, R., Khalil, M., Moussa, I. E.B., et al. (2020). Enzymatic degradation of *p*-nitrophenyl esters, polyethylene terephthalate, cutin, and suberin by sub1, a suberinase encoded by the plant pathogen *Streptomyces scabies*. *Microbes Environ 35*.https://doi.org/10.1264/jsme2.ME19086

Jaiswal, S., Sharma, B., & Shukla, P. (2020). Integrated approaches in microbial degradation of plastics. *Environmental Technology and Innovation, 17*, 100567. https://doi.org/10.1016/j.eti.2019.100567

Jambeck, J. R., Geyer, R., Wilcox, C., et al. (2015). Plastic waste inputs from land into the ocean. *Science, 347*, 768–771. https://doi.org/10.1126/science.1260352

Jia, H., Ben, H., Luo, Y., & Wang, R. (2020). Catalytic fast pyrolysis of poly (ethylene terephthalate) (PET) with zeolite and nickel chloride. *Polymers, 12*, 705. https://doi.org/10.3390/polym12030705

Jovanović, B. (2017). Ingestion of microplastics by fish and its potential consequences from a physical perspective. *Integrated Environmental Assessment and Management, 13*, 510–515. https://doi.org/10.1002/ieam.1913

Kalčíková, G., Skalar, T., Marolt, G., & Jemec Kokalj, A. (2020). An environmental concentration of aged microplastics with adsorbed silver significantly affects aquatic organisms. *Water Research, 175*, 115644. https://doi.org/10.1016/j.watres.2020.115644

Karan, H., Funk, C., Grabert, M., et al. (2019). Green bioplastics as part of a circular bioeconomy. *Trends in Plant Science, 24*, 237–249. https://doi.org/10.1016/j.tplants.2018.11.010

Karapanagioti, H. K., & Klontza, I. (2008). Testing phenanthrene distribution properties of virgin plastic pellets and plastic eroded pellets found on Lesvos island beaches (Greece). *Marine Environment Research, 65*, 283–290. https://doi.org/10.1016/j.marenvres.2007.11.005

Karbalaei, S., Hanachi, P., Walker, T. R., & Cole, M. (2018). Occurrence, sources, human health impacts and mitigation of microplastic pollution. *Environmental Science and Pollution Research, 25*, 36046–36063. https://doi.org/10.1007/s11356-018-3508-7

Kataoka, T., Nihei, Y., Kudou, K., & Hinata, H. (2019). Assessment of the sources and inflow processes of microplastics in the river environments of Japan. *Environmental Pollution, 244*, 958–965. https://doi.org/10.1016/j.envpol.2018.10.111

Kawai, F., Kawabata, T., & Oda, M. (2019). Current knowledge on enzymatic PET degradation and its possible application to waste stream management and other fields. *Applied Microbiology and Biotechnology, 103*, 4253–4268. https://doi.org/10.1007/s00253-019-09717-y

Kelly, M. R., Lant, N. J., Kurr, M., & Burgess, J. G. (2019). Importance of water-volume on the release of microplastic fibers from laundry. *Environmental Science and Technology, 53*, 11735–11744. https://doi.org/10.1021/acs.est.9b03022

Kershaw, P. J., Rochman, C. M. (2015). Sources, fate and effects of microplastics in the marine environment: part 2 of a global assessment. *Rep Stud - IMOFAOUnesco-IOCWMOIAEAUNUNEP Jt Group Experts Sci Asp Mar Environ Prot GESAMP Eng No 93*

Khoironi, A., Hadiyanto, H., Anggoro, S., & Sudarno, S. (2020). Evaluation of polypropylene plastic degradation and microplastic identification in sediments at Tambak Lorok coastal area, Semarang. *Indonesia. Mar Pollut Bull, 151*, 110868. https://doi.org/10.1016/j.marpolbul.2019.110868

Kole, P. J., Löhr, A. J., Van Belleghem, F. G. A. J., & Ragas, A. M. J. (2017). Wear and tear of tyres: A stealthy source of microplastics in the environment. *International Journal of Environmental Research and Public Health, 14*, 1265. https://doi.org/10.3390/ijerph14101265

Lapointe, M., Farner, J. M., Hernandez, L. M., & Tufenkji, N. (2020). Understanding and improving microplastic removal during water treatment: Impact of coagulation and flocculation. *Environmental Science and Technology, 54*, 8719–8727. https://doi.org/10.1021/acs.est.0c00712

Law, K. L., Morét-Ferguson, S., Maximenko, N. A., et al. (2010). Plastic accumulation in the North Atlantic Subtropical Gyre. *Science, 329*, 1185–1188. https://doi.org/10.1126/science.1192321

Lebreton, L., Egger, M., & Slat, B. (2019). A global mass budget for positively buoyant macroplastic debris in the

ocean. *Science and Reports, 9*, 12922. https://doi.org/10.1038/s41598-019-49413-5

Lithner, D., Larsson, Å., & Dave, G. (2011). Environmental and health hazard ranking and assessment of plastic polymers based on chemical composition. *Science of the Total Environment, 409*, 3309–3324. https://doi.org/10.1016/j.scitotenv.2011.04.038

Liu, Y., Huang, J., Jin, J., et al. (2020). The classification of micro-plastics and biodegradation of plastics/micro-plastics. 10

Lobelle, D., & Cunliffe, M. (2011). Early microbial biofilm formation on marine plastic debris. *Marine Pollution Bulletin, 62*, 197–200. https://doi.org/10.1016/j.marpolbul.2010.10.013

Lohmann, R. (2012). Critical review of low-density polyethylene's partitioning and diffusion coefficients for trace organic contaminants and implications for its use as a passive sampler. *Environmental Science and Technology, 46*, 606–618. https://doi.org/10.1021/es202702y

Losaria, P. M., & Yim, J.-H. (2020). Enhancement of strain-sensing performance through gas phase incorporation of siloxane into thermoplastic polyurethane-conducting polymer composite. *Macromolecular Chemistry and Physics, 221*, 2000155. https://doi.org/10.1002/macp.202000155

Lusher, A. L., McHugh, M., & Thompson, R. C. (2013). Occurrence of microplastics in the gastrointestinal tract of pelagic and demersal fish from the English Channel. *Marine Pollution Bulletin, 67*, 94–99. https://doi.org/10.1016/j.marpolbul.2012.11.028

Ma, Y., Yao, M., Li, B., et al. (2018). Enhanced poly(ethylene terephthalate) hydrolase activity by protein engineering. *Engineering, 4*, 888–893. https://doi.org/10.1016/j.eng.2018.09.007

Mahanty, M. M., Mohanty, P. K., Pattnaik, A. K., et al. (2016). Hydrodynamics, temperature/salinity variability and residence time in the Chilika lagoon during dry and wet period: Measurement and modeling. *Continental Shelf Research, 125*, 28–43. https://doi.org/10.1016/j.csr.2016.06.017

Masiá, P., Sol, D., Ardura, A., et al. (2020). Bioremediation as a promising strategy for microplastics removal in wastewater treatment plants. *Marine Pollution Bulletin, 156*, 111252. https://doi.org/10.1016/j.marpolbul.2020.111252

Mason, S. A., Garneau, D., Sutton, R., et al. (2016). Microplastic pollution is widely detected in US municipal wastewater treatment plant effluent. *Environmental Pollution, 218*, 1045–1054. https://doi.org/10.1016/j.envpol.2016.08.056

Matthews, M. E. (2018). *Crystallisation behaviour of commercial polyethylenes: A fundamental study*. Stellenbosch University.

Michielssen, M. R., Michielssen, E. R., Ni, J., & Duhaime, M. B. (2016). Fate of microplastics and other small anthropogenic litter (SAL) in wastewater treatment plants depends on unit processes employed. *Environmental Science: Water Research & Technology, 2*, 1064–1073. https://doi.org/10.1039/C6EW00207B

Moog, U., Blank, R., et al. (2019). Entwicklungsstörungen und Behinderungen. In G. F. Hoffmann, M. J. Lentze, & J. Spranger (Eds.), *Pädiatrie: Grundlagen und Praxis* (pp. 1–6). Springer.

Morgan, A., Cocca, M., Vega, K., et al. (2017). Vacuum UV photo-oxidation of poly(ethylene terephthalate). *Journal of Adhesion Science and Technology, 31*, 2542–2554. https://doi.org/10.1080/01694243.2017.1308994

Munn, S., & Goumenou, M. (2019). Thresholds for endocrine disrupters and related uncertainties. *Report of the Endocrine Disrupters Expert Advisory Group*

Murray, F., & Cowie, P. R. (2011). Plastic contamination in the decapod crustacean Nephrops norvegicus (Linnaeus, 1758). *Marine Pollution Bulletin, 62*, 1207–1217. https://doi.org/10.1016/j.marpolbul.2011.03.032

Napper, I. E., & Thompson, R. C. (2016). Release of synthetic microplastic plastic fibres from domestic washing machines: Effects of fabric type and washing conditions. *Marine Pollution Bulletin, 112*, 39–45. https://doi.org/10.1016/j.marpolbul.2016.09.025

Ogata, Y., Takada, H., Mizukawa, K., et al. (2009). International Pellet Watch: Global monitoring of persistent organic pollutants (POPs) in coastal waters. 1. Initial phase data on PCBs, DDTs, and HCHs. *Marine Pollution Bulletin, 58*, 1437–1446. https://doi.org/10.1016/j.marpolbul.2009.06.014

Ourmieres, Y., Mansui, J., Molcard, A., et al. (2018). The boundary current role on the transport and stranding of floating marine litter: The French Riviera case. *Continental Shelf Research, 155*, 11–20. https://doi.org/10.1016/j.csr.2018.01.010

Ozcan, S., Tor, A., & Aydin, M. E. (2013). Investigation on the levels of heavy metals, polycyclic aromatic hydrocarbons, and polychlorinated biphenyls in sewage sludge samples and ecotoxicological testing. *CLEAN – Soil Air Water 41*:411–418. https://doi.org/10.1002/clen.201100187

Parodi, E., Govaert, L. E., & Peters, G. W. M. (2017). Glass transition temperature versus structure of polyamide 6: A flash-DSC study. *Thermochimica Acta, 657*, 110–122. https://doi.org/10.1016/j.tca.2017.09.021

Peez, N., Janiska, M.-C., & Imhof, W. (2019). The first application of quantitative 1H NMR spectroscopy as a simple and fast method of identification and quantification of microplastic particles (PE, PET, and PS). *Analytical and Bioanalytical Chemistry, 411*, 823–833. https://doi.org/10.1007/s00216-018-1510-z

Pellini, G., Gomiero, A., Fortibuoni, T., et al. (2018). Characterization of microplastic litter in the gastrointestinal tract of Solea solea from the Adriatic Sea. *Environmental Pollution, 234*, 943–952. https://doi.org/10.1016/j.envpol.2017.12.038

Pellis, A., Malinconico, M., Guarneri, A., & Gardossi, L. (2021). Renewable polymers and plastics: Performance beyond the green. *New Biotechnology, 60*, 146–158. https://doi.org/10.1016/j.nbt.2020.10.003

Phelan, A., (Anya), Ross H, Setianto NA, , et al. (2020). Ocean plastic crisis—mental models of plastic pollution from remote Indonesian coastal communities. *PLoS ONE, 15*, e0236149. https://doi.org/10.1371/journal.pone.0236149

Picó, Y., & Barceló, D. (2019). Analysis and prevention of microplastics pollution in water: Current perspectives

and future directions. *ACS Omega, 4*, 6709–6719. https://doi.org/10.1021/acsomega.9b00222

Pinnell, L. J., & Turner, J. W. (2019). Shotgun metagenomics reveals the benthic microbial community response to plastic and bioplastic in a coastal marine environment. Front Microbiol 10.https://doi.org/10.3389/fmicb.2019.01252

Prata, J. C. (2018). Microplastics in wastewater: State of the knowledge on sources, fate and solutions. *Marine Pollution Bulletin, 129*, 262–265. https://doi.org/10.1016/j.marpolbul.2018.02.046

Prociak, A., Kurańska, M., Cabulis, U., et al. (2018). Effect of bio-polyols with different chemical structures on foaming of polyurethane systems and foam properties. *Industrial Crops and Products, 120*, 262–270. https://doi.org/10.1016/j.indcrop.2018.04.046

Qin, Y., Qu, M., Kaschta, J., et al. (2020). Studies on recycled polyester. In S. S. Muthu (Ed.), *Recycled polyester: Manufacturing, properties, test methods, and identification* (pp. 29–67). Springer.

Rahman, A., & Miller, C. D. (2017). Chapter 6 - Microalgae as a source of bioplastics. In R. P. Rastogi, D. Madamwar, & A. Pandey (Eds.), *Algal Green Chemistry* (pp. 121–138). Elsevier.

Rai, P. K., Lee, J., Brown, R. J. C., & Kim, K.-H. (2021). Micro- and nano-plastic pollution: Behavior, microbial ecology, and remediation technologies. *Journal of Cleaner Production, 291*, 125240. https://doi.org/10.1016/j.jclepro.2020.125240

Ribeiro, F., O'Brien, J. W., Galloway, T., & Thomas, K. V. (2019). Accumulation and fate of nano- and micro-plastics and associated contaminants in organisms. *TrAC, Trends in Analytical Chemistry, 111*, 139–147. https://doi.org/10.1016/j.trac.2018.12.010

Ritchie, H., & Roser, M. (2018). Plastic pollution. Our World Data

Roch, S., Friedrich, C., & Brinker, A. (2020). Uptake routes of microplastics in fishes: Practical and theoretical approaches to test existing theories. *Science and Reports, 10*, 3896. https://doi.org/10.1038/s41598-020-60630-1

Rodrigues, M. O., Abrantes, N., Gonçalves, F. J. M., et al. (2018). Spatial and temporal distribution of microplastics in water and sediments of a freshwater system (Antuã River, Portugal). *Science of the Total Environment, 633*, 1549–1559. https://doi.org/10.1016/j.scitotenv.2018.03.233

Roohi, B. K., Kuddus, M., et al. (2017). Microbial enzymatic degradation of biodegradable plastics. *Current Pharmaceutical Biotechnology, 18*, 429–440. https://doi.org/10.2174/1389201018666170523165742

Rummel, C. D., Jahnke, A., Gorokhova, E., et al. (2017). Impacts of biofilm formation on the fate and potential effects of microplastic in the aquatic environment. *Environmental Science & Technology Letters, 4*, 258–267. https://doi.org/10.1021/acs.estlett.7b00164

Sadri, S. S., & Thompson, R. C. (2014). On the quantity and composition of floating plastic debris entering and leaving the Tamar Estuary, Southwest England. *Marine Pollution Bulletin, 81*, 55–60. https://doi.org/10.1016/j.marpolbul.2014.02.020

Sánchez, C. (2020). Fungal potential for the degradation of petroleum-based polymers: An overview of macro- and microplastics biodegradation. *Biotechnology Advances, 40*, 107501. https://doi.org/10.1016/j.biotechadv.2019.107501

Sanchez-Vidal, A., Thompson, R. C., Canals, M., & de Haan, W. P. (2018). The imprint of microfibres in southern European deep seas. *PLoS ONE, 13*, e0207033. https://doi.org/10.1371/journal.pone.0207033

Satti, S. M., Shah, A. A., Marsh, T. L., & Auras, R. (2018). Biodegradation of poly(lactic acid) in soil microcosms at ambient temperature: Evaluation of natural attenuation, bio-augmentation and bio-stimulation. *Journal of Polymers and the Environment, 26*, 3848–3857. https://doi.org/10.1007/s10924-018-1264-x

Schaider, L. A., Balan, S. A., Blum, A., et al. (2017). Fluorinated compounds in U.S. Fast Food packaging. *Environmental Science & Technology Letters, 4*, 105–111. https://doi.org/10.1021/acs.estlett.6b00435

Seleem, S., Hopkins, M., Olivio, J., & Schiraldi, D. A. (2017). Comparison of thermal decomposition of polystyrene products vs. bio-based polymer aerogels. Ohio J Sci 117:50–60. https://doi.org/10.18061/ojs.v117i2.5828

Setälä, O., Fleming-Lehtinen, V., & Lehtiniemi, M. (2014). Ingestion and transfer of microplastics in the planktonic food web. *Environmental Pollution, 185*, 77–83. https://doi.org/10.1016/j.envpol.2013.10.013

Sharma, S., Basu, S., Shetti, N. P., et al. (2021). Microplastics in the environment: Occurrence, perils, and eradication. *Chemical Engineering Journal, 408*, 127317. https://doi.org/10.1016/j.cej.2020.127317

Shen, M., Zeng, G., Zhang, Y., et al. (2019). Can biotechnology strategies effectively manage environmental (micro) plastics? *Science of the Total Environment, 697*, 134200. https://doi.org/10.1016/j.scitotenv.2019.134200

Singh, B., & Sharma, N. (2008). Mechanistic implications of plastic degradation. *Polymer Degradation and Stability, 93*, 561–584. https://doi.org/10.1016/j.polymdegradstab.2007.11.008

Sol, D., Laca, A., Laca, A., & Díaz, M. (2020). Approaching the environmental problem of microplastics: Importance of WWTP treatments. *Science of the Total Environment, 740*, 140016. https://doi.org/10.1016/j.scitotenv.2020.140016

Steensgaard, I. M., Syberg, K., Rist, S., et al. (2017). From macro- to microplastics — analysis of EU regulation along the life cycle of plastic bags. *Environmental Pollution, 224*, 289–299. https://doi.org/10.1016/j.envpol.2017.02.007

Sternberg, J., Sequerth, O., & Pilla, S. (2021). Green chemistry design in polymers derived from lignin: Review and perspective. *Progress in Polymer Science, 113*, 101344. https://doi.org/10.1016/j.progpolymsci.2020.101344

Sun, Q., Aguila, B., & Ma, S. (2019). Opportunities of porous organic polymers for radionuclide sequestration. *Trends Chem, 1*, 292–303. https://doi.org/10.1016/j.trechm.2019.02.010

Sundbæk, K. B., Koch, I. D. W., Villaro, C. G., et al. (2018). Sorption of fluorescent polystyrene microplastic particles to edible seaweed Fucus vesiculosus. *Journal of Applied*

Phycology, 30, 2923–2927. https://doi.org/10.1007/s10811-018-1472-8

Thompson, R. C. (2015). Microplastics in the marine environment: Sources, consequences and solutions. In M. Bergmann, L. Gutow, & M. Klages (Eds.), *Marine Anthropogenic Litter* (pp. 185–200). Springer International Publishing.

Tien, C.-J., Wu, W.-H., Chuang, T.-L., & Chen, C. S. (2009). Development of river biofilms on artificial substrates and their potential for biomonitoring water quality. *Chemosphere, 76*, 1288–1295. https://doi.org/10.1016/j.chemosphere.2009.06.013

Tiwari, N., Santhiya, D., & Sharma, J. G. (2020). Microbial remediation of micro-nano plastics: Current knowledge and future trends. *Environmental Pollution, 265*, 115044. https://doi.org/10.1016/j.envpol.2020.115044

Tiwari, N., Bansal, M., & Sharma, J. G. (2021). 14 - Metagenomics: a powerful lens viewing the microbial world. In: Shah MP, Rodriguez-Couto S (eds) Wastewater Treatment Reactors. Elsevier, pp 309–339

Tsiota, P., Karkanorachaki, K., Syranidou, E., et al. (2018). Microbial degradation of HDPE secondary microplastics: preliminary results. In: Cocca M, Di Pace E, Errico ME, et al. (eds) *Proceedings of the International Conference on Microplastic Pollution in the Mediterranean Sea*. Springer International Publishing, Cham, pp 181–188

Tu, C., Zhou, Q., Zhang, C., et al. (2020). Biofilms of microplastics. In D. He & Y. Luo (Eds.), *Microplastics in terrestrial environments: Emerging contaminants and major challenges* (pp. 299–317). Springer International Publishing.

Turner, A., & Holmes, L. A. (2015). Adsorption of trace metals by microplastic pellets in fresh water. *Environmental Chemistry, 12*, 600–610. https://doi.org/10.1071/EN14143

Ufarté, L., Laville, É., Duquesne, S., & Potocki-Veronese, G. (2015). Metagenomics for the discovery of pollutant degrading enzymes. *Biotechnology Advances, 33*, 1845–1854. https://doi.org/10.1016/j.biotechadv.2015.10.009

Van Cauwenberghe, L., & Janssen, C. R. (2014). Microplastics in bivalves cultured for human consumption. *Environmental Pollution, 193*, 65–70. https://doi.org/10.1016/j.envpol.2014.06.010

van der Hal, N., Ariel, A., & Angel, D. L. (2017). Exceptionally high abundances of microplastics in the oligotrophic Israeli Mediterranean coastal waters. *Marine Pollution Bulletin, 116*, 151–155. https://doi.org/10.1016/j.marpolbul.2016.12.052

Vandermaesen, J., Horemans, B., Bers, K., et al. (2016). Application of biodegradation in mitigating and remediating pesticide contamination of freshwater resources: State of the art and challenges for optimization. *Applied Microbiology and Biotechnology, 100*, 7361–7376. https://doi.org/10.1007/s00253-016-7709-z

Varotsos, C. A., Krapivin, V. F., & Mkrtchyan, F. A. (2019). New optical tools for water quality diagnostics. *Water, Air, and Soil Pollution, 230*, 177. https://doi.org/10.1007/s11270-019-4228-4

Varotsos, C. A., Krapivin, V. F., Mkrtchyan, F. A., et al. (2020). A novel approach to monitoring the quality of lakes water by optical and modeling tools: Lake Sevan as a case study. *Water, Air, and Soil Pollution, 231*, 435. https://doi.org/10.1007/s11270-020-04792-8

Vasile, C. (2018). Polymeric nanocomposites and nanocoatings for food packaging: A review. *Materials, 11*, 1834. https://doi.org/10.3390/ma11101834

von Moos, N., Burkhardt-Holm, P., & Köhler, A. (2012). Uptake and effects of microplastics on cells and tissue of the blue mussel Mytilus edulis L. after an experimental exposure. *Environmental Science and Technology, 46*, 11327–11335. https://doi.org/10.1021/es302332w

Wang, J., Tan, Z., Peng, J., et al. (2016). The behaviors of microplastics in the marine environment. *Marine Environment Research, 113*, 7–17. https://doi.org/10.1016/j.marenvres.2015.10.014

Wang, Q., Adams, C. A., Wang, F., et al. (2021). Interactions between microplastics and soil fauna: A critical review. *Critical Reviews in Environment Science and Technology, 0*, 1–33. https://doi.org/10.1080/10643389.2021.1915035

Wei, R., & Zimmermann, W. (2017). Microbial enzymes for the recycling of recalcitrant petroleum-based plastics: How far are we? *Microbial Biotechnology, 10*, 1308–1322. https://doi.org/10.1111/1751-7915.12710

Wei, R., Oeser, T., Schmidt, J., et al. (2016). Engineered bacterial polyester hydrolases efficiently degrade polyethylene terephthalate due to relieved product inhibition. *Biotechnology and Bioengineering, 113*, 1658–1665. https://doi.org/10.1002/bit.25941

Wu, P., Cai, Z., Jin, H., & Tang, Y. (2019). Adsorption mechanisms of five bisphenol analogues on PVC microplastics. *Science of the Total Environment, 650*, 671–678. https://doi.org/10.1016/j.scitotenv.2018.09.049

Xie, W., Li, T., Tiraferri, A., et al. (2021). Toward the next generation of sustainable membranes from green chemistry principles. *ACS Sustain Chem Eng, 9*, 50–75. https://doi.org/10.1021/acssuschemeng.0c07119

Yu, F., Yang, C., Huang, G., et al. (2020). Interfacial interaction between diverse microplastics and tetracycline by adsorption in an aqueous solution. *Science of the Total Environment, 721*, 137729. https://doi.org/10.1016/j.scitotenv.2020.137729

Yuan, J., Ma, J., Sun, Y., et al. (2020). Microbial degradation and other environmental aspects of microplastics/plastics. *Science of the Total Environment, 715*, 136968. https://doi.org/10.1016/j.scitotenv.2020.136968

Zadjelovic, V., Chhun, A., Quareshy, M., et al. (2020). Beyond oil degradation: Enzymatic potential of Alcanivorax to degrade natural and synthetic polyesters. *Environmental Microbiology, 22*, 1356–1369. https://doi.org/10.1111/1462-2920.14947

Zettler, E. R., Mincer, T. J., & Amaral-Zettler, L. A. (2013). Life in the "Plastisphere": Microbial communities on plastic marine debris. *Environmental Science and Technology, 47*, 7137–7146. https://doi.org/10.1021/es401288x

Zhang, H., Wang, J., Zhou, B., et al. (2018a). Enhanced adsorption of oxytetracycline to weathered microplastic polystyrene: Kinetics, isotherms and influencing factors. *Environmental Pollution, 243*, 1550–1557. https://doi.org/10.1016/j.envpol.2018.09.122

Zhang, Q., Hu, X.-M., Wu, M.-Y., et al. (2018b). Effects of different catalysts on the structure and properties of polyurethane/water glass grouting materials. *Journal of*

Applied Polymer Science, 135, 46460. https://doi.org/10.1002/app.46460

Zhao, J., Ran, W., Teng, J., et al. (2018). Microplastic pollution in sediments from the Bohai Sea and the Yellow Sea, China. *Science of the Total Environment, 640–641*, 637–645. https://doi.org/10.1016/j.scitotenv.2018.05.346

Ziemert, N., Alanjary, M., & Weber, T. (2016). The evolution of genome mining in microbes – a review. *Natural Products Reports, 33*, 988–1005. https://doi.org/10.1039/C6NP00025H

Zou, W., Xia, M., Jiang, K., et al. (2020). Photo-oxidative degradation mitigated the developmental toxicity of polyamide microplastics to zebrafish larvae by modulating macrophage-triggered proinflammatory responses and apoptosis. *Environmental Science and Technology, 54*, 13888–13898. https://doi.org/10.1021/acs.est.0c05399

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2021 doi:10.20944/preprints202109.0199.v1

Big Five Personality Detection Using Deep Convolutional Neural Networks

Waiel Tinwala
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
Waieltinwala_2k18co399@dtu.ac.in

Shristi Rauniyar
Dept. of Computer Science and Engineering
Delhi Technological University
Delhi, India
Shristirauniyar_2k18co343@dtu.ac.in

***Abstract*—Personality is the most critical feature that tells us about an individual. It is the collection of the individual's thoughts, opinions, emotions and more. Personality detection is an emerging field in research and Deep Learning models have only recently started being developed. There is a need for a larger dataset that is unbiased as the current dataset that is used is in the form of questionnaires that the individuals themselves answer, hence increasing the chance of unconscious bias. We have used the famous stream-of-consciousness essays collated by James Pennbaker and Laura King. We have used the Big Five Model often known as the five-factor model or OCEAN model. Document-level feature extraction has been performed using Google's word2vec embeddings and Mairesse features. The processed data has been fed into a deep convolutional network and a binary classifier has been used to classify the presence or absence of the personality trait. Hold-out method has been used to evaluate the model, and the F1 score has been used as the performance metric.**

Keywords—Personality Detection, Big Five model, Five-factor Model, OCEAN, Deep Convolutional Neural Network, Mairesse Features, stream-of-consciousness

I. Introduction

Personality is defined as nature, characteristics or qualities that form an individual's character. It is the combination of the individual's thought process, opinions, emotions, fears, goals, and aspirations.

The personality of the individual is the factor that determines how an individual is going to be with their family and friends, what kind of career they will opt for, how they will perform at work and exist in the global community. When individuals are well-versed with their personality, they know about their strengths and weaknesses which will help in self-development.

Some widely used theories for classifying personality traits are the Big Five model, Myers-Briggs Type Indicator, Keirsey Temperament Sorter, Three-trait personality model and Eysenck Personality Questionnaire [1] .

In this paper, we are going to classify personalities on the Big Five classification. This classification remains relatively uniform throughout one's lifetime and is the most widely accepted personality classification theory adapted by psychologists and researchers worldwide.

The **Big Five** model classifies personalities under five broad umbrella terms that encompass more labels. The five personality categories are described below:

- **Conscientiousness (CON)**: This trait describes the individual's ability to control impulses for goal-oriented tasks. It describes the ability of the individual to be organized, responsible, hard-working and adhering to the rules and regulations. If this trait is dominant in an individual, the individual is competent, self-disciplined and driven to achieve their goals. If this trait is not dominant, then it tells us that the individual is careless, procrastinates and is disorganized.
- **Agreeableness (AGR)**: This trait is all about how an individual treats others in relationships. Individuals dominantly having this trait are perceived as straight-forward, modest, sympathetic, and forgiving. On the other hand, individuals having a low measure of this trait are perceived as demanding, insulting, stubborn and unsympathetic.

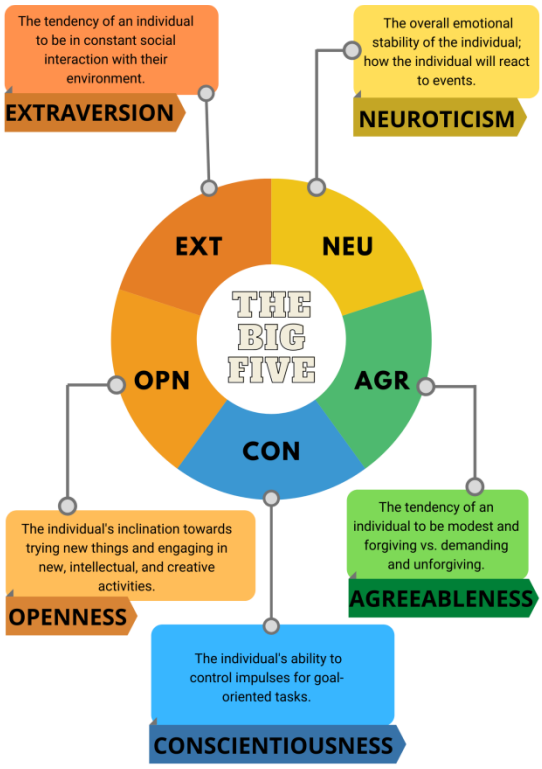


Fig. 1. The Big Five Model

- **Extraversion (EXT)**: This trait describes the tendency of an individual to be in constant social interaction and be interacting with their environment. Having a high measure of extraversion means the individual is very outgoing and enjoys being the centre of attention. Having a low measure of extraversion means that the individual is reserved in nature, prefers solitude over social interaction and does not want to be the centre of attention.
- **Openness (OPN)**: This trait reflects an individual's inclination towards trying new things and engaging in new, intellectual and creative activities. An individual having a high measure of openness means they are creative and think outside of the box. A low measure

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

© 2021 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2021 doi:10.20944/preprints202109.0199.v1

of this trait indicates that the individual prefers routine and traditional ways over new methods.

- **Neuroticism (NEU)**: This trait tells us about the overall emotional stability of the individual. It tells us how the individual is likely to interact and react to events. Having a high measure of this trait means that the individual is likely anxious, easily irritable and experiences mood shifts. An individual with a low measure of this trait is calm, confident and resilient.

The rest of this paper is organized as: related works, methodology, experimental results, conclusion, future work, acknowledgment and references.

II. Related Works

The first Personality Test ever conducted is commonly known as Woodworth Psychoneurotic Inventory and was developed during World War I. The U.S. military used it for screening army recruits. The screening was done for Post Traumatic Stress Disorder (PTSD). The Process Communication Model (PCM) was widely used. PCM was built by Taibi Kahler and was used for the selection of astronauts [1] .

The baseline methods of detecting personality are text, audio and visual cues. When using textual data, data preprocessing is a crucial step that directly affects the results. Generally, the textual features are extracted from the raw data and fed as input into Machine Learning models like Support Vector Machines (SVMs), Naïve Bayes Classifier, etc. The word embeddings are represented as vectors (Word2Vec, GloVe, etc.) and later more techniques are applied to yield results.

Kasula Chaithanya Pramodh et al. in 2016 [2] used the stream-of-consciousness and MyPersonality dataset. The MyPersonality dataset is a compilation of 250 users updating around 10,000 Facebook Status Updates. They have used Natural Language Toolkit for their model and their F1-scores are 0.665, 0.632, 0.625, 0.624 and 0.637 for OPN, CON, EXT, AGR and NEU traits respectively.

Carles Venture et al. [3] have used Convolutional Neural Networks in combination with Action Unit Recognition Systems and face detection to classify personality as per the Big Five model. They have used the First Impressions dataset that comprises of around 10,000 videos of various individuals.

Gokul K et al. [4] in 2018 have applied Bayes-Net classifier to classify personality as extrovert or introvert. They have used self-recorded audio samples as their dataset. Auditory Nerve Modeling that comprises of Voice Activity Detection in combination with Attribute Selection for dimensionality reduction was applied and they achieved an accuracy of 88.3%.

Bojan Simoski et al. [5] have developed a novel technique called Social Contagion Model for classifying personality on the Big Five Model. They used the responses of 25 individuals to the Big Five Questionnaire as their dataset.

Abir Abyaa et al. in 2018 [6] have used the StudentLife dataset that comprises of data from a group of 48 students. They have made use of supervised learning algorithms for the purpose of classification of personality as per the Big Five model. The used Support Vector Machines, Random forests, Logistic Regression, C4.5 Decision Tree and k-nearest neighbours algorithm.

Willy et al. in 2019 [7] have implemented C4.5 Decision Tree for classification of personality based on the Big Five Model. They have used the Twitter API to obtain the dataset that is roughly around 110 million tweets every day. They were able to obtain an accuracy of 64.30%.

In 2019, Tao Hong et al. [8] conducted a research to classify sentiment and detect personality. They have used the MDSTC dataset (comprises of speech, facial expressions and Galvanic Skin Response) and applied Deep Neural Network to classify personality. They achieved better results on performance metrics as compared to the state-of-the-art models.

Imanuel Buhapoda Drexel [9] has used Gaussian Naïve Bayes classifier to detect personality on the basis of Big Five Model. They have applied Word2Vec embeddings as well as FastText word embedding and used k-fold cross validation (k=5) for cross validation.

Songcheng Gao et al. in 2020 [10] have used a Multi-View Learning Approach that also includes a Multi-Task approach with StudentLife dataset that they obtains by developing a mobile application. They collected data from a group of 183 individuals from two different universities. The data collection is in the form of a questionnaire that one has to fill out on the mobile application. They have used mean absolute error and root mean squared error as the performance measure.

Marwa S. Salem et al. [11] have applied Multinomial Naïve Bayes, K-Nearest Neighbour, Support Vector Machine and Decision Trees on Egyptian Twitter Users Dataset to classify personality. According to their research, K-Nearest Neighbor was the best algorithm for Openness, Extraversion, Agreeableness and Neuroticism whereas Decision Tree performed best for Conscientiousness.

Kamal El - Demerdash et al. [12] have used Universal Language Model Fine-Tuning for personality trait detection on the Big Five model. They have applied this Model on the stream-of-consciousness dataset and they have managed to achieve around 1% better accuracy than most state-of-the-art models.

In 2017, a state-of-the-art research was performed by Navonil Majumder et al. in 2017 [13] where the researchers used a Deep Convolutional Neural Network to classify personality traits basis on the Big Five model. The model is coded in Python 2.7 and uses Google's word2vec embeddings and Mairesse features. This research was improved by Md. Abdur Rahman et al. in 2019 [14] , where the researchers applied different activation functions like Tanh, ReLU and leaky ReLU and observed significant improvement on the results. Even this was coded in Python 2.7.

We have coded the same model in Python 3.8 since Python 3.8 since python 2.7 is no longer supported (as of January 1, 2020). The methodology is explained in detail in the following section.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2021 doi:10.20944/preprints202109.0199.v1

III. Methodology

A Deep Convolutional Neural Network (CNN) has been trained to perform Personality Classification. This approach classifies five different personality traits: CON, AGR, EXT, OPN and NEU using the same network architecture. The methodology is described in detail in the following sections:

A. Dataset

The stream-of-consciousness essay dataset has been used. It is an assortment of 2,468 essays that are composed by people in a controlled environment. The writers also marked their personality traits themselves. The dataset was compiled by James Pennebaker and Laura King.

B. Data Preprocessing

Since the dataset contains of noise, we have performed cleaning by expanding abbreviations, removing unnecessary symbols and converting the text to lower-case. After performing cleaning, sentence-splitting was performed. The unique words are complied into a list that we have termed as vocabulary.

C. Deep CNN Architecture

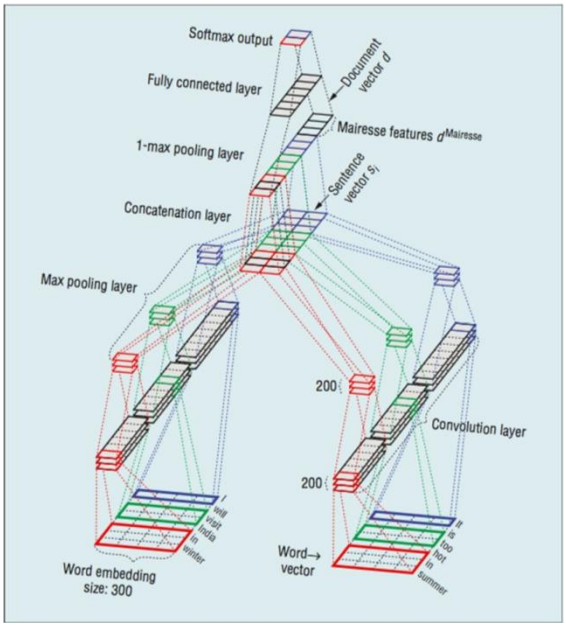


Fig. 2. Network Architecture [13]

The different layers are shown in Figure 1 and described below:

1) First layer—Word vectorization using Word2Vecs: Word-level feature extraction was performed using Google's Word2Vec Word embeddings library and the unknown words not found in Word2Vec were taken care of using uniform distribution. Equation (1) and Fig. 2 explain the Uniform Distribution Function.

$$F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases} \quad (1)$$

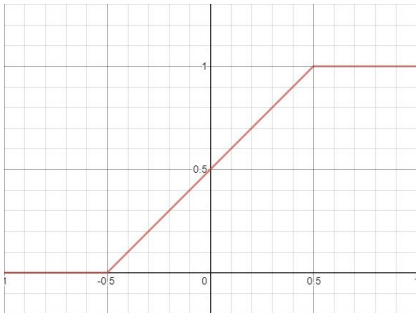


Fig. 3. Uniform Distribution Function

2) Second layer—Convolution: Convolutional filters along with max pooling were applied to the document in order to extract n-gram features from each sentence. The obtained feature vectors were concatenated into a sentence vectors.

3) Third layer—1-max pooling: The sentence vectors from Layer 3 were concatenated into document vectors. The dimension of a sentence vector is the multiple of n-grams. The document is considered only if it belongs to at least one feature.

4) Fourth layer—Concatenation: Mairesse features were concatenated to the document features obtained in Layer 3 were added to the document vector. Mairesse features are a set of document-level features such as Linguistic Inquiry and Word Count Features, utterance-type features, Medical Research Council features, and prosodic features.

5) Fifth layer—Activation: Sigmoid, Tanh and ReLU functions has been used as activation function. Sigmoid function is explained in (2) and Fig. 4. Tanh function is explained in (3) and Fig. 5. ReLU function is explained in (4) and Fig. 6.

$$S(x) = \frac{1}{1+e^{-x}} \quad (2)$$

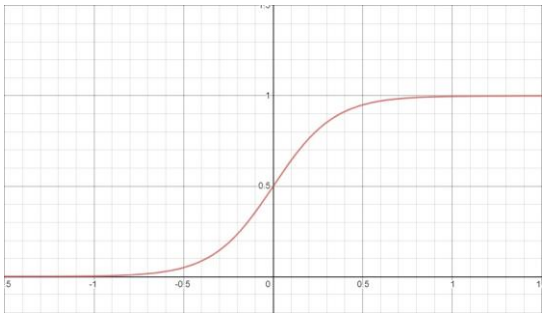


Fig. 4. Sigmoid Function

$$\text{Tanh}(x) = \frac{\sinh(x)}{\cosh(x)} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (3)$$

6) Sixth layer—Classification: The sixth and last layer is a binary classifier that classifies if the given personality trait is present or not by using softmax activation function (5) which is the same as sigmoid function for binary classifiers.

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_i}} \quad (5)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2021 doi:10.20944/preprints202109.0199.v1

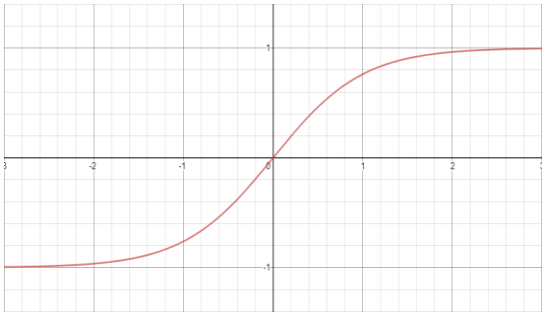


Fig. 5. Tanh Function

$$F(x) = \begin{cases} 0\,, & x < 0 \\ x\,, & x \geq 0 \end{cases} \quad (4)$$

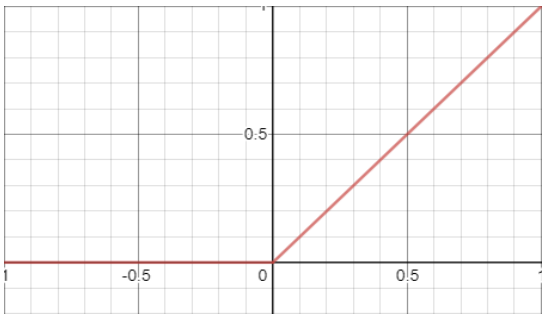


Fig. 6. ReLU Function

IV. Experimental Results

The constructed dataset is trained and evaluated using hold-out technique. The performance measure that is used is F1 score. It is the harmonic mean of Precision and Recall. True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) are the measures used to calculate Precision and Recall. F1-score is given by (8):

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

$$F1 - score = \frac{2*Precision*Recall}{Precision+Recall} \quad (8)$$

The results that we have obtained are described in Fig. 7.

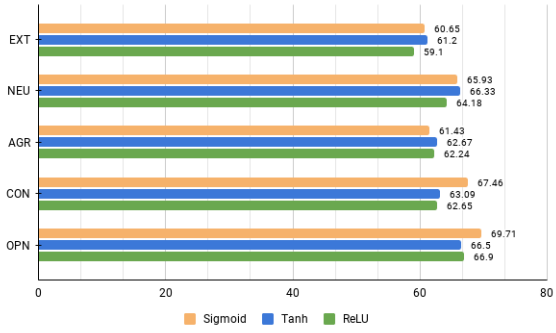


Fig. 7. F1-scores when applying various activation functions

We can observe that Tanh function performs best for traits Extraversion, Neuroticism and Agreeableness with F1-scores 61.2%, 66.33% and 62.67% respectively. Sigmoid performs best for Openness and Conscientiousness with F1-scores of 69.71% and 67.46% respectively.

V. Comparative Analysis

Activation functions are used in neural networks as outputs for the nodes in the layers. They are used to add non-linearity to the network. We have used 3 different activation functions for our model namely: Sigmoid, Tanh, and ReLU.

The state-of-the-art model has been trained for 6 epochs. The model was evaluated using k-fold cross-validation (k=10). Due to GPU and CPU constraints, we have trained our model for 50 epochs and evaluated our model using hold-out cross-validation technique with 9:1 as the train-test ratio.

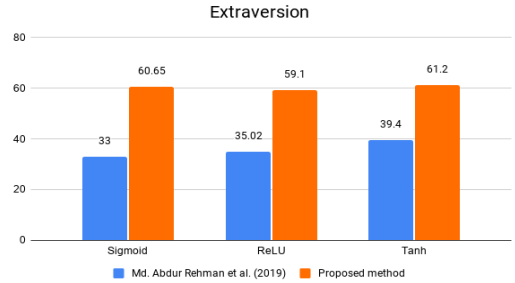


Fig. 8. F1-score comparison of Personality Trait: Extraversion

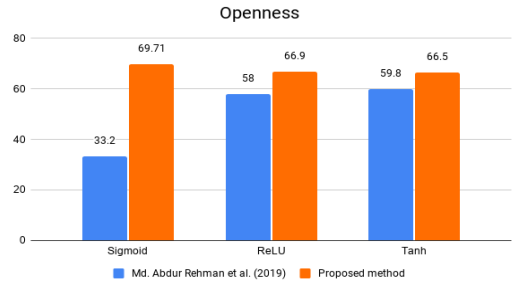


Fig. 9. F1-score comparison of Personality Trait: Openness

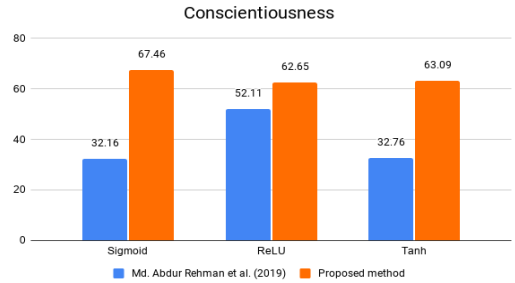


Fig. 10. F1-score comparison of Personality Trait: Conscientiousness

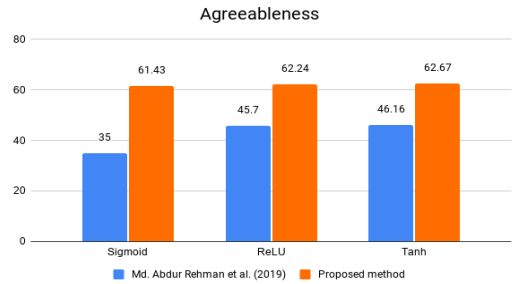


Fig. 11. F1-score comparison of Personality Trait: Agreeableness

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2021 doi:10.20944/preprints202109.0199.v1

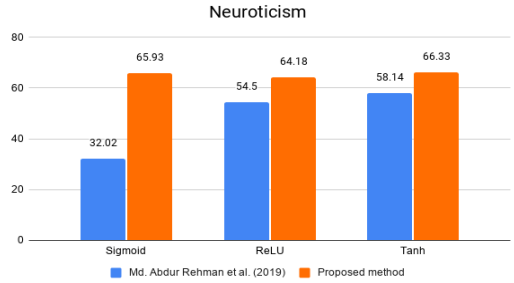


Fig. 12. F1-score comparison of Personality Trait: Neuroticism

Sigmoid and Tanh are non-linear activation functions whereas ReLU is a linear function. We have trained our model using these three activations and compared our results. The comparison for each of the five traits can be seen in Fig. 8, 9, 10, 11 and 12. We observe that the non-linear functions i.e.; Sigmoid and Tanh performed better than the linear activation ReLU in all cases. We also observe that our model performs better than the state-of-the-art model.

VI. Conclusion

We have been able to successfully code this model in Python 3.8 and have been able to optimize the previous code structure as well. We have been able to achieve better performance measures than previous state-of-the-art methods as can be observed in Fig. 5. The proposed model had used k-fold cross-validation technique (k=10) whereas we have used hold-out method because of limited resources and yet have achieved better results.

TABLE I. F1-scores comparison table

Personality Traits	Activation Function	Md. Abdur Rehman et al. (2019)	Proposed method
EXT	Sigmoid	33.00	60.65
	ReLU	35.02	59.10
	Tanh	39.40	61.20
NEU	Sigmoid	32.02	65.93
	ReLU	54.50	64.18
	Tanh	58.14	66.33
AGR	Sigmoid	35.00	61.43
	ReLU	45.70	62.24
	Tanh	46.16	62.67
CON	Sigmoid	32.16	67.46
	ReLU	52.11	62.65
	Tanh	32.76	63.09
OPN	Sigmoid	33.20	69.71
	ReLU	58.00	66.90
	Tanh	59.80	66.50

Fig. 13. F1 score Comparison Table between state-of-the-art method and our proposed model

VII. Future Work

As observed in this paper, personality detection is a vital field that requires more research. Application of personality detection ranges from enhancing voice assistants (Alexa, Siri, Google Assistant, Cortana, etc.) to areas like forensics and health care. It is also used for hiring purposes and psychological studies.

Less biased word embeddings could be explored and the accuracy of the new model could be compared to the current model.

We have also observed the need of larger and more accurate datasets. Most ways of collecting data for this field include handing out questionnaires. Individuals may or may not answer the questions honestly and there must be a way to check the credibility of the responses.

Acknowledgment

We take this opportunity to thank Dr. Ruchika Malhotra for her valuable guidance throughout this research, from identifying the problem statement to presenting the research. We would also extend our thanks to Delhi Technological University for providing us with all the resources that were required to complete this research successfully.

References

[1] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, “Recent trends in deep learning based personality detection,” *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2313–2339, 2020.

[2] K. C. Pramodh and Y. Vijayalata, “Automatic personality recognition of authors using big five factor model,” *2016 IEEE Int. Conf. Adv. Comput. Appl. ICACA 2016*, pp. 32–37, 2017.

[3] C. Ventura, D. Masip, and A. Lapedriza, “Interpreting CNN Models for Apparent Personality Trait Regression,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017, vol. 2017-July, pp. 1705–1713.

[4] K. Gokul and S. Lalitha, “Personality Identification Using Auditory Nerve Modelling of Human Speech,” *2018 Int. Conf. Adv. Comput. Commun. Informatics*, pp. 1731–1737, 2018.

[5] E. F. M. Araújo, B. Simoski, and M. Klein, “Applying machine learning algorithms for deriving personality traits in social network,” in *Proceedings of the ACM Symposium on Applied Computing*, 2018, pp. 346–349.

[6] A. Abyaa, M. Khalidi Idrissi, and S. Bennani, “Predicting the learner’s personality from educational data using supervised learning,” *ACM Int. Conf. Proceeding Ser.*, pp. 1–7, 2018.

[7] Willy, E. B. Setiawan, and F. N. Nugraha, “Implementation of Decision Tree C4.5 for Big Five Personality Predictions with TF-RF and TF-CHI2 on Social Media Twitter,” *2019 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Big Data Artif. Intell. IC3INA 2019*, pp. 114–119, 2019.

[8] T. Hong, X. Sun, F. Tian, and F. Ren, “Sentiment Classification And Personality Detection Via Galvanic Skin Response Based on Deep Learning Models,” *2019 5th Int. Conf. Big Data Comput. Commun.*, pp. 313–317, 2019.

[9] I. B. Drexel, “Feature Engineering and Word Embedding Impacts for Automatic Personality Detection on Instant Message,” in *Proceedings of 2019 International Conference on Information Management and Technology, ICIMTech 2019*, 2019, vol. 1, no. August, pp. 155–159.

[10] S. Gao, W. Li, L. J. Song, X. Zhang, M. Lin, and S. Lu, “PersonalitySensing: A Multi-View Multi-Task Learning Approach for Personality Detection based on Smartphone Usage,” pp. 2862–2870, 2020.

[11] M. S. Salem, S. S. Ismail, and M. Aref, “Personality Traits for Twitter Users in the Egyptian Dialect Writing Dataset,” *Icsie*, pp. 206–211, 2019.

[12] K. El-demerdash, R. A. El-khoribi, M. A. I. Shoman, and S. Abdou, “Psychological Human Traits Detection based on Universal Language Modeling,” *Egypt. Informatics J.*, no. xxxx, 2020.

[13] T. Yılmaz, A. Ergil, and B. İlgen, “Deep Learning-Based Document Modeling for Personality Detection from Turkish Texts,” *Adv. Intell. Syst. Comput.*, vol. 1069, pp. 729–736, 2020.

[14] M. A. Rahman, A. Al Faisal, T. Khanam, M. Amjad, and M. S. Siddik, “Personality Detection from Text using Convolutional Neural Network,” *1st Int. Conf. Adv. Sci. Eng. Robot. Technol. 2019, ICASERT 2019*, vol. 2019, no. Icasert, pp. 1–6, 2019.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 September 2021 doi:10.20944/preprints202109.0199.v1

Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Comparative Study of Nonlinear Controllers for Standalone PV System

Manisha[1], MD. Modassir Masoom[2] and Narendra kumar[3]
[1, 2, 3]Department of Electrical Engineering, Delhi Technological University, Delhi, India
[1]manisha_2k19ci03@dtu.ac.in, [2]mdmodassirmasoom_2k20phdee505@dtu.ac.in, and [3] narendrakumar@dtu.ac.in

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532701

Abstract: **The following paper shows the analysis of nonlinear controllers to deal with the nonlinearity present in the Photo-Voltaic (PV) system. The Output power of Photo-Voltaic (PV) panels changes if there are any changes in load or atmospheric conditions like temperature & irradiation level. A tracking technique is always required to obtain the utmost power from the panel. PV systems can be made more productive by simulating controllers to track maximum power point (MPP). The circuit simulated in this paper is having PV connected to a resistive load through DC/DC Boost converter, & a Sliding Mode Controller (SMC) and Back-Stepping Control (BSC) design is used to provide a dedicated control performance. The Sliding Mode and Backstepping controllers for obtaining the optimal power point and stability are based on Lyapunov function analysis and then efficiency with which these two controllers track the ideal power is obtained graphically. Model and the control methodologies are analyzed with a simulation model that is entirely done in Simulink/Matlab.**

Key Words- Sliding Mode Controller(SMC), Backstepping Controller (BSC), Lyapunov, DC/DC Converter, MPP, PV

I.INTRODUCTION

Too much dependency on non-renewable sources for meeting the energy demands has created too much adverse impact on the environment. Fossil fuels are having carbon components in them. They are unfriendly to our nature and also have a negative impact on health. Moreover, these resources are present in a very limited amount i.e. they are unsustainable and hence it cannot have so much dependency on them. Excessive usage of these can also cause an additional burden on our environment. As a result of which human tends to look upon the various energy sources to meet their requirements, these sources mainly depend on solar energy, wind energy, biomass, etc. These sources are proven to be more environmental friendly. Other advantages of relying on these renewable sources are our requirement can be accomplished without compromising the ability of the upcoming generation to fulfill their demands.

Solar energy is also can be used as an alternative to meet our energy demands. It is freely and plentifully available in nature. Other benefits of using solar energy are that it is having low cost maintain, no noise, etc. Moreover, solar energy can be easily converted to electrical energy by using PV modules. But the efficiency of these modules is very less and it depends on various factors like irradiation changes, temperature variations, and changes in load. The efficiency of solar panels can be increased to make them work at their Maximum Power Point(MPP). So, to make them work at MPP, it is required to use different kinds of techniques, which includes conventional method like Perturb and Observe(P&O), Incremental Conductance (InC) and other approaches like Fuzzy logic Control (FLC), Neural Network (NN). But, the characteristics of solar PV is non-linear so non-linear controller like Conventional Sliding Mode Controller (CSMC), advanced Sliding Mode Controller (ASMC), Backstepping Controller (BSC) are best suited to overcome the non-linearity of PV systems.

P&O includes the perturbation of duty cycle of DC-DC converter by a small amount which depends on the power variation. In this method, either current or voltage value is sensed. The exactness of this algorithm is based upon the amount of perturbation. Smaller perturbation can make this algorithm more accurate. InC methods use both current and voltage value as sensing parameters and variation in Power-Voltage (P-V) curve slope is observed. These two are very old methods and work great when irradiation and temperature levels are the same but in case of inevitable variation in these parameters, these two methods become least preferable.

Intelligent methods like Fuzzy controller depend on the information of a particular methodology to decided the rule base for optimizing the technique. Artificial Neural Network includes training of networks of neural. This method is having so much complexity and very costly implementation [1].

Non-Linear methods like SMC and BS are robust methods and can track the MPP and force the PV system to work at the MPP and hence are very effective methods to increase the efficiency of solar PV modules.

Section II of this paper demonstrates the model of PV [2], having an equivalent model & section III is having system specifications for both converter and PV module. IV shows the Converter topology and equations which define the system. The developed control techniques are briefly described in section V [3]. Section VI is having the observations achieved by both techniques. Section VII presents the results obtained

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

by SMC and BSC along with a comparison table that includes various performance parameters.

II. SOLAR CELL MODELLING

The solar cell can be represented as an electrical circuit & its modeling can be done as

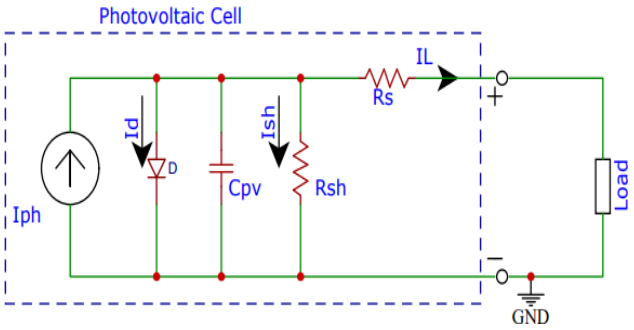


Fig.1.Solar Cell Equivalent

The above circuit can be implemented as a Solar cell [4].

The photovoltaic current expression is

$$I_{pv} = I_{ph} - I_D = I_{ph} - I_o\{\exp(^{e\,(V_{PV} + I_{PV}\,Rs)/\eta[KT)}) - 1\} \quad (1)$$

The output voltage of the PV module is given by

$$V_{pv} = V_{Thermal} * \ln\,(I_{ph} - I_{pv}/I_{saturation}\,) \quad (2)$$

From I_{PV} & V_{PV} equation it can be observed that there is a logarithmic relationship among the quantities involved in the above expressions.

III. SPECIFICATIONS

A.Converter Parameters

TABLE I. Converter Parameters

Parameters	Value
Cpv	100µF
L	20µH
Co	470µF
R	50Ω
Fsw(Switching Frequency)	5KHz
Vd(Diode Voltage drop)	0.82V

B.PV Module Parameters

TABLE II. PV Module Parameters

Parameters	Value
Series Cells	1
Parallel Cells	1
Voc	21.1V
Isc	3.8 A
Vmp	17.1 V
Imp	3.5 A
Maximum Power (MP)	59.85W

IV. BOOST CONVERTER DYNAMICS

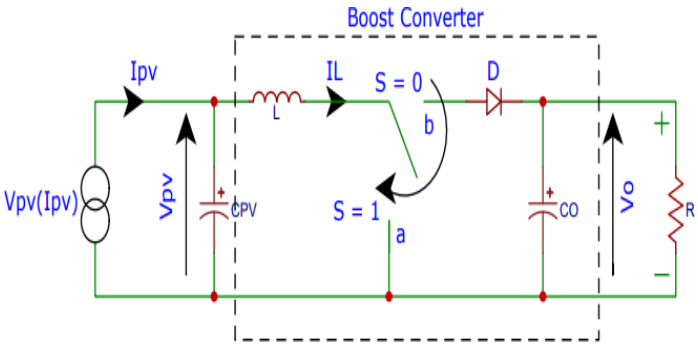


Fig.2. The system with Boot converter

When Switch 'S' is OFF, the system equations are- [5]-[6]

$$\frac{dI_L}{dt} = \frac{V_{pv}}{L} - \frac{V_o}{L}$$

$$\frac{dV_o}{dt} = \frac{I_L}{C_o} - \frac{V_0}{C_0 R}$$

When Switch 'S' is ON, the system equations are-

$$\frac{dI_L}{dt} = \frac{V_{pv}}{L}$$

$$\frac{dV_o}{dt} = \frac{I_L}{C_o}$$

'd' is the duty ratio and a control signal is given to the switching device, the system equations can be described as:

$$\frac{dI_L}{dt} = \frac{V_{pv}}{L} - (1-d)\frac{V_o}{L} \quad (3)$$

$$\frac{dV}{dt} = \frac{I_L}{C_o}(1-d) - \frac{V_o}{C_0 R} \quad (4)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

Where the value of 'd' is lie between 0 and 1

V. DESIGNING OF NON-LINEAR CONTROLLERS

A.SMC Controller

SMC approach includes two steps [4]-[5]: -
1.Selection of Sliding /Switching Surface
2.Implementation of proper control Signal

Sliding Surface can be obtained as [6]-

$$\frac{\partial P_{pv}}{\partial I_{pv}} = 0 \quad (5)$$

$$\frac{\partial P}{\partial I} = \frac{\partial (I_{PV}^2 * R)}{\partial I_{pv}} = I_{pv}\left(2R_{pv} + I_{pv}\frac{\partial R_{pv}}{\partial I_{pv}}\right)$$

Where $Rpv = Vpv/Ipv$

Derived switching Surface is

$$\sigma = 2R_{pv} + I_L \frac{\partial R_{pv}}{\partial I_L}$$

The modified duty cycle is (6)

$$d_{modified} = \begin{cases} d - \Delta d & ,\sigma < 0 \\ d + \Delta d & ,\sigma > 0 \end{cases} \quad (7)$$

Control equation can be obtained as

$$\sigma^{\cdot} = \frac{d\sigma}{dt} = 0$$

$$= \frac{d\sigma}{dt}\frac{dI_{pv}}{dt} = 0$$

$$= \frac{d\sigma}{dt} * I_L^{\cdot} = 0$$

$$= \frac{d\sigma}{dt} * \left[\frac{V_{pv}}{L} - (1 - deq)\frac{V_o}{L}\right] = 0$$

$$d_{eq} = 1 - \frac{V_{pv}}{V_o}$$

The control signal is given by [7]: -

$$d = \begin{cases} 1 & d_{eq} + K\,sign(\sigma) \geq 1 \\ deq + K\,sign(\sigma), & 0 < d_{eq} + K\,sign(\sigma) < 1 \\ 0 & d_{eq} + K\,sign(\sigma) \leq 0 \end{cases} \quad (8)$$

Where K is a positive constant.

B.BSC Designing

The backstepping approach is a recursive design method [7]-[8].
Designing of Backstepping controller involves the selection of Lyapunov function and then choosing an appropriate Control law to make first differentiation of scalar energy Lyapunov function negative [9]-[10]
These steps are followed while designing the controller-
1.Incremental Conductance method is used to find the error dynamics which is associated with voltage.
2.Based on the above-calculated error dynamics a suitable Lyapunov function in selected.
3.Examine the first differentiation of Lyapunov function, it must be negative.
4. Ensure the final Control law depends on the Lyapunov function.

The first error is defined as-

$$e_1 = V_{pv} - V^*_{pv} \quad (9)$$

Where Vpv and V^*pv are Vpv voltage and reference Voltage respectively.
Lyapunov Function can be defined as-

$$V_1 = \frac{1}{2}e_1^2 \quad (10)$$

Stabilization Function is defined as-

$$\alpha = I_{PV} + C_{PV}(K_1 e_1 - V_{PV}^{*\cdot}) \quad (11)$$

Next to tracking error can be defined as-

$$e_2 = I_L - \alpha \quad (12)$$

New Lyapunov Function can be-

$$V_2 = V_1 + \frac{1}{2}e_2^{\cdot 2}$$

Derivative of above Lyapunov Function (13)

$$V_2^{\cdot} = -K_1 e_1^2 - K_2 e_2 < 0$$

Control Law can be defined as-

$$d = 1 - \frac{1}{V_o}[V_{PV} - L\alpha^{\cdot} - L(\frac{1}{C_{PV}}e_1 - K_2 e_2)] \quad (14)$$

Where K_1 and K_2 are parameters of design.

The below Graph shows the step changes of irradiation at different time instances at the constant temperature of 25°C

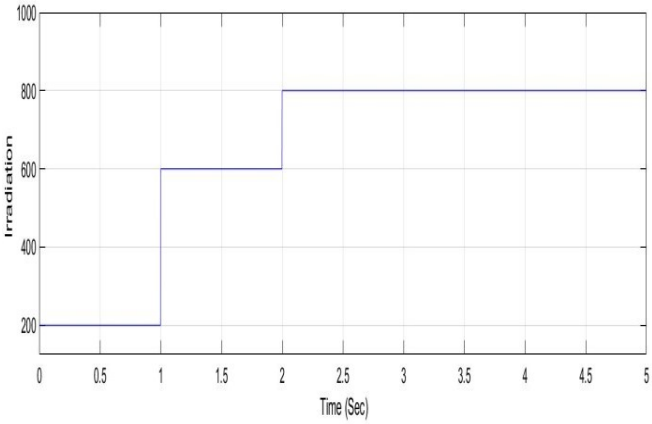


Fig. 3. Irradiation with Step variation

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

V. Observations With Two Approaches

A.SMC Approach

The below waveform shows the output power of the converter which is having higher oscillations initially [11]. Input Voltage at the input of Boost Converter is 17 V and this voltage is boosted to 45 V(for time 2 to 5 sec).

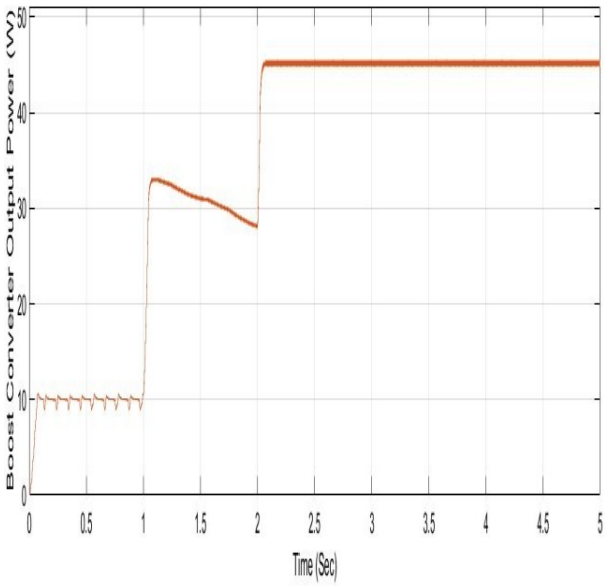


Fig.4.The waveform of Boost Converter Output Power

The below graph shows the power of PV which is obtained by using the SMC technique which loses its effectiveness for the time from 2 to 5 sec.

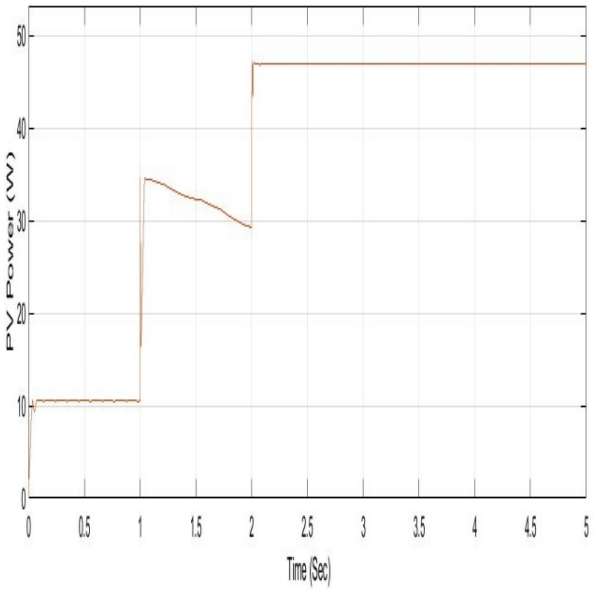


Fig.5. PV Output Power

B.BSC Approach

The below figure represents the Boost output power in which voltage at the input side of Converter is 17 and output of boost converter is 45V (for t = 2to 5 Sec).

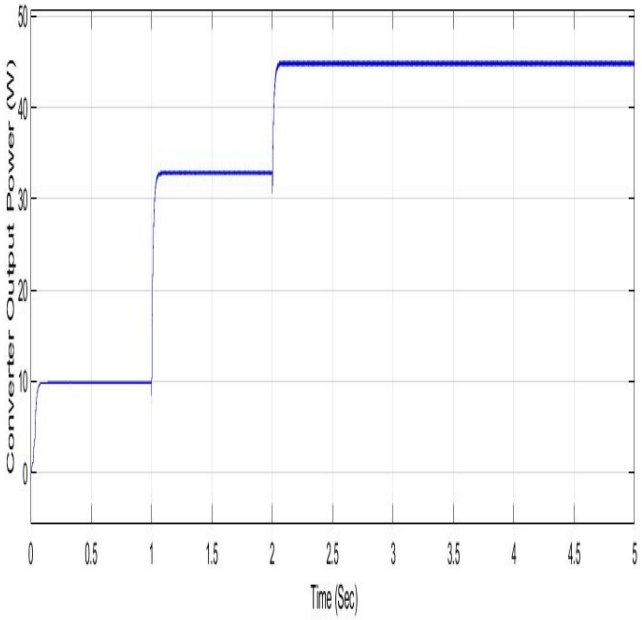


Fig.6.The waveform of Boost Converter Output Power

The below figure represents PV output power which is approximately free from Oscillations even at the points of step changes in irradiation [12].

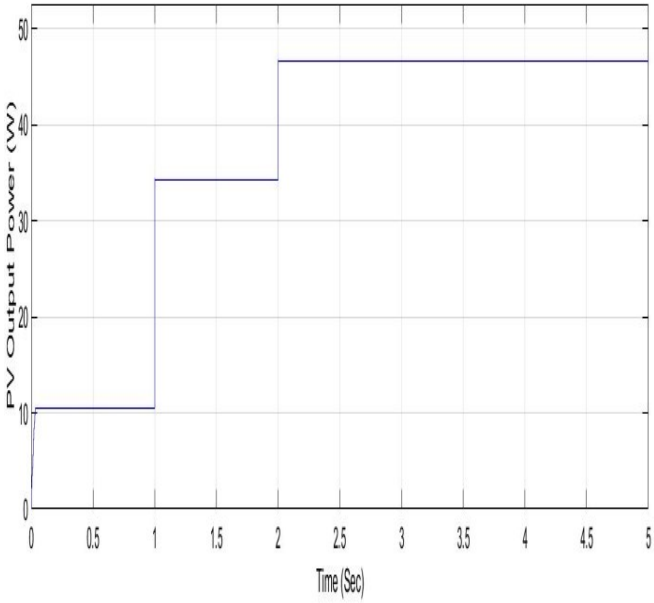


Fig. 7. PV Output Power

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

VI.RESULTS

A.SMC Approach

Below waveform shows that the boost converter's input power is approximately equal to its output Power but SMC is not working well for t= 2 to 5sec.

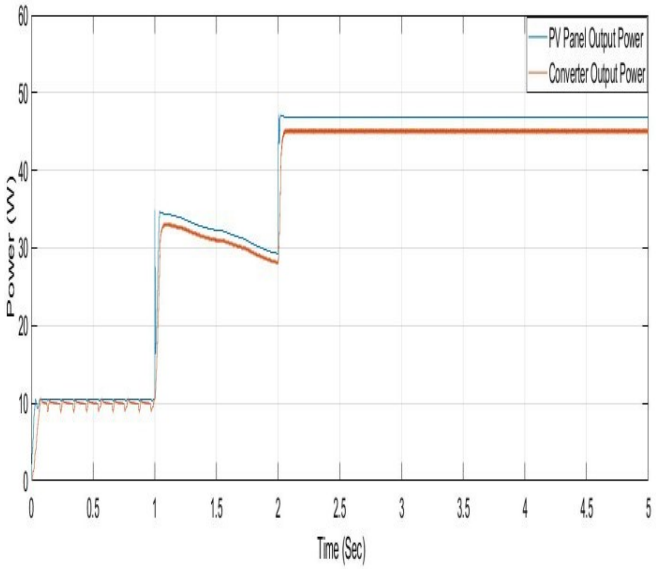


Fig. 8. PV and Converter Output Power

This graph shows the ideal and PV power curve. Here PV power is having oscillations for t = 0 to 1sec and for t=1 to 2 sec, the power tracking is not appropriate. For t =2 to 5 sec the SMC works better.

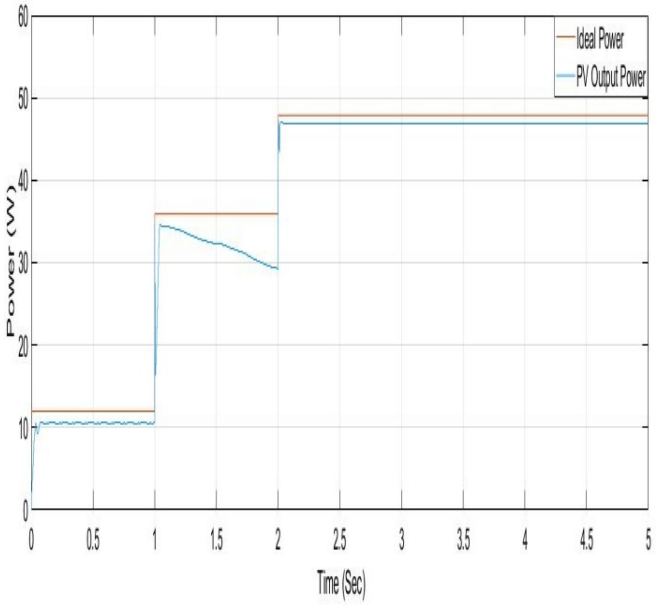


Fig. 9. Ideal and PV Power

This graph presents the efficiency of SMC in the tracking of ideal power. From t=0 to1 At t=1, the tracking plot shows a large decline.

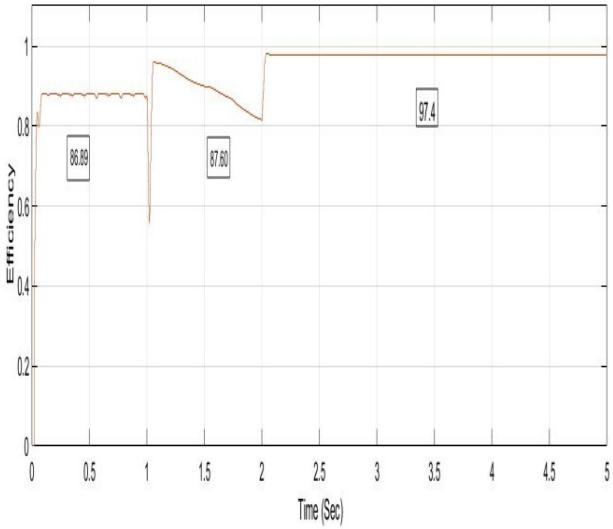


Fig. 10.The waveform of Tracking Efficiency

TABLE .III SMC Efficiency

Time (Sec)	% Efficiency
0-1	86.89
1-2	87.60
2-5	97.4

B.BSC Approach

Below waveform presents the converter and Power of PV. Here the converter's input power is approximately equal to its output power and there are no oscillations present [14]-[15].

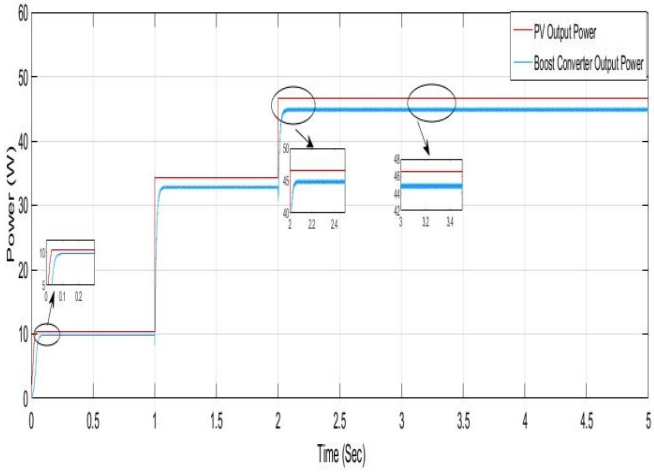


Fig. 11. PV and Converter Output Power

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

The below graph shows the ideal and output PV power plot. It can be observed that there are no oscillations present in output PV power and it is following the ideal power

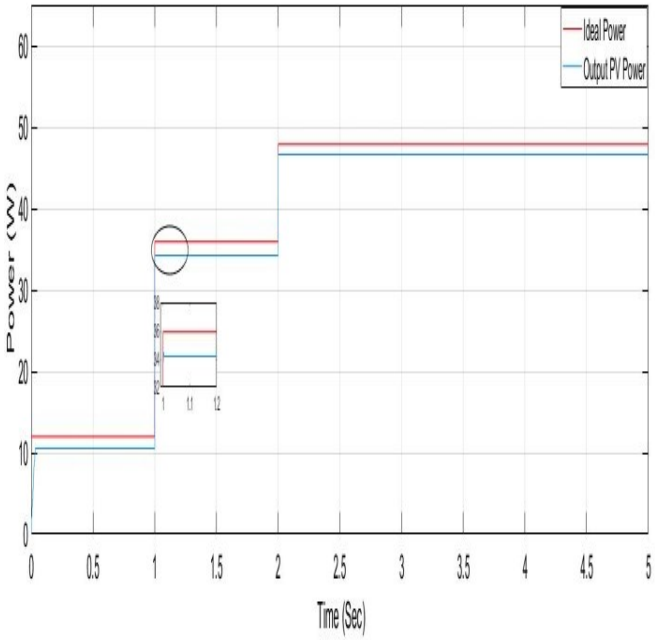


Fig. 12. Ideal and PV Power

The below graph shows the tracking efficiency of the BSC method which is proven to be more robust because it shows that the modeled system works effectively under all the irradiation whether it is constant or having steps changes.

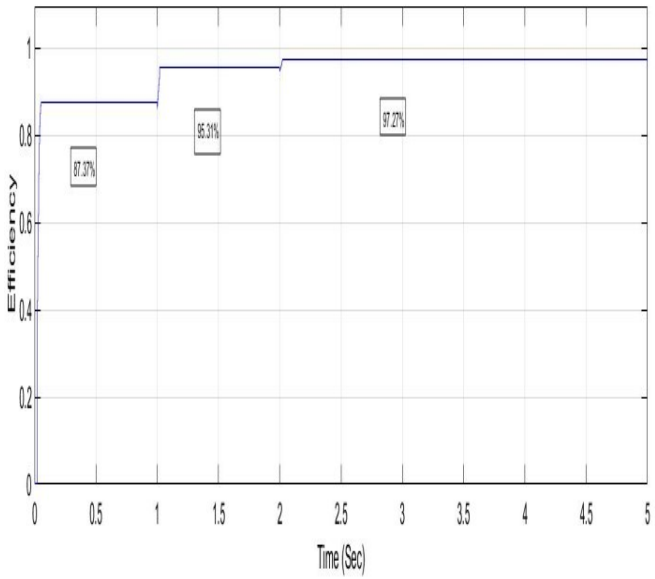


Fig. 13.The waveform of Tracking Efficiency

TABLE IV. BSC Efficiency

Time (Sec)	% Efficiency
0-1	87.37
1-2	95.31
2-5	97.27

TABLE V. Comparison

Performance Criteria	SMC	BSC
Simulation Complexity	Less	More
Accuracy of Controlling Technique(to step changes in irradiation)	Low	High
Tracking Efficiency	low	high

VII. CONCLUSION

This paper analyzes the behavior of solar panels connected to a converter. The switching of the converter is controlled through non-linear controllers i.e. SMC and BSC. It can examine that if there are step variations in the irradiation levels, the BSC can easily handle these variations with the least oscillations at the point of sudden changes in these irradiations whereas SMC works fine in case of constant irradiations but during the sharp changes in irradiation, this controller becomes less reliable. Furthermore, it can be easily analyzed with the help of an efficiency graph, BMC method is highly efficient, reliable and the tracking efficiency of the BSC technique does not drop down at the point of sharp changes in the irradiation level as compared with SMC where efficiency decreased to a greater extent during the sharp changes and the system become less reliable. Thus BSC method is superior to the SMC controller.

REFERENCES

[1] R. A. Shirazi, I. Ahmad, M. Arsalan, and M. Liaquat, "Integral Backstepping based MPPT controller for photovoltaic system using SEPIC converter," 2019 7th International Conference on Control, Mechatronics and Automation (ICCMA), 2019.

[2]H. Abouobaida, M. Cherkaoui and M. Ouassaid, "Robust maximum power point tracking for photovoltaic cells: A backstepping mode control approach," 2011 International Conference on Multimedia Computing and Systems, 2011.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

[3] K. Ali, Q. Khan, L. Khan, F. Hayat, U. A. Khan and A. Ullah, "MPPT for Photovoltaic System using non-linear Hybrid Robust Integral Backstepping Controller," 2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2020.
[4]A. Sharma, S. Gupta and N. Kumar, "Modelling and Comparison of Different MPPT Algorithms for a Standalone PV System," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020.
[5] H. Afghoul, D. Chikouche, F. Krim and A. Beddar, "A compararative study between sliding mode controller and P&O controller applied to MPPT," 2013 International Renewable and Sustainable Energy Conference (IRSEC), 2013.
[6]T. Zhao, N. Ji, H. Xu and T. Zhang, "Study on Variable Step Sliding-Mode Based MPPT for PV Array," 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2019.
[7] A. Sharma and N. Kumar, "Lyapunov Stability Theory based Non Linear Controller Design for a Standalone PV System," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020.
[8] Chandramouli, A., and V. Sivachidambaranathan. "Extract maximum power from PV system employing MPPT with FLC controller" 2019 Journal of Electrical Engineering and Automation,2019.
[9] M. M. Masoom, N. Kumar and A. R. Kulkarni, "MPPT Based Grid Connected Photovoltaic System Using Flyback Converter," IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, pp. 1-5,2020
[10]O. Barambones, P. Alkorta, and J. M. Gonzalez de Durana, “Sliding mode position control for real-time control of induction motors,” International Journal of Innovative Computing, Information and Control, Vol. 9, no. 7, pp. 2741-2754, 2013.
[11] Jamri, “Modeling and Control of a Photovoltaic Energy System Using the State-Space Averaging Technique,” American Journal of Applied Sciences, Vol. 7 No. 5, pp. 682-691, 2010
[12] A. Levant, “Principles of 2-sliding mode design,” Automatica, Vol. 43, Issue 4, pp. 576-586, April 2007
[13] K. M. Tsang and W. L. Chan, “Three-level grid-connected photovoltaic inverter with maximum power point tracking,” Energy Conversion and Management, Vol. 65, pp. 221-227, January 2013.
[14] H. Patel and V. Agarwal, “MATLAB-based modeling to study the effects of partial shading on PV array characteristics,” IEEE Transactions on Energy Conversion, vol. 23, no. 1, pp. 302-310, March 2008.
[15]C. M. P. Santana, “Backstepping Control of a Photovoltaic DC/AC Converter,” no. November, pp. 1–10, 2016.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:39:19 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Data rate over different applications in 5G and beyond Networks

Manpreet Kaur
Department of Computer Science & Engineering,
Delhi Technological University, Delhi, India
manpreettiet@gmail.com

Rajesh Kumar Yadav
Department of Computer Science & Engineering,
Delhi Technological University, Delhi, India
rkyadav@dtu.ac.in

***Abstract*—To address the explosive traffic demands, the capacity of the fading channel is increasingly becoming a prime concern in the designing of the wireless communication system. In this paper, we present a survey of the existing work related to the channel capacity (CC) model over various fading channels. With an elaborated explanation of the theory of CC, definitions of CC based on the channel state information are reviewed. Analytical expressions derived for the CC of the fading channels are reviewed based on the complexity of the expressions. To complement this, review of the diversity combining technique to enhance the CC is discussed. We also summarize the work related to CC in various types of wireless networks such as cognitive radio networks, cooperative networks, body area networks, hybrid networks and device-to-device networks etc. We finally cover the future research directions, including less explored aspects of the CC that can be studied to design efficient communication systems.**

Keywords- Data transmission, channel state information, diversity combining, WSN, VANET, hybrid system.

I. INTRODUCTION

With the recent advancements in wireless communication, there has been tremendous growth in the applications such as body area communication, vehicular communication and device-to-device communication etc., which require a high data rate with high reliability. In most of these future generation applications, the data is transferred from source to destination through a wireless link. The signal experience many phenomena like reflection, refraction, diffraction and scattering in its path traveling from transmitter to receiver [1]. These phenomena result in the fluctuations in the received signal strength which eventually degrades the quality of the signal and is termed as fading. Depending upon the scale of fading, it is classified as small scale (multipath) fading and large scale (shadowing) fading. The multipath fading results in the rapid fluctuation in the signal strength and shadowing give rise to the slow variation in the signal strength at the receiver. Therefore, the quality of service offered by the wireless communication system, in turn depends on the truthful assessment of the signal at the receiver [1]. These resultant phenomena may deteriorate the quality of radio links; reduce the overall reliability of the signal and the performance of the communication system. Apart from these deleterious effects, the general research of the communication system broadly covering the aspects of channel modeling and other signal improvement techniques. The prime-facie of any wireless communication system development is the accurate channel characterization. It helps us to acquire knowledge about the fading behavior of the channel, not only to analyze the performance but also to improve system design. Since the fading phenomenon observed in the wireless system is random, thus the effect can be modeled in terms of fundamental statistics, such as the density function, the cumulative distribution function (CDF), and the moment generating function (MGF). Among the aforementioned, the most commonly used fundamental statistics is the probability density function (PDF). The statistic can easily be used for analyzing the system with diversity and in error correction coding [1]. On the other hand, the remaining statistics also play a vital role in assessing the communication system performance metrics, namely, the outage probability (OP), the average output signal-to-noise-ratio (SNR), the channel capacity (CC) and the error probability etc.

The CC along with other performance parameters like the OP, and the error probability etc. is an essential parameter of concern in the designing of a communication system, which helps system engineers to provide quality of services (QoS) within the limited available spectrum to a large number of users. Also to address the explosive traffic demands, the capacity analysis over the fading channel is increasingly becoming a prime concern in the designing of the wireless communication system. The CC is an extremely important quantity since it allows the transmission of the data through the channel with an arbitrarily small probability of error.

The growing demand of high data rate in the future generation networks make it important to determine the capacity limits of the underlying channels. With the exponential rise in the services of the next-generation networks, it becomes extremely important to analyze the effect of several channel perturbations on the transmission rate. Capacity analysis based on Shannon's formula is investigated in the work of literature [2], [3]. CC is often expressed as a function of the average transmitted power. This function is obtained by optimizing the mutual information over all source distribution which have their average second moment, is either equal to the given power or upper bounded by this power. CC parameter governs how fast data can be transmitted through the channel; which eventually depends upon the bandwidth of the channel as well as the quality of the channel. Data transmitted over the wireless channel suffers from random fluctuations due to channel impairments, which tend to degrade the transmission capability of the channel. The CC analysis over various wireless channels under different propagation conditions has been analyzed in [4], [5], [6], [7], [8], [9]. Similarly, the secrecy capacity over Beaulieu-Xie fading channel has been well explored in [10]. Several works in literature are available that explores and review the implementation of CC analysis in different scenarios. For instance, a review of the multipath fading channels is presented in [11] where the performance comparison of the different models are presented. A conceptual explanation of the effective rate and its application to several delay-sensitive networks is comprehensively reviewed in [12]. Furthermore, a survey over the secrecy parameters of the wireless fading

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532961

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

channel is well explored in the literature [13], [14], [15]. In addition to this, a survey over the new era technologies and massive access design techniques to accommodate the surge in the capacity demands of the 5G and beyond networks is comprehensively provided in [16] and [17]. However, after a review of several literatures, to the best of the author's knowledge, it is found that there exists no work which comprehensively includes all the aspects of the CC and its modeling over the fading channels. Motivated by this gap, in this paper, we bring the review of the analysis of CC over fading channels as well as its implementation in various 5G and beyond wireless networks.

II. Theory of Transmission rate

Shannon derived the CC formula and showed that the reliable data transmission is possible if the transmitting rate does not exceed the CC. The reliable data transmission corresponds to achieve the arbitrarily small error probability. Shannon capacity with regards to wireless channels of interest is obscure and depends on the channel condition as well as on the channel information at transmitter and/or receiver. In wireless communication, having the information about the channel properties of the communication link is termed as channel state information (CSI). The information about how the signal propagates between transmitter and receiver and representation of various effects of fading can be estimated if CSI is known. So, it can be said that CSI makes it easier to design the system to adapt the current channel conditions and to provide an error-free transmission rate. Regarding this knowledge, three different scenarios, are considered to estimate the CC of the wireless channel.

A. Channel distribution information (CDI)

Under this scenario, the distribution of the channel gain is known to the transmitter as well as to the receiver. A series of contributions have been made in the literature on the study of capacity for the system having only CDI.

B. Receiver Channel State Information

When the channel state information is only known to the receiver and both transmitter and receiver are familiar with the distribution of the channel gain, there exist two definitions of the capacity that are relevant to the system design. The Ergodic capacity and capacity with outage come in this category.

C. CSI at transmitter and receiver

In case that both the transmitter and receiver have CSI, at that point the transmitter can adjust its transmission methodology comparative with this CSI, so a new dimension of freedom in terms of power control can be applied to calculate the Shannon capacity. Optimal power and rate adaption (OPRA), constant transmitting power also known as optimum rate adaptation (ORA), channel inversion with fixed-rate (CIFR) and truncated channel inversion with fixed-rate (TIFR) CC come under this category. A summary of the classification of CC works based on the CSI is provided in Table 1.

III. Transmission rate over Fading channels

A signal propagating through the wireless channel is a complicated phenomenon that is characterized by various effects such as multipath and shadowing which tend to degrade the signal. To model these effects, various mathematical distributions are explored in the literature. In the actual wireless propagation scenario, the signals are jointly affected by both phenomena; such fading channels are called composite fading channels. Depending upon these impairments, wireless fading channels are classified in the following category.

A. Multipath fading

It occurs when the transmitted signal divides or takes multiple paths due to the reflection, refraction, and scattering phenomena, resulting in a faded signal at the receiver. To model, these various fading models such as Rayleigh, Rician, Nakagami-*m*, Hyot and generalized fading models are widely explored in the literature. In [29], the performance analysis of the non-identical generalized Gamma fading channel is derived. This approach provides a simpler solution for the evaluation of CC in wireless communication system.

B. Shadowing model

The transmitted signal often suffers from shadowing due to the terrain, building structures, and trees, which in turn degrades the signal quality. To model this effect various distributions such as Lognormal, Gamma, inverse Gamma and inverse Gaussian are used. In [30], authors have presented an asymptotic analysis of the ergodic capacity for the Lognormal shadowed channel. The inverse Gaussian channel is studied in [31]. The benefits of the inverse gamma over Lognormal and gamma shadowing are discussed in [32] and its CC analysis is comprehensively studied. In [33], the analysis of CC over the lognormal channel is presented and the expression of CC is represented in a simpler form.

C. Composite Fading model

Degradation of the signal due to the simultaneous occurrence of multipath and shadowing effects is called as composite fading. The CC analysis over composite Nakagami-*m*/Gamma and Nakagami-*m*/inverse Gaussian channel is presented in [34] and [35]. The expressions of OP are also derived for both the cases in [34] and [35]. Using the incomplete generalized MGF, closed-form expressions for the OP of the maximum ratio combining (MRC) scheme are obtained over the κ-μ shadowed fading channels in [36]. Similarly, for α-κ-μ shadowed channel ACC is analyzed in [37], using the mixture gamma distribution.

Table 2, summarizes the capacity analysis review over different fading channels presented in the literature.

IV. Diversity Combining works on Transmission rate

In the communication system, diversity refers explicitly to a technique to improve the reliability of the data using two or more channels having different characteristics. By transmitting the signal over more than one path, the probability that each signal will receive the same deep fade is reduced. Various diversity combining techniques are popular in literature.

A. Selection Combining

Selection combining (SC) is based on the principle of selecting the branch having the maximum SNR among all the available SNR at the different branches at the receiver [48]. In this technique using the switch logic, the receiver monitors the receiver SNR at the input. The branch with the highest instantaneous SNR is connected to the output demodulator. SC performs detection, based on only one signal having the highest SNR. Therefore, it is the scheme that offers the least complexity. However, by selecting only one signal among the received branches, it does not fully exploit the diversity

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

Table 1. CC based on CSI condition

Reference	Year	CSI condition	Main Contribution
[18]	2001	No CSI	Proved that the capacity accomplishing distribution for independent and identically distributed Rayleigh fading channel is discrete with a limited number of points.
[19]	2016	No CSI	Presented the investigations over the block fading Rayleigh channel. Concluded that the pilot-based demodulations are imperfect and channel estimation is certainly performed by a capacity-optimal decoder
[20]	1999	No CSI	Analyzed the capacity of block fading multiple-input-multiple-output (MIMO) channel.
[21]	2019	No CSI	Used the first-order Gaussian Markov process for CC analysis for the system. Concluded that the channel rate information is a non-increasing function of the coherence channel co-efficient.
[22], [23]	1999, 2003	CSI known to receiver and transmitter	ORA, OPRA and CIFR analysis for Rayleigh and Nakagami-*q* fading channel is performed.
[24]	1997	CSI known to receiver only	For various fading channels expression for ergodic capacity is derived
[25], [26]	2005, 2007	CSI known to receiver only	Expression for the ergodic capacity over α-μ, κ- μ and η- μ fading channel is derived
[27], [28]	2010,1997	CSI known to receiver only	OP of Hoyt under Rayleigh interference and Lognormal shadowed Rician fading channel is studied

Table 2. Expression representation of CC over various fading Channels

Reference	Representation of Expression	Fading Channel	Parameter Analyzed	Comments
[26]	Meijer-G Function	η-μ and κ-μ	ACC	The expressions are presented in terms of the Meijer-G function. This function can be easily implemented in available software which makes the analysis simpler.
[25]		Nakagami-*m*, Rician , Weibull	ACC	
[38]		α-μ	OPRA, ORA	
[39]		α- η- μ and α – κ – μ	ACC	
[40]		Weibull/Lognormal	ACC	
[6]		Hoyt/Gamma	ORA	
[23]	Infinite series summation	Nakagami-*q*	OPRA, ORA, CIFR, and TIFR	The expressions are presented in the form of infinite series summation So, the truncation error need to be calculated to get the desired accuracy.
[33]		Lognormal	Ergodic CC	
[41]		Nakagami-*n*	OPRA, ORA, CIFR, and TIFR	
[42]		η-μ/inverse Gamma	Ergodic CC	
[43]	Bivariate Fox-H function	α -η-μ	OPRA, ORA, CIFR, and TIFR	It provides more precise analysis but at the cost of implementation complexity.
[44]	Fox-H function	α -η-κ-μ	ACC	The expressions are simple and easy to implement.
[45]	Hypergeometric function	η−μ/Gamma	OPRA, ORA	This function provides an approximation of the analysis and can be easily implemented in the available software packages.
[46]		Generalized-K	OPRA, ORA, CIFR, and TIFR	
[47]		κ − μ/inverse Gamma and η−μ/inverse Gamma	ORA	

B. Maximal ratio Combining

In MRC technique each branch signal is multiplied with a gain weight factor which is proportional to the SNR of the channel [49]. The channel having a high SNR will pose high weight factor. This is considered as most complex scheme as the SNR of all the branches is optimally combined at the receiver. Here all the signals are weighted according to their individual SNR and then summed.

C. Equal gain Combining

Equal gain combining (EGC) is the simplified version of the MRC technique. In this, the overall SNR at the output of the combiner is the addition of all branches SNR and the gain of all the branches are set to be one. Thus, it produces SNR at the combined output lower than that of the maximal ratio combiner. The SNR improvement of this technique is better than SC but less than MRC so it can also be considered as a compromise between SC technique and MRC technique.

The performance analysis of fading channels with diversity combining is well explored in the literature. Table 3 summarizes the CC work over various diversity combining techniques.

V. TRANSMISSION RATE WORKS OVER VARIOUS NETWORKS

The performance of the different wireless networks can be analyzed by estimating the CC, considering the various wireless design and network architectures. CC analysis over the cognitive radio network (CRN), cooperative network,

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

wireless sensor network (WSN) and the vehicular ad-hoc network (VANET) etc. have been extensively analyzed in the literature. In this section, we have reviewed the performance of CC over these networks.

Table 3. CC analysis based on diversity combining techniques

Reference	Year	Fading Channel	Diversity Combining
[50]	2020	B-X	MRC
[6]	2017	Hoyt/Gamma	MRC
[51]	2018	Hoyt/Lognormal	MRC
[52]	2019	Generic and composite	MRC
[53]	2008	Co-related Weibull	SC
[32]	2019	Inverse gamma shadowed	SC
[36]	2016	κ-μ Shadowed	SC
[54]	2007	Generalized-*K*	MRC, EGC, SC
[55]	2015	η –μ	MRC

A. Vehicular ad-hoc network

The concept of the intelligent transportation system in smart cities has been realized with the evolution of VANET [56]. For vehicle-to-vehicle (V2V) network performance of the communication channel is analyzed by the authors in [57]. In [58], authors have provided ergodic capacity expression for V2V network established at 28 GHz and then performed a diversity analysis using this expression experiencing α-η-κ-μ fading. In [59], OP is analyzed for a BS to vehicle communication at 60 GHz considering Rician fading channel. The analysis concludes that the reflected signal from the road infrastructure does put any impact on the communication. enable frequent handovers. In [60], authors have analyzed the transmission capacity over the VANET model undergoing Rayleigh fading. Nakagami-*m* fading model is an excellent model to represent the V2V communication environment. So, in [61], authors have analyzed the performance of full-duplex V2V communication under this environment. The expressions for the OP under the different scenarios are derived to analyze the performance.

B. Wireless Sensor network

Due to the latest communication services such as internet-of-thing (IoT), the implementation of wireless sensor networks is rapidly increasing, both in urban as well as rural areas. Several sensors that can collect information about the parameters of the field are deployed there. These sensor nodes wirelessly communicate and send the information of the collected data through the gateway on the internet. In [62], CC evaluation of the interference-limited WSN in the dense urban scenario is performed. In [63], authors have proposed an adaptive MIMO single value decomposition technique to enhance the CC and reduce the transmission energy of the WSN in case of imperfect CSI. Also, CC performance analysis for encoding methods over the Rayleigh fading channel is studied. In [64] and [65] authors have used space-time-block-codes (STBC) coding to improve the channel performance at the same transmission power of the WSN. Both utilize MIMO technique however, later uses only a maximum of two antenna elements in the STBC technique. Performance of WSN under the Rayleigh and Rician fading channels is studied in [66]. In [67], an analytical framework is presented to analyze the performance of WSN in a complex railway scenario. The analytical results are evaluated over the composite environment of Nakagami-*m* fading.

C. Cooperative network

Due to the ability of providing reliable data transmission, the communication over the cooperative network is gaining popularity among system designers. Authors in [68], have analyzed the capacity of Rayleigh faded cooperative network under various adaptive transmission schemes. Along with this, MRC and SC diversity combining techniques is also investigated. In [69] and [70] OP and bit error rate (BER) respectively, for decode and forward relaying is investigated over Rayleigh fading channel. For the same relaying technique capacity and outage performance of the underground tunnel system are studied in [71]. In [72], the performance of the amplify and forward (AF) technique implemented in a multi-relay system connecting multiple users via several relays to the destination is studied. The throughput performance of the full multi-packet reception model is explored in [73] over a slotted-time relay-assisted cooperative network. Some more existing work on the performance of the multi-hop cooperative network can be found in the literature [74], [75].

D. Cognitive Radio Networks

CRN is a promising technology for efficient utilization of the available spectrum and increases network capacity. It intelligently utilizes available side information about channel conditions, encoding strategies or transmitted data sequences of primary users with which it shares the spectrum. In recent years, a huge amount of work has been done to analyze the performance under the three paradigms of CRN, underlay, overlay, and interweave [76]. In [77], authors have analyzed the capacity of CRN undergoing multipath effect by estimating CC and BER. For the underlay CRN, closed-form expressions for ergodic capacity and outage capacity of SU are determined in [78]. Here, authors have considered a composite fading channel and shown that on increasing shadowing the outage performance degrades. Authors have analyzed the performance of the energy detection method for spectrum sensing model in CRN over different fading channels [79], [80], [81], [82], [83], [84], [85].

E. Unmanned Aerial Communication

UAV-assisted wireless communication has stroked a great interest by both the industrial as well as academic community due to its diverse application in entertainment, wildlife monitoring, military and surveillance fields. In [86], authors have derived the closed-form OP expression in air-to-ground and ground-to-air transmission links. In [87], authors have focused on the analysis of channel characteristics of the air-to-air UAV communication considering Ricean model. Using three dimensional geometrically based single bounce cylinder model channel for the UAV MIMO system, the influence of the UAV altitude parameters over the CC of the system is analyzed in [88]. The improvement in the performance of the high-altitude long-endurance UAV using the MIMO cooperative relay is analyzed in [89] considering Rayleigh fading channel.

F. Hybrid Satellite Terrestrial Links

A hybrid satellite-terrestrial network takes the advantage of both the terrestrial and satellite communication network to provide an effective solution for the realization of the foreseen heterogeneous global network system [90]. In [91], authors have derived an exact expression of ergodic CC for hybrid satellite-terrestrial single frequency network consisting of satellite as the source node, mobile as destination node and several gap filters as relay nodes on the ground. The performance of the various diversity combining techniques over the hybrid satellite-terrestrial link is studied in [92]. In this with the knowledge of CSI at the relay-destination link expression for the OP is derived over SC and MRC combining

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

techniques. In this series some significant work can be found in the literature [93], [94], [95].

G. Hybrid free-space optics (FSO) and Radio-frequency (RF) links

The hybrid FSO and RF links are an alternative to meet the growing demands of high data rates as it incorporates the properties of FSO as well as RF links. The performance of such communication link is analyzed in [96] considering FSO undergoing gamma-gamma fading and RF experiencing Rayleigh fading. Using three basic performance metrics OP, bit error rate (BER) and CC, the performance of this link is analyzed. In [97], performance is evaluated using SC diversity for hybrid systems. In [98], hybrid FSO/RF is tested by authors in a practical environment using adaptive combining. The receiver retrieved the original data out of two links by using an MRC technique. A dual-hop relay-assisted hybrid parallel FSO/RF system is investigated in [99], to analyze the performance closed-form expressions for BER and the OP is derived. In Gamma-Gamma atmospheric turbulence FSO link is considered with the impact of pointing error and the RF link is considered in Rayleigh fading.

H. Wireless body-area-network (WBAN)

This network typically consists of several body parameter monitoring sensors and actuators that are put at the different explicit locations inside and on the body and are connected through the wireless channel. The performance of the WBAN is strongly influenced by human movements and needs to be analyzed for the effective deployment of the on-body sensors. Numerical analysis of the OP, as well as asymptotic analysis, is provided to obtain the behavior of the network. With the recent advances IoT technologies have been integrated with the WBANs. In [100], the OP for three different protocols with energy harvesting is derived and formulated for the IoT-based health care system.

I. Other networks

There are many other emerging applications where the CC is considered as an important design parameter and is well studied. For example, the intelligent reflecting surface (IRS) provides the promising solution to enhance the spectral efficiency by smartly tuning the reflection of the signal through its passive reflecting elements. In this regards, IRS-assisted relay networks are being studied to improve the system performances. Similarly, the MIMO-OFDM network is being studied over various fading channels in terms of its throughput performance.

VI. Future scope and directions

With the increasing QoS in the wireless communication, the complexity of the network is also increasing, hence the CC analysis is becoming complex [101], [102]. After a thorough review of the existing literature, it is observed that there are potential future scopes that can be worked upon by the researchers, which are highlighted below.

A. CC analysis of the system without CSI

There is a lot of capacity analysis theory available in the literature with known CSI. In [20], using the information-theoretic considerations, performance criteria for the unitary space-time modulation are analyzed for the block fading Rayleigh channel having no channel coefficient information. Similarly, for multi-antenna system, asymptotic capacity analysis over the non-coherent block fading channel is studied in [103]. This analysis assumes no CSI at the transmitter and receiver. Some recent studies have been reported which considers the analysis of scenarios where obtaining CSI is a cumbersome process. In [21], the performance analysis of fading channels without CSI is presented. As wireless MIMO channel having random and deterministic nature, the spectral efficiency without CSI is derived in [104]. Over the block fading Rayleigh channel without CSI, CC performance is analyzed in [19]. In [105], the authors analyzed the cooperative non-orthogonal multiple access (NOMA) network over Nakagami-*m* fading channel having imperfect CSI. For the scenarios where the fading co-efficient vary rapidly and the transmitter and receiver cannot estimate the coefficients accurately, the performance of various fading channel can be explored by the researchers.

B. Channel capacity maximizing techniques

To improve the CC, various diversity combining techniques are well explored in the literatures. But most of the analysis considered MRC, SC and EGC whereas generalized selection combining (GSC) is less explored for the new fading channels. In [106], GSC performance analysis over digital modulation technique for the MIMO system is explored. In [107], the performance of GSC is analyzed over the correlated Nakagami-*m* fading channel. The application of GSC to maximize the performance of energy detector in cognitive radio network is presented in [108]. In [109], the authors have analyzed the performance of the GSC with Gaussian distributed weighting errors for physical-layer security of a wireless communication system by deriving the exact closed-form expressions for the secrecy performance. Similarly, in [110] the secrecy analysis of cognitive network implementing GSC is presented. Recently in [111], the performance of the NOMA networks with GSC is analyzed. As discussed earlier, the optimal resource allocation also leads to the performance enhancement and the inherent difficulty of resource allocation problems makes the use of machine learning tools appealing to solve the optimization problem. For instance, in [112], the deep neural networks and the theory of Lagrangian duality is used to solve the non-convex optimization problem, by formulating a finite-dimensional, unconstrained optimization problem in the dual domain. For a V2V system a decentralized resource allocation mechanism, based on the deep reinforcement learning is proposed in [113]. Although the concept is new and still need to be investigation over various networks. However, there is extensive scope in study of the diversity combining techniques, specifically, GSC over the various new existing fading channels as well as implementation of deep learning in optimizing the resource allocation.

C. CC analysis to guarantee QoS

It is observed from the trend that there will be continuing wireless data explosion and requiring high data rates with ultra-low latency [114], [115]. Evidently, it turns out to be more pivotal to investigate the ultimate capacity that a wireless channel can have under stringent delay constraints and to examine what delay limit might be accomplished in a particular wireless channel situation [116], [117]. The effective capacity (EC) model is widely explored in the literature which provides the capacity analysis incorporating QoS guarantee for the low-latency applications [117], [118], [119]. To improve the delay sensitive throughput in the vehicular network, the EC theory is applied in [120]. The effective-throughput maximization problem to optimize the transmission time and packet error rate for the IoT devices under various practical constraints is handled in [121]. Further, this concept can be explored to analyze the performance of several other heterogeneous networks which supports diverse delay-sensitive networks with different delay requirements.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

D. Energy-harvesting based CC Analysis

Conservation of energy in the future wireless system is a prime issue, as devices used have battery size as well as power constraint limitation. With the introduction of the energy harvesting battery modules, the performance of the wireless channel could be enhanced to a great extent. Limited work has been done in the literature to investigate the effect of energy harvesting on the achievable capacity. In energy harvesting channels [122], the transmitter is powered by energy, harvested from a random exogenous energy source. In [123], the authors have investigated the performance of a wireless power transfer based on the two-hop cooperative relaying systems in indoor channels experiencing log-normal fading. Authors in [124] have assumed the harvested energy to be Gamma distributed and have performed analysis by deriving the expression of outage probability and CC. Similarly, performance analysis of energy harvesting for V2V network implementing NOMA is provided in [125]. In [126], the performance of the MIMO system where transmitter harvesting energy from the various source in its environment is studied. The energy harvesting technique is a promising solution for the battery limited devices which is widely used in the real-world applications. Further investigations are required to utilize this technique to make the wireless system more efficient and robust.

VII. CONCLUSIONS

This paper presents a survey of the CC, which is the most important metrics to analyze the performance of a wireless system. The theory of Shannon's CC is discussed, and then depending upon the CSI CC classification is done. To elaborate on the concept, a review of the CC works over the various fading channels is presented. Analytical expressions derived for the CC over the fading channels are reviewed based on their complexity. Further, the diversity combining techniques used to maximize the CC is thoroughly reviewed. Furthermore, CC evaluation of various wireless networks such as cognitive radio networks, cooperative networks, body area networks hybrid networks and device-to-device networks etc. are added to show the implication of the concept. In the end, future research directions of the survey are provided, which can be further addressed by the researchers.

References

[1] M. K. Simon and M. S. Alouini, Digital Communications over Fading Channels: A Unified Approach to Performance Analysis, 2nd ed. NY: John Wiley & Sons, 2004.

[2] O. Dabeer and U. Madhow J. Singh, "Capacity of the discrete-time AWGN channel under output quantization," in IEEE International Symposium on Information Theory, Toronto, ON, 2008.

[3] C. Ling and Jean-C. Belfiore, "Achieving AWGN CC With Lattice Gaussian Coding," arXiv:1302.5906.

[4] M. S. Alouini A. Goldsmith, "Capacity of Nakagami multipath fading channels," in IEEE *47th Vehicular Technology Conference. Technology in Motion*, vol.1, pp. 358-362, 1997

[5] W. C. Y. Lee, "Estimate of CC in Rayleigh fading environment," IEEE Trans. Veh. Technol, vol. 39, pp. 187-190, Aug. 1990.

[6] S. Kumar, S. K. Soni, and P. Jain, "Micro-Diversity Analysis of Error Probability and CC over Hoyt/Gamma Fading," Radioengineering, vol. 26 ,no. 4 ,pp. 1096-1103, 2018.

[7] S.K. Soni, P. Jain S.Kumar, "Performance analysis of Hoyt-lognormal composite fading channel," in International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai,India, 2017, pp. 2503-2507.

[8] P. M. Shankar, Fading and shadowing in wireless systems, 1st ed. NY: Springer, 2012.

[9] P.S. Chauhan, S. Kumar, S.K. Soni, V.K. Upaddhyay, D. Pant, Average channel capacity over mixture Gamma distribution. International conference on Electrical & Electronics Engineering (ICE3-2020), 420-424.

[10] P.S. Chauhan, S. Kumar, S.K. Soni, On the Physical Layer Security over Beaulieu-Xie Fading Channel, *International Journal of Electronics and Communications*, 2020, 113(2020), 152940

[11] R.H. Chile and S.R. Sawant M.S. Chavan, "Multipath Fading Channel Modeling and Performance Comparison of Wireless Channel Models," International Journal of Electronics and Communication Engineering., vol. 4, no. 2, pp. 189-203, 2011.

[12] L. Musavian and M. H. Rehmani M. Amjad, "Effective Capacity in Wireless Networks: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3007-3038, Fourthquarter 2019,

[13] P. Yadav, S. Kumar, R. Kumar, A Comprehensive Survey of Physical-layer Security over Fading Channels: Classifications, Applications, and Challenges. Transaction on Emerging Telecommunication Technology, 2021, e4270.

[14] A. Khisti, C. Xiao, G. Caire, K. Wong and X. Gao Y. Wu, "A Survey of Physical Layer Security Techniques for 5G Wireless Networks and Challenges Ahead," IEEE Journal on Selected Areas in Communications, vol. 36, no. 4, pp. 679-695, April 2018.

[15] H. H. Chen, and L. Wang Y. Liu, "Physical layer security for next generation wireless networks: Theories, technologies, and challenges," IEEE Commun. Surveys Tuts., vol. 19, no. 1, pp. 347–376, 2017.

[16] A. Kak and S. Nie I. F. Akyildiz, "6G and Beyond: The Future of Wireless Communications Systems," IEEE Access, vol. 8, pp. 133995-134030 , 2020.

[17] D. W. Kwan Ng, Wei Yu,Erik G. Larsson, N. Al-Dhahir, and R. Schober X. Chen, "Massive Access for 5G and Beyond ," arXiv:2002.03491, 2020.

[18] M. D. Trott, and S. Shamai I. C. Abou-Faycal, "The capacity of discrete-time memoryless Rayleigh fading channels," IEEE Trans. Inform. Theory, pp. 1290-1301, 2001.

[19] M. Chowdhury and A. Goldsmith, "Capacity of block Rayleigh fading channels without CSI," IEEE Int. Symp. Info. Theory, pp. 1884-1888, 2016.

[20] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," IEEE Trans. Inf. Theory, vol. 45, no. 1, pp. 139–157, Jan. 1999., vol. 45, no. 1, pp. 139-157, 1999.

[21] X. Yang, "Capacity of Fading Channels without Channel Side Information," https://arxiv.org/pdf/1903.12360.pdf.

[22] M. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," IEEE Transactions on Vehicular Technology, vol. 48, no. 4, pp. 1165-1181, 1999.

[23] J. Cheng T. Berger, "Capacity of Nakagami-q (Hoyt) fading channels with channel side information," International Conference on Communication Technology Proceedings ICCT, vol. 2, pp. 1915-1918, 2003.

[24] A. Goldsmith and P. Varaiya, "Capacity of Fading Channels with Channel Side Information," IEEE Transactions on Information Theory, vol. 43, no. 6, pp. 1986-1992, 1997.

[25] G. S. Tombras and G. K. Karagiannidis N. C. Sagias, "New results for the Shannon CC in generalized fading channels," IEEE Communications Letters, vol. 9, no. 2, pp. 97-99, 2005.

[26] D. B. da Costa and M. D. Yacoub, "Average CC for Generalized Fading Scenarios," IEEE Communications Letters, vol. 11, no. 12, pp. 949-951, 2007.

[27] J. F. Paris and D. Morales-Jimenez, "OP analysis for Nakagami-q (Hoyt) fading channels under rayleigh interference," IEEE Transactions on Wireless Communications, vol. 9, no. 4, pp. 122-1276, 2010.

[28] Chin Choy Chai and X. Dong Tjeng Thiang Tjhung, "OP for lognormal-shadowed Rician channels," IEEE Transactions on Vehicular Technology, vol. 46, no. 2, pp. 400-407, 1997.

[29] P.S. Chauhan, V. Rana, S. Kumar, S.K. Soni, D. Pant, "Performance Analysis of Wireless Communication system over Non-identical Cascaded Generalised Gamma Fading Channels" *International Journal of Communication Systems*, 2019, vol. 32, no.13, pp. e4004

[30] I. S. Ansari and M. Alouini, "Asymptotic Ergodic Capacity Analysis of Composite Lognormal Shadowed Channels," in IEEE 81st Vehicular Technology Conference (VTC Spring) , Glasgow, 2015.

[31] A. Laourine, S. Affes and A. Stephenne I. Trigui, "The Inverse Gaussian Distribution in Wireless Channels: Second-Order Statistics and CC," IEEE Transactions on Communications, vol. 60, no. 11, pp. 3167-3173, 2012.

[32] P.S. Chauhan and S.K. Soni D. Pant, "Error probability and CC analysis of wireless system over inverse gamma shadowed fading channel with selection diversity," International journal of communication system , vol. 32, no. 16, pp. e4083, 2019.

[33] A. Stephenne and S. Affes A. Laourine, "On the capacity of log-normal fading channels," IEEE Transactions on Communications, vol. 57, no. 6, pp. 1603-1607, 2009.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

[34] I. M. Kostic, "Analytical approach to performance analysis for channel subject to shadowing and fading," IEE Proceedings - Communications, vol. 152, no. 6, pp. 821-827, 2005.

[35] M. S. Alouini, S. Affes and A. Stephenne A. Laourine, "On the Performance Analysis of Composite Multipath/Shadowing Channels Using the G-Distribution," in IEEE International Conference on Communications, Beijing, 2008.

[36] M. Shu, Y. Wang and C. Zhang C. Chen, "OP analysis for MRC in *k-μ* shadowed fading channels with co-channel interference," in IEEE International Conference on Information and Automation (ICIA), Ningbo, 2016.

[37] H. Al-Hmood, "A mixture gamma distribution based performance analysis of switch and stay combining scheme over a-k-mu shadowed fading," in Proc. IEEE Annual Conf. on New Trends in Info. & Commun, 2017.

[38] Amer M. Magableh and Mustafa M. Matalgah, "Closed-form expressions for the average CC of the a-μ Fading model under different adaptive transmission protocols," Wireless communications and Mobile Computing, vol. 15, no. 1, pp. 1-9, 2015.

[39] F. Yilmaz and M. Alouini, "A new simple model for composite fading channels: Second order statistics and CC," 7th International Symposium on Wireless Communication Systems, York, pp. 676-680, 2010.

[40] S. K. Soni, R.S. Raw and S.Kumar R. Singh, "A New Approximate Closed-Form Distribution and Performance Analysis of a Composite Weibull/Log-Normal Fading Channel," Wireless Pers Commun, vol. 92, pp. 883-900, 2017.

[41] P. Das, R. Ahmed and R. Subadar N. Barman, "CC of Adaptive Transmission Techniques over Rice (Nakagami-n) Fading Channels," Advanced Research in Electrical and Electronic Engineering , vol. 2, no. 13, pp. 18-21, 2015.

[42] P. C. Sofotasios et al., "Capacity analysis under generalized composite fading conditions," in Int. Conference on Advanced Communication Technologies and Networking (CommNet), Marrakech, 2018.

[43] A. M. Magableh, S. Mater and O. S. Badarneh T. Aldalgamouni, "Capacity analysis of α-η-μ channels over different adaptive transmission protocols," IET Communications, vol. 11, no. 7, pp. 1114-1122, 2017.

[44] J. M. Moualeu, D. B. da Costa, F. J. Lopez-Martinez and R. A. A. d. Souza, "On the Performance of α-η-κ-μ Fading Channels," IEEE Communications Letters, vol. 23, no. 6, pp. 967-970, 2019.

[45] J. Zhang, M. Matthaiou, Z. Tan and H. Wang J. Zhang, "Performance Analysis of Digital Communication Systems Over Composite η-μ/gamma fading channels. *IEEE Transactions on Vehicular Technology*, vol. 61, no. 7, pp. 3114 – 3124, 2012.

[46] N. Y. Ermolova and V. A. Aalo G. P. Efthymoglou, "CC and average error rates in generalised-K fading channels," IET Communications, vol. 4, no. 11, pp. 1364-1372, 2010.

[47] P. S. Chauhan, S. K. Soni and S. Naithani D. Pant, "CC Analysis of Wireless System under ORA scheme over κ–μ/Inverse Gamma and η-μ Inverse Gamma Composite Fading Models," in International Conference on Electrical and Electronics Engineering (ICE3), Gorakhpur, 2020.

[48] Rappaport TS., Wireless Communication—Principle and Practice. : Upper saddle River, New Jersy,: Prantice Hall PTR, 1996.

[49] R. Singh, S.K. Soni, P.K. Verma, S. Kumar, Performance Analysis of MRC Combiner Output in Log Normal Shadowed Fading, *Proceeding of IEEE International Conference on Computing, Communication and Automation (ICCCA - 2015),* pp. 1116-1120.

[50] M. Kaur and R.K. Yadav, "Performance analysis of Beaulieu-Xie fading channel with MRC diversity reception," Transactions on Emerging Telecommunication Technologies, vol. 31, no. 7, pp. e3949, 2020.

[51] S. Kumar, S. K. Soni and P. Jain, "Performance of MRC receiver over Hoyt-lognormal Composite Fading Channel," Internation Journal of Electronics, vol. 105, no. 9, pp. 1433-1450, 2018.

[52] P.S. Chauhan, S.Kumar and S.K. Soni "New approximate expressions of average symbol error probability, probability of detection and AUC with MRC over generic and composite fading channels," AEU - International Journal of Electronics and Communications, vol. 99, pp. 119-129, 2019.

[53] D. Milovic , A. Mitic and M. Jakovljevic M. Stefanovic, "Performance analysis of system with selection combining over correlated Weibull fading channels in the presence of cochannel interference.," AEU - Int. J. Electron. Communication, vol. 62, no. 9, pp. 695-700, 2008.

[54] P. T. Mathiopoulos and S. A. Kotsopoulos P. S. Bithas, "Diversity reception over generalized-K (KG) fading channels," IEEE Transactions on Wireless Communications, vol. 6, no. 12, pp. 4238-4243, 2007.

[55] E. Salahat and M. Qasaimeh, "Unified Analytical Modeling of the Error Rates and the Ergodic CC η-μ Generalized Fading Channels with Integer μ and Maximal Ratio Combining Receiver," https://arxiv.org/abs/1511.03039.

[56] Z. Li, H. Zhang,S. Wang and R. Bie W. Liang, "Vehicular Ad Hoc Networks: Architectures, Research Issues, Methodologies, Challenges, and Trends," International Journal of Distributed Sensor Networks, vol. 11, no. 8, pp. 1-4, 2014.

[57] V. D. Khairnar and K. Kotecha, "Performance of Vehicle-to-Vehicle Communication using IEEE 802.11p in Vehicular Ad-hoc Network Environment," International Journal of Network Security & Its Applications (IJNSA), vol. 5, no. 2, pp. 143-170, 2013.

[58] A. Ali , I. Khan, and K. Rabie O. A. Saraereh, "Interference Analysis for Vehicle-to-Vehicle Communications at 28 GHz," Electronics, vol. 9, 2020

[59] R. Verdone, "OP analysis for short-range communication systems at 60 GHz in ATT urban environments," in IEEE Transactions on Vehicular Technology, vol. 46, no. 4, pp. 1027-1039,. 1997.

[60] X. He, W. Shi and T. Luo, "Transmission Capacity Analysis for Vehicular Ad Hoc Networks," in IEEE Access, vol. 6, pp. 30333-30341, 2018.

[61] G. Kaddoum,K. B. Fredj,E. Soujer and F. Gagnon K. Eshteiw, "Performance Analysis of Full-Duplex Vehicle Relay-Based Selection in Dense Multi-Lane Highways," IEEE Access, vol. 7, pp. 61581-61595, 2019.

[62] F. Novillo, J. Gomez,and D. Dik C. Valdivieso, "Performance Evaluation of CC in Wireless Sensor Networks using ISM band in Dense Urban Scenarios," in IEEE Ecuador Technical Chapters Meeting (ETCM), Ecuador, 2016.

[63] Y. Kirsal and S. Redif D. Hassan, "CC improvement for cooperative MIMO wireless sensor networks via adaptive MIMO-SVD," , 2016.

[64] E. Dutkiewicz and X. Huang M. Ahmad, "Performance analysis of MAC protocol for cooperative MIMO transmissions in WSN," in Proceedings of International Symposium on Personal, Indoor and Mobile Radio Communications, France, 2008.

[65] V. K. Sachan, S. A. Imam and M. T. Beg, "Performance analysis of STBC encoded cooperative MIMO system for wireless sensor networks," 2012 IEEE International Conference on Signal Processing, Computing and Control, Waknaghat Solan, 2012, pp. 1-6.

[66] D. Puccinelli and M. Haenggi, "Multipath fading in wireless sensor networks: measurements and interpretation," in Proceedings of the International Conference on Wireless Communications and Mobile Computing, IWCMC 2006, Canada, pp. 3-6, 2006.

[67] T. Shi and X. Lv, "Transmission performance analysis of wireless sensor networks under complex railway environment," 29th Chinese Control And Decision Conference (CCDC), Chongqing, 2017, pp. 2970-2947.

[68] M. Torabi and D. Haccoun, "Capacity Analysis of Opportunistic Relaying in Cooperative Systems with Outdated Channel Information," IEEE Communications Letters, vol. 14, no. 12, pp. 1137-1139, 2010.

[69] N. C. Beaulieu and J. Hu, "A closed-form expression for the OP of decode-and-forward relaying in dissimilar Rayleigh fading channels," EEE Communications Letters, vol. 10, no. 12, pp. 813-815, 2006.

[70] I. Lee and D. Kim, "BER analysis for decode-and-forward relaying in dissimilar Rayleigh fading channels," IEEE Communications Letters, vol. 11, no. 1, pp. 52-54, 2007.

[71] I. F. Akyildiz and G. P. Hancke Z. Sun, "Capacity and Outage Analysis of MIMO and Cooperative Communication Systems in Underground Tunnels," IEEE Transactions on Wireless Communication, vol. 10, no. 11, pp. 3793-3803, 2011.

[72] A. Alfitouri, K. M. Rabie, B. Adebisi and K. A. Hamdi W. Gheth, "Performance Analysis of Cooperative Diversity in Multi-user Environments," in 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), Manama, Bahrain, 2019.

[73] I. Dimitriou and N. Pappas, "Performance Analysis of a Cooperative Wireless Network with Adaptive Relays," https://arxiv.org/abs/1710.05748.

[74] T.R. Rasethuntsa, M. Kaur, S. Kumar, P.S. Chauhan, K. Singh, On the performance of DF based multi-hop system over α–κ–μ and α–κ–μ-Extreme Fading Channels. Digital Signal Processing, 2021, vol. 109, pp. 102909.

[75] T.R. Rasethuntsa, M. Kaur, S. Kumar, On the Outage probability and BER of a DF based multi-hop system over α–κ–μ and α–κ–μ-Extreme Fading Channels International Journal of Electronics and Communications, vol. 124, pp. 153324, 2020.

[76] H. Venkataraman and Gabriel-Miro Muntean, Cognitive Radio and its Application for Next Generation Cellular and Wireless Networks.: Springer Netherlands, 2012.

[77] M. Premkumar and M. P. Chitra, "Analysis of Cognitive Radio Capacity in Fading Channels ," International Journal of Wireless Communications and Mobile Computing, vol. 6, no. 1, pp. 31-36, 2018.

[78] M. H. Khoshafa and S. Al-Ahmadi, "On the Capacity of Underlay Cognitive Radio Networks Over Shadowed Multipath Fading Channels," Arabian Journal for Science and Engineering, vol. 42, pp. 5191–5199, 2017.

[79] S. Kumar, "Performance of ED based spectrum sensing over α-η-μ fading channel," Wireless Personal Communications, vol. 100, no. 4, pp. 1845-1857, 2018.

[80] P.S. Chauhan, S. Kumar, V.K. Upaddhyay, R. Mishra, B. Kumar, S.K. Soni, Performance Analysis of ED over Air-to-Ground and Ground-to-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

Ground Fading Channels: A Unified and Exact Solution, International Journal of Electronics and Communications, pp. 153839, 2021.

[81] T.R. Rasethuntsa, S. Kumar, An Integrated Performance Evaluation of ED-Based Spectrum Sensing Over α–κ–μ and α–κ–μ-Extreme Fading Channels, *Transaction on Emerging Telecommunication Technology*, 30(5), pp. e3569, 2019.

[82] S. Kumar, P.S. Chauhan, P. Raghuwanshi, M. Kaur, K. Singh, ED performance over α–η–μ/IG and α–κ–μ/IG generalized fading channels with diversity reception and cooperative sensing: A unified approach *International Journal of Electronics and Communications*, vol. 97, pp. 273-279, 2018.

[83] S. Kumar, M. Kaur, N.K. Singh, K. Singh, P.S. Chauhan, Energy Detection based Spectrum Sensing for Gamma Shadowed α–η–μ and α–κ–μ Fading Channels. *International Journal of Electronics and Communications*, vol. 93, pp. 26-31, 2018.

[84] S. Kumar, Energy Detection in Hoyt-Gamma Fading Channel with Micro-Diversity Reception, *Wireless Personal Communications*, vol. 101, no. 2, p. 723-734, 2018.

[85] S. Kumar, P.K. Verma, M. Kaur, S.K. Soni, P. Jain, "Performance Evaluation of Energy Detection over Hoyt/gamma Channel with MRC Reception", *Journal of Electromagnetic Waves and Applications,* vol. 32 no. 16 pp. 2157-2166, 2018.

[86] M. Kim and J. Lee, "OP of UAV Communications in the Presence of Interference," in 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, UAE, pp. 1-6, 2018.

[87] N. Goddemeier and C. Wietfeld, "Investigation of Air-to-Air Channel Characteristics and a UAV Specific Extension to the Rice Model," in 2015 IEEE Globecom Workshops, San Diego, CA, pp. 1-5, 2015.

[88] Z. Chen, Y. Hu X. Gao, "Analysis of Unmanned Aerial Vehicle MIMO CC Based on Aircraft Altitude ," WSEAS Transactions on Information Science and Applications, vol. 10, no. 2, pp. 58-67, 2013.

[89] M. M. Noori and M. Madani M. Tarihi, "Improving the Performance of Hale UAV Communication Link Through MIMO Cooperative Relay Strategy," Wireless Personal Communications, vol. 113 ,pp. 1051–1071, 2020.

[90] B. Evans and M.D. Renzo A. Guidotti, "Integrated satellite-terrestrial networks in future wireless systems," International Journal of Satellite Communication and Networking, vol. 37, no. 2, pp. 73-75, 2019.

[91] L. Xie, H. Chen and K. Wang Y. Zhao, "Ergodic CC analysis of the hybrid satellite-terrestrial single frequency network," in IEEE PIMRC, Hong Kong, 2015.

[92] Z.Lin,M.Lin,Q.Huang and J.Ouyung G.Cheng, "OP Analysis for Hybrid Satellite and Terrestrial Network with Different Combining Schemes," WiSATS , LNICST , vol. 281, pp. 488–496, 2019.

[93] J.Ouyang, M. Lin, et al. K.An, "Outage analysis of multi-antenna cognitive hybrid satellite-terrestrial relay networks with beamforming.," IEEE Commun. Letter, vol. 19, no. 7, pp. 1157-1160, 2015.

[94] D.D. Vergados, A. Michalas N.I. Miridakis, "Dual-hop communication over a satellite relay and shadowed Rician channels," IEEE Trans. Vehicular Technology, vol. 64, no. 9, pp. 4031–4040 , 2015.

[95] T.Liang and K.An L.Zhao, "Performance Optimization of Hybrid Satellite-Terrestrial Relay Network Based on CR-NOMA," Sensors (Basel). , vol. 20, no. 18, pp. 1-14, 2020.

[96] S. Sheikh Muhammad, H.E. Nistazakis and G.S. Tombras H. Khalid, "Performance Analysis of Hard-Switching Based Hybrid FSO/RF System over Turbulence Channels," Computation, vol. 7, no. 28, pp. 1-10, 2019.

[97] W.M.R. Shakir, "Performance evaluation of a selective combining scheme for the hybrid FSO/RF system," IEEE Photonics , vol. 10, pp. 1-10, 2018.

[98] T. Rakia, H.C. Yang, M.S. Alouini, and F. Gebali, "Outage analysisof practical FSO/RF hybrid system with adaptive combining," IEEE Commun. Lett. 2015, 19, , vol. 19, pp. 1366–1369, 2015.

[99] M.A. Amirabadi, "Performance Analysis of a Novel Hybrid FSO/RF CommunicationSystem,"arxiv.org/ftp/arxiv/papers/1802/1802.07160.pdf

[100] M. S. Mahmoud, A. Alkhayyatand and Q H Abbasi A. A. Thabit, "Energy harvesting Internet of Things health-based paradigm: Towards OP reduction through inter–wireless body area network cooperation," Internation Journal of Distributed Sensor Networks, vol. 15, no. 10, pp. 1-12, 2019.

[101] J. I. Z. Chen, "5G Technology and Advancements in connected living-Comprehensive Survey," Journal of Electronics and Informatics, vol. 1, no. 2, pp. 71-79, 2019.

[102] S. Smys, H. Wang, A. Basar "5G Network simulation in smart cities using neural network algorithmTechnology and Advancements in connected living-Comprehensive Survey," Journal of Artificial Intelligence and capsule networks, vol. 3, no. 1, pp. 43-52, 2021.

[103] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: a geometric approach to the noncoherent multiple-antenna channel," IEEE Transactions on Information Theory, vol. 48, no. 2, pp. 359-383, 2002.

[104] A. Shukla D. Singh, "Performance Analysis of Channel Capacity of MIMO System Without CSI," in Information and Communication Technology for Intelligent Systems. ICTIS 2020. Smart Innovation, Systems. Singapore: Springer, 2021, pp. 529-535.

[105] X. Yue, F. Liu, X. Gong, "Performance Analysis of Cooperative NOMA Networks with Imperfect CSI over Nakagami-*m* Fading Channels," Sensors. 2020;20:424.

[106] P. Maheswaran and M. D. Selvaraj A. Ananth, "Generalized Selection Combining for Dynamic SSK-BPSK Systems," in National Conference on Communications (NCC), Banglore, 2019.

[107] R. K. Mallik and M. Z. Win, "Analysis of hybrid selection/maximal-ratio combining in correlated Nakagami fading," IEEE Transactions on Communications, vol. 50, no. 8, pp. 1372-1383, 2002.

[108] V. Kumar, R. Gangopadhyay, and S. Debnath D. C. Kandpal, "Performance of an energy detector with generalized selection combining for spectrum sensing," Cognitive Radio Oriented Wireless Networks.Springer International Publishing, pp. 375-384, 2016.

[109] W. Hamouda and F. Takawira J. M. Moualeu, "Secrecy Performance of Generalized Selection Diversity Combining Scheme with Gaussian Errors," in IEEE 88th Vehicular Technology Conference USA, 2018.

[110] A. Singh S.Thakur, "Secrecy analysis of cognitive radio network with MS-GSC/MRC scheme," J Wireless Com Network, vol. 114, 2018.

[111] B. Cardiff and M. F. Flanagan V. Kumar, "Performance analysis of NOMA with generalized selection combining receivers," Electronics Letters, vol. 55, no. 25, pp. 1-10, 2019.

[112] C. Zhang, L. F. O. Chamon, D. D. Lee and A. Ribeiro M. Eisen, "Learning Optimal Resource Allocations in Wireless Systems," IEEE Transactions on Signal Processing, vol. 67, no. 10, pp. 2775-2790, 2019.

[113] H. Ye and G. Y. Li, "Deep Reinforcement Learning for Resource Allocation in V2V Communications," in IEEE International Conference on Communications (ICC), Kansas City, 2018.

[114] M. Kaur and R.K. Yadav, "Effective Capacity Analysis over Fisher-Snedecor $\mathcal{F}$ Fading Channels with MRC Reception," Wireless Personal Communications, 2021, https://doi.org/10.1007/s11277-021-08692-w.

[115] P.S. Chauhan, S. Kumar, V.K. Upaddhyay, S.K. Soni, Unified approach to effective capacity for generalised fading channels. Physical Communication, 2021, vol. 45, 101278.

[116] P.S. Chauhan, S. Kumar, S.K. Soni, Performance analysis of non-identical cascaded α–μ fading channels. Wireless Personal Communications, vol. 116, pp. 3553-3566, 2021.

[117] P. Yadav, R. Kumar, S. Kumar, Analysis of EC over Gamma shadowed α–η–μ fading channel.Internation conference on computers, Management and Mathematical Sciences, IOP conference series: Materials Science & Engineering (ICCM-2020), pp. 1-6, 2020.

[118] P. Yadav, R. Kumar, S. Kumar, Effective Capacity Analysis over α-κ-μ/Gamma Composite Fading Channel IEEE International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)-2020, pp. 587-592, 2020.

[119] P. Yadav, R. Kumar, S. Kumar, Effective capacity analysis over generalized lognormal shadowed composite fading channels. Internet Technology Letters, vol. 3, no. 5, pp. e171, 2020.

[120] X. Ma, Y. Li and Z. Yang Yi Huang, "Effective Capacity Maximization in beyond 5G Vehicular Networks: A Hybrid Deep Transfer Learning Method," Wireless Communications and Mobile Computing, vol. 2021, pp. 1-12, 2021.

[121] Y. C. Liang, X. Kang and R. Zhang J. Chen, "Effective-Throughput Maximization for Wireless-Powered IoT Networks with Short Packets," in 2018 IEEE Globecom Workshops Abu Dhabi, UAE, 2018.

[122] J. Gómez-Vilardebó, "Competitive Design of Energy Harvesting Communications in Wireless Fading Channels," IEEE/ACM Transactions on Networking, vol. 25, no. 6, pp. 3863-3872, 2017.

[123] A. Salem, E. Alsusa and M. Alouini K. M. Rabie, "Energy-harvesting in cooperative AF relaying networks over log-normal fading channels," in IEEE International Conference on Communications, 2016.

[124] I. T. Castro, L. Landesa, A. Serna, "Modeling the Energy Harvested by an RF Energy Harvesting System Using Gamma Processes", Mathematical Problems in Engineering, vol. 2019, pp. 1-12, 2019.

[125] T. M. Hoang, X. N. Pham, P. T. Tran, B. C. Nguyen, "Performance Analysis of Energy Harvesting-Based Full-Duplex Decode-and-Forward Vehicle-to-Vehicle Relay Networks with Nonorthogonal Multiple Access," vol.19 , pp. 1-11, 2019.

[126] I. Chihaoui and M. L. Ammari, "Save and Transmit Scheme for Energy Harvesting MIMO Systems with TAS/MRC," Journal of Communications Software and Systems, vol. 15, no. 1, pp. 52-58, , 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:30:58 UTC from IEEE Xplore. Restrictions apply.

Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Deep Learning Based Thermal Object Recognition under Different Illumination Conditions

Rohini Goel
CSE Department,
Maharishi Markandeshwar (Deemed to be University),
Mullana, Ambala, Haryana, India.
27rohinisharma@gmail.com

Avinash Sharma
CSE Department,
Maharishi Markandeshwar (Deemed to be University),
Mullana, Ambala, Haryana, India.

Rajiv Kapoor
Department of ECE,
Delhi Technological University
Delhi, India.

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532715

***Abstract*—Object recognition has recently a tremendous development because of the utilization of deep learning network. The recognition accuracy frequently experiences some sources of variation that can be found in images. Probably the most difficult variations are actuated by changing illumination conditions. Numerous researchers are targeting improving performance of object recognition frameworks towards various illumination conditions. The better model understanding and contributions of deep network in the object recognition motivate to utilize the deep networks for object recognition in different illumination conditions. This paper presents an accurate and proficient approach for recognizing the objects under varying illumination conditions. Hence, this proposed technique presents deep network-based system for illumination invariant object recognition in thermal image dataset. The thermal image database is utilized for the training and validation of deep network. In this work, Faster R-CNN is utilized to adequately recognize objects in thermal images under various illumination conditions. The results exhibit that the proposed work can significantly upgrade the recognition performance, outperforming other state-of-the-art strategies. This work can be an incredible assistance in driver assistance system and surveillance frameworks.**

Keywords*—Deep Neural Network; Faster Region Based Convolutional Neural Network (Faster R-CNN); Region Proposal Network (RPN); Segmentation; Thermal Images.

I. INTRODUCTION

For effective surveillance and enhance security, the recognition of objects with high accuracy in different visibility conditions like haze, snow, rain and cloudy environment is imperative. The capacity to recognize an object and warn about apprehensive circumstance is very crucial for the surveillance framework. Thermal cameras are significant for security and surveillance because of the utilization in climate conditions where conventional cameras can't be utilized or provide bad results, e.g. in the darkness and night or in the haze and rain. Accordingly, it is a important matter to develop a strategy for detecting objects with high accuracy under all conditions. Numerous lucrative machine learning algorithm have been established for tracking and identifying objects like the face of human [1], [2] or the human figure [3] in RGB pictures. The motive for recognition of object is to classify and imprint their accurate location in the image or video frame. These days, the better object recognition performance is accomplished by convolutional neural networks (CNN) dependent models. The fame of CNNs and deep learning started with the extraordinary accomplishment of AlexNet for object recognition work [4]. After that, numerous effective CNN models for object recognition have been established, i.e., SSD [5], R-FCN [6], Fast R-CNN [7], Mask R-CNN [8] and YOLO [9]–[11].

Because of the distinctions in thermal and visual image features, the motive is to investigate the performance of deep learning strategies for recognition of object under various illumination conditions with thermal imaging. The utilization of convolutional neural networks (CNN) is the research area trend in object recognition utilizing thermal imaging cameras. Thermal images are for the most part used to identify the presence of individuals in various illumination conditions. D. Ghose et al. [12] illustrated to augment the use of thermal images along with saliency maps, to perform as an attention mechanism for the detection of pedestrian. D. Heo et al. [13] additionally presented the use of saliency maps and deep learning for pedestrian recognition at night. M. Ivasic-Kos et al. [14][15] introduced that the training of the YOLO detector was performed using thermal image dataset for person recognition and by examining distinctive climate conditions independently, testing is done on other databases and including potentially confusing objects. A. Gomez et al. [16] illustrated the utilization of LWIR thermal images for people counting in public areas, using small CNNs.

Aside from the utilization of CNNs for recognition of human in thermal images, Q. Liu et al. [17] proposed the utilization of CNNs with thermal images for object tracking. N. Shahid et al. [18] introduced a strategy based on CNNs and background modeling with thermal images. For real time execution, modified running Gaussian average is used to implement the background modeling and the human classification based on CNN is performed only for the identified foreground objects. X. Wang et al. [19] utilized deep CNNs for human detection on frames extracted from thermal video accordingly having some temporal data. Human recognition in thermal videos frames and images and their applications are as yet developing and confronting research domain in the field of computer vision and deep learning. The work presents an accurate as well as efficient system to recognize objects in thermal images under various illumination conditions to analyze its illumination invariance. Table I illustrates a comparative presentation of the related work in the area of object recognition in thermal images. In this paper, the proposed approach thus performs object recognition in adverse

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

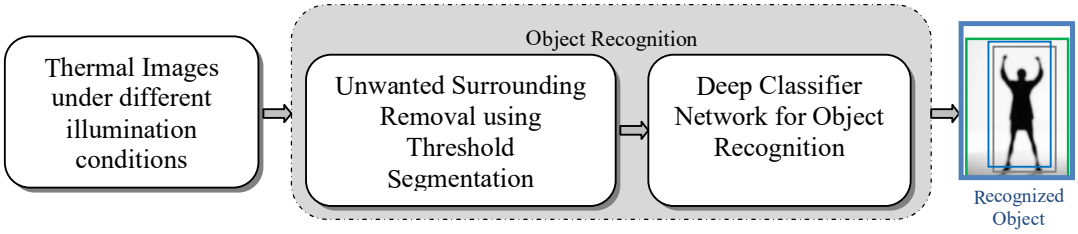


Fig. 1. Basic block diagram of the proposed method used to recognize the object under different illumination conditions.

TABLE I. A comparative presentation of thermal object recognition techniques

Year	Author	Method	Database	Application
2007	Li Zhang et. al. [22]	HOG+SVM, Adaboost	Thermal (OTCBVS)	Pedestrian
2009	Kai Jungling et al. [23]	SURF Feature Detector	Thermal (OTCBVS)	Human
2015	Eun Som Jeon et. al. [24]	Background Subtraction	Thermal (OTCBVS)	Human
2016	Eun JuLee et al. [25]	Boosted Random Forest +CNN	Pedestrian's unsafe behavior (PUB) dataset	Pedestrian
2017	Tao Yang et al. [26]	spatio-temporal tracking	OTCBVS Benchmark Dataset.	Pedestrian
2017	Duyoung Heo et al. [27]	YOLO + Saliency Map	KMU-PD dataset	Pedestrian
2017	Ya-Li Hou et al. [28]	SSD	KAIST Dataset	Pedestrian
2018	Andres Gomez et. al.[29]	CNN	Thermal (Real-time)	Human
2018	R.Soundrapandiyan et al.[30]	HoG+LDP and SVM	OTCBVS Benchmark Dataset.	Pedestrian
2019	Marina Ivasic-Kos et.al. [31]	YOLO	Thermal (Real-time)	Human
2019	Merzouk Younsi et. al. [32]	GMM, Particle Filter	AIC Thermal/ Visible night time dataset, OTCBVS Dataset.	Human, Animals
2019	Dai,Xiaobiao et al. [33]	CNN	Thermal Images	Pedestrian
2019	Nazeer Shahi et al. [34]	CNN	Thermal Images (Real-time)	Human
2020	MATE KRIŠTO et.al. [35]	YOLO	Thermal (OTCBVS Dataset)	Human
2020	A. Wolpert et al. [36]	Fusion CSPNet	KAIST Dataset	Pedestrian
2021	A. K. M. F. Rahman et al. [37]	Deep Saliency Map, Instance Segmentation	KAIST Dataset	Pedestrian

environmental conditions like rain, Haze, cloudy weather and darkness. The thermal OSU Pedestrian Dataset from OTCBVS Benchmark Database [20] is utilized to test the performance of the purposed work for pedestrian detection in various enlightenment conditions. In this methodology, a framework is intended to segment the objects from the images utilizing threshold segmentation. Finally, the Faster-RCNN is used to recognize the pedestrian in the thermal images under various illumination conditions. This paper is organized in different subsections as: section II presents the proposed technique exhaustively including threshold segmentation and deep network [21] (Faster R-CNN) for recognition of object. The results and discussions are illustrated in segment III. At last, the paper is concluded in the segment IV.

II. PROPOSED METHOD

The proposed technique is utilized to recognize the object in the thermal images captured under various illumination conditions. The fundamental block diagram of the proposed work is outlined in figure 1. The thermal images from thermal OSU Pedestrian Dataset from OTCBVS Benchmark Database [20] are utilized for the implementation of the proposed work. The different steps of the proposed approach for the recognition of the objects are delineated in algorithm 1 and discussed in the following subsections.

A. Object Recognition

For the identification of the objects in the thermal images under various illumination conditions, at first the image is pre processed [38] to eliminate noise (undesirable surroundings) from the image and afterward objects are distinguished utilizing Faster R-CNN [39] network.

1) Noise (unwanted surrounding) Removal Using Threshold Segmentation: The undesirable surroundings of the object like trees, poles and structures are considered as noise. Person and other living objects emit more noteworthy infrared radiations than other non-living items. As the thermal image database is utilized, so in the images the living objects seem brighter than others. Subsequently threshold segmentation [40] is applied to eliminate the less bright objects (undesirable surroundings) in the grayscale thermal images. By then, the lower limit T_{low} and upper limit T_{high} range of the white color is characterized in grayscale image. Thus, the thresholding is performed to separate the white color from the image that

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: Object Recognition under different illumination conditions

Input:
Thermal Images under different weather conditions.
Output:
Recognized obstacles.
begin
1. Extract frames *X(i,j)* from the video sequence.
2. Apply threshold segmentation on the input frame *X(i,j)* to remove noise from the frame.
3. Implement the Faster R-CNN to identify the object from the segmented frames.
4. The predicted output is unified with input image to generate the final output.
5. Repeat the steps 1 to 4 for the images under different illumination conditions.
End

tends to living objects; the segmented image is given as:

$$X_{Seg}(i,j) = \begin{cases} 1 & \left(if\ T_{low} \le X(i,j) \le T_{high}\right) \\ 0 & else \end{cases} \quad (1)$$

Where, '1' and '0' addresses white and dark color respectively, X(i,j) input grayscale image and $X_{Seg}(i,j)$ is output image after noise removal. At last, the objects are segmented from the image and rest of the surrounding objects is taken out in the output image. This improves the performance of the following stages of object recognition under various illumination conditions essentially.

B. Deep Network for Object Recognition

Faster Region based Convolutional Neural Network (Faster R-CNN) [39] is used to recognize objects in thermal images under various illumination conditions. It contains two modules: initial, deep fully convolutional network used to propose regions and second is the Fast R-CNN detector that recognizes the objects using region proposals. The Region Proposal Network (RPN) uses 'attention' framework to tell the Fast R-CNN detector network where to look. In this work, VGG-16 network model [41] is used in Faster R-CNN model. The smaller network is slide over the convolutional feature maps created by last shared convolutional layer. It creates the set of rectangular object proposals and introduced into fully connected layers – a classification layer which determines whether the object is a background or not and the regression layer shows that each window has an object by doing bounding box adjustment. In default arrangement, there are anchors at position of image having 3 scales and 3 ratios. After that the bunch of proposals are applied to ROI pooling layer which extracts predefined size feature maps for each anchor so every proposal has an object to classify. Finally, the Fast R-CNN network [42] is used for classification which has two fully connected layers. The one layer classifies the proposals in N+ 1 particular class (where N class and one background class for disposing of bad proposals). Another fully connected layer is used for better adjustment of bounding box for 'N' classes using regression prediction. In Faster R-CNN, two network modules are to be trained. The anchors those cover the ground truth with an IOU more than 0.7 are foreground and the anchors don't cover any ground truth object (for IOU under 0.3) are classified as background. The loss function to be restricted is given as:

$$f_{loss}(\{m_p\},\{n_p\}) = \frac{1}{C}\sum_p l_c(m_p, m_p^*) + \gamma \frac{1}{R}\sum_p m_p^* l_r(n_p, n_p^*) \quad (2)$$

Where, mp is the predicted anchor probability of anchor p being an object. Vector np signifies the parameterized coordinates of predicted bounding box. lr and lc are the regression and classification loss respectively. For classification, Cross Entropy Loss is used and for Bounding Box Regression smooth L1 loss is used. The normalization parameters C and R are mini batch size (i.e. C=256) and anchor

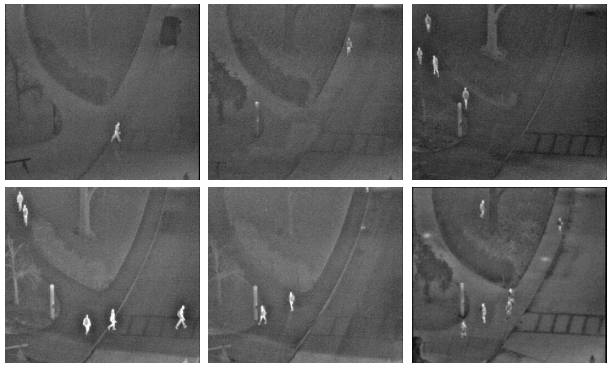


Fig. 3. Some thermal image OSU Thermal Pedestrian Database

locations (i.e. R=2400). The term 'γ' is the weight balancing parameters set to 10 (default). The Fast R-CNN network is trained similarly aside from classification it uses cross entropy loss in between of various object classes alongside end to end back propagation and Stochastic Gradient Descent with momentum having worth of 0.9.

III. RESULT AND DISCUSSION

This portion presents the experimental details of the proposed work to evaluate its exhibition. The experimental results are delineated as:

A. Experimental Setup

In this work, the Thermal OSU Pedestrian Dataset from OTCBVS Benchmark Database [20] is utilized to test the performance of the purposed work in various illumination conditions. The Raytheon 300D warm sensor was utilized to gather dataset at Pedestrian convergence on the Ohio State University grounds. The dataset contains 284 thermal images of 10 distinctive climate conditions like light rain, Haze, cloudy weather and fair sunlight. Some of the images of the database are shown in Fig. 3. The implementation of Faster R-CNN is performed utilizing system with configuration Intel(R) Core (TM) i5-1035G1CPU @ 1.00 GHz, 1.19 GHz, 16 GB RAM and NVIDIA MX230 GPU.

B. Object Recognition

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

The primary step is to recognize object in the thermal images under various light conditions. At first the input image is pre-processed to eliminate noise from it in terms of suppression of the undesirable surroundings of the objects. As discussed in segment 2, threshold segmentation approach is utilized for the removal of the undesirable surroundings. As of now discussed in the previous section, living objects seem more white than other objects when captured by the thermal camera because of the fact that the living objects emit more infrared rays than other objects. The white color is held in the image and rest all are eliminated from the image. For the masking of objects in the image two edges (T_{low} and T_{high}) are utilized for white color. As the pedestrians seem whiter than any remaining items in the scene, so the upper limit T_{high} is set to '1'. In the experiment, the lower limit T_{low} value is changed between ranges (0.0-1.0) to get the optimum segmentation result.

If lower threshold T_{low} is set at lower side of the range, the undesirable surroundings are not appropriately taken out from the scene and the necessary information is likewise removed when the T_{low} value is near to '1'. The segmentation results for distinct values of the lower limit T_{low} are analyzed. The optimum outcomes are recorded at threshold values (T_{low} =0.85, T_{high} =1) for thermal images containing pedestrian. At last, the Faster R-CNN network is applied to recognize pedestrians in the thermal images under various illumination conditions. The recognition performance is investigated with the Faster R-CNN network that was trained on thermal pedestrian image data. The 228 thermal images (80%) of the image database are used as training data and 56 thermal images (20%) of the image database are used as testing data from the Thermal OSU Pedestrian Dataset.

Figure 4(a) shows the thermal image under light rain conditions in the afternoon with temperature: 68°F, minimal dew point: 58°F, humidity: 70% and visibility: 9.0 miles. The segmented image as demonstrated in figure 4(b) is applied to the Faster R-CNN for the recognition of pedestrians. Figure 4(c) shows the output of the Faster R-CNN with recognized pedestrian. The final recognized pedestrian in the thermal image under light rain conditions are appeared in Fig. 4(d). The thermal image captured under Haze condition in the morning with temperature: 50°F, minimal dew point: 47°F, humidity: 89% and visibility: 2.5 miles is appeared in figure 5(a). Figure 5(b) shows the segmented image having the data about pedestrian is applied to the Faster R-CNN for the recognition of pedestrian. Figure 5(c) addresses the output of the Faster R-CNN with perceived pedestrian. The final recognized pedestrian in the input image of haze conditions is appeared in figure 5(d). After that the proposed work is tested for thermal image having scene captured under dark cloudy conditions.

Figure 6(a) shows the thermal image under dark cloudy conditions in the morning with temperature: 53°F, minimal dew point: 46°F, humidity: 77% and visibility: 8.0 miles. The segmented image as demonstrated in figure 6(b) that contains the information about pedestrian is applied to the Faster R-CNN for its recognition. Figure 6(c) represents the output of the Faster R-CNN with recognized pedestrian. The final recognized pedestrian in the input image of dark cloudy

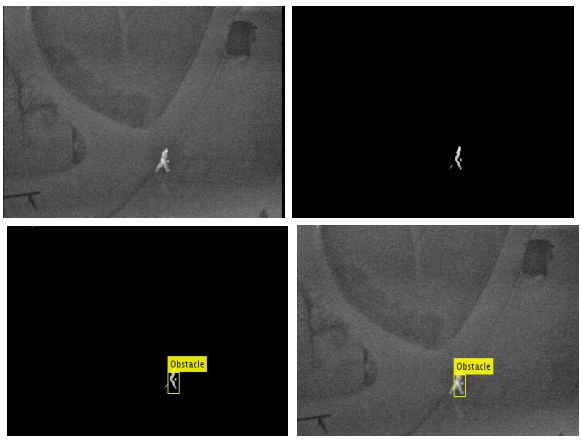


Fig. 4 (a) Input thermal image under light rain condition (b) The segmented image containing information of pedestrians (c) The output of the Faster R-CNN with recognized pedestrian (d) Final Recognized pedestrians in input thermal image

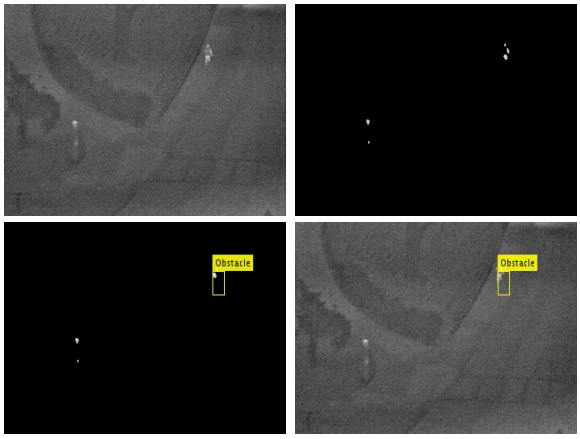


Fig. 5 (a) Input thermal image under haze condition (b) The segmented image containing information of pedestrians (c) The output of the Faster R-CNN with recognized pedestrian (d) Final Recognized pedestrians in input thermal image

is appeared in figure 6(d). The next illumination condition is considered as when climate is clear and sunny. The thermal image taken under reasonable bright conditions in the morning with temperature: 71°F, minimal dew point: 32°F, humidity: 45% and visibility: 10 miles is appeared in figure 7(a). Figure 7(b) shows the segmented image having the pedestrian information. At that point this thermal image is applied to the Faster R-CNN for the recognition of pedestrians. Figure 7(c) addresses the output of the Faster R-CNN with recognized pedestrian. The final recognized pedestrian in the image of clear and sunny weather condition is appeared in figure 7(d). Finally, the light condition in partly cloudy weather is thought of. The thermal image caught in partly cloudy weather with temperature: 37°F, minimal dew point: 30°F, humidity: 76% and visibility: 8 miles is appeared in figure 8(a). The segmented image that contains the pedestrian information is appeared in figure 8(b). At that point this is applied to the Faster R-CNN for the recognition of pedestrian. The output of

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

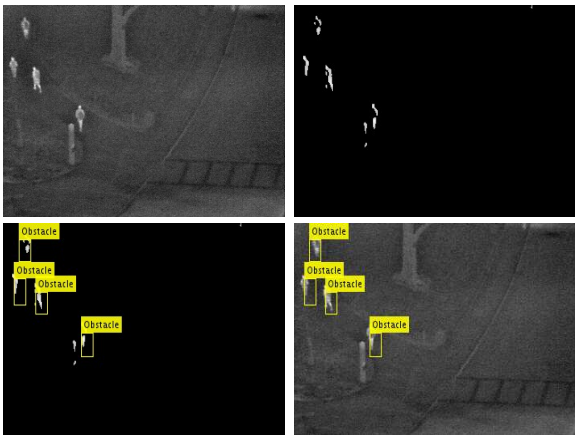


Fig. 6 (a) Input thermal image under dense cloudy condition (b) The segmented image containing information of pedestrians (c) The output of the Faster R-CNN with recognized pedestrian (d) Final Recognized pedestrians in input thermal image

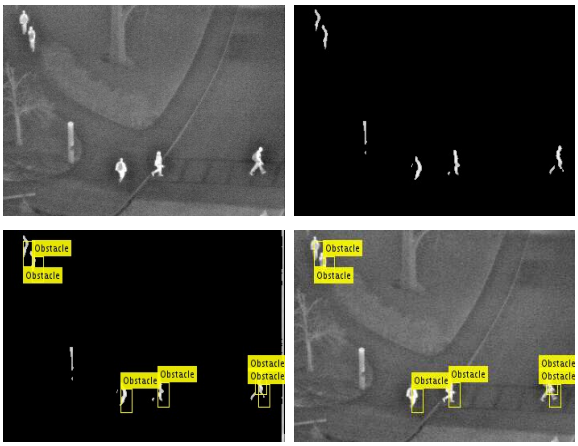


Fig. 7 (a) Input thermal image under fair and sunny condition (b) The segmented image containing information of pedestrians (c) The output of the Faster R-CNN with recognized pedestrian (d) Final Recognized pedestrians in input thermal image

the Faster R-CNN with recognized pedestrian is addressed in figure 8(c). Figure 8(d) shows the final recognized pedestrian in the thermal image of partly cloudy conditions. Subsequently the performance of the proposed work is analyzed under various illuminations in various climate conditions; it gives great illumination invariance by showing better performance in all illumination conditions.

C. Performance Evaluation

The performance of the proposed work can be assessed with the assistance of various parameters. In this work, the significant parameters which are used to analyze the performance are accuracy, precision and recall given as:

$$Accuracy = \frac{(True\ Positive + True\ Negative)}{Total\ Number\ of\ images} \quad (3)$$

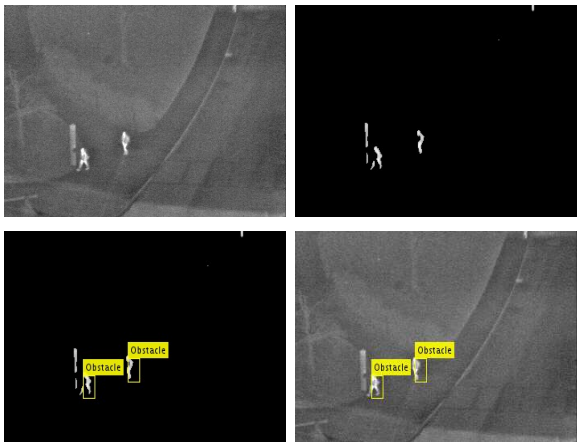


Fig. 8 (a) Input thermal image under partly cloudy condition (b) The segmented image containing information of pedestrians (c) The output of the Faster R-CNN with recognized pedestrian (d) Final Recognized pedestrians in input thermal image

$$Precision = \frac{True\ Positive}{True\ Positive + Fals\ \ Positive} \quad (4)$$

$$Recall = \frac{True\ Positive}{Number\ of\ images\ havin\ \ object\ class} \quad (5)$$

The accuracy [43] is the ratio of sum of true positive and true negative to total number of images in the data set. The precision [43] is the ratio of true positive to sum of true positive and false negative, while, the ratio of true positive to the total number of images having same object class is recall [43].The recognition performance of the proposed work is evaluated individually for each illumination condition in OSU database using evaluation metrics. Then, the overall performance is analyzed by taking the average of each evaluation metric calculated under each illumination condition as shown in Table II.

TABLE II. Performance of the proposed strategy for various illumination conditions in OSU database

Condition	Accuracy	Precision	Recall
Dense Cloudy	0.9348	0.9137	0.9203
Light Rain	0.9211	0.8874	0.9118
Partly Cloudy	0.9186	0.8816	0.9081
Haze	0.9274	0.9062	0.9147
Sunny	0.9172	0.8769	0.9073
Average	**0.9238**	**0.8932**	**0.9124**

1) Comparison with State-of-the-art Methods

To evaluate the performance of the proposed approach for object recognition in OSU Thermal Pedestrian Database from OTCBVS Benchmark Dataset Collection, it is contrasted with other state of the art recognition techniques with these parameters as demonstrated in Table III. At first, the performance of HOG+SVM [22], SURF [23], GMM+ Particle Filter [26], spatio-temporal [32] and YOLO [35] for recognition of obstacles is compared with proposed method. The proposed recognition framework outperforms all the strategies discussed in comparison based on accuracy, precision and recall.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

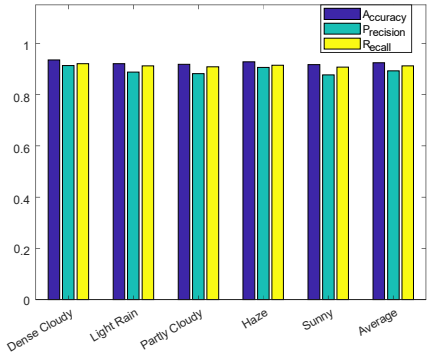


Fig. 9. Comparative Plot for performance of the proposed strategy for various illumination conditions in OSU database

TABLE III. Performance comparison of the proposed strategy for various state of the art for OSU database

Method	Accuracy	Precision	Recall
HOG+SVM+Adaboost[22]	0.7870	-	-
SURF [23]	-	-	0.70
Spatio-temporal tracking [26]	0.89	-	-
GMM, Particle Filter [32]	0.9145	-	-
YOLO [35]	-	0.84	0.87
Proposed Method	**0.9238**	**0.8932**	**0.9124**

IV. CONCLUSION AND FUTURE SCOPE

The proposed procedure is sufficiently adequate for the object recognition in thermal images under various illumination conditions. This paper presented a novel procedure based on deep learning network to recognize objects in images of in OSU thermal pedestrian data base. Accordingly, this work can be robust strategy to develop a proficient surveillance framework to hinder mishaps for the public safety upgrade in various lightening conditions. This methodology will be in likewise cost effective as it needn't any enormous change in the architecture. Later on, the system will be extended to effective recognition of other object classes to improve the adequacy of the system.

REFERENCE

[1] P. Viola and M. Jones, ‘‘Rapid object detection using a boosted cascade of simple features,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 1, Dec. 2001, pp. 511–518.

[2] C. K. Eveland, D. A. Socolinsky, and L. B. Wolff, ‘‘Tracking human faces in infrared video,’’ Image Vis.Comput.,vol.21,no.7,pp.579–590,Jul. 2003.

[3] N. Dalal and B. Triggs, ‘‘Histograms of oriented gradients for human detection,’’ in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR). Washington, DC, USA: IEEE Computer Society, vol. 1, Jun. 2005, pp. 886–893.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ‘‘ImageNet classification with deep convolutional neural networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A. C. Berg, ‘‘SSD: Single shot multi-box detector,’’ in Proc. Eur. Conf. Comput. Vis., 2016, pp. 21–37.

[6] J. Dai, Y. Li, K. He, and J. Sun, ‘‘R-FCN: Object detection via region based fully convolutional networks,’’ in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 379–387.

[7] R. Girshick, ‘‘Fast R-CNN,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1440–1448.

[8] K. He, G. Gkioxari, and R. Dollar, ‘‘Mask r-CNN,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2980–2988.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, ‘‘You only look once: Unified, real-time object detection,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 779–788.

[10] J. Redmon and A. Farhadi, ‘‘YOLO9000: Better, faster, stronger,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 7263–7271.

[11] J. Redmon and A. Farhadi, ‘‘YOLOv3: An incremental improvement,’’ 2018, arXiv:1804.02767.

[12] D. Ghose, S. M. Desai, S. Bhattacharya, D. Chakraborty, M. Fiterau, and T. Rahman, ‘‘Pedestrian detection in thermal images using saliency maps,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2019, pp. 1–10.

[13] D. Heo, E. Lee, and B. C. Ko, ‘‘Pedestrian detection at night using deep neural networks and saliency maps,’’ Electron. Imag., vol. 17, pp. 060403-1–060403-9, Jan. 2018.

[14] M. Ivasic-Kos, M. Kristo, and M. Pobar, ‘‘Human detection in thermal imaging using YOLO,’’ in Proc. 5th Int. Conf. Comput. Technol. Appl. (ICCTA), New York, NY, USA, 2019, pp. 20–24.

[15] M. Ivasic-Kos, M. Kristo, and M. Pobar, ‘‘Person detection in thermal videos using YOLO,’’ in Proc. SAI Intell. Syst. Conf. Cham, Switzerland: Springer, 2019, pp. 254–267.

[16] A. Gomez, F. Conti, and L. Benini, ‘‘Thermal image-based CNN’s for ultra-low power people recognition,’’ in Proc. 15th ACM Int. Conf. Comput. Frontiers. New York, NY, USA: ACM, May 2018, pp. 326–331.

[17] Q. Liu, X. Lu, Z. He, C. Zhang, and W.S. Chen, ‘‘Deep convolutional neural networks for thermal infrared object tracking,’’ Knowl.-Based Syst., vol. 134, pp. 189–198, Oct. 2017.

[18] N. Shahid, G.H. Yu, T. D. Trinh, D. S. Sin, and J.Y. Kim, ‘‘Real-time implementation of human detection in thermal imagery based on CNN,’’ J. Korean Inst. Inf. Technol., vol. 17, no. 1, pp. 107–121, Jan. 2019.

[19] X. Wang and S. Hosseinyalamdary, ‘‘Human detection based on a sequence of thermal images using deep learning,’’ Int. Arch. Photogramm., Remote Sens., Spatial Inf. Sci., vol. 42.2/W13, pp. 1–6, Jun. 2019.

[20] J. W. Davis, M. A. Keck, “A two-stage template approach to person detection in thermal imagery.” In Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION), vol. 1, 364–369, 2005.

[21] R. Goel, A. Sharma, and R. Kapoor, “Object Recognition Using Deep Learning,” Journal of Computational and Theoretical Nanoscience, vol. 16, pp. 4044–4052, 2019.

[22] Li Zhang, B. Wu and R. Nevatia, "Pedestrian Detection in Infrared Images based on Local Shape Features," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1-8.

[23] K. Jungling and M. Arens, "Feature based person detection beyond the visible spectrum," in 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009, pp. 30-37.

[24] E. Jeon, C. Jong-Suk, Ji. H. Lee, and K. Y. Shin, “Human Detection Based on the Generation of a Background Image by Using a Far-Infrared Light Camera,” Sensors (Basel, Switzerland), vol. 15, pp. 6763-88, 2015.

[25] E. J. Lee, B. C. Ko, and J. Y. Nam, “Recognizing pedestrian’s unsafe behaviors in far-infrared imagery at night,” Infrared Physics & Technology, vol. 76, pp. 261-270, 2016.

[26] T. Yang, D. Fu, and S. Pan, “Pedestrian tracking for infrared image sequence based on trajectory manifold of spatio-temporal slice,” Multimed Tools Appl, vol. 76, pp. 11021–11035, 2017.

[27] D. Heo, E. Lee, B. Ko, “Pedestrian Detection at Night Using Deep Neural Networks and Saliency Maps,” Journal of Imaging Science and Technology, vol. 61, 2017.

[28] Y. Hou, Y. Song, X. Hao, Y. Shen and M. Qian, "Multispectral pedestrian detection based on deep convolutional neural networks," in 2017 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), 2017, pp.1-4.

[29] A. Gomez, F. Conti, and L. Benini, “Thermal image-based CNN's for ultra-low power people recognition, in Proceedings of the 15th ACM International Conference on Computing Frontiers (CF '18). Association for Computing Machinery, New York, NY, USA, 2018, pp. 326–331.

[30] R. Soundrapandiyan, and P.V.S.S.R. ChandraMouli, “An Approach to Adaptive Pedestrian Detection and Classification in Infrared Images Based on Human Visual Mechanism and Support Vector Machine,” Arabian Journal for Science and Engineering, vol. 43, 2017.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

[31] M. Ivasic-Kos, M. Kristo, and M. Pobar, "Human Detection in Thermal Imaging Using YOLO," in Proceedings of 5th International Conference on Computer and Technology Applications (ICCTA 2019). Association for Computing Machinery, New York, NY, USA, 2019, pp. 20–24.
[32] Y. Merzouk, D. Moussa, S. Patrick, "Automatic multiple moving humans detection and tracking in image sequences taken from a stationary thermal infrared camera," Expert Systems with Applications, vol. 146, 2020.
[33] X. Dai, Y. Duan, J. Hu, S. Liu, C. Hu, Y. He, D. Chen, C. Luo, and J. Meng, "Near infrared nighttime road pedestrians recognition based on convolutional neural network, Infrared Physics & Technology, vol. 97, 2019, pp. 25-32.
[34] N. Shahid, G. Yu, T.D. Trinh, D. Sin, and J. Kim, "Real-Time Implementation of Human Detection in Thermal Imagery Based on CNN," Journal of KIIT. vol. 17, no. 1, Jan. 31, 2019, pp. 107-121.
[35] M. Kristo, M. Ivasic-Kos and M. Pobar, "Thermal Object Detection in Difficult Weather Conditions Using YOLO," in IEEE Access, vol. 8, pp. 125459-125476, 2020.
[36] A. Wolpert, M. Teutsch, M. Sarfraz, and R. Stiefelhagen, "Anchor-free Small-scale Multispectral Pedestrian Detection,"2020, ArXiv abs/2008.08418.
[37] A. K. M. F. Rahman, M. Raihan, S. M. Islam, S.M., "Pedestrian Detection in Thermal Images Using Deep Saliency Map and Instance Segmentation," International Journal of Image, Graphics and Signal Processing, vol. 13, no. 1, 2021, pp. 40-49.
[38] G. Ranganathan, "A Study to Find Facts Behind Preprocessing on Deep Learning Algorithms," Journal of Innovative Image Processing (JIIP), vol. 3, no. 01, 2021, pp. 66-74.
[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, pp. 1137–1149, 2017.
[40] K. Bhargavi, and S. Jyothi, "A Survey on Threshold Based Segmentation Technique," in Image Processing. International journal of innovative research and development, vol. 3, 2014.
[41] S. Tammina, "Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images," International Journal of Scientific and Research Publications, vol. 9, 2019.
[42] R. Girshick, "Fast R-CNN," in Proceedings of IEEE International Conference on Computer Vision (ICCV), 2015.
[43] C.B. Murthy, M.F. Hashmi, N.D. Bokde, and Z.W. Geem, "Investigations of Object Detection in Images/Videos Using Various Deep Learning Techniques and Embedded Platforms—A Comprehensive Review," Applied Sciences, vol. 10, no. 9, 2020.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:42:23 UTC from IEEE Xplore. Restrictions apply.

ISSN: 0974-5823 Vol. 6 No. 1 January-June, 2021

International Journal of Mechanical Engineering

Design and Analysis of a Novel Y-Type Hydraulic Braking System for Implementation in ATVs

Kunal Mathur
Mechanical Department, Delhi Technological University (DTU), Delhi, India, kunal.mathur98@gmail.com

Lalit Kumar Choudhary
Mechanical Department, Delhi Technological University (DTU), Delhi, India, lcdtume@gmail.com

Krovvidi Srinivas
Mechanical Department, Delhi Technological University (DTU), Delhi, India, srinivaskrovvidi@dtu.ac.in

Date of Submission: 08th December 2020 Revised: 21st February 2021 Accepted: 13th February 2021

How to Cite: **Mathur, K., Choudhary, L.K. and Srinivas, K. (2021). Design and Analysis of a Novel Y-Type Hydraulic Braking System for Implementation in ATVs. International Journal of Mechanical Engineering 6(1), pp.1-10.**

Abstract **- The aim of this paper is to design, develop and validate a 'Y-type' dual split hydraulic braking system for proper implementation in certain production based ATVs and other vehicles with similar specifications. This novel braking system is supposed to be inherently safer than the pre-existing ones being used by different ATVs. Various design approaches are followed in a calculative fashion along with some sound market selection strategies to develop and implement this braking system. Some braking components are designed and analysed using CAD and CAE software respectively while other components are selected from market to enhance standardization in the braking system. Components have been designed against static, thermal and fatigue loading. The entire braking system is validated for multiple ATV models manufactured by different OEMs using vehicle dynamics equations followed by systematic torque comparison and conformance with the government safety regulations.**

Index Terms – Ansys, ATV, FEA, NPT, SolidWorks, Y-type dual split hydraulic braking system.

INTRODUCTION

ATVs (All Terrain Vehicles) are used for multiple purposes such as recreation, motorsports or as commercial passenger vehicles. For such a wide range of customers of this product, occupant safety during dynamic conditions such as braking is a crucial need. This braking system is capable of locking a minimum of three wheels during a single circuit failure. Hence the brake circuit design provides a much safer and a reliable method of locking the wheels during a single brake circuit failure by reducing the chances of vehicle skid considerably.

In this paper, a Y-type dual split hydraulic braking circuit has been designed and analysed for failure in order to install in a particular group of ATVs. Unlike the common X-type or the horizontal split type circuit, this braking circuit locks three wheels in case of a single circuit failure, hence reducing the chances of skidding while braking. This sole factor makes the Y-type circuit a safer and a reliable option. A typical Y-type braking system involves each circuit to connect with three wheels/calipers at a time when two of the wheels/calipers belong on one axle (Figure 2). Specific ATVs are selected with similar specifications in this paper. The proposed Y-type dual split hydraulic braking system is designed for easy implementation, quick assembly/disassembly in such ATVs using standard available market brake components and connectors which are thoroughly selected according to their functions and compatibilities. This makes the system design highly modular, standardised and reduces the overall cost during a mass production process. The compatibilities of these components were confirmed after a series of calculations. The custom components are designed in SolidWorks® and analysed for the dynamic environment that the vehicle would be subjected to using CAE (Computer Aided Engineering) tools of Ansys 17.1®. These components are analysed for deformations, strength, fatigue life, thermal stresses and thermal fatigue. Possible alterations in the system that could occur in the implementation of this circuit design into ATVs not belonging to the group mentioned above are also discussed in this study. The entire developed

braking system is also tested for its conformability with the Indian government CMV Rules and standards.

Various research papers in the similar field are studied and referred for this study. The study by Jain PK et al. in [1] is about the development of a racing Go-kart in which the design and analysis of the braking system has been done. Multiple braking factors such as aerodynamic drag and rolling resistance are not taken into account for the deceleration effect. Moreover, the design of the pedal is not analysed for lateral deformation, which is required as the driver exerts an appreciable amount of force on the pedal with his/her foot in the lateral direction while exiting the vehicle. This factor has been taken into account in the current study. The selection of driver pedal effort is abstract in the studies of Jain PK et al. in [1] and Sharma A et al. in [2] and are not according to any anthropological survey standards. Mishra D et al. in [3] describes the design and analysis of the braking system of an ATV for Baja SAE. The use of a balance bar has been made which requires critical adjustment and its implementation according to the desired brake bias is often time consuming. We have deployed the standard proportioning valves that can be purchased as a whole unit in bulk for industrial mass production needs. These are relatively easy to operate and set. Moreover, the use of four different calipers has been done in [3] which offers difficulty in tuning the brakes for all wheels to lock simultaneously without skidding. Hence, the braking system developed by us has deployed a total of four disc brakes with same calipers. Shrivastava D. [4] and Sharma A et al. [2] have implemented the use of drum brakes in an ATV for Baja SAE, which are bulkier and have a lot of drawbacks as compared to the disc brakes. The work of Gupta US et al. [5] is based on the same study as that of Shrivastava D. [4] but the braking system of the ATV in [5] has been designed setting the frictional and the braking torques equal in magnitude. In practice, the braking torque would always be lesser in magnitude as compared to its computed value, owing to the irregularities in the braking fluid that violate Pascal's Law. Due to this reason, the braking torque has been kept always greater than the frictional torque in this paper.

DESIGN METHODOLOGY AND COMPONENT SELECTION

I. Selection of ATVs

The following ATVs, having similar GVW (Gross Vehicle Weight), axle weight distribution, tire diameters, maximum velocity, hydraulic disc brakes and maximum number of passengers (between one and two) have been selected.

- **ATV-1:** Polaris ACE 500®
- **ATV-2:** 2011 Polaris Ranger 500 EPI®
- **ATV-3:** Polaris ACE 570 SP®
- **ATV-4:** 2018 Kymco UXV 450i CAMO®
- **ATV-5:** 2018 Bennche Spire 150®



FIGURE 1
ATV-1 (POLARIS ACE 500) [6]

II. Brake circuit type

For modern front-wheel-drive cars the X-split circuit appears to have become an established practice, while the simple front-rear split circuit is still generally used on rear wheel drive cars. In off road vehicles usually ATVs or All-Terrain Vehicles, Y-split circuit should be used with the help of four disc brakes for efficient braking. This system is widely used nowadays in racing arena and SAE events for students. Y-split circuit is best suited for the ATVs [7]. Hence Y-type split hydraulic dual circuit is chosen for implementation in the selected ATVs. In this type of circuit design, two completely independent brake line circuits operate simultaneously, each having their own separate reservoirs. One circuit is attached to a disc D1 at front and both the discs (D3 and D4) at the rear of the vehicle. The other circuit connects D4 with both the discs at front (D1 and D2) as shown in figure 2. Hence in case of a single brake circuit failure due to leakage or contamination of brake fluid, at least three wheels would get locked, thereby diminishing the chances of vehicle skidding while braking.

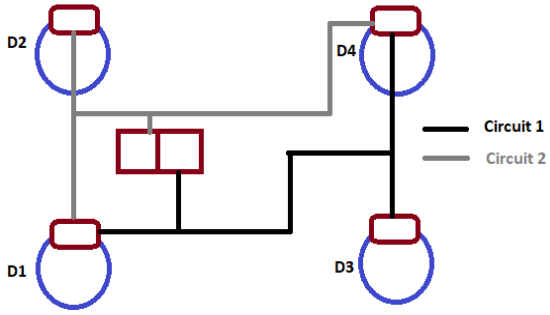


FIGURE 2
Y-TYPE DUAL SPLIT BRAKE CIRCUIT

III. Brake system components

The braking system for the ATVs will contain the following elements:

- Master cylinder and reservoirs (selected)
- Brake calipers (selected)
- Proportioning valves (selected)
- Unions/connectors (selected)
- Brake fluid (selected)

- Hoses (designed; only length)
- Brake pedal (designed)
- Brake pads (selected)
- Brake discs (designed)

At first, the selection of the standard selected components that are used in the braking circuit have been discussed below.

(a) Master Cylinder

Tandem master cylinders are preferred over single master cylinders for the following reasons. First, the tandem master cylinders have two separate piston chambers with two different inlet and outlet ports. Secondly, in order to implement a dual hydraulic braking system with independent circuits, one would have to use two single master cylinders which would have acquired space. This factor is eliminated by having one tandem master cylinder.

Wilwood's TM-1 tandem master cylinder is selected for its following features [8]:

- Lower cost
- Light weight since it is made of aluminium
- It comes with two separate reservoirs with respective mount holes. Both can be mounted on the cylinder itself.
- It can be mounted horizontally, thereby decreasing vertical space at the front of the vehicle.
- It comes with an attached push rod with fork ends for holding the pedal and transferring the pedal effort to the cylinder easily.
- It has a suitable stroke length that would easily accommodate the pedal travel by the driver while braking.



FIGURE 3
WILWOOD TM-1 MASTER CYLINDER [8]

(b) Brake Calipers

Calipers are used to support the brake pads that arrest the motion of the rotor to cause braking. They also support the pistons that transmit the force from the fluid to the pads. Wilwood's PS-1 caliper with one inch pistons is selected for the following reasons [8]:

- Compact body with appreciable amount of clamping force
- Light weight (aluminium body)
- Lower cost
- Different designs of the same product are available both for left and right wheels of the vehicle simplifying brake bleeding. Hence the use of different calipers for different wheels is not required. For this reason, PS-1 is selected for all the four wheels.
- Utilizes two stainless steel deep cup pistons to minimize heat transfer from the pads to the brake fluid.
- Only one bleeding valve is present on the caliper. The lesser number of leakage points decreases the chances of brake fluid leakage.
- Bleeding valve faces upwards for easy bleeding [3].
- The allowed thickness of the disc is lesser, which simplifies the manufacturing of the brake disc.



FIGURE 4
WILWOOD PS-1 CALIPER (RIGHT SIDED) [8]

(c) Proportioning Valves

Proportioning valves are selected over a balance bar to provide brake bias for the following reasons:

- Easy to use and install
- Balance bar requires the use of two single master cylinders.
- Balance bar implementation is not possible in a Y-type brake circuit since each circuit is connected to the wheels of both the axles at a time.

Wilwood's proportioning valve with M10x1 NPT (National Pipe Thread) inlet and outlet is selected for the following reasons [8]:

- They are much more compact, reliable and cheaper as compared to other companies.
- Ports have fine and tapered threads that prevent leakage of the pressurized fluid.
- Ports with M10x1 NPT are also used in the selected TM-1. Hence union connections of the same type could be used throughout the braking circuit. This reduces the number of different components used in the system, making it much more economical for bulk orders in a mass production process.
- Control in the amount of brake bias is achieved by just turning the knob of the proportioning valve (easy to use).
- Two lateral bolt holes are available for easy mounting.



FIGURE 5
WILWOOD M10X1 PROPORTIONING VALVES [8]

(d) Unions

Four types of unions/connectors are used to connect the different braking components of the circuit. They are as follows:

- **Simple unions:** have one inlet and outlet for connecting two brake components together. An M10x1 NPT union with both male threads is used here to join brake hoses with proportioning valves and master cylinders. Another union having one end as above and the other as male 1/8-27 NPT is used to join one brake hose and one caliper.
- **Tee unions:** are used to connect three brake circuit components for the transfer of pressurized brake fluid. One type of Tee connector has all three male M10x1 NPT connections for joining the corresponding female thread hoses. The other type of Tee union has one male 1/8-27 NPT and rest two M10x1 NPT male threads for connecting two brake hoses with a caliper (figure 2).
- **Cross unions:** used to connect four braking components for proper brake fluid transfer. Cross unions having four M10x1 NPT male threads have been used to connect four female thread brake hoses with each other.

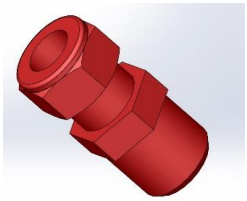


FIGURE 6(A)
SIMPLE UNION [8]

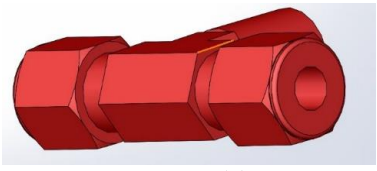


FIGURE 6(B)
TEE UNION [8]

(e) Brake fluid

Brake fluids are used to transmit pressure effectively throughout the brake hoses and connectors all the way up to the caliper pistons which arrest the motion of the brake disc by pushing the pads against it. Following are the required qualities of brake fluid [9]:

- High boiling
- Low freezing point
- Ability to not deteriorate metal or rubber brake parts
- Ability to absorb moisture

TABLE I
COMPARISON OF BRAKE FLUIDS [9]

DOT 3	DOT 4	DOT 5
Absorbs moisture the quickest	Lower absorbing power than DOT 3	Does not absorb moisture
Lowest boiling point of 205°C	Boiling point 230°C	Highest boiling point 260°C
Restricted for road use only	High temperature tolerance	Used in heavy vehicles
Lowest cost	Moderate cost	Highest cost

DOT 3, DOT 4 and DOT 5 are compared for the above properties. DOT 4 is chosen for the braking system after comparison on the basis of the above properties.

(f) Brake hoses

Standard flexible steel braided rubber hoses are to be used in our system. This prevents physical damage to the hose to a great extent and imparts flexibility to the brake lines for allowing them to connect at intricate corners in the vehicle. The length of these hoses can be customized by the hose supplier easily. Hoses with M10x1 NPT female threads would be used all across our circuit.

(g) Brake Pedal

It is a linkage that is used to provide mechanical advantage to the driver's pedal effort and transfers the force. The pedal force multiplied by the leverage ratio is applied to the master cylinder to produce a hydraulic pressure [10].

Class-2 lever mechanism (where the loads are at the same side of the pivot point) is chosen so as to push the pushrod in the same direction as the pedal travel. This would enable the OEM to install the master cylinder assembly behind the pedal, giving the driver room for his feet, hence preserving the ergonomic aspect of the vehicle. The design of the pedal is made in SolidWorks 2017®. The pedal would be pivoted on an extruded boss attached to its bracket and retracted with the help of a torsion spring. A small mount for the stop plug would be welded on the arm of the pedal so that the stop plug would restrict the pedal travel according to the pedal ratio and the maximum stroke of the master cylinder. The stop plug of the pedal arm would interfere with that on the bracket which would arrest further travel of the pedal. Figure 7(a) depicts the CAD model of the pedal.

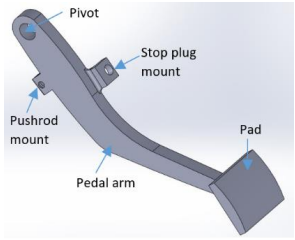


FIGURE 7(A)
CAD MODEL OF BRAKE PEDAL

Results based on FEA (Finite Element Analysis) using Ansys 17.1® helped to modify the design according to the software simulated conditions which are made very similar to that of the real dynamic environment of the vehicle.

For the 3D model of the pedal, two of the dimensions are very large in comparison to the third one, hence 2D meshing is carried out in the software [11]. The mesh model contains a fine mesh with 'quad' and 'tria' elements for accurate results [11]. The types of analysis performed on the mesh model are listed below:

- Deformation and vonMises stress on pedal effort
- Deformation and vonMises stress on lateral loading
- Fatigue analysis

The above analysis are completed after the input of required boundary conditions that are computed in the calculations section. Moreover, the pedal analysis are done when the pedal is in its maximum travel position (fully depressed).

Deformation and vonMises stress on pedal effort: Pedal effort (PE) is the normal force exerted by the driver on the pedal while braking. PE equals to 850N (calculated in the next section). The pivot hole of the pedal is set as a constraint while the load of 850N is applied normal to the pad of the pedal.

Aluminium-6061 is selected as the material of the pedal (owing to its light weight and satisfactory strength) and corresponding values are entered in the software. After the generation of the FE model, the software displays the deformation and stress contours.

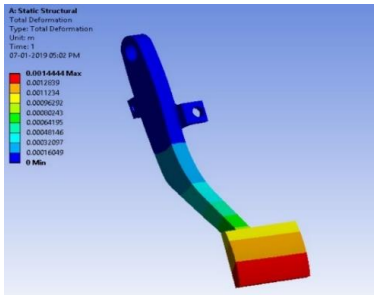


FIGURE 7(B)
TOTAL DEFORMATION UNDER NORMAL LOAD

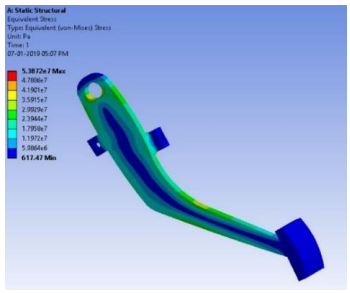


FIGURE 7(C)
VONMISES STRESS DISTRIBUTION FOR NORMAL LOAD

From figure 7(c), it can be concluded that the maximum equivalent stress comes out to be around 54MPa, whereas the yield strength of aluminium-6061 is 240MPa. Hence factor of safety (FOS) equals 4.44 which is suitable indeed.

Deformation and vonMises stress on lateral force: When the driver exits a vehicle, he/she might accidently apply some force on the lateral edge of the pedal. Hence the pedal is analysed for failure under lateral load. A lateral load of 300N is entered at the left edge of the pedal (assuming a right-hand drive vehicle). The constraint and material are same as above and are the boundary conditions of this FE model.

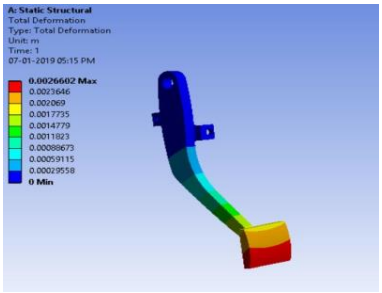


FIGURE 7(D)
TOTAL DEFORMATION FOR LATERAL LOAD

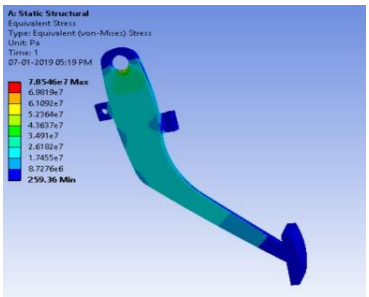


FIGURE 7(E)
VONMISES STRESS DISTRIBUTION FOR LATERAL LOAD

From figure 7(e), it is evident that the maximum vonMises stress in the pedal design upon lateral loading comes out to be around 78.6MPa. Hence the factor of safety FOS equals 3.0 which is acceptable.

Fatigue analysis: Since the proposed braking system has been designed for ATVs for public use, fatigue analysis is done to ensure the desired life of such components under prolonged load applications. Although the S-N curve of the chosen material does not have a well-defined knee point, after the input of the boundary conditions the design is found to be safe till 10^8 load cycles. The loading conditions entered are 0-850N of alternating loads and same constraints as those in the previous cases. S-N curve data of Al-6061 is also entered. The design is found to have a factor of safety (FOS) of 1.427 which is the ratio of the endurance strength according to 10^6 cycles and the maximum equivalent alternating stress. This suggests that the pedal design would not fail at given alternating loads before 10^8 such cycles which is more than favourable.

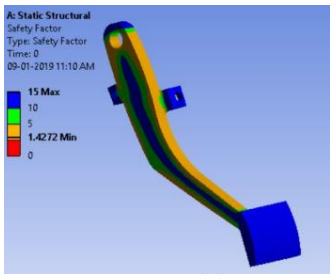


FIGURE 7(F)
FACTOR OF SAFETY AT 10^6 LOAD CYCLES

(h) Brake pads

The required properties of the brake pad here are long service life, high coefficient of friction between brake disc and the pad and the ability to recover easily from high temperatures and moisture [12].

The selected material is semi metallic brake pads, since they have a satisfactory service life and lesser noise generation than the metallic brake pads. They also do not cause thermal warpage unlike the ceramic brake pads [12].

(i) Brake disc

Discs over drum brakes are chosen for the current braking system because of the following factors:

- Lesser heat generated than drum brakes [5]
- Lesser vehicle stopping time than drum brakes [5]
- Lighter in weight and occupy lesser space
- Show lesser torque variation in the course of a stop as compared to the drum brakes which often exhibit a 'sag' in its torque curve [13]
- Have the ability to operate with little fade at high temperatures [14]
- Offer linear relationship between brake torque and pad/rotor friction coefficient [14].

Float type disc arrangement of the brake disc involves the outer brake disc riveted to a separate inner disc. On the other hand, the fixed type brake disc is a single brake disc in action. Float type discs are preferred over fixed type discs for the following reasons.

- They prevent the chances of thermal warpage of the brake disc during rubbing of the pads (unlike fixed discs).
- The floating brake disc can transmit mechanical vibrations to the riveted inner disc. Hence the brake disc is free from deformation due to the vibrations.

Figure 8(a) shows the CAD model of the brake disc that is made using SolidWorks 2017®. The model shows the outer brake discs riveted to an inner disc that has bolt holes for mounting on the hub. Slots and holes are present across the design for effective heat dissipation and weight reduction. The dimensions are taken in accordance with the standard design of the PS-1 caliper [8].
Thickness of disc = 0.19 inch = 0.004826m
Outer diameter of disc = 9 inch = 0.2286m
Brake pad width = 1 inch = 0.0254m
Effective radius of disc (r_{eff}):

$$r_{eff} = (9/2) - (1/2) = 4\ inch = 0.1016m \quad (1)$$

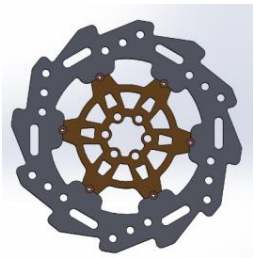


FIGURE 8(A)
CAD MODEL OF BRAKE DISC

Results based on FEA using Ansys 17.1® are helpful for improving the design according to the software simulated version of the real dynamic conditions that the design would be subjected to. Hence the following CAE simulations were done:

- Thermal analysis
- Fatigue analysis

According to the material data given in [15] for brake discs, out of Gray cast iron (GCI), Ti-alloy (Ti-6Al-4V), Ti-composite, Al-composite and Al-Cu alloy; the Al-Cu alloy has the best properties followed by GCI. The material properties considered are compressive strength, friction coefficient, wear resistance, thermal capacity and specific gravity. Since Al-Cu alloy has a higher cost, GCI is finally selected for our study due to economic factors and yet reasonable material properties.

Thermal analysis: This is done to ensure that the temperature of the rotor does not go above its melting point. Heat flux is calculated by converting the linear kinetic energy of the vehicle and the rotational kinetic energy of the wheels into heat energy. This heat energy is then divided by

the net disc-pad interface area and the time required to stop the vehicle. This time is obtained by using the parameters having the maximum value given in Table II to simulate for the 'worst case scenario'. Hence the total heat flux is obtained (597.5 kWm^{-2}) and entered into the software along with the initial temperature as 22°C.

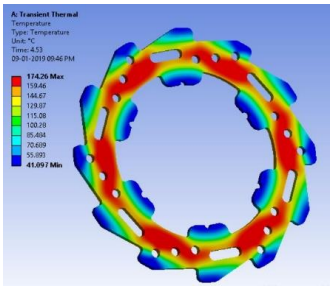


FIGURE 8(B)
TEMPERATURE DISTRIBUTION ACROSS THE BRAKE DISC

As shown in figure 8(b), the maximum temperature of the disc reaches 174.3°C during the stopping time i.e. 4.53 seconds. This temperature is below the melting point of the rotor. The FE model of only the outer disc is generated as the inner disc has been considered as a heat sink. This would give a slightly larger temperature than the actual one.

Fatigue analysis: As shown in figure 8(b), the maximum temperature is very much less than the melting point of grey cast iron (GCI). It is due to this reason that there is no need to perform the creep analysis. Moreover the compressive forces by pads will not lead to fatigue failure.

(j) Final assembly of the braking system

All the above selected and designed components of the braking system are finally assembled in order to form an effective 'Y-type' dual split hydraulic braking system. Figure 9 shows the entire braking circuit. A total of eight simple unions, four tee unions, one cross union, two proportioning valves, four PS-1 calipers, four brake discs (same design), eleven separate brake hoses, two reservoirs and one master cylinder are used, as depicted in figure 9.

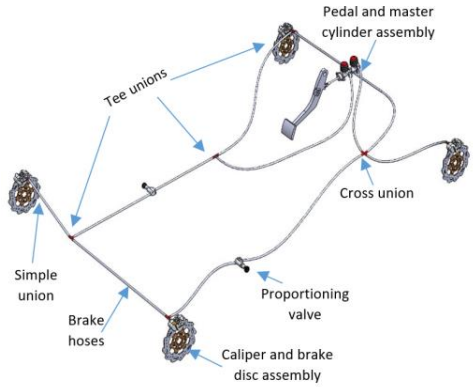


FIGURE 9
FULLY ASSEMBLED Y-TYPE BRAKING SYSTEM

IV. Validation of the braking circuit

The validity of the braking system formed by assembling the selected/designed components is tested for successful implementation in the above mentioned five ATVs.

The aim of the following calculative approach is for checking the validity or compatibility of the braking system with the selected ATVs in terms of performance and vehicle safety during dynamic conditions. This was done by:

- Ensuring compatibility with CMV (central motor vehicle) rules of the Indian government.
- Ensuring that the front and rear braking torques are greater than the respective front and rear frictional torques. They should be greater by a marginal amount, so that the wheels get locked at the required right time (not too soon for avoiding skidding and not too late).

TABLE II
LIST OF ATVS AND THEIR PARAMETERS [16] [6]

Parameters	ATV-1	ATV-2	ATV-3	ATV-4	ATV-5
x	0.4	0.4	0.4	0.45	0.4
y	0.6	0.6	0.6	0.55	0.6
W (N)	4498.2	4690	4429.6	4501	5311.6
h (m)	0.6985	0.762	0.6985	0.6985	0.7366
L (m)	1.5621	1.8288	1.5621	1.8288	1.651
r_f (m)	0.3048	0.3175	0.3175	0.3175	0.2413
r_r (m)	0.3048	0.3175	0.3175	0.3175	0.254
GC (m)	0.254	0.2667	0.2604	0.2667	0.2667
V (ms^{-1})	15.278	15.278	16.67	13.89	15.278

Physical quantities and their symbols to be used for the vehicle dynamics equations for system validation are as follows:

x: portion of weight at the front axle of the vehicle.
y: portion of weight at the front axle of the vehicle.
W: weight of the vehicle including the driver
h: height of the CG of vehicle from wheelbase
L: wheelbase of the vehicle
r_f, r_r: radius of front and rear tires respectively.
GC: ground clearance (used for checking with the rotor diameter)
V: velocity at which the vehicle brakes
μ: coefficient of friction between road and tire
μ= 0.7 [1]
Braking system parameters are as follows:
g = 9.8ms^{-2}
PR: pedal ratio = 6:1
d_{mc}: Diameter of master cylinder piston = 15.875mm = 0.015875m [8]
A_{mc}= Area of master cylinder piston

$$A_{mc} = \pi\, d_{mc}^2/4 \quad (2)$$

PE: Pedal effort of the driver
The values of the Knee Angle **(β)** and the Thigh Angle **(α)** are measured according to the vehicle. Figure 10 shows the

ninety fifth percentile male data chart from NASA [17] that gives the approximate value of the pedal effort. Hence PE equals to 850N.

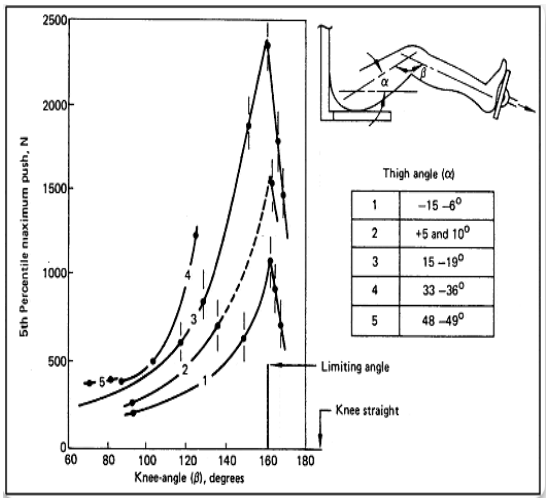


FIGURE 10
NINETY FIFTH PERCENTILE MALE DATA CHART FROM NASA [17]

F_P: Pushrod force

$$F_P = PR*PE \quad (3)$$

Where PR mechanical advantage by pedal or the pedal ratio

$$P = F_P/A_{mc} \quad (4)$$

F_P/A_{mc}= P
The above equation depicts Pascal's law.
P: brake line pressure
A_{cal}: Net caliper piston area
A_{cal}= 0.79in^2 = 0.00051m^2 (PS-1 properties) [8]
F_{cal}: Caliper clamping force

$$F_{cal\text{-}front} = P*A_{cal} \text{ [18]} \quad (5)$$

$$F_{cal\text{-}rear} = P*BR*A_{cal} \text{ [18]} \quad (6)$$

The above two equations are for front and rear calipers respectively.

BR is biasing ratio. Biasing ratio is the ratio between the brake line pressure towards the rear wheels and the front wheels. Since the front wheels require more braking force than the rear (due to dynamic load shift during braking towards front), the biasing ratio is introduced.

BR = 43/100 = 0.43 is taken since it is the maximum biasing limit of the selected proportioning valve [8].
μ_P : coefficient of friction between pad and brake disc
μ_P= 0.3 [1]
T_{bf}: braking torque at the front axle
T_{br}: braking torque at the rear axle

$$T_{bf} = 2*F_{cal\text{-}front}*r_{eff}*\mu_P \quad (7)$$

$$T_{br} = 2*F_{cal\text{-}rear}*r_{eff}*\mu_P \quad (8)$$

F_{bf}: braking force at the front wheels
F_{br}: braking force at the rear wheels

$$F_{bf} = T_{bf}/r_f \quad (9)$$

$$F_{br} = T_{br}/r_r \quad (10)$$

F_b: total braking force

$$F_b = F_{br} + F_{bf} + F_a \text{ [13]} \quad (11)$$

Where F_a is the aerodynamic drag force.

$$F_a = D_a V^2 \quad (12)$$

D_a: aerodynamic drag coefficient
D_a= 0.35 referring to vehicle parameters in [19].
d: deceleration of the vehicle

$$d = F_b*g/W \text{ [13]} \quad (13)$$

SD: stopping distance of the vehicle

$$SD = V^2/2d \quad (14)$$

In the above equation, time required for pedal travel and hydraulic pressure build up is neglected. The conformity of the stopping distance and the deceleration of the vehicle with the CMV (central motor vehicle) rules of the Indian government are also checked by relations given below:

$$SD_{max} = 0.15*V_o + V_o^2/130 \text{ [20]} \quad (15)$$

$$V_o = 0.8*V_{max}*5/18 \quad (16)$$

This is the maximum permitted stopping distance for a hydrostatic braking vehicle under the CMV Rules [20]. Hence if SD < SD_{max} and if d>d_{min}; then the braking design is compatible with the vehicle with respect to the CMV rules.

$$d_{min} = V^2/2SD_{max} \quad (17)$$

W_d: dynamic load shift while braking.

$$W_d = (d/g)*(Wh/L) \text{ [21]} \quad (18)$$

Value of d is determined after brake torque calculations.

$$W_f = W*x + W_d \quad (19)$$

$$W_r = W*y - W_d \quad (20)$$

The above equations compute the net normal loads on front and rear axles respectively while braking.
Calculating the frictional forces:

$$F_f = W_f * \mu \quad (21)$$

$$F_r = W_r * \mu \quad (22)$$

Frictional torques at respective axles:

$$T_f = F_f * r_f \quad (23)$$

$$T_r = F_r * r_r \quad (24)$$

If $T_f < T_{bf}$ and $T_r < T_{br}$; then the compatibility of the developed braking system with the vehicle requirements is ensured since the braking torques are greater than the frictional torques of the vehicle by a marginal amount for the wheels to lock.

Computed Constants		
μ	0.7	
g	9.8	m/s2
PE	850	N
PR	6	
F_p	5100	N
d_{mc}	15.875	mm
P	25.78	MPa
A_{cal}	509.6764	mm2
F_{cal}	13139.15	N
μ_p	0.3	
r_{eff}	0.1016	m
T_{bf}	800.955	Nm
BR	0.43	
D_a	0.35	
T_{br}	344.4107	Nm

FIGURE 11(A)
LIST OF BRAKING SYSTEM PARAMETERS

Computed Variables	ATV-1	ATV-2	ATV-3	ATV-4	ATV-5
F_{bf} (N)	2627.805	2522.693	2522.693	2522.693	3319.333
F_{br} (N)	1129.956	1084.758	1084.758	1084.758	1355.947
d (m/s2)	8.365	7.709	8.196	8.002	8.777
T_f(Nm)	750.197	758.571	761.966	762.115	717.358
T_r(Nm)	209.539	283.781	222.513	238.233	189.289
SD (m)	13.952	15.140	16.952	12.056	13.298
SD_{max} (m)	21.49305	21.49305	24.93186	18.31032	21.49305
d_{min} (m/s2)	5.430064	5.430064	5.572967	5.268398	5.430064

FIGURE 11(B)
LIST OF COMPUTED VARIABLES ACCORDING TO DIFFERENT ATVS

Data for Chart					
ATVs	ATV-1	ATV-2	ATV-3	ATV-4	ATV-5
T_f(Nm)	750.197	758.571	761.966	762.115	717.358
T_r(Nm)	209.539	283.781	222.513	238.233	189.289
T_{bf}(Nm)	800.955	800.955	800.955	800.955	800.955
T_{br}(Nm)	344.4107	344.4107	344.4107	344.4107	344.4107

FIGURE 11(C)
COMPARISON OF FRICTIONAL AND BRAKING TORQUE VALUES FOR BOTH THE AXLES

These calculations are solved in Microsoft Excel for the five selected ATVs, shown above. Figure 12 shows the marginal difference between the frictional and the braking torques with respect to both the axles. The red and the green lines show the braking torques of the braking system that remain constant throughout the five ATVs. The blue and the yellow lines show the frictional torques for both the axles for the five ATVs.

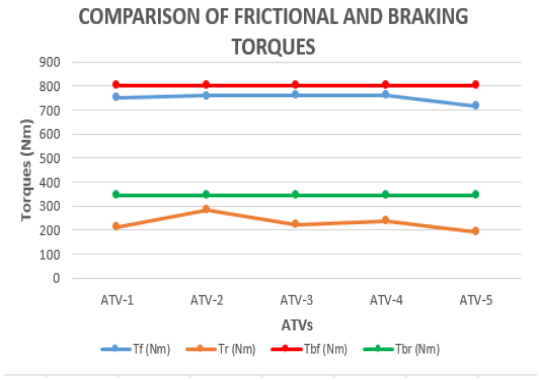


FIGURE 12
GRAPHICAL COMPARISON BETWEEN FRICTIONAL AND BRAKING TORQUES

Moreover, from figure 11(b) it is inferred that the stopping distance and the computed decelerations of the respective vehicles equipped with the new braking system fall within the limitations set by the CMV rules [20].

V. Limitations of the Braking System

The following are the projected limitations of the above designed braking system:

- The selected components will only work on the five ATVs and other ATVs with similar specifications. For heavier ATVs with larger tires and greater top speed, calipers with higher clamping forces can be used. Eg- Wilwood's DH4 dual inlet caliper [8] has two inlet ports that could eliminate the need of Tee unions attached to the caliper. In order to balance the braking force among wheels on the same axle, separate calipers would be used for those wheels where only one inlet pipe is required. The same strategies can be used to select/design the braking components for ATVs with different specifications, if not the same braking components.
- The 'Y-type' braking system involves the use of a larger number of components and a lot of hoses.
- Factors such as rolling resistance of the wheels, pressure build up time towards the caliper, pedal travel time, road gradient, pressure losses inside braking components, coefficient of friction on wet roads have been neglected in the vehicle dynamics calculations.

CONCLUSION

The components of the braking system are finally selected/designed and then assembled in order to form an effective 'Y-type' dual split hydraulic braking system. The designs of these components ware later analysed with the help of FEA using Ansys 17.1® for the dynamic conditions and prolonged application they would be subjected to while the vehicle operates. The compatibility of this braking system is ensured with the five selected ATVs (and similar such vehicles) by following a detailed calculative approach and checking the system's conformability with the Indian government CMV regulations for the dynamic safety of the vehicle. Moreover, the circuit design for the ATVs that do not fall into the above category is also discussed briefly. Overall it can be concluded that the selected five ATVs and vehicles with similar specifications can be fitted with the designed 'Y-type' braking system that would offer a safer braking in case of a single brake circuit failure by locking three wheels at a time.

ACKNOWLEDGMENT

The success of any research project stems from the support received by various well-wishers besides the efforts made by the authors. Apart from the unparalleled dedication portrayed by us, we would like to thank our colleagues, friends and families for their encouragement that had a vital involvement towards motivating us to strive for higher benchmarks within this current project.

REFERENCES

[1] Jain, P.K., Pankaj, Natu, A. et al., "Design, Development and Fabrication of a Go-kart". *International Journal of Advance Research and Innovation (IJARI)*, Vol.6, Issue 2, 2018, pp. 145-161, ISSN 2347–3258.

[2] Sharma, A., Amrute, P., Thakur, SS. et al., "Design, Analysis and Fabrication of Braking System with Rear Inboard Brakes in Baja ATV". *International Research Journal of Engineering and Technology (IRJET)*, e-ISSN: 2395-0056, Vol.5, Issue-05, 2018.

[3] Mishra, D., Nagubandi, R. et al., "Team Jaabaz Baja SAE (UTEP 2014) Design Report", *SAE International*, 2014.

[4] Shrivastava, D., "Designing of All Terrain Vehicle (ATV)". *International Journal of Scientific and Research Publications (IJSRP)*, Vol.4, Issue-12, 2014, pp. 1, ISSN: 2250-3153.

[5] Gupta, U.S., Chandak, S., Dixit, D. et al., "Design & Manufacturing of All Terrain Vehicle (ATV)- Selection, Modification, Static & Dynamic Analysis of ATV Vehicle". *International Journal of Engineering Trends and Technology (IJETT)*, Vol.20, No.3, 2015, ISSN: 2231-5381.

[6] Polaris off road vehicles [Internet]. Minnesota: Polaris; [cited 2018 Dec; updated 2020 Nov]. Available from: https://offroad.polaris.com/en-us/.

[7] Jain, P., Garani, H., "Analysis and Assessment of Dual Brake Circuits". *International Journal of Mechanical Engineering and Technology (IJMET)*, Vol.7, Issue-5, pp. 407–416, 2016, ISSN: 0976-6359.

[8] Wilwood disc brakes [Internet]. California: Wilwood; [cited 2018 Dec; updated 2020 Nov]. Available from: https://www.wilwood.com/.

[9] Verdone, P., "Chapter-7, Basic Hydraulic Systems- Peter Verdone Designs", 2013, pp. 64-65. Located at: https://www.peterverdone.com/archive/files/hydraulic%20system%20theory.pdf

[10] Shigley, J.E., Mischke, C.R., "Standard Handbook of Machine Design", *McGraw-Hill*, second edition, 1996, ISBN: 0-07-056958-4.

[11] Gokhale, N.S., Deshpande, S.S., Bedekar, S.V. et al., "Practical Finite Element Analysis", *Finite to Infinite*, first edition, 2018, ISBN: 978-81-906195-0-9.

[12] Owen, C.E., "Today's Technician: Automotive Brake Systems Classroom and Shop Manual", *Cengage learning*, fifth edition, 2010, pp. 27–28, ISBN: 978-1-4354-8655-3.

[13] Gillespie, T.D., Fundamentals of Vehicle Dynamics, USA: Society of Automotive Engineers (SAE), 1992, ISBN: 978-1-56091-199-9.

[14] Limpert, R., "Brake Design and Safety", *Society of Automotive Engineers (SAE)*, Second Edition, 1999.

[15] Maleque, M.A., Dyuti, S., Rahman, M.M. et al., "Material Selection Method in Design of Automotive Brake Disc", *Proceedings of the World Congress on Engineering*, Vol.3, 2010, ISSN: 2078-0966.

[16] ATV.com [Internet]. USA: atv.com; [cited 2018 Dec; updated 2020 Nov]. Available from: https://www.atv.com/.

[17] NASA website section-4, 95th percentile male data [Internet]. USA: NASA; [cited 2018 Dec]. Available from: https://msis.jsc.nasa.gov/sections/section04.htm

[18] Walker, J., "Technical whitepaper: the physics of braking systems". *Stoptech high performance braking systems*, 2018, pp. 11. Located at: https://www.apcautotech.com/getmedia/89aa6773-73d7-4f7f-b935-5e6bf2d28111/Centric_and_APC_Technical_Whitepaper_A1-The-Physics-of-Braking-Systems-8-2018_1.pdf

[19] Reif, K., "Brakes Brake Controls and Driver Assistance Systems Function, Regulations and Components", *Bosch professional automotive information*, 2014, ISBN: 978-3-658-03977-6.

[20] The Central Motor Vehicles Rules, 1989 [Internet]. India: Ministry of road transport and highways. Available from: http://chdtransport.gov.in/Forms/CMVR_1989.pdf

[21] Stone, R., Ball, J.K., "Automotive Engineering Fundamentals", *SAE (Society of automotive engineers)*, 2002, ISBN: 0-7680-0987-1.

Intelligent Service Robotics
https://doi.org/10.1007/s11370-021-00384-5

ORIGINAL RESEARCH PAPER

Design and development of a reconfigurable tiling robot

Rizuwana Parween[1] · Madan Mohan Rayaguru[2] · Shi Yuyao[1] · Borusu Charan Satya Chandra Sairam[1] · Mohan Rajesh Elara[1]

Received: 22 March 2021 / Accepted: 17 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
This paper presents a novel polyromb-based self-reconfigurable domestic floor cleaning robot called "hRhombo" that consists of four rhombus-shaped modules connected with three revolute joints. The mechanical design of the robot gives rise to an over-actuated mathematical model. The mechanical design, electronics layout, kinematics relations, reconfiguration theory and trajectory tracking methodology for the robot are discussed. The forward kinematics and inverse kinematics are derived in terms of the angle between the modules and the wheel steering angles. The tracking controller exploits a poly-cycle approach, where each wheel has a different reference path depending on the current configuration of the robot. The model predictive controller technique is used to design optimal controller for four unicycles, modeling the four wheels. The numerical simulations and experiments reveal the usefulness of the robot in the domestic environment with complex furniture settings.

1 Introduction

In recent years, research on application of autonomous devices is increasing over time. Instead of manually handling the time-intensive, laborious and repeating tasks, making use of autonomous devices is a more efficient choice. One of the problems faced by various industries is the area coverage problem. To tackle this specific area coverage problem, a lot of robots are developed based on coverage path planning. These robots are used in industries such as cleaning [1,2], inspection [3], photogrammetry [4] as well as antenna development [5].

Especially under the category of cleaning robot, area coverage is the prioritized goal due to the nature of this industry. The robots are expected to be deployed in an unknown workspace and navigate with maximum area coverage and minimum overlap. Massive researches and development works have been done, which leads to various commercial products from iRoomba, Samsung, Xiaomi, Honda Miimo, Aquabot, etc. According to [6,7], the process of the coverage path planning requires the robot to be aware of the boundary information of the environment in order to generate the map. After that, the traditional planning strategy is based on a fixed pattern, for example, spiral pattern [8] or zigzag pattern [9]. However, these strategies only work for known fixed situations and fall short while encountering unknown situations. As the real working scenario is often more complex than predefined situations, methods based on artificial intelligence are brought into play. For instance, M. A. Yakoubi and M.T. Laskri proposed a path planning method using genetic algorithms [10]. Although the method takes into consideration of dynamic obstacles and minimizes the distance travelling to increase efficiency, it does not account for the geometry of the robot which may lead to excess energy usage or lower area coverage under the real working scenario. P. C. Ku suggested a novel approach using graph-theory [11]. Moreover, as a majority of them are based on the grid information [10,12–14], more computational cost is required and resulted in slower execution speed.

Nevertheless, the geometric designs of cleaning robots can result in low area coverage as well. The most common shapes of the cleaning robots on market are circular, D-shape and

✉ Rizuwana Parween
rizuwana_parween@sutd.edu.sg

Madan Mohan Rayaguru
mmrayguru@dtu.ac.in

Shi Yuyao
yuyao_shi@sutd.edu.sg

Borusu Charan Satya Chandra Sairam
sairambcsc007@gmail.com

Mohan Rajesh Elara
rajeshelara@sutd.edu.sg

[1] Singapore University of Technology and Design, 8 Somapah Road, Singapore 487372, Singapore

[2] Electrical Engineering Department, Delhi Technological University, Delhi 110042, India

Published online: 09 September 2021

square shape. Each of them has its own advantages and also limitations due to the nature of their fixed morphology. This results in difficulties accessing narrow spaces, room corners or even around the legs of furniture. One of the possible solutions that have been brought up these years is to design the cleaning robot with self-reconfigurability. It is proven that with reconfigurable modules, the tiling robots are able to access the narrow spaces and show higher area coverage performance [1,15–22].

Apart from a suitable area coverage property, the reconfigurable cleaning robots should be able to follow the desired path for completing a specific cleaning task. Consequently, the path following controller design problem for cleaning robots is an essential aspect of their functionality. In general, the mathematical model for the motion of the reconfigurable mobile robots can vary depending on the steering mechanism. So far, the unicycle and differential drive mechanisms have been exploited in [19,22,23] for achieving the desired movements. The controller design techniques for the over-actuated mobile robots are discussed in [24]. Recently, the authors of [23] proposed an ICR-based differential drive method to control a tiling platform called hTetro. However, the stability and convergence aspects of the closed-loop dynamics are not analyzed. Paper [25] exploited a robust backstepping controller for a differential drive reconfigurable robot h-Trihex, but does not consider the actuator constraints. If the actuator saturation is not considered during the controller design, it may give rise to performance deterioration and instability [26]. Model predictive control (MPC)-based controller is suitable for dealing with the state and actuator constraints [27,28], but its implementation in the reconfigurable robots is not straightforward.

From the above discussions, it is apparent that a reconfigurable cleaning robot with a different form factor may give better accessibility and area coverage. Further, an optimal path following controller design while considering the actuator and state constraints can simultaneously minimize battery consumption and tracking error. These two issues are of significant practical interests, and finding their solutions is the primary motivation behind this paper. The major contributions of the paper are the following:

This paper presents a novel polyromb-based self-reconfigurable robot called "hRhombo" that consists of four rhombus-shaped modules connected with three revolute joints. The major contributions of the paper are as follows:

- The proposed robot design is capable of changing its morphology to any of the seven different configurations with an objective of maximizing the area coverage in an unstructured environment.
- The configurations formed by the hRhombo platform has non-prismatic shape with larger variations in the included angles, compared to [20,23]. The robot can access the narrow spaces with an included angle of less than 90 degree and irregular spaces with included angles more than 90 degrees. This feature helps the platform to adapt and clean the narrow and angular space.
- Additionally, the forward kinematics and inverse kinematics of the hRhombo are derived in terms of shape variables, and the resulting mathematical model is found to be over-actuated. The proposed controller design exploits the system's geometric parameters to decouple the platform's path tracking problem into four separate unicycle tracking problems. The control inputs for each module are computed by following the MPC strategy, which minimizes the tracking error while considering actuator limitations. The experimental results demonstrate the efficacy of the proposed scheme.

This paper outlines the following structure: Sect. 2 describes the polyromb tiling theorem as a path planning strategy for autonomous navigation. Section 3 presents the platform architecture with mechanical design and electronics layout. Sections 4 and 5 outline the kinematic modeling and controller design for tracking, respectively. Sections 6, 7 and 8 present the simulation results, experimental validation and conclusion, respectively.

2 Polyromb-based planar configuration

The hRhombo platform is designed for floor cleaning application in complex and unstructured domestic setting. It applies Polyromb-based tiling theorems for area coverage. A set of four rhombi, with each rhombus, has two 60 degree and two 120 degree angles, called tetrarhombi. A tetrarhombi adapts seven different configurations (forms) including 'I, D, L, S, Z, T, J,' as shown in Fig. 1. The different orientations of the tetrarhombi during clockwise and anticlockwise rotation are shown in Fig. 2. hRhombo platform is designed based on the different configurations of tetrarhombi.

3 hRhombo platform architecture

3.1 Mechanical design

The concept of self configurability is applied in developing a floor cleaning robot, called hRhombo. The platform consists of four blocks in rhombus shape ($block_i, i = 1, 2, 3$ and 4), and each block is connected to its adjacent block via an active hinge joint. Figure 3(a) shows the schematic representation of hRhombo platform with three hinged joints ($joint_i = s_i, i = 1, 2,$ and 3). Each hinge joint consists of a mechanical revolute joint with a servo motor attached to it. The servo motor provides planar rotation with one degree

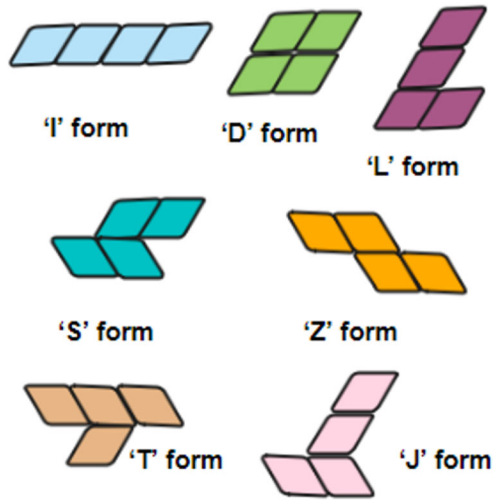


Fig. 1 Various configurations of the tetrarhombi

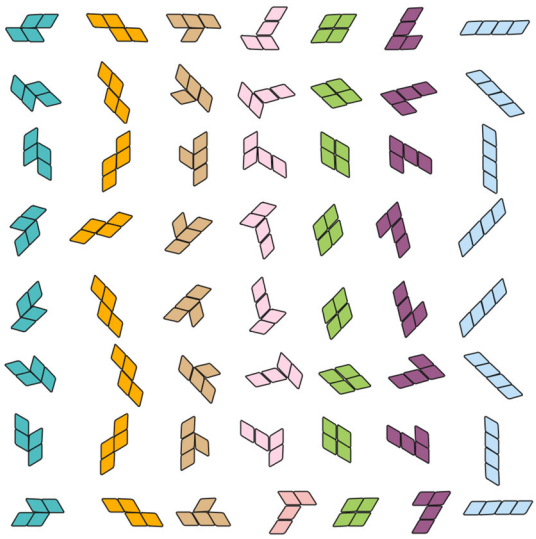


Fig. 2 Various forms of the tetrarhombi during clockwise and anti-clockwise rotation

of freedom to block during reconfiguration. The servo motor is a Herkulex with voltage 7.4V, current 670 mA and 2.2 kgcm torque capability. For the locomotion, each module is designed to have its drive module, as shown in Fig. 3(b). The drive unit consists of a steering module and wheel module shown in Fig. 3(c). The wheel module consists of a conventional fixed wheel with diameter 3.5 cm connected to a Pololu micrometal gear motor of 6 volts and 220 rpm. This wheel module is placed inside a steering module which is powered by a servo motor with specification of 670 mA, 7.4 V and 2.2 kgcm torque capacity. The servo motors are able to control the steering angle of each module, and by synchronizing the steering angle of each module, the platform can maneuver along a certain trajectory. The locomotion module of each block is controlled by a RoboClaw. The kinematics of anal-

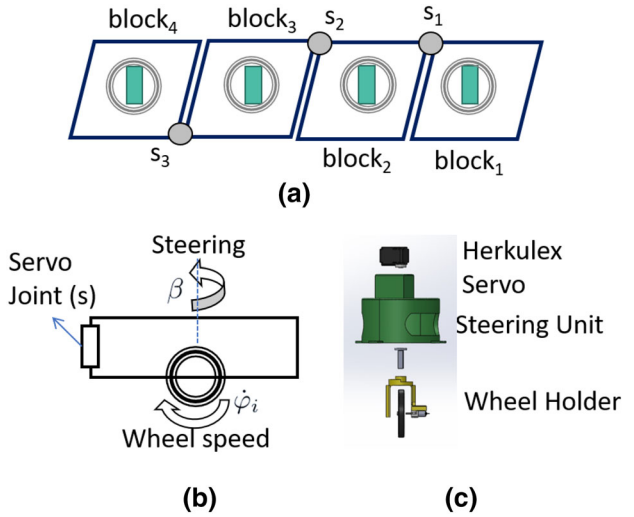


Fig. 3 **a** The schematic layout of the hRhombo Platform with joints and wheel at the center, **b** schematic of the wheel module, **c** components of the wheel modules

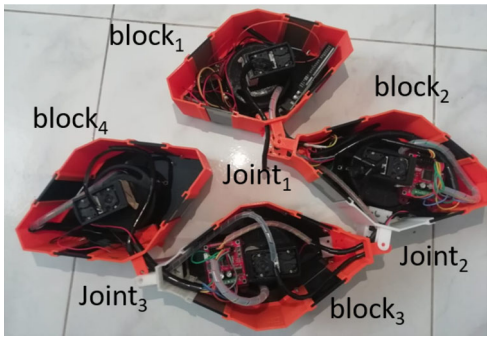


Fig. 4 The fabricated prototype of the platform

ysis of the platform will be discussed in the next section. Figure 4 shows the fabricated prototype of the platform. All the side walls of the blocks, the steering unit and hinges are fabricated using a 3D printer with PLA. The base of each block is fabricated using an acrylic plate. Under hinged dissection theorem, with four blocks modules, the robot can transform itself into seven unique configurations (I, D, L, S, Z, T and J forms) as shown in Fig. 5.

3.2 Electronics module

Figure 6 shows the electronics block diagram of hRhombo platform consisting of various hardware, drivers and microcontroller unit. An Arduino Atmega (Arduino Mega 2560) is used to process the algorithms developed for controlling various tasks including the locomotion, reconfiguration and sensing. A total of seven high-torque servo motors (Herkulex 0201), four dc motors with magnetic encoders, two motor drivers (RoboClaw $2x7A$), one current sensor (ACS715 current sensor), one $7.4v$ voltage regulator ($Pololu 7.5V, 2.4A$)



Fig. 5 Platform showing different configurations

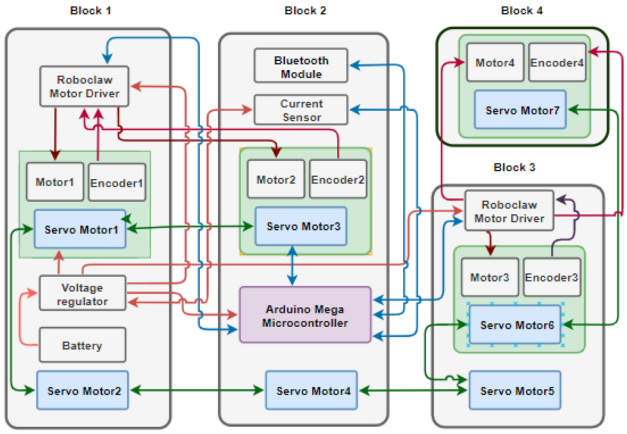


Fig. 6 The electronics layout of the hRhombo Platform

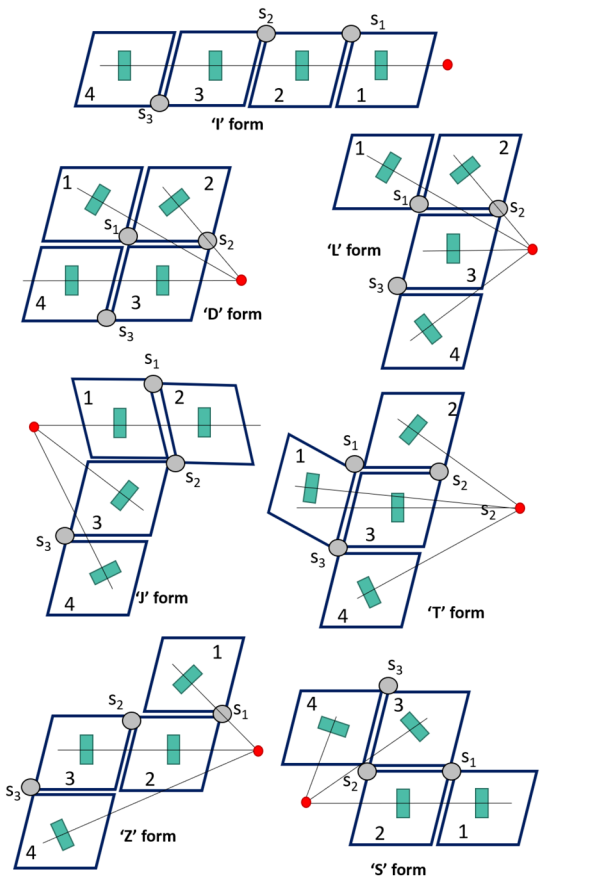


Fig. 7 Wheel steering for ICR for different configurations of hRhombo

and one Bluetooth module ($HC-06$) are incorporated in the platform. Two RoboClaw motor drivers are used to get the encoder reading and control the speed and direction of the motors. A battery of voltage 7.4V and current 4.5A is used to power the circuit, and it is connected to the voltage regulator to regulate the voltage. A current sensor is connected across the load to measure the current consumption during reconfiguration. A Bluetooth module is connected to the Arduino to control the robot wirelessly.

3.3 Autonomy module

Raspberry pi and a Marvelmind beacon system for localization was set up to check the position accuracy of the hRhombo on predefined path. Four stationary beacons were placed at the four corners of the working area, and a mobile beacon was connected to the raspberry pi and placed on the robot. The stationary beacons were placed in line of sight to each other stationary beacon in workstation to send signals to mark the position of other stationary beacons to make a virtual map in Marvelmind software, and the mobile beacon position can be tracked with respect to stationary beacons. A ROS node was created to publish the data, and python script was written in the ROS platform to collect the position data of rhombus but locomotion. The data received contain the planar coordinates of hRhombo with respect to the stationary beacon map.

4 Kinematic modeling

In Fig.7, $\mathcal{F}_I = (o_I, \boldsymbol{X_I}, \boldsymbol{Y_I}, \boldsymbol{Z_I})$ is the world frame, while $\mathcal{F}_b = (o_b, \boldsymbol{X_b}, Y_b, \boldsymbol{Z_b})$ is the robot frame, of which the origin is located at the geometric center of the current shape. Defining $\boldsymbol{\xi}_I = [x_i \quad y_i \quad \theta]^T$ as the position and orientation vector of the robot under the world frame and $\boldsymbol{\xi}_d = [x_d \quad y_d \quad \theta_d]^T$ to be the desired position and orientation.

For each configuration, the wheels have to be steered in a particular orientation such that each configuration has an instantaneous center of rotation (ICR). Figure 7 illustrates the ICR layout for each configuration.

In this section, the kinematic modeling of the hRhombo platform will be presented in detail. There are two assumptions during the modeling which are 1) the shape of the robot stays unchanged during the locomotion and 2) there are no slipping and no skidding. With these two assumptions,

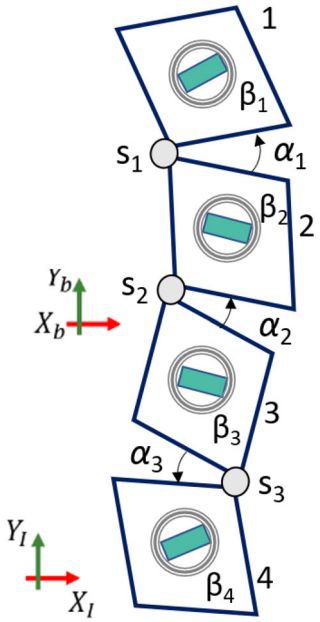


Fig. 8 Schematic diagram of kinematic diagram of the hRhombo

the robot is considered as a rigid body during locomotion. However, as the shape of the robot still changes while transformation happens, the orientation of each individual chassis changes as well. To simplify this issue, the orientation of the robot is defined as the orientation of the block 2 chassis. The position and velocity of the robot are defined to be that of the geometric center of the robot from the top view. Figure 8 shows the schematic diagram of hRhombo showing the included angle.

The steering angle of each locomotion module with respect to its chassis is represented using $\beta_i, i \in \{1, 2, 3, 4\}$. $\boldsymbol{\phi} = [\phi_1 \ \ \phi_2 \ \ \phi_3 \ \ \phi_4]^T$ is the actuation vector of each wheel. As the shape of the robot changes, $\alpha_j, j \in \{1, 2, 3\}$ describes the angle at each hinge. By separating the velocity of each module into the tangential velocity, $\boldsymbol{v}_t = [v_{t1} \ \ v_{t2} \ \ v_{t3} \ \ v_{t4}]^T$, and normal velocity, $\boldsymbol{v}_n = [v_{n1} \ \ v_{n2} \ \ v_{n3} \ \ v_{n4}]^T$ at the ground contact point of the driving wheel, the linear velocity of each module can be derived as:

$$\boldsymbol{v}_t = \mathbf{F}(\sigma_i)\mathbf{R}^{-1}(\theta)\dot{\boldsymbol{\xi}}_I - r_w\dot{\boldsymbol{\phi}} \tag{1}$$

$$\boldsymbol{v}_n = \mathbf{G}(\sigma_i)\mathbf{R}^{-1}(\theta)\dot{\boldsymbol{\xi}} \tag{2}$$

where $\mathbf{F}(\sigma_i) =$

$$\begin{bmatrix} -\sin(\sigma_1) & \cos(\sigma_1) & y_1^c \sin(\sigma_1) + x_1^c \cos(\sigma_1) \\ -\sin(\sigma_2) & \cos(\sigma_2) & y_1^c \sin(\sigma_2) + x_1^c \cos(\sigma_2) \\ -\sin(\sigma_3) & \cos(\sigma_3) & y_1^c \sin(\sigma_3) + x_1^c \cos(\sigma_3) \\ -\sin(\sigma_4) & \cos(\sigma_4) & y_1^c \sin(\sigma_4) + x_1^c \cos(\sigma_4) \end{bmatrix}$$

and $\mathbf{G}(\sigma_i) =$

$$\begin{bmatrix} -\cos(\sigma_1) & -\sin(\sigma_1) & -x_1^c \sin(\sigma_1) + y_1^c \cos(\sigma_1) \\ -\cos(\sigma_2) & -\sin(\sigma_2) & -x_1^c \sin(\sigma_2) + y_1^c \cos(\sigma_2) \\ -\cos(\sigma_3) & -\sin(\sigma_3) & -x_1^c \sin(\sigma_3) + y_1^c \cos(\sigma_3) \\ -\cos(\sigma_4) & -\sin(\sigma_4) & -x_1^c \sin(\sigma_4) + y_1^c \cos(\sigma_4) \end{bmatrix}.$$

σ_i is defined to be the combination of β_i and α_i with the relationship shown as:

$$\sigma_i = \begin{cases} \alpha_1 + \beta_1, & i = 1 \\ \beta_2, & i = 2 \\ \alpha_2 + \beta_3, & i = 3 \\ \alpha_2 + \alpha_3 + \beta_4, & i = 4. \end{cases}$$

Since the assumption of the modeling is non-slipping and non-skidding conditions, $\boldsymbol{v}_t$ and $\boldsymbol{v}_n$ are set to be zero, resulting in the inverse kinematics of the robot to be:

$$\dot{\boldsymbol{\phi}} = \frac{1}{r_w}\mathbf{F}(\sigma_i)\mathbf{R}^{-1}(\theta)\dot{\boldsymbol{\xi}}. \tag{3}$$

By taking the Moore–Penrose pseudo-inverse of the matrix $\boldsymbol{\omega}_\mathbf{i}$, the forward kinematics of the robot can be written as:

$$\dot{\boldsymbol{\xi}} = r_w\mathbf{R}(\theta)\mathbf{F}^{+}(\sigma_i)\dot{\boldsymbol{\phi}} \tag{4}$$

where $\mathbf{F}^{+}(\sigma_i) = \mathbf{F}^T(\sigma_i)\mathbf{F}(\sigma_i)\mathbf{F}^T(\sigma_i)$.

5 Tracking controller

The controller has to be designed in such a way that the center of mass of the platform follows a reference trajectory (x_d, y_d), with a desired heading (ϕ_d). The desired trajectories are decided by the conventional path planning algorithm. The kinematics model derived in the previous section constitutes an over-actuated system. As mentioned in "Introduction," the control allocation strategies are complex and the ICR-based method lacks stability assurance. We choose to apply multi-cycle approach [29], which transforms the tracking problem of the robot into four independent tracking problems of unicycles.

For the purpose, the desired trajectories for the wheels (x_{id}, y_{id}) in the four modules have to be derived for different configurations. Figure 9 shows this correlation using simple geometric calculations in two different configurations (D-form and L-form). The desired heading angle for each configuration is given by

$$\beta_{\mathrm{id}} = \frac{\dot{y}_{\mathrm{id}}}{\dot{x}_{\mathrm{id}}}.$$

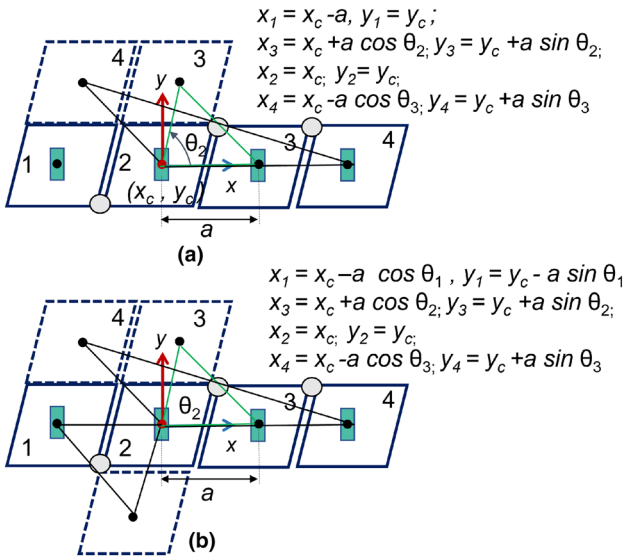


Fig. 9 Geometrical relation for wheels in two different configurations: **a** D-form, **b** L-form

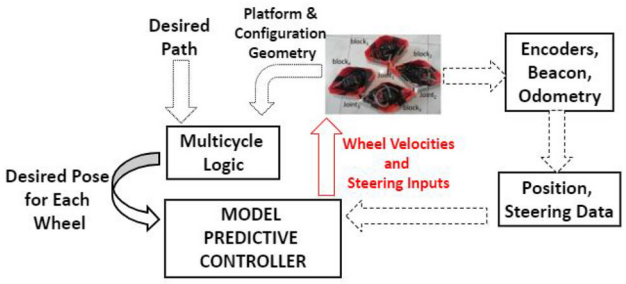


Fig. 10 Block diagram of the tracking controller

So the problem boils down to the trajectory tracking problem for the four wheels in respective modules of the platform. Figure 10 shows the block diagram of the tracking controller for the platform.

5.1 Model predictive control (MPC)

This subsection discusses one the most popular technique used in many industrial applications. The MPC design is chosen for its desirable qualities to handle MIMO systems, actuator constraints, state restrictions, etc. The unicycle model for each wheel in discrete time can be written as:

$$\begin{aligned} x_i(k+1) &= x_i(k) + v_i(k)\cos(\beta_i(k))T \\ y_i(k+1) &= y_i(k) + v_i(k)\sin(\beta_i(k))T \\ \beta_i(k+1) &= \beta_i(k) + \omega_i(k)T \end{aligned} \tag{5}$$

where T is the sampling time and $v_i(k)$ (linear velocity of wheel) and $\omega_i(k)$ (rate of steering) are the control inputs for the kth instant. The desired signals can be represented in an implicit form as:

$$\begin{aligned} x_{\mathrm{id}}(k+1) &= x_{\mathrm{id}}(k) + v_{\mathrm{id}}(k)\cos(\beta_{\mathrm{id}}(k))T \\ y_{\mathrm{id}}(k+1) &= y_{\mathrm{id}}(k) + v_{\mathrm{id}}(k)\sin(\beta_{\mathrm{id}}(k))T \\ \beta_{\mathrm{id}}(k+1) &= \beta_{\mathrm{id}}(k) + \omega_{\mathrm{id}}(k)T. \end{aligned} \tag{6}$$

The optimization problem for nonlinear systems like (5) is difficult to solve analytically, while considering all the actuator and state constraints. MPC is one of the potent numerical techniques to solve such nonlinear optimization problem, which transform the infinite horizon computation into a repetitive finite horizon computation. In general, the MPC algorithm solves a finite horizon open-loop optimal problem at each instant k. The first tread is then chosen as the control input for the next instant (k+1). The window/horizon is shifted to (k+1), and the same finite horizon cost function is optimized with the new control input. Fundamentally, the MPC algorithm solves an open-loop optimal control problem at each instant k, where the feedback comes from the application of that control law at next instant.

For a fixed horizon N and a control input $u(k) = [v_i(k), \omega_i(k)]^T$, the predicted state $X(k+j|k) = [x_i(k+j|k), y_i(k+j|k), \beta_i(k+j|k)]^T$ can be obtained for each $j = 0 : N-1$ from (5). Define two error signals

$$\begin{aligned} e(k+j|k) &= x(k+j|k) - x_d(k+j|k), \\ e_u(k+j|k) &= u(k+j|k) - u_d(k+j|k). \end{aligned}$$

For two positive definite matrices Q, R of suitable dimensions, the finite horizon optimal control problem can be stated as:

$$\begin{aligned} \min_{X,u}\ & l(k) = \sum_{j=1}^{N} (e(k+j|k)^T Q e(k+j|k) \\ & + u_e^T(k+j|k)^T R u_e(k+j|k)) \\ \text{s.t } & X(k|k) = X(k),\ \ Hu(k+j|k) \le h \\ & \text{and } X(k+j+1|k) = F(X(k+j|k), u(k+j|k)) \end{aligned} \tag{7}$$

where $F(.)$ is the compact expression representing the R.H.S of (5), $H = [I, I]^T$, and $h = [u_{max}, -u_{max}]^T$. The optimization problem is solved at each instant to get the optimal control inputs $u^*(k+j|k)$, from which $u^*(k|k)$ is chosen as the feedback for (k+1)th instant, and other values are discarded.

Note: The problem of applying MPC directly to our setup is that the kinematics equation is over-actuated. The inverse kinematics is not straightforward to obtain without approximation. Nonetheless, the actuator constraints are critical, and MPC is a very good alternative to handle it (constraint optimization problem). Therefore, we propose a new approach where MPC can be applied in this situation. The paper used

a multicycle approach to convert the tracking problem in the robot to four separate unicycle tracking problems and then applied MPC. To the best of the author's knowledge, the constrained optimal tracking controller design for the reconfigurable robots has not been addressed prior to this paper.

5.2 Stability

The nonlinear optimization problem described above cannot provide global stability because the control inputs are saturated, and the system dynamics is nonlinear. However, the local stability can be assured inside certain judiciously chosen compact sets. The tracking problem can be converted into a regulation problem by the error dynamics, and hence, stability analysis is similar. To simplify the explanation, the derivation discusses the regulation problem.

Let A_x, B_x denote the linearized system matrices for (5) around the equilibrium (0,0), and let K_x denote the optimal control gain matrix for an equivalent LQR problem. Consider a set defined as:

$$Z = \{X \in R^n,\ -u_{\max} \leq K_x X \leq u_{\max}\}.$$

Choose a positive definite symmetric matrix P such that

$$(A_x - B_x K_x)^T P(A_x - B_x K_x) - (1+\kappa)P + Q + K_x^T R K_x \leq 0 \tag{8}$$

where κ is a positive scalar. The stability result can be summarized in the following theorem.

Theorem 1 *The MPC control law $u^*(k), k = 1, 2...$, is derived from solving the nonlinear optimization problem mentioned in (7), and (8) holds true for a positive definite matrix P. Then, there exists a positive scalar α_x such that the convex set $\Theta = \{X^T P X \leq \alpha_x\} \subset Z$ is positively invariant for the system $X(k+1) = F(X(k), u^*(k))$.*

Proof The matrix $A_x - B_x K_x$ is an Hurwitz matrix in a strict sense. So, for a sufficiently small κ inequality (8) is automatically satisfied. For a choice of convex set $\Theta_\alpha = \{X^T P X \leq \alpha_x\}$, such that $\Theta_\alpha \subset Z$, the constraints of the optimization problem are satisfied for $\alpha_x \in (0, \alpha]$. For the MPC control law $u^*(k)$, the discrete-time closed-loop system becomes

$$X(k+1) = F(X(k), u^*(k)) = F(X(k), K_x X(k)).$$

Define a Lyapunov candidate $V = X^T P X$. The difference equation involving V(.) along the above closed-loop system satisfies

$$V(k+1) - V(k) \leq -X^T(k)(Q + K_x^T R K_x + \kappa P)X(k) + X^T(k)((A_x - B_x K_x)^T P + P(A_x - B_x K_x))\Delta_x(k) + \Delta_x^T(k) P \Delta_x(k) \tag{9}$$

where $\Delta_x(k) = F(x(k), -K_x X(k)) - (A_x - B_x K_x)$ is the nonlinear component of $F(x(k), u^*(k))$. It should be noted that $F(0,0) = 0$, and so $\Delta_x(0) = 0$. Further, the set Θ_α is convex and compact. Exploiting the smoothness of $F(.)$, it can be derived that there exists a Lipschitz constant L_1 for Δ_x inside Θ_α. For a sufficiently small α, the Lipschitz constant L_1 is very very small for Δ_x inside Θ_α. For small enough L_1,

$$V(k+1) - V(k) \leq -X^T(k)(Q + K_x^T R K_x + \frac{\kappa}{2} P)X(k) \tag{10}$$

for all $X(k) \in \Theta_\alpha$. The level set for $V(.)$ is Θ_α, and hence, the set is positively invariant. In other words, the closed-loop dynamics for MPC law is locally asymptotically stable (A.S) for Θ_α.

Choice of design parameters: The choice of finite horizon N and the matrices Q, R is not arbitrary. A longer horizon may offer a larger set to search for optimal control law, but it also increases the computation burden. Ideally, the computation time should be smaller than the sampling time to solve the MPC algorithm, and the horizon N should be selected, keeping this in mind. The matrices Q and R determine the weightage of state and control variables in the optimization problem. If the weight of Q is increased gradually with each iteration, the state penalty will increase in the cost function with each shift in horizon, which may lead to faster convergence. Increasing the weight of R matrix may deviate the convergence from the optimal trajectory as more penalty is associated with the input energy.

6 Simulation results

The simulation is performed using the MATLAB MPC Solver. The horizon N is selected as 15, and the sampling time is selected as 0.2 second. (Both N and sampling time are tuned for different trajectories by trial and error.) The control inputs are saturated at 0.55 for both v and ω. The matrices for the cost function are chosen as:

$$Q = \text{Diag}(1, 1, 0.3),\ \ R = \text{Diag}(0.3, 0.1).$$

The initial conditions are chosen as (0,0,0), and a square path of length 3m is chosen for the reference tracking. The simulation results for the trajectory tracking and the control inputs are shown in Figs. 11 and 12. From Fig. 11, small deviations

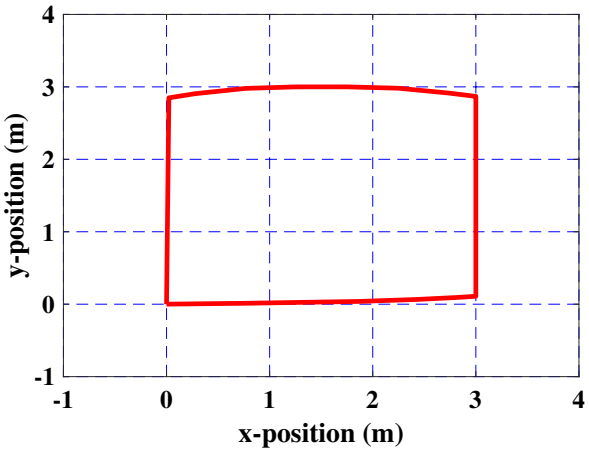


Fig. 11 Simulation for square path tracking

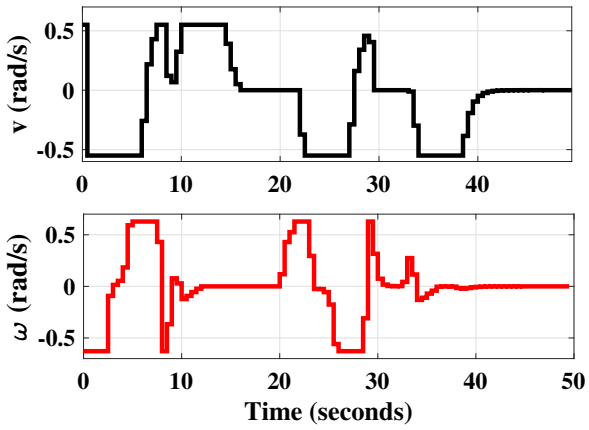


Fig. 12 Control inputs for square path tracking

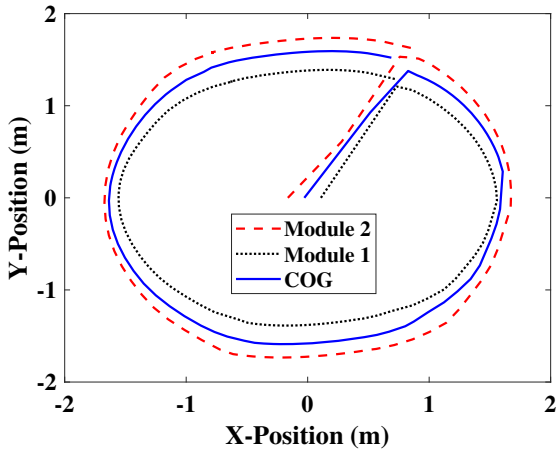


Fig. 13 Simulation result for circular path tracking of hRhombo in straight configuration

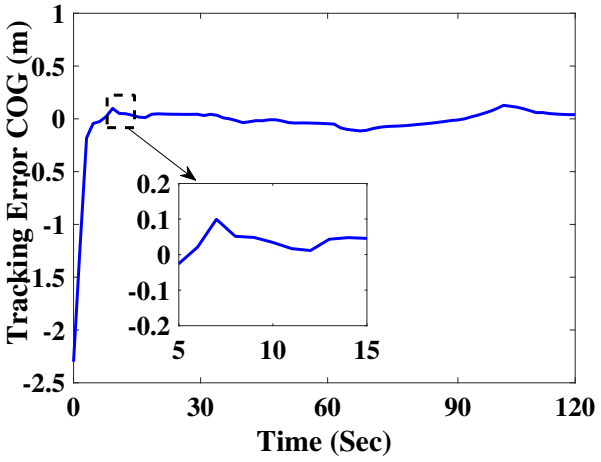


Fig. 14 Tracking Error of COG While Tracking the Circular Path

in the tracking can be noted around the corners. The deviations arise due to the sudden change in the reference position after a particular coordinate is reached. The deviations can be tuned by changing N and sampling time. We found that a smaller choice of N and sampling time reduces the deviations around the corners, but results in undesirable chattering. The control inputs shown in Fig. 12 saturate at 0.55. It is found that a larger actuation limit actually deteriorates the tracking performance.

The next set of simulations are performed to see the simultaneous circular path tracking performance of two unicycles. The control goal is chosen, such that the point S_2 of the platform tracks a circular path of radius 1.5 meter. For this purpose, we consider the unicycle models resulting from the wheels of modules 2 and 3 of the robot, whose wheels are separated by 20 cm in the straight configuration of hRhombo (I shape in Fig. 9). Two MPC algorithms are simultaneously solved for the wheels, and the results are presented in Figs. 13 and 14. The simulation presented here results from the choice of $N = 15$, $T = 1sec$. As observed from Fig. 14, the tracking error for the COG remains below 0.1 meters (less than 2 percent). To test the proposed multicycle MPC algorithm's performance for different configurations, three different configurations (I,D,L) are chosen to follow a desired circular path of 1.5 m. The path tracking results for these three different configurations are shown in figure 15. The differences in the tracking performance come from the variation in the shape variables for the different configurations. As the reference trajectory for these cases changes with each sampling time (time-varying signal), the numerical inaccuracies resulting from approximation do not get a chance to die down. This tracking errors can be decreased, if additional constraints like maximum deviation of the states are imposed on the optimization algorithm.

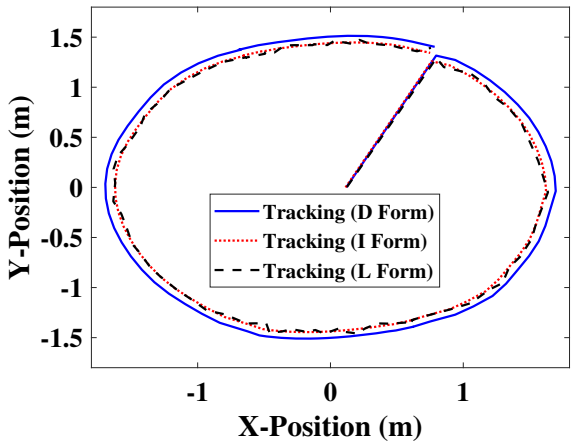


Fig. 15 Circular path tracking (COG) in different configurations

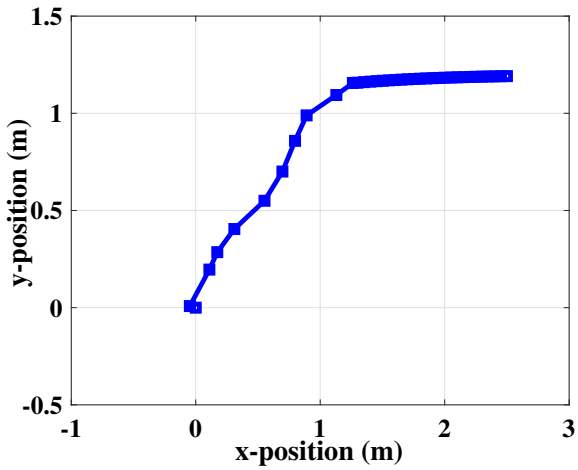


Fig. 16 Experimental result for ramp and straight-line tracking

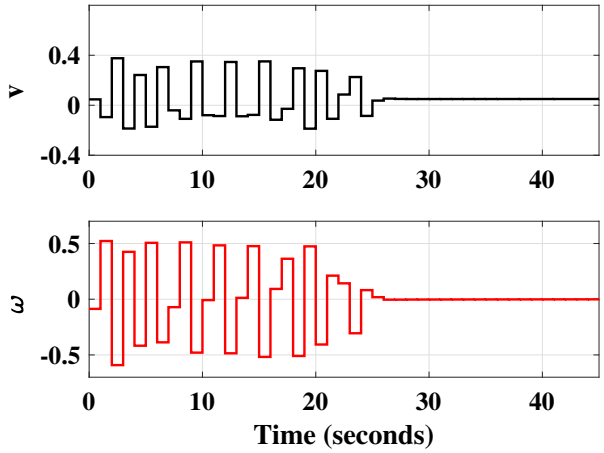


Fig. 17 Control inputs

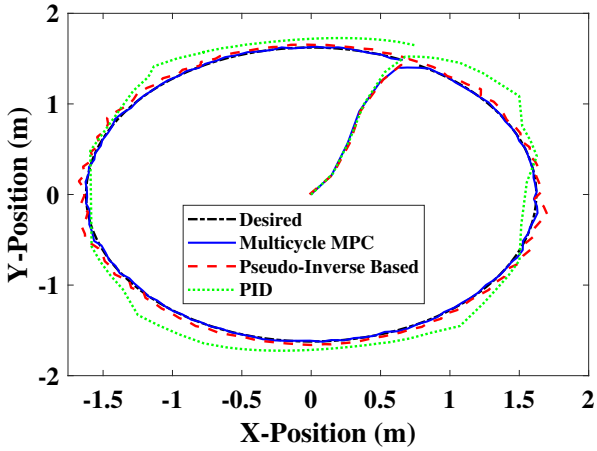


Fig. 18 Experimental results of circular path tracking for various approaches including the proposed method

7 Experimental results

The control inputs are generated using "quadprog" solver in MATLAB for, where the time step is chosen as 0.04 sec; the prediction and control horizon is selected as 7 and 2, respectively. The inbuilt MATLAB coder is used to generate the C++ code for execution. The localization is done using beacons, from which one is placed at the middle of the platform to provide its coordinates. The wheel rotations are collected using quadrature encoders and RoboClaw setup.

First experiment is performed for a desired ramp trajectory (0.05t,0.05t), and it is changed to a straight-line path (0.05t,0) once the platform reached (1.2,1.2) in the X–Y plane. The position measurements of COG are collected during the experiment and plotted in Figs. 16–17. Similarly, the second set of experiments is performed for a desired circular trajectory of 1.5 m. To show the effect of over-actuation in the robot kinematics on the closed-loop performance, a controller is implemented by using the pseudo-inverse of the Jacobian matrix $F(.)$ in the kinematics model (4). Apart from this controller, the performance is also compared with a conventional PID whose gains are tuned through the existing library functions of RoboClaw. The path tracking performance and the tracking errors for these control techniques are plotted in Figs. 18–19. The control inputs for the proposed controller are shown in Fig. 20. It can be observed from the figures that the response of the PID is inferior because it does not take the model knowledge into account. As the pseudo-inverse-based controller is affected by the over-actuation problem, the resulting wheel slippage deteriorates the tracking performance. The root mean square (RMS) of the path error is computed, and the results are presented in Table 1. It is calculated that the RMS error for the proposed controller is 58% less than the pseudo-inverse-based technique and more than 100% lesser than the conventional PID controller.

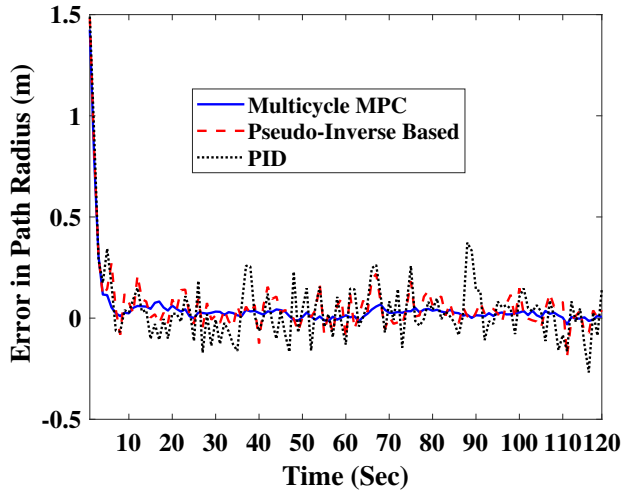


Fig. 19 Tracking errors resulting from circular path tracking experiments

Table 1 RMS path error (m) comparison

Controller	RMS Error
Multi.MPC	0.024
Inv. Jaco. based	0.038 (58% Larger)
PID	0.055 (> 100% Larger)

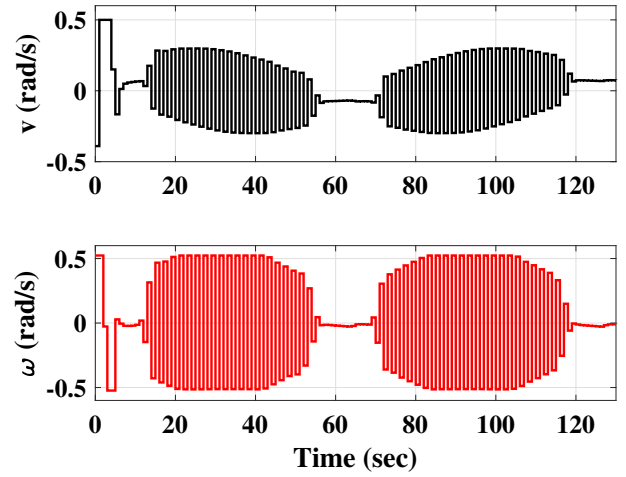


Fig. 20 Control inputs for circular tracking experiment with proposed approach

8 Conclusion

The work introduced a new reconfigurable robot platform named hRhombo. The mechanical design is inspired from polyrhombus principle which enables the platform to take seven different shapes. The platform's mechanism, electronics hardware, kinematics and control principle are described in detail. The forward kinematics and inverse kinematics are obtained assuming no slippage and skid. As the kinematics is over-actuated, a multicycle approach is proposed for decoupling the system dynamics. A model predictive control technique is used to ensure optimal path following, while satisfying actuator and state constraints. The robustness aspects and handling of other dynamical properties such as friction and gravity are the subjects of future research.

Acknowledgements This research was supported by the National Robotics Programme under its Robotics Enabling Capabilities and Technologies (Funding Agency Project No. 192 25 00051) and National Robotics Programme under its Robot Domain Specific (Funding Agency Project No. 192 22 00058) and administered by the Agency for Science, Technology and Research.

Author Contributions The authors declare that all the authors are equally involved in experimental design, writing and editing the manuscript.

Funding This research was supported by the National Robotics Programme under its Robotics Enabling Capabilities and Technologies (Funding Agency Project No. 192 25 00051) and National Robotics Programme under its Robot Domain Specific (Funding Agency Project No. 192 22 00058) and administered by the Agency for Science, Technology and Research.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

Consent to participate The article does not contain any studies that require any consent to participate.

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

Human and animal rights The authors declare that this article does not contain any studies including human participants or animal.

Informed consent: Not applicable.

References

1. Prabakaran V, Elara MR, Pathmakumar T, Nansai S (2018) Floor cleaning robot with reconfigurable mechanism. Autom Constr 91:155. https://doi.org/10.1016/j.autcon.2018.03.015
2. Vega-Heredia M, Mohan RE, Wen TY, SitiAisyah J, Vengadesh A, Ghanta S, Vinu S (2019) Design and modelling of a modular window cleaning robot. Autom Constr 103:268. https://doi.org/10.1016/j.autcon.2019.01.025
3. Mansouri SS, Kanellakis C, Fresk E, Kominiak D, Nikolakopoulos G (2018) Cooperative coverage path planning for visual inspection. Control Eng Pract 74:118. https://doi.org/10.1016/j.conengprac.2018.03.002
4. Cabreira TM, Franco CD, Ferreira PR, Buttazzo GC (2018) Energy-Aware spiral coverage path planning for UAV photogrammetric applications. IEEE Robot Autom Lett 3(4):3662
5. Anselmi N, Poli L, Rocca P, Massa A (2018) Diamond tiling optimization for hexagonal shaped phased arrays. In 12th European Conference on Antennas and Propagation (EuCAP 2018), pp 1–3

6. Miao X, Lee J, Kang B (2018) Scalable coverage path planning for cleaning robots using rectangular map decomposition on large environments. IEEE Access 6:38200
7. Khan A, Noreen I, Habib Z (2017) On Complete Coverage Path Planning Algorithms for Non-holonomic Mobile Robots: Survey and Challenges. J Inf Sci Eng 33(1)
8. Gonzalez E, Alvarez O, Diaz Y, Parra C, Bustacara C (2005) BSA: a complete coverage algorithm. In proceedings of the 2005 IEEE international conference on robotics and automation, pp 2040–2044
9. Lee TK, Baek S, Oh SY (2011) Sector-based maximal online coverage of unknown environments for cleaning robots with limited sensing. Robot Auton Syst 59(10):698. https://doi.org/10.1016/j.robot.2011.05.005
10. Yakoubi MA, Laskri MT (2016) The path planning of cleaner robot for coverage region using genetic algorithms. J Innov Digit Ecosyst 3(1):37. https://doi.org/10.1016/j.jides.2016.05.004
11. Cheng KP, Mohan RE, Nhan NHK, Le AV (2019) Graph theory-based approach to accomplish complete coverage path planning tasks for reconfigurable robots. IEEE Access 7:94642
12. Khan HRNLDZHA, Noreen I (2017) Online complete coverage path planning using two-way proximity search. Intell Serv Robot 10:229–240. https://doi.org/10.1007/s11370-017-0223-z
13. Zhang Y, Li S (2017) Distributed biased min-consensus with applications to shortest path planning. IEEE Trans Autom Control 62(10):5429
14. Viet MTCHH, Dange V (2013) BA*: an online complete coverage algorithm for cleaning robots. Appl Intell 39:217–235. https://doi.org/10.1007/s10489-012-0406-4
15. Tan N, Hayat AA, Elara MR, Wood KL (2020) A framework for taxonomy and evaluation of self-reconfigurable robotic systems. IEEE Access 8:13969
16. Prabakaran V, Elara MR, Pathmakumar T, Nansai S (2018) Floor cleaning robot with reconfigurable mechanism. 91:155
17. Parween R, Prabakaran V, Elara MR, Vengadesh A, Sivanantham V (2018) Application of tiling theory for path planning strategy in a polyiamond inspired reconfigurable robot. IEEE Access 7:6947
18. Parween R, Shi Y, Parasuraman K, Vengadesh A, Sivanantham V, Ghanta S, Mohan RE (2019) Modeling and analysis of hHoneycomb-A polyhex inspired reconfigurable tiling robot. Energies 12(13):2517
19. Parween R, Le AV, Shi Y, Elara MR (2020) System level modeling and control design of hTetrakis-A Polyiamond inspired self-reconfigurable floor tiling robot. IEEE Access 8:88177
20. Parween R, Heredia MV, Rayguru M, Abdulkader RE, Elara MR (2020) Autonomous self-reconfigurable floor cleaning robot. IEEE Access 8:114433–114442
21. Le AV, Parween R, Mohan RE, Enjikalayil NHKNR (2020) Optimization complete area coverage by reconfigurable hTrihex tiling robot. Sensors 20(11):3170
22. Veerajagadheswar P, Sivanantham V, Devarassu M, Elara MR (2019) Htetran–A polyabolo inspired self reconfigurable tiling robot. In 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS) (IEEE), pp 4877–4884
23. Shi Y, Elara MR, Le AV, Prabakaran V, Wood KL (2020) Path tracking control of self-reconfigurable robot hTetro with four differential drive units. IEEE Robot Autom Lett 5(3):3998
24. Johansen TA, Fossen TI (2013) Control allocation-A survey. Automatica 49(5):1087
25. Rayguru M, Kar I (2019) Contraction theory approach to disturbance observer based filtered backstepping design. J Dyn Syst Measurement Control 141(8):084501
26. Rayguru MM, Roy S, Kar IN (2019) Time-Scale Redesign-Based Saturated Controller Synthesis for a Class of MIMO Nonlinear Systems. IEEE Trans Syst Man Cybernet Syst 51:4681–4692
27. Rawlings JB, Meadows ES, Muske KR (1994) Nonlinear model predictive control: a tutorial and survey. IFAC Proc Vol 27(2):185
28. Morari M, Lee JH (1999) Model predictive control: past, present and future. Comput Chem Eng 23(4–5):667
29. Ploeg J, Vissers JP, Nijmeijer H (2006) Control design for an over-actuated wheeled mobile robot. IFAC Proc Vol 39(16):127

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

Available online at www.sciencedirect.com

ScienceDirect

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

Design and feasibility analysis of hydrogen based hybrid energy system: A case study

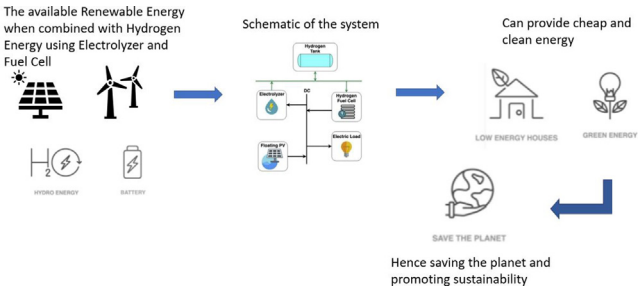
Srijit Basu [a], *Alfred John* [a], *Akshay* [a], *Anil Kumar* [a,b,*]

[a] *Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India*
[b] *Centre for Energy and Environment, Delhi Technological University, Delhi, India*

HIGHLIGHTS

- Analysis of a hybrid energy system with hydrogen storage for energy stabilization.
- Comparative study of solar-hydrogen, wind-hydrogen and hybrid energy models.
- Levelized cost of energy for hybrid model is $0.3387 and is most economical.
- CAPEX and OPEX of the hybrid model are $ 308,095.88 and $ 34,759.88 respectively.
- Hydrogen based energy storage system proves to be viable for the design constraint.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 30 May 2021
Received in revised form
27 July 2021
Accepted 7 August 2021
Available online xxx

Keywords:
Solar
Wind
Hydrogen
HOMER
Hybrid
Energy economics

ABSTRACT

Renewable energy sources can produce less carbon than conventional energy sources, which has the significant disadvantage of being intermittent, which triggers a stable storage system. This work focuses on the issues of hydrogen energy storage which can solve the fluctuating output power problem by simulating results on HOMER software. Three combinations of the Solar-Hydrogen system, Wind-Hydrogen system, and Solar-Wind-Hydrogen hybrid system are presented to find the most optimum one. Levelized Cost of Energy (LCOE) for Hybrid System has proven to be the most economical while the Wind Turbine cost 1.476% higher and the Solar Photovoltaics (PV) System costs 108.03% more. LCOE for Hybrid Model is $0.3387, while for Solar System it is $ 0.7046 and for Wind System it is $ 0.3437. These results show that a hydrogen-based energy storage system is viable for the considered.

© 2021 Hydrogen Energy Publications LLC. Published by Elsevier Ltd. All rights reserved.

* *Corresponding author*. Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India.
E-mail address: anilkumar76@dtu.ac.in (A. Kumar).

https://doi.org/10.1016/j.ijhydene.2021.08.036
0360-3199/© 2021 Hydrogen Energy Publications LLC. Published by Elsevier Ltd. All rights reserved.

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

Introduction

Energy, a central player of economic development, is imperative to support an advanced financial and social turn of events [1]. The increment in energy interest and environmental change issues are two of the present characterizing worldwide issues. Environmentally friendly power and energy effectiveness offer definite answers for decreasing greenhouse gases [2]. Oil costs for normal people are questionable and notably rising. Subsequently, comparing unpredictable energy costs affects energy security, equilibrium of installments, expansion, and financial development [3]. Along these lines, renewable energy and energy efficiency advancements permit energy to be low-carbon, spotless, protected, dependable, and autonomous on fuel costs, thus becoming essential [4]. Various arrangements effectively accessible to accomplish this job are as yet not palatable. In this manner, Hydrogen-based energy stockpiling advances show up and become current serious choices. Hydrogen has countless benefits; along these lines, it could diminish the discharges and reliance on imported oil compared to other options [5].

Hydrogen is the lightest, easiest, and generally copious of all synthetic components in the universe [6,7]. Hydrogen is anything but an essential wellspring of energy. Be that as it may, it turns into an appealing energy transporter when part from these different components by utilizing a wellspring of energy. Hydrogen is perfect as far as emanations at the mark of utilization. Sustainable hydrogen has a promising potential as an energy stockpiling medium. This way, it may be utilized to guarantee the energy framework security and cover the interest for energy in lower accessibility of variable renewable energy (VRE), particularly wind and sun-based energy [8]. Appropriate strategies for energy stockpiling help manage the discontinuity of renewable energy sources (RES) and add to numerous areas' decarbonization, including versatility and the building industry [9]. Understanding Hydrogen energy storage's stakes suggest incorporating the individual distinguishing proof of technological, financial, and social bottlenecks and building up a more fundamental methodology of the innovative framework rise. Such a capacity framework gives the electrical network improved force quality and strength by balancing out activity on the lattice during the critical time of supply-request irregularity. Monetarily, the put-away hydrogen with a 25 years life time is expressly moderate when utilized for power age in the framework. Long-haul power storage utilizing hydrogen-based innovations to guarantee framework energy effectiveness, dependability, and security with decreased fossil fuel byproducts at moderate expenses is presented along these lines.

In the past, Temiz et al. [10] designed and analyzed a PV and hydrogen-producing system on a floating platform with a Levelized cost of $0.612/kWh. It has also been concluded that the fuel cell generator with hydrogen production and accumulation unit tends to reduce unfulfilled electricity from 49.34% to 0.57%. Gökçek et al. [11] evaluated a hydrogen refueling station propelled by Wind-Photovoltaic hybrid power plant in Turkey with a Levelized cost of electricity as $0.146/kWh and Levelized cost of hydrogen was as $7.526/kg for Wind-Photovoltaic system with battery backup power system. They also concluded that the excess power generation was estimated as 41.11%. Sammy et al. [12] conducted a techno-economic analysis for rural electricity production in Egypt using multiple renewable energy sources. Optimum sizing of the Firefly Algorithm was found. Results were compared with those obtained from the Shuffled Frog Leaping Algorithm and the Particle Swarm Optimization. Al-Sharafi et al. [13] analyzed HOMER solar and wind energy and hydrogen production regions throughout Saudi Arabia. Minimum of COE 0.609$/kWh was found for power generation from combination of PV, wind turbine and battery as the energy storage system in the Yanbu area. However, the minimum COE in Abha was 1.208$/kWh depended on battery storage systems or hydrogen storage systems. Kalinci et al. [14] conducted an off-grid energy systems HOMER analysis on the island of Bozcaada, Turkey, and demonstrated which system component combination leads to minimal lowest energy cost (COE). Technical feasibility analysis of electricity and hydrogen production in Hendijan, Iran, has been shown in Qolipour et al. [15] using HOMER. Hendijan could economically achieve a hybrid energy system based on PV, wind and hydrogen. Ma et al. [16] examined the PV pumped storage pump system for Hong Kong off-grid power supply and found $0.289/kWh to be used for the optimal system.

Present study also focuses on a flawless design of a power station energized by Wind-hydrogen storage, Solar-hydrogen storage, and hybrid-hydrogen storage power system with the help of electrolyzer and fuel cell established in Lamba, Gujrat, as shown in Fig. 1.

The main novelty of the present work is the wind-solar hybridization with hydrogen energy and optimizing energy production with the lowest COE. The system is proposed for 150 houses in Lamba, Gujarat, with average energy consumption of 5.6 kWh/day. A research paper on such a topic and location has not been submitted before. The goal is to provide uninterrupted electricity with zero carbon emissions. Manuscript targets zero-emission during the mix of hydrogen creation dependent on water electrolysis, wind energy, and solar energy strategies with the chance of hydrogen executions for energy storage and transportation.

Simulation and modeling of hydrogen based hybrid energy system

Main purpose of the system is the electricity supply of a residential community and hydrogen production to stabilize the intermittency of renewable energy sources. Simulations were done using HOMER software for power stations energized by Solar-Hydrogen storage, Wind-Hydrogen storage, and Wind and Solar Hybrid-Hydrogen storage power system, as shown in Fig. 2. The software HOMER is used to detect the most feasible electrical load component sizes.

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036



Fig. 1 – Map of considered site.

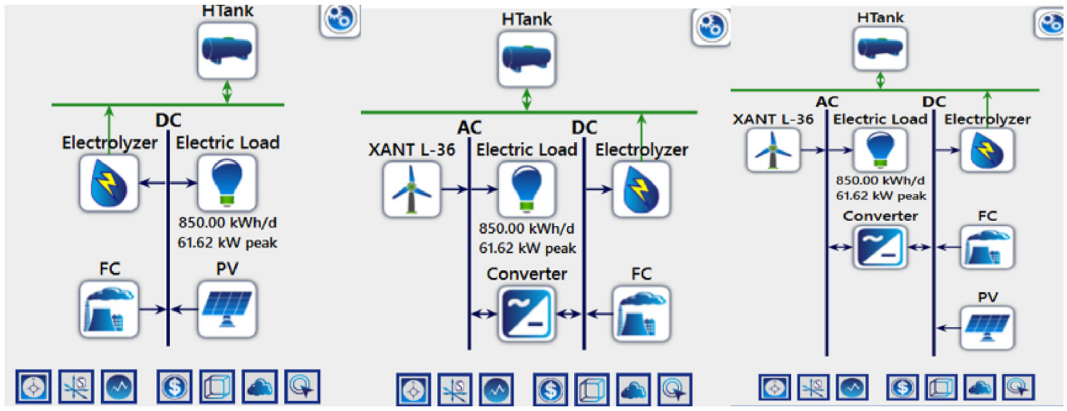


Fig. 2 – Layout of hybrid energy systems.

Solar PV system

Direct Current (DC) or Alternate Current (AC) electrical appliances are supplied with the help of Solar PV systems generally [17]. Solar PV system consisting of a PV solar module connected to a solar charge controller, converter or inverter and battery bank [18]. Sunlight is converted to direct current by solar panels with free power without excess electricity usage, supplied to the electrolyzer for producing hydrogen [19]. Fig. 3 depicts the monthly solar radiation, and from the following graph, we can find out that the yearly mean 4.63 kWh/m^2/day. As a monthly mean or as time series, HOMER accepts solar radiation data. The data sets TMY2 and TMY3 provided by the United States National Renewable Energy Laboratory are two of the best sources of solar radiation data (NREL). TMY2 and TMY3 files were directly imported into the solar resource input window of HOMER [20].

Electrical output of photovoltaic arrays is given in Equation (1).

$$P_{PV} = Y_{PV} \times f_{PV}\left(G_T/G_{T,STC}\right)\left[1 + P(T_C - T_{C,STC})\right] \quad (1)$$

Wind turbine system

HOMER determines the wind turbine yield with the aid of linear interpolation at points recorded to construct the power curve. The turbine power sum produced by the wind speed at the center point stature is known as power curve. The turbine output is zero outside the power curve [11]. When the breeze speed needed for the turbine's activity is too little to even think about creating energy, the turbine closes down to forestall harm. In this study, XANT L-36 turbine with a hub height of 55 m is used with a capacity of 300 kW. Wind power curve is given in Fig. 4 while Fig. 5 gives the mean wind speed. Mean

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

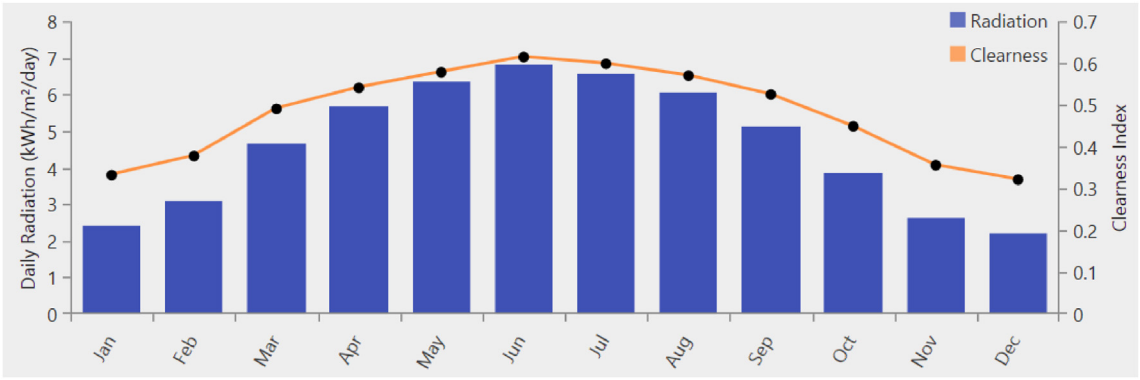


Fig. 3 – Monthly solar radiation.

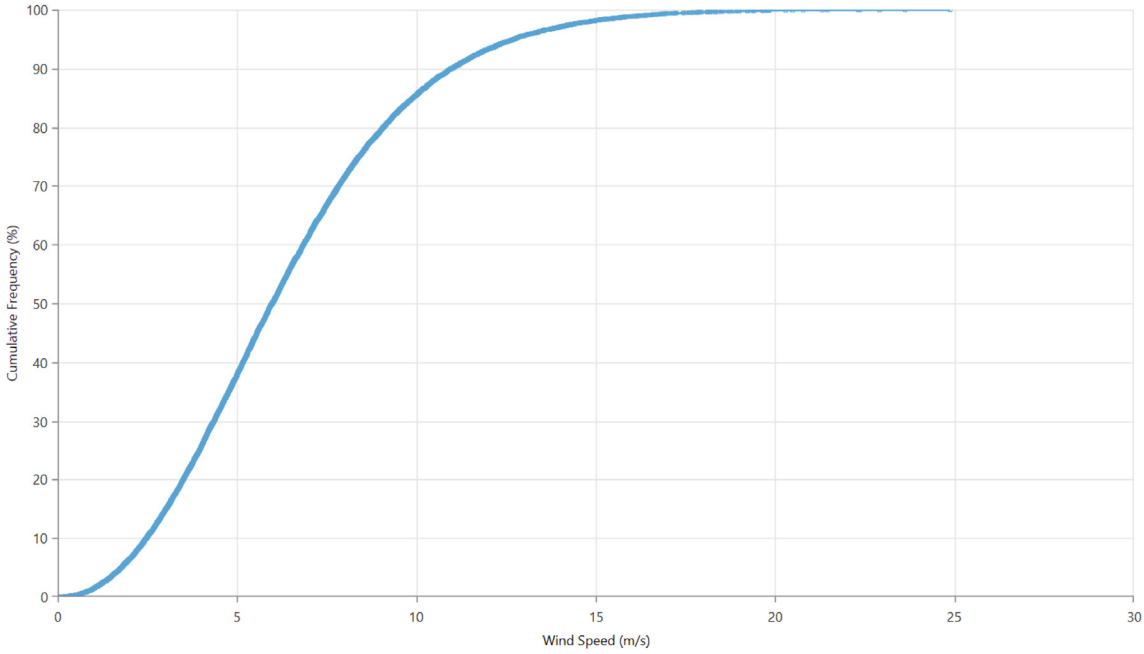


Fig. 4 – Wind power curve of XANT L-36.

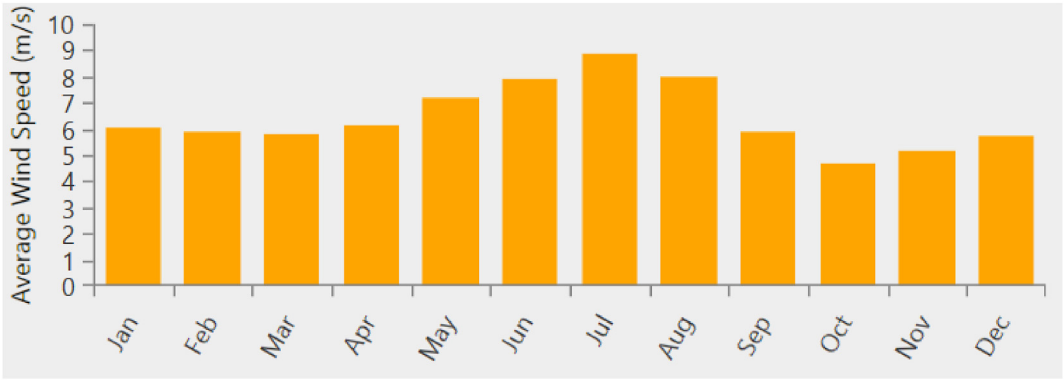


Fig. 5 – Monthly wind speed.

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

wind speed is 6.44 m/s at the considered region with an anemometer kept at the height of 10 m. Input data was downloaded from NASA Prediction of Worldwide Energy Resources [20].

Electrical output of wind turbine is given in Equations (2)–(4) under different conditions.

$$P_W = P_r \times \left[\left\{ U(t^3) - U_{ci}^3 \right\} / \left\{ U_r^3 - U_{ci}^3 \right\} \right], \; U_{ci} \le U \le U_r \tag{2}$$

$$P_w = P_r, \quad U_r \le U \le U_{co} \tag{3}$$

$$P_w = 0, \quad U_{co} \le U \le U_{ci} \tag{4}$$

In estimating V(t), the following Equation (5) is used.

$$V(t) = V_r \left(H_{WT} / H_r \right)^{\propto} \tag{5}$$

Electrolyzer

Electrochemical converter parts the water particle into its two segments, hydrogen and water. Electrolyzer produces hydrogen during the daytime, compacted and put away in the tank and shipped off the power module around evening time from demineralized water and photovoltaic energy.

Polymer electrolyte membrane (PEM) is an opposite interaction of PEM power module measure. Electrolyzer extracts hydrogen during daytime from demineralized water and solar power, compacted and preserved in the tank while being transported off the force module around evening time. PEM electrolysis is an alternate interaction of PEM power device measure [21]. Applying a DC voltage stronger than the thermoneutral voltage fragments water into oxygen, protons, and electrons on one cathode Protons, pass through the polymer electrolyte coating and bind with electrons on the cathode to generate hydrogen. Water transport connects the protons' segment via the plate using electroosmotic drag. It has porous anodes, a polymer film, stream fields, and plates like current, separator, end, and transport. The procedure guideline is the polar opposite of power system activity. In any case, the materials are ordinarily unique concerning PEM energy unit materials [22]. For the electrolyzer, capital and replacement cost was $80,000 and $44,926.71 with a lifespan of 15 years, respectively. Operation and maintenance cost was found as $30 with an efficiency of 85%.

Hydrogen tank

Hydrogen is an energy transporter and ought to be supplied to adapt to the occasional and day-by-day inconsistencies among power source attainable quality and the demand [23]. Capacity of hydrogen in a compressed gaseous structure has the most straightforward storing arrangement. For long-term use, the energy supply needs a storage system. Tank is shielded from direct sun-based radiation and set in an arid, cold, and aerated climate. Capital cost of hydrogen tank is taken as $90,000.

Fuel cell

It is an electrochemical instrument that changes chemical energy into direct current electrical energy to energize a load demand. Input into the cell is fuel and an oxidizer, which produces direct fuel, water and heat at the interior of the cell; the fuel and an oxidizer lap two electrodes, parted by the electrolyte for transfer of ions due to a reaction between the electrodes, closing the electric circuit. Wastewater is clean and non-contaminating [24]. When H_2 is brought into the cell, particles and electrons are initiated in the anode, and transfer of electrons occurs through electrolyte. Cathode completes the first two streams, the other using the formation of H^+ particles through the film, which was previously bound to the electrons traveling through the load associating the anodes (the client). On the cathode, the second phase of the redox reaction takes place. Electrons build up with the cations of hydrogen and a certain measure of oxygen, releasing water. Capital and replacement estimate of the fuel cell is $120,000 and $123,649.491, respectively, operation and maintenance costs are 0.10 $ per hour and age of 40,000 h with an efficiency of 50%.

Economic analysis

Economic assessment is crucial for durability; hence, elements between hybrid renewable energy systems (HRES) can ultimately be combined [25]. Net present cost (NPC) is the current absolute net present value for each fee, subtracted by the current virtue of the revenue to the complete amount earned up to its full life time [26]. NPC represents a total net current value of a provision. Costs include capital expenditure, substitution, operating expenses, fuel prices, pollution penalties and grid power purchase costs. Revenue comprises savings and sales revenue from the grid. HOMER analyses overall net cost provided in Equation (6).

$$C_{NPC} = C_{AT} / CRF\left(i_r, N_{proj}\right) \tag{6}$$

where C_{NPC} is taken as absolute NPC ($), C_{AT} is taken as absolute yearly cost ($/year). C_{RF} is taken as the capex recovery factor, i_r is taken as the interest rate, N_{proj} is taken as the whole project life time on a yearly basis that is 20 years shown in Equation (7).

$$CRF\left(i_r, N_{proj}\right) = i_r (i_r + 1)^{N_{proj}} / (i_r + 1)^{N_{proj}-1} \tag{7}$$

After moving to an initial cost then announced cost [27], the real annual interest rate is considered the bargain rate. Lifespan annual actual interest rate can be considered and presented as nominal interest rate in Equation (8).

$$i_r = i' - f / f + 1 \tag{8}$$

i' is minimal interest rate, f is annual rate of inflation. Overall annual cost of HRES is shown as follows, Equation (9).

$$C_{AT} = C_{A,Ca} + C_{A,Re} + C_{A,OM} + C_{A,Fu} \tag{9}$$

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

$C_{A,Ca}$ is taken as annual basis capital cost, $C_{A,Re}$ is yearly substitution cost, $C_{A,OM}$ is yearly maintenance cost, $C_{A,Fu}$ is yearly fuel cost of hybrid renewable energy sources parts. Yearly capital cost of a part is considered by using the following, Equation (10), [28].

$$C_{A,Ca} = C_{I,Ca} \times CRF(i_r, N_{proj}) \quad (10)$$

$C_{I,Ca}$ is taken as the capital cost of part at the beginning. Yearly substitution cost of a part is calculated by applying the following, Equation (11).

$$C_{A,Re} = C_{Re} \times F_{Re} \times SFF(i_r, N_{Comp}) - S \times SFF(i_r, N_{proj}) \quad (11)$$

C_{Re} is taken as the substitution cost of the part, F_{Re} is the substitution factor, SFF - sinking fund factor, s is considered as the salvage amount of the part. Levelized Energy Costs of production were determined as to assess the financial stability of the project given in Equation (12).

$$CoE = C_{AT}/(E_{PL} + E_{DL} + E_G) \quad (12)$$

where E_{PL} is considered as primary load supplied (AC and DC) in kWh/year, E_{DL} is deferrable load supplied in kWh/year, E_G is the absolute grid sales [29].

Results and discussion

HOMER has simulated the results for the best scenario at the following.

Solar PV system

PV system with a rated capacity of 741 kW and maximum output of 651 kW was simulated. PV system was operated for 4215 h in a year with mean output was 113 kW and a capacity factor of 15.3%, which contributes to a total production of 992,931 kWh/year. PV power output in kW is depicted in Fig. 6.

According to simulation results, the scale of the electrolyzer was 210 kW. It has been run for 2446 h in a year. In the electrolyzers, total energy consumption was 366.8 MWh/year with a mean power supply of 41.8 kW and a maximum power input of 210 kW. Hydrogen production of 0.90 kg/h was found. For an entire year, 7905 kg of hydrogen was produced. Average output of 0.902 kg/h hydrogen has been identified. Electrolyzer input in kW is depicted in Fig. 7 during the year. Generator capacity of the fuel cell is simulated as 60 kW. It worked for 5174 h a year and restarted 375 times a year. Fuel cell has produced electricity of 131,754 kWh/year.

Average power output is found to be 25.5 kW with a high potential output of 55.4 kW. Fig. 8 displays day-to-day fuel cell production by an hour in kW. 7905 kg were consumed, specific fuel consumption is 0.06 kg/kWh, 263.507 kWh/year fuel input, and mean electric power consumption of 50% is found. Principal function of the fuel cell is to compensate for electricity if no solar energy is sufficient. Operating hours have therefore not been continuous. As observed in Fig. 8, the power consumption is constant during the first hours of the day with approximately 12.0 kW, which is 20% of the total size. Capability factor is the proportion of electricity generated in a particular time period to the maximum electricity produced for the same time period. Simulation revealed a power factor of 25.1% for the hydrogen fuel cell. Service life of the fuel cell was estimated to be 7.73 years, depending upon these specific operating conditions. Simulation life is expected to be around 25 years; hence there have been 3 fuel cell substitutes. Hydrogen tank's size was 40 kg with overall power output as 1333 kWh. Fig. 10 indicates the amount of a hydrogen tank in kg, regularly by hour. Hydrogen tank autonomy was defined at 37.6 h, that is to say, the proportion of the maximum energy produced by the hydrogen tank with respect to the electrical demand load. The hydrogen storage tank levels for a complete year are observed in Fig. 9. Candle graphic representation shows the hydrogen levels. Each line reflects a month limit, daily maximum, weekly average, daily average low, monthly minimum and entire month minimum from top to bottom. Tank level can be identified for one full year, hourly and daily, in Fig. 9. Red area in the centre of Fig. 10 is elucidated as maximum electricity production by Solar PV system during the availability of maximum solar radiation. In contrast, blue, dark blue and black regions may be regarded as hours of work following a long solar intermittence during the day time, which could impact the population as the brief duration of energy.

A net 1124.684 MWh of electricity was produced for the whole year. Hydrogen fuel cell generates 11.7% of the overall

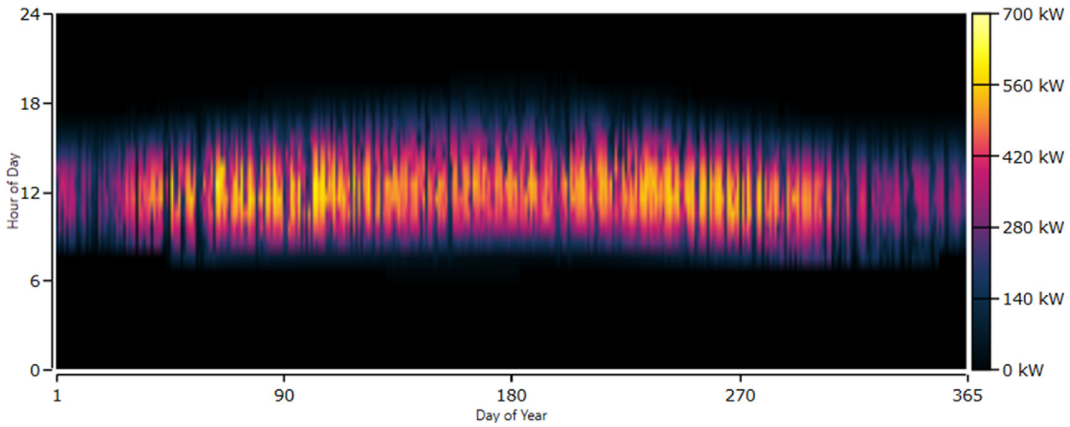


Fig. 6 – PV power output.

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

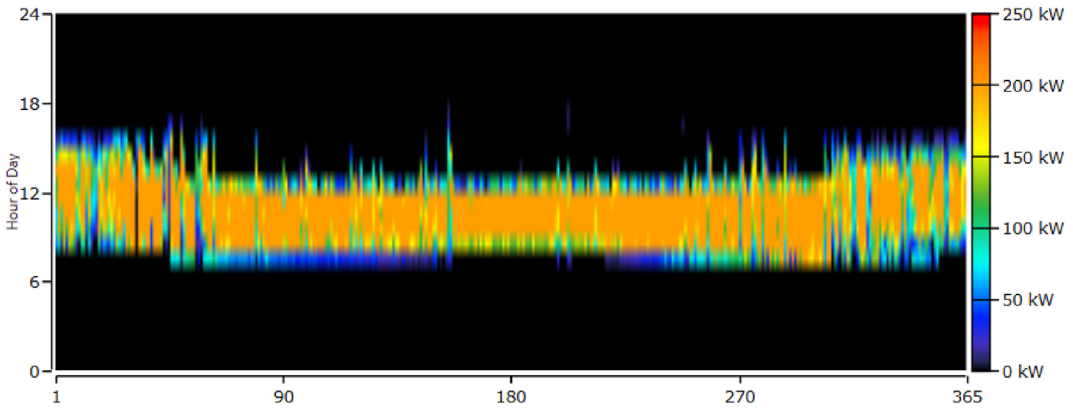


Fig. 7 – Electrolyzer input power (kW).

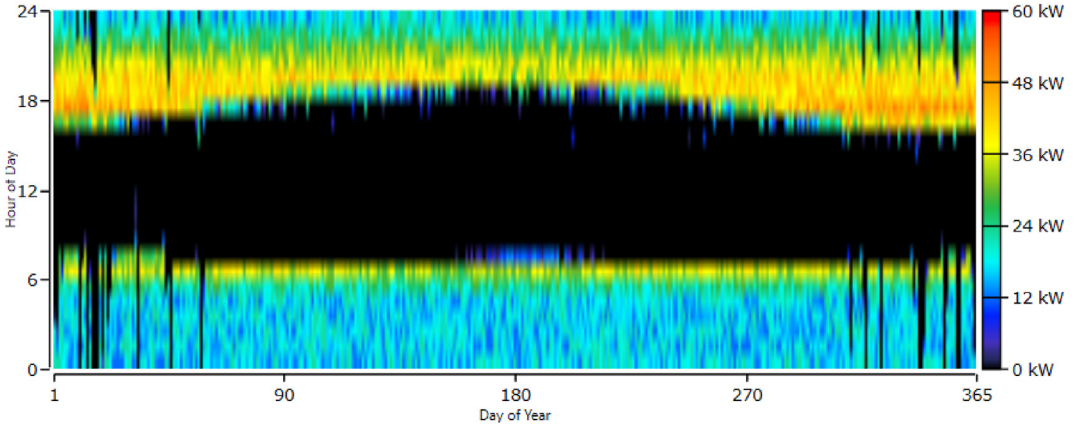


Fig. 8 – Fuel cell output (kW).

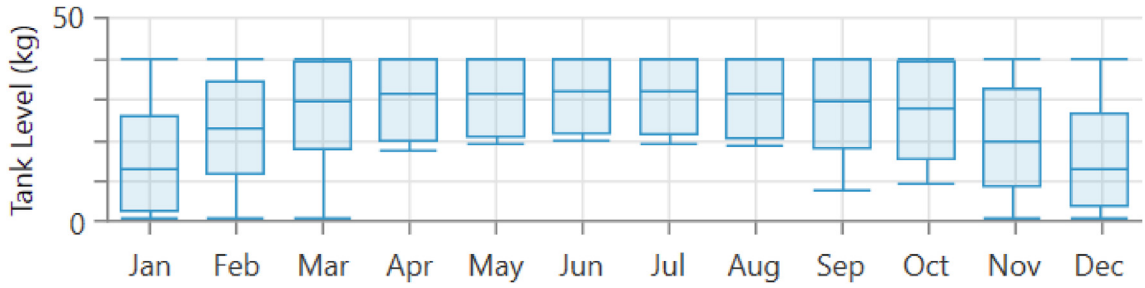


Fig. 9 – Monthly hydrogen tank level (kg).

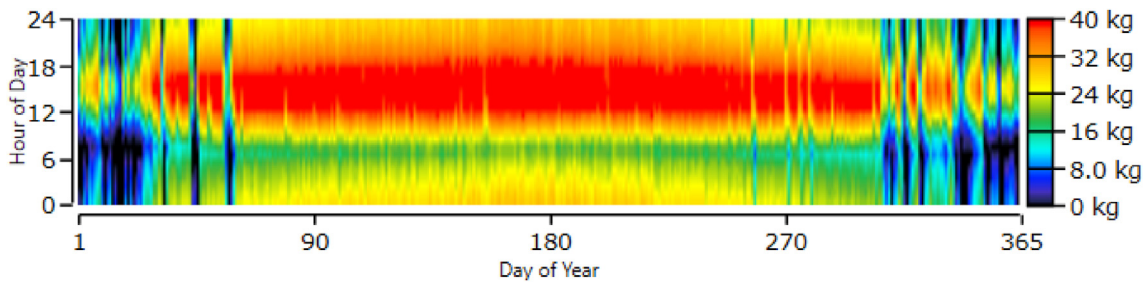


Fig. 10 – Hydrogen tank level (kg).

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

electricity generation, equivalent to 131.754 MWh/year. Photovoltaic system is the largest power producer of the system, with a share of 88.3′ which is 992.931 MWh per year. 59.69% of the overall power generation was consumed by electric load and electrolyzer. Electrical load consumption was found to be 304.489 MWh per year. Unmet electric load and capacity shortage of 5.761 MWh per year and 6.512 MWh per year, respectively, have been achieved. Capacity shortage and unmet electric load take a cumulative total of 1.82% of the total electricity load. System was detected at 453.342 MWh per year for the total surplus electricity.

Wind turbine system

Wind Turbine system with a rated capacity of 300 kW and maximum output of 300 kW was simulated. PV system was operated for 7836 h per year with mean output was 142 kW and a capacity factor of 47.2%, which contributes to a total production of 1,240,943 kWh/year. Wind power output in kW is depicted in Fig. 11.

According to simulation results, the scale of the electrolyzer was 40 kW. It has been run for 4075 h in a year. In the electrolyzer, total energy consumption was 132,469 kWh/year, with a mean power supply of 15.1 kW and a maximum power input of 39.6 kW. Hydrogen production of 0.853 kg/h was found. For an entire year, 2855 kg of hydrogen were produced. Average output of 0.326 kg/h hydrogen has been identified. Electrolyzer input in kW is depicted in Fig. 12 during the year. The generator capacity of the fuel cell is simulated as 40 kW. It worked for 3471 h a year and restarted 763 times a year. Fuel cell has produced an electricity of 46,077 kWh/year. Average power output is found to be 13.3 kW with a high power output of 40 kW. Fig. 13 displays day-to-day fuel cell production by hour in kW. In total, 2765 kg were consumed, where the specific fuel consumption is 0.06 kg/kWh, 92,154 kWh/year fuel input and means electric power consumption of 50% was found. Principal function of the fuel cell is to compensate for electricity if no solar energy is sufficient. Operating hours have therefore not been continuous. As observed in Fig. 13, the power consumption is constant during the first hours of the day with approximately 8.0 kW, 20% of the total size. Capability factor is the proportion of electricity generated in a particular time period to the maximum electricity produced for the same time period. Simulation revealed a power factor of 13.1% for the hydrogen fuel cell. Service life of the fuel cell was estimated to be 11.5 years, depending upon these specific operating conditions. Simulation life is expected to be around 25 years. Hence, there have been 2 fuel cell replacements. Hydrogen tank's size has been found to be 100 kg with overall power output of 3333 kWh. Fig. 15 indicates the amount of a hydrogen tank in kg, regularly by hour. Hydrogen tank autonomy was defined at 94.1 h, that is to say, the proportion of the maximum energy produced by the hydrogen tank to the electrical demand load. Hydrogen tank levels for the entire year are seen in Fig. 14. Candle graphic representation shows the hydrogen levels. The tank level can be identified for one full year, hourly and daily, in Fig. 14. A total of 1,287,020 kWh of electricity was produced for the whole year. Hydrogen fuel cell generates 3.58% of the overall electricity generation, equivalent to 46,077 kWh/year. Photovoltaic system is the largest power producer of the system with a share of 96.4%, which is 1,240,943 kWh per year. 34.16% of the overall power generation was consumed by electric load and electrolyzer. Electrical load consumption was found to be 307.261 MWh per year. The unmet electric load and capacity shortage of 2.989 MWh per year and 6.324 MWh per year, respectively, have been achieved per year. Capacity shortage and unmet electric load take a cumulative total of 2.11% of the total electricity load. System was detected at 838.013 MWh/year for the total surplus electricity.

Hybrid system

PV system in the hybrid system has a rated capacity of 16.7 kW and a maximum output of 14.6 kW was simulated. PV system was operated for 4215 h in a year with mean output was 2.55 kW and capacity factor of 15.3%, contributing to total production of 22,344 kWh/year. PV power output in kW is depicted in Fig. 16.

Wind Turbine system in the Hybrid System has a rated capacity of 300 kW with a minimum output of 0 kW and a maximum output of 300 kW was simulated. PV system was operated for 7836 h/year with mean output was 142 kW and capacity factor of 47.2%, contributing to a total production of

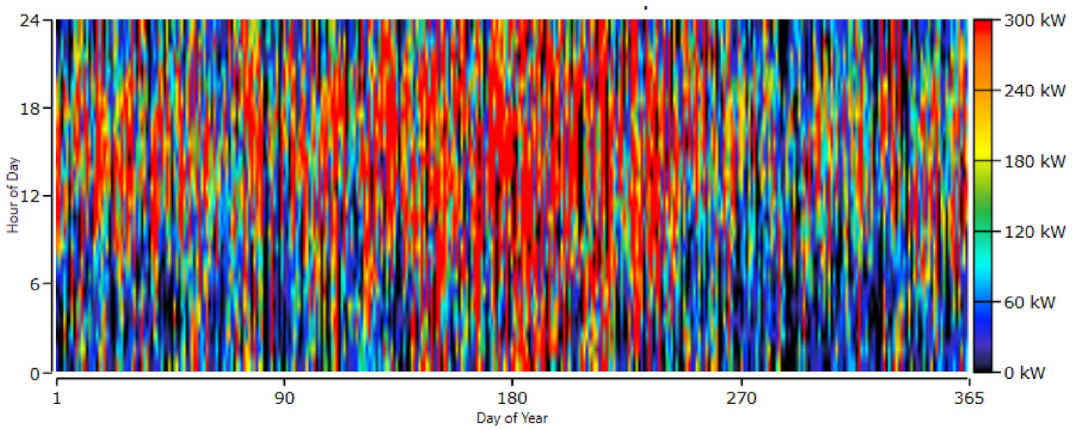


Fig. 11 – Wind turbine power output.

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

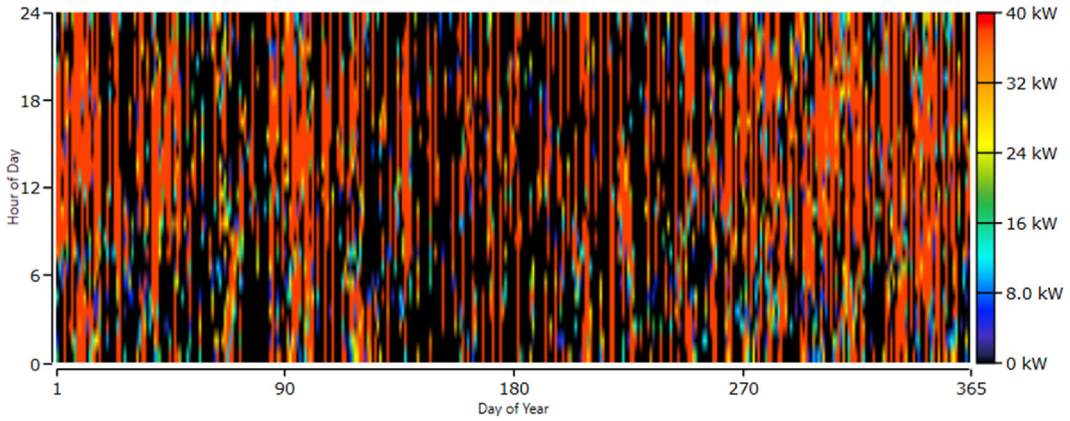


Fig. 12 – Electrolyzer input power (kW).

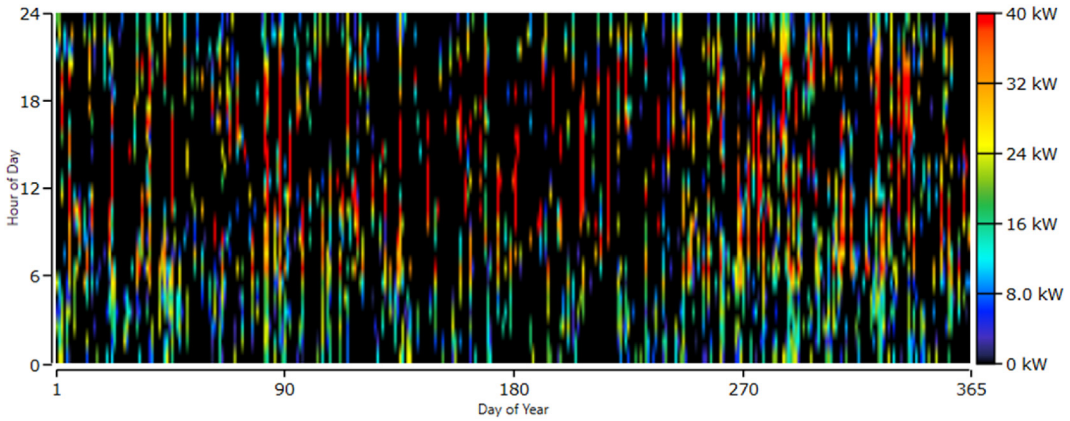


Fig. 13 – Fuel cell output (kW).

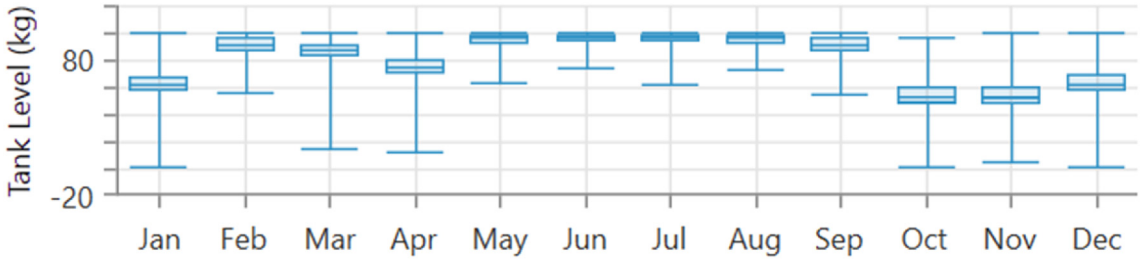


Fig. 14 – Hydrogen tank level (kg) by month.

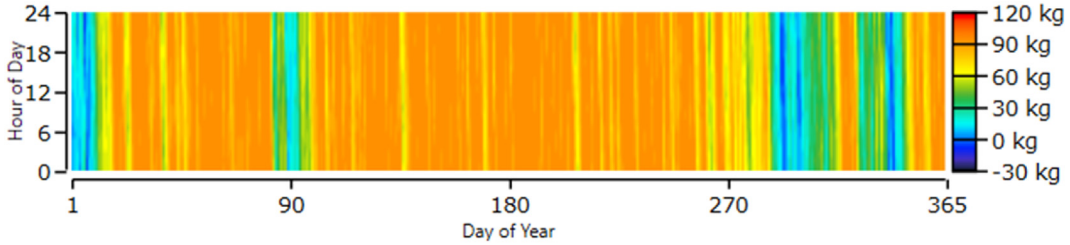


Fig. 15 – Hydrogen tank level (kg).

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

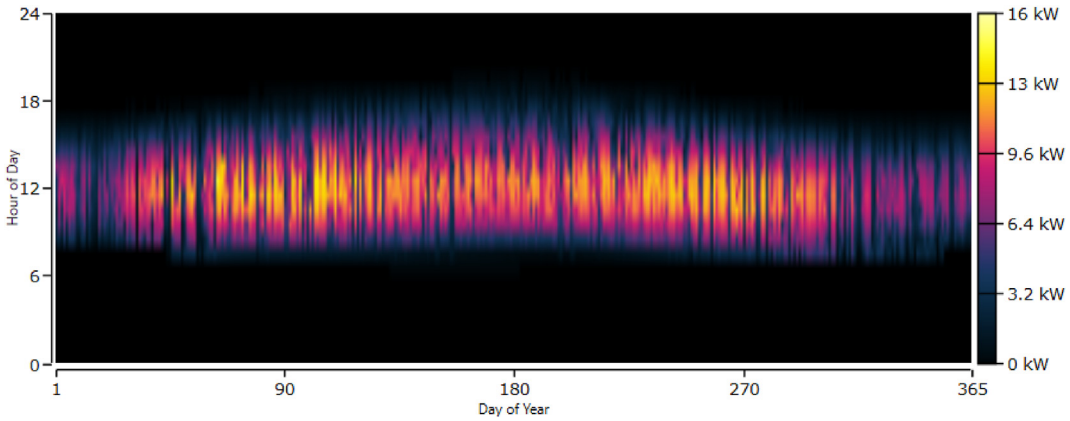


Fig. 16 – PV power output.

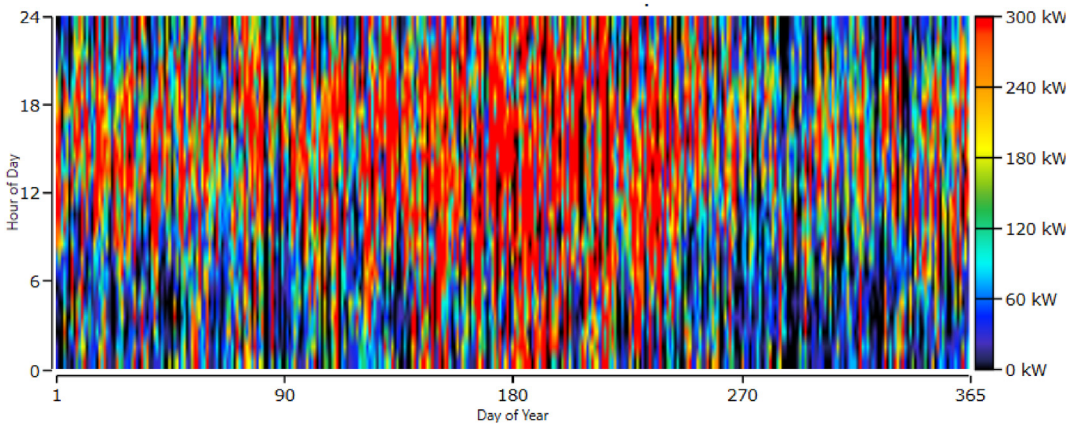


Fig. 17 – Wind turbine power output.

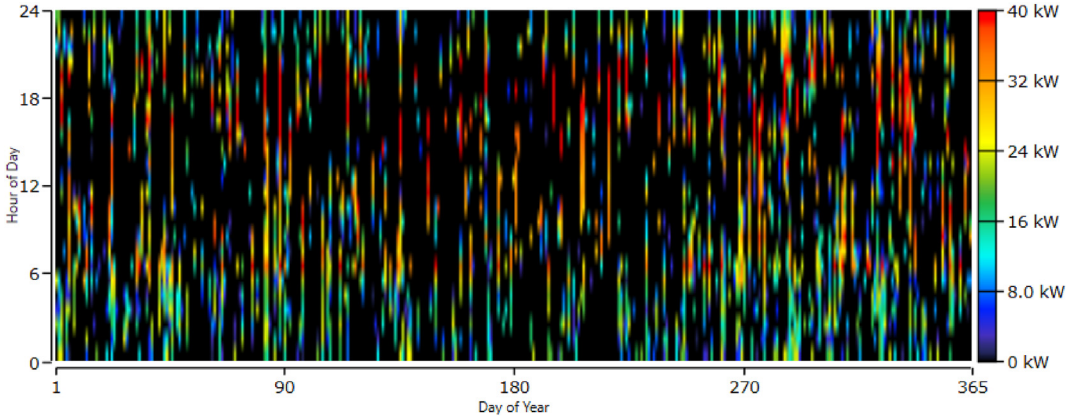


Fig. 18 – Electrolyzer input power (kW).

1,240,943 kWh/year. Wind Power output in kW is depicted in Fig. 17.

According to simulation results, the scale of the electrolyzer was 40 kW. It has been run for 3726 h in a year. In the electrolyzers, total energy consumption was found to be 117,569 kWh/year with a mean power supply of 13.4 kW and a maximum power input of 40 kW. Hydrogen production of 0.862 kg/h was found. For an entire year, 2534 kg of hydrogen was produced. Average output of 0.289 kg/h hydrogen has been identified. Electrolyzer input in kW is depicted in Fig. 18 during the year. Generator capacity of the fuel cell is simulated as 40 kW. It worked for 3348 h a year and restarted 773 times a year. Fuel cell has produced electricity of 41,325 kWh/year. Average power output is 12.3 kW with a high-power output of 40 kW and low output as 0.00652 kW. Fig. 19 displays day-to-day fuel cell production by hour in kW. In total,

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

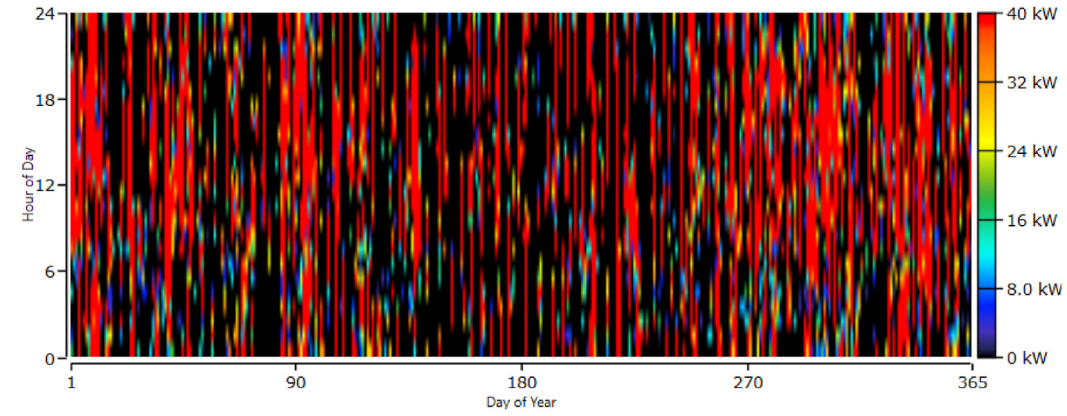


Fig. 19 – Fuel cell output (kW).

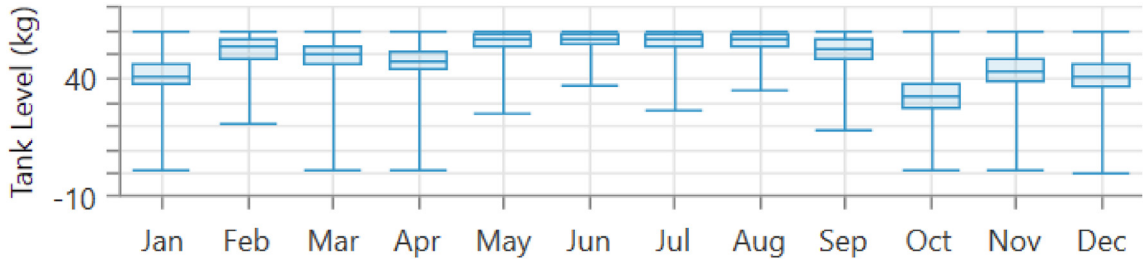


Fig. 20 – Hydrogen tank level (kg) by month.

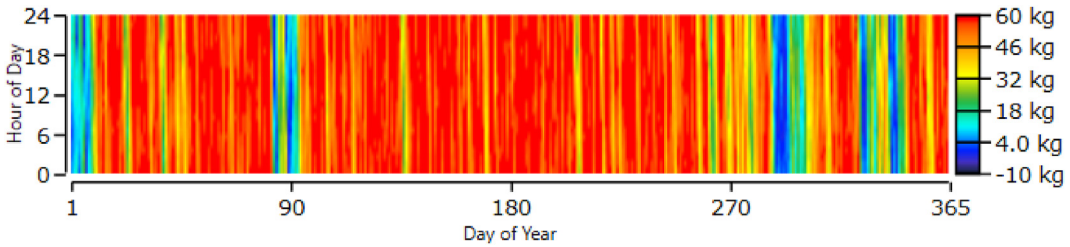


Fig. 21 – Hydrogen tank level (kg).

Table 1 – Economic summary from all the systems.

Economic parameter	Solar PV system	Wind turbine system	Hybrid system
Net present cost ($)	3,379,547.00	1,663,756.00	1,638,635.00
CAPEX ($)	605,683.59	313,220.66	308,095.88
OPEX ($)	78,610.08	35,034.72	34,759.88
LCOE (per kWh, $)	0.7046	0.3437	0.3387

2480 kg were consumed, where the specific fuel consumption is 0.06 kg/kWh, 82,651 kWh/year fuel input and mean electric power consumption of 50% was found. Principal function of the fuel cell is to compensate for electricity if no solar energy is sufficient. Operating hours have therefore not been continuous. As observed in Fig. 19, the power consumption is constant in the first hours of the day with approximately 8.0 kW, 20% of the total size. Capability factor is the proportion of electricity generated in a particular time period to the maximum electricity produced for the same time period. Simulation revealed a power factor of 11.8% for the hydrogen fuel cell. Service life of the fuel cell was estimated to be 11.9 years, depending upon these specific operating conditions. Life is expected to be around 25 years. Therefore, there have been 2 fuel cell replacements. Hydrogen tank's size has been found to be 60 kg with overall power output as 2000 kWh. Fig. 21 indicates the amount of a hydrogen tank in kg, regularly by hour. Hydrogen tank autonomy was defined at 56.5 h, that is to say, the proportion of the maximum energy produced by the hydrogen tank to the electrical demand load. Hydrogen tank levels for the whole year are observed in Fig. 20. Candle graphic representation shows the hydrogen

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

levels. Tank levels can be identified for one full year, hourly and daily, in Fig. 21. A total of 1,304,612 kWh of electricity was produced for the whole year. Hydrogen fuel cell generates 3.17% of the overall electricity generation, equivalent to 41,325 kWh per year. Wind turbine is the largest power producer of the system with a share of 95.1%, which is 1,240,943 kWh per year. The solar PV system produces 1.71% of the total production, which equals 22,344 kWh per year. 32.55% of the overall power generation was consumed by electric load and electrolyzer. Electrical load consumption was found to be 307.091 MWh per year. Unmet electric load and capacity shortage of 3.159 MWh per year and 6.459 MWh per year, respectively, have been achieved. Capacity shortage and unmet electric load take a cumulative total of 2.2% of the total electricity load. The system was detected at 871.968 MWh per year for the total surplus electricity.

Conclusion

In accordance with the simulation results by HOMER software, the Hybrid framework is found as the most practical where power demand is reasonable. LCOE for Hybrid System has been the cheapest while the Wind Turbine cost 1.476% higher and the solar PV System costs 108.03% more as observed from Table 1 economic summary from all the systems. Optimization shows that hydrogen production with hybrid sustainable energy is financially viable for the area relative to solar system and wind system hydrogen production. Modeling and designing three diverse frameworks have been simulated for 850 kWh/day in a residential area in Lamba, Gujarat. Hydrogen-generated electrolyzer and retained hydrogen have been used to substitute for electricity where the need cannot be met by renewable energy. Longevity of the system has been set to 25 years. The net present cost to be $3,379,547 and $1,663,756 for Solar PV System and Wind Turbine System respectively, while the Net Present Cost for Hybrid System is $1,638,635 (Table 1). Operating expense for solar PV System is $78,610.08 and for Wind Turbine System is $35,034.72, whereas for hybrid system it is $34,759.88. Levelized cost is $0.7046 and $0.3437 for solar PV system and wind turbine system, respectively, while for the Hybrid System, the LCOE is $0.3387.

For a remote residential environment, hydrogen and fuel cells provide a continuous energy supply with zero greenhouse gas emissions. Besides, it would be more economical as the machine is considered on a grid and sold surplus power in the grid. Technologies employed are still new, and Indian growth potentials are strong. If incentivized by the government, the technologies can provide sustainable and uninterrupted electrical power to the residents at a cheaper cost.

In addition to being an electricity generation solution and a sequence for energy, renewable hydrogen production has excellent ecological characteristics. Due to alternative energy sources, hydrogen can be produced in an ecologically friendly manner [30]. The results also revealed that an electrolyzer and hydrogen tank storage system could replace the battery bank. However, the cost is still high due to the higher initial investment involved in implementing such a plan. Hydrogen produced from renewable sources can be produced at lower costs by limiting its energy losses. Once the energy loss hurdle is overcome, a major source of future energy for transportation and power is green hydrogen.

Declaration of competing interest

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

Acknowledgment

The authors would like to express their deepest gratitude towards Centre for Energy and Environment, Delhi Technological University, Delhi (India), for the basic facilities provided in this manuscript compilation.

Nomenclature

AC	Alternate current
DC	Direct Current
PV	Photovoltaics
Y_{PV}	capacity of photovoltaic array
f_{PV}	Photovoltaic derating factor
G_T	Incident solar radiation at present time,
$G_{T,STC}$	Incident solar radiation at standard conditions
$_P$	Temperature coefficient of power
T_c	Photovoltaic cell temperature during present time,
$T_{C,STC}$	Photovoltaic cell temperature at standard conditions
P_w	Total power output of wind turbine
P_r	Rated power output of wind turbine
U_r	Rated speed of wind turbine
U_{ci}	Cut-in speed of wind turbine
U_{co}	Cut-out speed of wind turbine
$V(t)$	New speed of wind turbine at the hub height
H_{WT}	Hub height of wind turbine
H_r	Reference height
α	Friction coefficient
HRES	Hybrid Renewable Energy Systems
VRE	Variable Renewable Energy
RES	Renewable Energy Sources
C_{NPC}	Total Net Present Cost
C_{AT}	The total annualized cost
C_{RF}	The capital recovery factor
i_r	The interest rate
N_{proj}	The project life time in years
i'	The nominal interest rate
f	The annual inflation rate
$C_{A,Ca}$	The annualized capital cost
$C_{A,Re}$	The annualized replacement cost
$C_{A,OM}$	The annualized maintenance cost
$C_{A,Fu}$	The annualized fuel cost of HRES components
$C_{I,Ca}$	The initial capital cost of the component
C_{Re}	The replacement cost of the component
F_{Re}	The replacement factor
SFF	The sinking fund factor
s	The salvage value of the component

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

E_{PL}	The primary load served
COE	Cost of energy
$LCOE$	Levelised cost of energy
E_{DL}	The deferrable load served
E_G	The Total grid sales

REFERENCES

[1] Montoya FG, Aguilera MJ, Manzano-Agugliaro F. Renewable energy production in Spain: a review. Renew Sustain Energy Rev 2014;33:509–31. https://doi.org/10.1016/j.rser.2014.01.091.
[2] Kumar A, Kumar K, Kaushik N, Sharma S, Mishra S. Renewable energy in India: current status and future potentials. Renew Sustain Energy Rev 2010;14:2434–42. https://doi.org/10.1016/j.rser.2010.04.003.
[3] Klessmann C, Rathmann M, de Jager D, Gazzo A, Resch G, Busch S, et al. Policy options for reducing the costs of reaching the European renewables target. Renew Energy 2013;57:390–403. https://doi.org/10.1016/j.renene.2013.01.041.
[4] Nilsson M. Red light for green paper: the EU policy on energy efficiency. Energy Pol 2007;35:540–7. https://doi.org/10.1016/j.enpol.2005.12.023.
[5] Khalid F, Dincer I, Rosen MA. Analysis and assessment of an integrated hydrogen energy system. Int J Hydrogen Energy 2016;41:7960–7. https://doi.org/10.1016/j.ijhydene.2015.12.221.
[6] Dincer I. Technical, environmental and exergetic aspects of hydrogen energy systems. Int J Hydrogen Energy 2002;27:265–85. https://doi.org/10.1016/S0360-3199(01)00119-7.
[7] Othman E-SA, Nawar SK, Fahmy FH, El-Shafy A, Nafeh A. A new design of a hydrogen fueling station powered by renewable energy sources. IOSR J Electr Electron Eng Ver. II 2015;10. https://doi.org/10.9790/1676-1052116125. 2278–1676.
[8] Won W, Kwon H, Han JH, Kim J. Design and operation of renewable energy sources based hydrogen supply system: technology integration and optimization. Renew Energy 2017;103:226–38. https://doi.org/10.1016/j.renene.2016.11.038.
[9] Uddin MN, Daud WMAW. Technological diversity and economics: coupling effects on hydrogen production from biomass. Energy Fuels 2014;28:4300–20. https://doi.org/10.1021/ef5007808.
[10] Temiz M, Javani N. Design and analysis of a combined floating photovoltaic system for electricity and hydrogen production. Int J Hydrogen Energy 2020;45:3457–69. https://doi.org/10.1016/j.ijhydene.2018.12.226.
[11] Gökçek M, Kale C. Techno-economical evaluation of a hydrogen refuelling station powered by Wind-PV hybrid power system: a case study for İzmir-çeşme. Int J Hydrogen Energy 2018;43:10615–25. https://doi.org/10.1016/j.ijhydene.2018.01.082.
[12] Samy MM, Barakat S, Ramadan HS. Techno-economic analysis for rustic electrification in Egypt using multi-source renewable energy based on PV/wind/FC. Int J Hydrogen Energy 2020;45:11471–83. https://doi.org/10.1016/j.ijhydene.2019.04.038.
[13] Al-Sharafi A, Sahin AZ, Ayar T, Yilbas BS. Techno-economic analysis and optimization of solar and wind energy systems for power generation and hydrogen production in Saudi Arabia. Renew Sustain Energy Rev 2017;69:33–49. https://doi.org/10.1016/j.rser.2016.11.157.
[14] Kalinci Y, Hepbasli A, Dincer I. Techno-economic analysis of a stand-alone hybrid renewable energy system with hydrogen production and storage options. Int J Hydrogen Energy 2015;40:7652–64. https://doi.org/10.1016/j.ijhydene.2014.10.147.
[15] Qolipour M, Mostafaeipour A, Tousi OM. Techno-economic feasibility of a photovoltaic-wind power plant construction for electric and hydrogen production: a case study. Renew Sustain Energy Rev 2017;78:113–23. https://doi.org/10.1016/j.rser.2017.04.088.
[16] Ma T, Yang H, Lu L, Peng J. Pumped storage-based standalone photovoltaic power generation system: modeling and techno-economic optimization. Appl Energy 2015;137:649–59. https://doi.org/10.1016/j.apenergy.2014.06.005.
[17] Shiroudi A, Taklimi SRH. Demonstration project of the solar hydrogen energy system located on Taleghan-Iran: technical-economic assessments. In: Proc. World Renew. Energy Congr. – Sweden, 8–13 May, 2011, Linköping, Sweden, 57; 2011. p. 1158–65. https://doi.org/10.3384/ecp110571158.
[18] Silva SB, Severino MM, De Oliveira MAG. A stand-alone hybrid photovoltaic, fuel cell and battery system: a case study of Tocantins, Brazil. Renew Energy 2013;57:384–9. https://doi.org/10.1016/j.renene.2013.02.004.
[19] Khadem T, Billah SMB, Barua S, Hossain MS. HOMER based hydrogen fuel cell system design for irrigation in Bangladesh. In: 4th Int. Conf. Adv. Electr. Eng. ICAEE 2017. 2018-Janua; 2017. p. 445–9. https://doi.org/10.1109/ICAEE.2017.8255397.
[20] HOMER Energy LLC. HOMER pro version 3.7 user manual. HOMER Energy; 2016. p. 416. http://www.homerenergy.com/pdf/HOMERHelpManual.pdf.
[21] Alazemi J, Andrews J. Automotive hydrogen fuelling stations: an international review. Renew Sustain Energy Rev 2015;48:483–99. https://doi.org/10.1016/j.rser.2015.03.085.
[22] Mohammed OH, Amirat Y, Benbouzid M, Elbaset AA. Optimal design of a PV/fuel cell hybrid power system for the city of Brest in France. In: 2014 1st Int. Conf. Green Energy, ICGE 2014; 2014. p. 119–23. https://doi.org/10.1109/ICGE.2014.6835408.
[23] John A, Basu S, Akshay Kumar A. Design and evaluation of stand-alone solar-hydrogen energy storage system for academic institute: a case study. Mater Today 2021. https://doi.org/10.1016/j.matpr.2021.04.461 [in press].
[24] Hall JL. Cell components. Phytochemistry 1987;26:1235–6. https://doi.org/10.1016/s0031-9422(00)82398-5.
[25] Rohani A, Mazlumi K, Kord H. Modeling of a hybrid power system for economic analysis and environmental impact in HOMER. In: Proc. - 2010 18th Iran. Conf. Electr. Eng. ICEE 2010, 623; 2010. p. 819–23. https://doi.org/10.1109/IRANIANCEE.2010.5506962.
[26] Chauhan A, Saini RP. Techno-economic optimization based approach for energy management of a stand-alone integrated renewable energy system for remote areas of India. Energy 2016;94:138–56. https://doi.org/10.1016/j.energy.2015.10.136.
[27] Prasetyaningsari I, Setiawan A, Setiawan AA. Design optimization of solar powered aeration system for fish pond in Sleman Regency, Yogyakarta by HOMER software. Energy Procedia 2013;32:90–8. https://doi.org/10.1016/j.egypro.2013.05.012.
[28] Li C, Ge X, Zheng Y, Xu C, Ren Y, Song C, et al. Techno-economic feasibility study of autonomous hybrid wind/PV/battery power system for a household in Urumqi, China. Energy 2013;55:263–72. https://doi.org/10.1016/j.energy.2013.03.084.
[29] Shukla AK, Sudhakar K, Baredar P. Design, simulation and economic analysis of standalone roof top solar PV system in India. Sol Energy 2016;136:437–49. https://doi.org/10.1016/j.solener.2016.07.009.
[30] Chen F, Duic N, Manuel Alves L, da Graça Carvalho M. Renewislands-Renewable energy solutions for islands. Renew Sustain Energy Rev 2007;11:1888–902. https://doi.org/10.1016/j.rser.2005.12.009.

Please cite this article as: Basu S et al., Design and feasibility analysis of hydrogen based hybrid energy system: A case study, International Journal of Hydrogen Energy, https://doi.org/10.1016/j.ijhydene.2021.08.036

Arabian Journal for Science and Engineering
https://doi.org/10.1007/s13369-021-06157-7

RESEARCH ARTICLE-PHYSICS

Design and Investigation of Recessed-T-Gate Double Channel HEMT with InGaN Back Barrier for Enhanced Performance

Megha Sharma[1] · **Rishu Chaujar**[1]

Received: 2 May 2021 / Accepted: 29 August 2021
© King Fahd University of Petroleum & Minerals 2021

Abstract
In this paper, a novel recessed T-gate double channel HEMT with InGaN back barrier has been designed and studied. The double channel is formed by the AlN layer's insertion below the InAlN/GaN interface, enhancing the drain current density (2.5A/mm). However, double channel HEMT suffers from a lack of gate controllability, due to which device performance decreases. This problem has been resolved by using the InGaN as a back barrier which enhances the carrier confinement of 2DEG by raising the conduction band for the GaN buffer and significantly enhances the gate controllability over a lower channel. Precisely a brief comparison has been done of the proposed device with conventional double channel HEMT. Performance parameters like drain current (I_{ds}), transconductance (g_m), intrinsic gain (A_v), output conductance (g_d), early voltage (V_{EA}), TGF, parasitic capacitances (C_{gs}, C_{gd}), cut-off frequency (f_T), maximum oscillation frequency (f_{max}), gain frequency product (GFP), transconductance frequency product (TFP), and gain transconductance frequency product (GTFP) have been investigated. Further, the improvement in linearity parameters such as VIP2, VIP3, IIP3, and 1-dB compression point has been observed by varying the back barrier distance from the lower channel.

Keywords Analog · Back barrier · Double channel · Heterostructure · Linearity · Recessed gate

1 Introduction

In recent years, the III-Nitride-based HEMTs attracted extensive attention in high power RF applications. The GaN material shows a high power output density than the GaAs at high frequency operation due to its larger bandgap, higher saturation velocity, and higher electron mobility [1–3]. Several research papers have been published to enhance the performance of AlGaN/GaN HEMTs [1–8]. To improve the RF performance of GaN HEMTs, the scaling of device dimension is required. The reduction of AlGaN barrier layer produces the unwanted strain at the lattice mismatched AlGaN/GaN junction. The AlGaN layer successfully replaced by the lattice matched InAlN layer with Al composition 0.83 [9] and significantly enhances the device performance. To achieve finer RF performance, it is important to reduce the gate length as the gate length reduces the parasitic capacitance reduces (C_{gs}, C_{gd}), which significantly uplifts the RF parameters. Much work has already been done on T-gate single channel HEMT with different types of buffer and back barrier engineering. Dong Seup Lee et al. demonstrated the 65 nm gate length passivated by Al_2O_3 layer with a back barrier layer, exhibits the drain current 1.49 A/mm and g_m of 539 mS/mm [10]. Sarosij Adak et al. designed the p-GaN back barrier HEMT with 150 nm gate length and recorded the drain current density 1.24 A/mm, g_m of 589 mS/mm [11], 70 nm gate length HEMT with graded buffer designed by Tieching Han et al. recorded the drain current 1.4 A/mm [12]. Although using the back barrier enhances the RF performance, but it also reduces the drain current density.

Double channel HEMTs (DC-HEMT) [13, 14] have recently emerged as a better alternative to single-channel HEMT. L. Ravikiran et al. demonstrated the AlGaN/GaN double heterojunction HEMT and exhibited the f_T/f_{max} 22/25 GHz [15]. Nisha Chugh et al. show the polarization-dependent charge control model for AlGaN/GaN DC-HEMT [14]. A DC-HEMT comprises two heterointerfaces which lead to forming the two 2DEGs. However, the DC-HEMT shows high current drivability due to the formation of two

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Megha Sharma
meghasharma_phd2k18@dtu.ac.in

[1] Department of Applied Physics, Delhi Technological University, Delhi, India

Published online: 18 September 2021

channels which is advantageous in RF circuit design [16, 17]. However, DC-HEMT suffers from the poor confinement of 2DEG and poor gate controllability over the lower channel, leading to the severe short channel effect. We need some buffer engineering to restrict the spilling of 2DEG in the buffer. The InGaN back barrier engineering was very impressive [18, 19]. It effectively reduced the short channel effect by confining the 2DEG in channel region. The gate recessed engineering further enhances the gate controllability over the channel and shift the threshold voltage toward the positive side.

This paper has proposed the recessed T-gate InAlN/AlN double channel HEMT with InGaN back barrier (DC-BB-HEMT). The analog and RF performance of double channel HEMT with a back barrier and DC-HEMT without a back barrier has been investigated. It has been observed that the proposed device's analog and RF performance are improved by introducing the back barrier. The effect of back barrier distance from the lower channel on linearity parameters has also been investigated.

2 Device Structure

Figure 1 exhibits the schematic structure of recessed 70 nm T-gate InAlN/AlN DC-BB-HEMT. All the parameters are defined in Table 1. The Al_2O_3 layer passivates the device to reduce the parasitic capacitance. The drain and source regions are doped with Si of order 10^{20} cm^{-3} to reduce the contact resistance. The first AlN layer placed below the InAlN layer enhances the mobility of 2DEG [20]. The distance between the first and second AlN layer is 10 nm. The second AlN layer formed the second triangular quantum well at the interface of AlN (2)/GaN and formed the second channel. The 2 nm InGaN back barrier layer is placed below the lower channel to enhance the 2DEG confinement [18, 19].

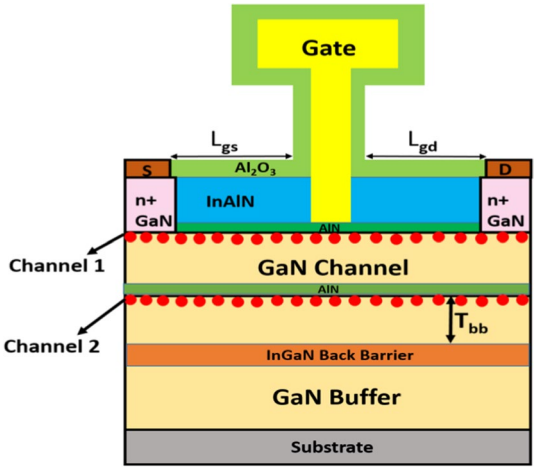


Fig. 1 Proposed structure of InAlN/AlN T-gate DC-BB-HEMT

Table 1 Design parameters of DC-BB-HEMT and DC- HEMT

Device Parameter	DC-HEMT	DC-BB-HEMT
Gate length (L_g)	70 nm	70 nm
Al_2O_3 passivation layer	10 nm	10 nm
Distance b/w two channels	10 nm	10 nm
Distance b/w lower channel and back barrier (T_{bb})	30 nm	30 nm
Source to gate length (L_{gs})	400 nm	400 nm
Gate to drain length (L_{gd})	1.5 µm	1.5 µm
Barrier layer (InAlN)	10 nm	10 nm
AlN layer	1 nm	1 nm
InGaN back barrier	–	2 nm
GaN buffer	1.5 µm	1.5 µm
Source/drain doping	1×10^{20} cm^{-3}	1×10^{20} cm^{-3}
Gate width	2×50 µm	2×50 µm
Gate stem height	100 nm	100 nm
Gate head	400 nm	400 nm

3 Experimental Calibration and Models

The Atlas-Silvaco simulator has been used for simulating the proposed structure [21]. The models which are used in simulating the proposed structure are low field mobility, Fermi–Dirac, high field mobility, polarization, and Shockley–Read–Hall model. The electron trapping mechanism has been calculated by the SRH model. The mobility models are used to calculate the scattering mechanism [21]. Both models are defined in Eq. (1) and (2), respectively.

1. High Field Mobility Model

$$\mu(E) = \frac{\mu(N,T) + V_{sat}\frac{E(N_1-1)}{E_C{}^{N_1}}}{1 + a_n\left(\frac{E}{E_C}\right)^{N_2} + \left(\frac{E}{E_C}\right)^{N_1}} \tag{1}$$

where E and v_{sat} are the electric fields and saturated velocity of electrons and E_C, a_n, N_1, N_2 are coefficients.

2. Low Field Mobility Model

$$\frac{1}{\mu(T,\ N)} = a\left(\frac{N}{N^{ref}}\right)\left(\frac{T}{300}\right)^{-15} \times \ln\left[1 + 3\left(\frac{T}{300}\right)^{2}\left(\frac{N}{N^{ref}}\right)^{\frac{-2}{3}}\right] + b\left(\frac{T}{300}\right)^{1.5} + \frac{c}{\left(e^{\frac{1065}{T}}\right) - 1} \tag{2}$$

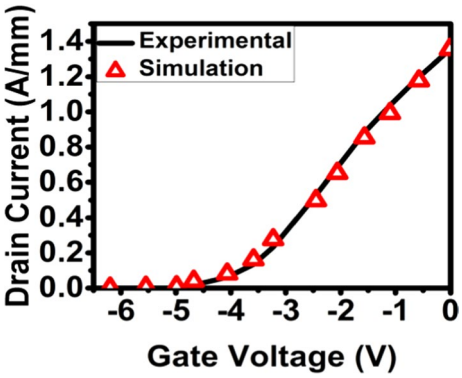


Fig. 2 Show the transfer characteristics of InAlN/AlN T-gate HEMT with $L_g = 70$ nm, $V_d = 5$ V and $W_g = 2 \times 40$ µm compare with the experimental data

where µ(T, N) is the mobility as a function of the doping and ambient temperature.

3. Total Polarization Model

In GaN-based devices, the two types of polarization have been induced, i.e., piezoelectric and spontaneous polarization. The total polarization model is defined in Eq. (3) [22]:

$$P_{total} = [(P_{PE} + P_{SP})(bottom)] - [(P_{PE} + P_{SP})(top)] \quad (3)$$

The above-mentioned models have been calibrated with the experimental paper of Han Tingting et al. [23]. Figure 2 exhibits the simulated and experimental I_d-V_g curve of 70 nm T-gate InAlN/GaN HEMT at $V_{ds} = 5$ V. It has been observed that the experimental result exhibits a good agreement with the simulated result.

4 Results and Discussion

4.1 Band Diagram and Electron Concentration

In this section, we have investigated the conduction band discontinuity and electron concentration of both devices. Figure 3 exhibits the conduction band diagram of both devices. Two quantum wells, i.e., primary quantum well and secondary quantum, are formed in the proposed device. The primary quantum well is generate at the interface of AlN (1)/GaN. This quantum depth is relatively deeper, which generates the major upper channel region and accumulates the highest electron concentration. The second quantum well is formed at the interface of AlN (2)/GaN, which generates the second channel. In Fig. 3, it can be seen that in DC-BB-HEMT, a sharp notch is formed at the back of the lower channel. This notch is formed by the induced polarization

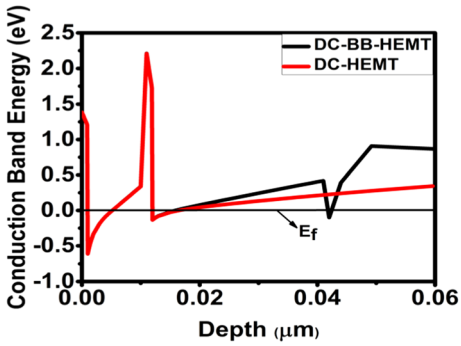


Fig. 3 Conduction band diagram of DC-BB-HEMT & DC-HEMT

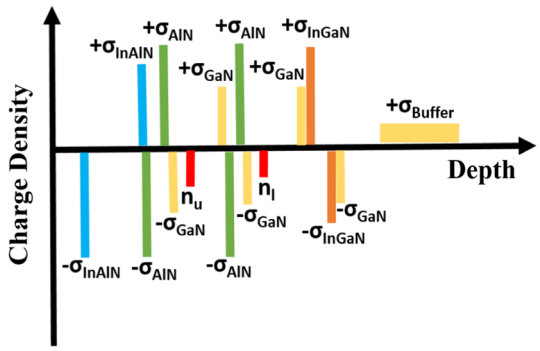


Fig. 4 Polarization charge distribution of each layer in DC-BB-HEMT

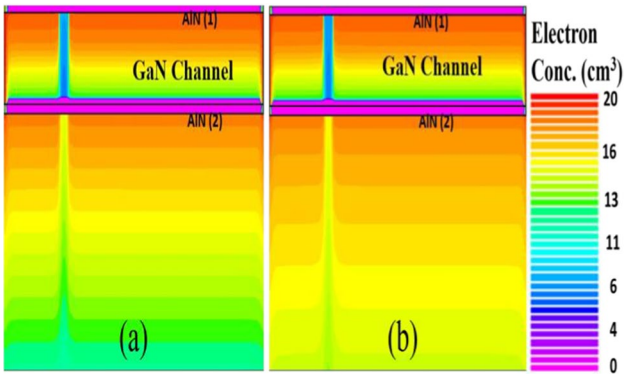


Fig. 5 Electron concentration contour plot of **a** DC-BB-HEMT **b** DC-HEMT

charges between the InGaN and GaN, which raises the potential barrier and confines the electron in the channel region, thereby enhancing the gate controllability over the lower channel.

Figure 4 exhibits the charge distribution of DC-BB-HEMT. Piezoelectric and spontaneous polarization charges are induced at the interface of layers. Due to the polarization effect, negative and positive polarization charges are induced on each

layer's opposite sides. The value of these polarization charges is calculated using [22]. Figure 5a and b show the electron concentration contour plot of DC-BB-HEMT and DC-HEMT. The higher electron concentration has been observed at AlN (1)/GaN heterojunction interface, which formed the major channel, whereas at the AlN (2)/GaN interface, the minor channel is formed. In Fig. 5a, the electron concentration below the lower channel is reduced more significantly than the Fig. 5b, which attributes to the fact that the insertion of InGaN layer restricts the diffusion and leakage of electron toward the buffer layer and considerably enhances the electron confinement.

4.2 Analog Performance of DC-BB-HEMT & DC-HEMT

Figure 6 reflects the drain current characteristics for DC-BB-HEMT and DC-HEMT devices. Initially, channel 2 is in ON-state due to the lack of gate controllability over channel 2 in DC-HEMT, and the negative threshold voltage is observed. As the gate voltage is increased, channel 1 is turned on, which is a major channel as it has a higher 2DEG density. When channel 1 is turned on, it screened channel 2 and saturates the 2DEG in channel 2. The InGaN back barrier's presence enhances the 2DEG confinement, increases the gate controllability over channel 2, and significantly shifts the threshold voltage toward the positive side. The drain current density in DC-BB-HEMT is (2.2A/mm) is obtained at $V_{gs}=3$ V, which is about 16.6% lower than the DC-HEMT (2.59A/mm). This decrement of current density due to the shift of threshold voltage of DC-BB-HEMT (0.3 V) toward the positive side about 1.5 V compared with DC-HEMT (− 1.8 V).

In this subsection, we have investigated the analog parameters such as g_m, TGF, g_d, V_{EA}, and A_v of the proposed device for the analog applications and compared them with the conventional device. The mathematical formulation of these parameters is shown below [24]:

$$g_m = \partial I_d/\partial V_{gs} \tag{4}$$

$$\mathrm{TGF} = g_m/I_d \tag{5}$$

$$g_d = \partial I_d/\partial V_{ds} \tag{6}$$

$$V_{EA} = I_d/g_d \tag{7}$$

$$A_v = g_m/g_d = (g_m/I_d) \times V_{EA} \tag{8}$$

Figure 7a shows the plot of transconductance with gate voltage for both devices. Transconductance is defined as the change in the drain current to the gate voltage at constant drain voltage. In Fig. 7a, double peaks, i.e., primary and secondary peaks, have been observed in the transconductance graph due to the formation of two channels. The higher primary peak is observed for the DC-BB-HEMT at $V_{gs}=0.8$ V due to the improvement of electron confinement and gate controllability for channel 2 by the InGaN back barrier. Simultaneously, the secondary peak for both the devices is comparable, which shows that the InGaN back barrier layer does not affect channel 1. Figure 7b exhibits the plot of transconductance generation factor (TGF) for both devices. It is an essential analog parameter which is defined as the ratio of transconductance and drain current. TGF also expressed as available gain/dissipate power [25, 26], where gain and power dissipation are represented by the g_m and drain current. The minimum value of TGF will lower the input driveability and dissipate the high power at the load. In Fig. 7b, at smaller $V_{gs,}$ high peak of TGF is obtained for DC-BB-HEMT (32 V^{-1}), which is about 10 times larger than the DC-HEMT (1.73 V^{-1}). The TGF value

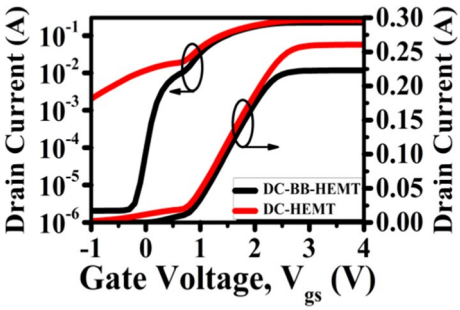


Fig. 6 Drain current characteristics of DC-BB-HEMT and DC-HEMT

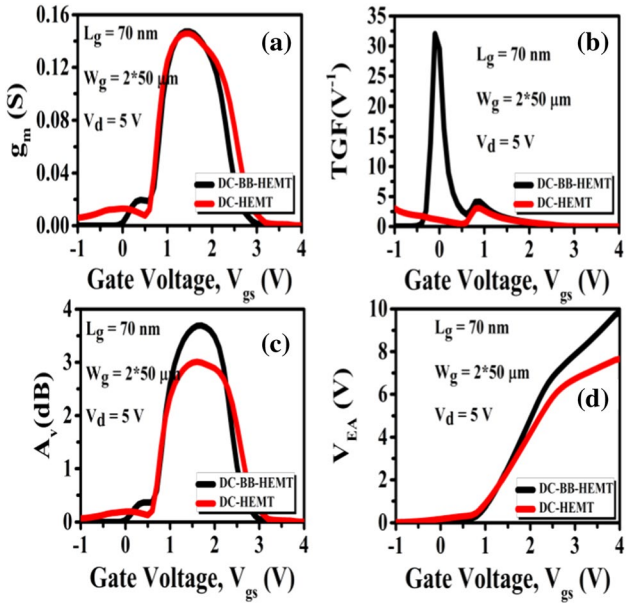


Fig. 7 **a** Transconductance **b** TGF **c** Intrinsic gain **d** Early voltage for both devices

is decreased as the gate bias increases, because of the enhancement of drain current at higher gate voltage.

Figure 7c shows the variation of intrinsic gain with gate bias. The high peak of intrinsic gain for DC-BB-HEMT (4 dB) observed at $V_{gs} = 1.5$ V. Intrinsic gain depends on the two factors: transconductance and output conductance. The decrement of output conductance significantly enhances the intrinsic gain of about 50% in the proposed device. Figure 7d exhibits the early voltage variation with gate bias for both structures. Early voltage is expressed as the ratio of drain current and output conductance. For better analog performance, both intrinsic gain and early voltage should be high. The early voltage for DC-BB-HEMT is 10 V observed, which is 53% higher than the DC-HEMT (6.5 V). This enhancement in the early voltage is because of the lower output conductance, which enhances the proposed structure's analog performance. The plot of output conductance for DC-BB-HEMT and DC-HEMT with drain bias is exhibited in Fig. 8. Output conductance calculates the driving ability of the device. At lower drain voltage, the higher value of output conductance is obtained. As the V_{ds} increases, the output conductance is decreased exponentially. Due to the higher gate controllability, minimum output conductance is observed for DC-BB-HEMT as compared to DC-HEMT.

4.3 RF Performance of DC-BB-HEMT and DC-HEMT

In this section, small-signal AC analysis has been performed with an operating frequency of 1 MHz for DC-BB-HEMT and DC-HEMT. All the RF parameters such as f_T, f_{max}, GFP, GTFP, and TFP have been investigated. The mathematical formulation of these parameters is listed below [26]:

$$f_T \approx g_m/2\pi(C_{gs} + C_{gd}) \tag{9}$$

$$f_{max} = f_T/\sqrt{4R_g(g_{ds} + 2\pi f_T C_{gd})} \tag{10}$$

$$\text{GFP} = (g_m/g_d) \times f_T \tag{11}$$

$$\text{TFP} = (g_m/I_d) \times f_T \tag{12}$$

$$\text{GTFP} = (g_m/g_d) \times (g_m/I_d) \times f_T \tag{13}$$

Figure 9a and b exhibit the plot of the gate to source capacitance (C_{gs}) and gate to drain capacitance (C_{gd}) with gate bias. For better RF performance, these parasitic capacitances should be low. At lower gate voltage, there is a slow increment in C_{gs} and C_{gd}, but at higher gate bias, there is a quick increment in C_{gs} and C_{gd} because of the enhancement of lateral electric field, which enhances the carrier movement from drain to the source side. The DC-BB-HEMT shows a lower value of C_{gs} and C_{gd} in comparison with the DC-HEMT. At the metal/InAlN interface, charges increase due to the higher sheet charge density in the quantum well, which enhances the capacitance for DC-HEMT.

Figure 9c exhibits the plot of cut-off frequency (f_T) with gate voltage. The cut-off frequency is defined as the frequency at which current gain reduces to unity. From Eq. (9), it is seen that f_T depends upon transconductance and parasitic capacitance. The double peak structure has been observed, the same as the transconductance curve Fig. 7a. The primary and secondary peaks are formed due to channel 2 and channel 1. The InGaN back barrier's presence increases the primary transconductance peak, which further enhances the cut-off frequency. The secondary transconductance peak is not affected by the InGaN back barrier, but the parasitic capacitance for the DC-BB-HEMT is lower, enhancing the cut-off frequency. The cut-off frequency obtained for DC-BB-HEMT (153 GHz) is 11.7%

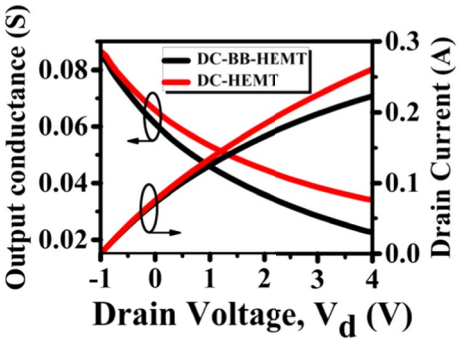


Fig. 8 Output conductance for both devices

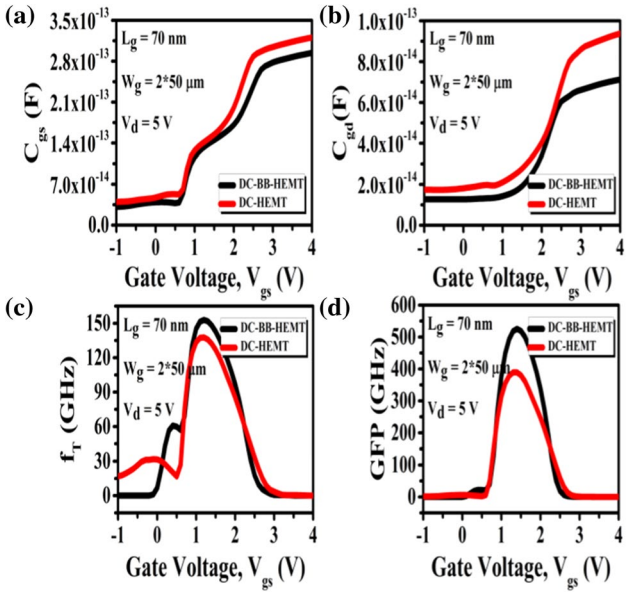


Fig. 9 Variation of **a** Cgs **b** Cgd **c** fT **d** GFP for both devices

higher than the DC-HEMT (136 GHz). In Fig. 9d, the gain frequency product (GFP) has been plotted against the gate bias. The GFP is another essential parameter for the high frequency application. Initially, with increasing the gate voltage, GFP increases linearly and holds an optimum value, after that, it decreases in the saturation region. The higher peak of GFP obtained for DC-BB-HEMT is 526 GHz which is about 35.2% higher than the DC-HEMT (389 GHz). GFP improvement in DC-BB-HEMT is mainly due to the increment in cut-off frequency, transconductance, and reduction of output conductance.

Variation of the drain to source conductance (g_{ds}) and maximum oscillation frequency (f_{max}) with gate bias is shown in Fig. 10a and b. The maximum oscillation frequency is defined as the frequency at which max unilateral gain reduces to unity. Equation (10) shows that the f_{max} depends upon parasitic capacitance, drain to source conductance (g_{ds}), and f_T. The 2 times higher f_{max} is obtained for DC-BB-HEMT (92 GHz) in comparison with DC-HEMT (36 GHz) due to the lower parasitic capacitance (C_{gd}), g_{ds}, and higher f_T.

Variation of transconductance frequency product (TFP) with gate bias is shown in Fig. 10c. TFP is utilized in moderate to high-speed design. In Fig. 10c, higher primary and secondary peaks are observed for the DC-BB-HEMT than the DC-HEMT due to the higher g_m, f_T, and lower I_d. Gain transconductance frequency product (GTFP) is another essential FOM that gives critical information regarding the circuit designing by determining the perfect region to achieve the tradeoff among the gain, transconductance, and device speed. In Fig. 10d, the higher value of GTFP is observed for DC-BB-HEMT (1310 GHz/V), which is about 57% higher than the DC-HEMT (828 GHz/V). The improvement of these parameters for DC-BB-HEMT is mainly due to the improvement of g_m and reduction of output conductance.

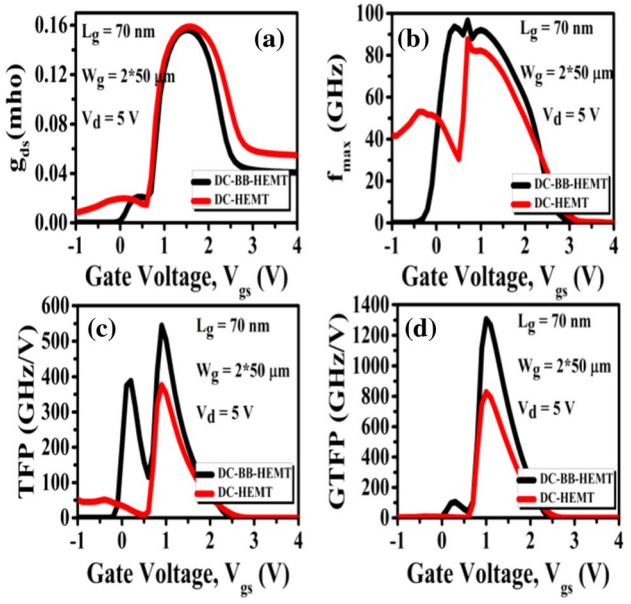


Fig. 10 Variation of **a** gds **b** fmax **c** TFP **d** GTFP with gate bias

4.4 Linearity Analysis of DC-BB-HEMT for Variable Tbb

Figure 11a and b exhibit the drain current and transconductance characteristics for different T_{bb} values. It has been observed that while increasing the BB distance from the lower channel, the drain current and transconductance characteristics improved. As the back barrier distance increases from 10 to 30 nm, the drain current density also increases. A higher value of drain current is observed 2.3 A/mm for 30 nm T_{bb}. On increasing the back barrier distance, the second quantum depth increases, which results in the enhancement of 2DEG concentration. Improvement in transconductance characteristics on increasing the back barrier distance is basically due to the lower scattering and higher mobility of 2DEG. When the InGaN back barrier is placed near the lower channel, the mobility of 2DEG is badly affected by the back barrier due to the alloy disorder scattering mechanism.

The investigation of nonlinear parameters is very important as they are the primary source of distortion. So to achieve high linearity, the nonlinear parameters should be low. The second- and third-order transconductance parameters (gm_2, gm_3) are the nonlinear parameters. Figure 11c and d exhibits the plot of nonlinear parameters with gate voltage for different back barrier distance. It has been observed that by increasing the T_{bb} from 10 to 30 nm, the nonlinear parameters are almost comparable, which shows increasing the back barrier distance does not affect the nonlinear parameters. The linearity parameters are mathematically expressed as [27].

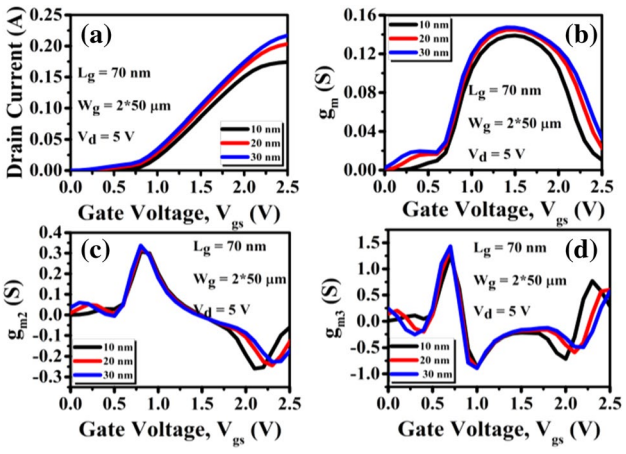


Fig. 11 **a** Transfer characteristics **b** transconductance **c** gm2 **d** gm3 for both devices

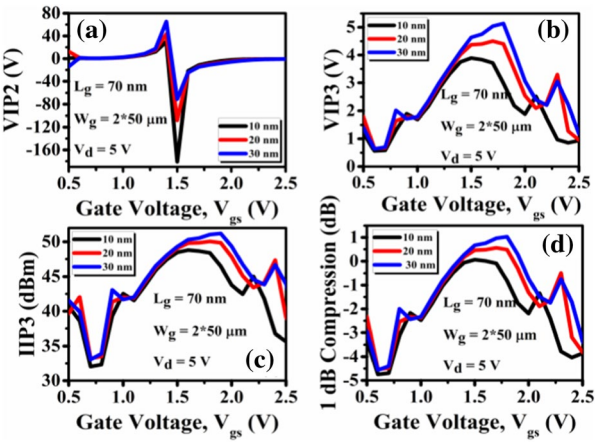


Fig. 12 Variation of linearity parameters with gate bias for different back barrier distance **a** VIP2 **b** VIP3 **c** IIP3 **d** 1-dB compression point

$$\mathrm{VIP2} = 4 \times \left(g_m/g_{m2}\right) \tag{14}$$

$$\mathrm{VIP3} = \sqrt{24 \times \left(g_m/g_{m3}\right)} \tag{15}$$

$$\mathrm{IIP3} = (2/3) \times \left(g_m/\left(g_{m3}R_s\right)\right) \tag{16}$$

$$1-\mathrm{dB\ compression\ point} = 0.22 \times \sqrt{\left(g_m/g_{m3}\right)} \tag{17}$$

VIP2 and VIP3 express the extrapolated gate voltage amplitude at which second- and third-order harmonics equal the fundamental tone in device current (I_{ds}). These FOMs can accurately investigate the distortion characteristics from DC parameters. In order to attain high linearity and low distortion, these parameters should be as high as possible. Figure 12a and b exhibit the graph of VIP2 and VIP3 with gate bias for different back barrier distance. It has been observed that by increasing the T_{bb} from 10 to 30 nm, VIP2 and VIP3 increase. The higher peak of VIP2 and VIP3 is observed 65.3 V and 5.1 V for 30 nm T_{bb} which is 2 times higher than the 10 nm T_{bb} (29 V and 2.9 V). As shown in Eqs. (14) and (15), VIP2 and VIP3 parameters depend upon the transconductance and the second- and third-order transconductance. Due to the lower scattering disorder and higher mobility of 2DEG, higher g_m has been obtained for 30 nm T_{bb}.

Third-order input intercept point (IIP3) and 1-dB compression point are other essential FOMs investigating linearity performance. IIP3 exhibits the extrapolated input power at which first and third harmonics are equal. Figure 12c shows the variation of IIP3 with gate bias for variable T_{bb}. The IIP3 parameter also increases on increasing the back barrier distance from the lower channel. The higher peak of IIP3 observed at $V_{gs} = 1.8$ V is 51 dBm for 30 nm T_{bb} which is about 10% higher than the 10 nm T_{bb} (46 dBm). 1-dB compression point is defined as the input power value at which the gain of the amplifier drops by 1-dB. The study of this parameter is essential for the amplifier circuit. It estimates the maximum input power that the circuit can handle by providing a fixed amount of gain. As the input power surpasses the 1-dB compression point, then gain starts decreasing. Thus, it is important to have a high 1-dB point. Figure 12d exhibits the higher peak of 1 dB compression point for 30 nm (1.3 dB), which is 8 times higher than the 10 nm (−8.1 dB) and emphasizes the superior linearity performance of DC-BB- HEMT with 30 nm T_{bb}. Thus, results reflect that RF harmonic distortion and linearity improve by increasing the InGaN back barrier distance from the lower channel.

5 Conclusion

The work presented in this paper concentrates on the impact of the InGaN back barrier on the analog/RF and linearity performance of the double channel T-gate HEMT. The proposed device exhibited improved results as compared to the conventional device. For analog performance, intrinsic gain increased by 50%, TGF increased by 477%, and early voltage was enhanced by 23%. The cut-off frequency increased by 11.7% for RF performance, maximum oscillation frequency raised by 10%, GFP, TFP, and GTFP increased by 36%, 44%, and 57%, respectively. The proposed device's linearity performance with varying back barrier distance shows the improved results for the 30 nm back barrier distance. Thus, inserting the InGaN back barrier in double channel HEMT enhances the device's analog/RF performance. The careful optimization of the back barrier distance further enhances the device's linearity performance.

Acknowledgements Authors are grateful to Microelectronics Lab, Applied Physics Deptt., DTU, to support the work.

Authors' Contribution All the authors to this study are conception and design.

Funding No funding has been received to carry out this study.

Availability of Data and Material The above-mentioned authors have all the relevant data associated with this research work and will be dedicated to share that, if they will be asked to do so in future.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

Ethical Approval All the Ethical Standards have been seen by the authors and will supposed to follow them in future as well.

References

1. Mishra, U.K.; Parikh, P.; Wu, Y.F.: AlGaN/GaN HEMTs - An overview of device operation and applications. Proc. IEEE **90**(6), 1022–1031 (2002). https://doi.org/10.1109/JPROC.2002.1021567
2. Mari, D.: Cermets and Hardmetals. Encycl. Mater. Sci. Technol. **96**(2), 1118–1122 (2001). https://doi.org/10.1016/b0-08-043152-6/00209-6
3. Frayssinet, E., et al.: High electron mobility in AlGaN/GaN heterostructures grown on bulk GaN substrates. Appl. Phys. Lett. **77**(16), 2551–2553 (2000). https://doi.org/10.1063/1.1318236
4. Feng, Z.H.; Zhou, Y.G.; Cai, S.J.; Lau, K.M.: Enhanced thermal stability of the two-dimensional electron gas in GaN/AlGaN/GaN heterostructures by Si3N4 surface-passivation-induced strain solidification. Appl. Phys. Lett. **85**(22), 5248–5250 (2004). https://doi.org/10.1063/1.1828231
5. Lin, Z.; Kim, H.; Lee, J.; Lu, W.: Thermal stability of Schottky contacts on strained AlGaN/GaN heterostructures. Appl. Phys. Lett. **84**(9), 1585–1587 (2004). https://doi.org/10.1063/1.1650875
6. Kumar, S. P.; Agarwal, A.; Chaujar, R.; Gupta, R.S.: Device linearity and intermodulation distortion comparison of dual material gate and conventional AlGaN/GaN high electron mobility transistor. *Microelectron*, 51, 587–596, 2011.
7. Ahmed, H.M.; Chaujar, R.: Rapid detection of biomolecules in a dielectric modulated GaN MOSHEMT. J. Mater. Sci.: Mater. Electron. **31**(19), 16609–16615 (2020)
8. Sano, S.; Ebihara, K.; Yamamoto, T.; Sato, T.; Miyazawa, N.: GaN HEMTs for wireless communication. SEI Tech. Rev. **86**, 65–70 (2018)
9. Neuburger, M.; Zimmermann, T.; and Kohn, E.: Unstrained InAlN/GaN HEMT Structure. 14(3), 785–790, 2004.
10. Lee, D.S.; Gao, X.; Guo, S.; Palacios, T.: InAlN/GaN HEMTs with AlGaN back barriers. IEEE Electron Device Lett. **32**(5), 617–619 (2011). https://doi.org/10.1109/LED.2011.2111352
11. Adak, S.; Sarkar, A.; Swain, S.; Pardeshi, H.; Pati, S.K.; Sarkar, C.K.: High performance AlInN/AlN/GaN p-GaN back barrier Gate-Recessed Enhancement-Mode HEMT. Superlattices Microstruct. **75**, 347–357 (2014). https://doi.org/10.1016/j.spmi.2014.07.036
12. Han, T.; Zhao, H.; Peng, X.; Li, Y.: Control of short-channel effects in InAlN/GaN high-electron mobility transistors using graded AlGaN buffer. Superlattices Microstruct. **116**, 207–214 (2018). https://doi.org/10.1016/j.spmi.2018.02.031
13. Zhang, W., et al.: Trap state analysis in AlGaN/GaN/AlGaN double heterostructure high electron mobility transistors at high temperatures. Appl. Phys. Lett. **110**(25), 252102 (2017)
14. Chugh, N.; Bhattacharya, M.; Kumar, M.; Deswal, S.S.; Gupta, R.S.: Polarization dependent charge control model for microwave performance assessment of AlGaN/GaN/AlGaN double heterostructure HEMTs. J. Comput. Electron. **17**(3), 1229–1240 (2018). https://doi.org/10.1007/s10825-018-1190-0
15. Ravikiran, L., et al.: Growth and characterization of AlGaN/GaN/AlGaN double-heterojunction high-electron-mobility transistors on 100-mm Si (111) using ammonia-molecular beam epitaxy. J. Appl. Phys. **117**(2), 025301 (2015)
16. Jha, S.K.; Surya, C.; Chen, K.J.; Lau, K.M.; Jelencovic, E.: Low-frequency noise properties of double channel AlGaN/GaN HEMTs. Solid. State. Electron. **52**(5), 606–611 (2008). https://doi.org/10.1016/j.sse.2007.10.002
17. Chu, R.; Zhou, Y.; Liu, J.; Wang, D.; Chen, K.J.; Lau, K.M.: AlGaN-GaN double-channel HEMTs. IEEE Trans. Electron Devices **52**(4), 438–446 (2005)
18. Lee, D.S.; Gao, X.; Guo, S.; Kopp, D.; Fay, P.; Palacios, T.: 300-GHz InAlN/GaN HEMTs With. IEEE Electron Device Lett. **32**(11), 1525–1527 (2011)
19. Palacios, T.; Chakraborty, A.; Heikman, S.; Keller, S.; DenBaars, S.P.; Mishra, U.K.: AlGaN/GaN high electron mobility transistors with InGaN back-barriers. IEEE Electron Device Lett. **27**(1), 13–15 (2006). https://doi.org/10.1109/LED.2005.860882
20. Shrestha, N.M.; Li, Y.; Chang, E.Y.: Simulation study on electrical characteristic of AlGaN/GaN high electron mobility transistors with AlN spacer layer. Jpn. J. Appl. Phys. **53** (2014)
21. ATLAS User's Manual: SILVACO International. CA, USA (2016)
22. Ambacher, O., et al.: Pyroelectric properties of Al (In) GaN/GaN hetero- and. J. Phys. Condens. Matter **14**, 3399–3434 (2002)
23. Wang, R., et al.: Gate-recessed enhancement-mode InAlN/AlN/GaN HEMTs with 1.9-A/mm drain current density and 800-ms/mm transconductance. IEEE Electron Device Lett. **31**(12), 1383–1385 (2010). https://doi.org/10.1109/LED.2010.2072771
24. Kumar, B.; Chaujar, R.: Analog and RF performance evaluation of junctionless accumulation mode (JAM) gate stack gate all around (GS-GAA) FinFET. SILICON **13**(3), 919–927 (2021)
25. Pardeshi, H.: Superlattices and microstructures analog/RF performance of AlInN/GaN underlap DG MOS- HEMT. Superlattices Microstruct. **88**, 508–517 (2015)
26. Chaujar, R.; Kaur, R.; Saxena, M.; Gupta, M.; Gupta, R.S.: Laterally amalgamated DUal material GAte Concave (L-DUMGAC) MOSFET for ULSI. Microelectron. Eng. **85**(3), 566–576 (2008). https://doi.org/10.1016/j.mee.2007.10.002
27. Kumar, B.; Chaujar, R.: TCAD Temperature analysis of gate stack gate all around (GS-GAA) FinFET for improved RF and wireless performance. SILICON (2021). https://doi.org/10.1007/s12633-021-01040-4

The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1742-7371.htm

Effect of data packet size on the performance of RIP and OSPF routing protocols in hybrid networks

Neha Jain and Ashish Payal
USICT, Guru Gobind Singh Indraprastha University, Delhi, India, and
Aarti Jain
Netaji Subhas University of Technology, East Delhi Campus, Delhi, India

Received 8 February 2021
Revised 3 April 2021
Accepted 23 April 2021

Abstract

Purpose – The purpose of this study is to calculate the effect of different packet sizes 256, 512, 1,024 and 2,048 bytes on a large-scale hybrid network and analysis and identifies which routing protocol is best for application throughput, application delay and network link parameters for different packet sizes. As the routing protocol is used to select the optimal path to transfer data packets from source to destination. It is always important to consider the performance of the routing protocol before the final network configuration. From the literature, it has been observed that RIP (Routing Information Protocol) and OSPF (Open Shortest Path First) are the most popular routing protocols, and it has always been a challenge to select between these routing protocols, especially for hybrid networks. The efficiency of routing protocol mainly depends on resulting throughput and delay. Also, it has been observed that data packet size also plays an essential role in determining the efficiency of routing protocol.

Design/methodology/approach – To analyse the effect of different packet sizes using two routing protocols, routing information protocol (RIP) and open shortest path first (OSPF) on the hybrid network, require detailed planning. Designing the network for simulate and then finally analysing the results requires proper study. Each stage needs to be understood well for work accomplishment. Thus, the network's simulation and evaluation require implementing the proposed work step by step, saving time and cost. Here, the proposed work methodology is defined in six steps or stages.

Findings – The simulation results show that both routing protocols – RIP and OSPF are equally good in terms of network throughput for all different packet sizes. However, OSPF performs better in terms of network delay than RIP routing protocol in different packet size scenarios.

Research limitations/implications – In this paper, a fixed network of 125 objects and only RIP and OSPF routing protocol have been used for analysis. Therefore, in the future, a comparison of different network sizes can be considered by increasing or decreasing the number of objects in the proposed network. Furthermore, the other routing protocols can be used for performance evaluation on the same proposed network.

Originality/value – The analysis can be conducted by simulation of the network, enabling us to develop a network environment without restricting the selection of parameters as it minimizes cost, network

This work is supported by the Guru Gobind Singh Indraprastha University, New Delhi, India, under the Indraprastha Research Fellowship No.182.

Declaration: We are enclosing herewith a manuscript entitled "Effect of data Packet Size on the Performance of RIP and OSPF Routing Protocol in Hybrid Networks" for publication in your Journal for possible evaluation. We confirm that this paper or any part of this paper has not been submitted or published elsewhere in any form.

Please address all correspondence concerning this manuscript to nehajain312@gmail.com.

Thank you for your consideration of this manuscript.

International Journal of Pervasive Computing and Communications
Vol. 17 No. 4, 2021
pp. 361-376
© Emerald Publishing Limited
1742-7371
DOI 10.1108/IJPCC-02-2021-0036

deployment overhead, human resources, etc. The results are analysed, calculated and compared for each packet size on different routing protocol networks individually and the conclusion is made.

Keywords Hybrid Network, RIP, OSPF, Packet size, Throughput, Delay

Paper type Research paper

1. Introduction

Internet communication has increased globally by using routers and internet protocol (IP) (Geng, 2020). IP allows us to communicate between nodes in different networks that are separated by thousands of kilometres and the networks can be wired, wireless or hybrid. These networks consist of various objects or components connected or distributed such as routers/switches, access points, sensors, power, monitor and control units (Wang *et al.*, 2018a, 2018b). Therefore these networks represent mechanics, electronics, electrical and computer engineering together in one system and the use of these systems is increasing exponentially for various applications used in daily life such as the Internet of Things (IoT) and Robotics (Belda and Rychnovsk`y, 2010).

In hybrid networks, routing is a meaningful, challenging and crucial task. Hybrid networks require the correct delivery of packets due to frequent topology changes between wired or wireless objects (Chauhan and Sharma, 2015; Wijaya, 2011; Alabady *et al.*, 2018; Chai and Zeng, 2019; Akyildiz *et al.*, 2005). In the whole process, routing protocols play a crucial role in ascertaining the network's information. The different routing protocols are routing information protocol (RIP), open shortest path first (OSPF), EIGRP, IS-IS and BGP (Jain and Payal, 2019; Samaan, 2018; Khan *et al.*, 2017; Jain and Payal, 2020). These routing protocols use algorithms to find optimal routing information for the source to destination delivery (Manzoor *et al.*, 2020; Dey *et al.*, 2016; Guo, 2010). This paper has used RIP and OSPF routing protocols in the proposed network as they are used extensively.

In the network, the data is encapsulated into packets where each packet contains the source address, destination addresses and other routing information (Rukmani and Ganesan, 2013; Alotaibi and Mukherjee, 2012). These packets are fixed or variable in size and their size can affect the network's performance. Thus, the change in packet size can change the overall network performance. Also, the performance depends on the maximum transfer unit (MTU) of packets; packets whose size is less than MTU create overhead in the network (Omar *et al.*, 2015). However, for the larger packets where size exceeds the network's maximum packet size, it requires a proper mapping function that operates segmentation and reassembly for appropriate packet transfer. Thus, in this study, it is found that the packet size affects routing as variable packets are generated and transmitted in the network. Therefore, it is to see when, where and why selecting a specific packet size may be advantageous, affecting network performance (Nedyalkov, 2019). In this paper, the network is considered and evaluated simultaneously with variable size packets of 256, 512, 1,024 and 2,048 bytes, consisting of both wired and wireless nodes. The relative analysis is done by simulating the network with each routing protocol – RIP and OSPF on Network Simulator (Netsim). The network is defined for the application constant bit rate (CBR) where the sender and receiver nodes transfer the traffic for this application. Based on packets sending and receiving, the application throughput and application delay have been calculated. The results are observed, evaluated, and compared for all packet sizes individually for different routing protocol networks.

The objectives of the paper are as follows:

- The proposed work calculates the effect of different packet sizes 256, 512, 1,024 and 2,048 bytes on a large-scale hybrid network.
- Analyse and identify which routing protocol is best for application throughput, application delay and network link parameters for different packet sizes.

The rest of the paper is organized as follows. Section 2 gives brief knowledge about routing protocols used in the simulation. Section 3 presents the methodology of the proposed work. Section 4 presents the literature survey of the related work. Section 5 presents the simulation design and result in analysis based on different packet sizes. Section 6 presents the conclusion and future work.

2. Routing information protocol and open shortest path first routing protocols

RIP and OSPF are two famous routing protocols and are widely used. RIP and OSPF both belong to the dynamic routing protocols where the router can gain knowledge from other routers in a network. They both use algorithms to find the shortest optimal path to route packets from source to destination. Whenever the topology of a network changes, RIP sends the complete routing table to the neighbours, while OSPF sends only the part of the routing table that has been changed. RIP is the primary and oldest routing protocol used for smaller simple networks and prevents loops in routing by limiting the number of hops. OSPF is used for more extensive complex networks that need fast convergence and preferable routing traffic (Waqas *et al.*, 2019; D'Arienzo and Romano, 2016; Hameed and Saleh, 2021). OSPF provides efficient, reliable transport and provides scalability but has configuring, managing and memory issues for large areas. The main differences between these two routing protocols are shown in Table 1.

3. Methodology

The performance evaluation requires complete planning to analyze or evaluate the effect of different packet sizes using two routing protocols, RIP and OSPF, on the hybrid network. Designing the network for simulation and then finally analysing the results requires a thorough study. Each stage needs to be understood well for accomplishing the work. Thus, the network's simulation and evaluation require step-by-step planning to save time and cost. Here, the proposed work methodology is defined in 6 steps or stages, which is shown in Figure 1.

4. Literature survey

Various studies have been done to analyse and compare the routing protocols simulated for wired, wireless and hybrid networks (Zhao and Raychaudhuri, 2009; Zoppi *et al.*, 2018).

In 2011, Ismail and Hassan used 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000 bytes different packet sizes on three different scenarios. Scenario 1–5 nodes Homogenous MANET, Scenario 2–7 nodes Heterogenous MANET and Scenario 3–8 nodes Heterogenous MANET. The network has used the AODV routing protocol and the network was simulated using OMNET++. The simulation parameters were throughput and PDR and it is found that the maximum throughput obtained at 1,000 bytes, the minimum has at 100 bytes. However, throughput becomes stable after a certain limit. For throughput, it has been found that scenario 1 has a 13% higher throughput than scenario 2 (consisting of 7 nodes) and 33% higher than scenario 3 (consisting of 8 nodes). For PDR, scenario 1 has 25% higher than scenario 2 (consisting of 7 nodes) and 45% higher than scenario 3 (consisting of 8 nodes). However, maximum PDR at 800 bytes for all scenarios (Ismail and Hassan, 2011). In 2012, Tolani and Mishra used 256, 512, 1,024, 2,048, 3,072, 4,096, 8,192 bits of packet sizes for a MANET network of 50 nodes. The different performance parameters – throughput, delay, data dropped, are calculated and the network is simulated using OPNET using different routing protocols – AODV, OLSR and DSR. It has been found that for congestion less medium, OLSR has best for Real-time applications with packet size 512 bits and AODV has best for high traffic conditions with 1,024 bits packet size. While for congestion less medium, OLSR is best for real-time applications with packet size 512 bits and AODV is best for high traffic conditions for 1,024 bit packet size (Tolani and Mishra, 2012). Mohapatra and

Parameters	RIP			OSPF	
Stands for	Routing Information Protocol			Open Shortest Path First	
Algorithm	Distance-Vector Algorithm/ Bellman-Ford Algorithm			Link State Algorithm/ Dijkstra Algorithm/Shortest Path Algorithm	
Versions	RIPv1, RIPv2, RIPng			OSPFv2, OPSFv3	
Metric	Hop Count			Cost=1/Bandwidth	
Hop Count	15-16			Unlimited	
Complexity	Simple			Complex	
Network	Suitable for small networks			Suitable for large networks	
Loop prevention features	Split Horizon, Route Poisoning, and hold-down timers			Do not have special features but done itself	
Convergence	Slow			Fast	
Subsystem specification	The autonomous system as one subsystem			The autonomous system divides into areas. Area 0/0.0.0.0 - Backbone area or default area	
Update and discovery	Periodically (every 30 seconds)			When required and uses link-state advertisements (LSA'S)	
Summarization	Auto			No-Auto	
Administrative Distance	120			110	
Load Balancing	Supports 6 identical cost paths			Supports 16 identical cost paths but challenging to implement	
Integrity	RIPv1	RIPv2	RIPng	OSPFv2	OSPFv3
	No authentication	Supports MD5	Supports MD5	Supports MD5	IPsec
Transmission	Broadcast Address	Multicast Address/ 224.0.0.9	Multicast Address/ FF02:9	Two IP Multicast addresses/ 224.0.0.5, 224.0.0.6	Two IP Multicast FF02::5, FF02::6
Addressing	Classful	Classless	Classless	Classless	Classless
Port	UDP-520	UDP-520	UDP-521	IP-89	IP-89
RFC	2453, 1388, 1058	2453, 1723, 1388	2080	1247, 1583	2740

Table 1.
Comparison between RIP and OSPF

Kanungo in the same year used 500 bytes packet size for MANET with four varying parameters. The first parameter has the different number of nodes – 10, 20, 30, 40, 50; the second parameter has pause time in seconds – 0, 30, 90, 120, 150, third has the area in square meter – 200, 400, 600, 800, 1,000 and the last has the type of application – CBR, transport – UDP. The network has been simulated using NS2 and different routing protocols used are AODV, DSR, OLSR and DSDV. It has been found that for high random mobile networks and areas less than 600 × 600 sqm, DSR performance is best. For high random mobile networks and areas above 600 × 600sqm, OLSR performance is best (Mohapatra and Kanungo, 2012). In 2015, Chauhan and Sharma used different packet sizes – 256, 512, 1,024, 2,058 bytes for 100 wireless node networks for application – CBR. The network has been simulated for different routing protocols – RIP, RIPng, OSPF, OSPFv3, using EXata/Cyber 1.1 Simulator. The various performance parameters – throughput, jitter, PDR, end-to-end delay, are simulated. It has been found that RIPng gives the best and OSPF gives the worst performance among all protocols for all performance parameters (Chauhan and Sharma, 2015). In 2016, Chandel and Sharma used 512 bytes packet size for a network of 50 nodes of

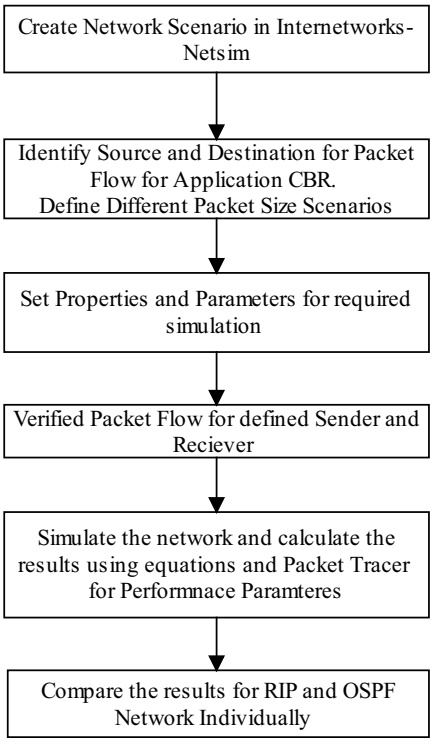


Figure 1.
Methodology flowchart

the wired, wireless and hybrid network. The CBR application is used. The network has been simulated for different routing protocols – RIP, OSPF, RIPng, OSPFv3, AODV using EXata/Cyber 1.1 Simulator. The different performance parameters simulated are – throughput, jitter, PDR, end-to-end delay. It is found that AODV performed best among all wired, wireless and hybrid networks for all performance parameters (Chandel and Sharma, 2016). In 2019, Kumari and Sahana calculated the PDR, end-to-end delay, throughput and packet-loss ratio performance parameters for AODV, DSR, OLSR routing protocols on Netsim software (Kumari and Sahana, 2019). In 2020, Azman *et al.* simulated the real-time network results for the oil and gas field process using the grid WSN network. The network is established for varying nodes, routing protocols and packet sizes to find its impact on research analysis performance. The different performance parameters evaluated are throughput, delivery ratio, routing overhead and fairness index and the network is simulated using NS2. The various networks consist of 32, 128, 512 bytes packet size, grids size – 6×4, 8×6, 10×8, 12×10, 14×12, 16×18 and number of nodes – 18×16 and 24, 48, 80, 120, 168, 224. The authors used the AODV and DSDV routing protocols. It is observed that performances or throughput start degrade as soon as the number of nodes goes beyond 80. As the packet size was small, there are fewer packet loss, better delivery ratio and transmission rate. However, as the packet size increased, the throughput also increased, but it also increased network overhead. DSDV performance is better than AODV in delivery ratio, routing overhead and the number of nodes (Azman *et al.*, 2020). In the same year, P. Sarao, the packet size selection has been made by an ANFIS model built on MATLAB and then it is verified by NS2.35 and applied to the AODV routing protocol. The different performance parameters evaluated are throughput, normalized routing load, data packets received, average end-to-end delay and packet size. The model has been checked for different packet sizes as 256, 512, 1,024, 2,048, 3,072, 4,096 and 4,500 bytes. It has been found that the proposed model gives better results for all parameters and packet size 1,024. The

model also gives optimized results in terms of maximum throughput and minimum end-to-end delay (Sarao, 2020). In 2021, various routing protocol performance has been evaluated in MANET's network (Haglan, 2021; Priyambodo *et al.*, 2021).

Above, the various routing protocol performances are evaluated for various packet sizes and the different simulation software is used for their evaluation. The performance parameters are compared for different routing protocols and these are throughput, end-to-end delay, PDR, control overhead. It is found from the study that considering different packet sizes on the network has an impact on its performance, especially on throughput and PDR. However, none have used the large-scale hybrid network for their analysis using both wired and wireless nodes. Therefore, in this paper, we have analysed the large-scale hybrid network performance in terms of application throughput and delay for four different standard packet sizes. The focus is on two main performance parameters as they are the critical parameters for estimating the quality of service (QoS). The network is simulated on Netsim, which provides the flexibility to use the RIP and OSPF routing protocols on a network.

5. Simulation design and result analysis

The simulation plays a significant role in testing, analysing and developing complex network scenarios before using it for real-time applications. Simulating a complex and massive network is a complicated task that requires more power, memory, space and time.

Netsim Standard v 11 (Network Simulator and Emulator) has been used in this work, which virtually allows users to build and study complex network behaviour. The packet flow visualization can be done through the packet animator and the results are analysed by output performance metric, the packet tracer and the event tracer (Lord and Memmi, 2008; tetcos, 2021).

As indicated earlier, this work's primary objective is to analyse the effect of different packet sizes – 256, 512, 1,024 and 2,048 bytes on a large-scale hybrid network using RIP and OSPF. The network scenario simulated is consisting of 125 objects builds on Netsim's GUI, shown in Figure 2. The proposed network consists of many routers, access points, wired and wireless nodes. In this network, the wired node topologies are connected through a router and switch with other node topologies. Similarly, wireless node topologies are connected through access points which are further connected to wired topologies through the router. These all are connected through bi-directional links. In the network below, the analysis is done where different packet size data has been sent from wired node 9 to wireless node 47. The four different CBR applications are defined for different packet sizes as App1_CBR, App2_CBR, App3_CBR and App4_CBR. All four applications are sent from node 9 to the destination node 47. The application flow can be seen in the figure from node 9 to node 47 and is represented by App3_CBR (Defining third CBR application for packet size 1,024). The various network specifications are required for optimal network survivability and efficient performance. The defined specifications and their detail or values for the proposed network are mentioned in Table 2.

The network is simulated for 100 s and results are analysed and calculated through performance metrics, packet tracer and event tracer records. The performance parameters such as application throughput and application delay are evaluated and calculated. These parameters are discussed below.

5.1 Application throughput

Application throughput is the total payload delivered to the destination. For each packet size, the total payload delivered is calculated using packet tracer and then application throughput has been calculated using equation (1).

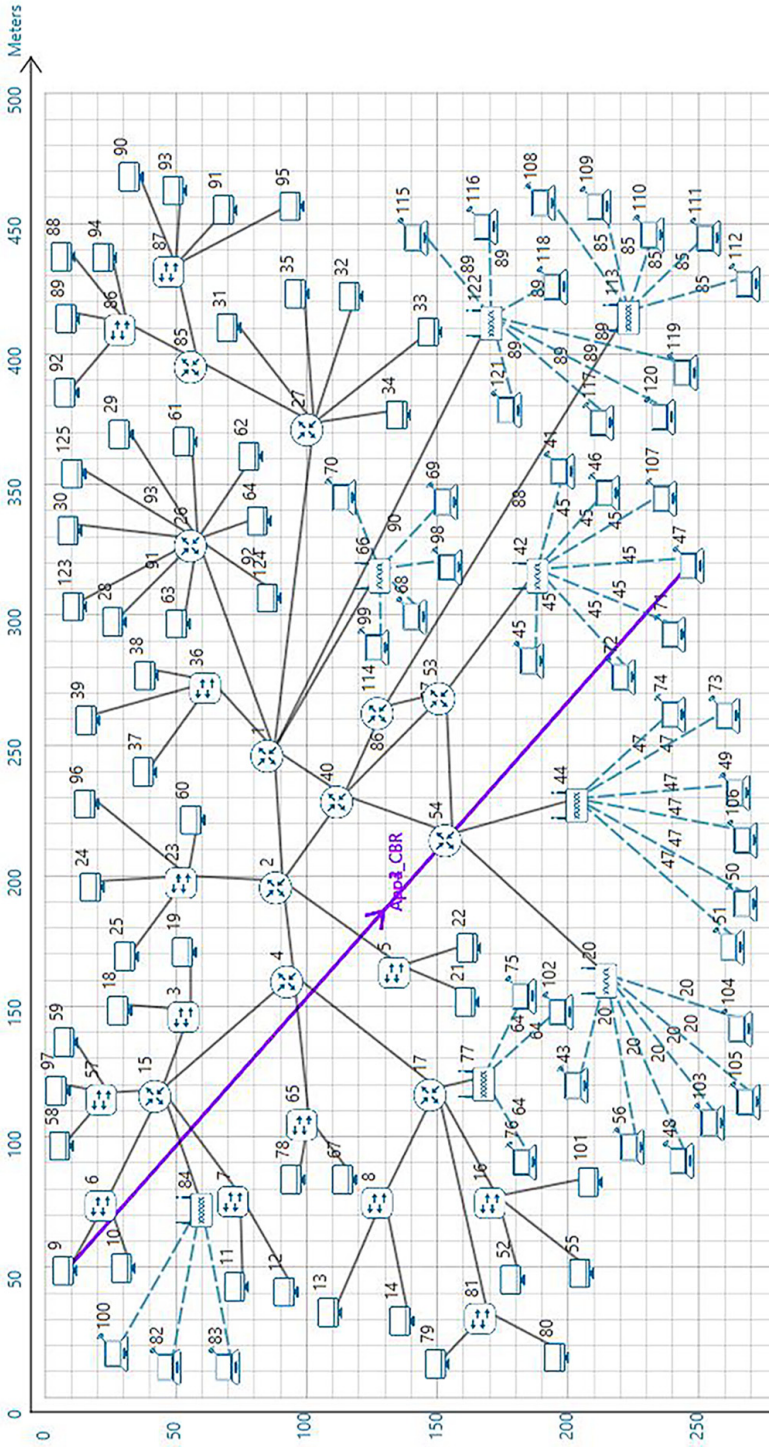


Figure 2.
Simulation network scenario

Specifications	Detail/value	Specifications	Detail/value
Simulator	Netsim	Output cost	100
Network	Both wired and wireless	Router dead interval	40 s
Routing protocol	RIP and OSPF for IPv4	Point to point interface router protocol	WAN
Application type	CBR	Wireless interface protocol access point protocol	IEEE 802.11
Number of objects	125	Transmission type	DSSS
Packets size in bytes	256, 512, 1,024 and 2,048	Transmission power	100 mW
Simulation time	100 s	Antenna height	1 metre
Packet distribution	Constant	CCA mode	Energy detection
Start timer	20 ms	Slot time	20 ms
End timer3	100 ms	Standard channel	1_2412 MHz
QoS	BE/RTPS	CW_min	31
Inter arrival time	20,000 ms	CW_max	1,023
Congestion control algorithm at the router	NEW_RENO	Retry limit	4
MSS bytes	1,460	RTS threshold	3,000
SIFS	10 ms	Buffer size	1 MB
Hello interval	10 s		

Table 2. Network specifications for simulation

$$Application\ \ Throughput = \frac{Total\ \ Payload\ \ deliverd\ \ to\ \ destnation\ \ \times 8}{Simulation\ \ Time} \tag{1}$$

In equation (1), the total payload delivered to the destination is the amount of traffic reached to the destination in bytes. Simulation time is the virtual time that begins from 0 to positive real numbers defined for network execution in microseconds (NetSim Support Portal, 2021). The simulation time should be taken longer as the routing table takes time for convergence. Here, the packet generation times are directly proportional to simulation time. The results are calculated for the application throughput for 256, 512, 1,024 and 2,048 bytes packets sizes using RIP or OSPF, are shown in Table 3. The results show that the RIP and OSPF protocol network have similar performance for application throughput. However, there are variations in payload delivered for the same packet size. The linear relation is obtained between packet size and throughput for both RIP and OSPF, which is shown in Figure 3, indicating, as the packet sizes doubles, throughput doubles.

The above results are verified from the application throughput graph obtained from the simulator. They are shown in Figure 4, indicating that the throughput for each packet size increases as the simulation time increases. Figure 4(a) shows that for packet

			RIP		OSPF	
Application	Packet size (Bytes), x	Simulation time (μsec)	Payload delivered (Bytes)	Application throughput (Mbps), y	Payload delivered (Bytes)	Application throughput (Mbps), y
App1_CBR	256	10^8	1,023,744	0.08	1,019,904	0.08
App2_CBR	512	10^8	2,046,464	0.16	2,041,344	0.16
App3_CBR	1,024	10^8	4,084,736	0.32	4,074,496	0.32
App4_CBR	2,048	10^8	8,167,936	0.65	8,158,336	0.65

Table 3. Application throughput for RIP and OSPF

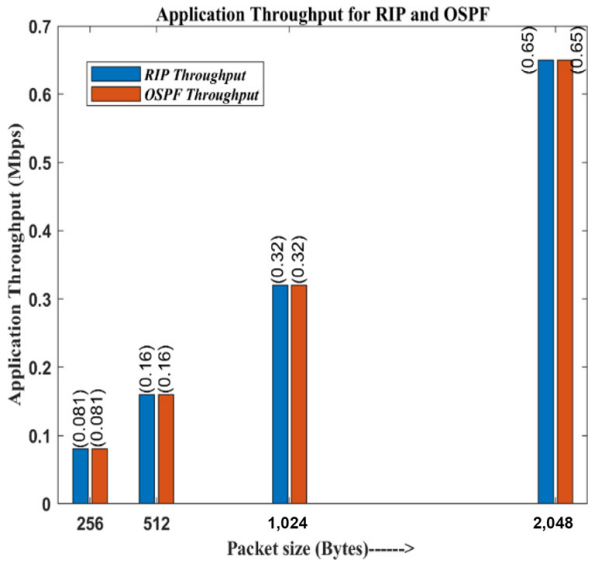


Figure 3.
Application throughput for RIP and OSPF

size 256 bytes, the maximum application throughput is 0.08 Mbps and Figure 4(b) shows that for packet size 512 bytes, the maximum application throughput is 0.16 Mbps. Figure 4(c) shows that for packet size 1,024 bytes, the maximum application throughput is 0.32 Mbps and Figure 4(d) shows that for packet size 2,048 bytes, the maximum application throughput is 0.65 Mbps. It can also be analysed from Figure 4 that the network has application throughput from 20 ms to 100 ms (start time = 20 ms and end time = 100 ms). This is because routing protocol takes time to converge. Thus, it is always advisable to choose a higher start time, as the higher the time, the better the network convergence for more extensive networks.

The perfect correlation of 0.999 is obtained between packet size and throughput for both RIP and OSPF network using equation (2) (Mukaka, 2012):

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left[n\left(\sum x^2\right) - (\sum x)2\right]\ \left[n(\sum y2) - (\sum y)2\right]}} \tag{2}$$

$$r = \frac{2395.904}{2396.003}$$

$$r = 0.999$$

where n = number of scenarios, x = packet sizes, y = application throughput and r = correlation coefficient (Table 2). RIP and OSPF have the same application throughput values; therefore, the coefficient correlation has a similar response.

5.2 Application delay

Application delay is the average amount of time taken for all the packets to reach the destination application layer from the source application layer. For each packet size, the delay is calculated using packet tracer and then application delay has been calculated using equation (3).

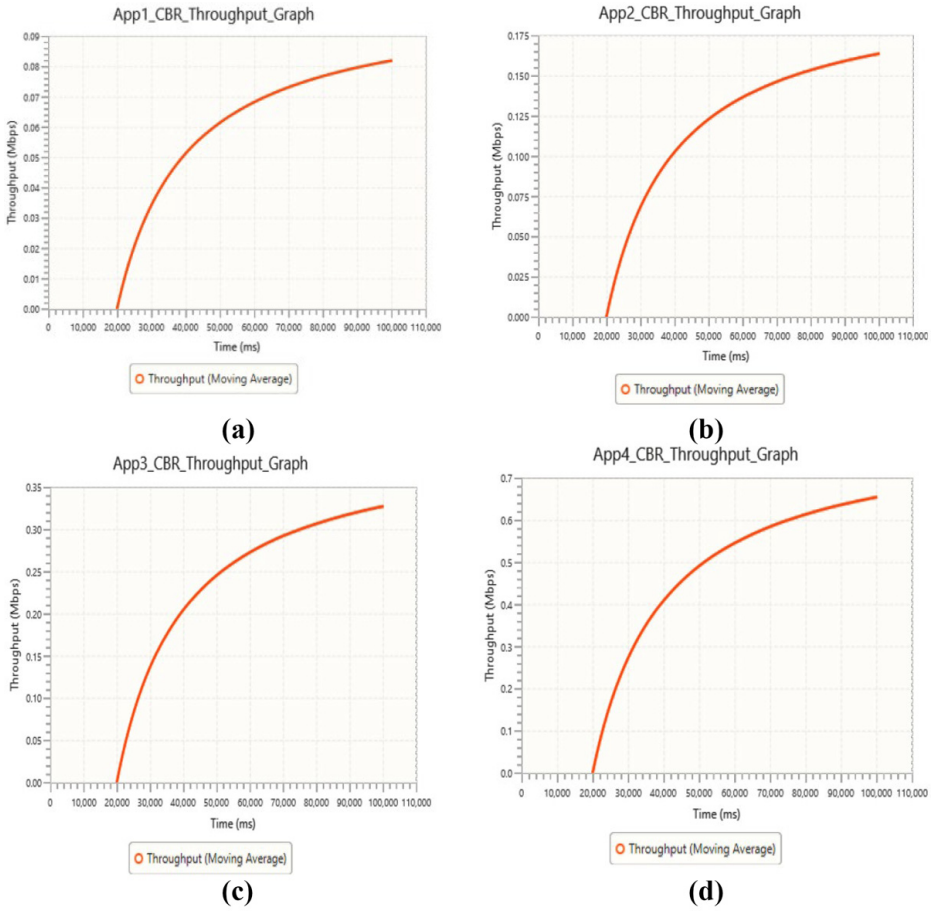


Figure 4.
Application throughput for different packet sizes (a) 256 bytes, (b) 512 bytes, (c) 1,024 bytes, (d) 2,048 bytes

Notes: (a) 256 bytes, (b) 512 bytes, (c) 1,024 bytes, (d) 2,048 bytes

$$Application\ Delay = \frac{Sum\ of\ Delay\ of\ all\ packets}{Number\ of\ packets} \tag{3}$$

In equation (3), the sum of the delay of all packets is obtained from the difference between the packet's end time and the packet's arrival time from the packet tracer. The end time and arrival time values are also calculated from the packet tracer and the calculation is done. The results show that the RIP and OSPF have different delays for each packet size; however, the delay in RIP is higher than OSPF due to the hop limit. Table 4 shows the comparison of application delay for 256, 512, 1,024 and 2,048 packet sizes on a network using either RIP or OSPF. It is found that as the packet sizes increase, the delay increases and shown in Figure 5.

Table 4.
Application delay for RIP and OSPF

Application	Packet size (Bytes)	RIP delay (μs)	OSPF delay (μs)
App1_CBR	256	976.4	975.5
App2_CBR	512	2,261.8	2,259.0
App3_CBR	1,024	3,917.0	3,909.5
App4_CBR	2,048	6,550.6	6,545.0

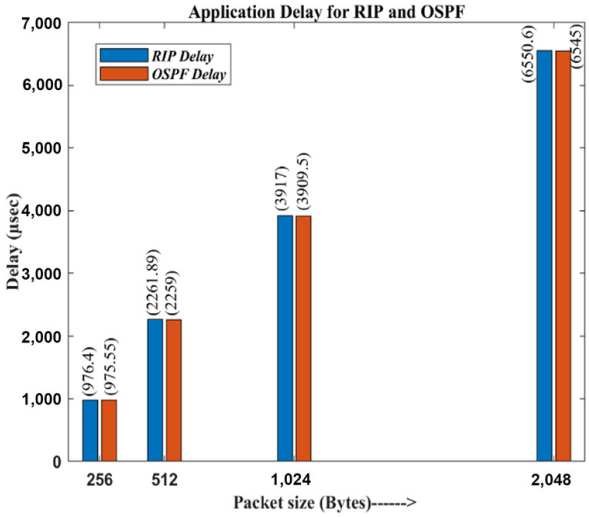


Figure 5.
Delay for RIP and OSPF

5.3 Network links

This parameter gives the values based on the overall network and links associated with it.

The values are obtained for RIP and OSPF networks with data encryption standard (DES) and without DES. The table has shown the various network link-related parameters. Table 5 shows the various parameters related to the 96 links in a network such as the data and control packets generated and transmitted from the source node, the data and control packets errored and collided during transmission. "E" in the table shows DES encryption. The table provides the complete network information about the packets. The different control packets used in the network are ARP request and reply, RTS, CTS, RIP packets, OSPF_Hello, TCP_ACK and TCP_SYN. The bytes transmitted, payload transmitted and overhead transmitted is also calculated for all links, which are detailed below.

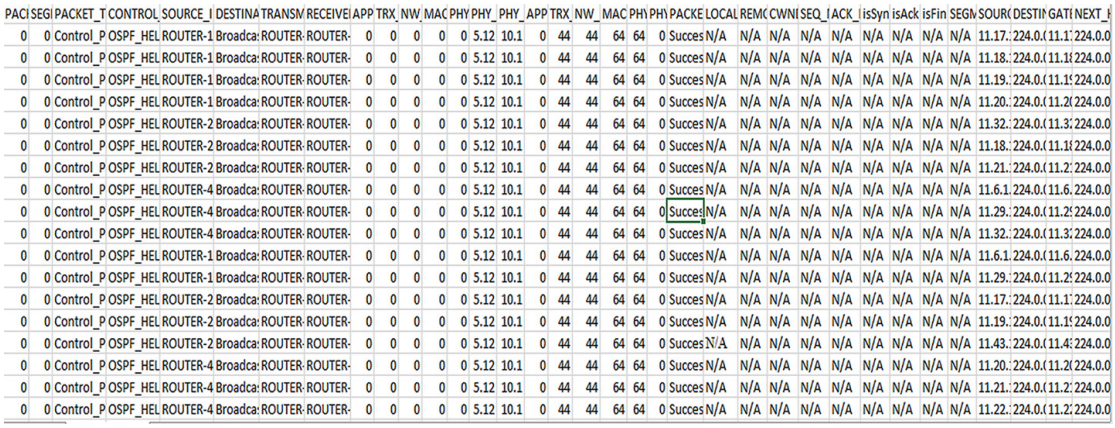
Bytes transmitted – It is defined as the total number of bytes sent in the link and calculated using equation (4). The packet tracer is used to estimate and verified the value through Figure 6.

$$Bytes\ \ Transmitted = Payload\ \ Transmitted + Overhead\ \ Transmitted \tag{4}$$

Payload transmitted – The total number of payloads sent in the link is calculated by equation (5), packet tracer is used to estimate and verify the value through Figure 6.

Routing protocol	Total links	Packet data (Bytes)	Packet control (Bytes)	Packet data errored (Bytes)	Packet control errored (Bytes)	Packet data collided (Bytes)	Packet control collided (Bytes)	Bytes transmitted (Bytes)	Payload transmitted (Bytes)	Overhead transmitted (Bytes)
OSPF(E)	96	159,423	20,341	113	5	0	0	129,557,962	122,307,264	7,250,698
RIP(E)	96	159,872	20,882	120	1	0	0	130,342,848	123,075,200	7,267,648
OSPF	96	159,555	20,346	109	6	0	0	129,674,972	122,427,776	7,247,196
RIP	96	159,836	20,882	116	1	0	0	130,304,540	123,052,224	7,252,316

Table 5.
Different network link parameters



PACI	SEGI	PACKET_T	CONTROL	SOURCE_I	DESTINA	TRANSM	RECEIVEI	APP	TRX	NW	MAC	PHY	PHY_	PHY_	APP	TRX	NW_	MAC	PHY	PH\	PACKE	LOCAL	REM(	CWNI	SEQ_I	ACK_i	isSyn	isAck	isFin	SEGN	SOUR(	DESTI	GATI	NEXT_I
0	0	Control_P	OSPF_HEL	ROUTER-1	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.17.	224.0.(	11.1	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-1	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.18.	224.0.(	11.1	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-1	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.19.	224.0.(	11.1	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-1	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.20.	224.0.(	11.2	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-2	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.32.	224.0.(	11.3	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-2	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.18.	224.0.(	11.1	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-2	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.21.	224.0.(	11.2	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-4	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.6.1.	224.0.(	11.6.	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-4	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.29.	224.0.(	11.2	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-4	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.32.	224.0.(	11.3	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-1	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.6.1.	224.0.(	11.6.	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-1	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.29.	224.0.(	11.2	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-2	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.17.	224.0.(	11.1	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-2	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.19.	224.0.(	11.1	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-2	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.43.	224.0.(	11.4	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-4	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.20.	224.0.(	11.2	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-4	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.21.	224.0.(	11.2	224.0.0
0	0	Control_P	OSPF_HEL	ROUTER-4	Broadca	ROUTER	ROUTER-	0	0	0	0	0	5.12	10.1	0	44	44	64	64	0	Succes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	11.22.	224.0.(	11.2	224.0.0

Figure 6.
Packet tracer readings

$$Payload\ \ Transmitted = Packet\ \ Tranmsitted\ \ \times Packet\ \ Size \tag{5}$$

Overhead transmitted – It is defined as the total number of overheads sent in the link and calculated from the link's layer-wise overheads and control packets. A packet tracer is used to estimate and verified the value.

Table 6 shows the comparison between the proposed method with other existing methods in terms of network application, the number of objects used in the network, different packets size considered, the different routing protocols used for analysis, the different performance parameters evaluated and the results obtained. The comparison highlights that our proposed network has a more number of objects than other research papers. Our research also presents the analysis for different data and control packets for the network links. However, it is noted that other research paper has not shown such detailed study about the links.

6. Conclusion and future work

The hybrid network consisting of 125 objects was simulated and compared on Netsim using RIP and OSPF routing protocol. The four scenarios – App1_CBR, App2_CBR, App3_CBR and App4_CBR were discussed in which each had a different packet size of 256, 512, 1,024 and 2,048 bytes and all using application CBR. The performance was evaluated for application throughput and application delay that defines QoS of the network. In RIP and OSPF networks, a similar application throughput performance was obtained for all four scenarios. For both networks, as the size of the packet increased, the application throughput increased. However, the throughout got double as packet sizes got double. It was also found that, as packet size increased, the application delay also increased. The application delays obtained were different for both RIP and OSPF networks. The network link performance parameters metric for 96 links were also calculated for the RIP and OSPF network with and without DES encryption.

It is concluded that either RIP or OSPF routing protocol can be used for application throughput performance as both give similar performance in terms of application throughput. However, the application delay in the RIP network is higher than in the OSPF network.

In future, a comparison of different network sizes can be made by increasing the number of objects in the proposed network. Furthermore, other routing protocols (like IS-IS, EIGRP, BGP) can be used for performance evaluation on the proposed network or

Methods	Application Used	Objects number	Packet size (Bytes)	Network	Routing protocol	Parameters	Results	Shortcomings
(Chauhan and Sharma, 2015)	CBR	100	256, 512, 1,024 and 2,048	Wireless network	RIP, RIPng, OSPF and OSPFv3	Throughput, jitter, PDR and end-to-end delay	–RIPng gives the best and OSPF gives the worst performance among all protocols for all performance parameters	Considers only wireless network; also, OSPF has not performed well in the network
(Chandel and Sharma, 2016)	CBR	50	512	Wired, wireless and hybrid network	RIP, OSPF, RIPng, OSPFv3 and AODV	Throughput, jitter, PDR and end-to-end delay	–AODV performance is best among all wired, wireless and hybrid networks for all performance parameters	Only 50 nodes network has been used with only one packet size
Proposed analysis	CBR	125	256, 512, 1,024 and 2,048	Hybrid network consisting of both wired and wireless nodes	RIP and OSPF	Application throughput, application delay and network links performance	–RIP and OPSF networks have the same application throughput for all packet sizes. –OSPF network has less delay than the RIP network for different packet sizes	–

Table 6.
Comparison between the proposed method with other existing methods

any other networks. The high-scale MANET networks can also be considered in the future.

References

Akyildiz, I.F., Wang, X. and Wang, W. (2005), "Wireless mesh networks: a survey", *Computer Networks*, Vol. 47 No. 4, pp. 445-487. March.

Alabady, S.A., Hazim, S. and Amer Mohammed Salih, A. (2018), "Performance evaluation and comparison of dynamic routing protocols for suitability and reliability", *International Journal of Grid and Distributed Computing*, Vol. 11 No. 7, pp. 41-52.

Alotaibi, E. and Mukherjee, B. (2012), "A survey on routing algorithms for wireless Ad-Hoc and mesh networks", *Computer Networks*, Vol. 56 No. 2, pp. 940-965. Elsevier B.V., 02-February.

Azman, A.S.A., Lee, M.Y., Subramaniam, S.K. and Feroz, F.S. (2020), "Performance evaluation of grid wireless sensor network with different packet size for pipeline in downstream of oil and gas industry", *MS&E*, Vol. 765 No. 1, p. 12069.

Belda, K. and Rychnovsk`y, V. (2010), "Wireless control communication for mechatronic systems", *Proceedings of Workshop on Systems and Control*, University of Pannonia, Veszprém, Vol. 68, p. 73.

Chai, Y. and Zeng, X.J. (2019), "Regional condition-aware hybrid routing protocol for hybrid wireless mesh network", *Computer Networks*, Vol. 148, pp. 120-128. January.

Chandel, S.T. and Sharma, S. (2016), "Performance evaluation of IPv4 and IPv6 routing protocols on wired, wireless and hybrid networks", *International Journal of Computer Networks and Applications*, Vol. 3 No. 3, pp. 57-62.

Chauhan, D. and Sharma, S. (2015), "Performance evaluation of different routing protocols in IPv4 and IPv6 networks on the basis of packet sizes", *Procedia Computer Science*, Vol. 46, pp. 1072-1078.

D'Arienzo, M. and Romano, S.P. (2016), "GOSPF: an energy efficient implementation of the OSPF routing protocol", *Journal of Network and Computer Applications*, Vol. 75, pp. 110-127. November.

Dey, G.K., Ahmed, M.M. and Ahmmed, K.T. (2016), "Performance analysis and redistribution among RIPv2, EIGRP and OSPF routing protocol", *1st International Conference on Computer and Information Engineering, ICCIE 2015*, pp. 21-24.

Geng, H., Zhang, H., Shi, X., Wang, Z., Yin, X., Zhang, J., Hu, Z. and Wu, Y. (2020), "A hybrid link protection scheme for ensuring network service availability in link-state routing networks", *Journal of Communications and Networks*, IEEE, Vol. 22 No. 1, pp. 46-60.,

Guo, L., Wang, X., Peng, Y., Wu, J., Zhang, X., Hou, W. and Zhao, J. (2010), "Design on routing protocol in hybrid wireless self-organizing networks", *2010 2nd IEEE International Conference on Network Infrastructure and Digital Content*, IEEE, pp. 569-573.

Haglan, H.M., Mostafa, S.A., Safar, N.Z.M., Mustapha, A., Saringatb, M.Z., Alhakami, H. and Alhakami, W. (2021), "Analyzing the impact of the number of nodes on the performance of the routing protocols in manet environment", *Bulletin of Electrical Engineering and Informatics*, Vol. 10 No. 1, pp. 434-440. doi: 10.11591/eei.v10i1.2516.

Hameed, S.S. and Saleh, H.M. (2021), "The impact of routing protocols and topologies on IOT based systems", *Journal of Physics: Conference Series*, Vol. 1804 No. 1, p. 12112.

Ismail, Z. and Hassan, R. (2011), "Effects of packet size on AODV routing protocol implementation in homogeneous and heterogeneous MANET", *2011 Third International Conference on Computational Intelligence, Modelling and Simulation*, pp. 351-356.

Jain, N. and Payal, A. (2019), "Comparison between IPv4 and IPv6 using OSPF and OSPFv3 on riverbed modeler", *2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, pp. 1-7.

Jain, N. and Payal, A. (2020), "Performance evaluation of IPv6 network for real-time applications using is-ISv6 routing protocol on riverbed modeler", *Procedia Computer Science*, Vol. 173.

Khan, A.A., Zafrullah, M., Hussain, M. and Ahmad, A. (2017), "Performance analysis of OSPF and hybrid networks", *2017 International Symposium on Wireless Systems and Networks (ISWSN)*, pp. 1-4.

Kumari, P. and Sahana, S.K. (2019), "Comprehensive survey and comparative experimental performance gain of AODV, DSR and OLSR in MANETs", *International Journal of Engineering and Advanced Technology*, Vol. 8 No. 5, pp. 1036-1045.

Lord, M. and Memmi, D. (2008), "NetSim: a simulation and visualization software for information network modeling", *2008 International MCETECH Conference on e-Technologies (mcetech 2008)*, pp. 167-177.

Manzoor, A., Hussain, M. and Mehrban, S. (2020), "Performance analysis and route optimization: redistribution between EIGRP, OSPF and BGP routing protocols", *Computer Standards and Interfaces*, Vol. 68, p. 103391. February.

Mohapatra, S. and Kanungo, P. (2012), "Performance analysis of AODV, DSR, OLSR and DSDV routing protocols using NS2 simulator", *Procedia Engineering*, Vol. 30, pp. 69-76.

Mukaka, M.M. (2012), "A guide to appropriate use of correlation coefficient in medical research", *Malawi Medical Journal*, Vol. 24 No. 3, pp. 69-71.

Nedyalkov, I. (2019), "Studying of a modeled IP – based network using different dynamic routing protocols", *2019 X National Conference with International Participation (ELECTRONICA)*, pp. 1-4.

NetSim Support Portal (2021), "What is simulation time? NetSim support portal", available at: https://support.tetcos.com/support/solutions/articles/14000057035-what-is-simulation-time-

Omar, M., Hedjaz, S., Rebouh, S., Aouchar, K., Abbache, B. and Tari, A. (2015), "On-demand source routing with reduced packets protocol in mobile ad-hoc networks", *AEU – International Journal of Electronics and Communications*, Vol. 69 No. 10, pp. 1429-1436. October.

Priyambodo, T.K., Wijayanto, D. and Gitakarma, M.S. (2021), "Performance optimization of MANET networks through routing protocol analysis", *Computers*, Vol. 10 No. 1, p. 2.

Rukmani, P. and Ganesan, R. (2013), "Scheduling algorithm for real time applications in mobile ad hoc network with opnet modeler", *Procedia Engineering*, Vol. 64, pp. 94-103.

Samaan, S.S. (2018), "Performance evaluation of RIPng, EIGRPv6 and OSPFv3 for real time applications", *Journal of Engineering*, Vol. 24 No. 1, pp. 111-122.

Sarao, P. (2020), "A packet size based intelligent model for AODV in MANET", *International Journal of Innovative Technology and Exploring Engineering*, Vol. 9 No. 3, pp. 2070-2079.

Tetcos (2021), "Netsim", available at: www.tetcos.com/

Tolani, M. and Mishra, R. (2012), "Effect of packet size on various MANET routing protocols", *International Journal of Applied Information Systems*, Vol. 4 No. 9, ISSN-2249-8068.

Tuteja, A., Gujral, R. and Thalia, S. (2010), "Comparative performance analysis of DSDV, AODV and DSR routing protocols in MANET using NS2", 2010 International Conference on Advances in Computer Engineering, pp. 330-333.

Wang, Q., Teng, L. and Zhao, S. (2018b), "A contextual awareness-learning approach to multi-Objective mobility management in 5G ultra-dense network", *Mechatronic Systems and Control*, Vol. 46 No. 2, pp. 82-91.

Wang, Q., Gao, L., Yang, Y., Zhao, J., Dou, T. and Fang, H. (2018a), "A load-balanced algorithm for multi-controller placement in software-defined network", *Mechatronic Systems and Control*, Vol. 46 No. 2, pp. 72-81.

Waqas, M., Malik, S.U.R., Akbar, S., Anjum, A. and Ahmad, N. (2019), "Convergence time analysis of OSPF routing protocol using social network metrics", *Future Generation Computer Systems*, Vol. 94, pp. 62-71. May.

Wijaya, C. (2011), "Performance analysis of dynamic routing protocol EIGRP and OSPF in IPv4 and IPv6 network", *Informatics and Computational Intelligence (ICI), 2011 First International Conference on*, pp. 355-360.

Zhao, S. and Raychaudhuri, D. (2009), "Scalability and performance evaluation of hierarchical hybrid wireless networks", *IEEE/ACM Transactions on Networking*, Vol. 17 No. 5, pp. 1536-1549.

Zoppi, S., Van Bemten, A., Gürsu, H.M., Vilgelm, M., Guck, J. and Kellerer, W. (2018), "Achieving hybrid wired/wireless industrial networks with WDetServ: reliability-based scheduling for delay guarantees", *IEEE Transactions on Industrial Informatics*, Vol. 14 No. 5, pp. 2307-2319.

About the authors



Neha Jain is currently a Research Scholar with the University School of Information, Communication and Technology, Guru Gobind Singh Indraprastha University, New Delhi 110078, India. She received the MTech degree in electronics and communication from Indra Gandhi Institute of Technology, Delhi and the BTech degree in electronics and communication from the Institute of Engineering and Technology, Rajasthan University. She is a member of IETE and has more than 10 years of teaching experience. Her research interests include computer networks, wireless sensor networks and the IoT. Neha Jain is the corresponding author and can be contacted at: nehajain312@gmail.com



Ashish Payal is currently an Assistant Professor with the University School of Information, Communication and Technology, Guru Gobind Singh Indraprastha University, New Delhi 110078, India. He received a PhD degree in the area of wireless sensor networks from Guru Gobind Singh Indraprastha University, New Delhi, India, in 2016. He received the MTech degree in computer science and engineering from Punjabi University, Patiala, Punjab, India and the MSc degree in applied physics from Guru Nanak Dev University, Amritsar, Punjab, India. He is a member of IEEE and ACM. His teaching and research interests include mobile *ad hoc* networks, wireless sensor networks, computer networks and the IoT.



Aarti Jain received the BTech degree in Electronics and Communication Engineering from Beant Government College of Engineering, Punjab in 2002, the ME degree in Electronics and Communication Engineering from Delhi College of Engineering, New Delhi, India, in 2009. She received her Doctorate in 2016 from Guru Gobind Singh Indraprastha University in the field of wireless sensor networks. She has more than 17 years of teaching experience. Her research interests include IoT, free-space optics, 5G communications, quality management in wireless sensor networks, fuzzy logistics and bio-inspired computing and its application. She is a senior member of IEEE and an associate member of IET. She served as chairperson and session chair in various international conferences. She has published several research papers in SCI-indexed journals and international conferences.

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**

Bull. Mater. Sci. (2021)44:237
https://doi.org/10.1007/s12034-021-02527-x

© Indian Academy of Sciences

Effect of metal doping on visible light absorption and optical properties of lithium niobate ($LiNbO_3$) crystal: a first-principles calculation

ASHISH RATURI[1], POORNIMA MITTAL[1,*] and SUDHANSHU CHOUDHARY[2]
[1]Department of Electronics and Communication Engineering, Delhi Technological University, Delhi 110042, India
[2]Department of Electronics and Communication Engineering, National Institute of Technology, Kurukshetra 136119, India
*Author for correspondence (poornimamittal@dtu.ac.in)

MS received 9 March 2021; accepted 18 May 2021

Abstract. Electronic and optical properties of pristine and metal-doped lithium niobate crystals are investigated by using first-principles DFT calculations. The results on optical properties suggest that there is a shift in the absorption edge towards visible region (red-shift) for metal-doped structures, in comparison with the pristine lithium niobite. A series of metals are used for doping and it is found that the absorption edge is shifted significantly to the visible region for the dopants; gold (Au), silver (Ag), aluminium (Al) and copper (Cu) due to surface plasma resonance. However, for other metal dopants like iron (Fe), manganese (Mn), molybdenum (Mo) and nickel (Ni), the absorption is slightly improved in the visible region and red-shift is observed. The bandgap of all the doped structures is found to be reduced, this might be proven advantageous for photovoltaic applications, which requires high optical absorption in the visible region. The dielectric constant and refractive index of the pristine and doped structures are also calculated and it is observed that the absorption trend is in accordance with dielectric constant. The optical absorption is enhanced in the visible region due to doping of selected metals (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni) making M-lithium niobite a promising material for optoelectronic- and photonic-based applications.

Keywords. Absorption; dielectric function; density function theory; doping; lithium niobate ($liNbo_3$); optoelectronic.

1. Introduction

In 1963, geologist Gustav Rose discovered the mineral $CaTiO_3$ (calcium titanate) in the mountains of Ural. The mineral was named as perovskite to award the gratitude to Lev Perovski, a distinguished mineralogist from Russia [1]. Any compound having formula ABC_3 is named as perovskite, where B ion surrounds the C ions. Numerous varieties of electric properties and solid-state behaviours including: insulating, semiconducting, metallic and superconducting are comprehended in the perovskite materials [2]. Therefore, these compounds are very fascinating for investigations and are applied to a large variety of applications. They also exhibit a variety of optical, magnetic and electronic properties. Most of the elements of the periodic table are found stable in the perovskite structure [3]. Studies of these compounds are of great interest owing to their wide applicability in the photochromic, image storage, electrochromic, filtering, switching and surface acoustic wave signal processing devices [4,5]. Likewise, they have a virtuous impact on numerous optoelectronic-, bio-sensing-, photo-electrolysis-based applications [6]. In recent years, a huge interest has developed in the studies of perovskite materials and investigating their properties for the photonic and optoelectronic applications. It is due to their adjustable optical and electrical properties, that too with mechanical flexibility [7,8]. Furthermore, their high-absorption coefficient, low diffusion length, excellent charge transportation and low non-radioactive emission exceed their utilization for photovoltaic- and optoelectronic-based devices [9–11].

Lithium niobate ($LiNbO_3$) is a notable member of the perovskite family. It exists in the ferroelectric state and liable for potential use in the electro-optical and optical systems due to its extraordinary piezoelectric, electro-optical, photovoltaic, photo elastic and nonlinear optical properties [12–14]. The study of $LiNbO_3$ is interesting for the researchers due to its unique properties that are mainly determined by its defect structure and composition. Its optical characteristics can be influenced by the significant amount of impurity, which can replace Li/Nb ion and can occupy the octahedral void in the crystal [15]. In the past, $LiNbO_3$ has attracted the attention of the researchers and manufacturers for optoelectronics-based applications. Javid

Published online: 30 August 2021

et al [16] reported the DFT study of electronic and optical properties of $LiNbO_3$. It has been concluded in the past studies that a major shortcoming associated with the $LiNbO_3$ is its large bandgap. Due to which, the $LiNbO_3$ crystal undergoes surface photo-activation under UV radiation alone [17]. However, UV light (300–400 nm), is a very small fraction of the sunlight (only 5%) and a large proportion of the sunlight will be left unutilized. So, it is interesting to shift the optical response of $LiNbO_3$ to the visible range of the optical spectrum [18]. This can be achieved by the metal-doping which leads an extension in the wavelength range towards the visible spectra and thus, improves the solar energy utilization efficiency [19,20].

Pathak *et al* [21] studied the effect of Eu^{3+} doping on the optical properties of $LiNbO_3$ and concluded that the optical properties are improved due to doping. Likewise, Hui *et al* [22] studied the effect of Fe and Mg doping on the optical properties of $LiNbO_3$. Bachiri *et al* [23] also analysed the effect of Ni doping on the properties of $LiNbO_3$. Metal doping is an effective way to improve the optical properties by lowering the bandgap of materials for photoexcitation (red-shift) and additionally reduce the recombination of charge carriers. In the category of dopants, metals can be further subcategorized based on their properties. Although there are various classes of metals, however, a superior class of metal is plasmonic (Ag, Au, Al and Cu) which exhibits the occurrence of surface plasmonic phenomena [24]. This phenomenon leads to the oscillation of electrons for the matched frequency of the incident radiation. The metals exhibiting surface plasmonic phenomena have unique optical properties and are exceptionally strong absorber of light. This makes them an appropriate dopant for tuning the optical absorption towards the visible region [25–28].

In this paper, the electronic and optical properties of the pristine $LiNbO_3$ crystal are investigated. Further, the crystal is doped with various metals (M = Fe, Mn, Mo and Ni) along with the metals exhibiting plasmonic effect (M = Au, Ag, Al and Cu). Subsequently, the effects of doping on electronic and optical properties are analysed. The bandgap of pristine and doped structures is calculated for analysing the electronic properties. The primary motivation of the doping is to reduce the large bandgap and to further tune the optical properties. For analysing the effect of doping on the optical properties, the energy and wavelength-dependent absorption coefficient are computed. Further, dielectric function and refractive index are also computed for the analysis. From the computed results, a significant improvement is observed in the optical properties due to metal-doping. These improvements are significant for the plasmonic metal dopant (M = Au, Ag, Al and Cu) [29]. However, for the other metal dopants like Fe, Ni, Mn and Mo, an improved absorption in visible spectra along with red-shift is observed. These findings suggest M-lithium niobite (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni) as a suitable candidate for photovoltaic and optical applications.

This paper is divided into four sections including 'Introduction' section. Further, the computational details for pristine and metal-doped $LiNbO_3$ are elaborated in section 2. Thereafter, section 3 presents the detailed structural geometry and properties, electronic properties and optical properties of the pristine and metal-doped $LiNbO_3$. Finally, important outcomes are summarized in section 4.

2. Computational details

The ground-state properties of any material can be efficaciously computed by utilizing first-principles calculations modelled on density functional theory (DFT). In this paper, the plane-wave DFT calculator is utilized to optimize the geometry and to compute the properties of $LiNbO_3$ and its metal-doped structures. The calculations are performed on the Quantum-ATK tool kit. In geometry optimization, to represent the exchange-correlation function, the generalized gradient approximation (GGA) of Perdew–Purke–Prnzerh is adopted [30,31]. The geometry optimization is performed by using Monkhorst–Pack k-point mesh with the $9 \times 9 \times 1$ k point sampling and density mesh cut off of 150 Hartree [32]. LBFGS optimization method of type quasi-Newton optimization algorithm is utilized for the geometry optimization until all forces acting on the atoms and stress became <0.02 eV $Å^{-1}$ and 0.004 GPa, respectively. To render steadiness with earlier research, GGA is adopted to compute the electronic properties [16,33]. Nonetheless, due to the use of effective approximation algorithms, the calculated results are in good agreement with the previously reported results [16,34]. The DFT in combination with meta-GGA (TB09) is used to calculate the optical properties. The $9 \times 9 \times 1$ k-points sampling is used for the Brillouin zone to calculate the optical properties [32].

The electronic and optical properties of the optimized $LiNbO_3$ hexagonal unit cell with various doping arrangements are calculated. Various metal dopants from the periodic table are chosen to regulate/tune the optical properties of the crystal. For the material, relationships among the dielectric constant (ε_r), polarizability (α), optical conductivity (σ) and susceptibility (χ) are given as [35]:

$$\varepsilon_r = 1 + \chi(\omega), \tag{1}$$

$$\alpha(\omega) = V\varepsilon_0\chi(\omega), \tag{2}$$

$$\sigma(\omega) = -\mathrm{img}(\omega\varepsilon_0\chi(\omega)). \tag{2}$$

The refractive index (η), extinction coefficient (k) and dielectric constant (ε_r) are related as:

$$\eta + i\kappa = \sqrt{\varepsilon_r}. \tag{4}$$

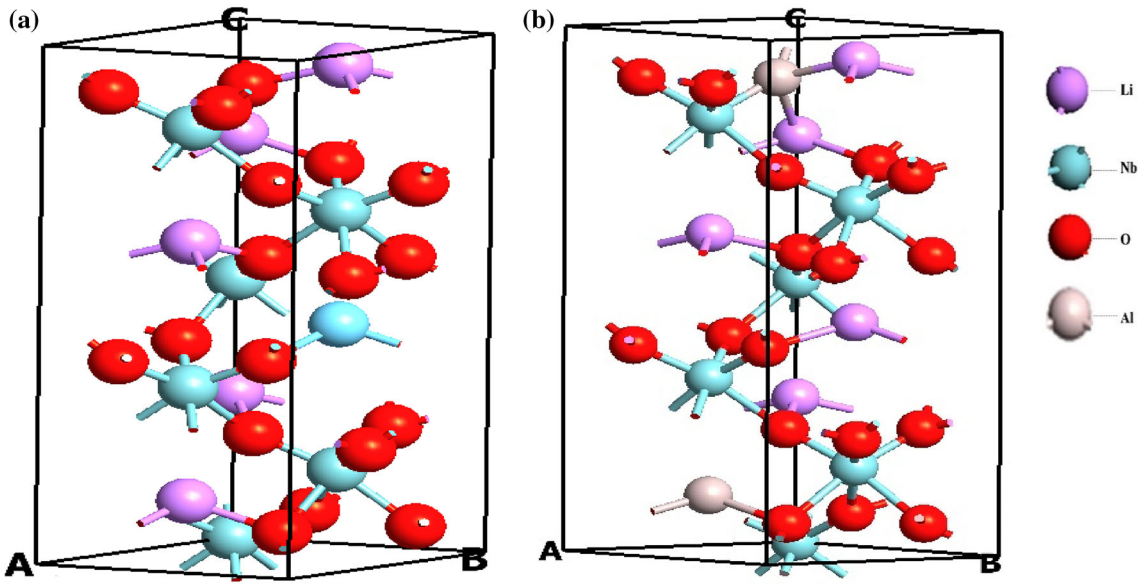


Figure 1. Hexagonal unit cell of (**a**) pristine $LiNbO_3$ and (**b**) metal (Al)-doped $LiNbO_3$.

Table 1. Lattice constants of pristine $LiNbO_3$.

a (Å)	b (Å)	c (Å)	Vol. (Å^3)	References
5.204	5.204	13.976	378.492	This work
5.220	5.220	14.093	384.011	Theor. [37]
5.057	5.057	13.942	356.542	Theor. [33]
5.147	5.147	13.849	366.880	Expt. [38]

Table 2. Lattice constant and bandgap of pristine $LiNbO_3$ and metal-doped structures.

Doped models	Lattice parameters			Volume (Å^3)	Bandgap (eV)
	a (Å)	b (Å)	c (Å)		
Pristine-LN	5.204	5.204	13.976	378.492	3.56
Ag-LN	5.551	5.551	14.711	453.298	1.78
Al-LN	5.505	5.505	15.068	456.635	2.02
Au-LN	5.505	5.505	14.346	434.755	1.36
Cu-LN	5.556	5.556	14.411	444.855	0.96
Fe-LN	5.483	5.483	14.333	430.897	1.38
Mo-LN	5.520	5.520	14.631	445.812	0.61
Mn-LN	5.517	5.517	14.390	437.992	0.72
Ni-LN	5.534	5.534	14.337	439.072	0.94

Refractive index (η) and extinction coefficient (k) can be represented in terms of real (ε_1) and imaginary parts (ε_2) of dielectric function, as:

$$\eta = \sqrt{\frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} + \varepsilon_1}{2}}, \tag{5}$$

$$\kappa = \sqrt{\frac{\sqrt{\varepsilon_1^2 + \varepsilon_2^2} - \varepsilon_1}{2}}. \tag{6}$$

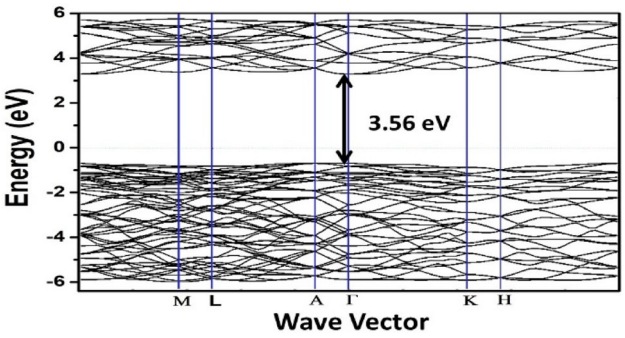


Figure 2. Band structure of pristine $LiNbO_3$ with bandgap, E_g = 3.56 eV.

Coefficient of optical absorption and coefficient of extinction are related as [36]:

$$\alpha_a = 2\frac{\omega}{c}\kappa. \tag{7}$$

Wavelength (λ) and photon energy (E) are related as:

$$\lambda(\mu\text{m}) = \frac{1.2398}{E(\text{eV})}. \tag{8}$$

The above equations are utilized to calculate the optical parameters from the computed values.

3. Results and analyses

In this section, structural, electronic and optical properties of pristine $LiNbO_3$ crystal are calculated and analysed in detail. The pristine $LiNbO_3$ is doped with variety of metal dopants (M = Al, Au, Ag, Cu, Fe, Mn, Mo and Ni). Further,

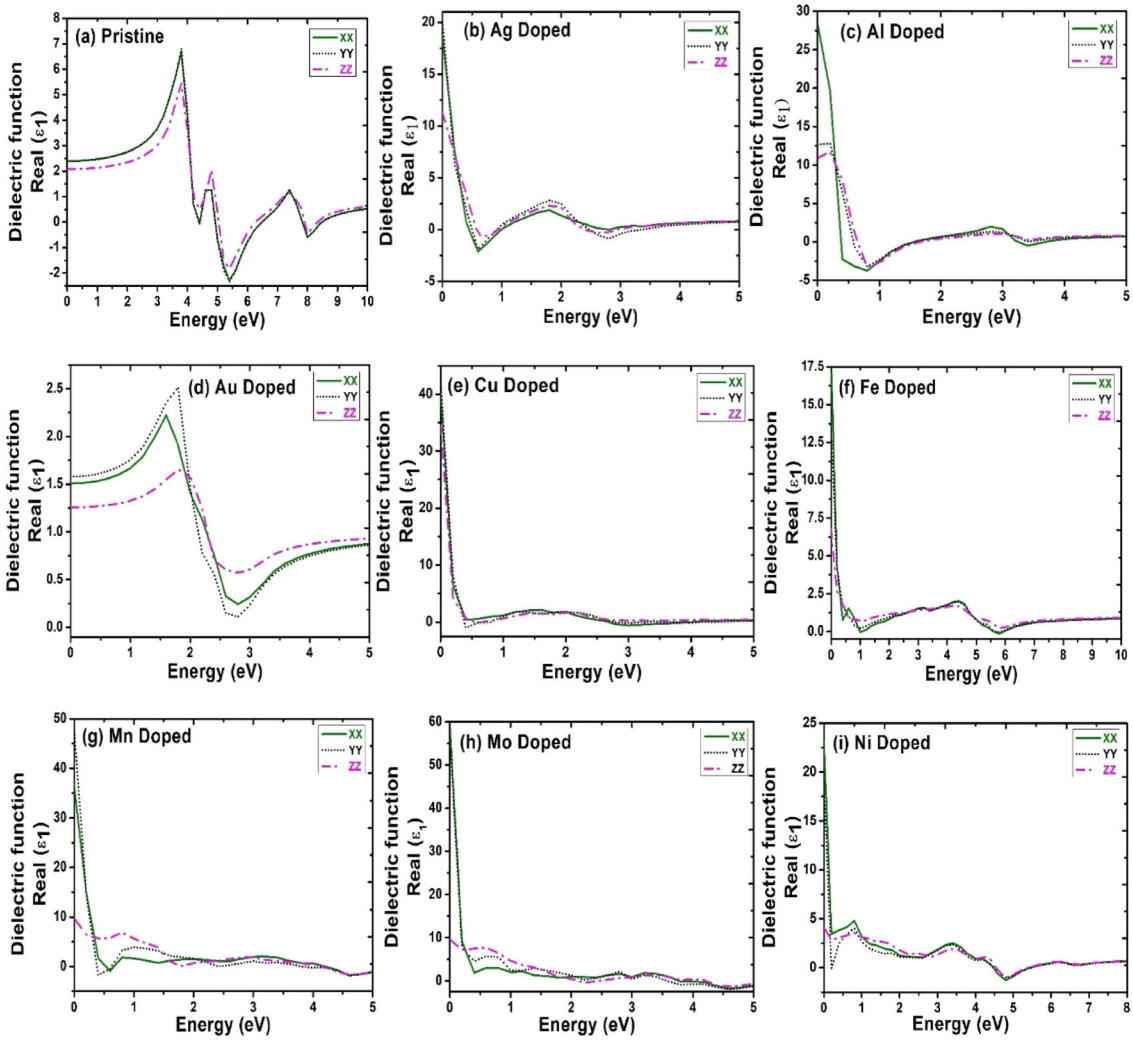


Figure 3. Energy-dependent real part of dielectric function for (**a**) pristine $LiNbO_3$, and doped with (**b**) Ag, (**c**) Al, (**d**) Au, (**e**) Cu, (**f**) Fe, (**g**) Mn, (**h**) Mo and (**i**) Ni, along the xx, yy and zz tensors.

the doped structures are simulated and their electronic and optical properties are calculated and analysed.

3.1 *Structural geometry and properties*

The hexagonal unit cell of $LiNbO_3$ crystal consists of six atoms of each lithium and niobate along with 18 atoms of oxygen (total 40 atoms). The $LiNbO_3$ crystal structure utilized for the calculations [37], is illustrated in figure 1a. In the doped structures, the hexagonal unit cell of $LiNbO_3$ containing 40 atoms is constructed, and two host atoms are replaced by dopant atoms. All the doped models are simulated with 5% doping concentration, as shown in figure 1b. The doping model in figure 1b is represented for the dopant; aluminium (Al) and same analogy is adopted for the other dopants.

For pristine $LiNbO_3$, the geometry is optimized and the lattice parameters of the optimized structure are found in good agreement with the previously reported results [16,34,37,38,39]. The compression of the experimental lattice parameter reported in earlier studies along with the values calculated in this study for pristine $LiNbO_3$ are depicted in table 1. The calculated lattice constant, lattice volume and bandgap for metal-doped structures are elaborated in table 2.

3.2 *Electronic structure and properties*

Band structure of any material is a representation of the band of eigen values, which can be utilized to understand the alignment of Fermi level. For any material, there is a noteworthy influence of the location of Fermi level on electronic and optical properties [40–44]. By identifying the location and energies of the dominant bands close to the Fermi level, the electronic and optical properties of the materials can be easily understood. The band structure of

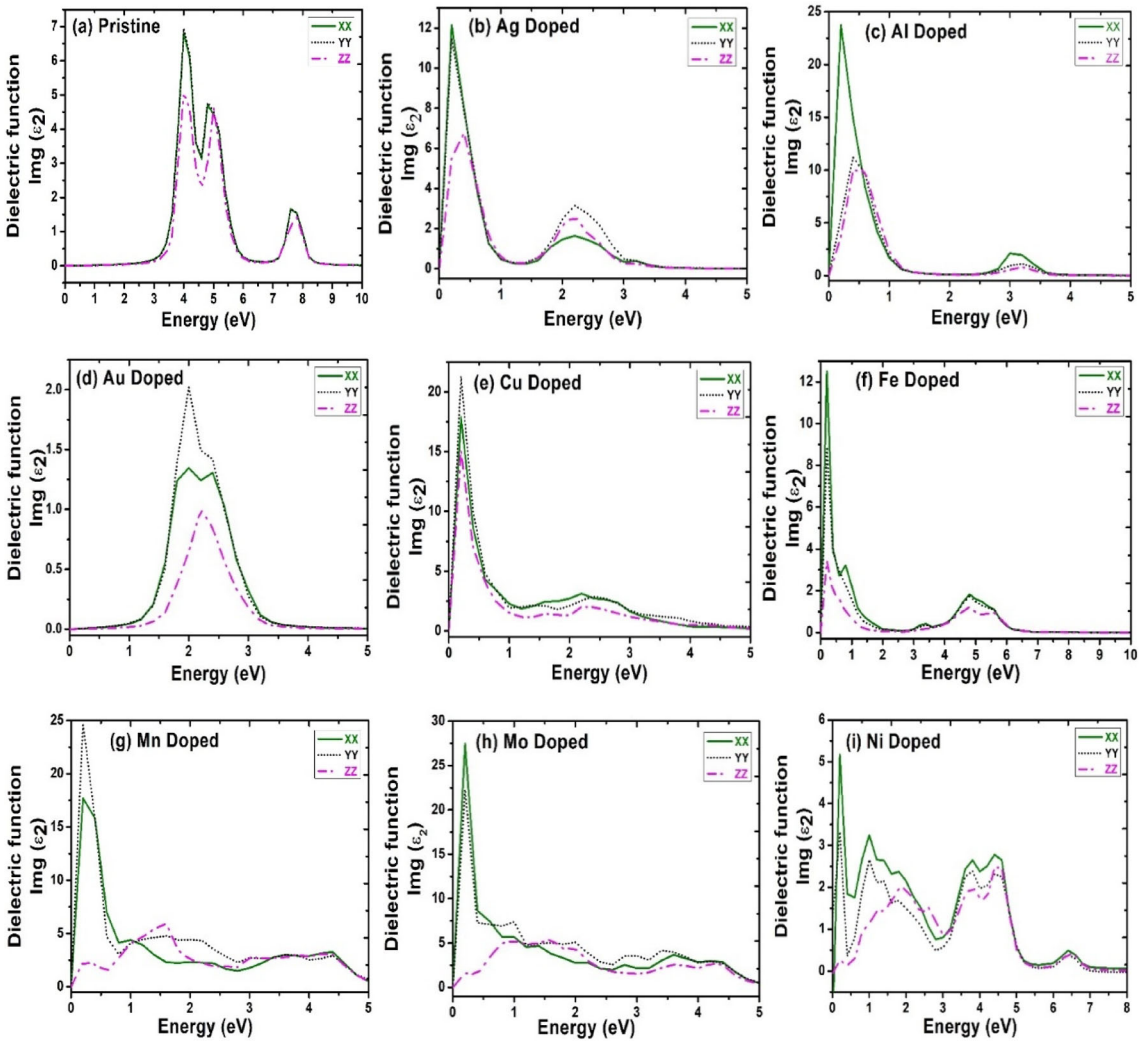


Figure 4. Energy-dependent imaginary part of dielectric function for (**a**) pristine $LiNbO_3$, and doped with (**b**) Ag, (**c**) Al, (**d**) Au, (**e**) Cu, (**f**) Fe, (**g**) Mn, (**h**) Mo and (**i**) Ni, along the xx, yy and zz tensors.

pristine $LiNbO_3$ crystal is computed using GGA and is presented in figure 2. It can be observed from the figure that the pristine $LiNbO_3$ crystal has a direct bandgap of 3.56 eV. The conduction band minima and valance band maxima are located at 2.94 and –0.62 eV. The valence band maximum and conduction band minimum occur at the same point Γ. The occurrence of valence band maximum and conduction band minimum at same point makes pristine $LiNbO_3$, a direct bandgap material [34]. The value of bandgap is in well agreement with the reported bandgap of 3.54 eV by Hossain *et al* [34], which validates the calculated results. The calculated value of the bandgap is also in good agreement with the other reported results in earlier theoretical and experimental studies [16,38].

To understand the effect of doping on the electronic and optical properties, M-$LiNbO_3$ (M = Al, Au, Ag, Cu, Fe, Mn, Mo and Ni) crystal is optimized. The calculated lattice constant, lattice volume and bandgap of pristine and metal-doped (M = Al, Au, Ag, Cu, Fe, Mn, Mo and Ni) structures after geometry relaxation are elaborated in table 2. A significant reduction in the bandgap is achieved for all the metal-doped structures (summarized in table 2) as compared to pristine $LiNbO_3$ due to the introduction of in-gap impurity states. It is determined from the calculated bandgap that the metal doping can be advantageous to achieve the improvement in the optoelectronic performance of $LiNbO_3$ due to the bandgap narrowing.

3.3 *Optical properties of pristine and doped structures*

In this section, the calculated optical properties including dielectric function, absorption coefficient and refractive index of pristine and M-$LiNbO_3$ (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni) crystals are analysed in detail. These analyses are very interesting and are essential for finding a notable doped structure of $LiNbO_3$ for optoelectronics-based applications. The values represented in this section

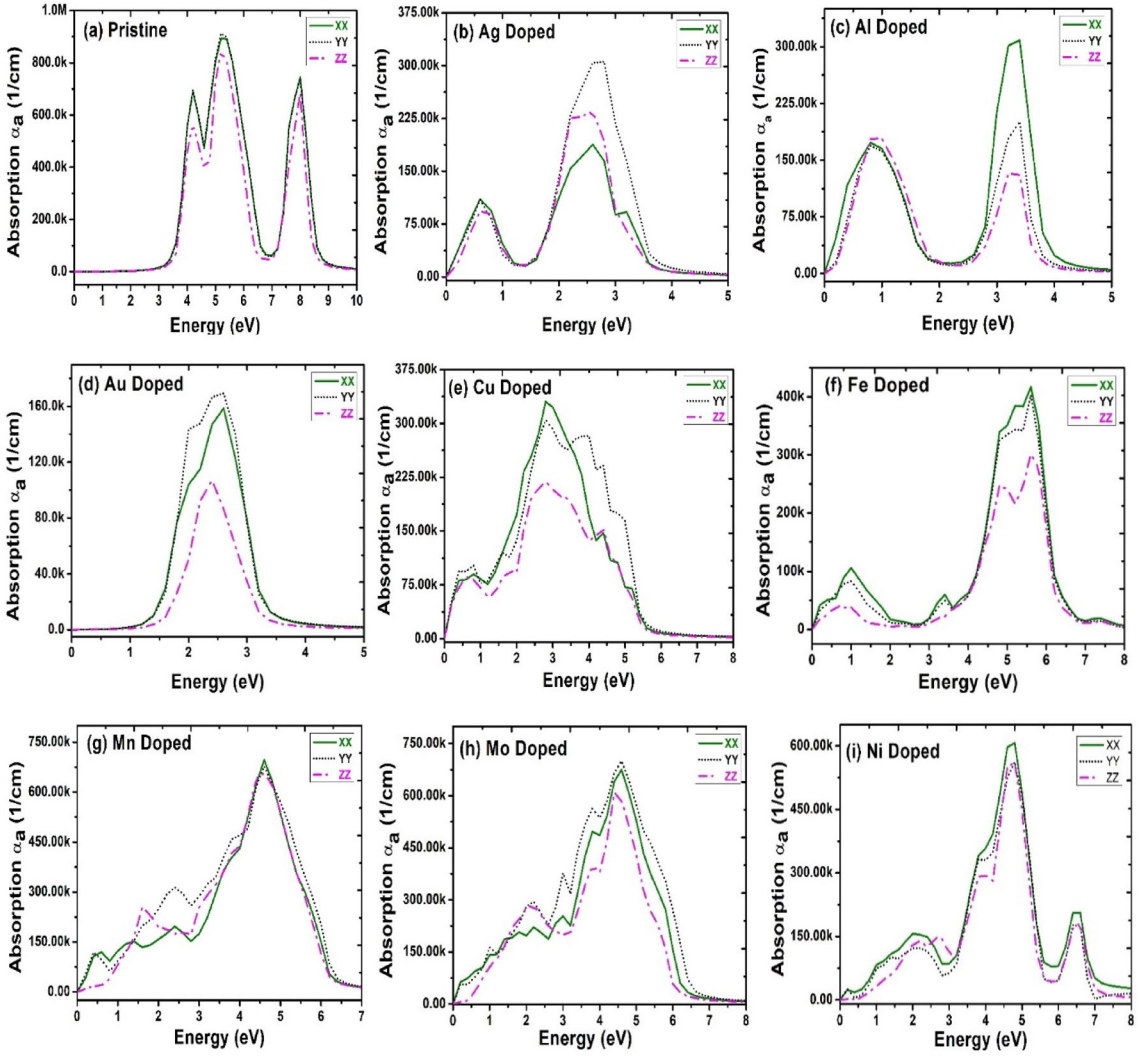


Figure 5. Energy-dependent absorption coefficient for (**a**) pristine $LiNbO_3$, and doped with (**b**) Ag, (**c**) Al, (**d**) Au, (**e**) Cu, (**f**) Fe, (**g**) Mn, (**h**) Mo and (**i**) Ni, along the xx, yy and zz tensors.

are calculated along the xx, yy and zz tensors. The results along the *x*-direction are discussed here, and the results in the other two directions may be interpreted in the same way [41]. The real part of dielectric function is illustrated in figure 3a for pristine $LiNbO_3$ and in figure 3b–h for M-$LiNbo_3$ (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni). The real part of dielectric function describes the phase lag between the driving and response frequency and represents the slowing down of the light in the material. For the doped $LiNbO_3$, the value of real part of dielectric function increases and the dielectric loss decreases with respect to pristine $LiNbO_3$. The decreased dielectric losses signify lesser slowing down of the light in the doped structures and better optical response in comparison to the pristine $LiNbO_3$.

The imaginary (ε_2) portion of the energy-dependent dielectric function for pristine and metal-doped $LiNbO_3$ are illustrated in figure 4. It is significant and fascinating to examine the imaginary (ε_2) portion of dielectric function due to its effectiveness in determining the optical absorption. The imaginary part of dielectric function is a representation of the damping and considered as an effective parameter to investigate the optical absorption. For pristine $LiNbO_3$, as shown in figure 4a, there are three peaks in the imaginary part of the dielectric function with the strongest one at around $\sim$3.6 eV and the subsequent-strongest one at $\sim$5 eV followed by the smallest one at $\sim$8 eV. Further, for all the metal-doped structures included in this analysis, the peaks are significantly shifted towards the lower energies and are broadly found in the range of $\sim$0–1 eV, as shown in figure 4b–i.

For pristine and all the metal-doped structures, the peaks of the imaginary part of the dielectric function are in consistency with the energy-dependent absorption coefficient as illustrated in figure 5. The presence of these peaks of dielectric function at low energies (for all the metal-doped structures) signifies a smaller excitation binding energy that in turn results in a higher absorption in the visible region of the optical spectrum.

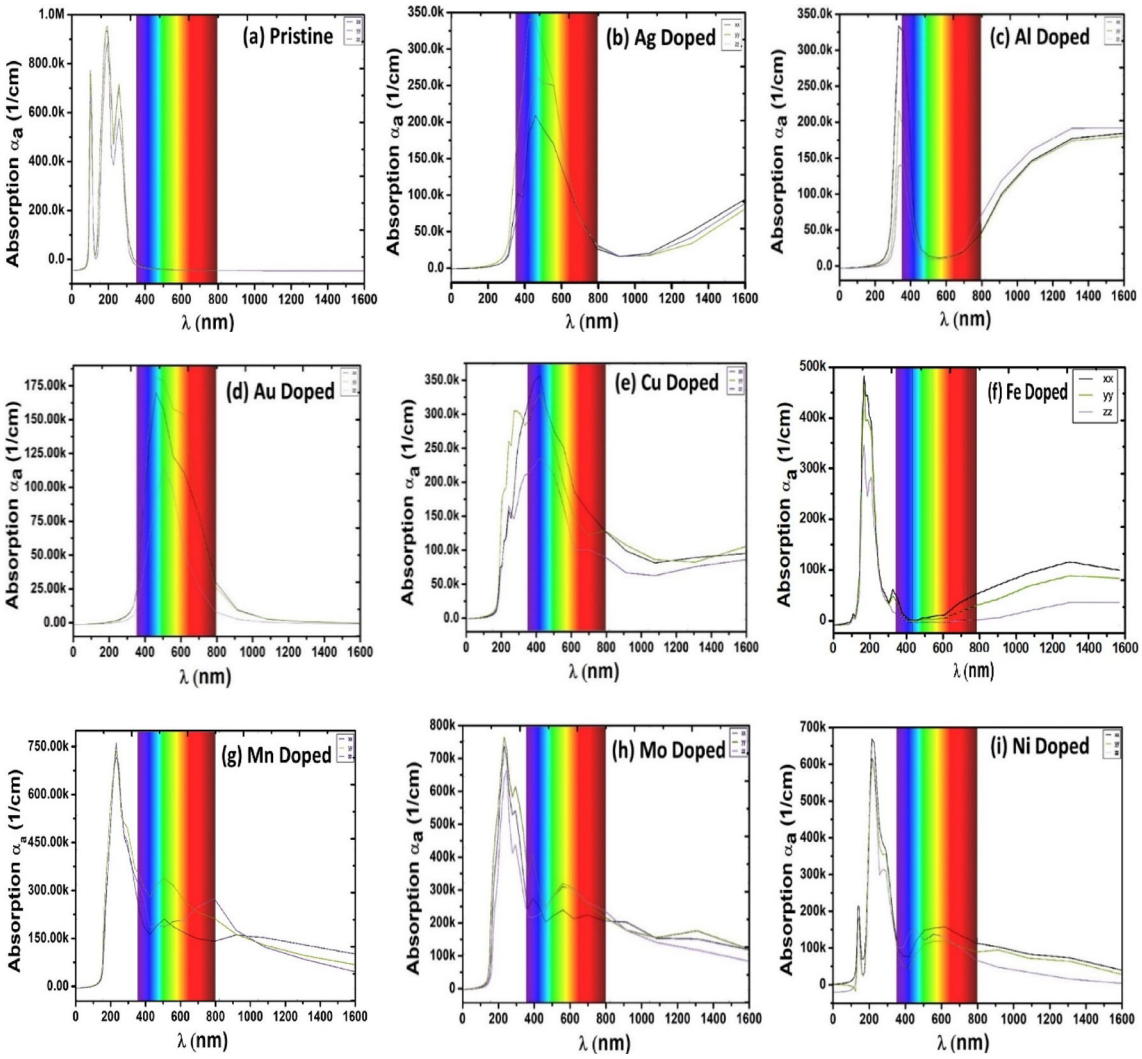


Figure 6. Wavelength-dependent absorption coefficient for (**a**) pristine $LiNbO_3$, and doped with (**b**) Ag, (**c**) Al, (**d**) Au, (**e**) Cu, (**f**) Fe, (**g**) Mn, (**h**) Mo and (**i**) Ni, along the xx, yy and zz tensors.

Further, the energy-dependent and wavelength-dependent absorption coefficients are illustrated in figures 5 and 6 for the pristine and metal-doped $LiNbO_3$, respectively. The absorption coefficient of any material is an effective indicative of the efficiency of solar energy conversion and signifies the propagation length of light in any material. As illustrated in figure 5a, the absorption starts from ∼3.6 eV which is in consistency with the bandgap calculated in the earlier section for the pristine structure. However, for metal-doped $LiNbO_3$, the absorption edge of absorption coefficient starts at lower energies (figure 5). Compared with that of pristine $LiNbO_3$ (3.6 eV), the absorption edge of metal-doped (M = Al, Au, Ag, Cu, Fe, Mn, Mo and Ni) shifts to the lower energy direction (long-wave direction). This shift predominantly improves the optical performance.

The wavelength-dependent absorption coefficient for the metal dopants: Ag, Al, Au, and Cu are illustrated in figure 6b–e. As shown in figure 6a, the absorption in the pristine structure lies completely in the UV region of the optical spectrum with absorption edge at ∼350 nm. This absorption edge is in strong consistency with the computed bandgap (E_g = 3.56 eV). It can be concluded from the absorption coefficient that there is a significant shift in the absorption edge towards the visible and near-visible regions (NVR) of the optical spectrum for the dopant Ag, Al, Au and Cu. It can be seen from figure 6b and d that the absorption has been completely shifted to 400–800 nm for the dopant Ag and Au. However, for dopant Al (figure 6b); the absorption is in near-visible region, i.e., around 400 nm and for dopant Cu (figure 6e), it is in 400–600 nm. This shifting of absorption in visible region for the dopants like Au, Ag, Cu and Al occurrs due to the surface plasmonic resonance phenomena. These dopants are in the category of plasmonic metal. The doping with these plasmonic metals leads to the oscillation of electrons matched to the frequency of incident radiation. Metals like Au, Ag, Cu and Al are considered as potential candidate for the plasmonic applications. Among these metals, Ag is superior for the

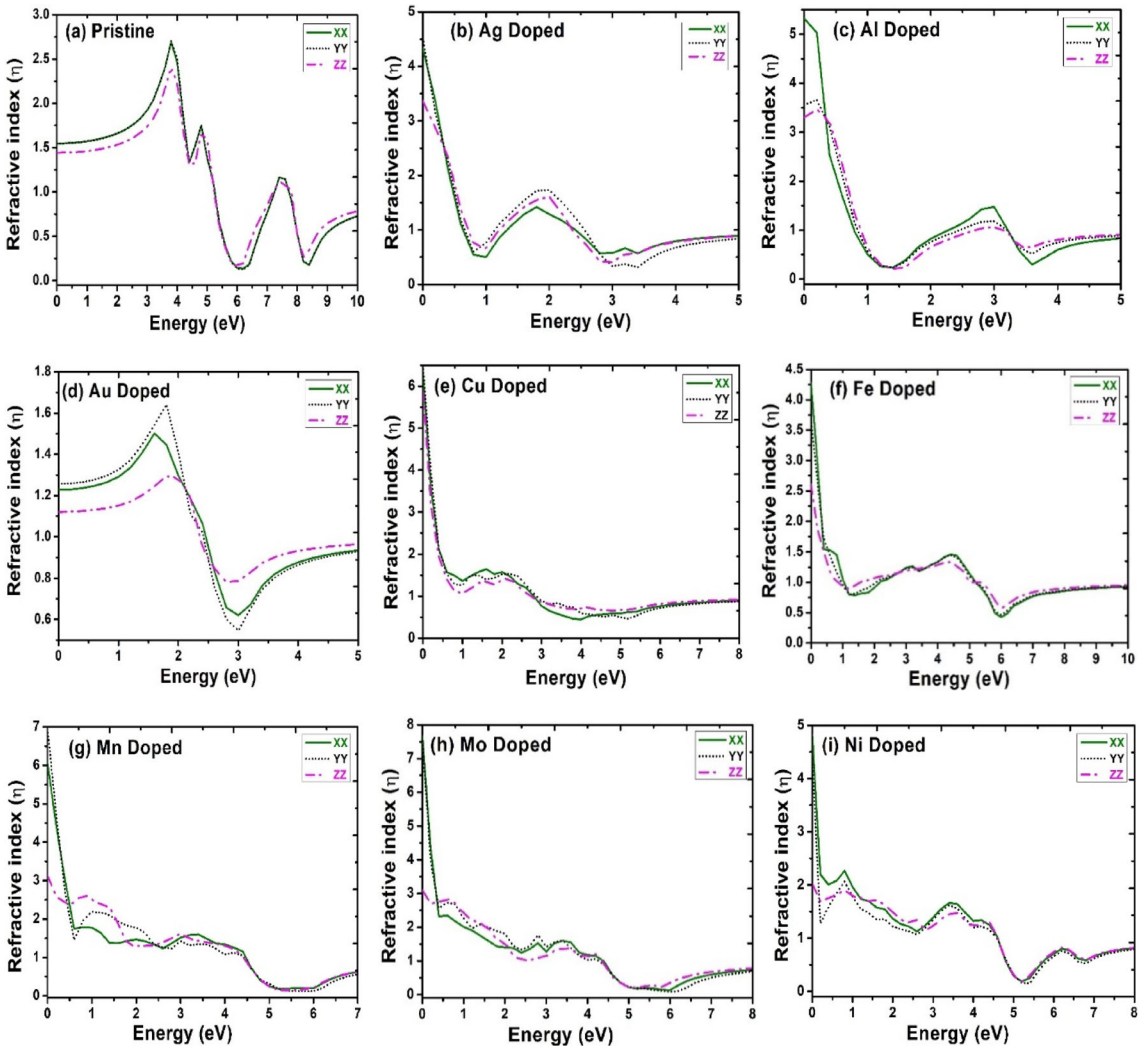


Figure 7. Energy-dependent refractive index for (**a**) pristine $LiNbO_3$, and doped with (**b**) Ag, (**c**) Al, (**d**) Au, (**e**) Cu, (**f**) Fe, (**g**) Mn, (**h**) Mo and (**i**) Ni, along the xx, yy and zz tensors.

Table 3. Comparison of the optical properties of pristine and metal-doped $LiNbO_3$.

Optical properties	Pristine $LiNbO_3$	Ag/Au-doped	Al/Cu-doped	Ni/Fe/Mn/Mo-doped
Optical absorption-UV region	High	Missing	Missing	High
Absorption-visible region	Missing	High	High in visible and NVR	Low
Suitability-optoelectronic applications	Not good	Best	Better	Good

plasmonic effect because of its least optical losses in the visible region. However, Ag oxidizes in a rapid manner and exhibits losses because of roughness at the surface. On the contrary, Au outperforms in the visible region that too with superior chemical stability. Conversely, the chemical instability of Cu and Al restricts them from the wide applicability.

The above facts are strongly supported by wavelength-dependent absorption coefficient, shown in figure 6b–e. It can be inferred from figure 6 that the optical spectrum is completely tuned in the visible region for the dopant Ag (figure 6b) and Au (figure 6d) due to their extraordinary absorbing and scattering efficiency. In these metals, the conduction electrons in the surface experience a collective oscillation for a specific wavelength during interaction with incident light. This oscillation may be identified as plasmon resonance, and thus, results in a higher absorption in the visible region. For the other plasmonic dopants; Al

(figure 6c) and Cu (figure 6e), the absorption is shifted towards visible and near-visible region (NVR). The value of absorption coefficient is the highest for Ag-doped $LiNbO_3$ among all the metal-dopants utilized in this study and thus, making it the best among all. Subsequently, for other metal-dopants like Fe, Mo, Mn, Mo and Ni, a red-shift along with slightly improved absorption in the visible spectra is observed, as illustrated in figure 6f–i.

Overall, the shifting of edges and improvement in absorption in the visible region for all the M-lithium niobite (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni) are attributed to the newly created energy levels near to the filled conduction band. This results in a red-shift as illustrated in figure 6. Further, it can also be concluded that due to metal doping, the equilibrium carrier concentration gets altered that led to the formation of traps. These traps ultimately reduce the charge carrier recombination rate and thus, improve the photovoltaic activity. As a conclusive remark, it can be stated that the doping with metal-dopant used in this study resulted in an improved absorption of visible light. The outcomes of the absorption coefficient are found consistent with the calculated dielectric function. Further, a broader range of absorption can easily be observed for the structures doped with plasmonic metals (M = Au, Ag, Al and Cu), from figure 6b–e. This broader range clearly signifies a stable and enhanced absorption in the visible range of the optical spectrum due to the doping of plasmonic metals. Owing to the improved visible light absorption, metal-doped $LiNbO_3$ structures (specifically plasmonic metal-doped) are proven a prominent candidate for the optoelectronics- and photonic-based applications.

The refractive index is computed from the real part of the dielectric function and is demonstrated in figure 7 for pristine and metal-doped $LiNbO_3$. The peaks in the refractive index follow the imaginary part of the dielectric function, shown earlier in figure 3, which illustrates the consistency of the refractive index with dielectric function. A significant improvement is observed in the refractive index for all the metal-dopants as shown in figure 7b–i than that of pristine $LiNbO_3$ shown in figure 7a. This implies comparatively a longer life of light in the crystal, resulting in the higher absorption of the photon in the respective spectrum range.

The comparison of optical properties and application validity of the pristine and M-doped $LiNbO_3$ (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni) is presented in table 3. It is clear from the summarized data that Ag- and Au-doped $LiNbO_3$ are the best suits for the optoelectronic- and photonic-based applications.

4. Conclusion

Here, the electronic and optical properties of pristine and M-doped $LiNbO_3$ (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni) are computed by using DFT-based simulations. Due to large bandgap, pristine $LiNbO_3$ cannot utilize the visible portion of the optical spectrum. The absorption in the visible region can be enhanced by means of doping with a suitable metal dopant. It is found from the calculated bandgap of the M-$LiNbO_3$ that the metal doping is helpful to improve the optoelectronic performance of $LiNbO_3$ owing to the narrowing of bandgap. In this study, the pristine $LiNbO_3$ is doped with the metals (M = Fe, Mn, Mo and Ni) along with a superior class of metal, i.e., plasmonic metal (M = Au, Ag, Al and Cu). In the M-$LiNbO_3$ (M = Au, Ag, Al and Cu), due to surface plasmonic resonance, the optical absorption is extraordinarily enhanced in the visible region. Among the plasmonic dopants, the enhancement is very much significant for the dopants Au and Ag, due to their extraordinary plasmonic properties. The other plasmonic dopants Cu and Al also showed an enhanced absorption in the visible region. However, for the other metal-dopants like Ni, Fe, Mn and Mo, a slightly increased absorption in the visible region along with red-shift is observed. It has been shown that the optical properties of $LiNbO_3$ are enhanced significantly by doping of metal dopants (M = Au, Ag, Al, Cu, Fe, Mn, Mo and Ni). Hence, it is envisaged that the metal-doped $LiNbO_3$ may find vital usage in photonic- and optoelectronics-based applications.

References

[1] Atta N F, Galal A, Ads E E and Pan L 2016 *Intech* **93052** 107

[2] Quan L N, Rand B P, Friend R H, Mhaisalkar S G, Lee T W and Sargent E H 2019 *Chem. Rev.* **119** 7444

[3] Huang H, Polavarapu L, Sichert J A, Susha A S, Urban A S and Rogach A L 2016 *NPG Asia Mater.* **8** 328

[4] Hu F, Zhang H, Sun C, Yin C, Lv B, Zhang C *et al* 2015 *ACS Nano.* **9** 12410

[5] Amri A M A, Cheng B and He J 2018 *IEEE Trans. Nanotechnol.* **18** 1

[6] Dogan F, Lin H, Viry M G and Pena O 2015 *Sci. Technol. Adv. Mater.* **16** 020301

[7] Negi S, Mittal P and Kumar B 2020 *J. Electron. Mater.* **49** 4610

[8] Quintero O A J, Sanchez R S, Rincon M and Sero I M 2015 *J. Phys. Chem. Lett.* **6** 1883

[9] Tan Z K, Moghaddam R S and Lai M L 2014 *Nat. Nanotech.* **9** 687

[10] Negi S, Mittal P and Kumar B 2020 *J. Inf. Disp.* **28** 1

[11] Negi S, Mittal P and Kumar B 2018 *Microsyst. Technol.* **24** 1

[12] Prezas P R and Graça M P F 2016 in *Applications of molecular spectroscopy to current research in the chemical and biological sciences*, Intech, **2016** 32 https://doi.org/10.5772/61896

[13] Klein R S, Kugel G E, Maillard A, Polgar K and Peter A 2003 *Optic. Mater.* **22** 171

[14] Sohler W, Hu H, Ricken R, Quiring V, Vannahme C, Herrmann H *et al* 2008 *Opt. Photonics News* **19** 24

[15] Abdi F, Fontana M D, Aillerie M and Bourson P 2016 *Appl. Phys. A* **83** 427
[16] Javid A, Khan Z, Mehmood Z and Nabi A 2018 *Int. J. Mod. Phys. B* **32** 1850168
[17] Sun B, Gou J, Wang J and Jiang Y 2019 *Proceedings of 9th international symposium on advanced optical manufacturing and testing technologies: optoelectronic materials and devices for sensing and imaging*
[18] Choudhary S and Garg A 2019 *IEEE Trans. Nanotechnol.* **18** 989
[19] Rahaman M and Hossain A K M 2018 *RCS Adv.* **8** 33010
[20] Song K and Han X 2013 *J. Alloys Compd.* **551** 118
[21] Pathak Nimai, Ghosh Partha, Mukherjee and Mandal Balaji 2020 *RSC Adv.* **10** 31070
[22] Hui H D, Sheng Y J, Long C Q, Jie W M, Qilong L, Liang S *et al* 2014 *Chin. Phys. Lett.* **31** 037103
[23] Bachiri A E, Hasnaoui M E, Bennani F and Bousselamti M 2016 *J. Mater. Environ. Sci.* **7** 3353
[24] Vilela D, González M C and Escarpa A 2012 *Anal. Chim. Acta* **751** 24
[25] Huang X and Sayed M A E 2010 *J. Adv. Res.* **1** 13
[26] Li M, Shi L, Xie T, Jing C, Xiu G and Long Y T 2017 *ACS Sensors* **2** 263
[27] Jabeen M and Haxha S 2018 *IEEE J. Quant. Electron.* **54** 1
[28] Negi S, Mittal P, Kumar B and Juneja P 2019 *Microelectron. Eng.* **218** 111154
[29] Sena P, Sen P K, Bhatt R, Kar S, Shukla V and Bartwal K S 2004 *Solid State Commun.* **129** 747
[30] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865
[31] Monkhorst H J and Pack J D 1976 *Phys. Rev. B* **13** 5188
[32] Kochar R and Choudhary S 2018 *IEEE J. Quant. Electron.* **54** 1
[33] Tran F and Blaha P 2009 *Phys. Rev. Lett.* **102** 226401
[34] Hossain M M 2019 *Heliyon* **5** 01436
[35] Martin R M 2014 *Basic theory and practical methods* (New York: Cambridge University Press)
[36] Griffithis D J 1999 *Introduction to electrodynamics* (New Jersey: Prentice Hall)
[37] Weis R S and Gaylord T K 1985 *Appl. Phys. A* **37** 191
[38] Liu X, Que W, He Y and Zhou H 2018 *J. Adv. Dielectr.* **8** 1820002
[39] Yu J and Liu X 2007 *Mat. Lett.* **61** 355
[40] Mittal P 2021 *J. Soc. Inf. Disp.* https://doi.org/10.1002/jsid.1007
[41] Fares N E H and Bouarissa N 2014 *Mater. Res.* **21** 20170964
[42] Kong L J, Liu G H and Zhang Y J 2016 *RSC Adv.* **6** 10919
[43] Chamola P and Mittal P 2020 *Int. J. Light Electron Opt.* **224** 165626
[44] Rizwan M, Ali A, Usman Z, Khalid N R, Jin H B and Cao C 2019 *Phys. Condens. Matter* **552** 52

Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Empirical Validation of cross-version and 10-fold cross-validation for Defect Prediction

Ruchika Malhotra
Department of Software Engineering
Delhi Technological University
Delhi, India
ruchikamalhotra2004@yahoo.com

Shweta Meena
Department of Software Engineering
Delhi Technological University
Delhi, India
shweta82336@gmail.com

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9533030

***Abstract*—Nowadays, the development of a prediction model is one of the most important research area. Prediction models are helpful in providing accurate results for unseen data or future data. The most important phase of the software development life cycle (SDLC) is testing. During testing, we face some uncertain behavior of the program due to the presence of defects. To remove these defects in the early stage of SDLC, we design Software Defect Prediction Model (SDPM). Although many SDPM have been developed using various Machine Learning (ML) and statistical technique, but the generalizability of results have prevailed because the developed models use the same dataset for training and testing. This study aims to develop SDPM using cross-version, which using two different versions of a project for training and testing. The historical labeled defect data of the previous version is used for the updated or upcoming version for defect prediction which is termed as Cross-Version Defect Prediction (CVDP). To complete experimentation, we have used 26 datasets from an open-source repository. The performance of SDPM is analyzed using performance metrics. The SDPM is also developed using 10-cross validation. In the end, the comparison of CVDP and 10-cross validation has been done using a statistical test. The aim of conducting this study is to analyze the applicability of cross-version defect prediction when a sufficient amount of data is not available for training and testing. According to statistical analysis, it has been observed that cross-version can be used if we have to test our prediction model for unseen projects or upcoming projects.**

Keywords—Defect Prediction, Machine Learning, Cross-Version, Cross-Validation, Cross-Project, Statistical Test

I. INTRODUCTION

The software project size is increasing everyday, it directly impacts the cost and complexity of software. Due to which the defects in the software are indispensable [1] [2]. The detection and correction of defects before the beta release of the software is not feasible to assure software quality of software. The defect prediction focuses on the identification of the defect-prone modules and files of a software project. Defect prediction helps researchers, practitioners, and testers to test suspicious modules on priority. The prediction models developed in existing studies are supervised classification models that are developed using labeled data of a project and can be used for prediction on unlabeled dataset of same project that is termed as Within-Project Defect Prediction (WPDP). There is a lack of data availability due to which an efficient and reliable defect prediction model could not be developed using WPDP. The model development using WPDP requires huge amount of training data to overcome this issue.

To resolve this limitation, an emerging concept of transfer learning is employed. In transfer learning [3], a prediction model is developed by utilizing the labeled data of other projects that is termed is Cross-Project Defect Prediction (CPDP). The performance of the model developed using the CPDP concept is not effective, due to data distribution differences across various projects. Existing studies in the literature created a prediction model by labeled data of earlier versions and predict the unlabeled module's data of upcoming or future versions of the same project that is termed as CVDP [4] [5]. In the upgraded version of any software some of the modules would be deleted, some of the modules would be added, some additional functionality would be added. But, in terms of data distribution difference, the differences across multiple versions of the same project are low in comparison to data distribution difference across cross-projects because the upcoming version carries a huge amount of information or data from the existing version of that software [6].

Consider Project X.1, Project X.2, and Project Y. If Project X.1 is used for training and testing then it comes under WPDP. If project X.1 is used for training and Project X.2 is used for testing then it comes under CVDP. If Project X is used for training and Project Y is used for testing, then it comes under CPDP. This study demonstrated the applicability of CVDP.

However, various studies exist in the literature for change and defect prediction using CVD [7] [8] [9], but no study discussed the applicability of CVDP for generalization. This paper aims at the investigation of CVDP for generalizability. In literature, existing studies developed defect prediction models using either cross-project or WPDP. But to enhance the generalizability of results and removing the bias this study analyzes the performance of cross-version and 10-fold cross-validation. Thus, for the sake of achieving our research goal, the formulated research questions are as follows:

RQ1: What is the predictive capability of defect prediction models developed using cross-version validation?

RQ2: What is the predictive capability of the defect prediction model developed using 10-fold cross-validation?

RQ3: What is the feasibility of implementing cross-version in comparison to the performance of the model developed using 10-fold cross-validation i.e. to what extent cross-version validation is feasible?

In this study, the SDPM is developed using various ML techniques. The seven ML techniques are used for developing prediction model such as Naïve Bayes (NB), Logistic Regression (LR), Multilayer Perceptron (MLP), AdaBoost (AB), Bagging, J48, and Random Forest (RF). The generalizability of the prediction model is analyzed using cross-version and 10-fold cross-validation. The organization of rest of the paper is as follows: Section II

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

discussed the related work. The independent variables, dependent variables used and the descriptive statistics of the dataset used in this study are summarized in Section III. In Section IV, the answers to research questions are discussed. Section V summarized the threats to the validity of this study. Section VI summarized the conclusion and future work.

II. RELATED WORK

To analyze the effect of WPDP, CPDP, HCPDP, and CVDP various studies have been conducted.

WPDP: In WPDP, the classifier model is developed for projects having a sufficient amount of labeled data for training and testing within the same project. The existing literature analyzed the performance of classifiers [10], [11] developed for WPDP, feature selection techniques [12], sampling methods [13], and model validation techniques [14].

CPDP: CPDP is mainly used for projects having lack of data for prediction or testing. The researcher proposed a transfer learning concept for the project with a lack of data. CPDP can be done using transfer learning by transferring knowledge from other projects i.e., source project for predicting other projects i.e., target project [15]. Instance transfer, function transfer, relational information transfer, and parameter transfer are types of transfer learning. Transfer Component Analysis (TCA) has been developed to balance the data distribution between source and target dataset [16]. A hybrid composition model has been developed by combining ensemble learning and genetic algorithm for CPDP [17].

HCPDP: In HCPDP, the feature sets of projects used for source and target datasets are different. Various studies exist for HCPDP with different approaches. Since the training and testing projects have a different feature set due to which data distribution is also different. The correlation analysis method has been demonstrated to construct mapping across source and target project data into a common feature space by focusing on similarity mapping [18]. Nam et al. conducted the study [19] in which the methods are employed to find out the features set having the most similarity and the features with the highest similarity are used for conducting HCPDP. According to the analysis of existing studies, WPDP performs better than CPDP and HCPDP. Due to data distribution between source and target dataset, HCPDP does not perform better than WPDP [20]–[23]. The HCPDP study conducted by Jing et al. [18] reported results only for 22 HCP pairs out of total of 90 pairs of source and target project and it was verified universally. The method proposed by Nam et al. [19] to find out the similarity matching pairs across source and target datasets failed in many cases due to low matching scores in comparison to the threshold.

CVDP: CVDP is used for projects having multiple versions. The experimentation has been done for CVDP, in which the prediction model utilizes the labeled modules of a prior version of the current project [24]. CVDP is a methodology between WPDP and HCPDP. Since software development is a dynamic process that results in changes of the data structure characteristics for each updated version, the upcoming or updated version utilizes a large amount of information from the existing version of the project. The data distribution difference across multiple versions of the same project is smaller than the data distribution across completely different projects. According to the above statement, CVDP is more beneficial for defect prediction across multiple versions of the same project.

III. INDEPENDENT AND DEPENDENT VARIABLES USED

All the variables used in this study are discussed in this section.

A. Independent Variables

Model development has two phases one is training and another one is testing. The dataset used for training a model comprises of independent/input/predictor (software metrics) and dependent/output/outcome variables (prediction of defect prone module/ prediction of change prone module). In this study, we have used 26 versions of 10 different software projects given by Madeyski and Jureczko [25]. Each project consists of a class file of JAVA projects. These modules consist of 20 software metrics and one binary predictor variable for the defect proneness of a module (if a module is defective, it is labeled as 1, if the module is non-defective it is labeled as 0). The detailed description of these modules is provided in an existing study [26]. Detailed statistics description of all the 26 versions of 10 projects is summarized in Table I. The independent variables in the dataset are Chidamber and Kemerer (C&K) object-oriented metrics.

Table I: Statistics of Dataset

Project	# of Modules	# of Defective Modules	% of Defective Modules
ant_ver1.6	352	92	26.10%
ant_ver1.7	745	166	22.30%
camel_ver1.0	329	13	3.95%
camel_ver1.2	608	216	35.50%
camel_ver1.4	872	145	16.60%
camel_ver1.6	965	188	19.50%
ivy_ver2.0	352	40	11.36%
jedit_ver3.2	272	90	33.08%
jedit_ver4.0	306	75	24.50%
jedit_ver4.1	312	79	25.30%
jedit_ver4.2	367	48	13.10%
jedit_ver4.3	492	11	2.20%
log4j_ver1.0	135	34	25.20%

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

log4j_ver1.1	109	37	33.90%
log4j_ver1.2	205	189	92.20%
lucene_ver2.0	195	91	46.70%
lucene_ver2.2	247	144	58.30%
lucene_ver2.4	340	203	59.70%
poi_ver2.5	385	248	64.42%
poi_ver3.0	442	281	63.60%
synapse_ver1.2	266	86	32.33%
xalan_ver2.4	723	110	15.20%
xalan_ver2.5	803	387	48.20%
xalan_ver2.6	885	411	46.40%
xalan_ver2.7	909	898	98.80%
xerces_ver1.4	573	426	74.35%

B. Independent and Dependent Variables

The independent variables are object-oriented metrics (OOM) in a used dataset. Any kind of fault and bug results in some uncertain situation or unexpected failure of software known as DEFECT. In this analysis, the dependent variable is a defect. This is a nominal variable that indicates the presence and absence of a defect in a module ("YES" or "1" represents defective modules and "NO" or "0" represents a non-defective module).

IV. RESULT ANALYSIS

The answers to the research questions formulated in Section I are addressed in this section to achieve the objective of this study.

A. RQ1: What is the predictive capability of defect prediction models developed using cross-version validation?

To answer this RQ, we have experimented using seven ML techniques on different versions of all 8 projects out of 10 projects, 2 projects have one version only. We have created a prediction model for all the 30 combinations and the AUC values have been noted down in Table II. From the AUC values, it has been observed that in some of the cases CVDP is outperformed. The results are statistically validated using the Friedman test followed by post-hoc analysis that is Wilcoxon signed-rank test. For the Friedman test, we have formed and tested the following hypothesis.

Null Hypothesis (H_o):- There is no significant difference among the performance of all ML techniques for CVDP.

Alternate Hypothesis (H_a): There is a significant difference among the performance of all ML techniques for CVDP.

The ranks obtained using Friedman test of all the ML techniques are shown in Table III. The significance value obtained using the Friedman test is 0.000 which is less than the significance value. Hence, H_o is rejected and alternate H_a is accepted.

Table II: AUC values for Cross-Version Defect Prediction

Techniques used	**NB**	**LR**	**MLP**	**AB**	**Bagging**	**J48**	**RF**
source dataset → target dataset							
ant_ver1.6 → ant_ver1.7	0.779	0.778	0.75	0.8	0.826	0.679	0.803
camel_ver1.0 → camel_ver1.2	0.49	0.586	0.56	0.589	0.558	0.544	0.574
camel_ver1.0 → camel_ver1.4	0.482	0.644	0.636	0.631	0.644	0.565	0.677
camel_ver1.0 → camel_ver1.6	0.552	0.63	0.618	0.639	0.638	0.552	0.651
camel_ver1.2 → camel_ver1.4	0.687	0.68	0.666	0.668	0.761	0.654	0.765
camel_ver1.2 → camel_ver1.6	0.642	0.589	0.593	0.619	0.656	0.621	0.677
camel_ver1.4 → camel_ver1.6	0.633	0.628	0.679	0.64	0.68	0.632	0.682
jedit_ver3.2 → jedit_ver4.0	0.751	0.776	0.788	0.755	0.8	0.756	0.824
jedit_ver3.2 → jedit_ver4.1	0.759	0.81	0.788	0.775	0.795	0.727	0.813
jedit_ver3.2 → jedit_ver4.2	0.785	0.805	0.787	0.804	0.825	0.738	0.833
jedit_ver3.2 → jedit_ver4.3	0.608	0.64	0.589	0.575	0.586	0.531	0.625
jedit_ver4.0 → jedit_ver4.1	0.799	0.811	0.763	0.808	0.839	0.667	0.836
jedit_ver4.0 → jedit_ver4.2	0.826	0.824	0.764	0.83	0.823	0.655	0.852
jedit_ver4.0 → jedit_ver4.3	0.628	0.702	0.672	0.645	0.631	0.611	0.617
jedit_ver4.1 → jedit_ver4.2	0.837	0.872	0.797	0.817	0.842	0.784	0.858
jedit_ver4.1 → jedit_ver4.3	0.629	0.669	0.643	0.586	0.599	0.606	0.688
jedit_ver4.2 → jedit_ver4.3	0.624	0.648	0.595	0.594	0.686	0.759	0.706
log4j_ver1.0 → log4j_ver1.1	0.84	0.802	0.802	0.803	0.783	0.748	0.785
log4j_ver1.0 → log4j_ver1.2	0.621	0.728	0.708	0.66	0.694	0.586	0.666

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

log4j_ver1.1 → log4j_ver1.2	0.635	0.626	0.593	0.64	0.504	0.413	0.516
lucene_ver2.0 → lucene_ver2.2	0.64	0.641	0.615	0.666	0.657	0.564	0.658
lucene_ver2.0 → lucene_ver2.4	0.68	0.704	0.65	0.671	0.702	0.636	0.699
lucene_ver2.2 → lucene_ver2.4	0.72	0.678	0.658	0.701	0.648	0.658	0.639
poi_ver2.5 → poi_ver3.0	0.766	0.677	0.597	0.613	0.614	0.548	0.619
xalan_ver2.4 → xalan_ver2.5	0.584	0.619	0.567	0.632	0.646	0.472	0.619
xalan_ver2.4 → xalan_ver2.6	0.756	0.702	0.586	0.707	0.693	0.46	0.69
xalan_ver2.4 → xalan_ver2.7	0.79	0.762	0.528	0.6	0.778	0.743	0.808
xalan_ver2.5 → xalan_ver2.6	0.581	0.635	0.71	0.688	0.725	0.672	0.688
xalan_ver2.5 → xalan_ver2.7	0.705	0.661	0.768	0.826	0.821	0.722	0.883
xalan_ver2.6 → xalan_ver2.7	0.835	0.841	0.778	0.9	0.903	0.868	0.908

Table III: Mean Rank of ML technique

ML Technique	Mean Rank
J48	1.93
MLP	3.23
NB	3.65
AB	4.22
LR	4.58
Bagging	4.82
RF	5.57

Wilcoxon signed-rank test is performed to validate the alternate hypothesis of the Friedman test. The Wilcoxon signed-rank test results are shown in Table IV. All the ML techniques are paired with other techniques, in such a manner, total of 21 pairs are formed. The S denotes a significant difference and NS denotes no significant difference among that pair. If the p-value is greater than the significance level (0.05), then it indicated that there is no significant difference among that pair. From Table IV, it has been observed that RF performed better than NB in CVDP. MLP and J48 performed better than LR. J48 and RF performed better than MLP. J48 performed better than Bagging and AB. Hence, RF performed better than J48 all other ML techniques for CVPD. Thus, RF can be used for CVDP in upcoming projects with multiple versions of the same project.

Therefore, for RQ1 we conclude that:
For CVDP, defect prediction models developed with the applications of J48, MLP, NB, AB, LR, bagging and RF gave AUC in the range of 0.49 to 0.908 overall 30 pairs. RF is outperformed in comparison to all other ML techniques used in a study for prediction defect-prone modules based on the Friedman test result.

Table IV: Wilcoxon Signed Rank test for CVDP

Technique	p-value	Difference (S/NS)
LR - NB	0.361	NS
MLP - NB	0.361	NS
ABABOOST - NB	0.361	NS
BAGGING - NB	0.045	NS
J48 - NB	0.009	NS
RANDOMFOREST - NB	0.002	S
MLP - LR	0.001	S
ABABOOST - LOG	1.000	NS
BAGGING - LOG	1.000	NS
J48 - LOG	0.002	S
RANDOMFOREST - LOG	0.265	NS
ABABOOST - MLP	0.100	NS
BAGGING - MLP	0.018	NS
J48 - MLP	0.000	S
RANDOMFOREST - MLP	0.002	S
BAGGING - ABABOOST	0.100	NS
J48 - ABABOOST	0.001	S
RANDOMFOREST - ABABOOST	0.026	NS
J48 - BAGGING	0.000	S
RANDOMFOREST - BAGGING	0.045	NS
RANDOMFOREST - J48	0.000	S

B. RQ2: What is the predictive capability of defect prediction model developed using 10-fold cross validation?

To answer this RQ, the performance of SDP is evaluated and analyzed on the same training and testing dataset. The performance of the evaluated model is noted down in Table V for all 26 datasets. NB does not perform better in comparison to other ML techniques. The AUC values obtained using NBL are low. Here, the prediction model is developed using 10-fold cross validation for all 26 datasets using 7 ML techniques. To validate the results are statistically validated using the Friedman test followed by post-hoc analysis that is Wilcoxon signed-rank test. The Friedman test is performed with a significance level of 0.05 and 6 degree of freedom. The hypothesis created and tested for the Friedman test is as follows:

Null Hypothesis (H_o):- For 10-fold cross validation, there is no significant difference in the efficiency of all ML techniques.

Alternate Hypothesis (H_a): For 10-fold cross validation, there is a significant difference in the efficiency of all ML techniques.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

Table V: AUC values for 10-fold cross validation

Technique used source dataset → target dataset	NB	LR	MLP	AB	Bagging	J48	RF
ant_ver1.6 → ant_ver1.6	0.823	0.844	0.807	0.82	0.882	0.82	0.904
ant_ver1.7 → ant_ver1.7	0.814	0.823	0.8	0.843	0.88	0.738	0.912
camel_ver1.0 → camel_ver1.0	0.834	0.736	0.735	0.883	0.825	0.68	0.863
camel_ver1.2 → camel_ver1.2	0.59	0.654	0.666	0.649	0.805	0.714	0.853
camel_ver1.4 → camel_ver1.4	0.678	0.678	0.637	0.652	0.701	0.592	0.73
camel_ver1.6 → camel_ver1.6	0.69	0.72	0.735	0.724	0.855	0.762	0.885
ivy_ver2.0 → ivy_ver2.0	0.809	0.827	0.803	0.83	0.879	0.8	0.916
jedit_ver3.2 → jedit_ver3.2	0.781	0.855	0.864	0.851	0.879	0.803	0.925
jedit_ver4.0 → jedit_ver4.0	0.769	0.811	0.832	0.813	0.88	0.804	0.914
jedit_ver4.1 → jedit_ver4.1	0.795	0.848	0.829	0.79	0.87	0.794	0.905
jedit_ver4.2 → jedit_ver4.2	0.845	0.864	0.838	0.873	0.906	0.772	0.928
jedit_ver4.3 → jedit_ver4.3	0.707	0.75	0.819	0.896	0.855	0.719	0.92
log4j_ver1.0 → log4j_ver1.0	0.863	0.821	0.846	0.84	0.861	0.767	0.903
log4j_ver1.1 → log4j_ver1.1	0.832	0.877	0.872	0.904	0.882	0.788	0.914
log4j_ver1.2 → log4j_ver1.2	0.792	0.81	0.851	0.823	0.887	0.798	0.936
lucene_ver2.0 → lucene_ver2.0	0.75	0.765	0.743	0.771	0.826	0.712	0.863
lucene_ver2.2 → lucene_ver2.2	0.624	0.643	0.706	0.699	0.757	0.687	0.826
lucene_ver2.4 → lucene_ver2.4	0.756	0.792	0.81	0.77	0.85	0.694	0.893
poi_ver2.5 → poi_ver2.5	0.773	0.849	0.879	0.888	0.918	0.875	0.944
poi_ver3.0 → poi_ver3.0	0.8	0.824	0.818	0.883	0.906	0.831	0.932
synapse_ver1.2 → synapse_ver1.2	0.768	0.789	0.81	0.823	0.843	0.757	0.892
xalan_ver2.4 → xalan_ver2.4	0.757	0.801	0.812	0.815	0.881	0.766	0.916
xalan_ver2.5 → xalan_ver2.5	0.604	0.695	0.719	0.71	0.838	0.759	0.89
xalan_ver2.6 → xalan_ver2.6	0.781	0.805	0.821	0.828	0.9	0.814	0.924
xalan_ver2.7 → xalan_ver2.7	0.925	0.933	0.822	0.93	0.894	0.811	0.958
xerces_ver1.4 → xerces_ver1.4	0.858	0.931	0.943	0.954	0.967	0.901	0.974

The ranks obtained using Friedman test of all the ML techniques are shown in Table VI. The computed p-value obtained using the Friedman test is 0.000 which is less than the significance value. Hence, H_o is rejected and alternate H_a is accepted.

Wilcoxon signed-rank test is performed to validate the alternate hypothesis of the Friedman test. Table VI shows the Wilcoxon signed-rank test results. All the ML techniques are paired with other techniques, in such a manner, total 21 pairs are formed. The S denotes a significant difference and NS denotes no significant difference among that pair. If the p-value is greater than the significance level (0.05), then it indicated there is no significant difference among that pair.

Table VI: Mean Rank of ML technique

ML Technique	Mean Rank
NB	2.28
J48	2.28
LR	3.25
MLP	3.37
AB	4.23
BAGGING	5.70
RF	6.88

From Table VI, it has been observed that LR, AB, Bagging, RF performed better than NB. Bagging and RF performed better than LR. Bagging and RF performed better than MLP. Bagging, J48, and RF performed better than AB. J48 and RF performed better than Bagging. RF performed better than J48. Hence, RF outperformed all other ML techniques for 10-fold cross-validation. Therefore, for RQ2 we conclude that:

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

For 10-fold cross-validation, defect prediction models developed with the applications of J48, MLP, NB, AB, LR, bagging and RF gave AUC in the range of 0.59 to 0.974 overall 26 pairs. RF outperformed in comparison to all other ML techniques used in this paper for the prediction of defect-prone modules based on the Friedman test result.

Table VII: Wilcoxon Signed Rank test for 10-fold cross validation

Technique	p-value	Difference (S/NS)
LR - NB	0.000	S
MLP - NB	0.201	NS
AB - NB	0.001	S
Bagging - NB	0.000	S
J48 - NB	0.855	NS
RF - NB	0.000	S
MLP – LR	0.584	NS
AB – LR	0.018	NS
BAGGING – LR	0.000	S
J48 – LR	0.100	NS
RF – LR	0.000	S
AB - MLP	0.201	NS
BAGGING - MLP	0.000	S
J48 - MLP	0.006	NS
RF - MLP	0.000	S
Bagging – AB	0.001	S
J48 – AB	0.001	S
RF – AB	0.000	S
J48 - Bagging	0.000	S
RF - Bagging	0.000	S
RF - J48	0.000	S

C. RQ3: What is the feasibility of implementing cross-version in comparison to the performance of model developed using 10-fold cross validation i.e. to what extent cross version validation is feasible?

The performance of prediction models developed using various ML techniques using cross-version validation should not vary significantly from the model developed using ten-fold cross-validation by applying the same ML techniques to research the applicability of cross-version methodology to recognize defective modules of the software. Cross-version validation applicability to defect, prediction models, is limited due to the significant amount of difference. The following hypothesis is established and tested to prove it:

Null Hypothesis (H_o): The performance of CVDP and the defect prediction model produced using 10-fold cross validation are not significantly different.

Alternate Hypothesis (H_a): The performance of CVDP and the defect prediction model established using 10-fold cross-validation differs significantly.

To determine the feasibility of cross-version validation, the findings are statistically validated. On the results of seven ML techniques, we used a pairwise Wilcoxon test with a significance level of 0.05. The performance of seven ML techniques is analyzed individually for both cross-version and10-fold cross-validation with respective AUC values. The pair-wise Wilcoxon test results are presented in Table VIII for cross-version and10-fold cross-validation. Table VIII depicts the p-value obtained from the pair-wise Wilcoxon test of seven ML techniques for cross-version and 10-fold cross-validation. From Table VIII it has been observed that the p-value is less than that of significance level (0.05). Therefore, the null hypothesis is rejected and the alternate hypothesis is accepted. Thus, there exists a significant difference among the performance of cross-version and 10-fold cross-validation. As a result of the Wilcoxon test results for RQ3, it is concluded that:

The performance of defect prediction modules developed using cross-version and 10-fold cross-validation differ significantly. Thus, it has been observed that usage of existing versions of software for upcoming software is not feasible in comparison to 10-fold cross-validation for the prediction of defects in upcoming versions of software.

Table VIII: Paired Wilcoxon test on CVDP and model developed using 10 fold cross-validation

Technique	p-value
$NB^{CV} - NB^{TF}$	0.002
$LR^{CV} - LR^{TF}$	0.000
$MLP^{CV} - MLP^{TF}$	0.000
$AB^{CV} - AB^{TF}$	0.000
$Bagging^{CV} - Bagging^{TF}$	0.000
$J48^{CV} - J48^{TF}$	0.000
$RF^{CV} - RF^{TF}$	0.000

V. THREATS TO VALIDITY

The validity threats of this empirical study are discussed in this section. In this study, the open-source datasets of various software are used for experimentation. This may affect the generalizability of upcoming software in different languages. Conclusion threats are the threats that establish an incorrect relationship among the experimentation and its output. In this study, we have analyzed the results using a statistical test to increase the validation of results and it reduces the threat to conclusion validity. Construct validity refers to the degree to which the independent and dependent variables measure the principles that they are supposed to measure. C&K metrics are used as independent variables in this analysis, and they have successfully demonstrated their ability to predict software defects in similar literature. Thus, the threat to construct validity is reduced in this study with respect to the independent variables.

VI. CONCLUSION AND FUTURE WORK

In this study, the feasibility of cross-version is analyzed using 26 versions of 10 different projects. Two experiments are conducted in this study. In the first

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

experiment, the defect prediction model is developed using the cross-version validation method. The defect prediction model is developed using 10-fold cross-validation in the second experiment, and the performance of this model is compared to that of a model developed using cross-version. The results are validated using statistical test. The results obtained from paired Wilcoxon test proved that it is not feasible to implement the CVDP model in comparison to the model developed using 10-fold cross-validation. Cross-version helps in improving the quality of software by using the existing version of software to identify defects in the updated version of that project. It has been observed that the training and testing dataset of the same project is efficient instead of using different versions of a project as a training and testing dataset.

We will try to explore this study further for large projects developed in various programming languages. In future, we will replicate this study by establishing similar data distribution among training and testing projects. To improve the viability of the CVDP model, we intend to investigate techniques such as heuristic-based and hybridized techniques. We will also analyze the feasibility of CVDP in comparison to CPDP.

REFERENCES

[1] N. E. Fenton and N. Ohlsson, "Quantitative analysis of faults and failures in a complex software system," *IEEE Transactions on Software Engineering*, vol. 26, no. 8, pp. 797–814, Aug. 2000.

[2] J. C. Knight, "Safety critical systems: Challenges and directions," in *Proceedings - International Conference on Software Engineering*, 2002, pp. 547–550.

[3] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10. pp. 1345–1359, 2010.

[4] E. Arisholm and L. C. Briand, "Predicting fault-prone components in a Java legacy system," in *ISESE'06 - Proceedings of the 5th ACM-IEEE International Symposium on Empirical Software Engineering*, 2006, vol. 2006, pp. 8–17.

[5] P. He, B. Li, X. Liu, J. Chen, and Y. Ma, "An empirical study on software defect prediction with a simplified metric set," *Information and Software Technology*, vol. 59, pp. 170–190, Mar. 2015.

[6] H. K. Dam, T. Tran, T. Pham, S. W. Ng, J. Grundy, and A. Ghose, "Automatic Feature Learning for Predicting Vulnerable Software Components," *IEEE Transactions on Software Engineering*, vol. 47, no. 1, pp. 67–85, Jan. 2021.

[7] R. Malhotra and M. Khanna, "Inter Project Validation for Change Proneness Prediction using Object Oriented Metrics," *Software engineering : An International Journal*, vol. 3, no. 1, pp. 21–31, 2013.

[8] R. Malhotra and A. J. Bansal, "Cross project change prediction using open source projects," in *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, 2014, pp. 201–207.

[9] Z. He, F. Shu, Y. Yang, M. Li, and Q. Wang, "An investigation on the feasibility of cross-project defect prediction," *Automated Software Engineering*, vol. 19, no. 2, pp. 167–199, Jun. 2012.

[10] B. Ghotra, S. McIntosh, and A. E. Hassan, "Revisiting the impact of classification techniques on the performance of defect prediction models," in *Proceedings - International Conference on Software Engineering*, 2015, vol. 1, pp. 789–800.

[11] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings," in *IEEE Transactions on Software Engineering*, 2008, vol. 34, no. 4, pp. 485–496.

[12] Z. Xu, J. Liu, Z. Yang, G. An, and X. Jia, "The Impact of Feature Selection on Defect Prediction Performance: An Empirical Comparison," in *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, 2016, pp. 309–320.

[13] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, "MAHAKIL: Diversity Based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction," *IEEE Transactions on Software Engineering*, vol. 44, no. 6, pp. 534–550, Jun. 2018.

[14] C. Tantithamthavorn, S. McIntosh, A. E. Hassan, and K. Matsumoto, "An Empirical Comparison of Model Validation Techniques for Defect Prediction Models," *IEEE Transactions on Software Engineering*, vol. 43, no. 1, pp. 1–18, Jan. 2017.

[15] S. Herbold, A. Trautsch, and J. Grabowski, "A Comparative Study to Benchmark Cross-Project Defect Prediction Approaches," *IEEE Transactions on Software Engineering*, vol. 44, no. 9, pp. 811–833, Sep. 2018.

[16] J. Nam, S. J. Pan, and S. Kim, "Transfer defect learning," in *Proceedings - International Conference on Software Engineering*, 2013, pp. 382–391.

[17] X. Xia, D. Lo, S. J. Pan, N. Nagappan, and X. Wang, "HYDRA: Massively compositional model for cross-project defect prediction," *IEEE Transactions on Software Engineering*, vol. 42, no. 10, pp. 977–998, Oct. 2016.

[18] X. Jing, F. Wu, X. Dong, F. Qi, and B. Xu, "Heterogeneous cross-company defect prediction by unified metric representation and CCA-based transfer learning," in *2015 10th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, ESEC/FSE 2015 - Proceedings*, 2015, pp. 496–507.

[19] J. Nam, W. Fu, S. Kim, T. Menzies, and L. Tan, "Heterogeneous Defect Prediction," *IEEE Transactions on Software Engineering*, vol. 44, no. 9, pp. 874–896, Sep. 2018.

[20] F. Rahman, D. Posnett, and P. Devanbu, "Recalling the 'imprecision' of cross-project defect prediction," in *Proceedings of the ACM SIGSOFT 20th International Symposium on the Foundations of Software Engineering, FSE 2012*, 2012, p. 1.

[21] T. Zimmermann, N. Nagappan, H. Gall, E. Giger, and B.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

Murphy, "Cross-project defect prediction: A large scale experiment on data vs. domain vs. process," in *ESEC-FSE'09 - Proceedings of the Joint 12th European Software Engineering Conference and 17th ACM SIGSOFT Symposium on the Foundations of Software Engineering*, 2009, pp. 91–100.

[22] B. Turhan, T. Menzies, A. B. Bener, and J. Di Stefano, "On the relative value of cross-company and within-company data for defect prediction," *Empirical Software Engineering*, vol. 14, no. 5, pp. 540–578, Oct. 2009.

[23] L. C. Briand, W. L. Melo, and J. Wüst, "Assessing the applicability of fault-proneness models across object-oriented software projects," *IEEE Transactions on Software Engineering*, vol. 28, no. 7, pp. 706–720, Jul. 2002.

[24] H. Lu, E. Kocaguneli, and B. Cukic, "Defect prediction between software versions with active learning and dimensionality reduction," in *Proceedings - International Symposium on Software Reliability Engineering, ISSRE*, 2014, pp. 312–322.

[25] L. Madeyski and M. Jureczko, "Which process metrics can significantly improve defect prediction models? An empirical study," *Software Quality Journal*, vol. 23, no. 3, pp. 393–422, Sep. 2015.

[26] M. Jureczko and L. Madeyski, "Towards identifying software project clusters with regard to defect prediction," in *ACM International Conference Proceeding Series*, 2010, p. 1.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:31:38 UTC from IEEE Xplore. Restrictions apply.

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-021-11300-5

Enhancing the wine tasting experience using greedy clustering wine recommender system

Rahul Katarya[1] · **Rajat Saini**[1]

Received: 6 July 2020 / Revised: 8 February 2021 / Accepted: 9 July 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Wine is not just about taste; it represents your class & overall personality. But tasting the same wine for a long time gets boring; sometimes, everyone needs something new, and for that, we usually get advice from friends. But it is not for sure that their taste will match yours and hence you might not like the wine they like. To resolve this issue, we propose the Greedy Clustering Wine Recommender System (GCWRS). This adaptive wine recommendation system uses principal component analysis (PCA) and K-Means clustering algorithms and a novel ranking algorithm using a greedy technique to recommend wines. Along with that, it also uses the elbow method, which helps find out the ideal number of clusters. Together with the clustering algorithm, the recommendation system is tailored in such a way to help all types of users (the first time tasters or regular wine drinkers) as well as the needs of the users (try something new or enhance their current preferences). The clustering model is trained on the wine dataset (with 6497 entries and 12 attributes) combined with the personal preferences of the user, results in an effective personalized recommendation system. The performance of the model is compared with other standard algorithms like K-means clustering, Agglomerative clustering, DBSCAN, Birch clustering algorithm, using performance evaluation functions like the Silhouette coefficient, Calinski-Harabasz Index, and Davies-Bouldin index. Results indicate that our Greedy Clustering Wine Recommender System (GCWRS) provides a better result than other standard algorithms with Silhouette coefficient=0.28, Calinski-Harabasz Index=2207.84, and Davies-Bouldin index=1.36.

Keywords Machine learning · Clustering techniques · Recommender system · Personalization · Wine tasting

✉ Rahul Katarya
rahuldtu@gmail.com

Rajat Saini
sainirajat500@gmail.com

[1] Big Data Analytics and Web Intelligence Laboratory, Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

Published online: 17 September 2021

1 Introduction

The term "wine" was originally from the Latin word "Vinum". Grape juice or must are fermented to form the alcoholic beverage that is wine. Added sugar, additional enzymes, acids, or any other nutrients are not added. Instead, it is formed when grape sugars are consumed by yeast and gives carbon dioxide and ethanol.

The United States, Australia, Chile, Argentina, and Europian countries like France and Italy are the popular wine-producing states. But the love for wine didn't occur overnight; it got a big history behind it. There are some myths of the origin of wine in the prehistoric days of ancient Romans and Greeks. The earliest evidence of winemaking and cultivating grapevines can be dated back to 5400 and 5000 BCE in Iran [38].

Wine is not just for luxury and pleasure; it also acts as medicine. Ancient Mesopotamian civilization endorses wine for treating coughs by mixing it with honey [54]. Ancient Egyptians use wine to cure weakness or help heal wounds [57].

And today, wine tasting is becoming more and more prominent. But the variety of wine in the market is enormous, and every wine tastes different. Some like one kind but do not like another kind. And some regular wine consumers are getting bored having the same kind of wine every time and want to try something new but do not know which one to select.

In this paper, we deal with the problem faced by the people while selecting wine. We introduced a Greedy Clustering Wine Recommender System (GCWRS) that recommends different kinds of wines using the PCA-K-Means clustering algorithm and a novel greedy approach based on recommending technique. Similar kinds of wines are clustered together to form one big cluster. And the wines which are different from one another are placed in different clusters. Similarly, several such clusters are formed using algorithm 2. And then these clusters are used by the greedy technique (running algorithm 3) to provide the user with the best possible suggestion according to their personalized taste.

In this paper, we deal with the problem faced by the people while selecting wine. We introduced a Greedy Clustering Wine Recommender System (GCWRS) that recommends different kinds of wines using the PCA-K-Means clustering algorithm and a novel greedy approach based ranking algorithm. Similar wines are clustered together to form one big cluster using a clustering algorithm. And the wines which are different from one another are placed in different clusters. Similarly, several such clusters are formed using algorithm 2. After which, these clusters are used by the novel ranking algorithm (algorithm 3) to provide the user with the best possible suggestion according to their personalized taste. The model also takes feedback scores from the user after every recommendation. Based on feedback scores from the user, GCWRS gets biased towards providing more personalized suggestions to the user rather than generic recommendations. Hence with the increasing number of users, GCWRS will provide even better and much more personalized suggestions.

1.1 What is a recommender system?

According to [12] "any system that produces individualized recommendations as output or has the effect of guiding the user in a personalized way to interesting or useful objects in a large space of possible options''. Or in simple words, we can say that Recommender Systems[83] are the systems that take a series of input and recommends output. There could be one or more recommendations as output by the system. Some examples of recommendation Systems are as follows:

(A) Movie recommendation system- A movie recommender systems [5, 17, 26, 41, 46] recommend the movie you might like according to the movies you watched in the past. E.g. Netflix, amazon prime uses a similar kind of system, amazon prime
(B) Book recommendation system- A book recommender systems [16, 43, 55, 63, 72, 76] recommend the book you might like to read next, based on the books you read in the past.
(C) Stocks recommendation system- A stocks recommender systems [3, 27, 79] recommend the stocks you should buy or sell to get marginal profits. It usually uses the historical stock dataset and news feeds to give out the best possible results.

There are some new emerging types of recommender systems as well like

(D) E-Commerce Recommender systems- An e-commerce recommender systems [35, 49] help in recommending the customer a highly customized list of objects to choose from (buy).
(E) E-government Recommender systems – An e-government recommender systems [2, 31, 52, 70] help the government takes better decisions for the progress of the state/country.
(F) E-tourism Recommender systems – An e-tourism recommender system [2, 39, 53, 69] helps tourists with customized information for a better experience.

Apart from all these different types of recommender systems, some new researches are emerging on recommender systems like in [59] and [47]. In [59], the author has proposed a recommender system that learns social representations using deep autoencoders. And in [47], the author has proposed a collaborative autoencoder for recommending top N choices using knowledge distillation.

1.2 Who will the proposed recommendation system benefit?

GCWRS will benefit every person who needs a suggestion while trying some wine.

As shown in Fig. 1, GCWRS can be used for all types of drinkers (regular ones as well as a first time drinkers).

(A) For the first time drinker - Recommends the wine they should try for the first time.
(B) For regular drinkers and want to taste something new - Recommends the wine which is completely different than what they currently consume.
(C) For regular drinkers and want to enhance the current preference - Recommends the wine, which is similar to the one they are having but have a better reviews from other users.

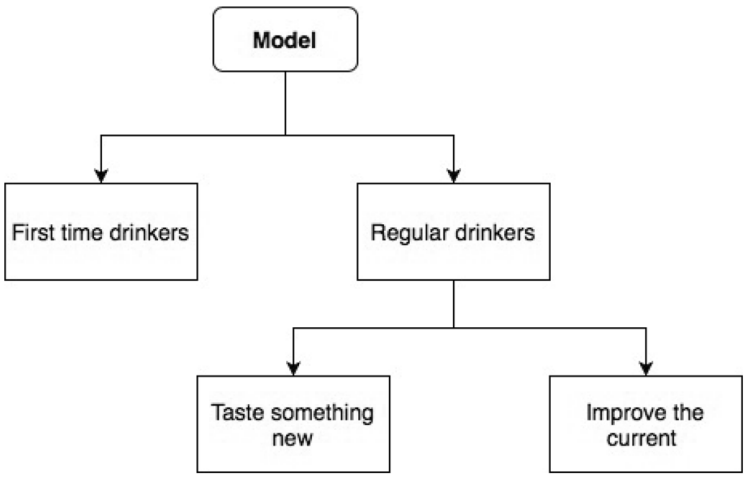


Fig. 1 Options provided by GCWRS to choose from

1.3 Why the clustering model?

While building a recommendation system, we need to choose (or build) the algorithm for the system. There are some recommender systems which is based on multi-agent [13, 77], some use graph-based approach [11, 33], some are based on fuzzy set theory [82], some are based on semi-supervised learning [6, 7], some are using mahout [37, 66], some are using neural networks [14, 74], some are based on Vague sets [81], some are using distributed genetic algorithms [56], some are using hybrid methods [12, 53], some are based on SVM [80], some are using decision trees [50, 51], some are using random forests [61], and some are using regression models [29, 85]. Some latest researches in the Clustering Optimization method has also emerged like [47] and [48] showing some astonishing results. In [48], the author has proposed a multi-objective differential evolution and firework algorithm named MASCC-DE/FWA which is used for automatic simultaneous classification and clustering algorithm. And in [48], the author has proposed a multi-objective optimization approach named MOSSA which is used to self-organize a limited-area sparse span array.

We propose PCA-K-Means clustering techniques to build the clustering model and then use a novel greedy approach based ranking algorithm to recommend different wines. The reason to opt for the clustering techniques instead of any other can be understood by understanding the Figs. 2, 3 and 4.

Figure 2 shows that the dataset is not exactly balanced as there are about 75.4% white wine instances, and only 24.6% instances are red wine (as shown in Fig. 2). As there are more instances of white wine, there is a chance that our model may get biased towards white wine over red wine.

More importantly, Fig. 3. represents that the quality rating given by the wine expert to the wines is mostly average ratings(that means very few wines got very low or very high rating). So if we predict the quality of wines, then the model will be biased, as most of the wine rating is in ranges from 5-7. Effectively the accuracy of the model might be on the higher side, but the practicality of the model will be minimal. To prevent the model from getting biased due to the above-mentioned reasons, we chose the clustering technique.

1.4 About greedy clustering wine recommender system (GCWRC)

The goal of GCWRC is to provide a common platform for all kinds of wine lovers where they can get recommendations according to their needs (explained in Sect. 1.2). Several concepts like Machine Learning, Clustering Techniques, Recommender System,

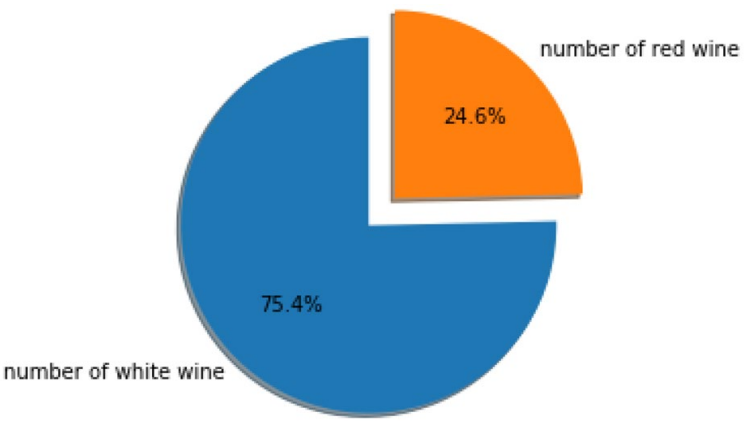


Fig. 2 Pie chart representing the distribution of a dataset between red wine instances and white wine instances

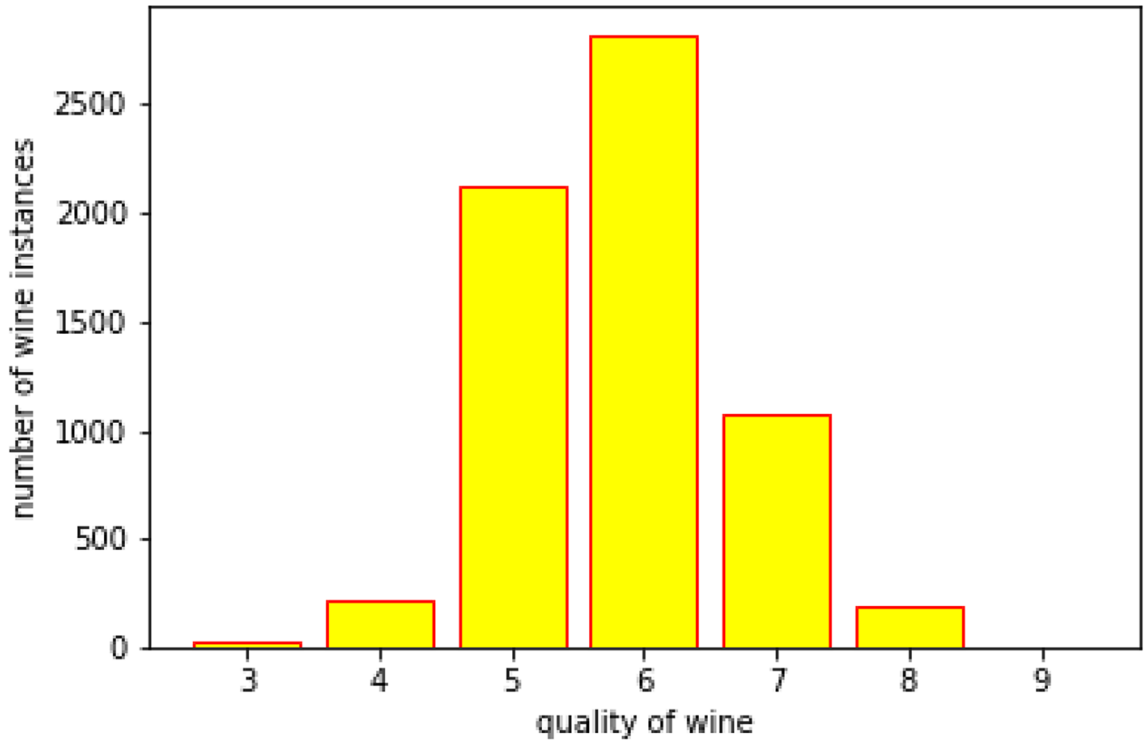


Fig. 3 Bar chart representing the number of wines obtaining the respective quality from the wine expert

Personalization, Wine Tasting are also used in building GCWRC to improve its performance and give the best experience to its users. The major contribution of GCWRC is to skillfully combine some of the best performing algorithms like Principal Component Analysis (PCA), K-Means clustering with the application of the elbow method along with a novel greedy approach based ranking algorithm. By building GCWRC, which uses some of the advanced techniques to recommend the users, we are providing a one-stop platform for all the wine consumers. This will save the time, efforts, and money of the wine consumers by preventing them to randomly try any kind of wine. Instead, it provides them a more personalized system to enhance their wine consuming experience.

The core working of GCWRC uses a clustering algorithm and a greedy approach based recommendation technique. Hence the GCWRC is divided into two parts that are part 1 and part 2. In part 1, clustering models are used which uses the proposed PCA-K-Means clustering algorithm to cluster the dataset. Then in part 2, we explain the novel greedy approach based ranking algorithm which uses the clusters produced in part 1 and provides the user with a list of recommended wines.

1.5 The overall flow of the paper

The overall flow of the paper is as follows:

In the first section, we introduced the uses and some background knowledge about the wine along with the introduction of recommendation systems. We also discussed who can benefit from this GCWRS. Further, we discussed how the clustering technique could be a better approach to the dataset we have. In the second section, we will discuss the related work. In the third section, we will discuss the proposed clustering model architecture (and

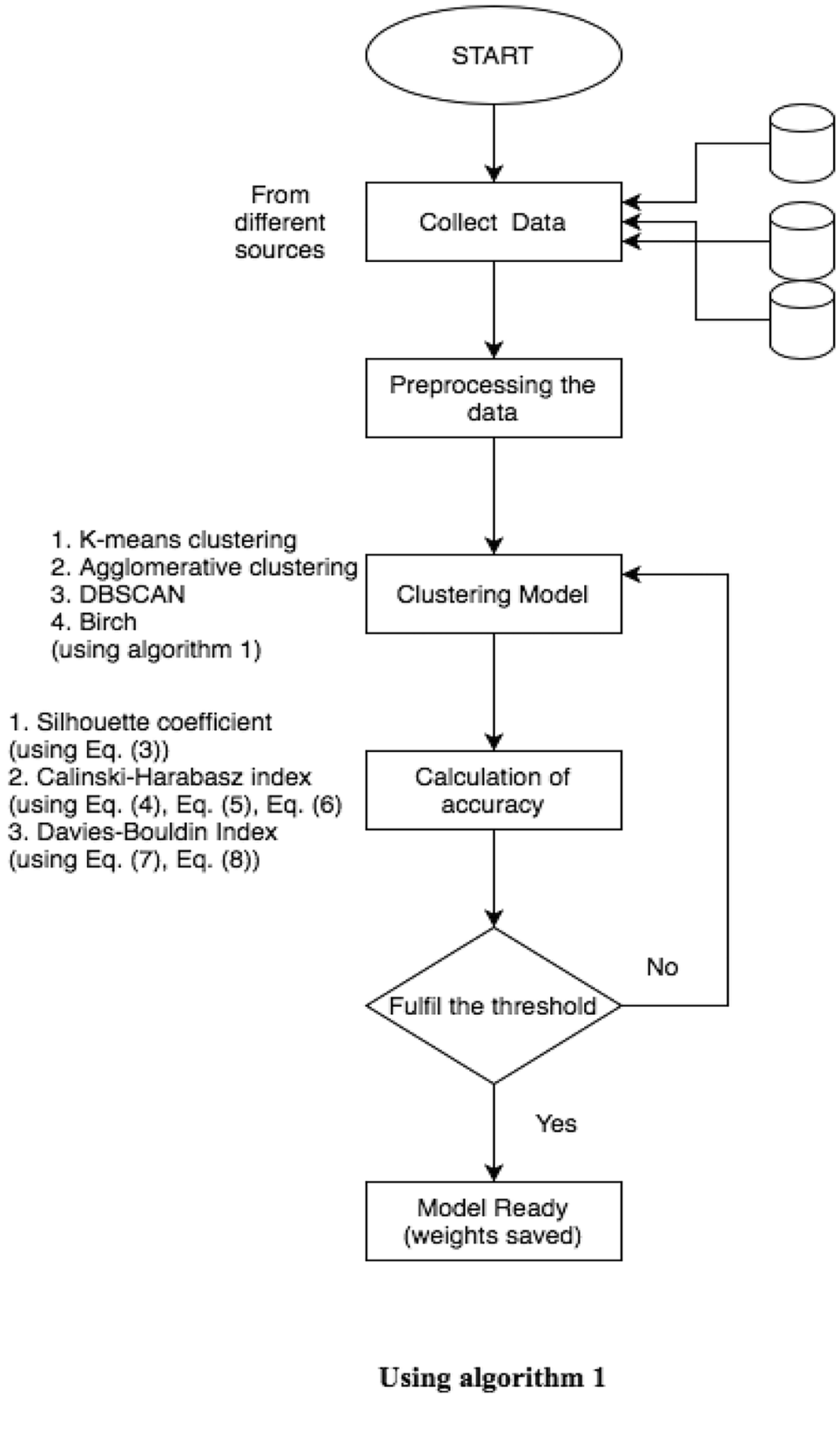


Fig. 4 Flowchart for training clustering model

its algorithm) along with some famous performance evaluation techniques. Next, in the fourth section, we focus on the proposed novel greedy approach based algorithm to provide a personalized list of wine recommendations. In the fifth section, we described the dataset and its attributes, and then the results are evaluated. In the last section, we discussed the future scope and conclusion of the paper. Hence, the fundamental goals of this paper are as follows:

- Wine data-based PCA-K-Means Algorithm proposed for optimal clustering (part 1 of GCWRS).
- A greedy approach based novel recommendation technique proposed (part 2 of GCWRS) to produce the recommendations for the user. Part 1 and part 2 combines to form our GCWRS.
- Performance Evaluation functions used to calculate the optimality of the clustering model.
- Experimental results representing improved performance as compared to the other standard algorithms.

2 Related work

In the present era, with the people getting more inclined towards entertainment and leisure, which created immense importance to the recommender systems. That is why more and more researchers are working on finding the model which gives personalized suggestion to the user.

In [20], the author proposed an approach that uses data mining techniques to predict human wine taste preferences. Models like Support Vector Machines (SVM), multiple regression techniques, and neural network methods were applied and compared. The paper concluded that the results from SVM give the best performance, outperforming the neural network method as well as multiple regression techniques.

In [74], the author proposed a three-layered artificial neural network (ANN) model with back-propagation (BP) to classify wine samples in six different regions. The classification was done using the measured trace amount of elements like Fe, Mn, Al, Cu, Sr, Ba, etc. An optical emission spectrometer (ICP-OES) was used to trace the amount of these elements. The results of the three-layered artificial neural network were compared with cluster analysis, Bayes discrimination method, and Fisher discrimination method. A very acceptable result of 100% accuracy was achieved using the jackknife leave-one-out procedure.

In [78], the author produced a synthetic data from the wine produced in a winery for three years. A neural network was successfully trained to predict the yeast-fermentation kinetics. The author also predicted the chemical and sensory properties of the finished wine. These results were reported based on the properties of the grapes and the intended processing. To accomplish these results, a hybrid neural network method, Stop Traning with Validation (STV) was developed. According to the author, using the general fermentation optimization method, which can be applied to the fermentation process can improve the productivity and quality of wine to a great extend.

In [45], 56 different samples of Italian wine were characterized using several physicochemical parameters (like density, alcohol content). A neural network was used, which took the data from an electronic tongue as the mapping of the parameters with a sensory taste panel was considered a difficult task by the author.

In [58], 54 different samples(among which 30 samples were from Tacoronte-Acentejo DO (class T), and 24 were from Valle de la Orotava DO (class O) wines) were discriminated into two classes(class T and class O). An optical emission spectrometer (ICP-OES) was used to trace the number of elements like Na, Sr, Fe, Al, Ba, Cu, Mn, Zn, Ca, K, and Mg). Methods like Linear discriminant analysis (LDA) and neural network techniques were applied, which results in 90% accuracy and 95%+ accuracy respectively.

In [19], the author applied techniques like support vector machine (SVM), multiple regression approaches, and the neural network to predict the meat preference. And concluded that the SVM outperforms NN and the regression approaches.

In [84], the author used the spectral measurements from 147 different wine bottles to predict three categories of rice wine age.

In [71], the author discusses the different needs of wine consumers along with exploring the difference between wine experts and novices. Wine apps and wine critics' recommendations were also compared. The author also combined different sensory evaluations and recommender systems. Many theoretical aspects of wine are discussed in this paper along with various studies done on wine.

In [44], the author conducted an online survey with Australian wine consumers(n=349) to define "Australian fine wine". Sensory attributes, grape variety, wine region, food pairing, etc were used for the survey. The major focus was to define how consumers have different opinions about different wine. The author concluded with a definition of Australian fine wine: "Australian fine wines are of high quality, good value for money, are easy to drink and consistently show balance, plus diversity, fruity and regional characteristics".

In [65], the author tries to find out which attributes of the wine are more desirable by the consumer and how it is attached to the lifestyle and consumption habits of the consumer. The author used a consumer survey (consists of 1004 entries) using a questionnaire and a rating technique which is a choice experiment to find out the preference (preferred attributes) of the consumer. The author concluded that the geographical site of production is of high importance to the consumer. But the largest consumer segment in the survey has a strong demand for information.

In [30], the author tries to find out the effect of expert ratings on the price of the wine. The rating was done on a scale of 100. The study shows that there is a sharp distinction in the quality and price of the wine if the wine is rated below 90 and the wine rated 90+.

In [34], the author gives an elaborate description of the alcoholic beverages(especially wine). The author also discussed the importance of wine in improving cardiovascular risk factors. The chemistry of wine along with the chemical composition, enological practice, and viticulture are also discussed in the paper. The author briefly talked about the effects of wine components as well.

In [62], the author tries to identify the preferences of the user for different wine attributes with the help of a web-based questionnaire (on 504 different consumers of wine from New York). The paper concluded that a mass marketing approach is no longer suitable, and there is an urgent need to adopt new marketing strategies to target different groups.

From the research work and Table 1, we understand that all the researchers did great work in achieving the specific goal they intended to. But these researches still lack a global goal of providing the consumer with better options and improving their experience with their proposed studies and models. Overall, we have recognized the following common problems of the existing approaches and also stated our contributions to overcome these obstacles.

Table 1 Comparative analysis of related work

Paper	Dataset	Algorithms and techniques used	Description	Advantages	Disadvantages
[21]	UCI machine learning repository dataset	SVM, multiple regression techniques, neural network	Different algorithms like SVM, NN, multiple regression techniques were applied to predict human wine taste preferences.	A better supervised technique to find the quality of a wine. Dataset; large dataset used; better efficiency	Only applicable for supervised data; minimal scope
[74]	ICP-OES measurement data	Clyster analysis, principal component analysis (PCA), discriminant analysis mentor and Neural Network with back-propagation	Different wine samples were classified into six different regions in this paper. And a whopping 100% accuracy was achieved.	High accuracy model; advanced techniques of ICP-OES used	Complex procedure os ICP-OES is followed; minimal scope; smaller dataset used
[78]	Synthesized data	Hybrid neural network, Stop Training, and validation(STV)	Yeast fermentation kinetics was successfully predicted using synthetic data and neural network model.	Improved wine productivity; improved wine quality	Lacks comparative studies; synthesized data used; time-consuming (to collect sufficient data)
[48]	Italian wine data(Synthesized data)	Neural network	Different Italian wine samples were characterized using a neural network model.	Advances tool like electronic tongue used; an advanced technique of neural network used	A very small dataset used; additional hardware needed(electronic tongue)
[59]	Synthesized data	Neural Network and Linear discriminant Analysis(LDA)	Methods like Linear Discriminant analysis(LDA) and neural network were used to classify the wine samples in two different classes namely class O and class T.	Impressive efficiency; Advances technique of ICP-OES(optical emission spectrometer) used	A very small dataset used; minimal scope; Complex procedure os ICP-OES is followed
[20]	Meat data	SVM, multiple regression techniques, neural network	This paper uses a neural network, SVM, and multiple regression approaches to predict meal preference.	Better recommendation system, an impressive comparative study performed	Recommends only meat
[71]	Not mentioned	Wisdom of the crowd(for wine app)	It is a theoretical paper that gives us in-depth knowledge about the wine and wine-related app. It also discusses various recommender systems used for wine and different studies on wine.	Impressive study; wine app comparative study; study about wine recommender systems	Limited application details on how to apply those techniques; Lacking models and algorithm comparative studies
[45]	Australian wine Data(synthesized data collected using Qualtrics)	Analysis techniques like SPSS 25.0 (IBM Corp) and XLSTAT Version 2017 (Addinsoft SARL, France).	The author of this paper tries to find out the preference of win in Australian win consumers. And tries to find the true meaning of “Australian fine wine”.	Impressive outcomes; simplified and defined procedure	A small dataset used; Synthesised data
[65]	Surveyed data(Synthesized data)	Questionnaire, rating technique	Consumer surveys and questionnaires are used to find out the preference of consumers for different attributes of wine.	Impressive results; better productions guidelines provided; very high scope for implementing in the real wine production process	Moderate sized dataset; lacks highly comparative studies with other similar models
[31]	Oregon’s Willamette Valley AVA, rated by the Wine Spectator between 1984 and 2008	empirical analysis using a regression model	This paper explains the price hike and the relationship between the expert rating and the price of the wine.	Splendid study results	Lacks a global objective of recommending
[62]	Surveyed data(Synthesized data)	Web-based questionnaire	The author tries to identify the preference of the consumer for different wine attributes using a web-based questionnaire.	impressive strategies proposed; simplified model used	A small dataset used; dataset is from a local area(New York) hence model would not scale globally; Lacks modern techniques and models

(A) **Minimal Scope:** In the previous studies, mostly the analysis was done to find out the attributes which are preferred by the user or to categorize the wine into different classes. But these studies lack a target audience that is the wine consumers. In some studies, the authors found out the attributes preference of the consumer but still lacks to present the user with the options(of wine) which the consumer will like according to their preferred attributes. To overcome this issue, our GCWRS will benefit a large group of people (Sect. 1.2) with its large scalability. Our system is easy to use as the user does not need to know the architecture of the overall model and just need to input their query to get the results.

(B) **Oversimplified techniques:** In some of the studies, an oversimplified technique of questionnaire and rating technique is used. These studies are somewhat efficient, but there are other major slightly more complex models we can apply to attain better results. To apply a better technique, we propose a combination of Principal Component Analysis (PCA) [1, 75] and K-Means clustering [10, 32, 36] to form the PCA-K-Means algorithm to produce better results than the standard algorithms.

(C) **Size of the dataset:** Most researchers used very small datasets (some are even less than 200 instances), which may provide us with higher accuracy during the training of the model but while testing the model with real-life instances probably will give inaccurate results. On the other hand, our GCWRS was trained on a very large dataset (with approx 6500 instances) as compared to previously researched studies.

(D) **Used traditional supervised learning approach:** Some of the researches use oversimplified techniques of a questionnaire and the remaining ones use traditional supervised learning approaches. Supervised learning techniques use the target attribute (like the quality of the wine), which limits our scope for applying unsupervised approaches. To resolve this issue, our GCWRS uses an unsupervised clustering algorithm which would significantly improve its scalability and the scope of the recommendation system.

(E) **Lacks global objective of recommending something to the consumer:** Most of the researchers focuses on trying to find out the preferred attributes of the consumers to propose better marketing strategies and targeting audience. But most of the researchers lack a global objective of recommending something to the consumer. On the other hand, our GCWRS majorly focuses on the objective to recommend the best possible wines to the consumer to enhance their experience.

3 Clustering study

In this section, we are going to discuss the clustering model and study some standard clustering algorithms. This section is further divided into two parts. The first part of this section describes the architecture of the clustering models which is used in many research work as well. In the second section, we discuss the algorithm (pseudocode) used to train and evaluate the clustering models.

Figure 4. represents the architecture of the clustering model using a flowchart. The flowchart is divided into different steps which are explained in details in Sect. 3.1.

3.1 Architecture

3.1.1 Data collection

In this step, we collect data from various sources. And then the collected data is stored in a standard format. Some of the standard file formats are .xml, .json, .csv, etc. Data in standard form is easy to understand and analyze.

3.1.2 Pre-Processing of data

Our data may contain various anomalies that need to be resolved. For that, we have to preprocess data before applying any machine learning algorithms. Some of the preprocessing steps are as follows:

(A) **Check for missing values:** There can be missing attributes/values in the dataset. Missing values results in problems that training the model. Hence, we need to remove the missing values from the dataset. There are various ways to do that:

- Remove the whole instance - We could directly remove that instance from our data if even one of its attributes is missing. But if the data is smaller in size, then we need to apply some other technique.
- Replace with an average value- One another way is to replace the null or empty value with the average of corresponding data value.
- Manual update- Another way is to manually update the missing value. But in most cases, it is not practical because of the large size of the dataset.

(B) **Standardize the data:** There is another possibility that the data is not in standard units (SI) or is not tidy. Then we have to process the dataset transform it into standard form.

(C) **Normalize the dataset:** The attribute of the dataset contains different information about the dataset. Some of the attributes have a higher range (e.g. the highest value attribute is 2200, and the lowest value is 1000, then the range will be 1200) and there are some attributes with comparatively lower range (e.g. the highest value attribute is 22 and the lowest value is 10, then the range will be 12). This high variation in the range of the attributes results in the model getting biased towards some attributes because of their higher values. To prevent this, normalization is done. Normalization is a process of converting all the attributes values in a range between 0 and 1 without distorting differences in the ranges of values. Eqs. (1) and (2) are used to normalize the dataset.

$$X(new) = (X - X.min)/(X.max - X.min) \tag{1}$$

$$X(scaled) = X(new) * (max - min) + min \tag{2}$$

Where min is the minimum value of the attribute and max is the maximum value of the attribute.

3.1.3 Clustering algorithms

Clustering is a process in which we cluster similar points (or data) together to form a cluster. Data points in a particular cluster are more similar than the points from a different cluster. A clustering algorithm is generally used with an unsupervised dataset. Some of the standard clustering algorithms are as follows:

Table 2 represents some of the standard clustering algorithms. Table 2 also represents the parameters and scalability of all the standard clustering algorithms.

3.2 K-means clustering algorithm

K-means clustering algorithm [10, 32, 36] is also called Lloyd's algorithm. It is the most famous clustering technique which is used in various researches as well. In this algorithm, we first randomly initialize k (number of clusters) different number of centroids in the space. For each data point 'A' in the dataset, we calculate the distance (manhattan distance [21], Euclidean distance [22, 64], or the Mahalanobis distance [23]) of that point 'A' from all the centroids. Then the data point is assigned to the centroid, which is nearest to that point. After that, the mean of each data point in the cluster is calculated and the centroid is shifted towards the new mean of the cluster. The deviation of the centroid (from its old coordinates to the new coordinates) is calculated. If the calculated deviation is less than the threshold, then the process is stopped. Otherwise, again all the data points are assigned to their new nearest cluster centroid.

Elbow method K-means clustering is a very efficient yet straight forwards and simple algorithm. The parameter we need to choose for our algorithm in k-means clustering is the value of k(number of clusters). But when the data is high dimensional (as in the case of our wine dataset), then providing the ideal value of the number of clusters for our dataset is a very difficult task. That is where the Elbow Method [8] becomes handy.

Elbow Method is used to help the programmer to choose the optimal number of clusters for our dataset by trying to fit the model with various values of k (number of clusters). After fitting the model, the error of the model is calculated and then visualization is done by plotting a line graph. The line graph will bear a resemblance to an arm, and then we

Table 2 Clustering algorithms and their respective parameters and scalability

S.no.	Name of the Method	Parameters	Scalability	
			For the number of samples	*For the number of clusters*
1	**K-means clustering**	Number of clusters	Very large	Medium
2	**Agglomerative clustering**	1. The number of clusters 2. Distance threshold 3. Types of linkage 4. Distance	Large	Large
3	**DBSCAN**	Size of the Neighbourhood	Very large	Medium
4	**Birch**	1. Branching factor 2. Threshold 3. Global cluster(optional)	Large	Large

need to find the "elbow" (the point of inflection on the curve) which will represent the ideal number of the cluster for our model.

As in Fig. 5, the line graph represents the number of clusters with their respective distortion. The graph resembles somewhat like an "arm" of a human; now we need to find the "elbow" to get the ideal value of k (number of clusters). According to Fig. 5, the distortion of the model is significantly high for K=1, but it drops drastically with K=2, and after that, it drops smoothly from 3 till 24. Analyzing the line graph in Fig. 5, we can conclude that the “elbow” of this graph is 3 or 4. Hence according to Elbow Method, the ideal number of clusters for our dataset is 3 or 4.

3.3 Agglomerative clustering algorithm

It is the most known kind of hierarchical clustering [18]. It is also used in various researches. It follows the algorithm in which in the beginning, all the data points are considered as singleton clusters. Then, step by step couple of clusters (similar ones) are merged one after the other until all the clusters get merged into one big cluster consisting of all objects. It looks like a tree where smaller clusters combine to form a bigger cluster and are named Dendrogram. It is a “bottom-up” approach as each object is initially an individual element (at the node) and then similar clusters combine to form a bigger cluster. If we take the exact opposite, starting with the root and divide the cluster into two sequentially, then this is known as Divisive Analysis (DIANA) which follows a “top-down” approach.

3.4 DBSCAN clustering algorithm

The algorithm we described in Sect. 3.1.3 and Sects. 3.2, 3.3 that is K-means clustering and agglomerative clustering (which is hierarchical clustering method) works well only with convex clusters or spherical-shaped clusters and is majorly affected by the outliers and noise in the data.

So to overcome these drawbacks DBSCAN [25, 68] is used. In DBSCAN for each point of a cluster, the points present in the neighborhood with a given radius must be equal to or more than the minimum number of points.

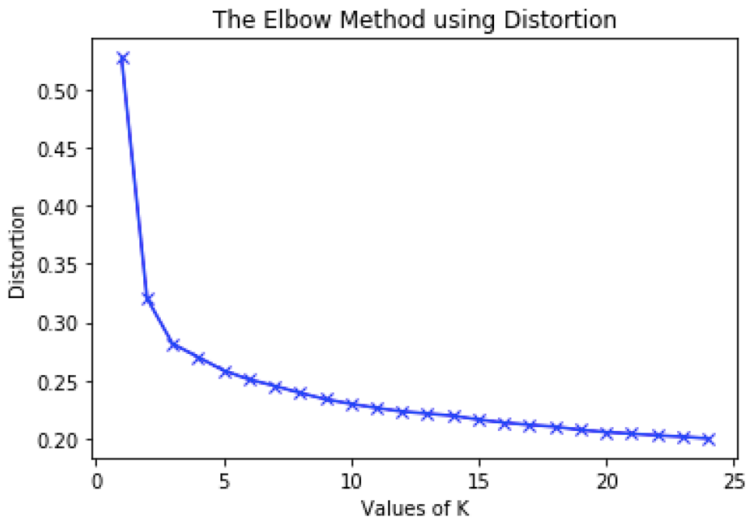


Fig. 5 Line graph for the elbow method with various values of k and the corresponding distortion

The parameters of the DBSCAN Clustering algorithm are as follows:

i. eps: It is used to define the radius for the neighborhood around a given data point. So, if two points are having distance less than or equal to the eps, then these points will be considered as neighbors.
ii. MinPts: It is the minimum number of data points within the eps radius. So we have to choose the MinPts very carefully as if the dataset consist of a very small number of data points then choosing a smaller value of MinPts is more favorable.

3.5 Birch clustering algorithm

When encountering machine learning/data science real-life problems, then we need to have a sufficient amount of data to build a model. But sometimes this required dataset is enormous and needs a huge amount of processing. If the resources are limited then it gets difficult to work with. And if the dataset increases even further then the accuracy of the results and the running time our model provides will get even worse. So there is a scalability issue with such huge datasets.

To overcome this issue, we can use BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) [73, 86], which can work on large datasets by doing lossy compression. The original dataset is converted to a summarized version (set of clustering feature nodes known as CF Nodes) of the original data which retains as much information as possible. Then the summarized data is clustered. Some of the methodologies of Birch clustering algorithms are as follows: Clustering Feature (CF): BIRCH algorithm can work on very large datasets as the original dataset is compressed and the information is summarized in dense regions as Clustering feature entries. CF-Tree: It a closed packed representation of the original dataset. Every entry in the leaf node is not just a single data point but instead, it is a sub-cluster.

3.5.1 Calculation of accuracy

In this step, we analyze how good our model is trained using the given machine learning model. Calculating the performance of a clustering model is not always that straight forward. There are some datasets in which the ground truth (actual labels of the instance) are available and there are some datasets in which these ground truths are absent. Similarity and Dis-similarities are used to calculate the accuracy of the model.

Suppose data point D1 and D2 belong to cluster C1. And data point D3 belongs to a cluster C2. Then:-

Similarity Data point D1 and D2 should be very similar to each other (the more similar they are, the better the clustering model performs).

Dis-similarity Data point D1 and D3 should be different from each other (the more different they are, the better the clustering model performs).

There are various ways to find out the performance of the clustering model as given in Fig. 6.

If the dataset contains the ground truth If the dataset contains the ground truth, then we can use functions like Adjusted Rand Index [42], Mutual Information based score [67], and Fowlkes-Mallows scores.

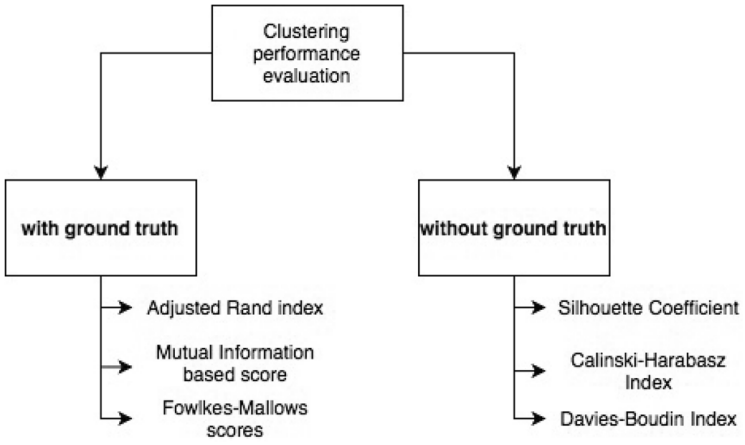


Fig. 6 Different ways to evaluate the performance of the clustering model

If the dataset does not contain the ground truth If the dataset does not contain the ground truths, then we can use the functions like Silhouette Coefficient, Calinski-Harabasz Index, and Davies-Bouldin Index.

Figure 6 represents a descriptive chart to show the evaluation function to be used. According to Fig. 6, if the dataset does contain the ground truth then they can be evaluated using functions like Adjusted Rand index, Mutual Information-based score, and the Fowlkes-Mallows scores. And if the dataset does not contain the ground truth, then they can be evaluated using functions like the Silhouette coefficient, Calinski-Harabasz, and Davies-Bouldin Index.

Our wine dataset does not contain the ground truth so; we are going to use the following functions to evaluate the performance of our model:

3.6 Silhouette coefficient

The Silhouette coefficient [15] technique is used when the truth labels are unknown. First the silhouette coefficient 's' is calculated using Eq. (1). The value of s can be in a range from -1 to +1. The value of s approaching -1 is considered incorrect clustering, the value of s in the neighborhood of 0 is considered overlapping clusters, and the value of s approaching +1 is considered highly dense clusters.

The equation used for the silhouette coefficient 's' is:

$$s = \frac{y - x}{max(x, y)} \tag{3}$$

Where x is the mean distance between the sample data point and all other points present in the same class, y is the mean distance between the sample and all other points in the next nearest cluster.

The major drawback of the silhouette coefficient is that it gives a higher value for convex clusters.

3.7 Calinski-Harabasz index

Calinski-Harabasz Index [40] technique is also used when the ground truth labels are unknown. It is also known as the Variance Ratio Criterion. If the Calinski-Harabasz score is high that refers to a model with more defined clusters (dense and well separated).

The equations used for Calinski-Harabasz score 's' are:

$$s = \frac{tr(B_k)}{tr(W_k)} \times \frac{n_E - k}{k - 1} \tag{4}$$

$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T \tag{5}$$

$$B_k = \sum_{q=1}^{k} n_q (c_q - c_E)(c_q - c_E)^T \tag{6}$$

where $tr(A)$ is the trace of matrix A, k is the number of clusters, E is the set of data, q is the center of the cluster, C_q are the points inside cluster q, c_E is the center of dataset E, ne is the size, n_q is the number of data points inside cluster q, (B_k) is group dispersion matrix, (W_k) is within-cluster dispersion matrix.

The major drawback of the Calinski-Harabasz Index is that it gives a higher value for convex clusters.

3.8 Davies- Bouldin Index

Davies- Bouldin Index [9] technique is also used when the ground truth labels are unknown. It evaluates the efficiency by which the separation between the clusters is done. First, the Davies-Bouldin Index DB is calculated using Eq. (6) with the help of Eq. (5). The value of DB is 0 or above 0. Values of DB close to 0 is considered better (that means better partition of the clusters). So the lower the value of DB, the better the clustering is done. The equations used for Davies-Bouldin Index 'DB' are:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}} \tag{7}$$

$$DB = \frac{1}{k} \sum_{i=1}^{k} \underset{i \neq j}{max} R_{ij} \tag{8}$$

Where s_i is the average distance between the respective point and the cluster centroid, d_{ij} is the distance between centroids of cluster i and j.

Although the Davies-Bouldin Index is faster than some of the very famous evaluation functions like Silhouette score, there are some major drawbacks as well:-

- Davies-Bouldin Index gives a higher value for convex clusters.
- As we are using centroid distance, hence it constraint us with the distance metric to Euclidean space.

3.8.1 Fulfilling the threshold

In Sect. 3.5.1, we calculated the accuracy of the model. To have a recommendation system or any machine learning model to work in a real-life problem, the accuracy of the model

must be above certain standards. To fulfill this task we need a threshold, the proposed model should give accuracy above a certain threshold to have a useful and good predicting/ recommendation model.

3.8.2 Saving the model

Once our model achieves a certain threshold then the model is ready to be used as a recommendation system. The model is simply a mathematical equation with certain weights. These weights are adjusted or updated while training the model (using the clustering algorithms discussed in Sect. 3.1.3). And once the model's accuracy is above a certain threshold then we save the weights of the equation. Now we can use these weights to predict the output instead of training the model again. Now, the model is ready for the prediction and accepts new queries. Using the new queries and the saved weights, the model can predict the output.

3.9 Algorithm for building clustering model

In this part of the section, we discuss the algorithm for building a clustering model using clustering algorithms.

Algorithm 1 Pseudocode for clustering model using 4 different clustering algorithms

Input: wine dataset (containing 12 attributes)
Output: K different clusters of wine
Steps:
1. begin
2. for ∀ instances in dataset:
Do:
2.1 If the value of any of its attribute is missing, then delete the instance.
2.2 If the value of the attribute is not within a certain range, then remove the instance.
2.3 Normalise the dataset
3. end for
4. For the complete dataset:
4.1 train model type-1 using the K-means clustering algorithm by the following steps:
4.1.1 Plot graph using elbow method for various values of K(that is for K=3, K=4, K=5, K=6, K=7, and K=8, where 'K' is the number of clusters)
4.1.2 Find out the "elbow" in the elbow method graph(Fig. 5). The point at which the graph forms an "elbow" will be an optimum value of K.
4.1.3 Train model type-1 using the optimum value of K evaluated using the elbow method graph.
4.2 train model type-2 using Agglomerative clustering algorithm by the following steps:
4.2.1 Select the value of the parameter, that is the number of clusters(k=3, k=4, k=5, k=6, k=7, and k=8).
4.2.2 Train different models for different values of k using Agglomerative clustering algorithm.
4.3 train model type-3 using DBSCAN algorithm by the following steps:
4.3.1 Select the value of the parameter, that is the value of eps(eps=0.30, eps=0.40, eps=0.50, eps=0.60, eps=0.70, eps=0.80, eps=0.90, **and** eps=1.0)
4.3.2 Train different models for different values of eps using DBSCAN clustering algorithm.
4.4 train model type-4 using Birch algorithm by the following steps:
4.4.1 Select the value of the parameter, that is the number of clusters(k=3, k=4, k=5, k=6, k=7, and k=8).
4.4.2 Train different models for different values of k using Birch clustering algorithm.
5. Calculate the efficiency of model type-1, model type-2, model type-3, and model type 4 using:
5.1 Silhouette coefficient: using Eq. (3), calculate the Silhouette coefficient for all the models.
5.2 Calinski-Harabasz Index: using Eq. (4), Eq. (5), and Eq. (6), calculate the Calinski-Harabasz Index for all the models.
5.3 Davies-Bouldin Index: using Eq. (7) and Eq(8), calculate the Davies-Bouldin Index for all the models.
6. Compare the efficiency of all the model and select the best model.
7. Save the model
8. end

Algorithm 1 describes the overall clustering model building process. The input of the model would be the wine dataset which contains 12 attributes. The output of this model would be K different clusters of wine where each cluster contains a different type of

wine. And each cluster is significantly different from another cluster. There are 5 different parts of algorithm 1:

- The first step is the input step, in which we input the wine dataset.
- The second step (that is from 2 to 3 in Algorithm 1), is a pre-processing step. In this step, all the missing values instances are deleted. We can also remove the outliers and normalize the dataset.
- In the next step (that is 5 and 6 in Algorithm 1), we are building model type-1, model type-2, model type-3, and model type-4 using the K-means clustering algorithm, Agglomerative clustering algorithm, DBSCAN algorithm, and the Birch clustering algorithm respectively.
- In the following step, all the models (that are model type-1, model type-2, model type-3, and model type-4) are evaluated using the Silhouette coefficient, Calinski-Harabasz Index and Davies-Bouldin Index to find out the best model among all of them.
- In the last step, that is the fifth step, the weights of the best performing model are saved for predicting outputs.

4 Proposed approach

In this section, we explain the proposed PCA-K-Means clustering algorithm and a novel greedy approach based recommending technique which is used to build our GCWRS. The complete architecture of GCWRS is as shown in Fig. 7. GCWRS is further divided into two parts. The first part is used to generate a clustering model using the proposed PCA-K-Means clustering algorithm. The overall flow chart of the first part of GCWRS remains the same as shown in Fig. 4, except the clustering algorithm used, is the new proposed PCA-K-Means clustering in GCWRS. And the second part of our GCWRS consists of the novel greedy approach based recommending technique to recommend the best wines to the consumer.

Figure 7 represents a complete flowchart of GCWRS. The flowchart is divided into two parts that are the clustering model (part 1 in Fig. 7) and the greedy approach-based recommendation technique (part 2 in Fig. 7). The first part(that is the clustering model) uses the proposed PCA-K-Means clustering algorithm (algorithm 2) and the second part uses a proposed novel greedy approach based on recommending algorithm(algorithm3) which recommends the best wine to the user. Combining part1 and part2 results in our GCWRS.

4.1 PCA-K-means clustering algorithm

PCA-K-Means algorithm is a combination of Principal Component Analysis (PCA) and K-Means algorithm. We have already discussed the K-Means algorithm in Sects. 3.1.3 and 3.2, and the Principal Component Analysis (PCA) is explained next sub-section (in Sects. 4.1 and 4.2). Then in the next sub-section (in Sects. 4.1 and 4.3) we will discuss the pseudocode for the PCA-K-Means clustering algorithm.

4.2 Principal component analysis (PCA)

Principal Component Analysis [1, 75] is a useful and very intuitive technique which is technique to transform the coordinates of the dataset. Suppose our data is in a 2-dimensional plane, then there will be originally two axes that are X-axis and Y-axis.

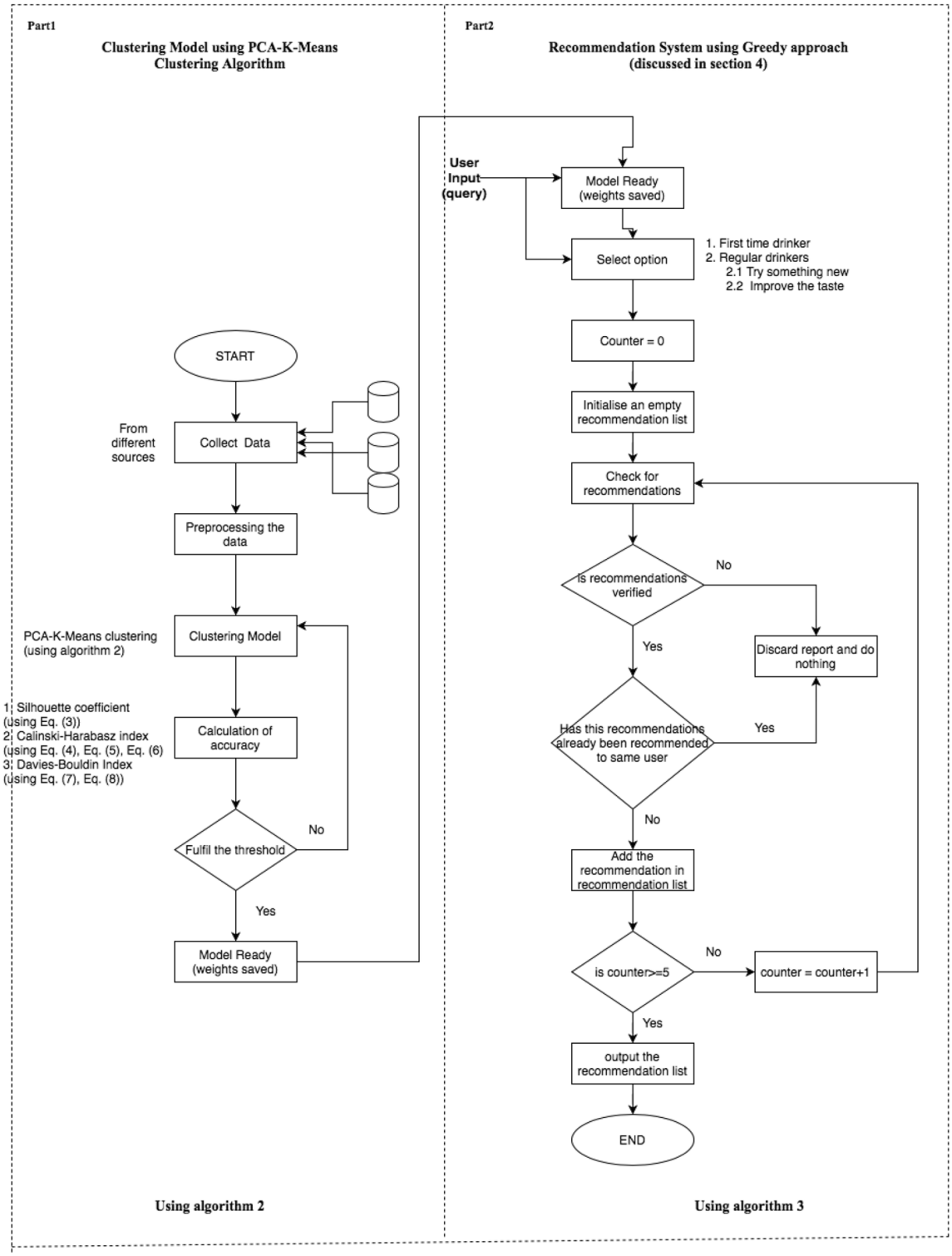


Fig. 7 Flow chart of Greedy Wine Clustering Recommender System (GCWRS)

And all the data points can be represented using these two axes. If we turn those two axes and then plot all those points according to these new coordinates, then this process is called coordinate transformation. PCA uses a similar approach. To select this new set of coordinates, PCA uses the eigenvectors [24, 28] of the dataset. And then the dataset can be plotted into the newly selected coordinate system (where the axes are the eigenvectors of the dataset).

PCA can also help to reduce the dimension of the dataset. So to reduce the dimension of the dataset we follow the following steps:

i. Find out the eigenvectors and corresponding eigenvalues of the complete dataset.
ii. Pair the eigenvectors with their corresponding eigenvalues.
iii. Sort all the pairs concerning the eigenvalues of the pair in descending order.
iv. Choose 'N' that is the new dimension of the dataset (which should be smaller than the original dimension of the dataset).
v. Now select the top 'N' pair of eigenvectors and eigenvalues.
vi. A New N-dimensional coordinate system will be generated where the top N selected eigenvectors will act as new axes.
vii. Transform the complete dataset into the new reduced dimensional coordinates.

So the ideology behind the PCA dimension reduction is that we select the component with the most amount of variance, as it is assumed that the component having higher variance contains more information and can help separate different types of datapoint in the dataset. As a result, PCA not only reduces the dimension of our dataset, but it also tries to retain maximum information while reducing the dimension of the dataset.

With the reduced dimension of the dataset, the model training time significantly improves, improves visualization, and reduces the possibility of overfitting. But the major drawback of PCA is some information is lost during the dimension reduction process.

4.3 PCA-K-means clustering pseudocode

PCA-K-Means clustering algorithm is a combination of PCA and K-Means algorithm. PCA is explained in Sects. 4.1 and 4.2 and K-Means algorithm is explained in Sects. 3.1A. The pseudocode for our PCA-K-Means Clustering is explained in Algorithm 2.

Algorithm 2 Pseudocode for PCA-K-Means algorithms

Input: wine dataset (containing 12 attributes)
Output: K different clusters of wine (trained with the PCA-K-Means clustering algorithm with the reduced dimensional dataset)
Steps:
1. begin
2. for ∀ instances in the dataset:
Do:
2.1 If the value of any of its attribute is missing, then delete the instance.
2.2 If the value of the attribute is not within a certain range, then remove the instance.
2.3 Normalise the dataset
3. end for
4. for ∀ **attributes** in the dataset:
Do:
4.1 Calculate the eigenvalues and eigenvectors of the dataset.
4.2 Pair the eigenvalues with their corresponding eigenvectors and insert them in a list.
4.3 Sort the list with respect to the eigenvalues in descending order.
5. end for
6. Select 'n' that is the number of eigenvectors we want to retain.
7. Now, the selected number of eigenvectors will become the new coordinate system of the dataset.
8. **Transform the data** into the selected number(n) of eigenvectors. Hence, the dimension of the dataset will reduce to n.
9. **Apply the K-Means algorithm** on the transformed data (with reduced dimension) for various values of K(that is for K=3, K=4, K=5, K=6, K=7, and K=8, where 'K' is the number of clusters)
10. Calculate the efficiency of the model using:
10.1 Silhouette coefficient: using Eq. (3), calculate the Silhouette coefficient for all the models.
10.2 Calinski-Harabasz Index: using Eq. (4), Eq. (5), and Eq. (6), calculate the Calinski-Harabasz Index for all the models.
10.3 Davies-Bouldin Index: using Eq. (7) and Eq(8), calculate the Davies-Bouldin Index for all the models.
11. Compare the efficiency of the model for various values of the parameter and select the ideal value of the parameter.
12. **Save the model**
13. end

Algorithm 2 explains the steps to be followed to train a clustering model using the PCA-K-Means algorithm. The input of the model is the complete dataset and the output of the algorithm 2 is the model trained with PCA-K-Means algorithm with K number of clusters. First, the preprocessing of data is done (as shown in step 2 in Algorithm 2). After that, the Principal Component Analysis (PCA) technique is applied to reduce the dimension of the dataset (from step 4 to step 8 in Algorithm 2). Then, K-Means clustering is applied to the reduced dimensional dataset. Next, different performance evaluation functions are applied to the model. The ideal value of the parameter is evaluated using results from performance evaluation functions. The weight of the best performing model is saved.

4.4 Greedy approach based recommendation technique

This part of the recommendation system provides the actual list of recommendations to the user. The model saved using algorithm 2 is used along with the required inputs (some of which are optional) to provide a personalized list of recommendations. Pseudocode for the greedy approach [4] based recommendation technique is shown in algorithm 3.

There are two inputs for the algorithm. The first (optional) input is the preferred wine of the user. For the second input, the user has to select an option (as discussed in Sect. 1.2). The output of the algorithm is a list of wine recommended to the user (sorted from most preferred to the least preferred wine). In step 2 of algorithm 3, we load the saved model (which was built using algorithm 2). Then we initialize a counter "i" to count the number of recommendations. Also, we initialize an empty recommendation list, which is used to store the recommendations provided by the recommender system.

If the user wants to try a new type of wine, then step 5.1 (of algorithm 3) will follow. Otherwise (if the user wants to enhance the existing preference), step 5.2 will follows.

In step 5.1 (of algorithm 3), if data for existing preferred wine is not provided by the user then the data is sorted for the quality attribute. And the best wines of different clusters are added on the recommendation list using step 5.1.1.3 (of algorithm 3). Then the counter variable is incremented until we get 5 different recommendations in the recommendation list (or the value of the variable counter is greater than or equal to five).

In step 5.1 (of algorithm 3), if data for existing preferred wine is provided by the user then also the data is sorted for the quality attribute. And the best wines of different clusters (excluding the cluster from which the existing preferred wine belongs to) are added on the recommendation list using step 5.1.2.3 (of algorithm 3). Then the counter variable is incremented until we get 5 different recommendations in the recommendation list (or the value of the variable counter is greater than or equal to five).

If the user chooses to enhance the existing preference then step 5.2 (of algorithm 3) will follow. Suppose the existing wine belongs to a cluster "A". Then we sort all the instances of the cluster "A" for the quality attribute. And then, we select the best 5 wine from the sorted list other than the existing preferred wine (as shown in step 5.2.3 of algorithm 3).

Then in the next step, the output list of recommendations is provided to the user.

4.5 Why divide the complete recommendation system into two parts?

We have divided our recommendation system into two parts. The first part is used to build a clustering model using the PCA-K-Means clustering algorithm. And the second part is

Algorithm 3 Pseudocode for customized wine recommendation system.

Input: 1. Previously preferred wine of the user(optional)
2. Select either option:

2.1 Try something new

2.2 Enhance the existing preference

Output: List of recommended wine ordered from most recommended to less recommended

Steps:

1. begin

2. Load the model built using **algorithm 2.**

3. Initialize counter $i = 0$.

4. Initialize an empty **recommendation list.**

5. Request the user to **select an option:**

5.1 Try something new: If the user selects the option to try a new type of wine then follows the following steps:

5.1.1. If the user has not provided previously preferred wine **then**

5.1.1.1 Pick a cluster from the saved model(using algorithm 2), select the instances of wine with the highest value of the "quality" attribute(greedy approach).

5.1.1.2. Sort the dataset with respect to "quality" attributes.

5.1.1.3. Select the instance io present on the top(with the highest quality).

If the cluster from which io belongs to is Ci and an instance from that cluster Ci is already in the recommendation list **then**

delete the instance io.

else

add the instance io in the recommendation list.

counter $i = i + 1$.

5.1.1.4. **if** $counter >= 5$ **then**

go to step 6

else

go to step 5.1.1.1

5.1.2. else

(Now we have the existing preferred wine of the user Lo**)**

5.1.2.1 Pick a cluster from the saved model(clustering model built using algorithm 2).

5.1.2.2. Sort the dataset with respect to quality attribute.

5.1.2.3. Select the instance io present on the top(with highest quality) (greedy approach).

If the cluster from which io belongs to is Ci and an instance from that cluster Ci is already in the recommendation list or Lo belongs to cluster Ci **then**

delete the instance io.

else

add the instance io in the recommendation list.

counter $i = i + 1$.

5.1.2.3. **if** $counter >= 5$ **then**

go to step 6

else

go to step 5.1.2.1

5.2 Enhance the existing preference.

(We have the existing preferred wine of the user Lo)

5.2.1. Let Lo belongs to the cluster Cj.

5.2.2. Sort the instances of cluster Cj with respect to the quality of the wine.

5.2.3. Select the top 5 wine from sorted list (greedy approach).

5.2.3.1 if sorted list does not contain Lo **then**

add all top 5 wine in recommendation list
go to step 6

5.2.3.2 else

remove the instance Lo and **then**

add the new top 5 wine in the recommendation list

go to step 6

6. Output the recommendation list to the user

7. end

using a greedy based approach technique that takes input query by the user and recommends the best wines using a greedy approach. The following are the advantages of dividing the system into two parts:

- **Simplification:** The complete model gets more simplified. As the clustering model is in one section and the recommended technique for the model is in another section, it makes things simpler to understand and implement.
- **Improve efficiency:** Separately the efficiency of both parts of the model can be enhanced as it would be easier to diagnose the lower performing aspects of the system.
- **Further Research:** As there are two different parts of the recommendation system, further research can be done on either of them to come up with a new recommendation system.

5 Experiment and result

5.1 About the dataset

The dataset used is the Wine Quality dataset [60]. There is a total of 6497 entries out of which 1599 instances are of red wine and 4898 instances of white wine. There are a total of 12 attributes that includes 1 output attribute which outputs the quality of the wine. All the input attributes (10 in total) are from objective tests and the Output attribute that is the quality of the wine is based on sensory data. Wine experts gave output about the quality of wine on the scale from 0(very bad) to 10(excellent) and the median is taken from at least 3 evaluations. Most of the attributes of the dataset are containing the chemical composition of the wine. And even a small change in chemical composition can result in a major change in the taste of the wine. A detailed description of all the attributes is in Table 3.

In Table 3, we described each attribute of our wine dataset along with the units in which the attribute is stored.

5.2 Results

Using algorithm 1, four different models were trained using k-means clustering, agglomerative clustering, DBSCAN, and Birch clustering algorithm. In this section, we applied three different performance evaluation functions (Silhouette score, Calinski-Harabasz Index, Davies Boudin Index). The performance for different values of parameters of the clustering model is also performed.

- Parameter for K-means clustering algorithm -- the value of k(that is the number of clusters).
- Parameter for agglomerative clustering algorithm -- the number of clusters.
- Parameter for DBSCAN clustering algorithm -- eps.
- Parameter for Birch clustering algorithm -- the number of clusters.
- Parameters for proposed PCA-K-Means algorithm — the number of eigenvectors and the number of clusters

Table 3 Description of the attributes of the dataset

S.no.	Input variable	Unit	Description
A	**Fixed acidity**	g/dm^3	It is the amount of tartaric acid which does not readily evaporate.
B	**Volatile acidity**	g/dm^3	It is the amount of acetic acid. It should not be way too high or else the wine would taste like vinegar.
C	**Citric acid**	g/dm^3	It is the amount of citric acid in the wine, which gives freshness to the wine and adds a flavor to the wine.
D	**Residual sugar**	g/dm^3	It is the quantity of sugar left after the fermentation process stops. Usually, it ranges between 1gram/liter and 45grams/liter.
E	**Chlorides**	g/dm^3	It is the amount of sodium chloride(salt) in the wine.
F	**Free sulphur dioxide**	mg/dm^3	There is an equilibrium between the free sulfur dioxide form and molecular sulfur dioxide. It also helps in preventing the growth of micro-bacterias and prevents the oxidation of the wine.
G	**Total sulphur dioxide**	mg/dm^3	It is the total amount of sulfur dioxide in the wine. If the concentration of total sulfur dioxide is high, then it can be sensed from the nose as well as the taste of the wine.
H	**Density**	g/cm^3	Usually, the density is very close to the density of water but it depends on the alcohol and sugar content in the wine.
I	**pH**	pH value	It ranges between 0(very acidic) and 14(very basic).
J	**Sulphates**	g/dm^3	It is the amount of potassium sulfate in the wine.
K	**Alcohol**	% by volume	It is the amount of percentage of alcohol content present in the wine.
L	**Quality**	Between 0-10	It is an attribute that explains the quality of the drink.

The results of the first four algorithms are tabulated in Tables 4 and 5.

In Table 4, we tabulated the results of algorithms like K-means clustering and agglomerative clustering. Silhouette score, Calinski-Harabasz Index, and Davies Bouldin Index are used to evaluate the performance of the clustering algorithm for various parameters. The parameter for K-means clustering is the number of clusters (K). And the parameter for the agglomerative clustering is also the number of clusters (k). The results are calculated for parametric value k=3,k=4,k=5,k=6,k=7, and k=8.

In Table 5, we tabulated the results of algorithms like DBSCAN and Birch clustering. Silhouette score, Calinski-Harabasz Index, and Davies Bouldin Index are used

Table 4 Results after applying various performance evaluation functions on k-means clustering and agglomerative clustering algorithm

Accuracy model	Name of the algorithms											
	K-means clustering						*Agglomerative clustering*					
Parameter	*k=3*	*k=4*	*k=5*	*k=6*	*k=7*	*k=8*	*k=3*	*k=4*	*k=5*	*k=6*	*k=7*	*k=8*
Silhouette score	0.2354	0.2394	0.1873	0.1914	0.1809	0.1599	0.2186	0.2179	0.1590	0.1601	0.1383	0.1175
Calinski-Harabasz Index	2025.19	1710.04	1516.91	1351.26	1227.59	1124.94	1894.50	1527.37	1297.95	1156.22	1039.17	953.76
Davies Boudin Index	1.51	1.56	1.79	1.59	1.64	1.77	1.58	1.66	2.09	1.79	1.91	1.90

Table 5 Results after applying various performance evaluation functions on DBSCAN and Birch clustering algorithm

Accuracy model	Name of the algorithms													
	DBSCAN								*Birch*					
Parameter	***eps=0.30***	***eps=0.40***	***eps=0.50***	***eps=0.60***	***eps=0.70***	***eps=0.80***	***eps=0.90***	***eps=1.00***	***k=3***	***k=4***	***k=5***	***k=6***	***k=7***	***k=8***
Silhouette score	-0.07	-0.05	-0.03	-0.01	0.01	0.03	0.02	-0.04	0.2095	0.2003	0.1326	0.1356	0.1313	0.1331
Calinski-Harabasz Index	2.74	2.80	2.98	3.13	3.42	4.01	5.13	6.47	1781.76	1463.17	1273.11	1131.95	1026.56	943.19
Davies Boudin Index	1.10	1.10	1.11	1.13	1.16	1.19	1.21	1.22	1.55	1.66	1.97	1.72	2.02	1.89

to evaluate the performance of the clustering algorithm for various parameters. The parameter for DBSCAN is eps(discussed in Sects. 3.1.3 and 3.4). And the parameter for the Birch is the number of clusters (k). The results are calculated for parametric value eps=0.30, eps=0.40, eps=0.50, eps=0.60, eps=0.70, eps=0.80, eps=0.90, eps=1.00 for DBSCAN algorithm and k=3,k=4,k=5,k=6,k=7, and k=8 for Birch clustering algorithm. Now, for our proposed PCA-K-Means clustering algorithm, we are going to apply PCA for reducing the dimension of our dataset. But as we have already discussed (in Sect. 4.1) that reducing the dimension of the dataset results in loss of information, hence we need to retain a sufficient amount of information while applying PCA.

Like our dataset contains 13 different attributes hence our data is 13 dimensional, but if we select 12 different eigenvectors and transform our data into a new set of selected coordinates (where these 12 selected eigenvectors will be our new coordinate axes) then some information is lost. But the time taken to train the model will be improved as the dimension of the model is reduced and the clustering will be efficient. So we want our model to lose the least amount of information and still reduced the dimension of the dataset.

Table 6 represents the amount of information retained with the number of eigenvectors. With 1 eigenvector, we can retain a maximum of 29.51% of the information. Similarly, if we select 2 eigenvectors, then we can retain 49.98% information at max. With 3 eigenvectors we can have 62.60%, with 4 eigenvectors we can have 70.92%, with 5 eigenvectors we can have 77.46%, with 6 eigenvectors we can have 82.55%, with 7 eigenvectors we can have 86.96%, with 8 eigenvectors we can have 90.96%, with 9 eigenvectors we can have 94.58%, with 10 eigenvectors we can have 96.92%, with 11 eigenvectors we can have 98.89%, with 12 eigenvectors we can have 99.81%, with 13 eigenvectors we can have 100% information retained.

Figure 8 graphical represents the information in Table 6. Figure 8 represents the variation of the amount of information retained (explained variance ratio) with respect to the principal component index (number of eigenvectors). So we want to select the number of eigenvectors such that sufficient information is retained. Ideally, we should retain at least 85% of the information. From Table 6 and Fig. 8, we can conclude that to retain at least

Table 6 Amount of information retained with the selected number of eigenvectors

Number of Eigen vectors	Information retained (Approx.)	Number of Eigen vectors	Information retained (Approx.)
1	29.51%	8	90.96%
2	49.98%	9	94.58%
3	62.60%	10	96.92%
4	70.92%	11	98.89%
5	77.46%	12	99.81%
6	82.55%	13	100%
7	86.96%		

85% of the information, the number of eigenvectors should be greater than or equal to 7. The performance of the PCA-K-Means model with eigenvectors greater than or equal to 7 is represented in Table 7.

Table 7 represents the results of different evaluation functions with varying the parameters (number of eigenvectors and the number of clusters). We took results for number of clusters k=3, k=4, k=5, k=6, k=7, and k=8 and the number of eigenvectors n=7, n=8, n=9, n=10, n=11, n=12. Silhouette score, Calinski-Harabasz Index, and Davies Bouldin Index are used to evaluate the performance of the clustering algorithm for various parameters.

Figure 9 compares the performance of PCA-K-Means model with varying the number of clusters(k=3, k=4, k=5, k=6, k=7, and k=8) along the x-axis and varying the number of eigenvectors(n=7, n=8, n=9, n=10, n=11, and n=12) along the y-axis. Figure 9 shows that the performance of the dataset transformed into 7 eigenvectors gives the best result in the Silhouette score, Calinski-Harabasz Score as well as with Davies Bouldin Score. This results in the conclusion that the performance of the model with 7 numbers of eigenvectors performs the best. So we will select the PCA-K-Means model with 7 eigenvectors for building the clustering model for our GCWRS.

Table 8 represents the results of performance evaluation functions for the proposed PCA-K-Means algorithm for GCWRS with 7 numbers of eigenvectors and varying the number of clusters (K=3, K=4, K=5, K=6, K=7, and K=8).

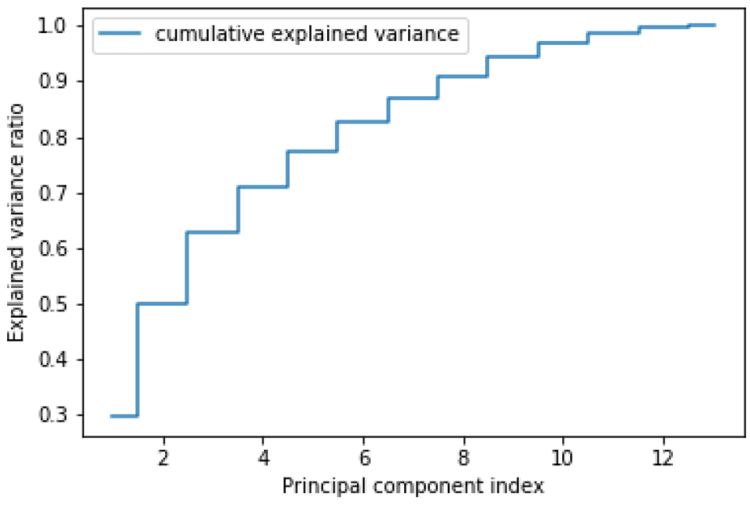


Fig. 8 Line graph representing explained variance ratio (ratio of information retained) vs principal component index (number of eigenvectors)

Table 7 Results after applying various performance evaluation functions on PCA-K-Means algorithm varying parameters

Number of clusters	Number of Eigen vectors																	
	7			8			9			10			11			12		
	Performance Evaluation function			Performance Evaluation function			Performance Evaluation function			Performance Evaluation function			Performance Evaluation function			Performance Evaluation function		
	Silhouette Score	Calinsk-Harabaz Score	Davies-Bouldin Score	Silhouette Score	Calinsk-Harabaz Score	Davies-Bouldin Score	Silhouette Score	Calinsk-Harabaz Score	Davies-Bouldin Score	Silhouette Score	Calinsk-Harabaz Score	Davies-Bouldin Score	Silhouette Score	Calinsk-Harabaz Score	Davies-Bouldin Score	Silhouette Score	Calinsk-Harabaz Score	Davies-Bouldin Score
3	0.27	2546	1.32	0.26	2354	1.38	0.25	2208	1.43	0.24	2122	1.46	0.23	2058	1.49	0.23	2031	1.51
4	0.28	2207	1.36	0.26	2021	1.43	0.25	1880	1.48	0.24	1801	1.52	0.24	1741	1.54	0.23	1715	1.56
5	0.22	1992	1.56	0.21	1814	1.63	0.20	1678	1.70	0.19	1603	1.74	0.18	1546	1.77	0.18	1522	1.79
6	0.21	1756	1.53	0.21	1613	1.52	0.20	1500	1.51	0.19	1429	1.55	0.17	1352	1.78	0.19	1353	1.64
7	0.21	1644	1.41	0.20	1474	1.52	0.19	1370	1.55	0.18	1301	1.59	0.18	1255	1.60	0.18	1225	1.69
8	0.19	1517	1.55	0.18	1363	1.62	0.17	1260	1.67	0.16	1194	1.73	0.16	1149	1.74	0.16	1129	1.76

5.3 Evaluations

In this section, Tables 4, 5 and 8 are further evaluated and compared to evaluate the best clustering model.

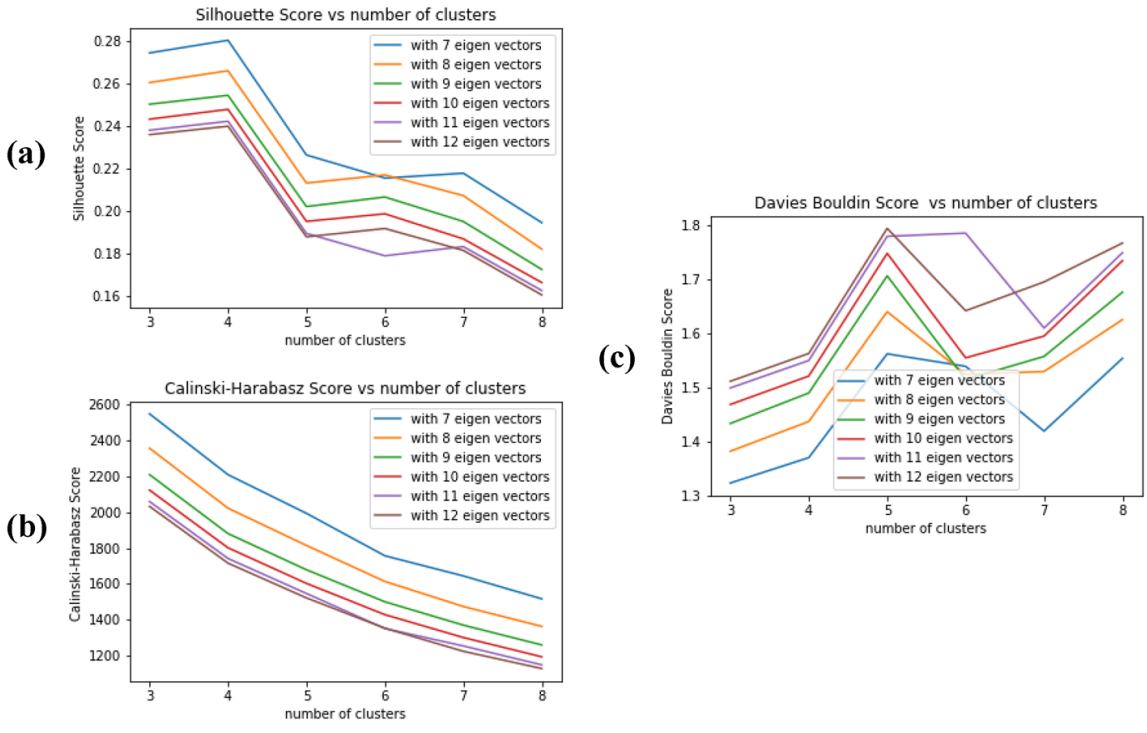


Fig. 9 Performance of PCA-K-Means algorithm with varying number of clusters and the number of eigenvectors using (**a**) Silhouette score (**b**) Calinski-Harabasz score (**c**) Davies-Bouldin score

Table 8 Results after applying various performance evaluation functions on PCA-K-Means with 7 eigenvectors

Accuracy model	Name of the algorithms					
	Proposed PCA-Kmeans Algorithm for GCWRS					
Parameter	***k=3***	***k=4***	***k=5***	***k=6***	***k=7***	***k=8***
Silhouette score	0.27	0.28	0.22	0.21	0.21	0.19
Calinski-Harabasz Index	2546	2207	1992	1756	1644	1517
Davies Boudin Index	1.32	1.36	1.56	1.53	1.41	1.55

The variation of the evolution function score with different values of parameters and for different algorithms is represented in Fig. 10. The number of clusters taken into considerations is 3 to 8, whereas the value of eps for DBSCAN is considered from 0.1 to 1.00. Using Fig. 10, we can analyze that the performance of DBSCAN is very poor as compared to all other algorithms(that is k-means clustering, agglomerative clustering, Birch clustering, and PCA-K-Means clustering algorithm) according to all the performance evaluation functions(that is Silhouette score, Calinski-Harabasz Score, Davies Bouldin Score). But the performance of other algorithms seems comparable, hence for further analysis; we need to compare the remaining algorithms even further using the comparative study as shown in Figs. 11 and 12.

5.4 Evaluating which algorithm performs best for our dataset

According to the Silhouette evaluation function, the silhouette score should be as high as possible to have highly dense clustering. In the comparative study done in Figs. 11a and 12a, the value of the silhouette score is maximum for our proposed clustering algorithm even with all of the value of the parameter (for all classes from 3 to 8). According to the Calinski-Harabasz evaluation function, the value of the Calinski-Harabasz Score should be as high as possible to have dense and well-separated clusters. In the comparative study done in Figs. 11b and 12b, the value of the Calinski-Harabasz Score is maximum for our proposed clustering algorithm even with all of the values of the parameters (for all classes from 3 to 8). According to the Davies-Bouldin evaluation function, the value of the Davies-Bouldin index score should be close to zero to have a better partitioning of the clusters. In the comparative study done in Figs. 11c and 12c, the value of the Davies-Bouldin evaluation function is minimum for our proposed clustering algorithm even with all of the values of the parameters (for all classes from 3 to 8). If we look at the overall score of all these three algorithms, then our proposed algorithm performs significantly better than every other standard algorithm.

5.5 Evaluating the ideal number of clusters

Using the Silhouette evaluation function: In the comparative study done in Figs. 11a and 12a, we can analyze that the value of the Silhouette score rapidly drops if the number

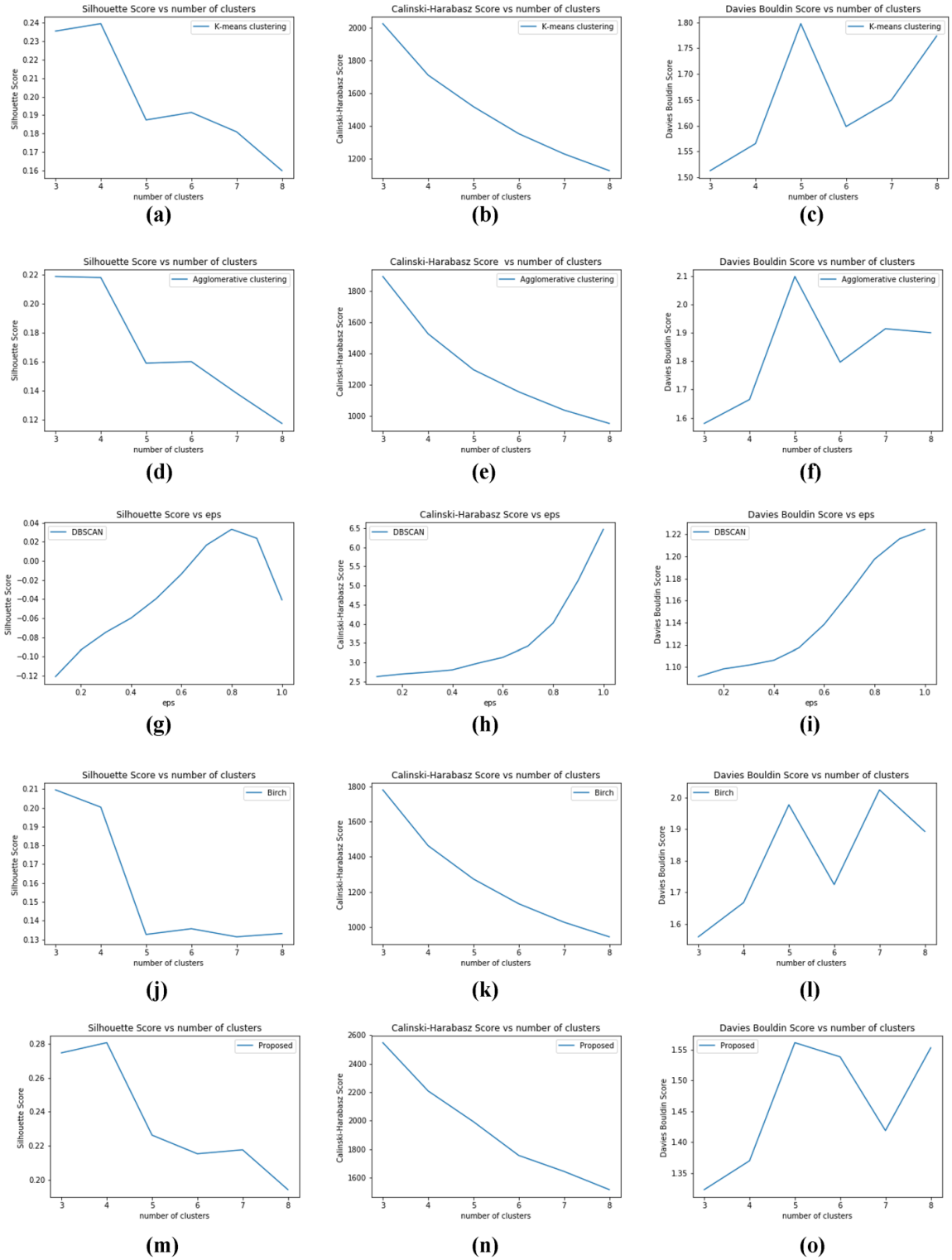


Fig. 10 Line graph to represent the variation of (**a**)Silhouette score with the number of cluster in k-means clustering, (**b**) Calinski-Harabasz Score with the number of clusters in k-means clustering, (**c**) Davies Bouldin Score with the number of clusters in k-means clustering, (**d**)Silhouette score with the number of cluster in agglomerative clustering, (**e**) Calinski-Harabasz Score with the number of clusters in agglomerative clustering, clustering, (**f**) Davies Bouldin Score with the number of clusters in agglomerative clustering, clustering, (**g**)Silhouette score with eps in DBSCAN clustering, (**h**) Calinski-Harabasz Score with the eps in DBSCAN clustering, (**i**) Davies Bouldin Score with the eps in DBSCAN clustering, (**j**)Silhouette score with the number of cluster in Birch clustering, (**k**) Calinski-Harabasz Score with the number of clusters in Birch clustering, (**l**) Davies Bouldin Score with the number of clusters in Birch clustering, (**m**)Silhouette score with the number of cluster in proposed algorithm, (**n**) Calinski-Harabasz Score with the number of clusters in proposed algorithm, (**o**) Davies Bouldin Score with the number of clusters in proposed algorithm

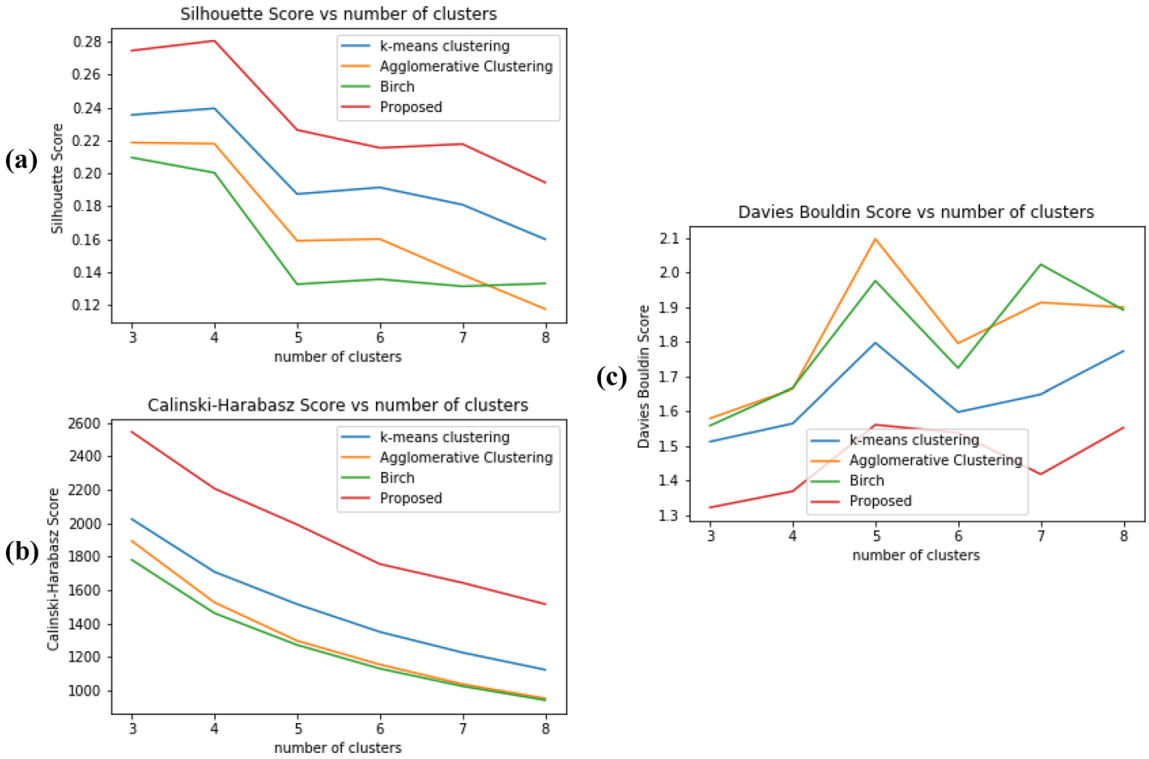


Fig. 11 Comparative study of the k-means clustering algorithm, Agglomerative clustering, Birch clustering algorithm, and proposed algorithm using (**a**) Silhouette Score, (**b**) Calinski-Harabasz Score, (**c**) Davies Bouldin Score

of clusters is 5 or more. So according to the Silhouette score evaluation function, we conclude that the ideal number of clusters should be 3 or 4. Using the Calinski-Harabasz evaluation function: In the comparative study done in Figs. 11b and 12b, we can analyze

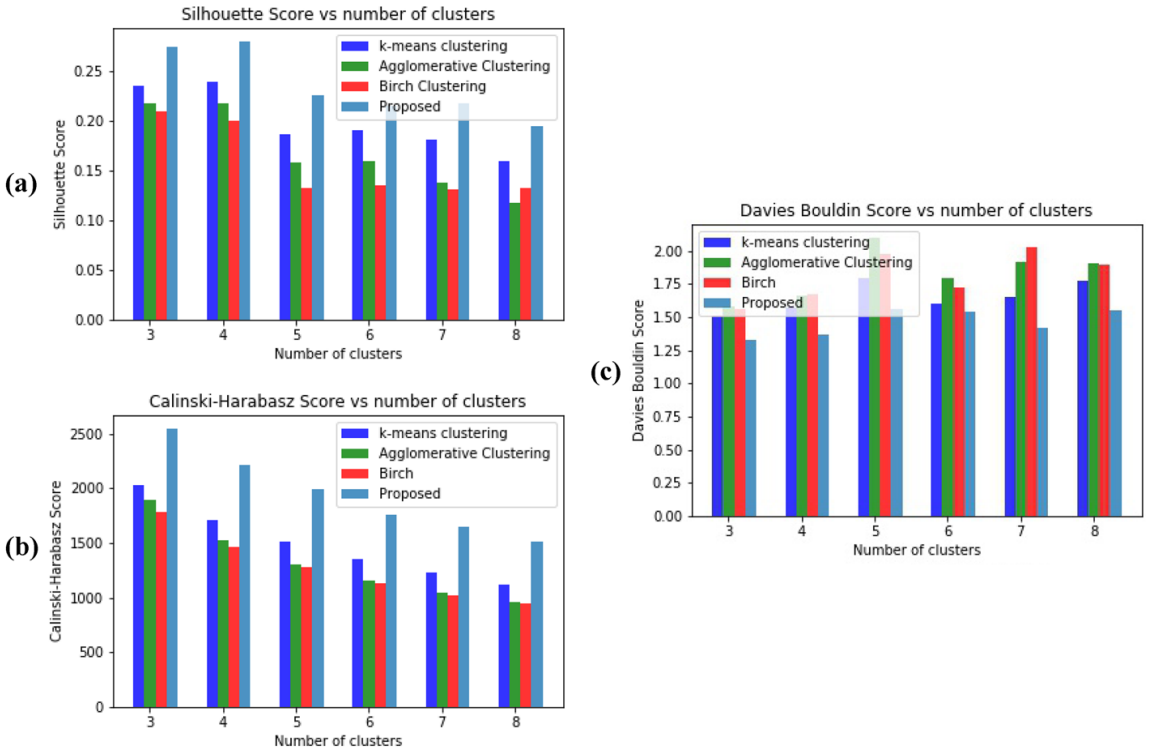


Fig. 12 Comparative study of the k-means clustering algorithm, Agglomerative clustering, Birch clustering algorithm, and proposed algorithm using (**a**) Silhouette Score, (**b**) Calinski-Harabasz Score, (**c**) Davies Bouldin Score

that the value of the Calinski-Harabasz score drops consistently from 3 to 8. So according to the Calinski-Harabasz evaluation function, we concluded that the ideal number of clusters should be as low as possible (like 3 or 4). Using Davies-Bouldin's evaluation function: In the comparative study done in Figs. 11c and 12c, we can analyze that the value of the Davies-Bouldin score steadily increases with the increase in the number of clusters with an exceptional drop at the number of cluster=6. So according to the Davies-Bouldin evaluation function, we concluded that the ideal number of clusters should be 3, 4, or 6.

6 Conclusion and future direction

In this paper, we used five different clustering algorithms (K-means clustering, agglomerative clustering, DBSCAN, Birch clustering, and the proposed PCA-K-Means clustering algorithm) to use in the first part of our two parted GCWRS. The performance of each algorithm is evaluated using performance evaluation functions like the Silhouette score, Calinski-Harabasz index, and Davies-Bouldin Index. And the results from the performance evaluation functions were evaluated and concluded that the proposed PCA-K-Means (with 7 number of eigenvectors) clustering algorithm gives the best result among all other algorithms with the number of clusters equal to 3 to 4. But if we select three clusters, our number of recommendations will get limited, so it is better to take four clusters to provide the user with more recommended wines to choose from. The second part of our two parted GCWRS uses a novel greedy approach based on recommending technique that follows a greedy approach to recommend different wines to the consumer according to their requirements(try a new kind of wine or enhance the existing preference). We proposed a new recommendation system GCWRS that uses the PCA-K-Means clustering algorithm along with a novel greedy recommendation technique. The main asset of the proposed system is that as the number of users increases, GCWRS will get even better with much more personalized suggestions. For future directions, we can make changes in the first part of GCWRS by using different other clustering algorithms or by proposing a deep learning model to enhance the performance of the clustering algorithm.

References

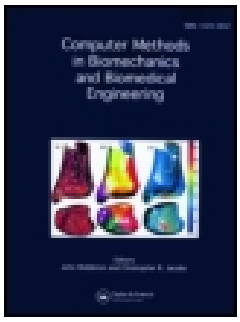
1. Abdi H, Williams LJ (2010) Principal component analysis, Wiley Interdiscip. Rev Comput Stat. https://doi.org/10.1002/wics.101
2. Al-Hassan M, Lu H, Lu J (2015) A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system. Decis Support Syst. https://doi.org/10.1016/j.dss.2015.02.001
3. Al-Naymat G, Al-Kasassbeh M, Sober Z (2018) Pairs trading strategy: a recommendation system. Int J Comput Appl. https://doi.org/10.1080/1206212X.2018.1493074
4. Asuncion A, Newman DJ (2007) {UCI} Machine Learning Repository, \emph{https://www.ics.uci.edu/$/sim$mlearn/MLRepository.html}. University of California, Irvine, School of Information and Computer Sciences
5. Azaria A, Hassidim A, Kraus S, Eshkol A, Weintraub O, Netanely I (2013) Movie recommender system for profit maximization, in: AAAI Work. Tech Rep
6. Basu S, Banerjee A, Mooney R (2002) Semi-supervised clustering by seeding. Proc 19th Int Conf Mach Learn

7. Basu S, Bilenko M, Mooney RJ (2004) A probabilistic framework for semi-supervised clustering. KDD 2004 Proc Tenth ACM SIGKDD Int Conf Knowl Discov Data Min. https://doi.org/10.1145/1014052.1014062
8. Bholowalia P, Kumar A (2014) EBK-means: A clustering technique based on elbow method and K-means in WSN
9. Blackledget JM (2006) Eigenvalues and Eigenvectors. Digit Signal Process. https://doi.org/10.1533/9780857099457.2.255
10. Boehmke B, Greenwell B, Boehmke B, Greenwell B (2020) K-means Clustering. Hands-On Mach Learn with R. https://doi.org/10.1201/9780367816377-20
11. Brandes U, Gaertler M, Wagner D (2003) Experiments on graph clustering algorithms. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinform). https://doi.org/10.1007/978-3-540-39658-1_52
12. Burke R (2002) Hybrid recommender systems: Survey and experiments. User Model User Adapt Interact. https://doi.org/10.1023/A:1021240730564
13. Buşoniu L, Babuška R, De Schutter B (2010) Multi-agent reinforcement learning: An overview. Stud Comput Intell. https://doi.org/10.1007/978-3-642-14435-6_7
14. Caron M, Bojanowski P, Joulin A, Douze M (2018) Deep clustering for unsupervised learning of visual features. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinform). https://doi.org/10.1007/978-3-030-01264-9_9
15. Cengizler C, Kerem-Un M (2017) Evaluation of Calinski-Harabasz Criterion as Fitness Measure for Genetic Algorithm Based Segmentation of Cervical Cell Nuclei. Br J Math Comput Sci. https://doi.org/10.9734/bjmcs/2017/33729
16. Chandak M, Girase S, Mukhopadhyay D (2015) Introducing hybrid technique for optimization of book recommender system. Procedia Comput Sci. https://doi.org/10.1016/j.procs.2015.03.075
17. Choi SM, Ko SK, Han YS (2012) A movie recommendation algorithm based on genre correlations. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2012.01.132
18. Contreras P, Murtagh F (2015) Hierarchical clustering. Handb Clust Anal. https://doi.org/10.1201/b19706
19. Cortez P, Portelinha M, Rodrigues S, Cadavez V, Teixeira A (2006) Lamb meat quality assessment by Support Vector Machines. Neural Process Lett. 24:41–51. https://doi.org/10.1007/s11063-006-9009-6
20. Cortez P, Cerdeira A, Almeida F, Matos T, Reis J (2009) Modeling wine preferences by data mining from physicochemical properties. Decis Support Syst 47:547–553. https://doi.org/10.1016/j.dss.2009.05.016
21. Craw S (2017) Manhattan Distance. Encycl Mach Learn Data Min. https://doi.org/10.1007/978-1-4899-7687-1_511
22. Danielsson PE (1980) Euclidean distance mapping. Comput Graph Image Process. https://doi.org/10.1016/0146-664X(80)90054-4
23. De Maesschalck R, Jouan-Rimbaud D, Massart DL (2000) The Mahalanobis distance. Chemom Intell Lab Syst. https://doi.org/10.1016/S0169-7439(99)00047-7
24. Eigenvalues and Eigenvectors (1973) Math Sci Eng. https://doi.org/10.1016/S0076-5392(08)60533-0
25. Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, pp 226–231
26. Eyjolfsdottir E, Tilak G, Li N (2010) MovieGEN: A Movie Recommendation System. Comput Sci Dep
27. Fasanghari M, Montazer GA (2010) Design and implementation of fuzzy expert system for Tehran Stock Exchange portfolio recommendation. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2010.02.114
28. Feo TA, Resende MGC (1995) Greedy Randomized Adaptive Search Procedures. J Glob Optim. https://doi.org/10.1007/BF01096763
29. Gaffney S, Smyth P (1999) Trajectory clustering with mixtures of regression models. https://doi.org/10.1145/312129.312198
30. Gokcekus O, Gokcekus S (2019) Empirical evidence of lumping and splitting: Expert ratings' effect on wine prices. Wine Econ Policy 8:171–179. https://doi.org/10.1016/j.wep.2019.09.003
31. Guo X, Lu J (2007) Intelligent e-Government services with personalized recommendation techniques. Int J Intell Syst. https://doi.org/10.1002/int.20206
32. Hartigan A, Wong MA (1979) A K-Means Clustering Algorithm. J R Stat Soc. https://doi.org/10.2307/2346830
33. Hartuv E, Shamir R (2000) Clustering algorithm based on graph connectivity. Inf Process Lett. https://doi.org/10.1016/S0020-0190(00)00142-3
34. Haseeb S, Alexander B, Santi RL, Liprandi AS, Baranchuk A (2019) What's in wine? A clinician's perspective. Trends Cardiovasc Med 29:97–106. https://doi.org/10.1016/j.tcm.2018.06.010
35. Hwangbo H, Kim YS, Cha KJ (2018) Recommendation system development for fashion retail e-commerce. Electron Commer Res Appl. https://doi.org/10.1016/j.elerap.2018.01.012

36. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett. https://doi.org/10.1016/j.patrec.2009.09.011
37. Jain E, Jain SK (2015) Using Mahout for clustering similar Twitter users: Performance evaluation of k-means and its comparison with fuzzy k-means. Proc 5th IEEE Int Conf Comput Commun Technol ICCCT 2014. https://doi.org/10.1109/ICCCT.2014.7001465
38. Johnson H (1989) Vintage: The story of Wine. ISBN 10: 0671791826 / ISBN 13: 9780671791827, Simon & Schuster, NY
39. Kabassi K (2010) Personalizing recommendations for tourists. Telemat Inform. https://doi.org/10.1016/j.tele.2009.05.003
40. Karkkainen I, Franti P (2000) Minimization of the value of davies-bouldin index. Proc IASTED Int Conf Signal Process Commun
41. Katarya R, Verma OP (2017) An effective collaborative movie recommender system with cuckoo search. Egypt Informatics J. https://doi.org/10.1016/j.eij.2016.10.002
42. Kraskov A, Stögbauer H, Grassberger P (2004) Estimating mutual information. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top. https://doi.org/10.1103/PhysRevE.69.066138
43. Kurmashov N, Latuta K, Nussipbekov A (2016) Online book recommendation system. Proc 2015 12th Int Conf Electron Comput Comput ICECCO 2015. https://doi.org/10.1109/ICECCO.2015.7416895
44. Kustos M, Goodman S, Jeffery DW, Bastian SEP (2019) Using consumer opinion to define New World fine wine: Insights for hospitality. Int J Hosp Manag 83:180–189. https://doi.org/10.1016/j.ijhm.2019.04.018
45. Legin A, Rudnitskaya A, Lvova L, Vlasov Y, Di Natale C, D'Amico A (2003) Evaluation of Italian wine by the electronic tongue: Recognition, quantitative analysis and correlation with human sensory perception. Anal Chim Acta. https://doi.org/10.1016/S0003-2670(03)00301-5
46. Lekakos G, Caravelas P (2008) A hybrid approach for movie recommendation. Multimed Tools Appl. https://doi.org/10.1007/s11042-006-0082-7
47. Li H, He F, Chen Y (2020) Learning dynamic simultaneous clustering and classification via automatic differential evolution and firework algorithm. Appl Soft Comput 96:106593, ISSN 1568-4946. https://doi.org/10.1016/j.asoc.2020.106593
48. Li H, He F, Chen Y, Luo J (2020) Multi-objective self-organizing optimization for constrained sparse array synthesis. Swarm Evol Comput 58:100743, ISSN 2210-6502. https://doi.org/10.1016/j.swevo.2020.100743
49. Li YM, Te Wu C, Lai CY (2013) A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. Decis Support Syst. https://doi.org/10.1016/j.dss.2013.02.009
50. Liu B, Xia Y, Yu PS (2000) Clustering through decision tree construction. https://doi.org/10.1145/354756.354775
51. Liu B, Xia Y, Yu PS (2005) Clustering via decision tree construction. https://doi.org/10.1007/11362197_5
52. Lu J, Shambour Q, Xu Y, Lin Q, Zhang G (2010) BizSeeker: A hybrid semantic recommendation system for personalized government-to-business e-services. Internet Res. https://doi.org/10.1108/10662241011050740
53. Lucas JP, Luz N, Moreno MN, Anacleto R, Figueiredo AA, Martins C (2013) A hybrid recommendation approach for a tourism system. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2012.12.061
54. Lucia SP (1963) A History of Wine As Therapy, J. B. Lippincott Company; 1st edition
55. Mathew P, Kuriakose B, Hegde V (2016) Book Recommendation System through content based and collaborative filtering method. Proc 2016 Int Conf Data Min Adv Comput SAPIENCE 2016. https://doi.org/10.1109/SAPIENCE.2016.7684166
56. Maulik U, Bandyopadhyay S (2000) Genetic algorithm-based clustering technique. Pattern Recognit. https://doi.org/10.1016/S0031-3203(99)00137-5
57. McGovern PE, Katz SH, Fleming SJ (2003) The origins and ancient history of wine: Food and nutrition in history and antropology. https://doi.org/10.4324/9780203392836
58. Moreno IM, González-Weller D, Gutierrez V, Marino M, Cameán AM, González AG, Hardisson A (2007) Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks. Talanta. 72:263–268. https://doi.org/10.1016/j.talanta.2006.10.029
59. Pan Y, He F, Yu H (2019) A novel Enhanced Collaborative Autoencoder with knowledge distillation for top-N recommender systems. Neurocomputing 332:137–148, ISSN 0925-2312. https://doi.org/10.1016/j.neucom.2018.12.025
60. Pan Y, He F, Yu H (2020) Learning social representations with deep autoencoder for recommender system. World Wide Web 23:2259–2279. https://doi.org/10.1007/s11280-020-00793-z
61. Perbet F, Stenger B, Maki A (2009) Random forest clustering and application to video segmentation. Br Mach Vis Conf BMVC 2009 Proc. https://doi.org/10.5244/C.23.100
62. Pomarici E, Lerro M, Chrysochou P, Vecchio R, Krystallis A (2017) One size does (obviously not) fit all: Using product attributes for wine market segmentation. Wine Econ Policy 6:98–106. https://doi.org/10.1016/j.wep.2017.09.001

63. Priyanka K, Tewari AS, Barman AG (2015) Personalised book recommendation system based on opinion mining technique. Glob Conf Commun Technol GCCT. https://doi.org/10.1109/GCCT.2015.7342668
64. Rédei GP (2008) Euclidean distance. Encycl Genet Genom Proteom Inform. https://doi.org/10.1007/978-1-4020-6754-9_5603
65. Risius A, Klann BO, Meyerding SGH (2019) Choosing a lifestyle? Reflection of consumer extrinsic product preferences and views on important wine characteristics in Germany. Wine Econ Policy 8:141–154. https://doi.org/10.1016/j.wep.2019.09.001
66. Rong C, Esteves RM (2011) Using Mahout for clustering Wikipedia's latest articles: A comparison between k-means and fuzzy c-means in the cloud. In Proceeding 2011 3rd IEEE International Conference on Cloud Computing Technology and Science. https://doi.org/10.1109/CloudCom.2011.86
67. Rousseeuw PJ (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. https://doi.org/10.1016/0377-0427(87)90125-7
68. Schubert E, Sander J, Ester M, Kriegel HP, Xu X (2017) DBSCAN Revisited, Revisited. ACM Trans Database Syst. https://doi.org/10.1145/3068335
69. Sebastia L, Garcia I, Onaindia E, Guzman C (2009) E-Tourism: A tourist recommendation and planning application. Int J Artif Intell Tools. https://doi.org/10.1142/S0218213009000378
70. Shambour Q, Lu J (2011) Government-to-Business personalized e-services using semantic-enhanced recommender system. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinform). https://doi.org/10.1007/978-3-642-22961-9_16
71. Smith BC (2019) Getting More Out of Wine: wine experts, wine apps and sensory science. Curr Opin Food Sci 27:123–129. https://doi.org/10.1016/j.cofs.2019.10.007
72. Sohail SS, Siddiqui J, Ali R (2013) Book recommendation system using opinion mining technique. Proc 2013 Int Conf Adv Comput Commun Inform ICACCI. https://doi.org/10.1109/ICACCI.2013.6637421
73. Steinley D (2004) Properties of the Hubert-Arabie adjusted Rand index. Psychol Methods. https://doi.org/10.1037/1082-989X.9.3.386
74. Sun LX, Danzer K, Thiel G (1997) Classification of wine samples by means of artificial neural networks and discrimination analytical methods. Fresenius J Anal Chem 359:143–149. https://doi.org/10.1007/s002160050551
75. Syms C (2018) Principal components analysis. Encycl Ecol. https://doi.org/10.1016/B978-0-12-409548-9.11152-2
76. Tewari AS, Kumar A, Barman AG (2014) Book recommendation system based on combine features of content based filtering, collaborative filtering and association rule mining. Souvenir 2014 IEEE International Advance Computing Conference. https://doi.org/10.1109/IAdCC.2014.6779375
77. Tweedale J, Ichalkaranje N, Sioutis C, Jarvis B, Consoli A, Phillips-Wren G (2007) Innovations in multi-agent systems. J Netw Comput Appl. https://doi.org/10.1016/j.jnca.2006.04.005
78. Vlassides S, Ferrier JG, Block DE (2001) Using historical data for bioprocess optimization: Modeling wine characteristics using artificial neural networks and archived process information. Biotechnol Bioeng. https://doi.org/10.1002/1097-0290(20010405)73:1%3c55::AID-BIT1036%3e3.0.CO;2-5
79. Wang W, Mishra KK (2018) A novel stock trading prediction and recommendation system. Multimed Tools Appl. https://doi.org/10.1007/s11042-017-4587-z
80. Winters-Hilt S, Merat S (2007) SVM clustering. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-8-S7-S18
81. Xu C, Zhang P, Li B, Wu D, Fan H (2013) Vague C-means clustering algorithm. Pattern Recognit Lett. https://doi.org/10.1016/j.patrec.2012.12.001
82. Yang MS (1993) A survey of fuzzy clustering. Math Comput Model. https://doi.org/10.1016/0895-7177(93)90202-A
83. Yang Z (2013) Recommendation Systems. Appl Data Min. https://doi.org/10.1201/b15027-15
84. Yu H, Lin H, Xu H, Ying Y, Li B, Pan X (2008) Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy. J Agric Food Chem. https://doi.org/10.1021/jf0725575
85. Zhang B (2003) Regression clustering. Proceeding IEEE International Conference on Data Mining. https://doi.org/10.1109/icdm.2003.1250952
86. Zhang T, Ramakrishnan R, Livny M (1996) BIRCH: An efficient data clustering method for very large databases. SIGMOD Rec (ACM Spec Interes Gr Manag Data) https://doi.org/10.1145/235968.233324

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



Taylor & Francis
Taylor & Francis Group

Computer Methods in Biomechanics and Biomedical Engineering

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gcmb20

Finite element study on the influence of pore size and structure on stress shielding effect of additive manufactured spinal cage

Vijay Kumar Meena, Parveen Kalra & Ravindra Kumar Sinha

To cite this article: Vijay Kumar Meena, Parveen Kalra & Ravindra Kumar Sinha (2021): Finite element study on the influence of pore size and structure on stress shielding effect of additive manufactured spinal cage, Computer Methods in Biomechanics and Biomedical Engineering, DOI: 10.1080/10255842.2021.1970142

To link to this article: https://doi.org/10.1080/10255842.2021.1970142

View supplementary material

Published online: 23 Sep 2021.

Submit your article to this journal

Article views: 26

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=gcmb20

COMPUTER METHODS IN BIOMECHANICS AND BIOMEDICAL ENGINEERING
https://doi.org/10.1080/10255842.2021.1970142

Taylor & Francis
Taylor & Francis Group

Check for updates

Finite element study on the influence of pore size and structure on stress shielding effect of additive manufactured spinal cage

Vijay Kumar Meena[a,b], Parveen Kalra[b] and Ravindra Kumar Sinha[c]

[a]Biomedical Instrumentation, Central Scientific Instruments Organisation, Chandigarh, India; [b]Department of Production Engineering, Punjab Engineering College, Chandigarh, India; [c]Department of Applied Physics, Delhi Technological University, New Delhi, India

ABSTRACT

The stress shielding effect occurs when the orthopedic implant reduces the load delivered to the bone, causing inefficient stress transfer to the host bone. The usage of porous additive manufactured (AM) cages reduces the stress shielding effect and promotes bone ingrowth also. The purpose of this work is to study the stress and deformation on porous hybrid spinal cages under different loading conditions using Finite Element Analysis (FEA). The spinal cages consisting of three porous structures with pore sizes ranging from 0.4 to 0.6 mm were investigated for stress shielding and fatigue strength. The results showed a significant reduction in stress shielding for the studied designs and conclude that the pore size has a greater significant effect on stress shielding than the porous structure in spinal cages.

ARTICLE HISTORY
Received 2 March 2021
Accepted 16 August 2021

KEYWORDS
Additive manufacturing; porous structure; pore size; FEA; fatigue

1. Introduction

Spinal injuries and spinal disorders are becoming increasingly common ailments in modern life. One of the most common diseases/ailments is Degenerative Disc Disease, popularly known as DDD (Pannell et al. 2015). The cure involves surgical procedures of degenerated disc and insertion of appropriate implants, namely spinal cages between the vertebrae. Titanium alloy (Ti6Al4V) is the preferred choice for spinal cages due to its high tensile strength, good biocompatibility, good fatigue strength, and corrosion resistance (Ramakrishna et al. 2001). These spinal cages are generally made of solid dense metals/polymers e.g. titanium, Carbon Fiber Reinforced PEEK, etc. Young's modulus of these materials is much higher than human bone Young's modulus. The elastic modulus of bone varies between 1 and 20 GPa whereas the elastic modulus of titanium is 110 GPa. Due to this vast difference in elastic modulus, the loads are not transferred from the implant to adjacent bone tissue, resulting in stress shielding between the host bone and the implant. This leads to adaptive resorption of bone tissue and a decrease in mechanical rigidity of the bone as per Wolff's law (Chen et al. 2010). Similarly, with reduced stress shielding, bone tissues are known to generate deposition of new bone, which increases mechanical rigidity (Stock 2018). Also, the smooth and shiny surface of solid metal implants makes it difficult to integrate with the host bone. This causes amyotrophy and osteonecrosis of bones around the implants, loosening of the implant, distortion of bones, etc (Haibo et al. 2012).

To overcome the issues related to metallic implants, researchers are developing implants with bioactive surfaces or porous structured to promote cellular adhesion and bony in-growth (Kanayama et al. 2000; Herrera et al. 2014). Major parameters influencing the formation of new bone, are implant material-bone apposition, the stability of the porous implant-bone interface, and availability of the surrounding bone. However, the porosity percentage, pore size, and the implant material have significant effects on the performance of an implant (Cook et al. 1985; Hofmann et al. 1997; Itälä et al. 2001). The implant must have adequate porosity, in between 35 and 70% to achieve bone ingrowth (osseointegration) into the implant porous surface. Animal implantation experimentation showed that pore size affects the bone ingrowth potential significantly. Different researchers have recommended different pore sizes for bone ingrowth varying from 50 µm to 600 µm (Bobyn et al. 1980; Barth et al. 1986; Bobyn et al. 1999; Melikyan and Itin 2002; Taniguchi et al. 2016).

CONTACT Vijay Kumar Meena vijay.kumar.meena@gmail.com, vijaykumar@csio.res.in

Supplemental data for this article is available online at https://doi.org/10.1080/10255842.2021.1970142.

© 2021 Informa UK Limited, trading as Taylor & Francis Group

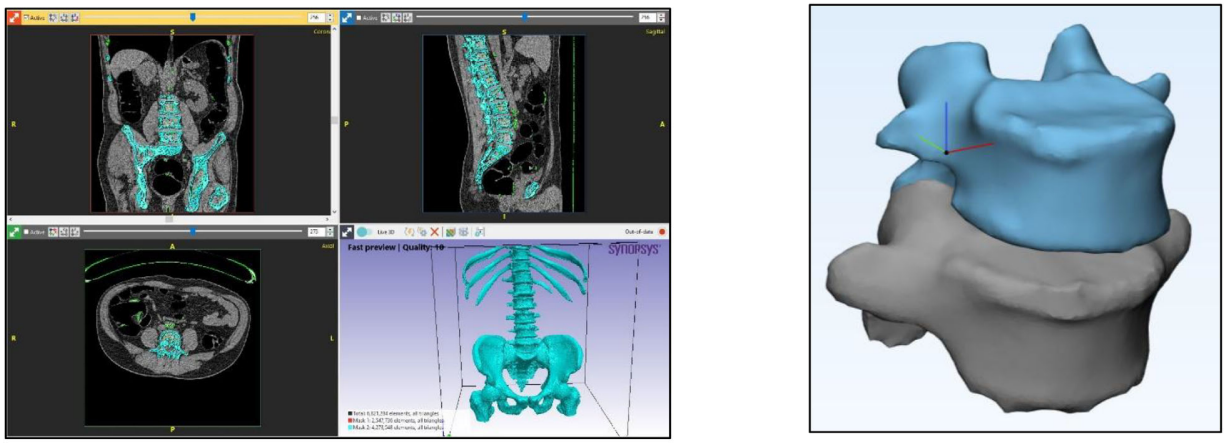


Figure 1. (a) Demo model of lumbar spine segment, (b) CAD model of L4-L5 vertebrae.

Tsuruga et al. reported the pore size range of 300–400 µm is optimum for the growth of osteoblasts and vascularization in porous implants (Tsuruga et al. 1997). Clemow et al. also reported similar results for optimum size of pores i.e. 200–500 µm in porous surface coated Ti6Al4V specimens implanted into the femoral medullary canal of dogs (Clemow et al. 1981). However, when the pore size exceeds above 1.0 mm, it supports the formation of fibrous tissues more than the growth of bone (Itälä et al. 2001). Lee et al. developed spinal Posterior Lumbar Interbody Fusion (PLIF) cages with 1 mm diameter having 50% porosity and performed Finite Element Analysis (FEA) under the different orientations of the cage with and without bone fusion (Lee et al. 2016). Chuah et al. showed a reduction in stress shielding using a topology-optimized spinal cage (Chuah et al. 2010). Zhang et al. compared the biomechanical properties of porous and solid Ti cages with PEEK cages using FEA (Wu et al. 2013). The Young's modulus of the porous structure was calculated mechanically and assigned the properties of the porous structure to the solid region for performing FEA.

The reported research studies mostly emphasized the bone ingrowth feature of porous metals. However, the lattice structure has a greater effect on load transfer to the host bone and stress shielding reduction. The study of the stress shielding effect of porous structures combined with pore sizes for hybrid spinal cages under different loading conditions has not been investigated in detail. In previous studies, the FEA studies have been performed either by assigning porous metal properties to the solid region or without any convergence analysis study (Zhang et al. 2018; Zhu et al. 2020). Fatigue analysis using FEA for hybrid spinal cages has not been performed in any previously published works to the best of the authors' knowledge. In the present work, hybrid PLIF spinal cages with both porous and solid regions were designed and the effect of various pore structures and pore sizes on stress shielding under different loading modes was studied using FEA. The pore sizes used in this study have the potential for bone ingrowth as recommended by previous literature (Frosch et al. 2004; Taniguchi et al. 2016). Fatigue analysis of spinal cages with the least stress shielding was also simulated to ascertain the life cycle of the porous spinal cages.

2. Methodology

2.1. Spine model development

A demo 3 D CAD model of the lumbar spine segment was obtained from the software Simpleware ScanIP 2019.09 (Synopsys, CA, USA) as shown in Figure 1(a). CAD model of L4-L5 vertebrae was separated from the rest of the spine in the software. Smoothening and wrapping of the model were done in the 3-Matic module of Mimics software version 13.0 (Materialise NV, Leuven, Belgium). The final model of L4-L5 vertebrae taken for the study is shown in Figure 1(b).

2.2. Hybrid PLIF cage model development

Solid PLIF cages were designed in Solidworks 2018 software (Dassault Systèmes SolidWorks Corporation, MA, USA) shown in Figure 2. To model a hybrid cage consisting of both solid and porous regions, a design space for the porous region was defined in the solid cage (Figure 3). This design space was used for

the generation of Diamond, Gyroid, and Schwarz W porous structures with pore sizes of 0.4 mm, 0.5 mm, and 0.6 mm in Simpleware ScanIP software using CAD + module. Figure 4 shows the pore structure geometry of the designs used in the study.

2.3. FEA mesh generation & boundary conditions

Before meshing the actual assembly model, a convergence analysis study has been performed. For this, a simple assembly model consisting of a hybrid spinal implant sandwiched between two cylindrical plates is designed (Figure 5) and meshed with 3 D tetrahedra elements in Simpleware software. Simpleware software used a grading scale ranging between −50 to +50 to define the coarseness and fineness of the mesh. Going from left to right on the scale, the mesh got finer. Four grade values of −20, −15, −10, and −5 for the meshing were used to study the convergence. The four output meshes are imported in Abaqus 2020 software (Dassault Systèmes Simulia Corp., RI, USA) and FEA is performed simultaneously. A uniaxial load of 400 N is applied on the top surface of the upper plate and the bottom surface of the lower plate is given fixed constraints for all the models. The number of elements and nodes with the von-Mises stress values for each grade value is tabulated in Table 1.

A graph is plotted between grade value for meshing and von-Mises stress value (Figure 6) and from the graph, it is concluded that the stress values are converging for −10 and −5 grades. So, to mesh the actual model −10 grade value is used as for the −5 grade value the simulation time will be more.

Each designed hybrid PLIF spinal cage (total nine) and one solid PLIF cage was assembled with the L4-L5 vertebras and meshed with 3 D tetrahedra elements using Simpleware ScanIP software. The number of nodes and elements for each assembly model was mentioned in Table 2.

The assembly mesh was then imported into Abaqus software for performing FEA. The meshing and FEA were performed on the FUJITSU workstation having a system configuration of Intel(R)Xeon(R)CPUE5-2600v4@2.00 GHz processor, 192 GB RAM, and 8 GB graphics card.

The assembly was constrained from the lower surface of the L5 vertebra. A physiological load of approximately 55–60% of the total body weight, that is, about 400 N for the standard body weight of 700 N was applied on the top surface of the L4 vertebra for

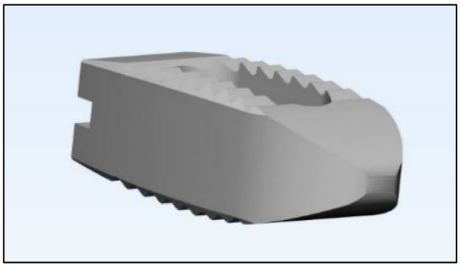


Figure 2. Solid PLIF cage.

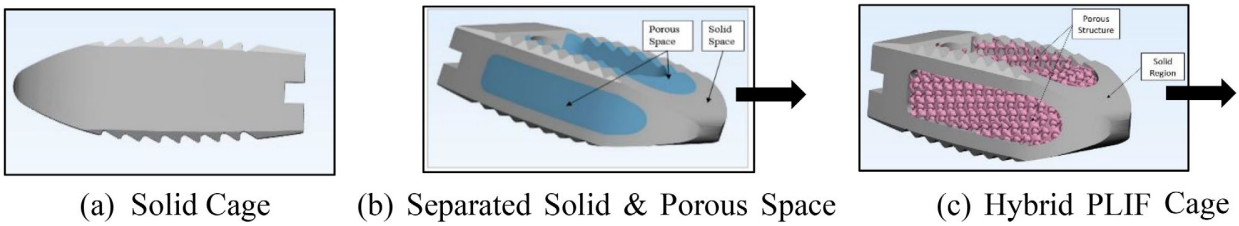


Figure 3. Porous cage design workflow.

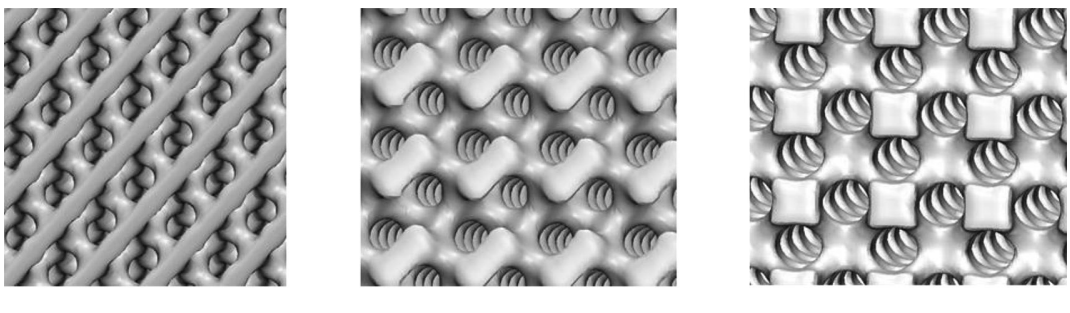


Figure 4. CAD representation of different porous structures.

uniaxial compression testing (Kurutz 2010). A moment of 7.5 Nm was applied about all the three axes (x, y, and z) respectively for the study of flexion/extension, bending, and torsion (Figure 7), and is consistent with the previous studies (Zhang et al. 2018; Zhu et al. 2020). All the connections of the assembly were assigned tie constraints including between the facet joints of vertebrae L4 and L5, simulating the effect of screw and rod (Epasto et al. 2019). Due to fixed constraints between the facet, relative movements are restricted between the vertebrae along with the screw-rod system. As the relative motion between the vertebrae was constrained, the FE model was developed without any ligament (Zhang et al. 2009; Choi et al. 2013). It is worth mentioning that the assumptions mentioned above can be considered as a limitation of the study. Boundary conditions used were consistent with data in the works of literature (Goel et al. 1995; Epasto et al. 2019). The loading and boundary conditions were kept the same for all the design configurations.

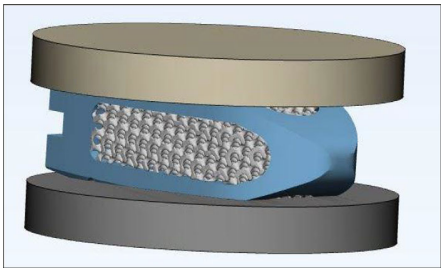


Figure 5. Assembly model for convergence analysis.

Table 1. Number of nodes and elements with maximum von-Mises stress values of assembly model used for convergence analysis.

Grade value for meshing	Number of nodes	Number of elements	Maximum von-Mises stress on assembly (MPa)
−20	399,612	1,218,805	257.9
−15	440,519	1,348,806	395
−10	545,558	1,707,447	647
−5	1,425,073	5,170,833	655

2.4. Material property assignment

For material assignment to vertebrae, each vertebra was divided into two parts: cancellous bone and cortical bone (Figure 8). The thickness of the outer shell cortical bone was kept at 1.0 mm (Ambati et al. 2015). The material of the vertebrae is assumed to be isotropic. For the solid region of the hybrid PLIF spinal cage, the material properties of biocompatible grade Ti6Al4V ELI alloy were selected. Since with porous structure, Young's modulus of metal decreases, it had to be experimentally calculated. To do so, samples of size 10 mm x 10 mm x 20 mm for each type of porous structure and pore size were modeled in Simpleware ScanIP software. 5 samples of each design (total 45 samples) were additively manufactured in an EOS M290 metal printer using EOS Ti6Al4V ELI biocompatible grade powder. Table 3 shows the fabricated sample of each porous structure and pore size. The samples were compression tested at a constant strain rate of 0.005 in/in/min (m/m/min) on a Universal Testing Machine (DAK System Inc Series 7200) according to the standard ISO 13314. The Young's modulus values were computed automatically using the embedded software of the UTM machine. The stress-strain graphs obtained from the compression test for different specimens are shown in Supplementary File I. The experimentally determined average values of Young's Modulus for different structures with different pore sizes are shown in Table 4. Respective Young's modulus values were used for assigning material properties of porous structures while performing FEA analysis. Poisson's ratio value is taken 0.30 for each porous structure (Zhang et al.

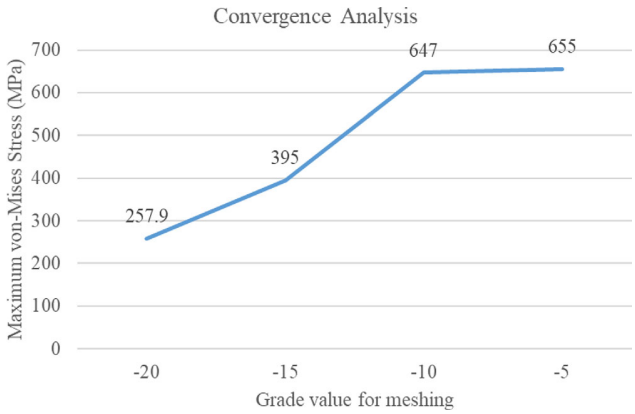


Figure 6. Graph of grade value for meshing vs maximum von-Mises stress.

Table 2. Numbers of nodes and elements for all the studied assembly models for FEA.

Assembly model type	Number of nodes	Number of elements
Solid spinal cage assembly	698,049	2,924,111
Hybrid spinal cage assembly (Porous structure: Gyroid, Pore size: 0.4 mm)	1,436,793	5,671,617
Hybrid spinal cage assembly (Porous structure: Gyroid, Pore size: 0.5 mm)	1,340,727	5,170,617
Hybrid spinal cage assembly (Porous structure: Gyroid, Pore size: 0.6 mm)	1,238,856	4,613,916
Hybrid spinal cage assembly (Porous structure: Diamond, Pore size: 0.4 mm)	1,346,228	5,293,319
Hybrid spinal cage assembly (Porous structure: Diamond, Pore size: 0.5 mm)	1,343,925	5,094,849
Hybrid spinal cage assembly (Porous structure: Diamond, Pore size: 0.6 mm)	1,382,151	5,120,517
Hybrid spinal cage assembly (Porous structure: Schwarz W, Pore size: 0.4 mm)	1,344,770	5,407,024
Hybrid spinal cage assembly (Porous structure: Schwarz W, Pore size: 0.5 mm)	1,253,088	4,928,177
Hybrid spinal cage assembly (Porous Structure: Schwarz W, Pore Size: 0.6 mm)	1,220,761	4,720,936

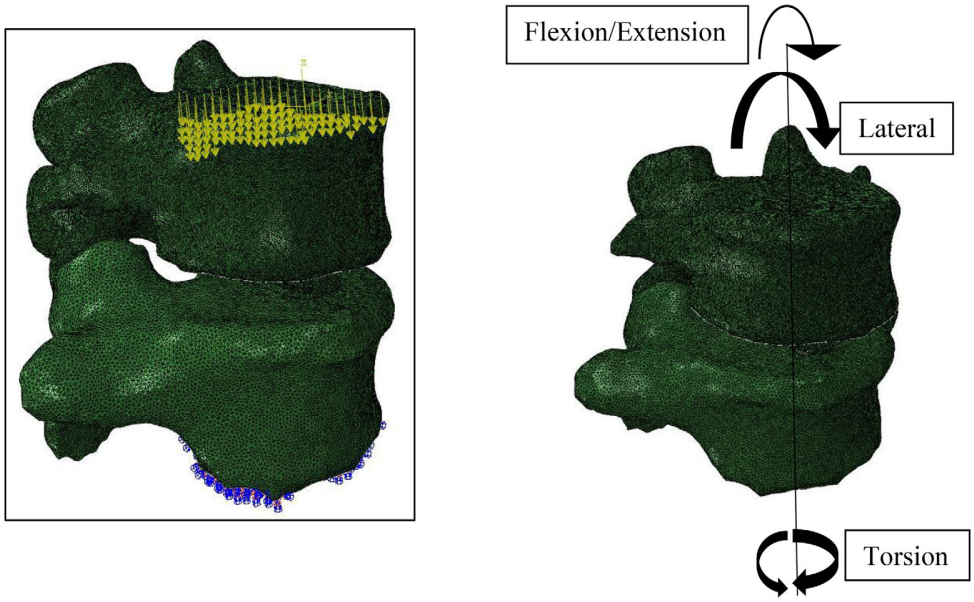


Figure 7. Loading type and boundary conditions for (a) uniaxial compression, (b) flexion/extension, lateral bending, and torsion.

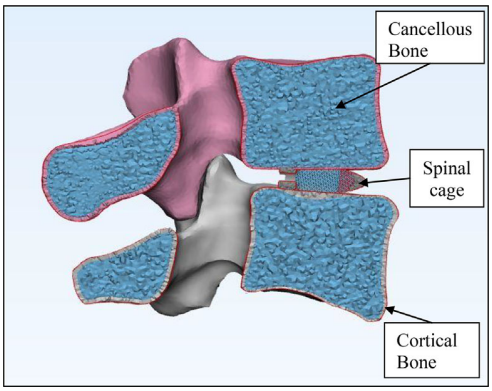


Figure 8. Section cut of the assembly model.

2018). All other material properties used for performing FEA are listed in Table 5.

2.5. Fatigue analysis

With the help of fe-safe 2020 software (Dassault Systèmes Simulia Corp., RI, USA), compressive fatigue analysis of the hybrid spinal cages with the least stress shielding has been simulated for 3000 N at 5,000,000 cycles as per ASTM 2077 (ASTM F2077 2020; Guidance for Industry 2020). The Brown-Miller algorithm and the Morrow mean stress criterion have been used for performing the fatigue simulation. The Brown-Miller approach calculates the probability of fatigue by using standard material data and critical plane techniques to obtain accurate stress-strain spectra for non-uniform loading situations (fe-safe 2019 2018). This algorithm assumes that the plane subjected to the maximum shear stress amplitude

Table 3. Fabricated samples of studied porous structures and pore sizes.

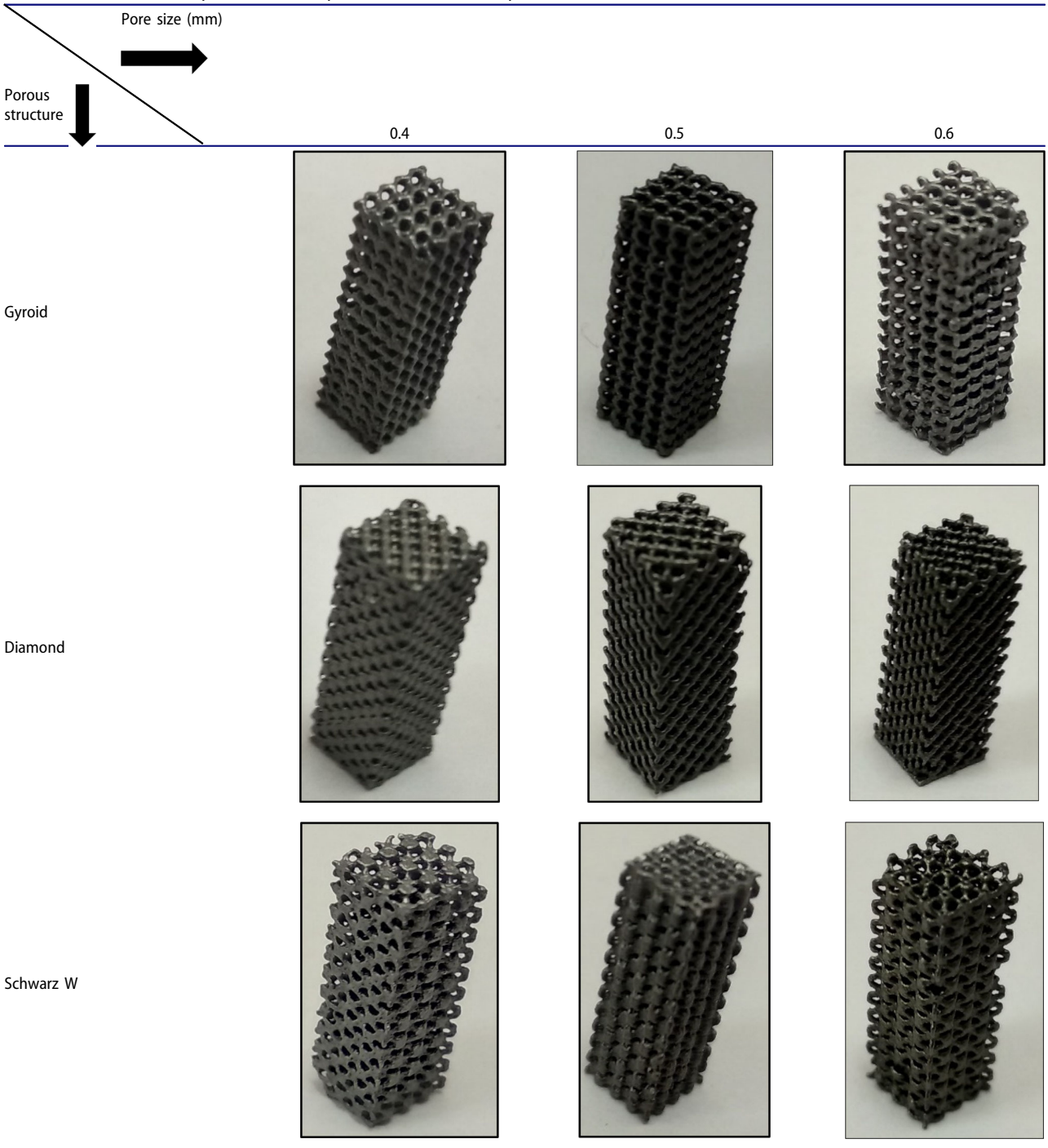


Table 4. Material properties of porous samples.

Structure	Pore size (mm)	Young's modulus (MPa)
Gyroid	0.40	3353.83
	0.50	2757.51
	0.60	2549.97
Diamond	0.40	3275.31
	0.50	2696.88
	0.60	2423.43
Schwarz W	0.40	3237.93
	0.50	2677.72
	0.60	2411.15

experiences the most fatigue damage and that fatigue damage is caused by the maximum shear strain and the co-action of the principal strain in this plane. It is required to alter the algorithm to account for the influence of average stress and strain. The Morrow method corrects the elastic stress by subtracting the average stress per cycle, bringing the projected fatigue life closer to the real fatigue life (Morrow 1968). The minimum fatigue life cycle is calculated from the log-

Table 5. Material properties of components in FE models (Shirazi-Adl et al. 1986; Chosa et al. 2004; Zhong et al. 2006; Shirazi-Adl et al. 2012).

Components	Young's modulus (MPa)	Poisson ratio	References
Cortical bone	12,000.00	0.30	Wu et al. (2013); Zhang et al. (2018)
Cancellous bone	100.00	0.20	Wu et al. (2013); Zhang et al. (2018)
Titanium	110,000.00	0.30	Frosch et al. (2004); Zhu et al. (2020)

Table 6. Maximum values of von-mises stress and deformation on the solid cage for different load conditions.

Type of load	Maximum stress on full cage (MPa)	Maximum stress on L5 (MPa)	Maximum displacement of full cage (μm)	Maximum displacement of L5 (μm)
Uniaxial compression	260.70	16.29	9.63	6.36
Flexion/Extension	180.70	13.99	6.31	20.20
Bending	84.40	11.73	1.54	19.18
Torsion	114.00	12.91	8.53	15.48

Table 7. Maximum values of von-mises stress and deformation on the gyroid porous structured hybrid cage and L5 vertebra for different load conditions.

Pore size	Type of load	Maximum stress on full cage (MPa)	Maximum stress on L5 (MPa)	Maximum displacement of full cage (μm)	Maximum displacement of L5 (μm)
0.4	Uniaxial Compression	255.50	39.39	33.54	7.93
	Flexion/Extension	223.80	21.77	19.63	21.16
	Bending	93.79	16.26	3.95	19.08
	Torsion	174.10	17.08	20.27	16.15
0.5	Uniaxial Compression	305.50	42.44	42.47	8.47
	Flexion/Extension	261.90	24.07	25.04	21.35
	Bending	96.29	18.04	4.08	19.10
	Torsion	219.30	19.79	21.55	16.15
0.6	Uniaxial Compression	326.30	45.11	53.51	8.82
	Flexion/Extension	293.00	26.07	31.60	21.50
	Bending	98.16	19.51	4.14	19.09
	Torsion	239.50	23.28	22.96	16.11

Table 8. Maximum values of von-mises stress and deformation on the diamond porous structured hybrid cage and L5 vertebra for different load conditions.

Pore size	Type of load	Maximum stress on full cage (MPa)	Maximum stress on L5 (MPa)	Displacement of full cage (μm)	Displacement of L5 (μm)
0.4	Uniaxial Compression	274.40	39.70	39.65	8.20
	Flexion/Extension	222.50	21.94	27.55	21.30
	Bending	94.38	15.28	6.19	19.09
	Torsion	173.00	17.74	20.65	16.14
0.5	Uniaxial Compression	316.40	42.59	48.12	8.89
	Flexion/Extension	251.50	25.42	32.53	21.73
	Bending	97.21	16.28	4.11	19.09
	Torsion	222.40	19.94	22.27	16.14
0.6	Uniaxial Compression	329.90	46.25	56.50	9.25
	Flexion/Extension	282.60	26.68	33.19	21.74
	Bending	98.62	18.05	4.16	19.09
	Torsion	242.00	23.66	23.24	16.14

Table 9. Maximum values of von-mises stress and deformation on Schwarz W porous structured hybrid cage and L5 vertebra for different load conditions.

Pore size	Type of load	Maximum stress on full cage (MPa)	Maximum stress on L5 (MPa)	Displacement of full cage (μm)	Displacement of L5 (μm)
0.4	Uniaxial Compression	306.50	39.88	34.11	8.06
	Flexion/Extension	226.60	22.04	19.96	21.23
	Bending	93.69	18.03	3.96	19.08
	Torsion	174.30	17.64	20.35	16.16
0.5	Uniaxial Compression	325.50	42.79	45.55	8.69
	Flexion/Extension	267.10	24.21	26.65	21.49
	Bending	96.68	19.50	4.09	19.09
	Torsion	222.60	19.95	21.94	16.15
0.6	Uniaxial Compression	334.00	45.41	53.87	9.17
	Flexion/Extension	284.20	26.04	31.58	21.71
	Bending	98.18	20.60	4.13	19.09
	Torsion	242.40	23.69	22.90	16.14

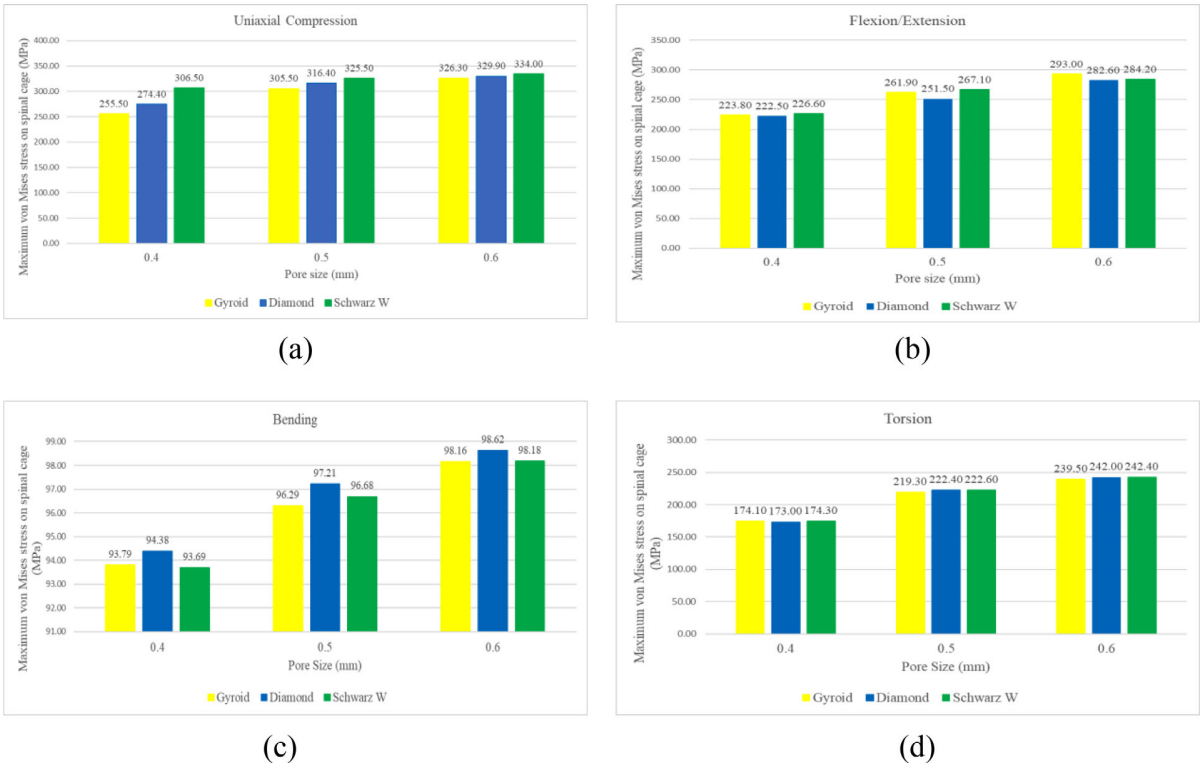


Figure 9. Comparative graphical representation of maximum von-Mises stress on the hybrid spinal cage with different pore structures under load type of (a) compression, (b) flexion/extension, (c) bending, (d) torsion.

life repeats obtained from the software using the following formula (fe-safe 2019: SIMULIA-dassault systemes 2018):

$$Fatigue\ Life\ Cycle = \frac{10^{(Log-Life\ \ Repeats)}}{2} \qquad (1)$$

3. Results

FEA analysis was performed with solid as well as hybrid cages with different pore structures and different pore sizes in Abaqus software. The results showed that the hybrid PLIF spinal cages reduce the stress shielding effect compared to the solid PLIF cage, which complies with previous literature (Zhang et al. 2018). The FEA results of different cages are discussed below:

3.1. Solid PLIF cage

The maximum von-Mises stress and deformation values for different loading conditions are listed in Table 6. The stress and deformation distribution results on the solid cage and L5 vertebrae obtained from FEA are shown in Supplementary File II, Table 1. The maximum stress on the cage is approximately 260.7 MPa and the maximum displacement of the cage is 9.63 µm for uniaxial compression. The maximum stress transferred to L5 due to uniaxial compression load application was observed to be 16.29 MPa and the maximum displacement of the L5 vertebra is 6.36 µm.

3.2. Hybrid PLIF cages

Under similar loading conditions, FEA analysis of hybrid porous PLIF cages with different structure and pore sizes were performed. The results are summarized as follows:

A. **Gyroid porous structured PLIF cage:** The maximum values of von-Mises stress and deformation on the cage and adjacent L5 vertebra are listed in Table 7. Pictorial representation of von-Mises stress and deformation distribution on gyroid porous structured PLIF cage and adjacent L5 vertebra is shown in Supplementary File II, Table 2.
B. **Diamond porous structured PLIF cage:** The maximum values of von-Mises stress and deformation on the cage and adjacent L5 vertebra are listed in Table 8. Pictorial representation of von-Mises stress and deformation distribution on

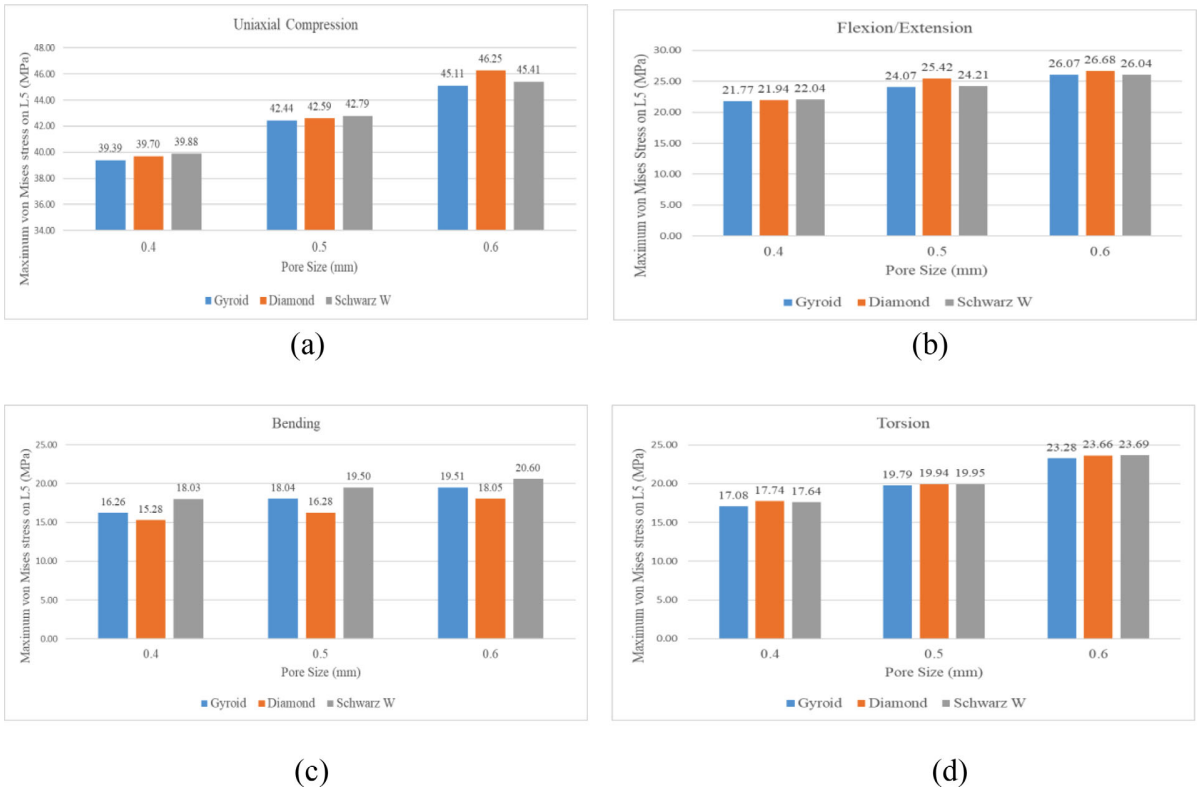


Figure 10. Comparative graphical representation of maximum von-Mises stress on L5 vertebra with different pore structures under load type of (a) compression, (b) flexion/extension, (c) bending, (d) torsion.

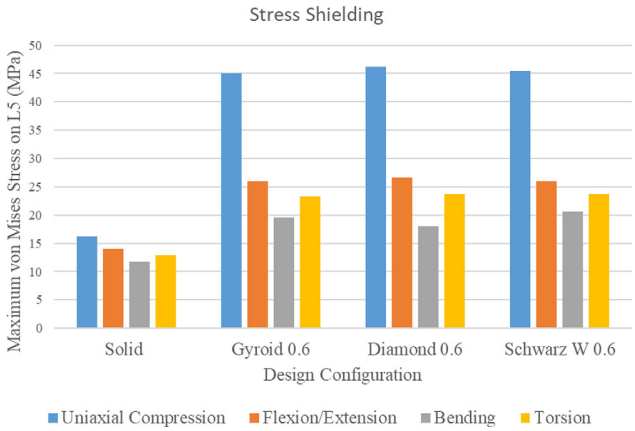


Figure 11. Reduced stress shielding by hybrid spinal cages as compared with solid spinal cage.

diamond porous structured PLIF cage and adjacent L5 vertebra is shown in Supplementary File II, Table 3.

C. **Schwarz W porous structured PLIF cage:** The maximum values of von-Mises stress and deformation on the cage and adjacent L5 vertebra are listed in Table 9. Pictorial representation of von-Mises stress and deformation distribution on Schwarz W porous structured PLIF cage and adjacent L5 vertebra is shown in Supplementary File II, Table 4.

4. Discussion

The biomechanical effect of various hybrid spinal cages on stresses on the cage and adjacent vertebra in four loading modes, namely uniaxial compression, flexion/extension, bending, and torsion, was investigated using a FE model. Such an FE investigation is not plausible in a cadaver model as to estimate the stresses in various spinal structures is not feasible (Vadapalli et al. 2006).

A considerable Young's modulus mismatch between the metal implant and the adjacent natural

Table 10. Minimum Log-life repeat values as derived from FEA results and minimum fatigue life of PLIF cages.

PLIF cage type	Minimum Log-life repeats	Minimum fatigue life in cycles
Gyroid porous structured with a pore size of 0.6 mm	7.106	6,382,194
Diamond porous structured with a pore size of 0.6 mm	7.204	7,997,790
Schwarz W porous structured with a pore size of 0.6 mm	7.065	5,807,243

bone causes an irregular stress distribution at the bone-implant interface, resulting in a stress shielding effect (Alvarez and Nakajima 2009; Yan et al. 2015). Stress shielding increases the risk of fracture and implant loosening by causing bone tissue absorption around the metal implant (Huiskes et al. 1992). As a result, the ideal orthopedic implant material has Young's modulus that is similar to human bone tissue. The porous structures used in this study have this property similar to that of the natural human bone which is in the range of 0.5–20 GPa (Bonfield et al. 1998). Junchao et al. conducted a study on the compressive strength of titanium materials based on porosity and found that when porosity increases, the elasticity increases too (Junchao et al. 2016). The porous titanium alloy structures in this study can be identified with similar results that have the high elasticity to reduce the stress shielding effect.

A graphical representation of maximum von-Mises stress values on hybrid spinal cages with studied porous structures and pore sizes, and on the L5 vertebra under all types of load conditions have been shown in Figures 9 and 10 for easy interpretation. It can be visualized from the figures that as the pore size increases, the stress transfer to the L5 vertebra increases which helps in reducing stress shielding caused due to the implant. The hybrid spinal cages with a pore size of 0.6 mm showed high elasticity and reduced stress shielding effect. This result is consistent throughout regardless of the type of the porous structures.

Solid spinal cage fusion has been demonstrated in certain studies to have a high-stress shielding effect, which may hasten intervertebral disc degeneration at surrounding levels (Kim et al. 2014; Zhang et al. 2016). A comparative graph of stress transferred on the L5 vertebra in the case of the solid cage and hybrid spinal cages having a pore size of 0.6 mm with different loading conditions is shown in Figure 11. The graph clearly shows a reduction in stress shielding on adjacent vertebra with hybrid spinal cages as compared to the conventional solid spinal cage.

From these results, it can be predicted that if the pore size is same, the difference in porous structures does not significantly influence the load-bearing strength and the stress shielding effect of the hybrid spinal cage. The pore size of 0.6 mm is found to be the best suited for reducing the stress shielding effect by transferring the maximum stress from L4 to L5 under various types of load applications. In vitro investigations by Frosch et al. revealed that porous materials with pore sizes of 0.4–0.6 mm had more osteoblasts than those with pore sizes of 0.3 mm and 1 mm (Frosch et al. 2004).. Taniguchi et al. in their in vivo experiments also found that samples with a pore size of 0.6 mm were best suited for bone-tissue ingrowth among the samples with pore sizes 0.3/0.6/0.9 mm (Taniguchi et al. 2016). Large pore sizes are more favorable to capillary development and nutrition penetration, but they may compromise the implant's mechanical qualities (Li et al. 2019). As a result, the pore size of 0.6 mm is simply the point at which various influencing factors are balanced (Taniguchi et al. 2016).

Compressive fatigue testing simulation as per ASTM F2077 to determine the performance of hybrid porous spinal cages of 0.6 mm pore size was performed. Table 10 listed the minimum values of log life repeats and fatigue cycles. The log-life repeats representations are shown in (Supplementary File II, Figure 1). There was no change or breakage in the cages after 3000 N to 5,000,000 cycles. So, the hybrid spinal cages can be considered structurally stable, since they sustained 5,000,000 cycles under the standardized load.

The use of a unique lumbar model, simplifying the material properties of some components, and disregarding the role of muscles are some limitations of this finite element study. The geometric model of the lumbar spine, such as disc height and joint space, differs from person to person. But in this study, just one lumbar spine model is used. Secondly, although the lumbar spine's components are nonlinear in nature, their material properties were simplified in this work to be linear elastic. However, to improve calculation efficiency, many FEA on the lumbar spine assumed that the components were linear (Grauer et al. 2006; Vadapalli et al. 2006; Zhong et al. 2006; Kim et al. 2010; Choi et al., 2017; Zhang et al. 2018).

5. Conclusion

PLIF spinal cages with three different structures Gyroid, Diamond, and Schwarz W having pore sizes

0.4 mm, 0.5 mm, and 0.6 mm were designed and studied by subjecting different loading conditions such as uniaxial compression, flexion/extension, bending, and torsion on vertebra-spinal cage assembly using FEA. The results were observed for the stress shielding effect due to implantation by studying the maximum stress transfer from the L4 vertebra to L5 vertebrae. With an increase in pore size, the stress shielding effect is decreasing due to the reduced stiffness of the cage. From the results obtained in this study, it can be safely predicted that if the pore size is same, the difference in porous structures does not significantly influence the load-bearing strength and the stress shielding effect of the spinal cage. PLIF spinal cages with 0.6 mm pore size were found to have the ability to reduce the stress shielding effect in all the studied structures and it has the advantage of better osseointegration as discussed in previous studies. All the spinal cages having a pore size of 0.6 mm fulfilled the requirement of relevant FDA standard fatigue life. This study will help neuro/orthopedic surgeons in the selection of appropriate spinal cages in terms of pore sizes and pore structures.

6. Future work

To fully satisfy the mechanical as well as biological requirements of a spinal cage, bone ingrowth studies in samples need to be clinically validated using animal models.

Acknowledgment

The authors acknowledge the support provided by CSIR, New Delhi, India.

Disclosure statement

The authors declare there is no conflict of interest.

References

Alvarez K, Nakajima H. 2009. Metallic scaffolds for bone regeneration. Materials. 2(3):790–832.

Ambati DV, Wright EK, Jr, Lehman RA, Jr, Kang DG, Wagner SC, Dmitriev AE. 2015. Bilateral pedicle screw fixation provides superior biomechanical stability in transforaminal lumbar interbody fusion: a finite element study. Spine J. 15(8):1812–1822.

ASTM F2077: Testing intervertebral body fusion devices. 2020. [accessed 2020 Sep 9]. https://www.element.com/nucleus/2019/11/21/19/25/astm-f2077-testing-ibfds#:~:text=ASTM%20F2077%20is%20the%20standard,properties%20of%20intervertebral%20spinal%20implants.

Barth E, Ronningen H, Solheim LF, Saethren B. 1986. Bone ingrowth into weight-bearing porous fiber titanium implants. Mechanical and biochemical correlations. J Orthop Res. 4(3):356–361.

Bobyn JD, Pilliar RM, Cameron HU, Weatherly GC. 1980. The optimum pore size for the fixation of porous-surfaced metal implants by the ingrowth of bone. Clin Orthopaed Relat Res. 150:263–270.

Bobyn JD, Stackpool GJ, Hacking SA, Tanzer M, Krygier JJ. 1999. Characteristics of bone ingrowth and interface mechanics of a new porous tantalum biomaterial. J Bone Joint Surg Br Vol. 81-B(5):907–914.

Bonfield W, Wang M, Tanner K. 1998. Interfaces in analogue biomaterials. Acta Mater. 46(7):2509–2518.

Chen J-H, Liu C, You L, Simmons CA. 2010. Boning up on Wolff's Law: mechanical regulation of the cells that make and maintain bone. J Biomech. 43(1):108–118.

Choi J, Shin D-A, Kim S. 2017. Biomechanical effects of the geometry of ball-and-socket artificial disc on lumbar spine: a finite element study. Spine (Phila Pa 1976). 42(6):E332–E339.

Choi K-C, Ryu K-S, Lee S-H, Kim YH, Lee SJ, Park C-K. 2013. Biomechanical comparison of anterior lumbar interbody fusion: stand-alone interbody cage versus interbody cage with pedicle screw fixation - a finite element analysis. BMC Musculoskelet Disord. 14(1):220–229.

Chosa E, Goto K, Totoribe K, Tajima N. 2004. Analysis of the effect of lumbar spine fusion on the superior adjacent intervertebral disk in the presence of disk degeneration, using the three-dimensional finite element method. Clin Spine Surg. 17(2):134–139.

Chuah HG, Rahim IA, Yusof MI. 2010. Topology optimisation of spinal interbody cage for reducing stress shielding effect. Comput Methods Biomech Biomed Eng. 13(3):319–326.

Clemow AJT, Weinstein AM, Klawitter JJ, Koeneman J, Anderson J. 1981. Interface mechanics of porous titanium implants. J Biomed Mater Res. 15(1):73–82.

Cook SD, Walsh KA, Haddad RJ. 1985. Interface mechanics and bone-growth into porous Co-Cr-Mo alloy implants. Clin Orthopaed Relat Res. 193:271–280.

Epasto G, Distefano F, Mineo R, Guglielmino E. 2019. Subject-specific finite element analysis of a lumbar cage produced by electron beam melting. Med Biol Eng Comput. 57(12):2771–2781.

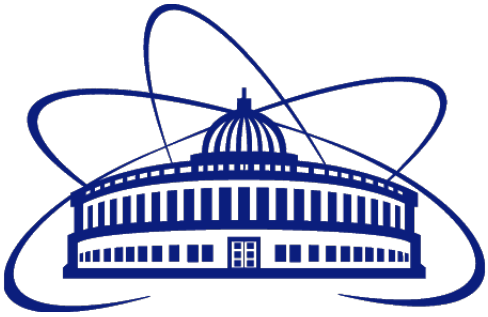
fe-safe 2019 FATIGUE THEORY REFERENCE: Dassault Systemes. pp. 123–176. [Online]. Available: https://help.3ds.com/2019/English/DSSIMULIA_Established/FesafeFatigueTheoryPdf/FesafeFatigueTheory.pdf. [Accessed 2020 Feb 26].

Frosch K-H, Barvencik F, Viereck V, Lohmann CH, Dresing K, Breme J, Brunner E, Stürmer KM. 2004. Growth behavior, matrix production, and gene expression of human osteoblasts in defined cylindrical titanium channels. J Biomed Mater Res. 68A(2):325–334.

Goel V, Monroe B, Gilbertson L, Brinckmann P. 1995. Interlaminar shear stresses and laminae separation in a disc: finite element analysis of the L3-L4 motion segment subjected to axial compressive loads. Spine. 20(6):689–698.

Grauer JN, Biyani A, Faizan A, Kiapour A, Sairyo K, Ivanov A, Ebraheim NA, Patel TC, Goel VK. 2006. Biomechanics of two-level Charité artificial disc

placement in comparison to fusion plus single-level disc placement combination. Spine J. 6(6):659–666.
Guidance for Industry and FDA Staff Spinal System 510(k)s. 2020. [accessed 2020 Sep 9]. https://www.fda.gov/media/71604/download.
Haibo H, Huiqun L, Jieen Danqing W, Shang F, Wuling S. 2012. Research progress of biomedical porous titanium and its alloys. Mater Rev. 26(19):262–270.
Herrera A, Yánez A, Martel O, Afonso H, Monopoli D. 2014. Computational study and experimental validation of porous structures fabricated by electron beam melting: a challenge to avoid stress shielding. Mater Sci Eng C Mater Biol Appl. 45:89–93.
Hofmann AA, Bloebaum RD, Bachus KN. 1997. Progression of human bone ingrowth into porous-coated implants. Rate of bone ingrowth in humans. Acta Orthop Scand. 68(2): 161–166. 1997/01/01
Huiskes R, Weinans H, Van Rietbergen B. 1992. The relationship between stress shielding and bone resorption around total hip stems and the effects of flexible materials. Clin Orthopaed Relat Res. 274:124–134.
Itälä AI, Ylänen HO, Ekholm C, Karlsson KH, Aro HT. 2001. Pore diameter of more than 100 microm is not requisite for bone ingrowth in rabbits. J Biomed Mater Res. 58(6):679–683.
Junchao L, Yanyan Z, Wei W. 2016. Elastic modulus and stress analysis of porous titanium parts fabricated by selective laser melting. J Harbin Inst Technol. 23:46–50.
Kanayama M, Cunningham BW, Haggerty CJ, Abumi K, Kaneda K, McAfee PC. 2000. In vitro biomechanical investigation of the stability and stress-shielding effect of lumbar interbody fusion devices. J Neurosurg. 93(2 Suppl):259–265.
Kim H-J, Kang K-T, Chang B-S, Lee C-K, Kim J-W, Yeom JS. 2014. Biomechanical analysis of fusion segment rigidity upon stress at both the fusion and adjacent segments: a comparison between unilateral and bilateral pedicle screw fixation. Yonsei Med J. 55(5):1386–1394.
Kim K-T, Lee S-H, Suk K-S, Lee J-H, Jeong B-O. 2010. Biomechanical changes of the lumbar segment after total disc replacement: charite(r), prodisc(r) and maverick(r) using finite element model study. J Korean Neurosurg Soc. 47(6):446–453.
Kurutz M. 2010. Finite element modelling of human lumbar spine. Finite Element Analysis 209–236. https://www.intechopen.com/chapters/11994.
Lee Y-H, Chung C-J, Wang C-W, Peng Y-T, Chang C-H, Chen C-H, Chen Y-N, Li C-T. 2016. Computational comparison of three posterior lumbar interbody fusion techniques by using porous titanium interbody cages with 50% porosity. Comput Biol Med. 71:35–45..
Li P, Jiang W, Yan J, Hu K, Han Z, Wang B, Zhao Y, Cui G, Wang Z, Mao K, et al. 2019. A novel 3D printed cage with microporous structure and in vivo fusion function. J Biomed Mater Res A. 107(7):1386–1392.
Melikyan ML, Itin VI. 2002. Dynamics of bone tissue mineralization in porous titanium and the mechanical properties of a titanium-bone tissue composite. Tech Phys Lett. 28(8):673–674.
Morrow J. 1968. Fatigue properties of metals. In: Richard C. Rice (Ed.), Fatigue design handbook; p. 21–30. USA: Society of Automotive Engineers.
Pannell WC, Savin DD, Scott TP, Wang JC, Daubs MDJTSJ. 2015. Trends in the surgical treatment of lumbar spine disease in the United States. 15(8):1719–1727.
Ramakrishna S, Mayer J, Wintermantel E, Leong KW. 2001. Biomedical applications of polymer-composite materials: a review. Compos Sci Technol. 61(9):1189–1224.
Shirazi-Adl A, Ahmed AM, Shrivastava SC. 1986. Mechanical response of a lumbar motion segment in axial torque alone and combined with compression, (in Eng). Spine (Phila Pa 1976). 11(9):914–927.
Stock JT. 2018. Wolff's law (bone functional adaptation). In: Wenda Trevathan, Matt Cartmill, Darna L. Dufour, Clark Spencer Larsen, Dennis H. O'Rourke, Karen Rosenberg, Karen B. Strier (Eds.), The international encyclopedia of biological anthropology. p. 1–2. USA: Wiley Online Library.
Taniguchi N, Fujibayashi S, Takemoto M, Sasaki K, Otsuki B, Nakamura T, Matsushita T, Kokubo T, Matsuda S. 2016. Effect of pore size on bone ingrowth into porous titanium implants fabricated by additive manufacturing: an in vivo experiment. Mater Sci Eng C Mater Biol Appl. 59:690–701.
Tsuruga E, Takita H, Itoh H, Wakisaka Y, Kuboki Y. 1997. Pore size of porous hydroxyapatite as the cell-substratum controls BMP-induced osteogenesis. J Biochem. 121(2): 317–324.
Vadapalli S, Sairyo K, Goel VK, Robon M, Biyani A, Khandha A, Ebraheim NA. 2006. Biomechanical rationale for using polyetheretherketone (PEEK) spacers for lumbar interbody fusion-a finite element study. Spine (Phila Pa 1976). 31(26):E992–E998.
Wu S-H, Li Y, Zhang Y-Q, Li X-K, Yuan C-F, Hao Y-L, Zhang Z-Y, Guo Z. 2013. Porous titanium-6 aluminum-4 vanadium cage has better osseointegration and less micromotion than a poly-ether-ether-ketone cage in sheep vertebral fusion. Artif Organs. 37(12):E191–E201. [24147953]
Xiao Z, Wang L, Gong H, Zhu D. 2012. Biomechanical evaluation of three surgical scenarios of posterior lumbar interbody fusion by finite element analysis. Biomed Eng Online. 11(1):31.
Yan C, Hao L, Hussein A, Young P. 2015. Ti-6Al-4V triply periodic minimal surface structures for bone implants fabricated via selective laser melting. J Mech Behav Biomed Mater. 51:61–73.
Zhang QH, Zhou YL, Petit D, Teo EC. 2009. Evaluation of load transfer characteristics of a dynamic stabilization device on disc loading under compression. Med Eng Phys. 31(5):533–538.
Zhang Z, Li H, Fogel GR, Liao Z, Li Y, Liu W. 2018. Biomechanical analysis of porous additive manufactured cages for lateral lumbar interbody fusion: a finite element analysis. World Neurosurg. 111:e581–e591.
Zhang Z, Sun Y, Sun X, Li Y, Liao Z, Liu W. 2016. Recent advances in finite element applications in artificial lumbar disc replacement. JBiSE. 09(10):1–8.
Zhong Z-C, Wei S-H, Wang J-P, Feng C-K, Chen C-S, Yu C-h. 2006. Finite element analysis of the lumbar spine with a new cage using a topology optimization method. Med Eng Phys. 28(1):90–98.
Zhu H, Zhong W, Zhang P, Liu X, Huang J, Liu F, Li J. 2020. Biomechanical evaluation of autologous bone-cage in posterior lumbar interbody fusion: a finite element analysis. BMC Musculoskelet Disord. 21(1):379.



JOINT INSTITUTE FOR NUCLEAR RESEARCH

Veksler and Baldin laboratory of High Energy Physics

FINAL PROJECT REPORT ON THE INTEREST PROGRAMME

Generation and analysis of events for Heavy Ions collisions using the MC generator - Therminator 2

Anish Kalsi

Bachelor of Technology, Engineering Physics,

Delhi Technological University, India

Supervisor: **Mr. Krystian Roslon**

24 MAY - 2 July 2021, WAVE 4

Abstract

The Study of heavy-ion collisions can help us unravel the history of the Universe and understand the properties of matter moments after the Big Bang. Quark-Gluon plasma is believed to be formed in the early Universe and can be found in extreme Astrophysical conditions. In this project, we have used the same QGP model to study some heavy ion collisions. Using Therminator 2 we have generated events for Au-Au using hypersurfaces and velocity profiles describing data obtained by RHIC at $\sqrt{s_{NN}} = 200GeV$ and at centrality 40%-50%. Then we analysed the femtoscopic correlation function for identical particle pairs like pions ($\pi^+\pi^+$ and $\pi^-\pi^-$), kaons (K^+K^+ and K^-K^-) and protons, Also for non-identical particle pairs pions-kaons (π^+K^- and $K^+\pi^-$), pions-protons (π^+p and π^-p) and Kaons-protons (K^+p and K^-p).

Acknowledgement

I would like to extend my gratitude towards my project supervisor, **Mr. Krystian Roslon**, who guided and helped me undertake this project, especially during these difficult times. I would like to take this opportunity to thank him for his invaluable guidance, constant encouragement, and comments, which help sustained my efforts at all stages of this project work.

I would like to thank my fellow project participant, **Kitzia Michelle Curiel**, for being such a sport, helping and motivating me through out.

A sincere thanks to **JINR INTEREST Team** also, to introduce this program and for providing me and students all around the globe with this opportunity and for the smooth overseeing of the program. Over the course of our project work I have gained immense knowledge which would help us in the coming years.

Contents

Abstract **i**

Acknowledgement **ii**

1 Introduction **1**

1.1 High Energy Physics 1

1.2 Quark-Gluon Plasma 1

1.3 Heavy Ion Collision 2

1.4 Femtoscopy 3

1.4.1 Correlation Function 4

2 THERMINATOR 2 **5**

2.1 Installation 5

2.1.1 Events Generation 6

3 Femtoscopic Analysis using Therminator 2 **7**

3.1 Analysis of Generated data 8

3.1.1 Identical Particles 10

3.1.2 Non-identical Particles 11

3.2 ϕ Meson Observation 13

4 Conclusion **16**

References **17**

Chapter 1

Introduction

1.1 High Energy Physics

High Energy Physics is one of the vast domains of Physics, intending to determine the most fundamental building blocks of matter and the forces interacting through these particles. Relativistic Heavy Ion Collisions is the part of High Energy Physics which deals with Nuclear Physics of the colliding nucleus of atoms with higher atomic number, at relativistic speeds.

Such high energy levels can be achieved via colliding accelerated heavy ions and studying them. Collisions of such magnitude are carried on at two colliders around the world, the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory (BNL) in Upton, New York and the Large Hadron Collider (LHC) at the France-Switzerland border near Geneva. By studying of these collisions, recreate in the laboratory, we can understand nuclear matter state, close to the one that was present shortly after the Big Bang.

1.2 Quark-Gluon Plasma

Heavy ion collisions with the energy in the order of 10-100 A.GeV can also produce QGP for a very small amount of time [1]. These conditions can then be studied by different methods of analysis; one of them is called Femtoscopic Correlation. Different

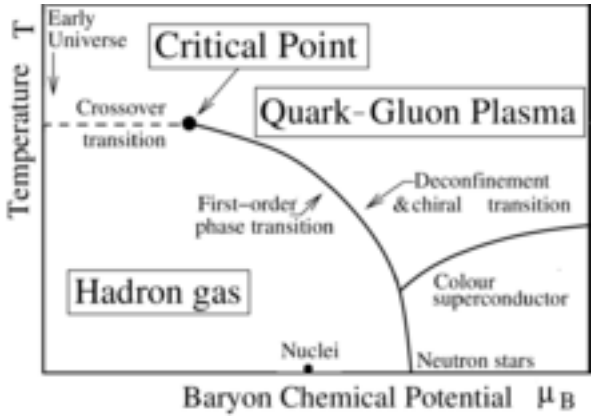


Figure 1.1: Phase diagram of Strong Interacting matter(QGP)

accelerators study the various domains of the QGP to get a better understanding of the early universe. Analyzing these particles that result from heavy ions collision represents a promising way to understand QGP and our universe's evolution, in addition to recognizing the dynamics of formation of matter. So it is generally considered that acceleration and collision of these nuclei is a simulation or regeneration for the same events that happened around 14 billion years ago, for sure this helps in discovering the secrets of the early universe.

1.3 Heavy Ion Collision

Heavy ions collision enables us to study the behaviour of matter under high density and pressure. At free collision of nuclei, there is no creation of new particles, as they must have the threshold value of energy ranging from small values of MeV to TeV for production of new particles. As the energy of collided particles increase, the number of particles and anti-particles produced will increase too. Collision type can be classified due to the value of the energy of collided ions. Intermediate heavy ions collision has range of 10 - 100 MeV, while relativistic has a range of 100 MeV - 10 GeV and ultra-relativistic that starts from 10 GeV and above, at 10 GeV the formation of QGP is possible [2].

Using statistical mechanics, we can describe matter as a system in thermal equilibrium, with global observables like net charge, temperature, and energy density. The system can display multiple phases of matter when these variables take on

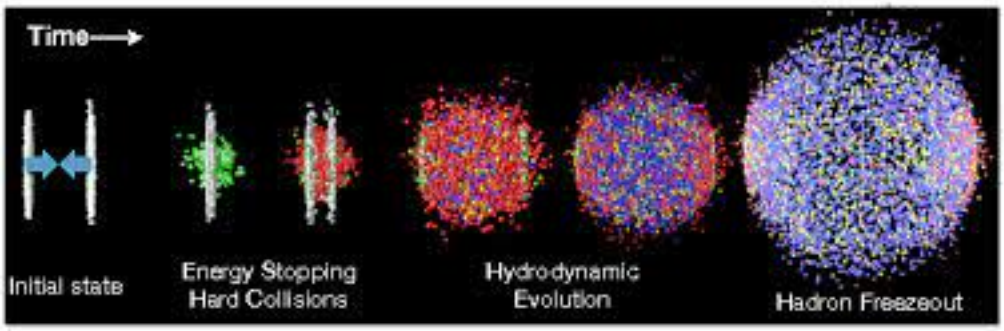


Figure 1.2: Relativistic Heavy-ion collision

changing values. A phase transition from hadronic matter to QGP should occur at high temperatures/densities.

1.4 Femtoscopy

Femtoscopy correlation is a technique used in relativistic heavy-ion collisions that can give us an insight of the space-time characteristics of particle production at the femtometer scale(10^{-15} m) using particle correlations [3]. For identical and non-identical particles, there are two sorts of correlations. The first can be used to investigate the size of two ion collision systems, and the non-identical ones are utilised to investigate the time scale of nucleus fragments created during the collision [4].

The centrality, which is closely related to the impact parameter and represents the size of the overlap region that corresponds to a different number of nucleons that participate in the collision (participants) and those that do not, is another important physical concept used in the analysis of heavy-ion collisions (spectators) [5].However, because the impact parameter of a collision cannot be directly measured in experiments, but because it is an important metric for event selection and analysis, final state observables such as charged particle multiplicity and energy deposited by spectators are used to define centrality classes based on theoretical model predictions.The centrality class allows us to have an estimate of how many nucleons take part in the collision.

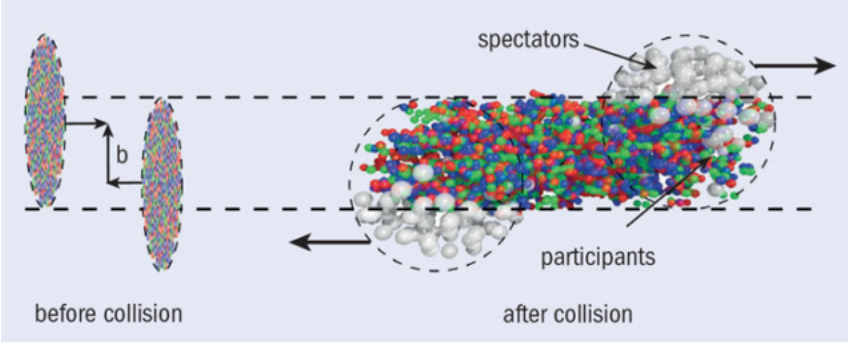


Figure 1.3: Semi-peripheral Heavy-ion collision

1.4.1 Correlation Function

The femtoscopic correlation function is a probabilistic function which determines how closely the particles are packed in the QGP. The correlation function for two identical particles is given by equation (29) in THERMINATOR2 manual [6]. The correlation function of two particles p1 and p2 can be given by the following equation:

$$C(p1, p2) = \frac{\mathcal{P}(p1, p2)}{\mathcal{P}(p1)\mathcal{P}(p2)} \tag{1.1}$$

Here, the numerator represents the measured momentum distribution difference ($q = p1 - p2$) of the two particles (p1,p2) from the same events and the denominator the distribution of two particles taken from different events.The results are normalized so the correlation function tends to 1 when there is no correlation between the particles.e the average transverse momentum of the pair will be denoted by K_T. Calculating this correlation function for different particle pairs was the main objective of this project.

Chapter 2

THERMINATOR 2

THERMINATOR 2 [6], **THERM**al heavy **I**oN gener**ATOR** 2, is a Monte Carlo event generator used to study the statistical production of subatomic and elementary particles in relativistic heavy ion collisions. It is written in C++ and using the standard CERN ROOT environment. This not only help us in model applications, but also, the code can be easily adapted for purposes directly linked to experimental data analysis, detector modeling, or estimates for the heavy-ion experiments at RHIC, LHC, SPS, FAIR, or NICA.

It uses a library of standard hypersurfaces and velocity profiles corresponding to Au-Au, Pb-Pb and p-Pb collisions at RHIC and LHC. The package includes the tools for femtoscopic analysis of events. Which can help us carry out analysis of correlation function of HBT radii of the generated events among other auxiliary results. A description of the tasks performed, from the installation of the program to the analysis of the results, is given in the following subsections.

2.1 Installation

THERMINATOR 2 uses the ROOT CERN Package, available for various operating systems. However, it is recommended to use Linux. Moreover, a C++ compiler is also needed. For this project, Ubuntu 18.04.5 was installed on a virtual machine, followed by the ROOT environment (the latest version-v6.24.00). A full guide on how to install ROOT is available at the official website [7].

The working and installation, as well as the latest version THERMINATOR 2,

can be found at the official website. The major steps regarding the installation and setup of the program are clearly given in the manual [6], but one must be aware that the last version of THERMINATOR 2 is a decade old and the installation requires some modifications to the code. Once the documentation has been downloaded and unzipped, before compiling using **make** command, add the line **using namespace std;** in the **build/src/therm2_events.cxx** file after all the function declaration. Also, the 119,123,127 lines of **Makefile** in the main directory in the given form, otherwise it will hinder with the compilation process.

$\$(LD)\ \$(LFLAGS)\ \$\hat{}\ -o\ \$@\ to\ \$(LD)\ \$\hat{}\ -o\ \$@\ \$(LFLAGS)$

2.1.1 Events Generation

To check, if the installation and setup was done properly, one can a test run script called ***./runall.sh***. This command run displays all basic functionality of the package, including the femtoscopic analysis, producing the plots with the physical results and more.

In this project, I generated Au-Au collision, using data from the RHIC, using command ***./Therm2_events***. Its important characteristics are the energy of the collision, which is at $\sqrt{s_{NN}} = 200GeV$ and the centrality (40%-50%), total 20,000 events were generated and 40 event*.root files were obtained, containing data of 500 events per file. Other events are available as well. To generate a type of events other than that set by default, one must open the **therminator2/fomodel/lhyquid2dbi** folder and pick the file with the wanted parameters. Copy and paste the name of the file in line 42 of the **fomodel/lhyquid2dbi.ini** file and run the event. A new folder bearing the name of the file will be created.

It is important to generate enough events to obtaining good statistics. The number of events can be increased either by running the same event several times, or directly by modifying the Number of Events line 55 in the **events.ini** file. In this case, I have generated 20,000 events (20,000 - 50,000 events are preferable).

Chapter 3

Femtoscopic Analysis using Therminator 2

The THERMINATOR2 package includes a femtoscopic analysis tool called **therm2_femto** which can be used to calculate the femtoscopic correlation function for identical and non-identical particle pairs [8]. After generating the events as described before, the output was saved in the events folder that contains multiple root files with the name event*.root with the information obtained from the simulation. These event files can be used to carry our multiple physical analysis. here we will use it to obtains correlation function of the acquired data.

The main source file that performs all the calculation is **therm2_femto.cxx**. All the physical processes according to the formalism mentioned earlier are defined there and some changes are necessary in order to run the program for different particle pairs. Some of those changes are mentioned in the following sections. The corresponding initialization file is **femto.ini** where some initial parameters are specified, like the type of particle pairs to correlate, number of particles to mix and enable the Coulomb Interactions.

The command to run femtosopic analysis is as following:

$./Therm2_femto <$ KTBIN > < Event Directory > < Total number of event files >

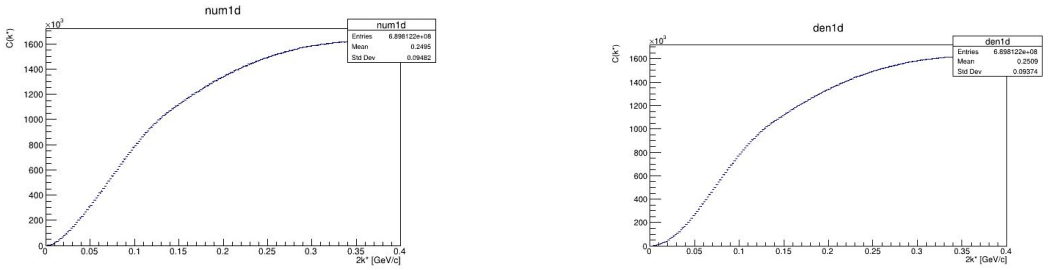
Where, the transverse momentum bin of the pair (k_T), KTBIN has the range from 0 to 3, where each value corresponds to a k_T value range:

- $k_T = 0 : (0.15, 0.25)$ GeV/c
- $k_T = 1 : (0.25, 0.35)$ GeV/c
- $k_T = 2 : (0.35, 0.45)$ GeV/c
- $k_T = 3 : (0.45, 0.6)$ GeV/c

The output files consist of the femto*.root files that contain the one-dimensional histograms of the numerator and denominator.

3.1 Analysis of Generated data

To obtain the correlation function we divided the 1D histograms for the denominator and numerator included on the **femto*.root** for each k_T bin. See fig 3.1 below.



(a) Numerator (b) Denominator

Figure 3.1: One dimensional histograms included in the femto*.root files (these are from femtopipi0a.root).

The division and graphs were done using a root macro. This division is conducted through the construction of a macro file written in C++ as following:

```
#include <iostream>
#include <fstream>
#include <sstream>
#include <TH1D.h>
#include <TH3D.h>
#include <TFile.h>
```

```
#include <TGraph.h>
#include <TPad.h>
#include <TCanvas.h>
#include <TImage.h>
#include <TMath.h>
#include <TDatime.h>
#include <math.h>
using namespace std;

TFile* tInRootFile;
TH1D* numq;
TH1D* denq;
TH1D* ratq;

void Correlation(){
tInRootFile = new
TFile("path of the root file/ name of the root file");
numq = new TH1D(*((TH1D *) tInRootFile->Get("num1d"))); //nominator
denq = new TH1D(*((TH1D *) tInRootFile->Get("den1d"))); //denominator
ratq = new TH1D(*numq);
ratq->Reset("ICE");
ratq->Divide(numq, denq, 1.0, 1.0); //Divison Process
ratq->SetName("Name");
ratq->SetTitle("Title");
ratq->Draw();   //Plotting
}
```

Which can be called using command **root correlation**, keep in mind, that the macro file and the void function must have same name. This code is very basic and can be modified to plot correlation function for all k_T bin values for a particular particle pair, which is been used in this project and resulting plots can be seen in next sections.

3.1.1 Identical Particles

It is possible to obtain the correlation function for both identical particles and non-identical particle pairs. Particles with the same type, mass and charge are referred as identical particles, like the pairs π^+-π^+, K^+-K^+, p-p. Identical particle femtoscopy can be useful to determine the relation between specific space-time information and the source radii. Below are the correlation functions obtained for identical particles.

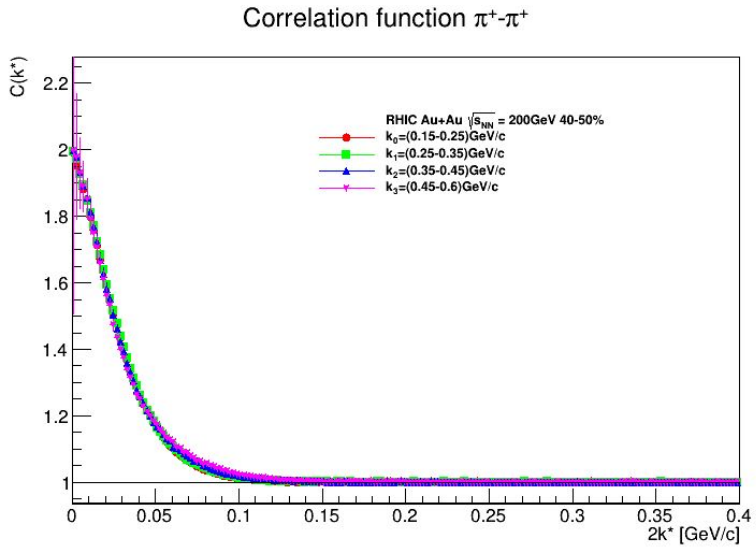


Figure 3.2: Correlation Function for Pion pairs (π^+-π^+ and π^--π^-) for various k_T bin values

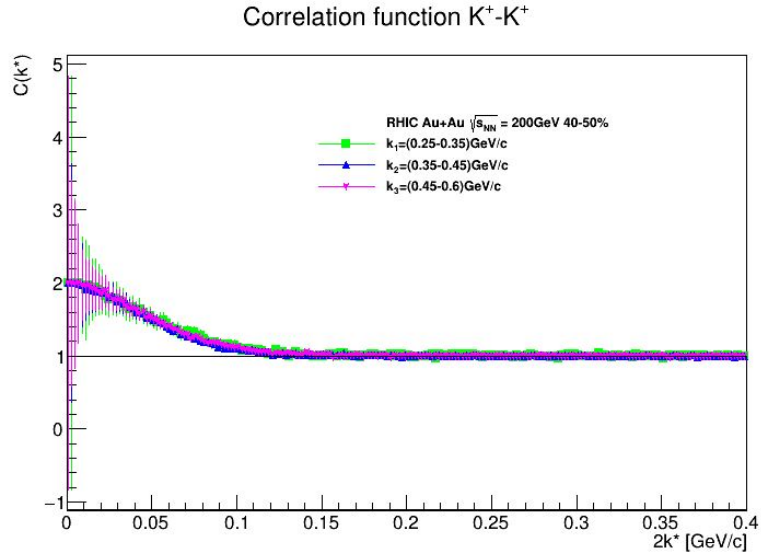


Figure 3.3: Correlation Function for Kaon pairs (K^+K^+ and K^-K^-) for various k_T bin values

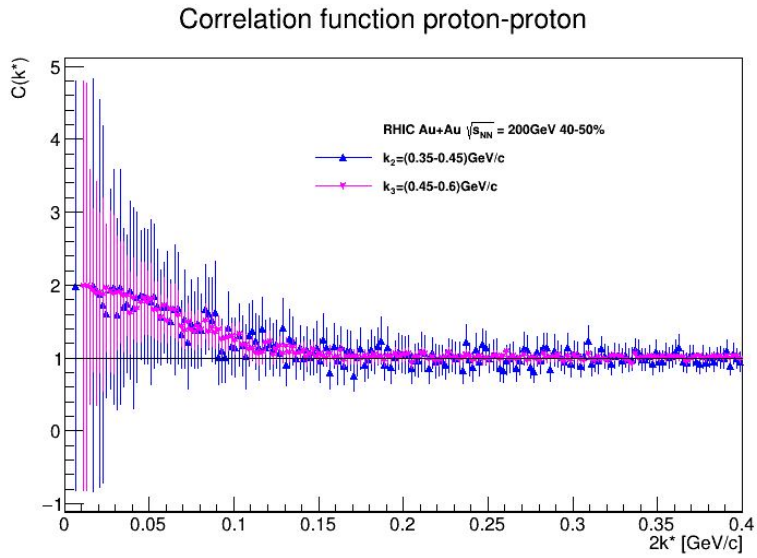


Figure 3.4: Correlation Function for proton pairs (pp) for various k_T bin values

3.1.2 Non-identical Particles

For non-identical particles, we have considered particle pairs of different types, π^+-K^+,or different charge like K^+-K^-. For unlike particles correlation is dominated by Coulomb Interactions [10] which is considered in this analysis, furthermore can be enabled by setting **docoulomb=1** in the **therm2_femto.cxx** file or can be set in **femto.ini** file.

To analyse non-identical particles using therm2_femto, we need to make various changes in source code file (**therm2_femto.cxx**). Below are the changes made for analysis of π^+K^- and π^-K^+.

1. First define the missing PID codes for π^- (#define PINPID -211)and K^- (#define KNPID -321).

2. the section ”Read parameters” we added the missing cases according to the pairs to use: ”else if (tPair == ”piplus-kminus”) pairtype = 3;”.

3. In ”Pair and particle cuts” we noticed that some of the particles combinations in the ”switch(pairtype)” were already defined so we added the ones that were not like the following:

```
    case 3:
ptmin1 = PTMIN;
```

```
ptmin2 = PTMIN;
ptmax1 = PTMAX;
ptmax2 = PTMAX;
break;
```

and at the end of that part add:

```
    else if (pairtype == 3){
partpid = PIPID;
partpid2 = KNPID;
}
```

4. Then in ”Save Histograms to file”, to save the data on under the given root file, add

```
else if (pairtype == 3)
sprintf(bufs,"%sfemtopi+k-%i%s.root",sEventDir.Data(), nbin,
onlyprim ? "p" : "a");
```

5. Finally, in ”Temporary file” define:

```
    else if (pairtype == 3) {
pionac = -248.519 / 0.197327;
partpid = PIPID;
partpid2 = KNPID;
}
```

6. It is also necessary to change the line 32 in the **femto.ini** file, from the name of the pair we are interested in and that we defined in the ”Read Parameters” part.

If after all these changes the desired output is not obtained or an error comes up, we tried commenting a if condition named **if (fabs(mKStarSigned)<0.1)** in ”non-identical particle correlation” part. And closed an open file in ”save histogram file” section by adding **ofile->Close();** at the end.

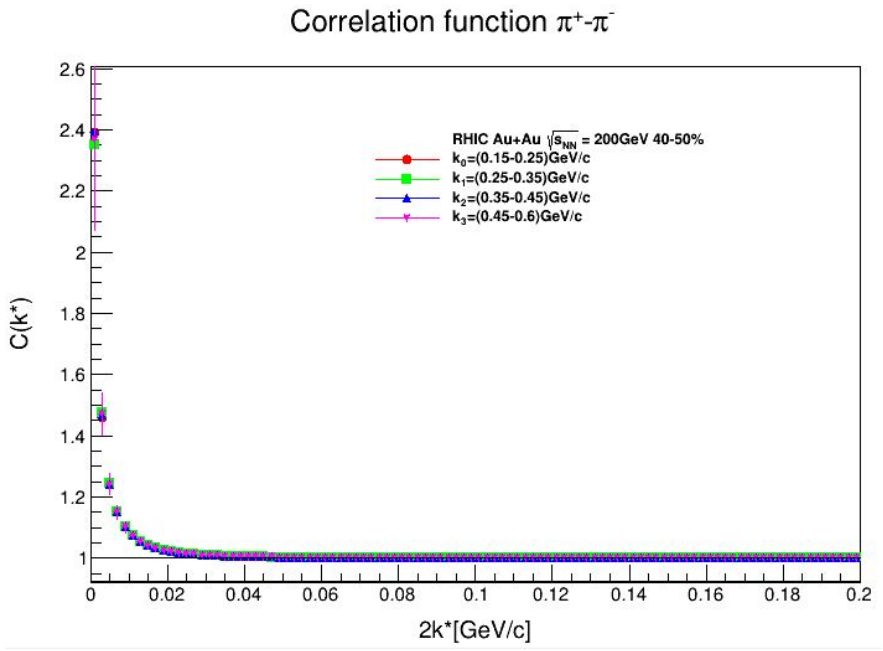


Figure 3.5: Correlation Function for Pion pairs ($\pi^+\pi^-$ and $\pi^-\pi^+$) for various k_T bin values

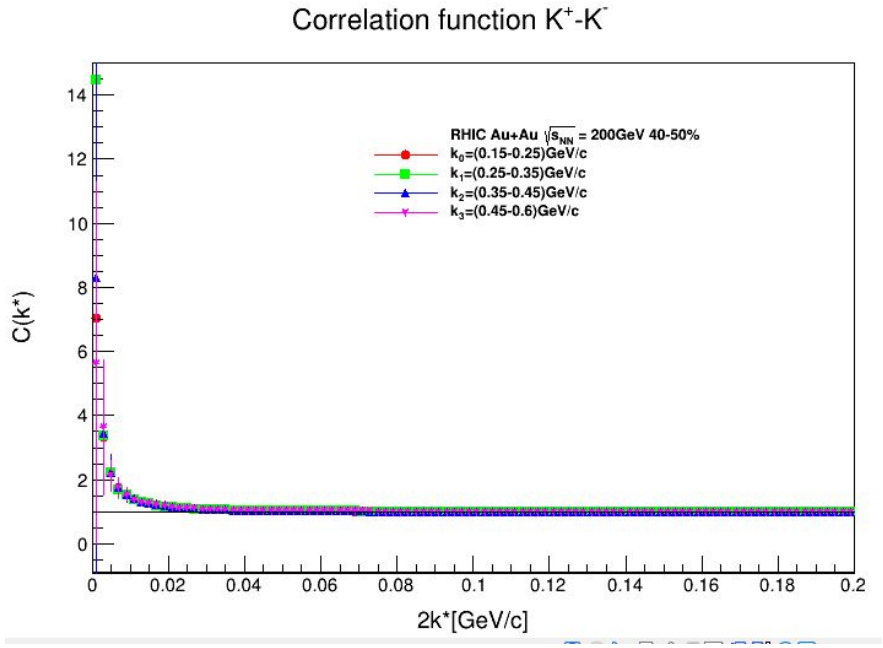


Figure 3.6: Correlation Function for kaon pairs (K^+K^- and K^-K^+) for various k_T bin values

3.2 ϕ Meson Observation

The ϕ meson is considered as a good probe to study the QCD matter created in relativistic collisions [12] and according to [11] the ϕ meson may be observed in the correlation for K^+K^- since it can decay in the channel $\phi \rightarrow K^+K^-$. Using

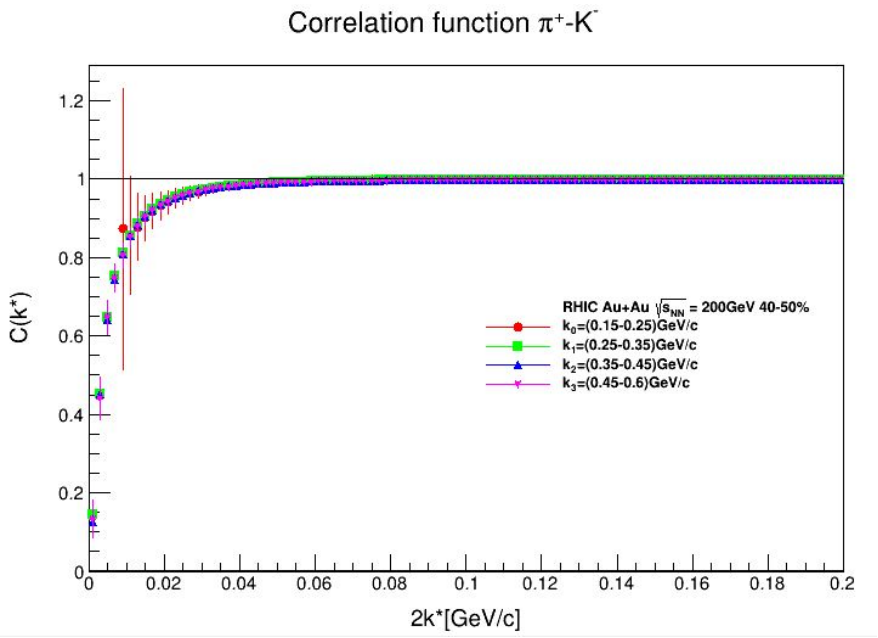


Figure 3.7: Correlation Function for Pion-kaon pairs (π^+-K^- and $\pi^- K^+$) for various k_T bin values

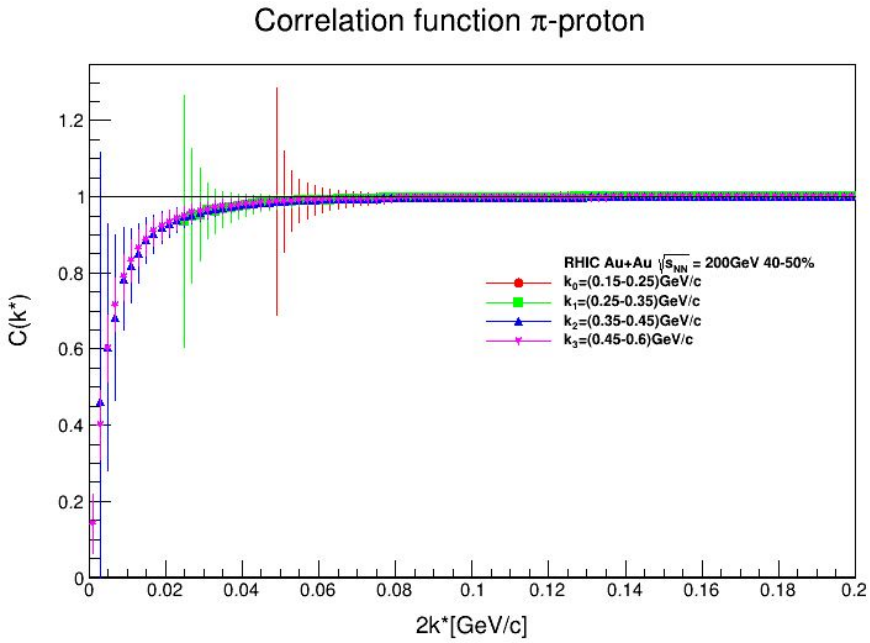


Figure 3.8: Correlation Function for Pion-proton pairs ($\pi^+ p$ and $\pi^- p$) for various k_T bin values

the version of the code given in the previous section but selecting the histograms "num1dptrue;1" instead of "num1d;1" and "den1dptrue;1" instead of "den1d;1", it is possible to observe the ϕ meson peak, as shown in fig 3.10.

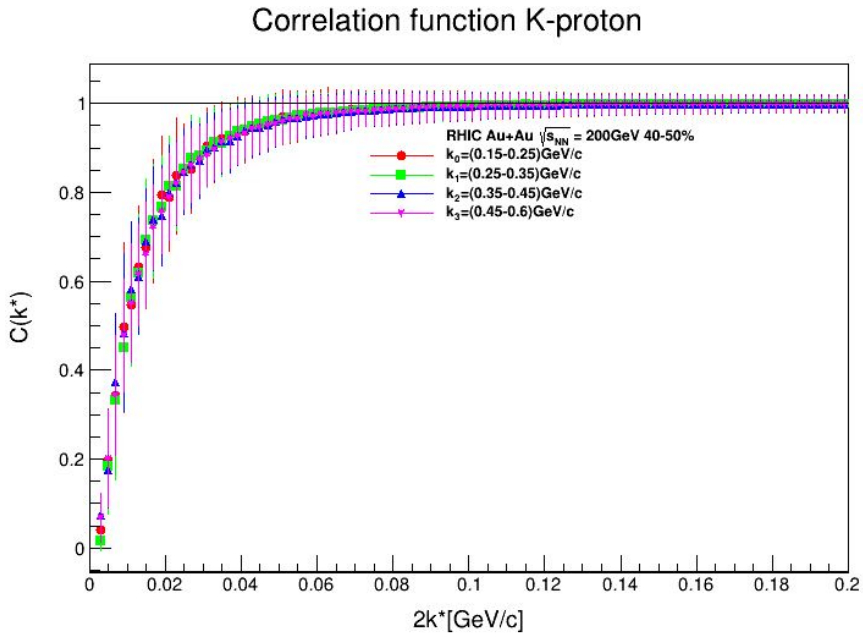


Figure 3.9: Correlation Function for kaon-proton pairs (K^+p and K^-p) for various k_T bin values

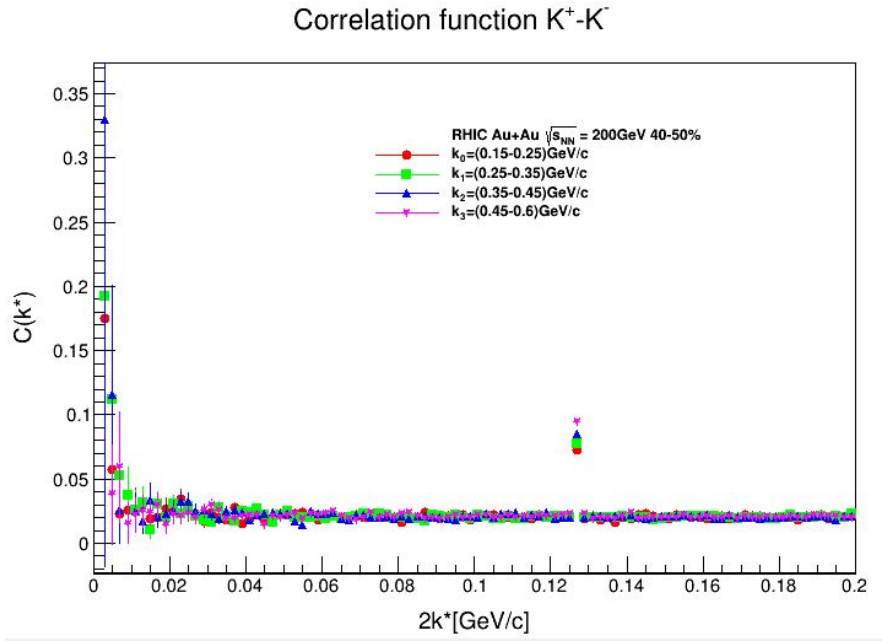


Figure 3.10: ϕ meson peak in K^+K^- Correlation Function.

Chapter 4

Conclusion

In conclusion, events for Au+Au collisions were generated with Therminator 2 for a centrality class of 40%-50% corresponding to semi-peripheral collisions. Using femtoscopy analysis, the correlation function for identical particle pairs ($\pi^+\pi^+$, K^+K^+ and pp) as well as for non-identical particle pairs were obtained. A special remark is made in the case of the K^+K^- correlation function since it can be used to search for the presence of the ϕ meson.

As a further extension of the project, a deeper analysis could be made comparing the results with the ones produced at other centrality ranges. It also would be interesting to perform the HBT analysis using the code implemented in Therminator 2.

References

[1] Connecting QGP-Heavy Ion Physics to the Early Universe, https://arxiv.org/pdf/1306.2471.pdf.

[2] Heavy Ion Collisions: The Big Picture, and the Big Questions,https://arxiv.org/pdf/1802.04801.pdf.

[3] Kaon femtoscopy in pb-pb collisions at $\sqrt{s_{NN}} = 2.76tev$. Phys. Rev. C, 96:064613, Dec 2017.

[4] Adam Kisiel. Nonidentical-particle femtoscopy at $\sqrt{s_{NN}} = 200GeV$ in hydrodynamics with statistical hadronization. Phys. Rev. C, 81:064906, Jun 2010.

[5] Centrality determination of Au + Au collisions at 1.23A GeV with HADES. https://arxiv.org/pdf/1712.07993.pdf.

[6] M.Chojnacki, *THERMINATOR 2: THERMal heavy IoN generATOR 2*, https://arxiv.org/pdf/1102.0273.pdf.

[7] https://root.cern/

[8] Richard Lednicky. Femtoscopic correlations of nonidentical particles. Acta Phys. Polon. B, 40:1145–1154, 2009.

[9] Monte Calro particle numbering Scheme, https://pdg.lbl.gov/2007/reviews/montecarlorpp.pdf.

[10] Michael Annan Lisa, Scott Pratt, Ron Soltz, and Urs Wiedemann. Femtoscopy in relativistic heavy ion collisions. Ann. Rev. Nucl. Part. Sci., 55:357–402, 2005.

[11] B. I. Abelev et al. Measurements of phi meson production in relativistic heavy-ion collisions at RHIC. Phys. Rev. C, 79:064903, 2009.

[12] James S. Lindsey and Gerald A. Smith. Production Properties and Decay Modes of the phi Meson. Phys. Rev., 147:913–922, 1966.

Journal of Religion and Health
https://doi.org/10.1007/s10943-021-01419-y

ORIGINAL PAPER

Gratitude Questionnaire (GQ-6)—Exploring Psychometric Properties in India

Naval Garg[1] · **Nancy Katiyar**[2] · **Mehak**[3]

Accepted: 28 August 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Due to culture-based differences in conceptualization, expression, and gratitude experience, the GQ-6 scale has been validated in different countries. The present study examines the adaptability of the Gratitude Questionnaire (GQ-6) scale in India. Two studies are conducted to evaluate the psychometric properties of GQ-6 for Indian college students. The first study explores factorial validity, reliability, and measurement invariance of the scale. Appropriate factor loading in exploratory factor analysis and model fit indices in confirmatory factor analysis (CFA) provides empirical evidence of the scale's factorial validity. The second study investigates the convergent and criterion validity of the scale. A positive correlation with positive affect (PANAS scale) and life satisfaction (Wheel of Life Scale) and a negative association with negative affect (PANAS scale) establish criterion validity of GQ-6. The results conclude adequate suitability of the GQ-6 scale for Indian college students.

Keywords Gratitude · Reliability · Validity · India

Introduction

Several researchers have reported the issue of extreme concentration of researchers, practitioners, and psychologists on studying unhealthy aspects of one's personality (Diener et al., 1999; Schrank et al., 2014). Elements such as dark personality traits, deviance, stress, and anxiety were at the forefront for quite some time. However, with the emergence of "Positive Psychology," the emphasis has now shifted to individual goodness (Gouveia et al., 2019). Contemporary studies,

✉ Naval Garg
naval.garg@dtu.ac.in

[1] University School of Management and Entrepreneurship, Delhi Technological University, New Delhi, India

[2] Dayalbagh Institute of Technology, Agra, India

[3] IISC, Banglore, India

Published online: 03 September 2021

for example, look at the potentialities, virtues, and values of human beings. Thus researchers have now started the scientific examination of topics such as love, trust, wisdom, appreciation, gratitude, humor, optimism, and happiness. In particular, research on the subject of gratitude has flourished in the last decade (Kong et al., 2017).

Gratitude has been conceptualized as an emotion, a mood, a trait, a virtue, a habit, an affective characteristic (Emmons & McCullough, 2003). Watkins (2016) argue that "gratitude can refer to an emotional state, an emotional expression, a character trait or even a virtue" (p. 13). The diverse conceptualization of gratitude leads to the emergence of multiple definitions and constructs of gratitude. Watkins (2016) recommended Rosenberg's (1998) level of analysis to understand the complexities of gratitude. Rosenberg (1998) suggested that affective phenomenon such as gratitude works at three levels, i.e., emotion, mood, and trait. Rosenberg also argued that if the affective phenomenon is studied on a continuum, emotion and trait rest at two extreme ends, and mood lies somewhere in the middle, emotions are short-lived and transitory. Emotions are brief "psychophysiological changes" (Rosenberg, 1998; p. 250). As an emotion, gratitude is experienced when "something good has happened to them, and they recognize that someone else is largely responsible for this benefit" (Watkins, 2016, p. 17). Moods are relatively longer than emotions; however, they are general. Moods are "wax and wane, that fluctuates throughout or across days" (Rosenberg, 1998, p. 250). Traits are usually permanent and stable. As a trait, gratitude is defined as a significant life orientation toward noticing and appreciating life's positives (Jans-Beken et al., 2019). As a trait, it is regarded as a "general tendency to recognize and respond with grateful emotions to the roles of other people's benevolence in the positive experiences and outcomes that one obtains" (McCullough et al., 2002, p. 112). The trait perspective recommends gratitude as a personality attribute.

Medical practitioners, psychologists, social scientists, and HR managers recommend several gratitude interventions and practices to deal with modern-day challenges. Gratitude interventions improve the physical health of an individual (Redwine et al., 2016; Jackowska et al., 2016; Emmons & McCullough, 2003; Krause et al., 2014). For example, Redwine et al. (2016) conducted an experimental study in which heart failure patients were asked to write about things they were grateful for each day for eight weeks. They found that the experimental group showed better biomarkers than the control group. Patients in the experimental group also reported more parasympathetic heart rate variability, which is a sign of improving heart health. McCraty et al. (1995) also reported a positive relationship between gratitude and heart rate variability of patients. Gratitude also aids better quality and longer sleep (Mills et al., 2015). A higher level of dispositional gratitude also leads to lower stress, anxiety, and depression (Petrocchi & Couyoumdjian, 2015; Disabato et al., 2016). Gratitude can also motivate individuals to maintain a healthy lifestyle. For example, Hills et al. (2013) concluded that grateful people are more likely to engage in healthy activities. Participants of the experimental group engaged in "gratitude journaling" for ten weeks reported spending significantly more time exercising than the control group asked to write five daily events (Emmons & McCullough, 2003). Lambert et al. (2012) suggested that the experience of gratitude may cancel out

negative emotions. Garland et al. (2010) clarified that positive emotions like gratitude lessens one's attention toward negativity, as positive and negative emotions are incompatible.

Numerous other studies have also discussed the physical, psychological, social, and organizational benefits of gratitude practices and interventions. For example, individuals showing high levels of gratitude also report higher levels of life satisfaction (McCullough et al., 2002; Wood et al., 2008a, 2008b), happiness (Morgan et al., 2017; Watkins et al., 2003), optimism and hope (McCullough et al., 2002), vitality and resilience (Garg & Sarkar, 2020), positive affect (Hill & Allemand, 2011), pleasure and self-realization (Nezlek et al., 2016), trust (Drążkowski et al., 2017), self-esteem (Lin, 2015), empathy ((Krumrei-Mancuso, 2016), prosocial behavior (Bono et al., 2017), and patience (Dickens & DeSteno, 2016). Gratitude helps fulfill three basic psychological needs, namely relatedness, autonomy, and competence. Lee et al., (2015) elaborated that there "exists a dual upward spiral between gratitude and specific basic psychological needs." Benefits associated with gratitude have also encouraged practitioners and researchers to examine the workplace implications of gratitude. A feeling of gratefulness for coworkers, supervisors, clients, company, and job negatively associates burnout (Lanham et al., 2012). Chan (2011) reported that dispositional gratitude reduces burnout among school teachers. Grateful employees reported a greater sense of responsibility toward colleagues and societal issues (Andersson et al., 2006). Other studies found empirical evidence of linkages between gratitude and organizational citizenship behavior (Spence et al., 2013; Garg et al., 2020). Garg et al. (2020) also reported that the experience of gratitude might reduce employees' turnover intentions.

Although India has a rich historical tradition of gratitude, systematic studies on gratitude are rare in the Indian context. Gratitude has been revered as a divine virtue in Indian epics and scriptures (Shankar, 2013). Bhagavad Gita says, "*Yadrccha-labha-santustoDvandvatitoVimatsarah*" (verse 4.22). It preaches one to remain grateful for all material and spiritual possessions. Most Hindu *mantras* either start or end with the word "N*amah*," which is nothing but thankfulness (Tiwari, 2015). "*NamoBharantaEmasi*" (*Rig Veda,* verse 1.1.7) utters gratitude for all *gurus* (teachers) for their knowledge and wisdom. "*Om NamahShivaya*" expresses gratitude to the almighty for its unconditional love for humanity. The ancient notion of "pious obligations" called *Ṛṇas* could explain the Indian perspective of gratitude. According to Hindu scriptures, a man must perform five holy duties (Pandey & Navare, 2018). These five pious obligations are *Deva-yajna, Pitṛ-yajna, Bhūta-yajna, Manuṣya-yajna,* and *Brahma-yajna.Deva-yajna* suggests prayer, spiritual ceremonies, meditation, fasting, and self-introspection to express one's gratitude to God. *Pitr-yajna* inspires one to remain grateful to parents and ancestors. India celebrates a fifteen-day long festival known as "*shradh"* to remember and honor their forefathers. *Bhūta-yajna* is an obligation toward all nonhuman forms of life (Pandey & Navare, 2018). Hindu culture offers prayer and respect to plants, animals, and birds as an expression of gratitude. *Brahma-yajna* recommends utmost devotion to teachers and knowledge. There are specific days to honour the teacher (G*uru Purnima),* knowledge (*Basant Panchmi),* and skills (*Vishwakarma day). Manuṣya-yajna* is a duty toward fellow human beings. It offers thankfulness to others for their valuable

contributions to one's life. The notion of five sacred duties seems to have 360-degree coverage, wherein gratitude is expressed for almost everyone and everything. Hindu tradition provides one of the most holistic views of gratitude (Garg, 2020).

There is a severe shortage of scientific studies on gratitude in the Indian context (Garg and Mehak, 2021). Developing a reliable and valid measurement tool (scale) is one of the primary requirements for systematic empirical investigation of a phenomenon in any society. Although several scales likeGratitude Adjective Checklist (GAC), Gratitude Questionnaire-6 (GQ-6), Gratitude Resentment and Appreciation Test (GRAT), Short-GRAT, and Multi-Component Gratitude Measure (MCGM) are available to measure gratitude, their adaptability in Indian culture is yet to be established. These questionnaires were developed and validated in non-Eastern settings. Since gratitude is culture-sensitive, these questionnaires require validation in different cultural contexts.

Previous research has reported a different cultural-based manifestation of gratitude. Appadurai (1985) argued that anthropologists and sociologists found cultural variations in the expression of gratitude. Ide (1998) suggested that the experience of gratitude varies with culture. Sommers and Kosmitzki (1988) conducted a comparative study among citizens of the USA and Germany. The results reported that male participants of Germany experience gratitude more frequently than that of the USA. Similarly, Morgan et al. (2014) found that citizens of the USA experience gratitude less frequently than citizens of the UK. Tudge et al. (2016) examined gratitude among American, Brazilian, Chinese, and Russian children in a comprehensive study. They reported that Russian children were least likely to express "concrete gratitude," whereas American children were most likely to express "concrete gratitude." It is being stated that "gratitude may contain a common core with culturally ubiquitous features, in addition to socially constructed elements that change depending on the culture being studied" (Morgan et al., 2014, p. 23). Sommers and Kosmitzki (1988) also argued that experience and expression of gratitude vary with individuals' culture and nationality(Sommers & Kosmitzki, 1988). Kraus et al. (2015) observed that religiosity is related to gratefulness.

The aforementioned discussion clarifies two things: Gratitude is culture-sensitive, and it enjoys a privileged position in Indian tradition. It strongly recommends the investigation of existing gratitude scales in the Indian context for two fundamental reasons. Firstly, it will encourage other researchers to explore individual, group, and societal benefits of gratitude for Indians. Similarly, it might promote Indian HR practitioners to derive benefits from different gratitude interventions like "counting blessing," "gratitude letter," "gratitude meditations." Secondly, the Indian perspective on gratitude would surely enrich the existing literature and knowledge on gratitude.

Therefore, this paper aims to examine the adaptability of the Gratitude Questionnaire (GQ-6) scale in India. The scale has been extensively used by previous researchers (Wood et al., 2008a, 2008b; Froh et al., 2009; Hill & Allemand, 2011; Sun & Kong, 2013; Kong et al., 2017; Hill & Allemand, 2011; Nezlek et al., 2016; Drążkowski et al., 2017). The GQ-6 is a six-item scale that measures one's proneness to gratitude in everyday life. It is a self-report questionnaire that rates six items on a seven-point rating scale ranging from 1 (strongly disagree) to 7 (strongly agree).

The six statements of the scale are as follows: "I have so much in life to be thankful for," "If I had to list everything that I felt grateful for, it would be a very long list," "When I look at the world, I do not see much to be grateful for," "I am full to a wide variety of people," "Long amounts of time can go by before I feel grateful to something or someone," and "As I get older, I find myself to be more able to appreciate people, events and situations that have been part of my life history." Two statements are reverse-scored to avoid participants' errors. The authors reported acceptable psychometric properties of the scale. The internal consistency reliability (Cronbach's alpha) values for six statements of GQ-6 ranged from 0.76 to 0.84. Also, the scale showed criterion validity with a positive correlation with "grateful," "thankful," and "appreciative" rankings.

The present study evaluates the adaptability of GQ-6 in Indian settings with the help of two studies. The first study explores factorial validity and reliability with exploratory factor analysis, confirmatory factor analysis, and reliability estimates. The second study investigates the validity of GQ-6 among the Indian population using convergent and criterion validity. The measurement variance across gender is also studied with the help of multi-group factor analysis.

Study1: GQ-6 in India: Factorial Validity and Reliability

Study1 aimed to test the factorial validity by identifying a one-factor solution with the help of exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). The scale's reliability was assessed using Cronbach's alpha, composite reliability, and McDonald's omega values.

Methodology

Participants

A total of 534 undergraduates, graduates, postgraduates, and Ph.D. students participated in the study. The sample mean age was 23.3 years. Male students comprised 54.80% of the sample, and the remaining were female students (45.20%). The course-wise distribution of the sample is as follows: science (54.80%), commerce and business management (26.20%), and humanities and others (19%). Also, participants belong to different states of India. A detailed sample description is given in Table 1. A convenience sample was used to collect the data. The participants were requested to complete questionnaires created using Google Forms.

Instruments and Procedure

First, participants were asked to share their demographic details like gender, age, academic stream, and residence state. Second, McCullough et al.'s (2002) Gratitude Questionnaire (GQ-6) was used to measure one's experience of gratefulness in

Table 1 Sample description. *Source* Primary data

Variable	Category	Sample 1	Sample 2
Gender	Male	293	297
	Female	241	237
Age	17–19 years	103	115
	20–22 years	214	227
	23–25 years	164	135
	Above 25 years	53	57
Qualification	UG	229	245
	PG	213	197
	PhD	59	45
	Other courses	33	47
University	Central	123	116
	State	244	270
	Private	167	148
Course	Commerce	293	281
	Science	140	150
	Humanities	101	103
	Overall	534	534

everyday life. Sample items include "when I look at the world, I don't see much to be grateful for" and "I have so much in life to be thankful for." The scale comprises a seven-point rating scale that ranges from 1 (strongly disagree) to 7 (strongly agree).

Procedure

The participants were first made aware of the purpose of the study, and then informed consent was obtained. Participants were also ensured of confidentiality and anonymity of the data. The participants were requested to read the statements carefully and respond appropriately.

Data Analysis and Results

The sample was divided randomly into two subsamples. The first subsample of 282 participants was subjected to exploratory factor analysis (EFA) using varimax rotation. SPSS statistics package 21 was used for data analysis. Before EFA, the Kaiser–Meyer–Olkin (KMO) test and Bartlett's test of sphericity were conducted to verify the factorability of the data. Acceptable values of the KMO measure of sampling adequacy (0.805, greater than 0.70) and Bartlett's sphericity test, $\chi^2(15) = 694.675$, $p < 0.001$ (significant at 0.001), concluded appropriateness of data for EFA. Table 2 depicts the results of EFA. It revealed acceptable factor loading (greater than 0.50) for all six statements, which ranges from 0.58 (item 03) to 0.82 (item 02). The six items explained 56.06% variance and had an eigenvalue of 3.364.

Table 2 Factor structure of the Gratitude Questionnaire (GQ-6). *Source* Primary data, based on a sample of 282 participants

Items	Loadings	h^2
04 "I am grateful to a wide variety of people	.81*	.65
02 I had to list everything that I felt grateful for; it would be a very long list	.82*	.68
03 When I look at the world, I don't see much to be grateful for. (R)	.58*	.34
05 As I get older, I find myself more able to appreciate the people, events, and situations that have been part of my life history	.79*	.63
06 Long amounts of time can go by before I feel grateful for something or someone. (R)	.64*	.42
01 I have so much in life to be thankful for"	.80*	.64
Number of items	6	
Explained variance (%)	56.06	
Eigenvalues	3.364	
Cronbach's α	.84	
McDonald's Ω	.84	
Composite Reliability	.881	

The uni-dimensionality of the scale was corroborated by communalities (h^2), whose values range from 0.42 (item06) to 0.68 (item02). These results provide empirical evidence of the factorial validity of GQ-6 in the Indian context.

The cross-validation of the EFA results was done with the help of confirmatory factor analysis (CFA). The CFA was applied to the second sample of 252 participants. The CFA test was conducted using AMOS software. Acceptable values of model fit indices are RMSEA and RMR ≤ 0.08, GFI, CFI, NFI, and RFI ≥ 0.90 (Bentler, 1990; Browne and Cudeck, 1993).CFA results demonstrated a good model fit with following values:$\chi^2/df = 6.245$, GFI = 0.929, CFI = 0.928, and RMSEA = 0.074. Figure 1 reports that the regression weights of the factors range from 0.61 (item 03) to 0.80 (item 02).

Internal consistency reliability was accessed with the help of Cronbach's alpha values. It represents the average of all possible split-half correlations for a set of items. The Cronbach's alpha value for the GQ-6 scale in Indian settings was 0.84, which confirmed the scale's internal reliability. Cronbach and Shavelson (2004) suggested reliance on composite reliability value to establish internal reliability decisively. The value of composite reliability was 0.881. Both values are greater than 0.70, which confirmed the internal consistency reliability of the GQ-6 scale. Further, recent research has recommended McDonald's omega value to reassure the reliability of any scale. Previous researchers have argued that omega values are more realistic reliability estimates because they consider the strength of association between items and constructs and item-specific measurement errors. Again, an acceptable value (0.84, greater than 0.70) concludes the reliability of GQ-6 for Indian participants.

We measured GQ-6 invariance across gender (male and female) with the help of three models: configural, metric, and scalar invariance model. The measurement variance was tested using change in CFI (ΔCFI) and RMSEA (ΔRMSEA)

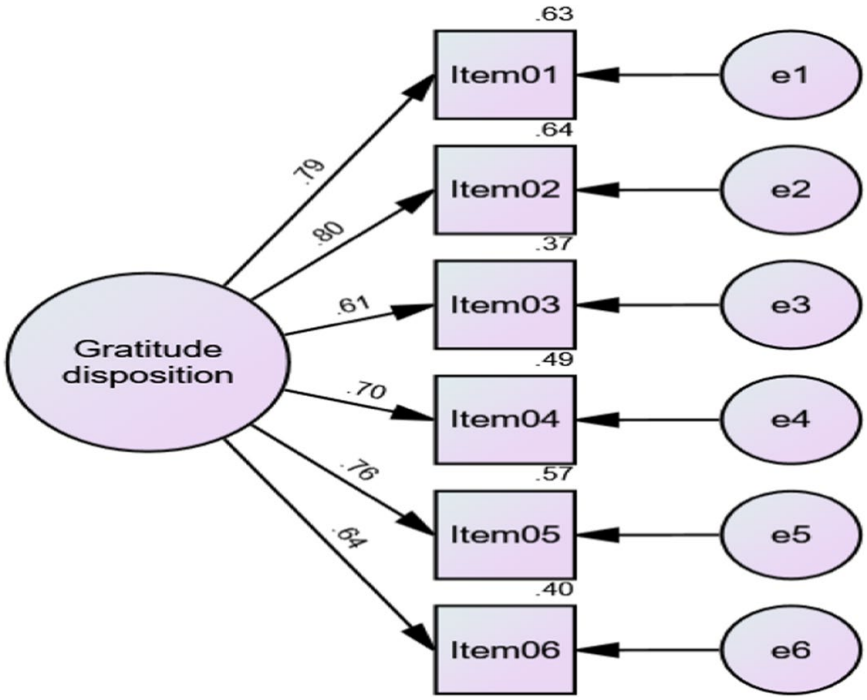


Fig. 1 Factor structure of GQ-6 Questionnaire. *Source*: Primary Data, Based on a sample of 232 participants

values. According to Chen et al. (2016), the values of ΔCFI and ΔRMSEA must be less than 0.010 and 0.015 to validate measurement invariance. First, a multigroup CFA was conducted to investigate configural invariance of the one-factor GQ-6 scale across males and females. Table 3 reported a good fit for the configural invariance model (model-1) with CFI and RMSEA values of 0.928 and 0.775, respectively. It meant that the one-factor model (GQ-6) is suitable for both male and female groups. Further, results showed that the metric invariance model (model 2 in Table 3) also fit well with CFA = 0.924 and RMSEA = 0.068. On comparing model 1 and model 2, ΔCFI and ΔRMSEA were 0.004 and 0.007, respectively. It confirmed metric measurement invariant across genders. The scalar invariance model (model 3 in Table 2) also fitted well with CFI and RMSEA values of 0.921 and 0.060. Also, a comparison of model 2 and model 3 revealed acceptable values of ΔCFI and ΔRMSEA. It concluded scalar invariance of GQ-6 across Indian males and females. Considering the results of all three models, it could be concluded that GQ-6 is equally suitable for men and women in India.

Table 3 Measurement of invariance of GQ-6 across gender. *Source* Primary data, Δ = differences between the current and previous model, based on a sample of 584 samples

Measurement Invariance	CFI	ΔCFI	RMSEA	ΔRMSEA
Configural invariance (Model 1)	0.928	–	0.075	–
Metric invariance (Model 2)	0.924	0.004	0.068	0.007
Scalar invariance (Model 3)	0.921	0.003	0.060	0.008

Study 2: GQ-6 in India: Criterion and Convergent Validity

Study 2 examined the convergent validity and criterion validity of the GQ-6 scale in the Indian context.

Methodology

Participants

A convenience sample was used to collect data from a different set of 534 undergraduates, graduates, postgraduates, and Ph.D. pursuing students. Data were collected using an online questionnaire. There were 297 (55.66%) male participants and 237 (44.34%) female participants. The mean age of the sample was 22.5 years. Also, participants enrolled in science, commerce and business management, and humanities courses were 281, 150, and 103, respectively. Like study 1, participants of the study belonged to different states of the country. A detailed sample description is given in Table 1.

Instruments

The questionnaire consisted of four sections. The first section inquired about demographic information, like gender, age, stream of the study, and domicile state. The second section collected gratitude scores using the GQ-6 scale. Third, Watson et al.'s (1988) twenty-item PANAS (Positive Affect Negative Affect Schedule) scale was used to access positive affect and negative affect scores of the participants. The items were rated on a 5-point rating scale ranging from 1 (very slightly affected) to 5 (extremely affected). The PANAS scale was selected among various scales (e.g., Daniels, 2000; Russell & Daniels, 2018) to measure affect for two reasons. First, PANAS is a comprehensive and structurally efficient measure of affect. It measures positive and negative affect states with the help of ten items for each affective state. Items of the positive state include attentive, excited, enthusiastic, determined, inspired, active, proud, interested, alert, and strong. Negative affect captures hostility, irritability, ashamed, guilt, distress, upset, scared, afraid, jittery, and nervous. Moreover, these scales of two states are statistically uncorrelated and stable at appropriate levels over two months (Watson et al., 1988). Second, the PANAS exhibits cross-cultural validity and reliability (von Humboldt et al., 2017; Weidong et al., 2004). Owing to its robust psychometric properties, researchers consider it a reliable and valid instrument across varying cultures. More importantly, previous researchers, including Dahiya and Rangnekar (2019) and Pandey and Srivastva (2008), reported good reliability and validity of PANAS for Indians.

And, lastly, Whitworth et al.'s (1998) Wheel of Life Scale (WLS) was used to assess participant's satisfaction with essential life domains. It is a tool to determine the participant's satisfaction/dissatisfaction in the different life domains: money and

finance, career and work, health and fitness, fun and recreation, environment, community, family and friend, partner and love, growth and learning, and spirituality. The scale was chosen for its comprehensiveness and resemblance with the Hinduism notion of *Purusartha* (broad aims of life). These are *Dharma* (righteousness), *Kāma*(sensory pleasure), *Artha* (economic and social means), and *Mokṣa (*freedom from bondage). According to the notion of *Purusartha,* one should try to fulfill these four aims of life. The ten items of the WLS scale seem an elaborated version of four *Purusartha.*

Procedure

Participants were made aware of the study's purpose, and informed consent was obtained. Confidentiality and anonymity of the data were also assured. Different cover stories were given to GQ-6, PANAS, and WLS scales to avoid common method bias. The participants were told that the data collected from three different scales belonged to three separate studies, and there was no linkage between these three scales. Different cover stories create psychological separations between the scales. Hence it reduces common method bias (Podsakoff et al., 2003).

Data Analysis and Results

The convergent validity was ensured using average variance extracted (AVE) values. Here, the AVE value of the GQ-6 scale in India was 0.554 (greater than 0.50), so it confirmed convergent validity. The criterion validity of GQ-6 in the Indian context was evaluated through analyzing correlation coefficient values between the gratitude scale and two scales measuring positive affect, negative affect, and satisfaction with life. Previous research notes how individuals showing high levels of gratitude showed higher satisfaction in life and possessed a more positive attitude toward life. Such individuals were also supposed to have a lesser negative affect. The findings of study 2 validate such predictions pointing that gratitude scores were positively correlated with the Wheel of Life ($r = 0.486$, $p < 0.01$) and positive affect ($r = 0.465$, $p < 0.01$), and negatively associated with negative affect ($r = -\ 0.190$, $p < 0.01$). These results confirmed the validity of the GQ-6 scale for the Indian population.

Discussion

Due to culture-based differences in conceptualization, expression, and gratitude experience, the GQ-6 scale has been validated in different contexts (King et al., 2017). Chen et al. (2012) reported the adaptability of scale in Taiwan. Yuksel and Oguz Duran (2012) recommended a single-factor scale composed of five items for Turkey. One item was deleted from GQ-6 for its unsuitability. Jan-Beken et al. (2015) concluded sufficient reliability and validity of GQ-6 among citizens of the Netherlands. Gouveia et al. (2019) reported good factorial and criterion validity in Brazil and suggested measurement variance across gender. Langer et al. (2016)

also concluded good reliability (CA = 0.83) of GQ-6 in Chile. They also confirmed the measurement invariance of the scale across male and female participants. Kong et al. (2017) also observed factorial validity, criterion-related validity, and measurement variance across the gender of the scale in a sample of Chinese students.

Acknowledging these studies, the present paper investigated the adaptability of GQ-6 in an Indian context. To the best of our knowledge, the present study is the first study that explored factorial and criterion validity and reliability of GQ-6 across Indian male and female students. Our paper comprises two studies. Study 1 examined factorial validity, reliability, and measurement invariance of the scale. Appropriate factor loading in EFA and model fit indices in CFA provided empirical evidence of the scale's factorial validity. Three estimates (Cronbach's alpha, composite reliability, and McDonald's Omega) recommended the reliability of GQ-6 for the Indian population. The measurement invariance across gender was studied using multi-group CFA. The configural, metric, and scalar invariance model concluded the suitability of the scale for males and females. In addition, AVE values confirmed convergent validity. Also, a positive correlation with positive affect (PANAS scale) and life satisfaction (Wheel of Life Scale) and a negative association with negative affect (PANAS scale) established criterion validity of GQ-6. In summary, the results show adequate suitability of the GQ-6 scale in Indian settings.

Theoretical and Practical Contributions

The present study offers several significant theoretical and practical contributions. As discussed, India has a long tradition of cherishing gratitude in every sphere of life. One of the most revered Indian scriptures, *Ramayan,* stated gratitude as the essence of Hindu dharma. *Ramayan*states, "Krte ca pratikartavyameshamdharmahsanatanah," which means "to repay a good deed of other" is the core of human life. There are other multiple references of gratitude in *Vedas, Bhagavad Gita, Upanishads,* and other religious Indian texts. The notion of five pious obligations is one of such an Indian conceptualization of gratitude. But, ironically, most scientific studies on gratitude are confined to western settings (Garg, 2020). Only a handful of researchers like Sarkar and Garg (2020), Saxena et al. (2020), Singh et al. (2014), Mahipalan and Sheena (2018) have studied Indian's experience of gratitude. The current study's findings are poised to motivate contemporary Indian researchers to understand the Indian perspective of gratitude. It may lead to empirical studies on conceptualization, experiencing, and expression of gratitude in India. A reliable and valid measurement tool of gratitude would promote a systematic and scientific examination of traditional Indian wisdom. For instance, Indian researchers may examine the role of gratitude in satisfying the four *Purusartha* of human life (*Dharma, Kāma, Artha,* and *Mokṣa)* using the GQ-6 scale. Although Indian sages and *gurus* suggest that a grateful person is more likely to derive life satisfaction through these four broad aims of life, the GQ-6 scale might help in the scientific investigation of this proposition. Similarly, the relevance of the idea of five *rnas* in the twenty-first century may be explored using the GQ-6 scale. Such efforts are beneficial. Firstly, it will provide scientific evidence to Indian thought on gratitude.

Secondly, existing knowledge on gratitude would be significantly enhanced with ancient Indian insights.

Empirical evidence of the adaptability of GQ-6 in India may promote cross-cultural studies. It may encourage a comparative analysis of the gratitude experience of people of different cultural settings. One may be interested in comparing the gratitude level in western and eastern settings. The findings of the study will also promote situation-specific usage of different gratitude interventions in India. Western researchers and practitioners use a wide range of gratitude practices and interventions depending upon their previous experiences and judgments. The utility of these interventions and those derived from the Indian stance on gratitude may now be analyzed empirically.

Emmons and Crumpler (2000) and McCullough et al. (2001) have argued that psychologists and HR managers had neglected gratitude. Other researchers like Fredrickson et al. (2003) and Watkins (2016) have often recommended organizational usage and application of gratitude. A gratuitous workplace may bring vitality and psychological capital among employees (Sarkar and Garg, 2020). In this context, the modern-day HR manager or organizational development (OD) practitioner may use the reliable and valid GQ-6 in organizational settings. Depending on their gratitude score, employees may be trained in different gratitude practices, like gratitude letters, self-introspection, and reciprocal thinking.

Limitations and Future Research

Finally, some important limitations need to be considered. First, the study sample was limited to age groups between 18 and 32 years, so results may not be generalized for other age groups. Second, participants may falsify their responses due to the usage of self-report measures, which may influence the findings of the study. Future researchers may rely on double-source data for better results. Third, only internal consistency reliability is ensured. Future scholars may evaluate inter-rater, parallel form, and test–retest reliability of the scale. Also, additional data may be collected to access the discriminant and concurrent validity of the scale. Fourthly, data related to gratitude, positive and negative affects, and satisfaction with life were collected simultaneously. It may lead to common method bias. Future researchers are requested to rely on longitudinal research design for better generalizations of the results.

Conclusion

Gratitude has historical relevance for Indians. Numerous references related to gratitude are reported in Indian scriptures and spiritual texts. However, the availability of a reliable and valid scale is a prerequisite for any scientific study. This study examines the adaptability of GQ-6 with the help of rigorous statistical measures. The results conclude that the scale is reliable and valid for Indian college students.

Acknowledgements We acknowledge the invaluable contribution of Dr. Jason Palframan, Faculty of Business and Hospitality, Athlone Institute of Technology, Ireland, for proofreading the manuscript.

Funding No funding from any source.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

Informed consent Participants were duly informed of the study's purpose, and consent was procured from all the participants. Also, we declare that the datasets generated during and/or analyzed during the current study are not available publicly due promise of confidentiality to participants.

References

Andersson, L. M., Giacalone, R. A., & Jurkiewicz, C. L. (2006). On the relationship of hope and gratitude to corporate social responsibility. *Journal of Business Ethics, 70*(4), 401–409. https://doi.org/10.1007/s10551-006-9118-1

Appadurai, A. (1985). Gratitude as a social mode in South India. *Ethos, 13*(3), 236–245. https://doi.org/10.1525/eth.1985.13.3.02a00020

Bono, G., Froh, J. J., Disabato, D., Blalock, D., McKnight, P., & Bausert, S. (2017). Gratitude's role in adolescent antisocial and prosocial behavior: A 4-year longitudinal investigation. *The Journal of Positive Psychology, 14*(2), 230–243. https://doi.org/10.1080/17439760.2017.1402078

Chan, D. W. (2011). Burnout and life satisfaction: Does gratitude intervention make a difference among Chinese school teachers in Hong Kong? *Educational Psychology, 31*(7), 809–823. https://doi.org/10.1080/01443410.2011.608525

Chen, L. H., Chen, M. Y., & Tsai, Y. M. (2012). Does gratitude always work? Ambivalence over emotional expression inhibits the beneficial effect of gratitude on well-being. *International Journal of Psychology, 47*(5), 381–392. https://doi.org/10.1080/00207594.2011.632009

Chen, L. H., Wu, C. H., & Chang, J. H. (2016). Gratitude and athletes' life satisfaction: The moderating role of mindfulness. *Journal of Happiness Studies, 18*(4), 1147–1159. https://doi.org/10.1007/s10902-016-9764-7

Dahiya, R., & Rangnekar, S. (2019). Validation of the positive and negative affect schedule (PANAS) among employees in Indian manufacturing and service sector organisations. *Industrial and Commercial Training, 51*(3), 184–194. https://doi.org/10.1108/ict-08-2018-0070

Dickens, L., & DeSteno, D. (2016). The grateful are patient: Heightened daily gratitude is associated with attenuated temporal discounting. *Emotion, 16*(4), 421–425. https://doi.org/10.1037/emo0000176

Diener, E., Suh, E. M., Lucas, R. E., & Smith, H. L. (1999). Subjective well-being: Three decades of progress. *Psychological Bulletin, 125*(2), 276–302. https://doi.org/10.1037/0033-2909.125.2.276

Disabato, D. J., Kashdan, T. B., Short, J. L., & Jarden, A. (2016). What predicts positive life events that influence the course of depression? A longitudinal examination of gratitude and meaning in life. *Cognitive Therapy and Research, 41*(3), 444–458. https://doi.org/10.1007/s10608-016-9785-x

Drążkowski, D., Kaczmarek, L. D., & Kashdan, T. B. (2017). Gratitude pays: A weekly gratitude intervention influences monetary decisions, physiological responses, and emotional experiences during a trust-related social interaction. *Personality and Individual Differences, 110*, 148–153. https://doi.org/10.1016/j.paid.2017.01.043

Emmons, R. A., & Crumpler, C. A. (2000). Gratitude as a human strength: Appraising the evidence. *Journal of Social and Clinical Psychology, 19*(1), 56–69. https://doi.org/10.1521/jscp.2000.19.1.56

Emmons, R. A., & McCullough, M. E. (2003). Counting blessings versus burdens: An experimental investigation of gratitude and subjective well-being in daily life. *Journal of Personality and Social Psychology, 84*(2), 377–389. https://doi.org/10.1037/0022-3514.84.2.377

Fredrickson, B. L., Tugade, M. M., Waugh, C. E., & Larkin, G. R. (2003). What good are positive emotions in crisis? A prospective study of resilience and emotions following the terrorist attacks on

the United States on September 11th, 2001. *Journal of Personality and Social Psychology, 84*(2), 365–376. https://doi.org/10.1037/0022-3514.84.2.365

Froh, J. J., Yurkewicz, C., & Kashdan, T. B. (2009). Gratitude and subjective well-being in early adolescence: Examining gender differences. *Journal of Adolescence, 32*(3), 633–650. https://doi.org/10.1016/j.adolescence.2008.06.006

Garg, N. (2020). Exploring role of gratitude in developing teacher leadership in Indian universities. *International Journal of Educational Management, 34*(5), 881–901. https://doi.org/10.1108/ijem-07-2019-0253

Garg, N., & Sarkar, A. (2020). Vitality among university students: Exploring the role of gratitude and resilience. *Journal of Organizational Effectiveness: People and Performance, 7*(3), 321–337. https://doi.org/10.1108/joepp-03-2020-0041

Garg, N., Jain, A., & Punia, B. (2020). Gratitude, social intelligence, and leadership among university teachers: Mediation and moderation analysis. *International Journal of Organizational Analysis, 29*(2), 368–388. https://doi.org/10.1108/ijoa-01-2020-2016

Garland, E. L., Fredrickson, B., Kring, A. M., Johnson, D. P., Meyer, P. S., & Penn, D. L. (2010). Upward spirals of positive emotions counter downward spirals of negativity: Insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. *Clinical Psychology Review, 30*(7), 849–864. https://doi.org/10.1016/j.cpr.2010.03.002

Gouveia, V. V., Ribeiro, M. G. C., de Aquino, T. A. A., Loureto, G. D. L., Nascimento, B. S., & Rezende, A. T. (2019). Gratitude Questionnarie (GQ-6): Evidence of construct validity in Brazil. *Current Psychology, 40*(5), 2481–2489. https://doi.org/10.1007/s12144-019-00197-x

Hill, P. L., & Allemand, M. (2011). Gratitude, forgivingness, and well-being in adulthood: Tests of moderation and incremental prediction. *The Journal of Positive Psychology, 6*(5), 397–407. https://doi.org/10.1080/17439760.2011.602099

Ide, R. (1998). 'Sorry for your kindness': Japanese interactional ritual in public discourse. *Journal of Pragmatics, 29*(5), 509–529. https://doi.org/10.1016/s0378-2166(98)80006-4

Jackowska, M., Brown, J., Ronaldson, A., & Steptoe, A. (2016). The impact of a brief gratitude intervention on subjective well-being, biology and sleep. *Journal of Health Psychology, 21*(10), 2207–2217. https://doi.org/10.1177/1359105315572455

Jans-Beken, L., Jacobs, N., Janssens, M., Peeters, S., Reijnders, J., Lechner, L., & Lataster, J. (2019). Gratitude and health: An updated review. *The Journal of Positive Psychology, 15*(6), 743–782. https://doi.org/10.1080/17439760.2019.1651888

Jans-Beken, L., Lataster, J., Leontjevas, R., & Jacobs, N. (2015). Measuring gratitude: A comparative validation of the dutch gratitude questionnaire (GQ6) and short gratitude, resentment, and appreciation test (SGRAT). *Psychologica Belgica, 55*(1), 19–31. https://doi.org/10.5334/pb.bd

Kong, F., You, X., & Zhao, J. (2017). Evaluation of the gratitude questionnaire in a Chinese sample of adults: Factorial validity, criterion-related validity, and measurement invariance across sex. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2017.01498

Kraus, R., Desmond, S. A., & Palmer, Z. D. (2014). Being thankful: Examining the relationship between young adult religiosity and gratitude. *Journal of Religion and Health, 54*(4), 1331–1344. https://doi.org/10.1007/s10943-014-9923-2

Krause, N., Emmons, R. A., & Ironson, G. (2015). Benevolent images of god, gratitude, and physical health status. *Journal of Religion and Health, 54*(4), 1503–1519. https://doi.org/10.1007/s10943-015-0063-0

Krumrei-Mancuso, E. J. (2016). Intellectual humility and prosocial values: Direct and mediated effects. *The Journal of Positive Psychology, 12*(1), 13–28. https://doi.org/10.1080/17439760.2016.1167938

Lambert, N. M., Fincham, F. D., & Stillman, T. F. (2012). Gratitude and depressive symptoms: The role of positive reframing and positive emotion. *Cognition & Emotion, 26*(4), 615–633. https://doi.org/10.1080/02699931.2011.595393

Langer, L. I., Ulloa, V. G., Aguilar-Parra, J. M., Araya-Véliz, C., & Brito, G. (2016). Validation of a Spanish translation of the Gratitude Questionnaire (GQ-6) with a Chilean sample of adults and high schoolers. *Health and Quality of Life Outcomes*. https://doi.org/10.1186/s12955-016-0450-6

Lanham, M., Rye, M., Rimsky, L., & Weill, S. (2012). How gratitude relates to burnout and job satisfaction in mental health professionals. *Journal of Mental Health Counseling, 34*(4), 341–354. https://doi.org/10.17744/mehc.34.4.w35q80w11kgpqn26

Lee, L. N., Tong, E. M. W., & Sim, D. (2015). The dual upward spirals of gratitude and basic psychological needs. *Motivation Science, 1*(2), 87–97. https://doi.org/10.1037/mot0000018

Lin, C. C. (2015). Self-esteem mediates the relationship between dispositional gratitude and well-being. *Personality and Individual Differences, 85*, 145–148. https://doi.org/10.1016/j.paid.2015.04.045

McCraty, R., Atkinson, M., Tiller, W. A., Rein, G., & Watkins, A. D. (1995). The effects of emotions on short-term power spectrum analysis of heart rate variability. *The American Journal of Cardiology, 76*(14), 1089–1093. https://doi.org/10.1016/s0002-9149(99)80309-9

McCullough, M. E., Emmons, R. A., & Tsang, J. A. (2002). The grateful disposition: A conceptual and empirical topography. *Journal of Personality and Social Psychology, 82*(1), 112–127. https://doi.org/10.1037/0022-3514.82.1.112

McCullough, M. E., Kilpatrick, S. D., Emmons, R. A., & Larson, D. B. (2001). Is gratitude a moral affect? *Psychological Bulletin, 127*(2), 249–266. https://doi.org/10.1037/0033-2909.127.2.249

Mills, P. J., Redwine, L., Wilson, K., Pung, M. A., Chinh, K., Greenberg, B. H., Lunde, O., Maisel, A., Raisinghani, A., Wood, A., & Chopra, D. (2015). The role of gratitude in spiritual well-being in asymptomatic heart failure patients. *Spirituality in Clinical Practice, 2*(1), 5–17. https://doi.org/10.1037/scp0000050

Morgan, B., Gulliford, L., & Kristjánsson, K. (2014). Gratitude in the UK: A new prototype analysis and a cross-cultural comparison. *The Journal of Positive Psychology, 9*(4), 281–294. https://doi.org/10.1080/17439760.2014.898321

Morgan, B., Gulliford, L., & Kristjánsson, K. (2017). A new approach to measuring moral virtues: The multi-component gratitude measure. *Personality and Individual Differences, 107*, 179–189. https://doi.org/10.1016/j.paid.2016.11.044

Nezlek, J. B., Newman, D. B., & Thrash, T. M. (2016). A daily diary study of relationships between feelings of gratitude and well-being. *The Journal of Positive Psychology, 12*(4), 323–332. https://doi.org/10.1080/17439760.2016.1198923

Pandey, A., & Navare, A. V. (2018). Paths of Yoga: Perspective for workplace spirituality. In S. Dhiman, G. Roberts, & J. Crossman (Eds.), *The Palgrave handbook of workplace spirituality and fulfillment.* Palgrave Macmillan. https://doi.org/10.1007/978-3-319-62163-0_4

Pandey, R., & Srivastava, N. (2008). Psychometric evaluation of a Hindi version of positive-negative affect schedule. *Indian Psychiatry Journal, 17*(1), 49–54.

Petrocchi, N., & Couyoumdjian, A. (2015). The impact of gratitude on depression and anxiety: The mediating role of criticizing, attacking, and reassuring the self. *Self and Identity, 15*(2), 191–205. https://doi.org/10.1080/15298868.2015.1095794

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Redwine, L. S., Henry, B. L., Pung, M. A., Wilson, K., Chinh, K., Knight, B., Jain, S., Rutledge, T., Greenberg, B., Maisel, A., & Mills, P. J. (2016). Pilot randomized study of a gratitude journaling intervention on heart rate variability and inflammatory biomarkers in patients with stage B heart failure. *Psychosomatic Medicine, 78*(6), 667–676. https://doi.org/10.1097/psy.0000000000000316

Rosenberg, E. L. (1998). Levels of analysis and the organization of affect. *Review of General Psychology, 2*(3), 247–270. https://doi.org/10.1037/1089-2680.2.3.247

Saxena, A., Garg, N., Punia, B., & Prasad, A. (2020). Exploring role of Indian workplace spirituality in stress management: A study of oil and gas industry. *Journal of Organizational Change Management, 33*(5), 779–803. https://doi.org/10.1108/jocm-11-2019-0327

Schrank, B., Brownell, T., Tylee, A., & Slade, M. (2014). Positive psychology: An approach to supporting recovery in mental illness. *East Asian Archives of Psychiatry, 24*(3), 95–103.

Singh, M., Khan, W., & Osmany, M. (2014). Gratitude and health among young adults. *Indian Journal of Positive Psychology, 5*(4), 465. https://doi.org/10.15614/IJPP/2014/V5I4/88528

Sommers, S., & Kosmitzki, C. (1988). Emotion and social context: An American–German comparison. *British Journal of Social Psychology, 27*(1), 35–49. https://doi.org/10.1111/j.2044-8309.1988.tb00803.x

Spence, J. R., Brown, D. J., Keeping, L. M., & Lian, H. (2013). Helpful today, but not tomorrow? Feeling grateful as a predictor of daily organizational citizenship behaviors. *Personnel Psychology*. https://doi.org/10.1111/peps.12051

Sun, P., & Kong, F. (2013). Affective mediators of the influence of gratitude on life satisfaction in late adolescence. *Social Indicators Research, 114*(3), 1361–1369. https://doi.org/10.1007/s11205-013-0333-8

Tudge, J. R. H., Payir, A., Merçon-Vargas, E., Cao, H., Liang, Y., Li, J., & O'Brien, L. (2016). Still misused after all these years? A reevaluation of the uses of bronfenbrenner's bioecological theory of

human development. *Journal of Family Theory & Review, 8*(4), 427–445. https://doi.org/10.1111/jftr.12165

Watkins, P. C. (2016). *Gratitude and the good life: Toward a psychology of appreciation*. Springer. (Softcover reprint of the original 1st ed. 2014 ed.).

Watkins, P. C., Woodward, K., Stone, T., & Kolts, R. L. (2003). Gratitude and happiness: Development of a measure of gratitude, and relationships with subjective well-being. *Social Behavior and Personality: An International Journal, 31*(5), 431–451. https://doi.org/10.2224/sbp.2003.31.5.431

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Whitworth, L. (1998). *Coactive coaching*. Davies Black Publishing.

Wood, A. M., Joseph, S., & Maltby, J. (2008a). Gratitude uniquely predicts satisfaction with life: Incremental validity above the domains and facets of the five factor model. *Personality and Individual Differences, 45*(1), 49–54. https://doi.org/10.1016/j.paid.2008.02.019

Wood, A. M., Maltby, J., Stewart, N., Linley, P. A., & Joseph, S. (2008b). A social-cognitive model of trait and state levels of gratitude. *Emotion, 8*(2), 281–290. https://doi.org/10.1037/1528-3542.8.2.281

Yüksel, A., & Oguz Duran, N. (2012). Turkish adaptation of the gratitude questionnaire. *Eurasian Journal of Educational Research, 46*, 199–215.

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

Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Low Light GAN-Based Photo Enhancement

Ashish Chopra
Samsung Electronics
Noida, India
Ashish.c2@samsung.com

Akash Maan
Delhi Technological University
Delhi, India
maanakashdtu@gmail.com

Anmol Kesharwani
National Institute of Technology
Patna, India
anmolkesharwani095@gmail.com

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532798

***Abstract*— This paper is based on a novel learning-based pipeline that replaces the conventional image processing pipeline in order to enhance performance of current multi camera devices for low light photography. This paper proposes a new suggestive algorithm to suggest a better position to capture image based on region segmentation (object-based segmentation), camera calibration, parameter selection, Frame Regeneration (Merging and alignment of frames) and then using Generative Adversarial Network (GAN) for image enhancement. As Low Light Image enhancement is one of the biggest digital image processing problem due to factors like noise, low exposure and incorrect edge detections of the object in low light so image enhancement is difficult these images, this research work proposes a low light image enhancement model by comparing it with current models. Also, the proposed research work is mainly focused on mobile devices as they are currently used by over 90% users to capture the images. The proposed method assists in increasing the SNR by 7.63 % over the other existing approaches.**

Keywords*— *Low light photography, segmentation, frame regeneration, generative adversarial networks, machine learning, signal to noise ratio.

I. INTRODUCTION

Whenever images are captured in low light, noisy images are produced by our smartphone cameras. Even with current advancements on both sensor and camera ISP side it is still tough to obtain a great image from smartphones in low light. Machine learning networks can be used to enhance the picture for low light by increasing the signal to noise ratio. One of the main reasons for this are low photon count and low SNR. The low SNR is due noise sources like shot noise, read noise. Over the years there are many advancement in improving low light photography.

Quality of low light photograph is dependent on two factors sensor size and exposure time. For better image quality the sensor size must be large or exposure time must be large. Both directly measures how much light is entering while capturing the photograph. Google and apple have given more preference to the second option and dealt with the shaking hands by adding some clever computational tricks, while the Huawei flagships have given more weight to the first option, going with a larger sensor.

Once the sensors measure the amount of light entering the camera lens smartphone automatically detects whether to switch to night mode or not. This is based on the threshold values of lux measured by sensors. This threshold values changes from manufacturers to manufacturers based on algorithm used in further processing.

In this paper, region-based segmentation is used along with GAN. The region-based segmentation method[1]looks for similar attributes between adjacent pixels and groups them into unique regions. This technique[2] was used earlier also for color information retrieval. It can be used for semantic interpretation[3] of an image and also for segmentation of an n-dimensional image[4].This technique can also be used for multi-focus image fusion[5] which is combining complementary information from multiple images of the same scene into an image, so that the resultant image contains a more accurate description of the scene.

For capturing the light photons exposure time is increased i.e. slow shutter speed. But while capturing the for longer duration slight movement in the hand create huge distortion so to avoid this dynamic frames capturing techniques are applied to capture variable number of frames at different point of time to compensate for those blurred frames.

Once this is done now comes a problem which is quite common in mobile photography i.e. over and under exposed images. To tackle this problem different images covering different dynamic range are captured and finally superimposed for better results. Lastly white balance and tone mapping is performed to convert image to bring life images. Fig. 1. provides high level overview of what current low light photography pipeline look is like.

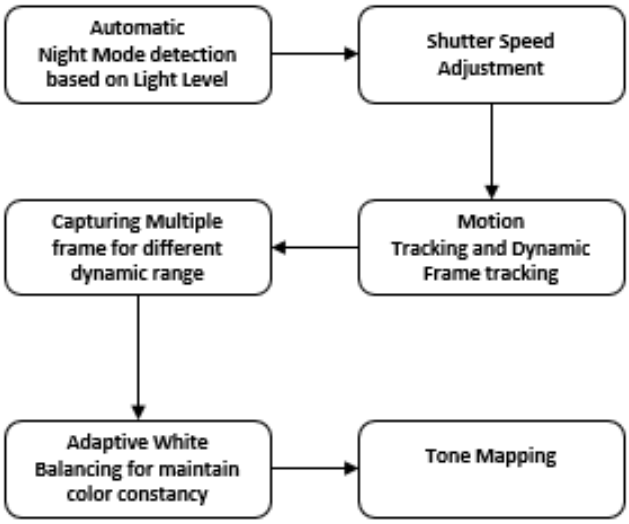


Fig. 1. Low Light Photography Pipeline

II. LITERATURE SURVEY

As per the current scenario, there are various existing methodologies for low-light image enhancement such as L. Tao, et. al [6] proposed a CNN (Convolutional Neural Networks) based approach by utilizing multi-scale feature maps and used SSIM loss function to train their model whereas deep hybrid network-based method[7] was suggested by W. Ren et al for enhancing the visibility of low-light degraded images. The method proposed by [7] used recurrent neural network (RNN) as an edge stream to model edge details and performed better as compared with [6]. An illumination Map Estimation method[8] was given by X. Guo et. al for low-light image enhancement but then a

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

better approach was proposed by L. Li et. al [9] for both denoising and contrast enlarging of low-light images. As compared with the approach of [8], L. Li et. Al [9] used a different concept of image segmentation by dividing it into super-pixels and then a smooth base layer is adaptively extracted by using the BM3D filter. A yet another approach based on pipeline neural
network was proposed by Y. Guo et. al [10] for low-light image enhancement using DWT (discrete wavelet transforms) and CNN.

In this paper, we have used GAN (Generative Adversarial Networks)[11] for generative modelling using deep learning methods. It is an unsupervised learning task which automatically discovers and learns the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset. Ian J. Goodfellow et. al[12] proposed a new framework in which they simultaneously trained two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G.A further improvement in GAN came in 2017 for image-to-image translation in [13] and [14] and for crowdsourcing[15] in 2016.

III. METHODOLOGY

We propose a calibration network combined with a GAN based image processing pipeline which helps in enhancing images in low light photography mode. The calibration network predicts the camera settings required for each camera on the multi camera system based on object of interest. The settings are used during the capture to generate a series of frame which are super imposed using a merger and aligner. And this super imposed image is checked for any Arti-crafts and enhanced by a GAN based model to obtain the final image. The GAN based model directly operates on the output from camera by passing the traditional camera ISP pipeline for end to end enhancement. Figure 2.1 presents the process flow of the methodology of our approach and detailed explanation is covered in next section.

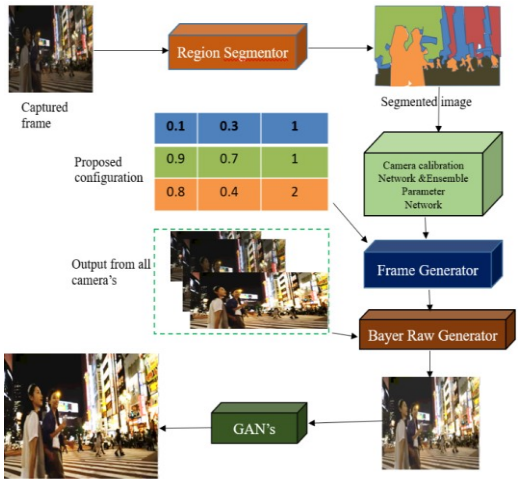


Fig. 2. General Methodology

In Step 1, we process input from wide angle view and perform regions segmentation based both on structure and light intensity. In Step 2, region segmented image is passed to camera calibration network and mapping to ISO, Shutter Speed, and Aperture is performed. In Step 3, 4 generated frames are captured at different possible parameter mapping obtained from previous Steps. We collect results for high edge, contrast and background noise. In Step 5 we super impose frames, aligning and merging is also done in this Step. In Step 6 image enhancement is performed based on dynamic decision of Generator. In final Step i.e. Step 7 filter is applied to images to reduce Glowing noise for Light sources.

A. Region Segmentation

The model analyzes the scene by considering the wide-angle camera's view. Then it segments the image into regions based on their structure and light intensity values. Object segmentation with K-means for covering all cases (K value is dynamically decided to map into region segmentation) is done at the end. Fig. 3, explains how k is selected and output after segmentation.

B. Region to Parameter Mapping

Generated image segment is then passed through camera calibration network which maps ISO and shutter speed values corresponds to segmented regions and aligning it with fixed apertures. Whenever two regions use same aperture there are some intersection regions to enhance them combination of parameters are being used. Camera Calibration Network is trained over wide range of individual image subjects to better predict the ISO, Aperture, and Shutter Speed as per our constructed scale. It is trained on millions of region of images over different ISO, Aperture and shutter speed.

Unlike traditional training of neural network in which the input is augmented we are augmenting the output with enhancement ratio as weightage factor to capture better images.

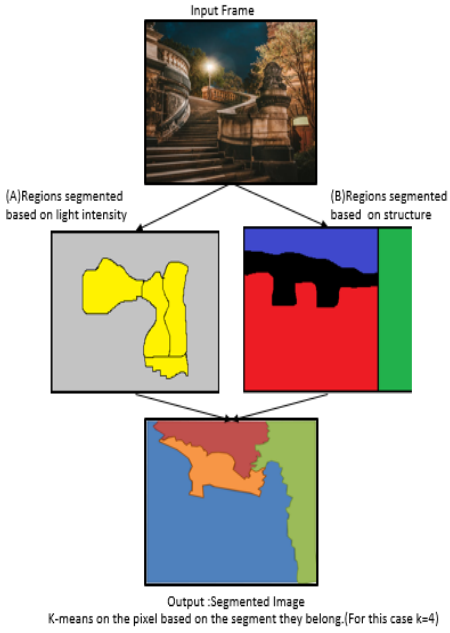


Fig. 3. Region segmentation demonstration.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

C. Camera Calibration Network

Dataset is prepared by capturing images of various subjects each with various settings using the camera which has similar properties of our mobile device. Later these images are stored along with their parameters used during capture – (ISO, shutter speed, Aperture values). Now these, images are segmented using an object detector to find the regions, these regions are passed to $Edge_{index}$, $Contrast_{index}$, $Background_{index}$ and enhancement ratio is calculated which acts as weightage factor of parameters used during the capture of the image.

In other words the enhancement ratio assesses the quality of parameters for segmented images. These labels are embedded in segmented image as pixel value of image. These segments are each divided into individual image. Now these individual images are taken as Input (X) vector. The Scaled versions of ISO, Aperture, Shutter Speed (normalized and multiplied with area ratios) are added as elements of Output(Y) vector. The camera calibration networks architecture is designed under shared feature concept where features used for one model can be used for other model. Also this helps us in reducing the number of computations.

Now, the final layers of the model are selected with consideration of final prediction parameters. ISO mostly

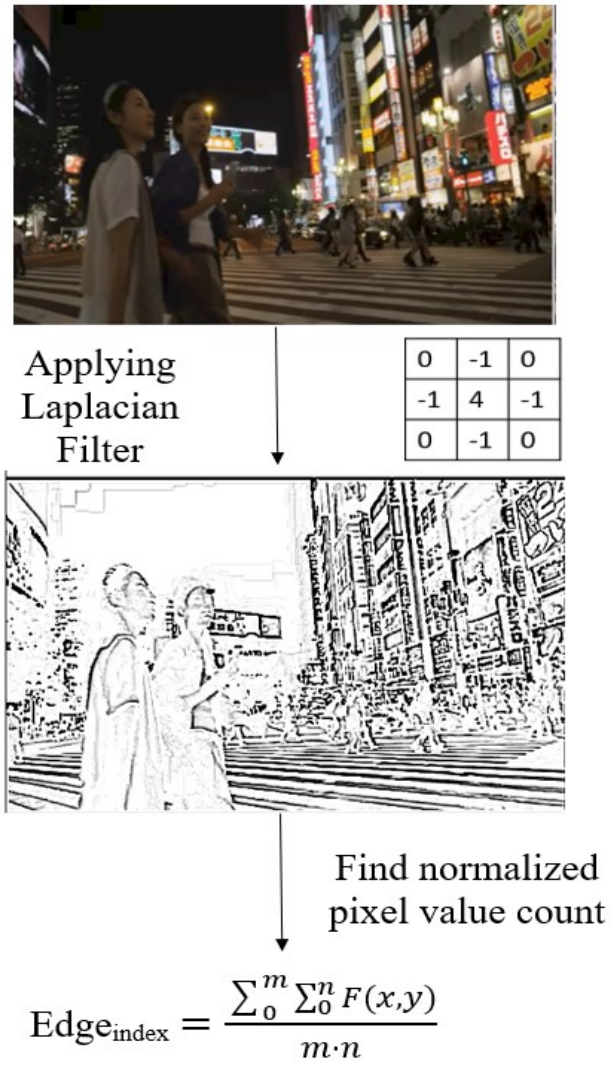


Fig. 4. $Edge_{index}$ Process

depends on the color of image than shape. So a single convolution block is used. In case of Aperture and shutter speed shape and edges play an import role as for aperture area of focus matters.For which two convolution blocks are used for shutter speed distinguishing the edges and high level features is necessary so residual block has been used. The res block help us to retain most of the input image while still being able to learn new features. The convolution blocks helps us to learn filters with decrease in input operation.

$$\text{Enhancement Ratio} = (\text{Edge}_{\text{index}} + \text{Contrast}_{\text{index}} + \text{Background}_{\text{index}})/3 \quad (1)$$

In case of low light photography, the images produced generally suffer from blurriness, dull and with noisy backgrounds. In order to minimize these articrafts we have considered edge index for selecting sharp images, contrast index for selecting bright images, background index to select images with better background.

- $Edge_{index}$: For selecting sharp images

$$\text{Edge}_{\text{index}} = \frac{\sum_0^m \sum_0^n F(x,y)}{m \cdot n} \quad (2)$$

Where, m = breadth of image, n = length of image, F(x,y) - pixel value at location (x,y)

- $Contrast_{index}$: For selecting images with balanced contrast.

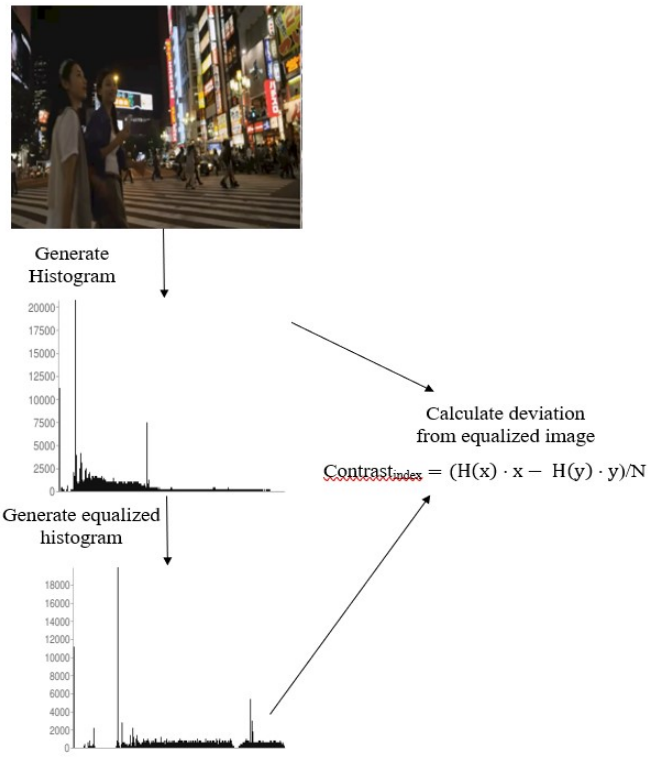


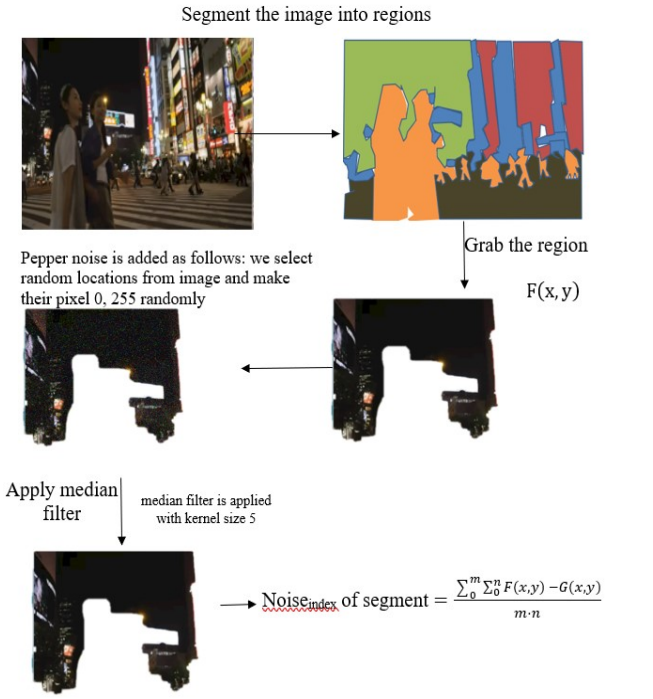
Fig. 5. $Contrast_{index}$

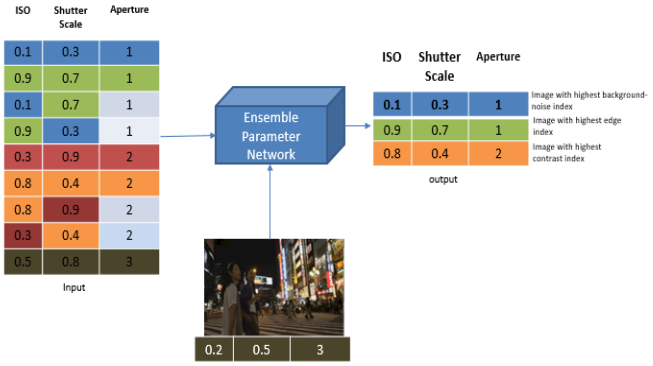
$$\text{Contrast}_{\text{index}} = (H(x) \cdot x - H(y) \cdot y)/N \quad (3)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

Where, H(x) - count of pixel value, N - Number of pixels

Fig.6. $Noise_{index}$





- $Noise_{index}$ of Segment : For selecting images with less noisy segments

$$\text{Noise}_{index}\text{ of segment} = \frac{\sum_0^m \sum_0^n F(x,y) - G(x,y)}{m \cdot n} \quad (4)$$

Where, m = breadth of image, n = length of image, G(x,y) – Image added with salt and pepper noise then passed from median filter.

$$\text{Loss}\,(L(x, y)) = \frac{\lambda_1 |x-y|^2}{2} + \lambda_2 ln(cosh(x-y)) \quad (5)$$

$$y_{iso} = r.Y_{captured_iso} \quad (6)$$

$$y_{aperture} = r.Y_{aperture} \quad (7)$$

$$Y_{shutter_speed} = r.Y_{captured_shutter_speed} \quad (8)$$

$$r = \frac{e_i + b_i + c_i}{3} \quad (9)$$

Where, $e = edge_{index}$, $b = background_{index}$, $c = contrast_{index}$

$$\text{Updated Weight}\,(W_{new}) = W_{old} - n \cdot \frac{\partial L}{\partial \omega} \quad (10)$$

Fig. 7. Parameter Selection.

D. *Parameter Selection*

Parameters generated from the region are later passed through an ensemble network to select the camera parameters with best possible index using Ensemble parameter network. The main objective of ensemble parameter network is to estimate the edge, contrast, background noise index values of images whose settings (iso, shutter speed, aperture) are known but are not yet captured. We want to calculate these values as it's faster compared to capture the images with the known settings and select from them.

E. *Frame Generation*

Once capture button is pressed frames are captured with camera settings from parameter table generated previously. These selected frames are later passed into super imposer for aligning and merging them into one frame. The aligner and merger operate on the Bayer raw format of the images to get a image without any preprocessing.

These frames are later combined by aligner and merger to produce the final Bayer raw. Super imposition consists of mainly two processing stages.



Fig. 8. Frame regeneration.

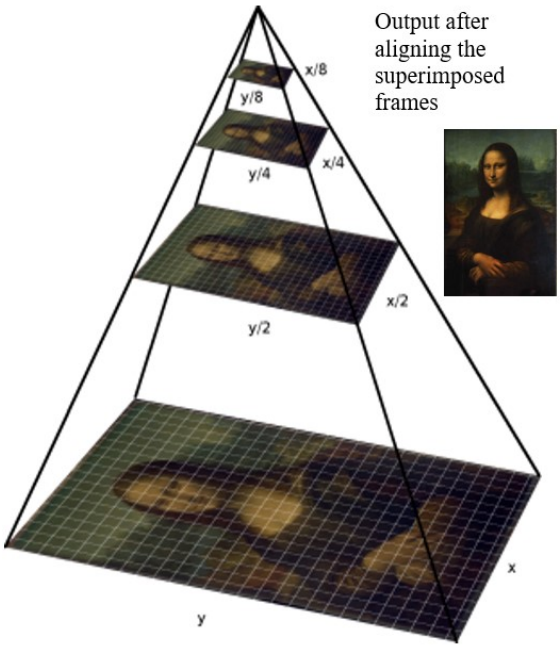


Fig. 9. Aligner Process

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

1. Aligner.

   A reference frame from given multiple images is selected based on sharpness using technique of lucky imaging. Later the respective frames are cropped as per the reference frame size and aligning is done at course level using Gaussian pyramids and tile level using hierarchical alignment.



Tile 1 of image Tile 2 of image

Later these images are applied with Laplacian filters at each level.
Resulting in the images shown below:
The Laplacian filtered images are mapped with their original counter parts for the next merging step.

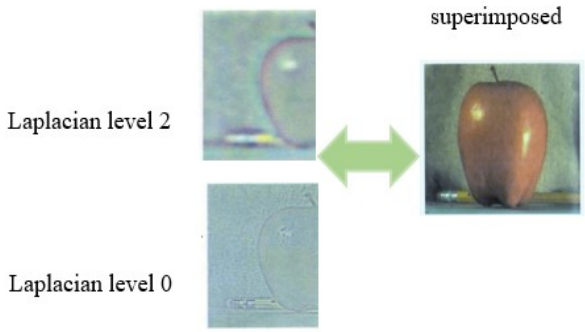


Fig. 10. Alignment

2. Merger.

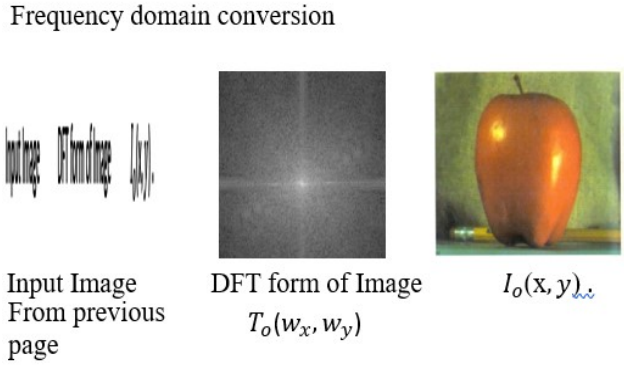
   Now these aligned frames are merged using stack of DFT's based on a merger. First the frames are de-noised using a pairwise temporal filter on tile basis for each frame based on deviation from reference frame. Later these de-noised tiles are combined with reference frame with weights from the pairwise temporal filters.

   The main advantage of this step is we are operating completely on the Bayer images which provides us with 2 or 4 bits more precision compared to normal 8 bit data done by isp.

   Now image tiles with less than 5 percent difference at each level in their Laplacian form are selected for the merging step. For merging these images are converted into frequency domain using 2D DFT(discrete Fourier transform).

Let for image I(x,y) in frequency domain be T(w_x, w_y). The final output of merged image is as follows:

$$T_o(w_x, w_y) = \frac{1}{N}\sum_{z=0}^{N-1}[T_z(w_x, w_y) + K_z(T_r(w_x, w_y) - T_z(w_x, w_y))$$



(11)

Fig. 11. Merger Operations

Here,

$T_o(w_x, w_y) \rightarrow DFT$ of the final merged image.

$T_r(w_x, w_y) \rightarrow DFT$ of the reference image selected based on
edge index.

$T_z(w_x, w_y) \rightarrow DFT$ of the image being merged.

$K_z \rightarrow$ is the proportional factor determined by the average pixel difference b/w reference image and given image.

Now this final $T_o(w_x, w_y)$ is converted from frequency domain to image domain using inverse Fourier transform to get final output $I_o(x, y)$.

F. Image Enhancement Using GAN

We used Adaptive GAN based model to account for the realistic nature of the produced images which is ensured by the adversarial loss term. For generator instead of encoder decoder based we have chosen U-Net as it compensates the information loss faced during progressive down sampling across the layers. There are many corrections like black level adjustment, color correction, gamma correction, edge enhancement etc., that need to perform on the Bayer raw to convert it final jpeg image.

These are compensated by the filtering operations performed at each level of U-Net model. In the generator more convolution blocks are used to pass the information at various resolutions of the input image being passed. One of the main factors we use multivariate loss function based on number of layers in generator is this dynamic nature that helps us to find the optimal number of layers depending on the mobile device the model is being incorporated. Once the outputs are calculated the loss terms are calculated the weights of the discriminator network are updated first. Next the weights of generator network are updated. Also note that while calculating the loss term of the generator we made sure that contribution of individual losses is maintained with proper weights λ1, λ2.

$$W_{new} = \text{new updated weights} \quad (12)$$

$$W_{new} = W_{old} - \boldsymbol{n} \cdot \frac{\partial L}{\partial \omega} \quad (13)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

$\mathbf{L_D^{CGAN}}$ = total loss of discriminator,
$\mathbf{L_G^{CGAN}}$ = total loss of generator,
λ_1, λ_2 = hyper parameters

$$\mathbf{L_G^{CGAN}} = E[\log(D(x,G(x))] + \lambda_1 L_1[G(x),y] + \lambda_2 L_2[G(x),y] + \lambda_3 \alpha T_{batch} \quad (14)$$

Where, α = Number of downsample layers, T_{batch} = Average processing time for batch.

$$\mathbf{L_D^{CGAN}} = \mathbf{E}\left[\log\left(\mathbf{D}(\mathbf{x},\mathbf{y})\right)\right] + \mathbf{E}[\log(\mathbf{1} - \mathbf{D}(\mathbf{x},\mathbf{G}(\mathbf{x}))] \quad (15)$$

The role of discriminator is to classify whether the given image pair is real pair (i.e., sample from training data) or fake pair (i.e., sample generated using generator). In other words the discriminator acts as an image classifier over a pair of images to label them as fake or real. Coming to architecture the initial residual blocks are chosen to convert the Bayer and RGB format inputs to common format with equal number of channels. Now, the input from both formats at this stage is concatenated for further processing. In the final layers the residual block and convolution blocks are later added get the salient features for final labeling. For training the forward pass should be performed in two steps: First the generator is passed with Bayer input to get a rgb generated output. Later the discriminator is passed with real pair i.e., Bayer input and its corresponding rgb output as label(1) and the rgb generated by generator and original Bayer as fake pair with label(0). Now both the generator and discriminator are passed with input_x and 3 outputs are calculated i.e.,

$$\boldsymbol{O_G^{CGAN}} = \boldsymbol{G(x)} \quad (16)$$

where, $\boldsymbol{O_G^{CGAN}}$ = output of generator

$$\boldsymbol{O_{DR}^{CGAN}} = \boldsymbol{D(x,y)} \quad (17)$$

where, $\boldsymbol{O_{DR}^{CGAN}}$ = output of Discriminator (real pair)

$$\boldsymbol{O_{DF}^{CGAN}} = \boldsymbol{D(x,G(x))} \quad (18)$$

Where, $\boldsymbol{O_{DF}^{CGAN}}$ = output of Discriminator (fake pair)

Compared to normal conditional GAN our model doesn't introduce a random vector along with input. This randomness is ensured by superimposed step as multiple images are aligned and merged whose properties vary significantly.

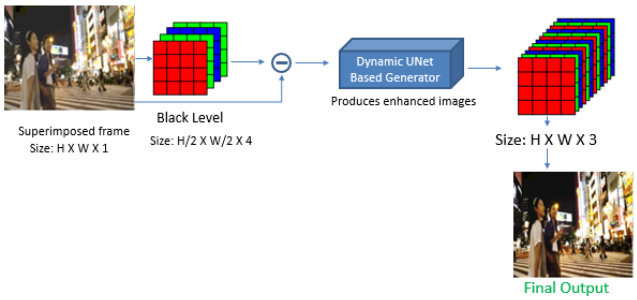


Fig. 12. GAN Enhancement

IV. RESULT

Results obtained from the Model is compared with other famous Lowlight image enhancement model such as Huawei Night Mode, Google Pixel Night Vision and Original Scene with Professional DSLR Camera in order to obtain the Signal to noise ratio with respect to the Original Scene captured by DSLR Camera. Device used is the Samsung S10 with Exynos 9820 with Octa-core architecture CPU with frequency upto 2.73 GHz and Mali-G76 MP-12 GPU with frequency upto 0.7 GHz and 12 core architecture.



Fig. 13. Real Time Model Testing

A real time low light scene is captured as shown in Fig. 14. and tested by the proposed model on Samsung Device. Whereas results produced by compilation of various models is Fig. 13. Table I. and II. shows comparison with several in market models for low photography.

Proposed Model	Google Night Sight	Huawei P30 Pro
Bayer raw i.e., direct sensor output image (10-12bit format) is used as input .	They operate on the processed RGGB image which suffers from noise effects of isp.	They operate on the processed RYYB image which is complete different stack.
We use input from multiple cameras to gather more photon information related to scene.	Multiple frames from same camera are used this introduces errors like motion blur and ghosting	Multiple frames from same camera are used but the RYYB sensor captures more photons than traditional RGGB sensor
Learning based pipeline using Gans to ensure the realistic nature produced image	Hybrid of traditional pipeline and learning based pipeline which take more time to produce output	Traditional isp pipeline but for RYYB which introduces complexity for real implementation
S/N - 29.6dB obtained	S/N - 27.5dB obtained	S/N - 22.3dB obtained

TABLE I. MODEL COMPARISION

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

Device	SNR Ratio
Proposed Method	29.6 dB
Huawei Night Mode	22.3 dB
Google Night Sight	27.5 dB
Oneplus Nightscape	20.7dB
Vivo Super Night Mode	22.1 dB
Oppo Ultra Night Mode	21.7 dB
Apple Night Mode	25.9 dB

TABLE II. RESULT COMPARISION



FIG. 14. RESULTS

V. CONCLUSION

Hence, after observation of result table and general metrics we came to conclusion that our proposed model performs better than the other low light algorithm. To ensure our result credibility Signal to Noise Ratio (SNR) is evaluated with respect to the original scene image. As known previously greater the Signal to Noise Ratio (SNR) better the resultant
image would be, resultant image with larger SNR ratio would capture grater details i.e. image sharpness and color tone would be better than low SNR ratio image with less noise and would have less distortion i.e. no blurry edges.

We conclude our research as our proposed model performs better than other low light image capturing algorithms. We have compared our Low Light image enhancement model against the current best model by Google, Huawei. Shortcoming of various GAN based image processing models is that it can provide a Deep fakes and entirely fictional image after enhancement, whereas segmentation algorithm can overlook object which are blurred in low light these limitations are addressed in our proposed models by using various parameters in GAN and segmentation algorithms are modified accordingly to improve accuracy.

VI. REFERENCES

[1] L. LALAOUI and T. MOHAMADI, "A comparative study of Image Region-Based Segmentation Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 6, 2013, doi: 10.14569/IJACSA.2013.040627.

[2] Y.-I. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," *Computer Graphics and Image Processing*, vol. 13, no. 3, Jul. 1980, doi: 10.1016/0146-664X(80)90047-7.

[3] T. Kanade, "Region segmentation: Signal vs semantics," *Computer Graphics and Image Processing*, vol. 13, no. 4, Aug. 1980, doi: 10.1016/0146-664X(80)90030-1.

[4] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images." doi: 10.1109/ICCV.2001.937505.

[5] S. Li and B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," *Image and Vision Computing*, vol. 26, no. 7, Jul. 2008, doi: 10.1016/j.imavis.2007.10.012.

[6] L. Tao, C. Zhu, G. Xiang, Y. Li, H. Jia, and X. Xie, "LLCNN: A convolutional neural network for low-light image enhancement," Dec. 2017. doi: 10.1109/VCIP.2017.8305143.

[7] W. Ren *et al.*, "Low-Light Image Enhancement via a Deep Hybrid Network," *IEEE Transactions on Image Processing*, vol. 28, no. 9, Sep. 2019, doi: 10.1109/TIP.2019.2910412.

[8] X. Guo, Y. Li, and H. Ling, "LIME: Low-Light Image Enhancement via Illumination Map Estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, Feb. 2017, doi: 10.1109/TIP.2016.2639450.

[9] L. Li, R. Wang, W. Wang, and W. Gao, "A low-light image enhancement method for both denoising and contrast enlarging," Sep. 2015. doi: 10.1109/ICIP.2015.7351501.

[10] Y. Guo, X. Ke, J. Ma, and J. Zhang, "A Pipeline Neural Network for Low-Light Image Enhancement," *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2891957.

[11] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Dec. 2018, Accessed: May 17, 2021. [Online]. Available: https://arxiv.org/abs/1812.04948

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

[12] I. J. Goodfellow *et al.*, "Generative Adversarial Networks," Jun. 2014, Accessed: May 17, 2021. [Online]. Available: https://arxiv.org/abs/1406.2661

[13] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks," Mar. 2017, Accessed: May 17, 2021. [Online]. Available: http://arxiv.org/abs/1703.10593

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," Nov. 2016, Accessed: May 17, 2021. [Online]. Available: http://arxiv.org/abs/1611.07004

[15] A. Khetan and S. Oh, "Achieving Budget-optimality with Adaptive Schemes in Crowdsourcing," Feb. 2016, Accessed: May 17, 2021. [Online]. Available: https://arxiv.org/abs/1602.03481

[16] Andrade, Marcos Blandim, Fernanda Formiga da Silva, Lucas MC e Martins, Hector W. dos S. Oliveira, Fábio LL de Mendonça, and Rafael T. de Sousa. "IoT Fog-based Image Matching Monitoring System for Physical Access Control through IP Camera Devices." In 2020 Workshop on Communication Networks and Power Systems (WCNPS), pp. 1-6. IEEE, 2020.

[17] Manoharan, Samuel. "Improved Version of Graph-Cut Algorithm for CT Images of Lung Cancer With Clinical Property Condition." Journal of Artificial Intelligence 2, no. 04 (2020): 201-206.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:38:40 UTC from IEEE Xplore. Restrictions apply.

Multilevel profiling of situation and dialogue-based deep networks for movie genre classification using movie trailers

Dinesh Kumar Vishwakarma, Mayank Jindal, Ayush Mittal, Aditya Sharma
Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi, India

Abstract

Automated movie genre classification has emerged as an active and essential area of research and exploration. Short duration movie trailers provide useful insights about the movie as video content consists of the cognitive and the affective level features. Previous approaches were focused upon either cognitive or affective content analysis. In this paper, we propose a novel multi-modality: situation, dialogue, and metadata-based movie genre classification framework that takes both cognition and affect-based features into consideration. A pre-features fusion-based framework that takes into account: situation-based features from a regular snapshot of a trailer that includes nouns and verbs providing the useful affect-based mapping with the corresponding genres, dialogue (speech) based feature from audio, metadata which together provides the relevant information for cognitive and affect based video analysis. We also develop the English movie trailer dataset (EMTD), which contains 2000 Hollywood movie trailers belonging to five popular genres: Action, Romance, Comedy, Horror, and Science Fiction, and perform cross-validation on the standard LMTD-9 dataset for validating the proposed framework. The results demonstrate that the proposed methodology for movie genre classification has performed excellently as depicted by the F1 scores, precision, recall, and area under the precision-recall curves.

Key Words: Movie Genre Classification, Convolutional Neural Network, English movie trailer dataset, Multi modal data analysis.

1. Introduction

Movies are a great source of amusement for the audience, impacting society in numerous ways. Identifying the genre of a movie manually may vary due to an individual's taste. Hence, automated movie genre prediction is an active area of research and exploration. Movie trailers are becoming a useful source for predicting the genres of the movie. They provide useful insights into the movie in a very short duration of time. Movie trailers consist of two types of content: cognitive content and affective content.

Cognitive content describes the composition of the events, objects, and persons in a particular video frame of the movie trailer, while Affective content describes the types of psychological features such as feelings or emotions in a movie trailer [1]. Examples of cognitive content comprise a playground, a building, a man, a dog, etc. Examples of affective content are feelings/emotions such as happiness, sadness, anger, etc. Both the cognitive and affect-based content provide prominent features for predicting the genres of the movie.

In this paper, we propose a novel multi-modality situation, dialogue, and metadata-based movie genre classification framework, which aims to predict movie genres using video, audio and metadata (plot/description) content of movie trailers. Our novel framework focuses on extracting both the cognitive and affective features from the movie trailer. For achieving this, a sentence (generated from situations) composed of relevant nouns and verbs is extracted from the video frame. Nouns give the relevant information about the cognitive content of the trailers, and verbs provide useful affect-based mapping with the corresponding genres. For example, the verbs such as laughing, giggling, tickling, etc. provide an affect-based mapping with the 'comedy' genre. The verbs such as attacking, beating, hitting, etc. provide an affect-based mapping with the 'action' genre. Along with situations, dialogue and metadata-based features additionally contribute to cognitive and affective content as they include event descriptions (cognitive content) and psychological features (affective content).

Just like the standard machine learning process, the work is carried out in multiple phases. The 1st phase is the dataset generation phase, where we generate the EMTD, which contains 2000 Hollywood movie trailers belonging to 5 popular genres: Action, Romance, Comedy, Horror, and Science Fiction. The 2nd phase involves pre-processing of video trailers where all repeated frames are removed and resized. The sentences containing important nouns and verbs are extracted from the useful frames. We also prepare the audio transcripts of movie trailers to get dialogues from trailers. In the 3rd phase, we design and train the proposed architecture, which extracts and learn the important features from the trailers. Finally, in the 4th phase, the performance of our proposed architecture is evaluated using the Area under the Precision-Recall Curve (AU (PRC)) metric. The following are the significant contributions of our work:

- We propose a novel EMTD (English Movie Trailer Dataset) containing Hollywood movie trailers of the English language belonging to five popular and distinct genres: Action, Romance, Comedy, Horror, and Science Fiction.

- This work proposes a novel approach to predict movie genres using cognitive and affect-based features. None of the previous literature has focused on a combination of dialogue, situation, and metadata-based features extracted from the movie trailers to the best of our knowledge. Hence, we perform: situation-based analysis using nouns and verbs, dialogue-based analysis using speech recognition, and metadata-based analysis with metadata available with trailers.
- The proposed architecture is also evaluated by performing cross-dataset testing on the standard LMTD-9 [2] dataset. The results show that the proposed architecture has performed excellently and demonstrates the superior performance of the framework.

The remaining portion of the paper is organized as: In Section 2, the past literature on movie genre classification is reviewed, and the motivation behind the proposed work is highlighted. In Section 3, we discuss the proposed EMTD. In Section 4, we provide a detailed description of the proposed architecture. In Section 5, we evaluate the performance of the proposed framework and validate it against two different datasets. The paper is concluded in Section 6.

2. Background and Related Work

This section discusses the past methodologies for movie genre classification and the motivations behind our study. Video content is majorly partitioned to (1) Video frames (Images) and (2) Audio (Speech {dialogues} + Non speech {vocals}). To analyze the video content, various studies are done in the past, focusing chiefly upon cognitive [3]–[7] or affective [8] levels individually. For a more effective study, both levels need to be taken into account to perform better in a genre classification task.

In the past studies, many cognition-based approaches have been proposed based upon low-level features, including visual disturbances, average shot length, gradual change in light intensity in video frames, and peaks in audio waveform [3], to capture scene components [4]. Other features used for cognitive classification include RGB colors in frames [6], film shots [7], shot length [9], type of background in scenes (dark/non-dark) [6], etc. Similarly, some approaches are proposed for only affective analysis [8].

A movie can have multiple genres depicting a lot of information to viewers thus also serve as a task to recommend a movie to a viewer. Jain et al. [5] used 4 video features (shot length, motion, color dominance, lighting key) and 5 audio features to classify movie clips using only 200 training samples. They used complete movie clips to predict genres. However, the study

uses only 200 training samples for training their model. Accordingly, the accuracy reported by them might be due to over-fitting. Also, the study focused only on single-label classification. Huang et al. [4] proposed the Self Adaptive Harmony Search algorithm with 7 stacked SVM's that used both audio and visual features (about 277 features in total) on a 223 sized dataset. Ertugrul et al. [10] used low-level features, including the movies' plot, by breaking the plot into sentences and classifying sentences into genres and taking the final genre to be one with maximum occurrence. Pais et al. [11] proposed to fuse image-text features by relying on some important words from the overall synopsis and performed movie genre classification based on those features. The model was tested on a set of 107 movie trailers. Shahin et al. [12] used movie plots and quotes and proposed Hierarchical attention networks to classify genres. Similarly, Kumar et al. [13] proposed to use movie plots to classify genre using hash vectorization by focusing on reducing overall time complexity. The above-mentioned studies rely on low-level features and do not capture any high-level features from movie trailers, thus cannot be relied upon for a good level recognition system.

From more recent studies, many researchers used deep networks for movie genre classification tasks. Shambharkar et al. [14] proposed a single label 3D CNN-based architecture to seize the spatial and temporal features. Though spatial and temporal features are captured in this, the model is not robust due to single-label classification. Some researchers have worked on movie posters to classify movie genres. Chu et al. [15] formulated a deep neural network to facilitate object detection and visual appearances. Though work captured a lot of information from posters, the poster itself is not enough to completely describe a movie. Simoes et al. [16] proposed a CNN-Motion that included scene histograms provided by the unsupervised clustering algorithm, weighted genre predictions for each trailer, along with some low-level video features. This provided a major group of features from a video but lacked some affective and cognitive-based features to classify genre.

Thus, from the past literature, it is evident that major information should be extracted from the video trailers for cognitive as well as affective study. So, our motivation behind the work is to device an approach relying on both levels of video content analysis as in [1]. We believe that the proposed architecture and the model are novel and robust and can be used in the future for various research perspectives.

3. EMTD Dataset

The datasets in previous literature lack the uniform composition of movie genres. Hence, we propose an EMTD (English Movie Trailer Dataset) consisting of around 2000 unique Hollywood movie trailers downloaded from IMDB[1]. EMTD contains 2000 unique trailers of 5 genres namely: action, comedy, horror, romance, science fiction. The dataset is extracted from IMDB by web scrapping procedure as follows: (1) fetch the list of movie titles available on IMDB (with at least 1 genre common to one mentioned above), (2) scrape metadata corresponding to each movie title including trailer link to download, and (3) download the trailers (.mp4) corresponding to the link into a folder, and list down all the information/meta-data about the movie including trailer name, descriptions, plot, keywords, and genres in the form of a CSV file. In this work, the dataset is partitioned into train set (1700 trailers), validation set (300 trailers) as shown in Table 1.

The study is conducted with the above genres only because mostly these genres are observed in the movies. We also want to explore the performance of our architecture first on a small set of genres, so we go for choosing only 5 genres instead of going towards a broad set of genres.

Table 1: Dataset Composition

Genres / No. of trailers	Action	Comedy	Horror	Romance	Science Fiction
Train set (1700)	671	637	422	388	555
Validation set (300)	126	135	49	67	99

4. Proposed Methodology

We propose a Multi-modality-based M_{SD} framework (Pre-fused feature model) to perform the multi-label movie genre classification task. The Proposed model framework uses textual features i.e., situation-based features from video frames, dialogues extracted from audio, and meta-data from the dataset, as shown in Fig. 1. Majorly our framework is focused upon three features from movie trailers as also depicted in Fig. 1 (1) Descriptions, (2) Dialogues and (3) Situations (from frames). In the subsequent sections, the framework is discussed in detail.

4.1 Descriptions

Movie plot/descriptions are an important feature to describe a movie. In most cases, the plot mentioned for a movie being released is either too short or not mentioned in some cases.

[1] www.imdb.com

Considering this, we choose to use the descriptions concatenated with the dialogues extracted from movie trailers to finally predict the movie genre, as discussed in Section 4.2 in detail. Descriptions are fetched from the IMDB website as metadata as already mentioned in Section3.

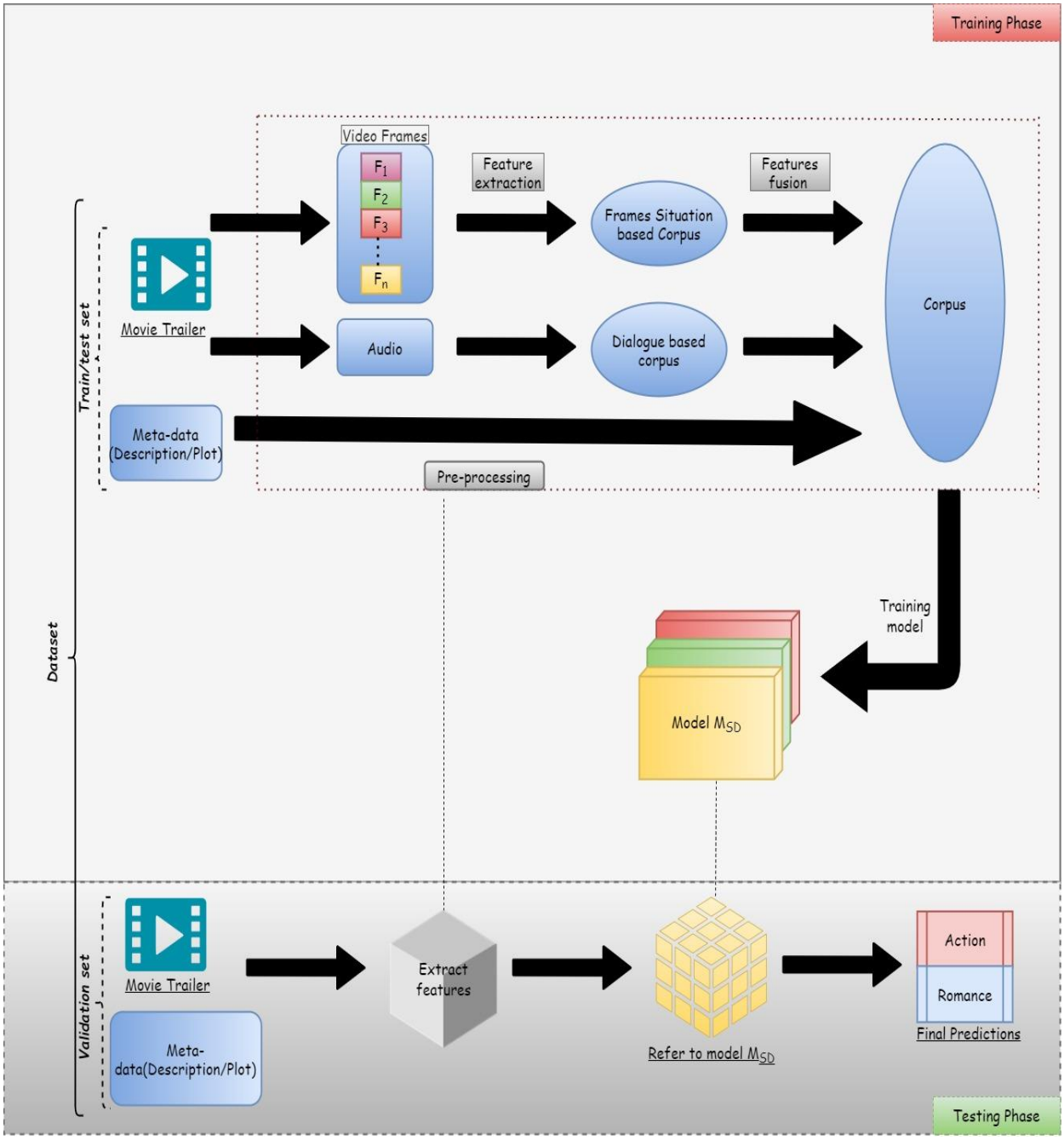


Fig. 1: Pipeline of the framework

4.2 Dialogue

In this section, we propose an architecture to process a list of dialogues from the trailer's audio (concatenated description/plot to dialogues) to predict movie genres. Significant steps for this stream include: (1) Extract speech (dialogue) from movie trailer and (2) Design a model to predict genres on the basis of speech and metadata.

4.2.1 Data pre-processing

The audio files in (.wav) format are extracted from the (.mp4) video trailers. Next, the audio file is split into small audio clips and converted to dialogues as proposed in [17]. All the text is collected to form an input corpus. Description/plot (if available in metadata) is also merged to this corpus. Our study is targeted for the English language trailers only. Just like movie plots, the speech extracted from the trailers can work as a supplement to our text corpus, which can help in better understanding of the relation between the text context and the genre of the movie. After generating the corpus comprised of a single record for each trailer in our training/testing phase the following pre-processing steps were conducted: converting all the text to lowercase, eliminating digits, punctuations, stop-words, and web-links. Text obtained above is used to feed as an input to the model/pre-trained model for training/testing.

Table 2: Abbreviations with their meanings

Abbreviations	Meanings
EMTD	English Movie Trailer Dataset
LMTD-9	Labeled Movie Trailer Dataset – 9
M_S	Situation based Model
M_D	Dialogue based Model
M_{SD}	Situation and Dialogue based Model
CNN	Convolution Neural Network
ANN	Artificial Neural Network
AU (PRC)	Area under the Precision–Recall curve
ADAM	Adaptive Moment Estimation
Bi-LSTM	Bidirectional Long Short-Term Memory
TF-IDF	Term Frequency - Inverse Document Frequency
E-Bi LSTM	Embedding-Bidirectional LSTM
ECnet	Embedding – Convolution network
TFAnet	Term Frequency Artificial Neural Network

4.2.2 Feature Extraction (Dialogue)

The large audio file is split into smaller chunks based on silence. These chunks are then processed, and the final text is obtained by concatenating all processed chunks. An audio file containing a sequence of 'n' chunks can be represented as $A = \{c_1, c_2, c_3, \dots\dots c_n\}$, where c_i $(1 \leq i \leq n)$ is the i^{th} chunk in audio – A. Let $f(A)$ be the function that takes audio as an input and generates dialogues from it defined as Eq. (1) :

$$f(A) = f(c_1, c_2, c_3, \dots\dots c_n) = concat(\widehat{W}_{c_i}), 1 \leq i \leq n \quad (1)$$

$\widehat{W}_{c_i}$ is the maximum a posteriori probability $P(W_{c_i}|X_{c_i})$ of predicting the optimal word sequence W_{c_i}, given the speech signal X_{c_i} for each audio chunk c_i.

$$\widehat{W}_{c_i} = arg\max_{w} P(W_{c_i}|X_{c_i}) \tag{2}$$

Where $P\left(W_{c_i}\middle|X_{c_i}\right)$ is defined using Bayes' Rule as Eq. (3).

$$P\left(W_{c_i}\middle|X_{c_i}\right) = \frac{p(X_{c_i}|W_{c_i})P(W_{c_i})}{p(X_{c_i})} \tag{3}$$

In case of conditional independence, Eq. (3) can be written as Eq. (4).

$$P\left(W_{c_i}\middle|X_{c_i}\right) \propto p\left(X_{c_i}\middle|W_{c_i}\right)P\left(W_{c_i}\right) \tag{4}$$

Substituting the value of conditional independence in Eq. (2), it can be written as Eq. (5).

$$\widehat{W}_{c_i} = arg\max_{w} p(X_{c_i}|W_{c_i})P(W_{c_i}) \tag{5}$$

$p(X_{c_i}|W_{c_i})$ is the acoustic model likelihood and $P(W_{c_i})$ is the language model probability as mentioned in [17]. For each audio chunk c_i, when the time sequence is expanded and the observations $x_{c_{i_t}}$ are assumed to be generated by hidden Markov models (HMMs) with hidden states $\alpha_{c_{i_t}}$, and can be written as Eq. (6).

$$\widehat{W}_{c_i} = arg\max_{w} P(W_{c_i}) \sum_{\alpha} \prod_{t=1}^{T} p\left(x_{c_{i_t}}\middle|\alpha_{c_{i_t}}\right) P\left(\alpha_{c_{i_t}}\middle|\alpha_{c_{i_{t-1}}}\right) \tag{6}$$

where α belongs to the set of all possible state sequences for the transcription W_{c_i}. Substituting $\widehat{W}_{c_i}$ from Eq. (6) in Eq. (1), we get Eq. (7).

$$f(A) = concat\left(arg\max_{w} P\left(W_{c_i}\right) \sum_{\alpha} \prod_{t=1}^{T} p\left(x_{c_{i_t}}\middle|\alpha_{c_{i_t}}\right) P\left(\alpha_{c_{i_t}}\middle|\alpha_{c_{i_{t-1}}}\right)\right), 1 \le i \le n \tag{7}$$

Eq. (7) represents the final dialogues generated from audio.

4.2.3 ECnet (Embedding – Convolution network)

To build cognitive-based genre detection architecture, the crucial features of the trailer in the form of a text corpus needs to be learned by a model. This can be achieved by using a combination of Embedding and CNN (Convolution Neural Network) layers. The layers of the

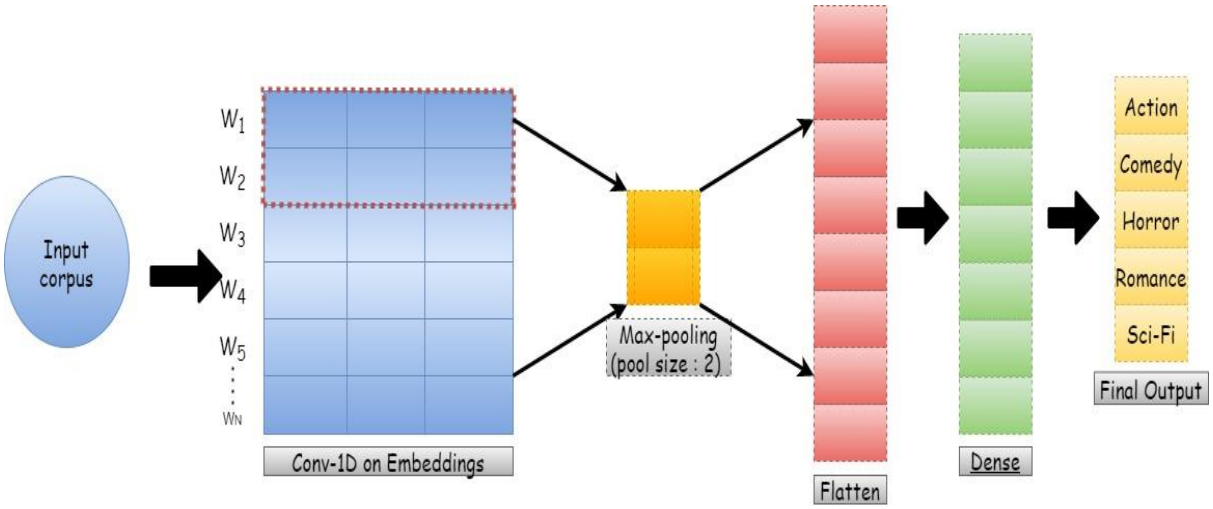


Fig. 2: ECnet architecture

multi-label classification network is depicted in Table 3. Embedding is one of the popular techniques used in NLP problems for converting words into mathematical representation in the form of numerical vectors.

Before actually sending input to the architecture, the vocabulary needs to be designed and the size of a corpus for each data point needs to be fixed. A vocabulary of size 10,395 words is designed and the maximum length of the number of words in each corpus is set to be the length of the longest sentence in our training corpus, which is 330 in our case. If the number of words in a corpus is less than the max length, the corpus is padded with 0's. For a 2-3-minute movie trailer, 330 words are found to be sufficient as in some parts of the trailer there may be no speech (only vocals might be present).

Now for each corpus in the input data, we are having an input of shape (330,) (330 is the number of words in each data point), which is fed to the first layer of our architecture as in Fig. 2**,** i.e., embedding layer. The embedding layer gives an output of dimension (330, 64,) as the length of embedding for each word is taken to be 64 in our proposed architecture.

Table 3: Parameters of ECnet architecture

Layer Name: Type	Description	Input Shape	Output Shape	Parameters #
embedding: Embedding	Embedding Layer with input length = 330	(330,)	(330, 64,)	665280
conv1d: Conv1D	Conv1D Layer with 64 filters	(330, 64,)	(330, 64,)	12352
max_pooling1d: MaxPooling1D	MaxPooling1D Layer with pool size = 2	(330, 64,)	(165, 64,)	0
flatten: Flatten	Flatten Layer	(165, 64,)	(10560,)	0
dense_1: Dense	Dense Layer with 32 neurons	(10560,)	(32,)	337952
dense_2: Dense	Dense Layer with 5 neurons	(32,)	(5,)	165

After the embedding layer, a 1-D convolution layer is fed with the output of the embedding layer. Again, the convolution layer gives an output shape of (330, 64,). To get the same output, we apply the padding uniformly to the input of the convolution layer. Next, a max-pooling layer is used to reduce the dimension of data from (330, 64,) to (165, 64,). The architecture is followed by a flatten layer to transform the two-dimensional data into one-dimensional data, to further send the output to a dense layer.

As depicted in Table 3**,** the flatten layer gives an output of shape (10560,) which is fed to a dense layer as input and giving an output shape of (32,). Finally, the final dense layer is applied to the architecture returning the output shape of (5,) denoting our five genres. In our architecture's final dense layer, we use "sigmoid" as an activation function best suited for our multi-label classification problem.

4.3 Situation

This section includes the work we proposed on visual features from movie trailers. Primary steps for this stream include: (1) fetch video frames from the trailer, (2) extract situations from the frames and (3) build architecture to finally classify the trailers into genres.

A novel situation-based video analysis model is proposed by extracting the situations and events based on each frame extracted from the video for visual features. Thus, a corpus is created to train/test the model by collecting them together.

To the best of our knowledge, we are proposing a novel framework by fusing the situation, event, and dialogue analysis for genre classification. More details about the framework are described in the below sections.

4.3.1 Frame Extraction from Video

Consider a movie trailer video V_t, it can be expressed as a sequence of frames and represented as Eq. (8).

$$V^t = \{F_1^t, F_2^t, F_3^t, \dots\dots\dots F_n^t\}, \tag{8}$$

where n is the total number of frames in trailer t and F_i is the i^{th} frame of trailer t.

After various experimentation using some subset of movie trailers, it is found that taking every 10^{th} the frame is beneficial to avoid redundancy in frames (consecutive frames from a video appear to be similar). Hence, after discarding the redundant frames, the final video frames considered can be expressed as Eq. (9):

$$(V^t)_f = \{F_1^t, F_{21}^t, F_{31}^t, \ldots\ldots F_m^t\} \quad (9)$$

In the subsequent sections, we consider these frames for every trailer.

4.3.2 Feature Extraction (Situation)

In this work, we aim to extract a mixture of cognitive and affective features from the video frames. There has been a lot of work on identifying the objects from the frames, as discussed in 2, which provides only cognitive features. For situation analysis, we extracted features from Grounded Situation Recognition [18] to identify both cognitive and affective features. Situation-based features describing the frame are extracted for each frame from the set $(V^t)_f$, which contains verbs depicting affect-based features, nouns (agent, item, part, place, stage, tool) depicting cognitive ones, and the verb definitions as depicted in Fig. 3.



Fig. 3: Situations for above frames: (a) the soldiers marches in outdoor. (b) haunted lady with blood on face standing along tree (c) people celebrates birthday at a room. (d) a man sprints at a racetrack

To extract features from frames of trailer, the work proposed in [18] to produce situations from images is used. Every image can be expressed in a combination of associated verb and semantic roles present in an image. For every verb, a semantic role can be formed describing the situation like, for eating it can be $Role_{crying} = \{Agent\ (man), Place(road)\}$, where $Role_{crying}$is semantic role for verb crying. Hence, a situation in image I can be expressed as Eq. (10).

$$S = \{\ v,\ A_{(v,I)}\ \}\ ,\ v\ is\ verb\ expressed\ in\ situation\ S \quad (10)$$

where, $A_{(v,I)} = \{\ (r_i, n_i)\ , where\ r_i \in\ Role_v\ , n_i\ \in Nouns\ 1 \leq i \leq |Role_v|\ \ \}$

And the probability that situation S belongs to an image I can be denoted as in Eq. (11).

$$P(S/I, \alpha)\ =\ P(\ v, A_{(v,I)}\ , \alpha) \quad (11)$$

We can replace $A_{(v,I)}$ in Eq. (11) from Eq. (10). Hence, $P(S/I, \alpha)$ can further be expressed as in Eq. (12).

$$P(S/I,\alpha) = P\big(v, (r_1,n_1),(r_2,n_2),(r_3,n_3), \ldots\ldots\ldots, (r_{|Roles_v|}, n_{|Role_v|})|I;\ \alpha\ \big) \quad (12)$$

α denotes the parameter for our neural; network. Now, we can define the semantic roles in an image in a particular order. Thus further, the Eq. (12) be reduced to Eq. (13).

$$P(S/I,\alpha) = P\big(v, n_1, n_2, n_3, \ldots\ldots, n_{|Role_v|}|I;\ \alpha\ \big) \quad (13)$$

Eq. (13) can be further simplified as Eq. (14).

$$P(S/I,\alpha) = P(v/I)\ \prod_{j=1}^{|Role_v|} P(\ n_j\ |v, n_1, n_2, n_3, \ldots\ldots, n_{j-1}\ |I;\ \alpha\ ) \quad (14)$$

For a given particular image/frame, the situation having maximum value probability defined in Eq. (14) will be considered for that image.

The situation for every frame is extracted using the above method, $S^t = \left\{S_{F_1^t}, S_{F_{21}^t}, \ldots\ldots, S_{F_m^t}\right\}$, F_i^t is defined in Section 4.3.1 where S^t is the set of the situation extracted for every frame. Finally, the corpus situation text C^t is formed by the concatenation of individual situations from frames as in Eq. (15).

$$C^t = \ concat(\ S_i\ ),\ \ where\ i\ \in \left[(V^t)_f\right] \quad (15)$$

Now the task is converted to a text classification task for which we propose the model architecture as discussed in upcoming sections. Before proceeding to the next step, text pre-processing is conducted: converting all the text to lowercase, eliminating digits, punctuations, and stop-words, as mentioned in Section 4.2.1**.** These same steps are performed in the testing procedure to predict the movie trailer genre.

4.3.3 TFAnet (Term Frequency Artificial Neural Network)

After extracting visual features, a robust architecture is required to classify the final genres for the trailers. This model is different from the model we proposed in the dialogue stream. Here, TFAnet (Term Frequency Artificial Neural Network) is proposed consisting of a deep network of dense and dropout layers as depicted in Fig. 4.

Before coming to the proposed architecture, we will discuss the text representation using TF-IDF in [19]. For this architecture, it is proposed to use in the word count in the corpus of each data point. Hence, we use the word count from the corpus as features for classifying the movie trailer genres. In order to get a large number of words included as features in our

vocabulary set, trailers from a large range of released dates are used in our EMTD to get a huge corpus available with us while training the model. A combination of unigrams, bigrams and trigrams is used from our corpus as the features and TF-IDF (term frequency-inverse document frequency) algorithm represents our text in a numerical form. The total n-grams features taken are around 34,684. Now our text-based features are transformed into mathematical form, so next (artificial neural network) is trained to classify the genres of the trailer.

Table 4: Parameters of TFAnet

Layer Name: Type	Description	Input Shape	Output Shape	Parameters #
dense_1: Dense	Dense Layer with 64 neurons	(34684,)	(64,)	2219840
dropout_1: Dropout	Dropout Layer with rate = 0.4	(64,)	(64,)	0
dense_2: Dense	Dense Layer with 32 neurons	(64,)	(32,)	2080
dropout_2: Dropout	Dropout Layer with rate = 0.2	(32,)	(32,)	0
dense_3: Dense	Dense Layer with 5 neurons	(32,)	(5,)	165

The architecture of TFAnet (Term Frequency Artificial Neural Network) is depicted in Table 4. The input shape, as discussed above, is (34684,). This input is given to a dense layer, which gives an output of shape (64,). Then a dropout layer is applied to reduce over-fitting with a rate of 0.4. Again, a dense layer is applied, and we obtain an output of shape (32,), followed by a dropout layer with a rate of 0.2. Finally, a dense layer is applied, which gives an output of shape (5,) to finally predict five genres, with sigmoid as an activation function.

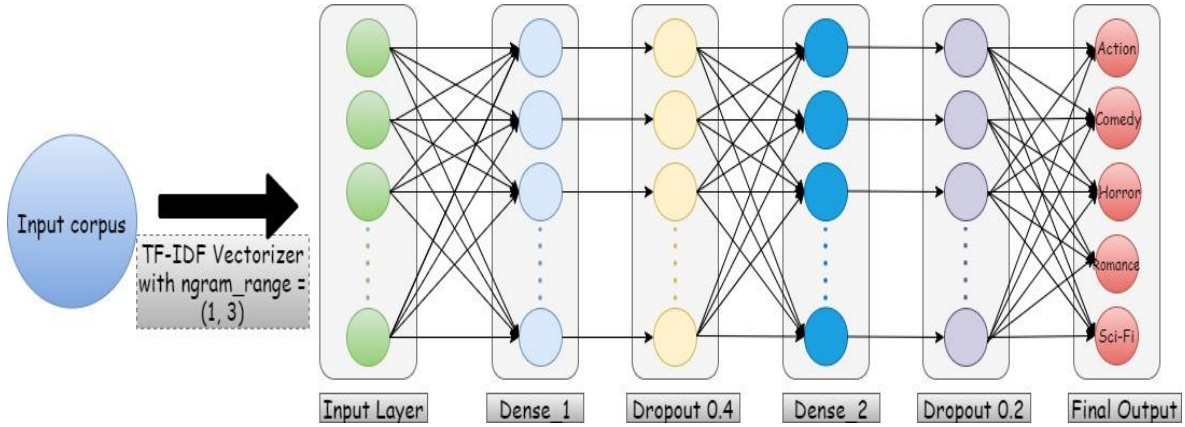


Fig. 4: TFAnet Architecture

The algorithm of the training phase of M_{SD} model is written as Algorithm 1.

Algorithm 1: Training phase of M_{SD} model
INPUT: Set of movie trailers along with meta-data for the training of **M_{SD}** **OUTPUT:** Trained **M_{SD}** with learned weights ready to predict multi-label genres. **START** 1 $C_F = \Phi$

2 for every trailer in training dataset:
3 *Situation extraction*
4 $C_S = \Phi$
5 C_M = description of trailer from metadata
6 for every 10^{th} frame in trailer:
7 resize frame to 299*299*3
8 extract situations from frame
9 append situation to corpus C_S
10 *Dialogues extraction*
11 A_d = fetch audio from trailer
12 C_D = get speech from A_d
13 *Corpus fusion*
14 $C_{SD} = C_D + C_S + C_M$
15 Append C_{SD} to C_F
16 design ECnet architecture
17 train M_{SD} using C_F
END

C_F: list of input corpora
C_S: situation corpus for a trailer
C_D: dialogue for a trailer
C_M: description from metadata
C_{SD}: a concatenated corpus for a trailer
A_d: audio from a trailer

The process of the testing phase can be understood with Algorithm 2.

Algorithm 2: Testing phase of M_{SD} model

INPUT: Set of movie trailers along with meta-data for the test set.
OUTPUT: Genres predicted to test set movie trailers.

START
1 for every trailer in the test set
2 C_{SD} = extract features of the trailer using step 3 to 14 of algorithm 1
3 P_t = get a prediction from pre-trained model M_{SD}
4 output predicted genres
END

C_{SD}: a concatenated corpus for a trailer
P_t: predictions corresponding to trailer t

4.4 M_{SD} Framework

M_{SD} framework is based on the pre-feature fusion of all 3 modalities as described in the previous sections. It uses a similar architecture as the one proposed in Section 4.2.3. Only a certain number of parameters are changed, like the size of vocabulary, which in this case is 14172, and the maximum length of the corpus for each data point is 1661. Algorithm1 and Algorithm2 depict the framework M_{SD}. The final feature fusion is expressed as Eq. (13).

$$F_{SD} = F_S \parallel F_D \parallel F_M \quad (13)$$

where $\boldsymbol{F_S}$, $\boldsymbol{F_D}$ and $\boldsymbol{F_M}$ represent the situation, dialogue, and metadata features, respectively. F_{SD} is obtained by concatenating F_S, F_D and F_M for each trailer.

5. Experiments

In this part, we will examine various model architectures on different modalities and pre-feature fused models. Later, we verify our work by validating it on the standard LMTD-9 dataset as well as on our proposed dataset. Finally, a comparative study is discussed to explore our model robustness. All the experiments are performed on GPU workstations with 128 GB DDR4 RAM and Nvidia Titan RTX (24 GB) GPU configuration.

5.1 Datasets

To verify our framework, we utilize our proposed dataset and standard LMTD-9 [2] dataset. Comprehensive details are mentioned as follows:

5.1.1 English movie trailer dataset (EMTD)

EMTD: Our proposed dataset contains a separate training set of 1700 unique trailers and a validation set of 300 unique trailers, all taken from IMDB, as mentioned in Section 3.

5.1.2 Labeled movie trailer dataset (LMTD-9)

LMTD [16], [20] is a multi-label large-scale movie trailer dataset including trailer link, trailer metadata, plot/summary, unique trailer id consisting of around 9k movie trailers belonging to 22 distinct labels/genres. For verification purposes, a validation set (subpart) of LMTD-9 [2] is used that only includes the Hollywood trailers released after 1980 and trailers specific to our genre list. The dataset contains varying length trailers with different video quality and aspect ratios.

5.2 Classification results on different models

In this section, we will discuss our experiments with different framework variations. We experimented with 3 different frameworks based on separate modalities and pre-fused features.

- **M_S (Video frames analysis):** Model considering the only Situation based features from video frames.
- **M_D (Dialogues-metadata analysis):** Model considering dialogues from audio and descriptions from metadata as features.
- **M_{SD} (Multi-modality analysis):** Model considering situation-based features from video

frames, dialogues from audio and descriptions from metadata as features.

For M_S, we build an ECnet architecture using Keras framework in python as per the architecture defined in Section 4.2.3. The data is split into 85:15 ratios for training and testing. The model is trained on 50 epochs with a learning rate of 0.001 using Adam [21] optimizer to get a training accuracy of 83% and testing accuracy of 70% in 50 epochs.

For M_D, we build a TFAnet (Term Frequency Artificial Neural Network) architecture using Keras framework in python using the layers defined in Section 4.3.3. The data is split into 85:15 ratios for training and testing. The model is trained on 50 epochs with a learning rate of 0.001 using Adam [21] optimizer to get a training accuracy of 80% and testing accuracy of 60% in 50 epochs. After training, M_S and M_D are validated on both datasets.

Table 5: Classification results of M_{SD} on EMTD and LMTD-9

Model		**M_{SD}**	
Genres \ **Dataset**		**EMTD**	**LMTD-9**
Action	P	0.88	0.89
	R	0.89	0.89
	F1	0.88	0.89
Romance	P	0.79	0.71
	R	0.76	0.77
	F1	0.77	0.73
Horror	P	0.88	0.89
	R	0.81	0.87
	F1	0.84	0.88
Sci-fi	P	0.89	0.68
	R	0.86	0.80
	F1	0.87	0.75
Comedy	P	0.86	0.85
	R	0.85	0.85
	F1	0.85	0.84

The architecture proposed in Section 4.2.3 with pre-fused features is used for M_{SD}. However, the input corpus is slightly modified. The corpus defined in Section 4.4 is used for M_{SD}. Precision, Recall, and F1-score for M_{SD} on LMTD-9 and EMTD is depicted in Table 5. However, the AU (PRC) comparison of M_{SD} with M_S and M_D is discussed in the upcoming section.

Some variation can be seen among the performance of different genres. Most of the trailers belonging to major genres are being classified precisely (with an F1 score of 0.84 and above), which shows that the proposed model is performing well. The action genre was the best performing genre among five with an F1-score of 0.88 and 0.89 on EMTD and LMTD-9 respectively. The romance genre was seen to be the least performing genre among all genres

in terms of the F1-score. It is observed that many romance genre trailers are being misclassified into comedy as both these genres are dominated by similar words like happy, smile, laugh, etc.

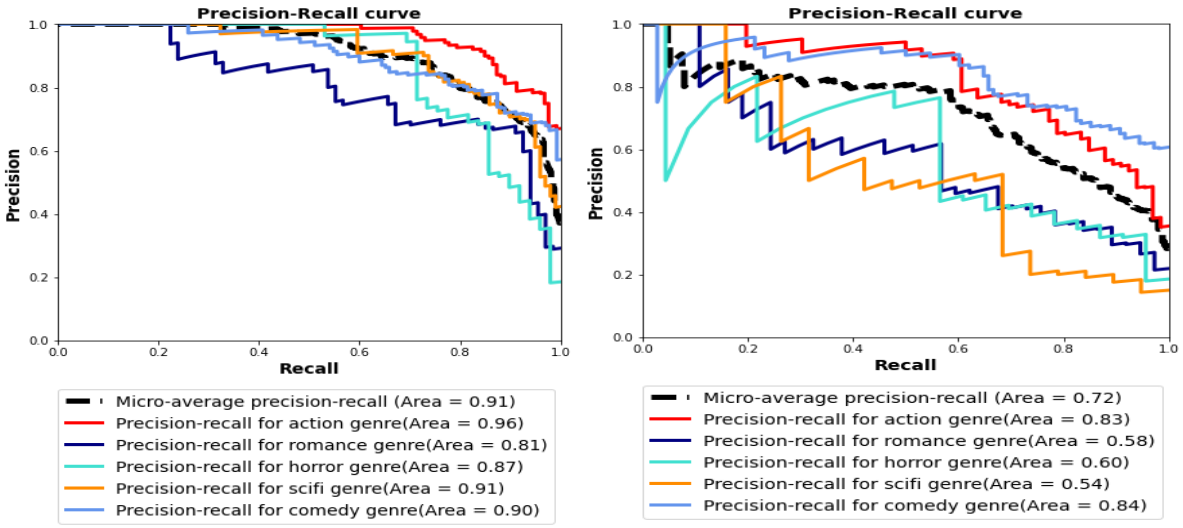


Fig. 5: Precision recall curve for M_D A) EMTD B) LMTD-9

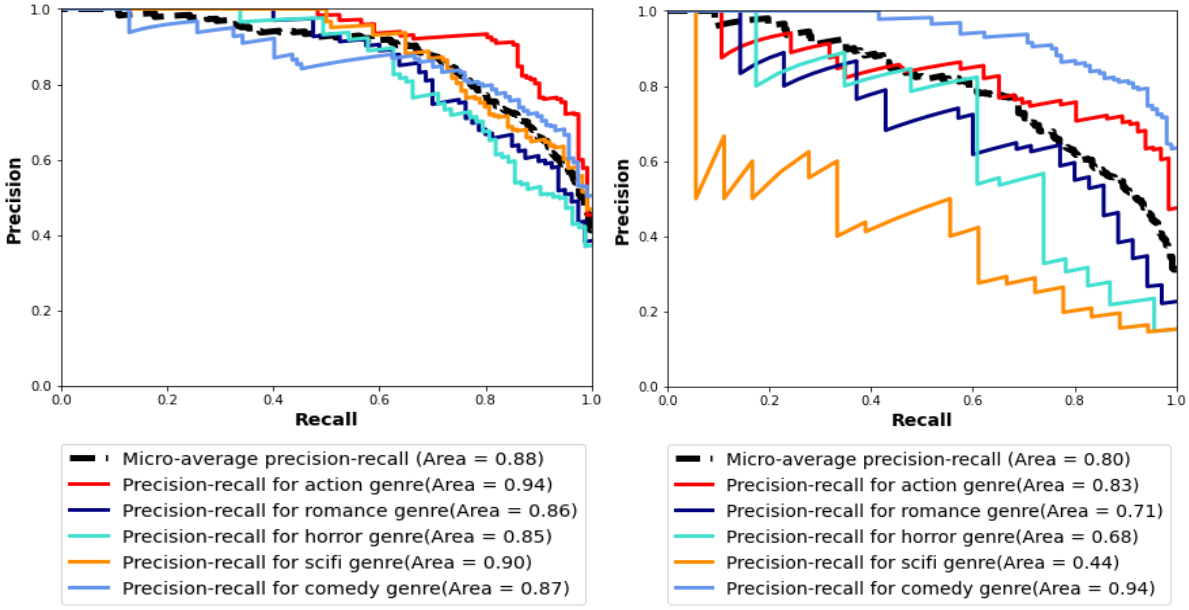


Fig. 6: Precision recall curve for M_S A) EMTD B) LMTD-9

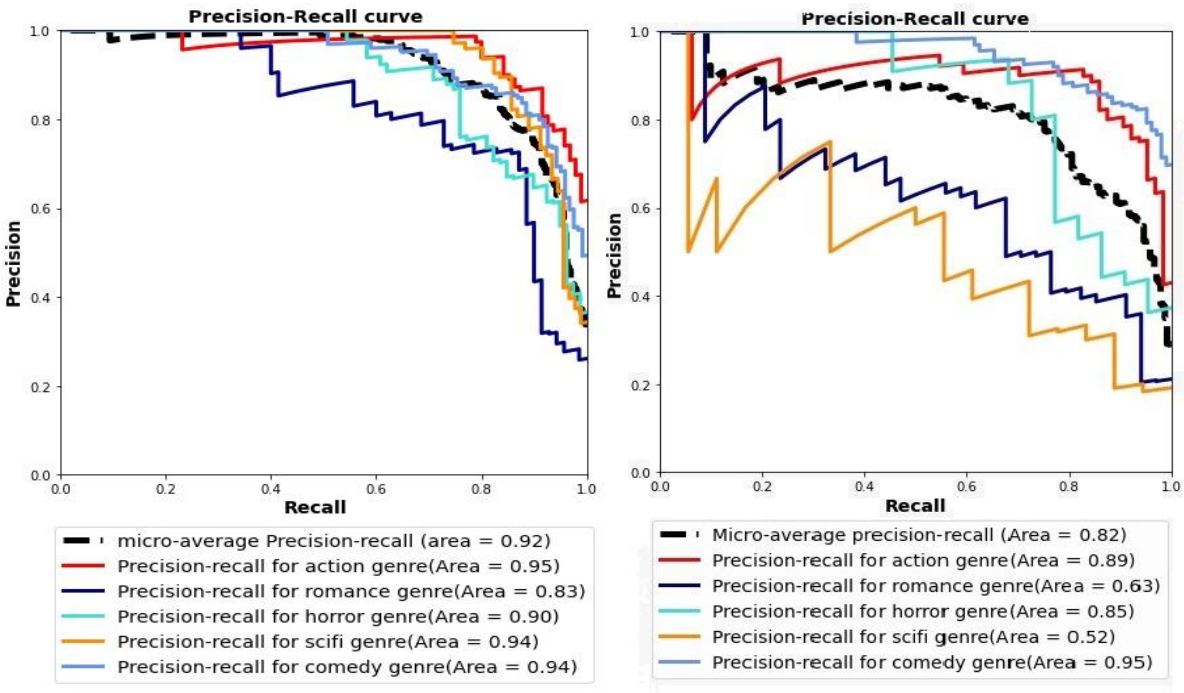


Fig. 7: Precision-Recall Curves M_{SD} A) EMTD B) LMTD-9

5.3 AU (PRC) Comparison

The AU (PRC) i.e., area under the precision-recall curve, is calculated to compare our classification results, as we are dealing with the multi-label classification problem. AU (PRC) measure helps to compare the actual performance of our model, compensating for the noise effect due to class imbalance in the multi-label dataset. The AU (PRC) curves are created for all 3 models on both datasets as depicted in Fig. 5, Fig. 6, and Fig. 7. On the validation set of EMTD, we found almost similar AU (PRC) values 92%, 91%, 88% on M_{SD}, $M_{D,}$ and $M_{S,}$ respectively. However, we found that our M_{SD} gives the 82% AU (PRC) values on the LMTD-9 dataset, which is greater than the other two models i.e., 72% and 80% AU (PRC) of M_D and M_S respectively as in Table 6.

Table 6: AU (PRC) on different models

Features	Model	Dataset	
		EMTD	LMTD-9
Dialogue (M_D)	E-Bi LSTM	0.87	0.66
	ECnet	0.91	0.72
Situation (M_S)	ECnet	0.86	0.75
	TFAnet	0.88	0.80
Fused Features (M_{SD})	ECnet	**0.92**	**0.82**

However, for overall comparison with some other models that we experimented with within our study, we mention their results in Table 6. For choosing the best architecture, the models are compared in terms of AU (PRC) on both of the validation datasets. The implementation of

all the mentioned models helps us in deciding the best model for the fused features. Although M_D has comparable AU (PRC) values with M_{SD} on EMTD but on LMTD-9, M_{SD} outperformed M_D. Similar is the case with M_S on LMTD-9. While M_{SD} performed simultaneously well on both datasets, which is not true in the case of M_S and M_D individually. So, by cross dataset validation M_{SD} proves to be a more robust one. We conclude that the proposed M_{SD} is the best performing model.

5.4 Baseline comparison

In this section, we validate the performance of our proposed model by performing the state of art comparison with the previous approaches for movie genre classification using the AU (PRC) metric for each genre separately as depicted in Table 7. All the results mentioned in Table 7 are shown up to two decimal places and are based on the standard LMTD-9 dataset except for Fish et. al. [22], whose results are based on MMX trailer-20 dataset. It does not consider the romance genre in its study. However, for the other genres, the difference in the AU (PRC) values of Fish et. al [22] and M_{SD} is worth noting. M_{SD} outperforms it by 20% on average. Low-level visual features based classification [23] is based on 24 low-level visual features, SAS-MC-v2 [24] uses only the synopsis for trailer classification, Fish et. al. [22] and CTT-MMC-TN [25] are based on high-level features. Comparing to low-level feature approaches [23], [24], M_{SD} on an average outperforms by 10%, and by comparing from approaches using high-level features [22], [25], it outperforms by 8% on an average for each genre. It is also observed comedy genre performed well in most works as compared to the other four genres while science-fiction has relatively lower AU (PRC) values. This could be due to the unavailability of proper distinction in the science-fiction genre, as its features overlap with some other similar genres (like action).

Table 7: Comparison of the proposed model with similar state-of-the-arts using AU (PRC)

S. No.	Methods	Action	Romance	Sci-fi	Horror	Comedy
1	Low-level visual features [23]	0.85	0.46	0.19	0.42	0.87
2	SAS-MC-v2 [24]	0.68	0.48	0.52	0.58	0.74
3	Fish et. al. [22]	0.62	----	0.50	0.56	0.71
4	CTT-MMC-TN [25]	0.83	0.45	0.40	0.66	0.87
5	M_{SD}	**0.89**	**0.63**	**0.52**	**0.85**	**0.95**

The comparative study demonstrates that the proposed model is robust as it outperforms existing approaches and gives excellent results. The better performance is due to the reason that the architecture proposed includes both cognitive and affective features, helping the model to learn substantial characteristics of each genre, hence predicting genres more precisely.

6 Conclusion

This work extends the idea of a novel holistic approach to the movie genre classification problem that includes affective and cognitive levels by considering multiple modalities, including situation from the frame, dialogues from speech, and meta-data (movie plot and description). We also built a Hollywood English movie trailers dataset EMTD that includes around 2000 trailers from 5 genres, namely action, comedy, horror, romance, science fiction, to pursue this study. We experimented with various model architectures as discussed in Section 5.2 and also validated our final framework on EMTD and on standard LMTD-9 [2] that achieves AU (PRC) values of 0.92 and 0.82 respectively. Our study's main aim is to build a robust framework to classify a movie genre from its short clip i.e., trailer. Although our study includes English speech as a feature, it can also be applied to some Non-English trailers. For Non-English ones, our model can incorporate the video features only, so on the basis of that, predictions can be made by our architecture.

For extension of our proposed model, background audio studies based on vocals can also be incorporated. Hence, in the future, we plan to build a framework considering background vocals in audio along with our current framework to better extract and use most features from movie trailers. We also plan to add some more genres to our study for multi-label classification.

7 References

[1] A. Hanjalic and L. Q. Xu, "Affective video content representation and modeling," *IEEE Trans. Multimed.*, vol. 7, no. 1, 2005.

[2] J. Wehrmann and R. C. Barros, "Convolutions through time for multi-label movie genre classification," in *Proceedings of the ACM Symposium on Applied Computing*, 2017, vol. Part F1280, pp. 114–119.

[3] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, Jan. 2005.

[4] L. H. Chen, Y. C. Lai, and H. Y. Mark Liao, "Movie scene segmentation using background information," *Pattern Recognit.*, vol. 41, no. 3, 2008.

[5] S. K. Jain and R. S. Jadon, "Movies genres classifier using neural network," 2009.

[6] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, 2013.

[7] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, 2013.

[8] A. Yadav and D. K. Vishwakarma, "A unified framework of deep networks for genre classification using movie trailer," *Appl. Soft Comput. J.*, vol. 96, 2020.

[9] K. Choroś, "Video genre classification based on length analysis of temporally aggregated video shots," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018, vol. 11056 LNAI, pp. 509–518.

[10] A. M. Ertugrul and P. Karagoz, “Movie Genre Classification from Plot Summaries Using Bidirectional LSTM,” in *Proceedings - 12th IEEE International Conference on Semantic Computing, ICSC 2018*, 2018, vol. 2018-January.
[11] G. Païs, P. Lambert, D. Beauchêne, F. Deloule, and B. Ionescu, “Animated movie genre detection using symbolic fusion of text and image descriptors,” 2012.
[12] A. Shahin and A. Krzyżak, “Genre-ous: The Movie Genre Detector,” in *Communications in Computer and Information Science*, 2020, vol. 1178 CCIS.
[13] N. Kumar, A. Harikrishnan, and R. Sridhar, “Hash Vectorizer Based Movie Genre Identification,” in *Lecture Notes in Electrical Engineering*, 2020, vol. 605.
[14] P. G. Shambharkar, P. Thakur, S. Imadoddin, S. Chauhan, and M. N. Doja, “Genre Classification of Movie Trailers using 3D Convolutional Neural Networks,” 2020.
[15] W. T. Chu and H. J. Guo, “Movie genre classification based on poster images with deep neural networks,” 2017.
[16] G. S. Simões, J. Wehrmann, R. C. Barros, and D. D. Ruiz, “Movie genre classification with Convolutional Neural Networks,” in *Proceedings of the International Joint Conference on Neural Networks*, 2016, vol. 2016-October.
[17] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, “Chapter 2 - Fundamentals of speech recognition,” in *Robust Automatic Speech Recognition*, J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, Eds. Oxford: Academic Press, 2016, pp. 9–40.
[18] S. Pratt, M. Yatskar, L. Weihs, A. Farhadi, and A. Kembhavi, “Grounded Situation Recognition,” in *Computer Vision -- ECCV 2020*, 2020, pp. 314–332.
[19] B. Beel, Joeran and Langer, Stefan and Gipp, “TF-IDuF: A Novel Term-Weighting Sheme for User Modeling based on Users’ Personal Document Collections,” *Proc. iConference 2017*, 2017.
[20] J. Wehrmann, R. C. Barros, G. S. Simoes, T. S. Paula, and D. D. Ruiz, “(Deep) Learning from Frames,” 2017.
[21] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2015.
[22] E. Fish, A. Gilbert, and J. Weinbren, “Rethinking movie genre classification with fine-grained semantic clustering,” *arXiv Prepr. arXiv2012.02639*, 2020.
[23] F. Álvarez, F. Sánchez, G. Hernández-Peñaloza, D. Jiménez, J. M. Menéndez, and G. Cisneros, “On the influence of low-level visual features in film classification,” *PLoS One*, vol. 14, no. 2, 2019.
[24] J. Wehrmann, M. A. Lopes, and R. C. Barros, “Self-attention for synopsis-based multi-label movie genre classification,” 2018.
[25] J. Wehrmann and R. C. Barros, “Movie genre classification: A multi-label approach based on convolutions through time,” *Appl. Soft Comput. J.*, vol. 61, 2017.

Nonlinearity Analysis of Quantum Capacitance and its Effect on Nano-Graphene Field Effect Transistor characteristics

Declaration

I am in full confidence conform that for this manuscript we do not any conflict of interest

Munindra, Deva Nand

Electronics and Communication Engineering, Delhi Technological University, Delhi, India

Bawana Road, Shahbad Daulatpur Village, Rohini, Delhi, 110042

About Authors



MUNINDRA was born in Roorkee, Uttrakhand-India in August 1990. He received the B. Tech. and M. Tech. degrees in Electronic and Communication engineering from the Uttrakhand Technical University, Dehradun, in 2011 and National Institute of Technology (NIT) Delhi in 2016 respectively. Currently he is pursing the Ph.D. degree in Electronic and Communication engineering from Delhi Technological University, Delhi-India since 2017.



Deva Nand has completed his Ph.D. from Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. At present he is an Assistant Professor in Department of ECE, DTU, Delhi, India. A life member of ISTE, member of IAENG and IEEE. His research interests include Microelectronics, Analog VLSI Design.

Nonlinearity Analysis of Quantum Capacitance and its Effect on Nano-Graphene Field Effect Transistor characteristics

Munindra and *Deva Nand

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India
kumar.muninder90@gmail.com, * devkamboj07@gmail.com

Abstract: A simple, compact, and fundamental physics-based quasi-analytic model for Single layer graphene field effect transistors (GFETs) with large area graphene is presented in which the quantum mechanical density gradient method is utilised. The basic device physics of the two-dimensional (2D) graphene channel is studied analytically. This modeling leads to the precise drain current calculation of the GFETs. The drain current calculation for GFETs starts from charge carrier concentration, its density of states and quantum capacitance(QC). QC depends on the channel voltage as a function of gate to source voltage V_{gs} and drain to source voltage V_{ds} primarily. The formulation of the drain current with velocity saturation has been done by the Monte Carlo simulation method. The performance of the analytical GFETs model is present the precise values of QC, its impact on drain current and transfer as well as output characteristics. The impact of QC at nanometer technology adds the nonlinearity to characteristics curves. The proposed method provides better results as compared with the previous analytical and simulated results.

Keywords – quasi-analytical model, quantum capacitance, GFETs, nonlinearity

I. INTRODUCTION

Heart of the electronic devices is always the ability of a particular material to control its electronic properties in the presence of another electric field. Since the semiconductor industry is already out of scaling limits for silicon-based devices. So, this industry needs the next excellent alternate of silicon. Graphene, after its first physically stable noble experimentation in 2004 is promising to be the next alternate of silicon. Graphene is a single atomic (carbon) thin or thick layer (~3.34 x 10^{-10} m) so it is called as 2-D material [1-5], along with high charge carrier mobility [2] of value more than 10^5 cm^2/V-sec and even thermodynamically stable. The CVD grown graphene layer and graphene growth over base other then silicon could be a possible option for high frequency and analog electronic devices and circuits [4-5]. The very first physical demonstration of field emission effect in graphene layers comes in the form of an experimental device in 2007 [6]. As the technology advancement is very fast, in few next years interesting GFETs models with the intrinsic transit frequency and the maximum frequency of GFETs operation is comparable and higher than similar size CMOS nanometer technology [7-10]. As the current saturation effect in a single layer [11] and bilayer layer graphene [12], with a high value of transconductance and transistor resistance directly affect the GFETs performance. Gate/channel length of graphene has an impact on the operating frequency of the proposed GFETs model, therefore, besides the high resistance and

transconductance values, graphene RF circuits find their application in defense in real life, so this is the main focus of this particular work. Other important, compacts and fundamental physics-based quasi-analytic modelling for (GFETs) with large area graphene is also given in literature [13]. Some other work like GFETs modelling based on basic physics explores drain current, capacitance, and charge parameters for graphene channel [14-17]. Explicit compact model of drain current [14] has been extended in [15] to explore AC and transient behaviour of the GFETs. Both of these physical framework is field effect model based on drift-diffusion charge carrier transport mechanism. The Ward-Dutton charge partition scheme is implemented to model the source-drain charge together with transconductance and self-capacitance. These [16-17] compact model is benchmarked with measured current-voltage characteristic with accuracy for prototype devices and predictive behaviour. Also, few comprehensive simples and compact mathematical model provides graphene field effect transistor's physical parameters closed form formula's like g_m, r_0, $C_{gs,}$ and C_{gd}. Many other comprehensive, compact and analytical modeling of GFETs gate capacitance and quantum capacitance as dependent modeling parameter already done. But all of them either are limited to modelling and simulation of GFETs characteristics like some latest GFETs models [18-20], or are complex or are practically absurd [21-23] with no future application in device and circuits. Latest GFETs model is [24].The proposed model is keep its nature lucid mathematically, present the explicit quantum capacitance, drive the drain current formulation using it and also present the nonlinearity of characteristic curves of the GFETs at nanometer technology. Since noise is always a key element to disturb the performance of the FETs and GFETs in real world electronic device and circuits, so the study of nonlinearity to the transfer as well as output characteristics is our addition to the modelling and simulation of the GFETs. Dual gate control along with the nonlinearity study has been given in transfer characteristic curves in result and discussion section. Dual gate control to large area graphene GFETs is also a new concept for the proposed GFET modeling at 300nm technology.

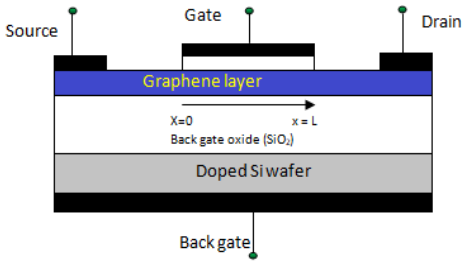
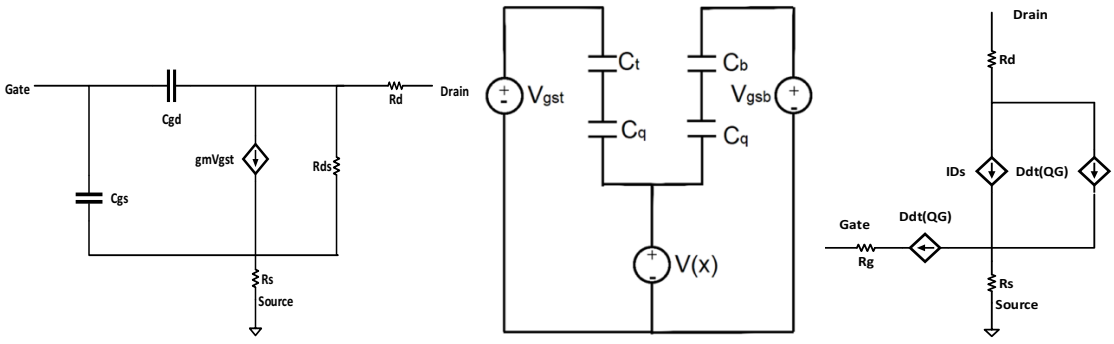


Figure 1 2-D Cross-sectional view of the proposed GFETs.



(a) Large signal (b) Capacitance (c) Small signal

Figure 2 (a), (b), (c) Equivalent circuit of the proposed GFETs (all h-parameter symbols have their usual meaning).

The structure of this paper is as follows, the detailed explanation of physic-based GFET model starts from charge carrier concentration and density of states in the section II analytical model. Furthermore, section II explains quantum capacitance which leads to the drain current formulation and help to plot the graphical results for the GFET at 1000nm and 300nm channel length. Where this is noticeable that the source is grounded. Section III discussed and compare the experimental results with the proposed model results. Finally, section IV concludes this GFET model.

II. ANALYTICAL MODEL

A. Device structure of GFETs

The proposed compact model is based on well-established dual-gate MOS structure which is shown in Fig. 1 as a 2-D cross-sectional view of proposed GFETs, which consists two gates (top and back gate), a large area graphene sheet as channel placed on the thick SiO_2 back-gate oxide layer (300 nm and 1000nm), where heavily doped silicon wafer act as a back gate. This article emphasis on p-type GFETs, while a similar approach is applicable to the n-type GFETs also. The C_{gs} and C_{gd} are gate to source and gate to drain capacitance. R_{ds}, R_d and R_s are the drain to source resistance, drain resistance and source resistance respectively as shown in Fig. 2.(a), all of these parameters nonlinear effect at nanometer technology is considered minimum since these can be control externally by suitable parameter modelling. V(x) is shown above in Fig. 2 (b) as channel voltage, which is variable and depends on x's value from 0(0) to L(V_{Ds}). While C_t, C_b and C_q are top gate capacitance, back gate capacitance and, quantum capacitance respectively and V_{gs_t} and V_{gs_b} applied internal voltages. V_{ds} and V_{BG} are the external applied drain to source and back gate voltage. Small signal equivalent circuit of the proposed GFETs is also shown in Fig.2 (c) where R_s, R_d and R_g are the source, drain and gate resistance respectively.

B. Formulation of charge carrier sheet density (ρ_{gr}) and quantum capacitance calculation (C_q)

2-D Energy and wave vector relation for the mobile π electrons in graphene can be formulated [25] as $E(k) = s\ h\ v_f|k|$ while s is equal to +1 in the conduction band and s equal to -1 in valance band for the first Brillouin Zone (BZ) in graphene layers. Where (h) is the reduced form of Planck's constant and Fermi velocity (v_f) is of high magnitude equal to $1x10^8$cm/s along with $|k|$ a 2D x-y plane wave vector at the origin $|k| = 0$ is referred to "Dirac point." Each Dirac point is double fold spin degeneracy g_s = 2 and contains double valley g_v = 2, thus the number of state in a graphene layer is

$$N(E) = g_s g_v \frac{Ak^2}{4\pi(\hbar v_f)^2} \tag{1}$$

here A is an area of the particular layer since the linear charge density of states [22] for an intrinsic graphene layer is a differential form of the number of the state of the particular layer with respect to the energy of the particular states, which can be expressed as,

$$\rho_{gr} = \frac{1}{A}\frac{dN}{dE} = \frac{2}{\pi}\frac{|E|}{(\hbar v_f)^2} \tag{2}$$

and so, to calculate the charge carriers (hole) concentration in 2D graphene sheet based on fundamental physics can be formulated as

$$p = \int_{-\infty}^{Ecv} \rho_{gr}(E)[1 - f(E)]dE \tag{3}$$

after replacing f(E) as the Fermi-Dirac distribution function and $\rho_{gr}(E)$ too, the no. of holes in the graphene sheet can be written as follows, where Ecv work as reference energy (i.e., it will be taken $Ecv = 0$)

$$p = \int_{-Ecv}^{\infty} \frac{2}{\pi} \frac{|E|}{(\hbar v_{\mathrm{f}})^2} \left[\frac{1}{exp\frac{(E+E_F)}{K_B T} + 1} \right] dE \tag{4}$$

similarly, no. of the electron in graphene channel with respect to field dependent charge carrier can be represented as

$$n = \int_{0}^{\infty} \frac{2}{\pi} \frac{|E|}{(\hbar v_{\mathrm{f}})^2} \left[\frac{1}{exp\frac{(E-E_F)}{K_B T} + 1} \right] dE \tag{5}$$

thus the total charge (Q_{net}) is equal to $(p - n)$ and could be utilized to estimate total charge sheet density Q_{gr}, which is equal to $\mathrm{q} * (\mathrm{p} - \mathrm{n})$ as follows

$$Q_{gr} = \frac{2}{\pi} \frac{q|E|}{(\hbar v_{\mathrm{f}})^2} \int_{0}^{\infty} \left[\frac{1}{exp\frac{(E-E_F)}{K_B T} + 1} - \frac{1}{exp\frac{(E+E_F)}{K_B T} + 1} \right] dE \tag{6}$$

the quantum capacitance of the proposed device is $\mathrm{C_q}$ which can be calculated [22] as the differential of the channel sheet charge density Q_{gr} with respect to V_{gr}, where V_{gr} is graphene channel voltage

$$\mathrm{C_q} = -\frac{\mathrm{d}Q_{gr}}{\mathrm{d}V_{gr}} \tag{7}$$

while minus sign here reflects the impact of more and more positive voltage applied to gate directly increase the positive channel voltage i.e. more negative charge results in the graphene layer.

$$\mathrm{C_q} = -\frac{\mathrm{d}}{\mathrm{d}V_{gr}} \left\{ \frac{2}{\pi} \frac{q|E|}{(\hbar v f)^2} \int_{0}^{\infty} \left[\frac{1}{exp\frac{(E-E_F)}{K_B T} + 1} - \frac{1}{exp\frac{(E+E_F)}{K_B T} + 1} \right] dE \right\} \tag{8}$$

To solve the integral term in the above equation take constants outside and substitute exponential term $exp\frac{(E \pm E_F)}{K_B T} \pm 1$ by x and dE by $\frac{K_B T}{x} dx$ we get simplified equation as

$$C_q = -\frac{d}{dV_{gr}}\left(\frac{2}{\pi}\frac{q}{(\hbar v f)^2}\right)\left\{\left[\int_A^{\infty}|K_B T ln x - E_F|\frac{1}{x} * \frac{K_B T}{x}\right]dx - \left[\int_B^{\infty}|K_B T ln x + E_F|\frac{1}{x} * \frac{K_B T}{x}\right]dx\right\} \quad (9)$$

a simplified quantum capacitance formulated above in Eqn. (9) can be rewritten as

$$C_q = -\frac{d}{dV_{gr}}\left(\frac{2}{\pi}\frac{q}{(\hbar v f)^2}\right)\{part\ A' + Part\ B'\} \quad (10)$$

by dividing integral into two parts and naming as $part\ A'$and $Part\ B'$ as

$$\int_A^{\infty}|K_B T ln x - E_F|\frac{1}{x} * \frac{K_B T}{x}dx \quad (11)$$

$$\int_B^{\infty}|K_B T ln x + E_F|\frac{1}{x} * \frac{K_B T}{x}dx \quad (12)$$

where A is $\exp(E_F/K_BT + 1)$ and B is $\exp(-(E_F/K_BT + 1))$, so after putting limits in exponential term and replacing x = 0 and by solving Eqn. (11) and (12) by two-part integral substitution method we get combined solution for both $part\ A'$and $Part\ B'$ and further depends on positive and negative sign as given in Eqn. (13), where $K_B T ln x \pm E_F$ as part I and $\frac{1}{x} * \frac{K_B T}{x}$ as part II,

$$\int_{A,B}^{\infty}|K_B T ln x \pm E_F| * \frac{1}{x} * \frac{K_B T}{x}dx \quad (13)$$

Now assuming E_F as constant since $E_F = q.V_{ch}$ and by following standard mathematical rules to solve this integral and after applying limits we get solved part as given below

$$\frac{(K_B T)^2}{A}(lnA + 1) - E_F\frac{(K_B T)}{A} \quad (14)$$

similarly, after solving part II

$$\frac{(K_B T)^2}{B}(lnB + 1) + E_F\frac{(K_B T)}{B} \quad (15)$$

Now by sum-up $part\ A'$and $Part\ B'$, one could represent a solved integral in quantum capacitance as follows

$$C_q = -\frac{d}{dV_{gr}}\left(\frac{2}{\pi}\frac{q}{(\hbar v f)^2}\right)\left\{\left[\frac{(K_B T)^2}{A}(lnA + 1) - E_F\frac{(K_B T)}{A}\right] - \left[\frac{(K_B T)^2}{B}(lnB + 1) + E_F\frac{(K_B T)}{B}\right]\right\} \quad (16)$$

By substituting the value of A and B, and after performing some basic algebraic mathematics, the quantum capacitance can be represented in a simple formula as given below

$$C_q = -\frac{d}{dV_{gr}}\left(\frac{2}{\pi}\frac{q}{(\hbar v f)^2}\right)\left\{\left[\frac{(K_B T)^2}{exp\frac{(E_F)}{K_B T}+1} - \frac{(K_B T)^2}{exp\frac{(-E_F)}{K_B T}+1}\right]\right\} \quad (17)$$

By making further math's effort to the above exponential terms and replacing by the hyperbolic trigonometric formulas quantum capacitance could be present as

$$\mathrm{C_q} = -\frac{2q}{\pi}\frac{(K_BT)^2}{(\hbar vf)^2}\frac{d}{dVgr}\left[\frac{2sinh\frac{qVgr}{K_BT}}{2\left(1+cosh\frac{qVgr}{K_BT}\right)}\right] \tag{18}$$

By using reverse engineering, we can write quantum capacitance [18] even more simplified as given in log term

$$\mathrm{C_q} = -\frac{2q}{\pi}\frac{qVgrK_BT}{(\hbar vf)^2}\ln\left[2\left(1+cosh\frac{qVgr}{K_BT}\right)\right] \tag{19}$$

For the $q\,V_{gr} \gg K_BT$ condition, Eqn. (19) will be simplified to the following single term formula

$$\mathrm{C_q} = -\frac{2q^2}{\pi}\frac{q|Vgr|}{(\hbar vf)^2} \tag{20}$$

As like total charge density in Eqn. (6) has been used to formulate quantum capacitance accurately, channel voltage V_{gr} and quantum capacitance $\mathrm{C_q}$ have useful significance to formulate ρ_{gr}. So, that by using basic mathematics technique of two equations and two variable drain current can be formulated by using ρ_{gr}.

$$Q_{gr} = -\int \mathrm{C_q} * dV_{gr} = -\frac{1}{2}\,\mathrm{C_q}V_{gr} \tag{21}$$

$$\rho_{gr} = \frac{Q_{gr}}{q} = \left|\frac{-\frac{1}{2}CqV_{gr}}{q}\right| \tag{22}$$

while channel voltageV_{gr} directly can be calculated by using basic technique KVL and KCL form the given Fig. 2 (c) and can be expressed as

$$V_{gr} = \frac{1}{\mathrm{C_t} + \mathrm{C_b} + 0.5\,\mathrm{C_q}}\{[V_{gs_t} - V(x)]C_t - [V_{gs_b} - V(x)]C_b\} \tag{23}$$

The pre-factor of $\frac{1}{2}$ with C_q is because of its dependency on V_{gr} and now by putting Eqn. (23) in Eqn. (22) we get a simple a second order algebraic Eqn.

$$C_t{}^2 + (C_t + C_b)C_q - [V_{gs_t} - V(x)]C_t - [V_{gs_b} - V(x)]C_b = 0 \tag{24}$$

Thus by rearrange the equation we get a quadratic equation of C_q in terms of V_{gs_t}, V_{gs_b} and $V(x)$. Where $V(x)$ is the channel voltage depend on x (0 to L) and the value of $V(x)$ of at (x = 0) is 0V, while at (x = L) is V_{ds}V.

C. Drain current calculation

For drain current calculation at very first we need to focus on charge charier velocity $v = v(x)$ mainly through the channel (x =0 to L) from (1). A low saturation velocity characteristic is shown by Monte Carlo simulation [26] at steady state is

$$v = \frac{\mu E}{1+\frac{\mu|E|}{v_{sat}}} \tag{25}$$

Where E represents the electric field, μ is the carrier mobility and v_{sat} is saturation velocity at low field. The carrier velocity, especially v_{sat} of graphene channel at the SiO_2 platform is strongly affected by rare interfacial scattering of phonons. Experimental work has been talked about the two dominated optical phonons at SiO_2 and graphene interface. They have been shown a magnitude of 59meV and 155meV respectively but one can describe this phenomenon by a single phonon only with enough accuracy and with the energy of $\hbar\Omega$. By the approach used [27] to calculate v_{sat}, v_{sat} is depends on charge carrier sheet density (ρ_s) and the phonons effective energy Ω.

$$v_{sat} = \frac{\Omega}{\sqrt{\pi\rho_{gr}}} \tag{26}$$

Charge carrier-interaction phenomenon in graphene-MOSFET channel introduces a new term as $0.5 + A.V^2(x)$ [26] to correct the calculation of v_{sat} to one decimal point at the channel drain end. Now Eqn. (26) will become as,

$$v_{sat} = \frac{\Omega}{(\pi\rho_{gr})^{0.5+A.V^2(x)}} \tag{27}$$

here A is a dimensionless important factor of order 10^{-3}, when ρ_s is of magnitude 10^{12} cm^{-2} and the final unit will be cm/s. Now using E = $(-\frac{dV}{dx})$ and Eqn. (20) drain current (I_d) becomes

$$I_d = -q\rho_{gr}\frac{\mu(-\frac{dV}{dx})}{1+\frac{\mu(-\frac{dV}{dx})}{v_{sat}}}W \tag{28}$$

after integrating left side 0 to x and 0 V to V_{ds} V right side, since $V = V(x)$, is a function of the channel length and by changing the variable of separation I_d becomes

$$I_d = q\mu W\frac{\mu\int_0^{V_{ds}}\rho_{gr}dV}{L-\mu\int_0^{V_{ds}}\frac{1}{v_{sat}}dV} \tag{29}$$

$$I_d = -\frac{\mu W}{2L_{eff}} * \frac{C_t}{C_t + C_b + 0.5\,C_q}\{(V_{gs_t} - V(x)/2) - (V_{gs_b} - V(x)/2)\} \tag{30}$$

$$L_{eff} = L - \frac{\mu}{\Omega}\left\{\left[\log(\frac{\pi C_q}{2q}) + \log[(V_{gs_t} - V_{ds})C_t + (V_{gs_b} - V_{ds})C_b]\right] * \left[0.5\,V_{ds} + \frac{A\,{V_{ds}}^3}{3}\right] + \int_0^{V_{ds}}\{(\frac{2q}{\pi C_q}) + \frac{(C_t+C_b)}{[(V_{gs_t} - V_{ds})C_t + (V_{gs_b} - V_{ds})C_b]}\} * \left[0.5\,V_{ds} + \frac{A\,{V_{ds}}^3}{3}\right] \tag{31}$$

III. Results and Discussion

To investigate the quantum capacitance effect on gate capacitance, a basic MOS structure without the back gate of oxide thickness (300nm) and top gate oxide of the thickness (3nm,), zero voltage as V_{gs_t}, V_{gs_b} and V_{Ds} is arranged for the proposed GFETs model. Quantum

capacitance of the amplitude of the 10th multiple micron is enhanced overall gate capacitance with respect to the relation $C_q = C_{ox} * C_q/(C_q + C_{ox})$.

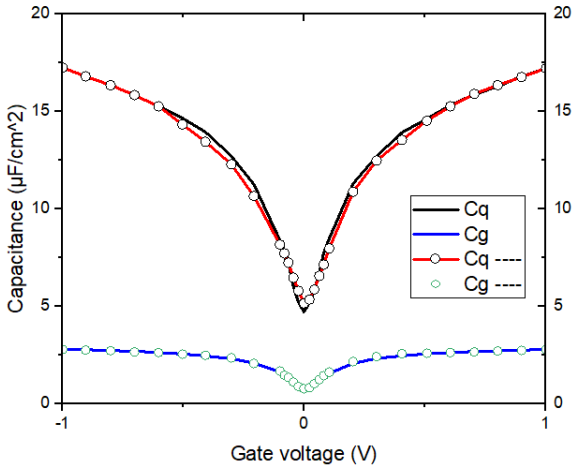


Figure 3 Quantum capacitance effect on overall gate capacitance vs gate voltage. Line graph is simulated proposed model results, while symbolic results graph are analytical results and this is same for all presented result graphs.

One of the important aspects of nanometer technology GFETs is that its quantum capacitance plays an important role in terms of device characteristics. Fig. 3 elaborate graphically how the quantum capacitance has an impact on the total gate capacitance at 300nm channel length. The proposed model provides better results analytical (with line) and the simulated (with symbol) as compared with [11] for QC as well as gate capacitance at 300nm. Table 1 explains the numerical values of all important parameters and dimensions of the GFETs model.

Table 1 Dimensions and parameters of the modeled GFETs devices

Parameter	**GFET 1 [11]**	**GFET 300nm(proposed)**
L (μm)	1	.3
W (μm)	2.1	2.1
t_{ox} (nm)	15	15
t_{box} (nm)	285	285
$V_{gs\text{-}top}$ (V)	1.45	.95
$V_{gs\text{-}back}$ (V)	2.7	2.7
p	$1.5*10^{12}$	$3*10^{12}$
μ_n (cm^2/V.s)	1500	2500
μ_p (cm^2/V.s)	1500	2500
H (meV)	56	99
λ (nm)	10-100	100-300

Fig. 4 and Fig. 5 have been explained graphically the transfer characteristic and output characteristic of the proposed GFETs model and compare it with the experimental work [11] mentioned as GFET1 in table 1. GFET1 featuring a high dielectric 2D gate material HfO_2 on exfoliated graphene channel and SiO_2 is the back-gate dielectric. The output characteristic of the proposed model in Fig. 5, has been shown coherence with the [11], with the applied voltage at source (0V), negative drain voltage (-2 to 0V), top-gate (sweep from -0.75 to 0.75 V) as it

can easily view in a graph with the back-gate voltage of -40 V. In the same manner transfer characteristic of the GFETs both analytical and simulated is compared with [11] in Fig. 5.
The nonlinearity characteristic behavior of the GFET at 300nm channel has been shown in Fig. 4,5 and 6. Dirac-point shift has been shown in Fig.4 after applying external gate to source voltage for single layer graphene, while for the conventional MOSFETs Dirac point has same contact point of Conduction and valance band. This Dirac-point shift in transfer characteristic and kink presented in the output characteristics is a proof of nonlinear behavior of the GFETs. This nonlinear behavior is tending to enhanced when the channel length of the GFETs is being decreased from 1000nm to 100nm and below that value as this is given in the latest accepted paper [24]. The dual gate control has been graphically present in the Fig. 6 at same time nonlinear behavior of the GFETs can be observe very easily from Fig 6. e and f. Here we concentrate on the nonlinearity of the transfer characteristics of the Proposed model. A flat line drain current has been shown in the fig. 6.e for -0.1 to 0.1V gate to source voltage and beyond this it has shown gradual increment, which is normal in electronic device characteristic curves but a flat drain current shows a nonlinearity in the transfer characteristic curve.

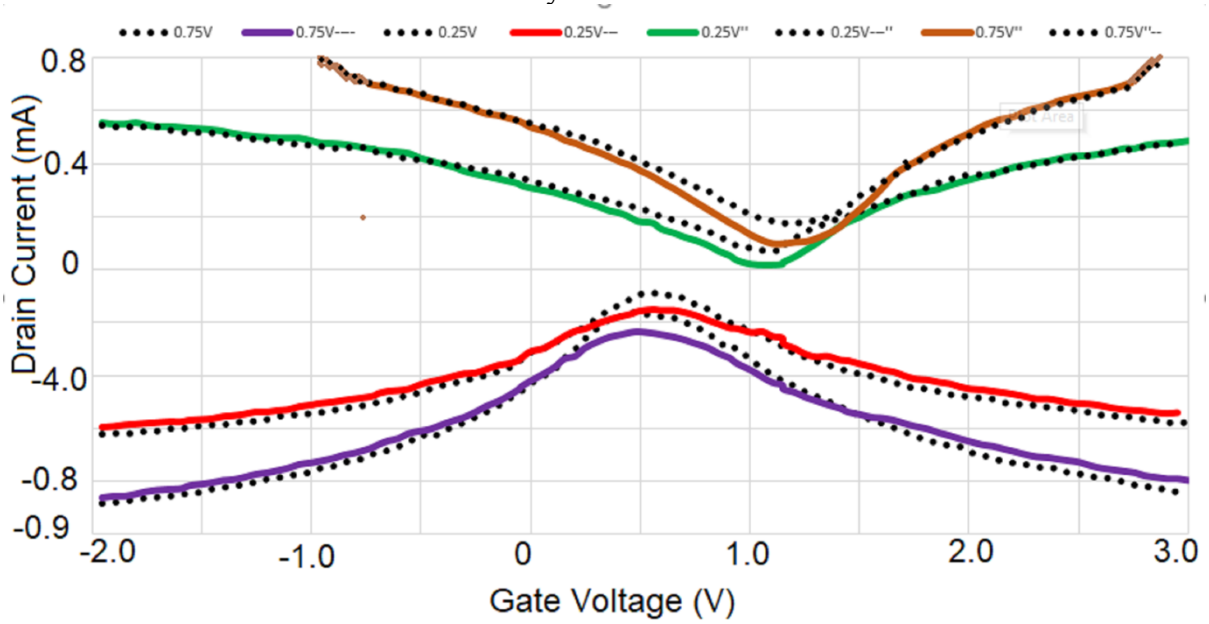


Figure 4 Transfer characteristic of the proposed model i.e. drain current vs top-gate voltage for the proposed GFETs model.

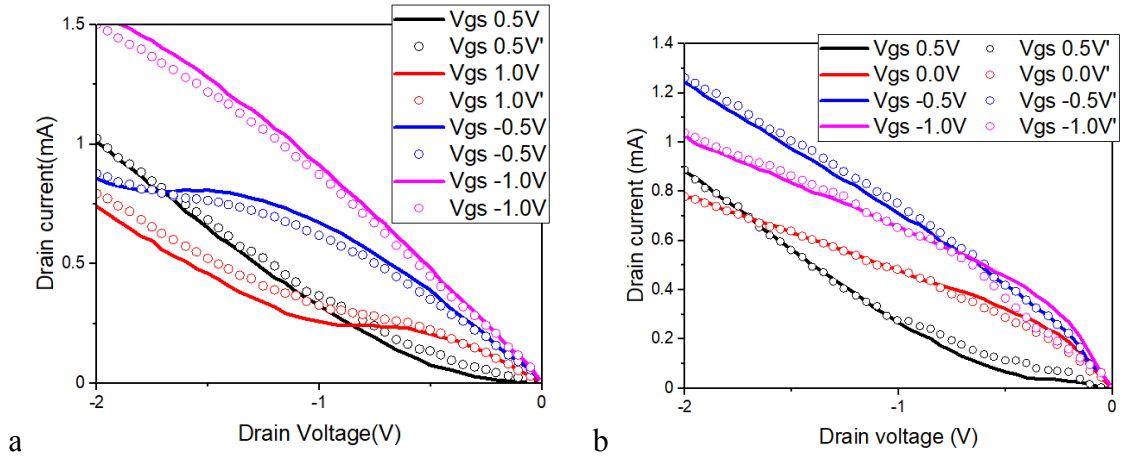


Figure 5.a and 5.b: Output characteristic of the proposed model i.e. drain current vs drain to source voltage for the GFETs model at 1000nm in 5.a and 300nm in 5.b.

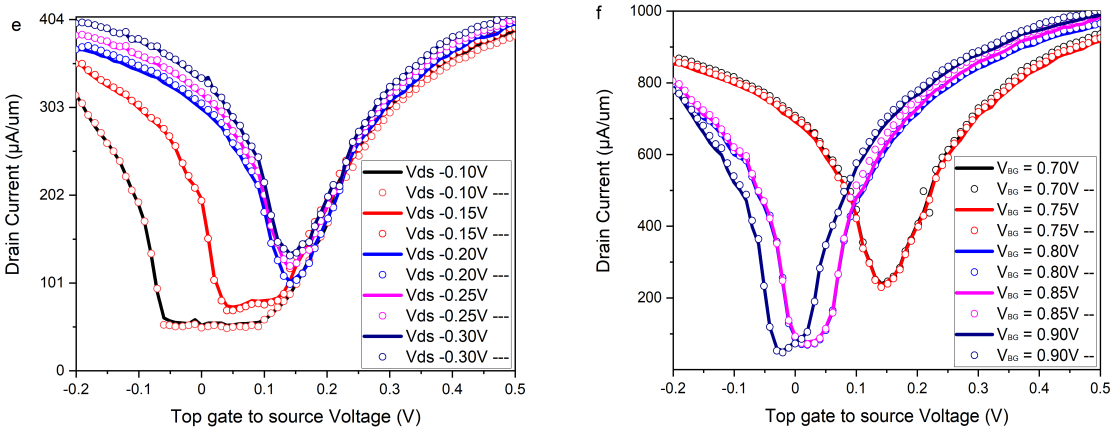


Figure 6.e and 6.f: Transfer characteristic of the proposed model i.e. drain current vs drain to source voltage for the GFETs model. Top gate control by V_{ds} and back gate control by V_{BG} over drain current is presented in this graph (while for this dual gate control an exception is considered that the graphene sheet is not a large area sheet). V_{ds} and V_{BG} are the external applied drain to source and back gate voltage.

IV. Conclusion

A quasi-analytic model for single layer GFETs with large area graphene is presented. Drain current formulation with velocity saturation effect and role of quantum capacitance over the total capacitance is explained with the precise output and transfer characteristics of the proposed GFET model, which facilitate the readers a simple mathematical GFET model. The effect of enhanced quantum capacitance value at nanometer technology adds nonlinearity to the transfer and output drain characteristics of the proposed GFETs model has been presented. Thus the modeling and simulation of GFETs model has present and compare the results with the well established experimental work and it will surly befits the reader to understand basic physic of the GFETs.

References

[1] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, and S. V. Dubones, I. V. Grigorieva and A.A. Firsov "Electric Field Effect in Atomically Thin Carbon Films," Sci. Report, vol. 306, no. 5696, pp. 666–669, 2004.

[2] K. S. Novoselov and A. K. Geim, "The rise of graphene," Nat. Mater., vol. 6, no. 3, pp. 183-191, 2007.

[3] F. Schwierz, "Graphene transistors," Nat. Nanotechnol., vol. 5, no. 7, pp. 487–496, 2010.

[4] A. K. Geim, "Graphene: status and prospects" Science, vol. 324, no. 5934, pp. 1530-1534, 2009.

[5] K. S. Novoselov, V. I. Falko, L. Colombo, P. R. Gellert, M. G. Schwab and K. kim "A roadmap for graphene," Nature, vol. 490, no. 11, pp. 192–200, 2012.

[6] M. C. Lemme, T. J. Echtermeyer, M. Baus, and H. Kurz, "A graphene field-effect device," *IEEE Electron Device Lett.*, vol. 28, no. 4, pp. 282–284, 2007.

[7] S. J. Han, K. A. Jenkins, A. Valdes Garcia, A. D. Franklin, A. A. Bol, and W. Haensch, "High-frequency graphene voltage amplifier," *Nano Lett.*, vol. 11, no. 9, pp. 3690–3693, 2011.

[8] Y. Wu, D. Farmer, A. Valdes-Garcia, W. Zhu, K. Jenkins, C. Dinmitrakopoulos, et al. " Record High-frequency graphene voltage amplifier," *IEEE IEDM.*, vol. 11, no. 11, pp. 528–530, 2011.

[9] Y. M. Lin, D. B. Farmer, K. A. Jenkins Y. Wu., J. L.Tedesco, et al., "Enhanced Performance in epitaxial graphene FETs with optimized channel morphology," *IEEE*

Electron Device Lett., vol. 32, no. 10, pp. 1343–1345, 2011.
[10] L. Liao, Y. C. Lin, M. Bao, R. Cheng, J. Bai, Y. Liu, et al. "High-speed graphene transistors with a self-aligned nanowire gate," Nature, vol. 467, no. 9, pp. 305–308, 2010.
[11] I. Meric, M. Y. Han, A. F. Young, B. Ozyilmaz, P. Kim, and K. L. Shepard, "Current saturation in zero-bandgap, top-gated graphene field-effect transistors," *Nat. Nanotechnol.*, vol. 3, no. 11, pp. 654–659, 2008.
[12] B. N. Szafranek, G. Fiori, D. Schall, D. Neumaier, and H. Kurz, "Current Saturation and Voltage Gain in Bilayer Graphene Field Effect Transistors," vol. 12, no. 3, pp. 1324–1328, 2012.
[13] S. A. Thiele, J. A. Schaefer, and F. Schwierz, "Modeling of graphene metal-oxide-semiconductor field-effect transistors with gapless large-area graphene channels," *J. Appl. Phys.*, vol. 107, no. 9, pp. 094505-1–8, 2010.
[14] D. Jiménez and O. Moldovan, "Explicit drain-current model of graphene field-effect transistors targeting analog and radio-frequency applications," IEEE Trans. Electron Devices, vol. 58, no. 11, pp. 4049–4052, 2011.
[15] D. Jiménez, "Explicit drain-current, Charge and Capacitance model of graphene field-effect transistors," IEEE Trans. Electron Devices, vol. 58, no. 12, pp. 4377–4383, 2011.
[16] I. Z. Gennady, A. T. Alexander, K. B. Daria and V. M. Evgeny "Small-Signal Capacitance and Current parameter Modeling in Large-Scale Hgh-Frequency Graphene Field Effect Transistors," IEEE Trans. Electron Devices, vol. 60, no. 6, pp. 1799–1806, 2013.
[17] J. D. A. Morales, S. Frégonèse, C. Mukherjee, C. Maneux and T. Zimmer, "An Accurate physics-based Compact Model for Dual-gate Bilayer Graphene FETs," IEEE Trans. Electron Devices, vol. 62, no. 12, pp. 4333–4339, 2015.
[18] Abhishek Kumar Upadhay, Ajay kumar Kuashwaha and Santosh Kumar Vishvakarma "A Unified Scalable Quasi-Ballistic Transport Model of GFET for circuit Simulations," IEEE Trans. Electron Devices, vol. 65, no. 2, pp. 739-746, 2018.
[19] Abhishek Kumar Upadhay, Ajay kumar Kuashwaha, Priyank Rastogi, Yogesh Singh Chauhan and Santosh Kumar Vishvakarma "Explicit Model of Channel Chagre, Backscattering, and Mobility for Graphene FET in Quasi-Ballistic Regime," IEEE Trans. Electron Devices, vol. 65, no. 10, pp. 739-746, 2018.
[20] Nagothu Karmel Kranthi, Abhishek Mishra, Adil Meersha, and Mayank Shrivastava , "On the ESD Behavior of Large-Area CVD Graphene Transistors: Physical Insights and Technology Implications ," IEEE Trans. Electron Devices, vol. 66, no.1, pp. 743-751, 2019.
[21] F. Pasadas, and D. Jimenez, "Large-signal Model of Graphene Field Effect Transistor-Part I: Compact modeling of GFET Intrinsic Capacitance," IEEE Trans. Electron Device, vol. 63, no. 7, pp. 2936–2941, 2016.
[22] F. Pasadas, and D. Jimenez, "Large-signal Model of Graphene Field Effect Transistor-Part II: Circuit Performance Benchmarking," IEEE Trans. Electron Device, vol. 63, no. 7, pp. 2942–2947, 2016.
[23] J. J. Ebers and J. L. Moll, "Large-Signal behavior of junction transistors," *Proc. IRE*, vol. 42, no. 12, pp. 1761–1772, 1954.
[24] Munindra and Deva Nand, "Nonlinearity and scaling trends of quasiballistic graphene field-effect transistors targeting RF applications," *Journal of Coputational Electronics*, 10.1007/s10825-021-01772-x, Aug 2021
[25] T. Fang, A. Konar, H. Xing, and D. Jena, "Carrier statistics and quantum capacitance of graphene sheets and ribbons," Applied Physic Letters, Vol. 91, no.092109-1-8, 2007.
[26] A. Akturk and N. Goldsmen, "Electron transport and full-band electron-phonon interactions in graphene," Journal of Applied Physics, Vol. 103, no. 0537021-1-7, 2008.

[27] I. Meric, N. Baklitskaya, P. Kim and K. L. Shepard, “RF performance of top-gated, zero-bandgap graphene field-effect transistors,” IEDM, Vol. 21, no. 2, pp.15-18 2008.

Silicon
https://doi.org/10.1007/s12633-021-01395-8

ORIGINAL PAPER

Numerical Study of JAM-GS-GAA FinFET: A Fin Aspect Ratio Optimization for Upgraded Analog and Intermodulation Distortion Performance

Bhavya Kumar[1] · Rishu Chaujar[1]

Received: 1 September 2021 / Accepted: 13 September 2021
© Springer Nature B.V. 2021

Abstract
This paper optimizes the fin aspect ratio (AR) of Junctionless Accumulation Mode Gate Stack Gate All Around (JAM-GS-GAA) FinFET with constant conducting channel area for upgraded static, analog, and distortion performance. The crucial static and analog parameters, for instance, electric field, potential, band diagram, electron concentration, transconductance, quality factor, drain current, early voltage, intrinsic gain, output resistance, and output conductance, are evaluated and analyzed. The simulated results reveal that the static and analog performance of JAM-GS-GAA FinFET enhances with the increase in the fin AR. Besides, linearity and distortion parameters like voltage intercept points (VIP2, VIP3), 1-dB compression point, zero crossover point, harmonic distortions (HD2, HD3), and total harmonic distortion are explored. It has been noticed that a low fin AR lessens the distortion in the device and improves the device's linearity performance. Moreover, the zero crossover point decreased significantly for the device with low fin AR, which reduces the device operation optimum bias point. Thus, the findings of this paper can help engineers to design 3-D devices according to their needs.

Keywords Distortion Performance · Gate All Around (GAA) FinFET · Junctionless Accumulation Mode (JAM) · Fin Aspect Ratio (AR) · Gate Stack (GS) · Analog Performance

1 Introduction

In modern years, the desire for higher battery backup and portable electronic devices has increased significantly, leading to a gradual reduction in the transistor size [1]. The electronic device's efficiency reduces severely with the transistor size reduction due to unwanted short channel effects (SCEs) [2–4]. Researchers over the years have proposed several multi-gate (MG) device structures, and FinFET has appeared as one of the most encouraging devices to alleviate the SCEs [5]. However, at such small dimensions, the fabrication of sharp junctions is not a straightforward task. The junctionless field-effect transistors (JLFETs) are economical and easy to fabricate [6]. In JLFETs, the problem arises in depleting the channel completely to achieve excellent turn-off characteristics during the OFF state [7]. The relatively low doping concentration in the channel region resolves the issue, but it also decreases drain current accompanied by increased source/drain (S/D) series resistance. Thus, Junctionless Accumulation Mode (JAM) FETs with enhanced source/drain doping are considered [8, 9].

During complementary metal oxide semiconductor (CMOS) scaling, reducing gate oxide thickness beyond a certain level leads to an enhancement in the tunneling gate current [10, 11]. The application of high-κ dielectric materials on a silicon substrate is the best technical solution, although it contains drawbacks like mobility degradation, threshold voltage instability, etc. [12, 13]. Therefore, a Gate Stack (GS) configuration providing a narrow SiO_2 layer to stabilize the high-κ dielectric and silicon substrate without affecting the device's effective oxide thickness is implemented [14]. The device's electrostatic control increases with the addition of the number of gates over the channel. Thus, Gate All Around (GAA) arrangement was introduced to enhance the device performance and sub-threshold characteristics [15, 16]. As a consequence, we have proposed a JAM-GS-GAA FinFET.

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in

Bhavya Kumar
bhavyakumar_phd2k18@dtu.ac.in

[1] Department of Applied Physics, Delhi Technological University, Delhi, India

Published online: 18 September 2021

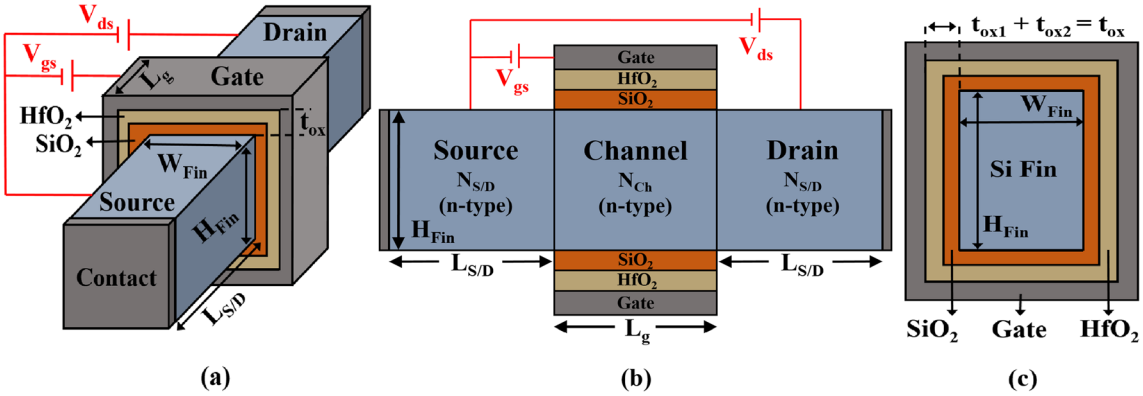


Fig. 1 (**a**) The simulated JAM-GS-GAA FinFET 3-dimensional structure (**b**) horizontal (**c**) vertical 2-dimensional view slit through the device's silicon fin

The System-on-Chip (SoC) technology can handle the increased passive components and transistor density in a single chip to achieve suitable device performance [17]. The system elements are independently incorporated on a semiconductor chip, and optimization of the device is an uphill task with the SoC concept. Therefore, this study aims to determine the optimized fin aspect ratio for upgraded analog and intermodulation distortion performance of the JAM-GS-GAA FinFET. In recent studies, the researchers have demonstrated that alteration of the geometrical parameters such as oxide thickness (t_{ox}), fin width (W_{Fin}), gate length (L_g), fin height (H_{Fin}) primarily influence the FinFET performance [17–19]. The aspect ratio (AR) is an essential geometric parameter in FinFET technology. It is described as the ratio of W_{Fin} to H_{Fin} in a few research papers [17, 20], but we have considered the more precise definition: the ratio of H_{Fin} to W_{Fin}. Because when H_{Fin} is greater than W_{Fin}, the sidewall surface dominates over the top surface, and the device acts as a 3D device. In contrast, the device acts as a planar device when W_{Fin} is greater than H_{Fin} due to the dominance of top surface orientation [21]. In the research conducted on the fin AR, the effective channel area ($H_{Fin} \times W_{Fin}$) has not been kept constant to a particular value while varying W_{Fin} and H_{Fin}. In this scenario, all the related electrical parameters and the channel's conductance will be altered [22]. Therefore, we have kept the device's effective channel area constant to get the exact effect of the fin AR on crucial static, analog, and intermodulation distortion parameters while optimizing the fin AR.

This work is arranged as Section 2 outlines the device structure, Section 3 describes the simulation framework and physical models, Section 4 discusses the experimental calibration and fabrication feasibility of the proposed JAM-GS-GAA FinFET. Section 5 analyses the influence of the fin aspect ratio on the device's various static, analog, and distortion performance parameters with a conclusion note in Section 6 validating the uniqueness of the work.

2 Device Structure

The JAM-GS-GAA FinFET 3-dimensional structure is presented in Fig. 1(a). In contrast, the horizontal and vertical 2-dimensional view slit through the device's silicon fin is represented in Fig. 1(b) and (c). Silicon material is used in the entire fin region. The length of the source/drain regions ($L_{S/D}$) and gate (L_g) are kept constant at 10 nm and 7 nm, respectively. The thickness of the oxide (t_{ox}) is 1 nm with t_{ox1} = 0.5 nm (SiO_2, κ = 3.9) and t_{ox2} = 0.5 nm (HfO_2, κ = 25) [23] are employed to stack the gate oxide. The concentration of the doping in the channel region (N_{Ch}) is 1×10^{16} cm^{-3}, while it is 1×10^{19} cm^{-3} in the source/drain regions ($N_{S/D}$). Every region is doped uniformly with n-type. Due to the presence of the poly-depletion effect in polysilicon gates, metal gates are used these days. Besides that, metal gates have high thermal stability, high purity, and they are also compatible with CMOS processing [24, 25]. Therefore, titanium nitride (TiN) is used as the metal gate with a work function (ϕ_m) equal to 4.65 eV [26]. The effective channel area

Table 1 Different device configurations used for simulation

Config. No.	Fin Height, H_{Fin} (nm)	Fin Width, W_{Fin} (nm)	Effective Channel Area $H_{Fin} \times W_{Fin}$ (nm^2)	Fin Aspect Ratio, AR (H_{Fin}/W_{Fin})
C1	10	8	80	1.25
C2	16	5	80	3.2
C3	20	4	80	5

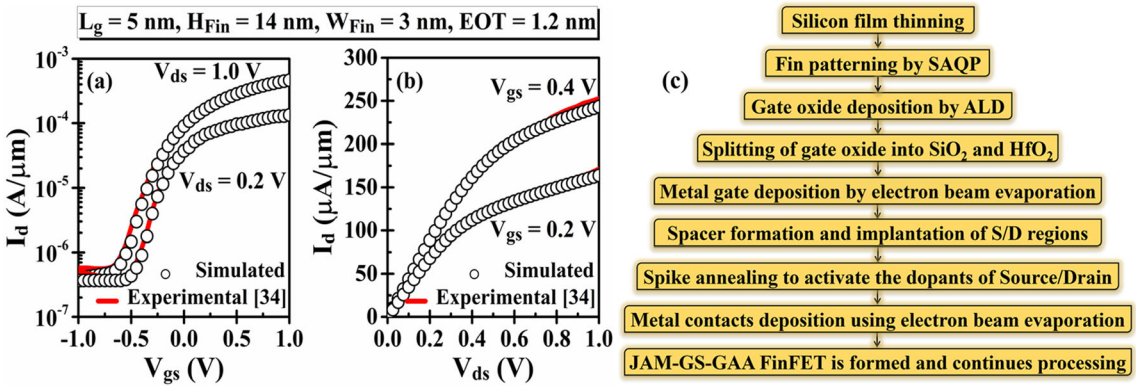


Fig. 2 Experimental and simulated (**a**) transfer (**b**) output characteristics of the All-Around-Gate (AAG) Si FinFET [34] (**c**) JAM-GS-GAA FinFET step-by-step fabrication outline

is constant at 80 nm^2 while varying the fin width and height. Table 1 portrays the three different configurations used for simulation placed in increasing fin aspect ratio order.

3 Simulation Framework and Physical Models

The simulations have been performed in the SILVACO ATLAS 3D simulator [27]. The gate-source voltage (V_{gs}) is varied from 0 to 1 V, whereas drain-source voltage (V_{ds}) and temperature (T) is fixed at 0.5 V and 300 K, respectively, during the entire simulations. Equations (1–5) describes the different physical models included in the device simulation to achieve accurate results.

A. Shockley-Read-Hall (SRH) recombination model includes the effect of recombination and generation [28, 29].

$$R_{SRH} = \frac{pn - n_{ie}^2}{\tau_p\left[n + n_{ie}\exp\left(\frac{E_{Trap}}{kT_L}\right)\right] + \tau_n\left[p + n_{ie}\exp\left(\frac{-E_{Trap}}{kT_L}\right)\right]} \quad (1)$$

where k signifies the Boltzmann's constant, τ_n and τ_p denotes the electron and hole lifetimes, E_{TRAP} is the difference between intrinsic Fermi level and trap energy level, and T_L is the lattice temperature.

B. Arora analytical model correlates the low-field carrier mobility with impurity concentration and temperature [30].

$$\mu_n = 88\left(\frac{T_L}{300}\right)^{-0.57} + \frac{1252\left(\frac{T_L}{300}\right)^{-2.33}}{1 + \frac{N}{1.432\times10^{17}\left(\frac{T_L}{300}\right)^{2.546}}} \quad (2)$$

where N is the total local dopant concentration and T_L is the lattice temperature.

C. Klaassen band to band tunneling model accounts for the electrons tunneling between valence and conduction band [31].

$$G_{BBT} = D \times BB.A \times E^{BB.GAMMA}\exp\left(-\frac{BB.B}{E}\right) \quad (3)$$

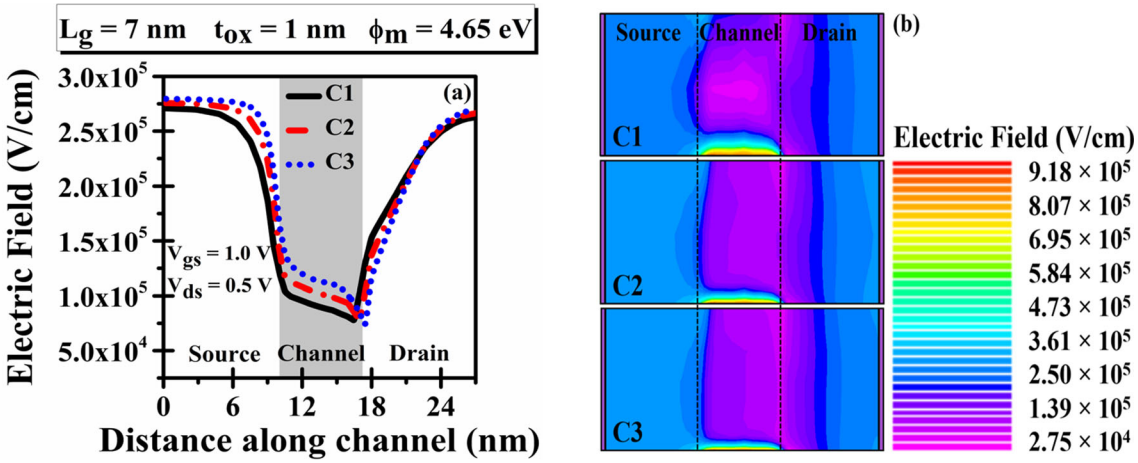


Fig. 3 (**a**) Change in the electric field with the channel distance (**b**) electric field contour plot for each configuration considered

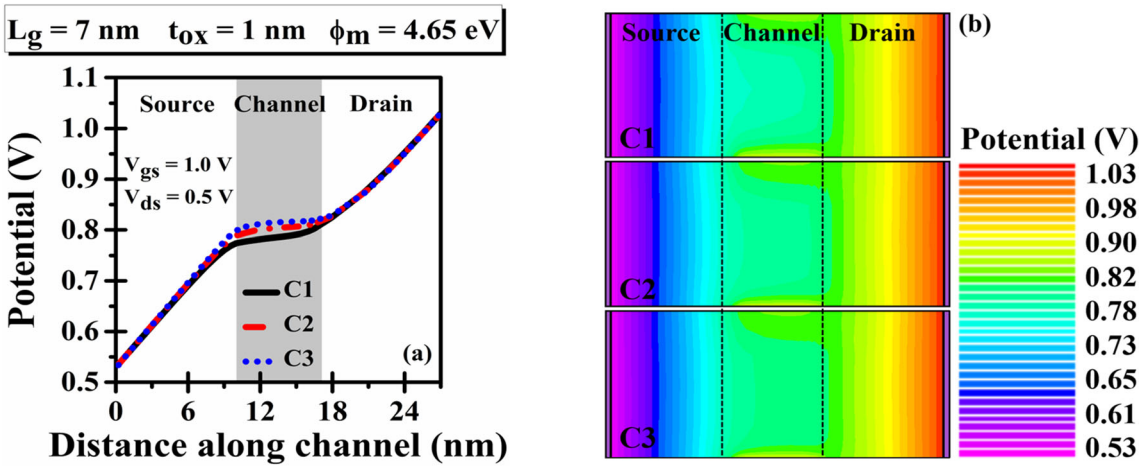


Fig. 4 (**a**) Potential profile variation along the channel distance (**b**) potential contour of all three configurations

where E signifies the electric field magnitude, D denotes statistical factor, and BB.A, BB.B, BB.GAMMA, are user-definable parameters.

D. Crowell-Sze impact ionization model introduces the impact ionization effects [32].

$$\alpha_{n,p} = \frac{1}{\lambda}\exp\left[C_0(r)+C_1(r)x + C_2(r)x^2\right] \quad (4)$$

where λ denotes the carrier mean free path for optical phonon generation.

E. Fermi-Dirac statistics model enhances the result accuracy [33].

$$f(\varepsilon) = \frac{1}{1+\exp\left(\frac{\varepsilon - E_F}{kT_L}\right)} \quad (5)$$

where E_F indicates the Fermi level and ε represents energy.

Additionally, Newton and Gummel's methods are used to achieve a solution [27].

4 Experimental Calibration and Fabrication Feasibility

The All-Around-Gate (AAG) Si FinFET is calibrated with the experimental data extracted from Hyunjin Lee et al. [34] to validate the above-discussed physical models. We have calibrated the experimental data by considering silicon material in the entire fin region with fixed device dimensions (L_g = 5 nm, H_{Fin} = 14 nm, and W_{Fin} = 3 nm) as mentioned in the paper to authenticate the simulations. Figure 2(a) and (b) describes the experimental and simulated transfer and output characteristics of the All-Around-Gate (AAG) Si FinFET, respectively. The selection of simulation models is validated due to the close agreement between the experimental and simulated data sets.

Furthermore, the step-by-step device fabrication outline of JAM-GS-GAA FinFET is displayed in Fig. 2(c) to demonstrate the fabrication feasibility of the proposed device. The initial step is to thin the silicon film followed by fin patterning using the self-aligned quadruple patterning (SAQP) method [35]. When two self-aligned double patternings (SADP) are applied in a row to enhance the feature density, it is known as

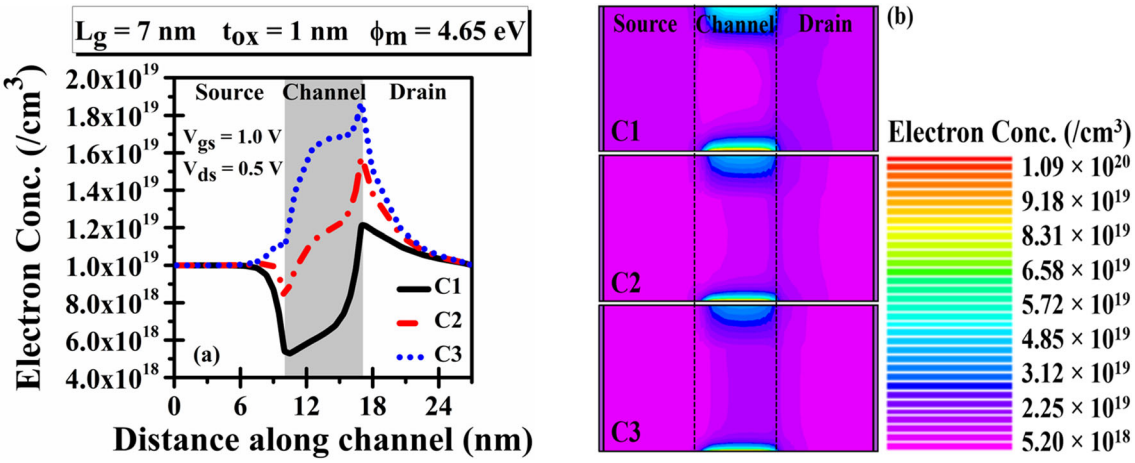


Fig. 5 (**a**) Electron concentration plot against the distance along the channel (**b**) electron concentration contour profile of each simulated device

SAQP. The gate dielectric (SiO_2/HfO_2) is deposited on the silicon interfacial layer by atomic layer deposition (ALD) [36]. Then TiN metal gate is deposited using electron beam evaporation at room temperature on the top of the gate dielectric [36]. The source and drain regions are implanted and then spike annealed to activate the source and drain region dopants. The source/drain metal contacts are deposited by electron beam evaporation followed by lift-off. The JAM-GS-GAA FinFET is formed and continues processing.

5 Results and Discussion

5.1 Fin Aspect Ratio Optimization for Static Performance

It is essential to evaluate the vital static parameters like the electric field, energy band profiles, potential, electron concentration, etc., to precisely examine the device performance. Therefore, in this subsection, discussed static parameters are examined to optimize the fin AR for superior device performance. The electric field alteration with the channel distance for each configuration is displayed in Fig. 3(a). A lower electric field is acquired at the drain end compared to the source end. In addition, because of the almost same electric field at both drain and source end for each configuration, the channel region charge carriers accelerate, resulting in an enhanced electron injection velocity from the source to the channel region [37–39]. A higher electric field is witnessed for the C3 configuration compared to their counterparts in the channel region. Figure 3(b) reveals the contour plot of the electric field for all three configurations, which confirms the rise in the electric field with the fin AR increase.

The potential profile along the channel distance for the respective configurations is plotted in Fig. 4(a). Due to the lower electric field at the drain end compared to the source end, potential improves significantly in the channel region for the C3 configuration, indicating a boost in the device potential with the increase in the fin AR. The equivalent potential contour for all three configurations is presented in Fig. 4(b), which confirms the increase in potential with the rise in the fin AR. Figure 5(a) portrays the plot of electron concentration against the distance along the channel for each mentioned configuration. It is evident from the graph that for the C3 configuration, electron concentration increases massively in the channel region due to the decreased electric field at the drain end compared to the source end. This surge in the concentration of electrons inside the channel region with the fin AR enhances the current drivability and improves the leakage current. The electron concentration contour profile of each simulated device is presented in Fig. 5(b).

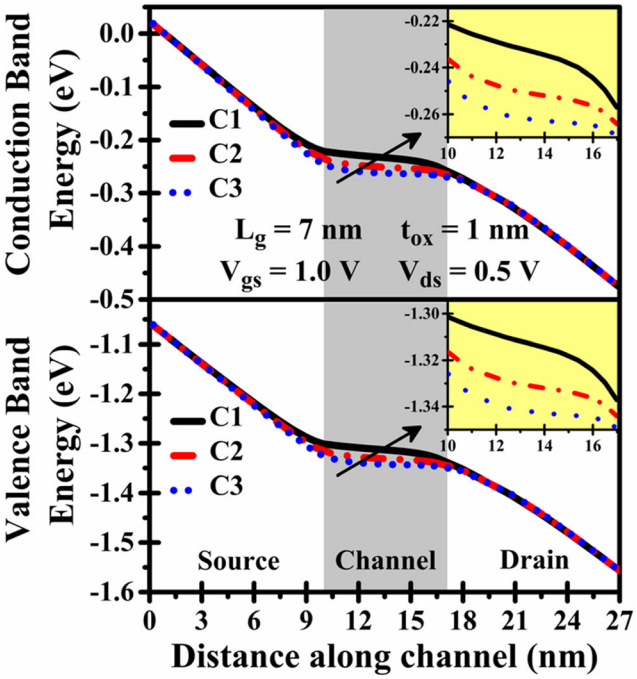


Fig. 6 Conduction and valence band energy against the distance across the channel for all the three simulated devices

Conduction and valence band energy is extracted and represented against the distance across the channel for all the three simulated devices in Fig. 6. It is observed that in the channel region, conduction and valence band energy reduces with the rise in the fin AR. The channel region inset graph clearly demonstrates the decrease in the conduction and valence band energy with fin AR, indicating improvement in the ON and OFF current of the device. Thus, discussed static parameters improve appreciably with the fin AR increase, i.e., enhanced device performance is obtained for the C3 configuration.

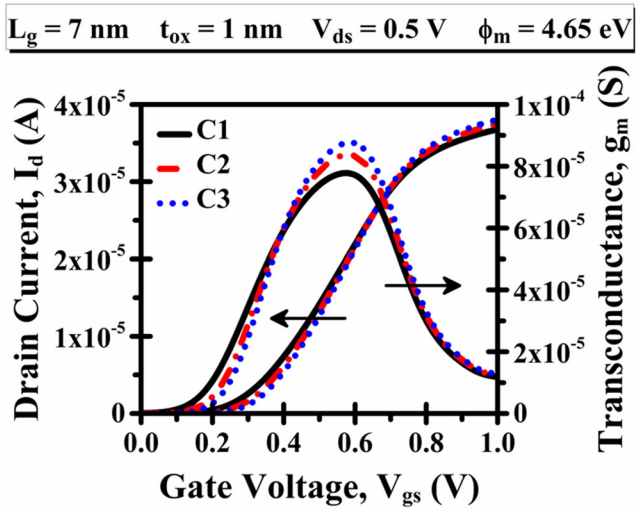


Fig. 7 Plot of I_d and g_m against V_{gs} at a static V_{ds} for each configuration

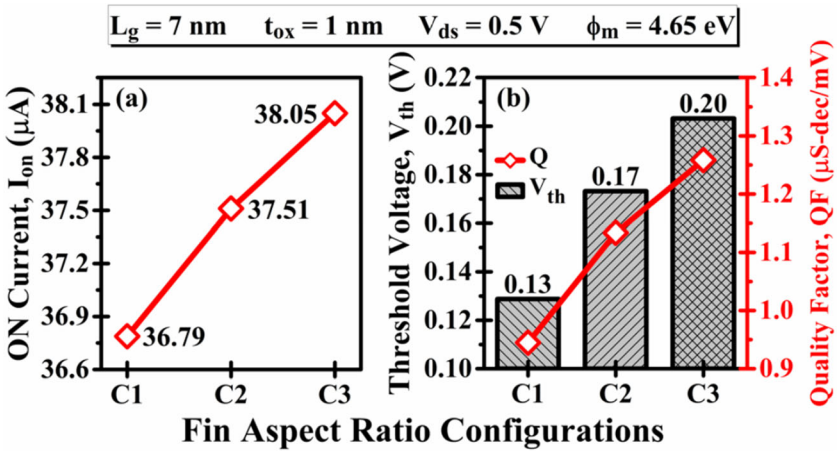


Fig. 8 Change in (**a**) ON current (**b**) threshold voltage and quality factor against different fin aspect ratio configurations

5.2 Fin Aspect Ratio Optimization for Analog Performance

This section optimizes the fin AR for enhanced device analog performance by exploring crucial analog parameters, for instance, transconductance (g_m), quality factor (QF), drain current (I_d), early voltage (V_{EA}), intrinsic gain (A_v), etc. The collaborated plot of I_d and g_m against V_{gs} at a static V_{ds} for each configuration is displayed in Fig. 7. It is observed that compared to the other two simulated structures, the C3 configuration acquires the maximum value of drain current because the current drivability of the device rises with the increase in fin AR. Transconductance (g_m) measures the deviation in the drain current to the gate-source voltage alteration at uniform drain-source voltage, i.e., $g_m = \partial I_d/\partial V_{gs}$ [40]. Figure 7 confirms that transconductance rises with the fin AR increase because of the enhanced drain current and lower electric field at the drain end. Consequently, the C3 configuration displays a more excellent transconductance value than the C1 and C2 configurations.

Figure 8(a) displays the ON current (I_{on}) of the respective configurations, increasing with the fin AR rise. The I_{on} observed for the C3 configuration is 38.05 μA, whereas it is 36.79 μA for the C1 configuration. Thus, when fin AR changes from 1.25 to 5, the device acquires a 3.42 % increase in ON current. Furthermore, Fig. 8(b) demonstrates the threshold voltage (V_{th}) change against the fin AR. The threshold voltage obtained for C1, C2, and C3 configurations is 0.13 V, 0.17 V, and 0.20 V. This points out that V_{th} rises sequentially with the increase in the fin AR, thus enhancing the subthreshold device characteristics. Quality factor (QF) is an additional vital parameter that primarily defines the device switching behavior and is given as QF = g_m/SS, where SS denotes subthreshold swing [41, 42]. The SS attained for C1, C2, and C3 configurations are 81.63 mV/dec, 73.47 mV/dec, and 69.46 mV/dec. The maximum value of g_m (obtained at V_{gs} = 0.6 V) has been considered for the QF evaluation. Figure 8(b) exhibits the QF for all three simulated devices and shows that QF rises with fin AR. This surge in the QF value with fin AR increase is due to the enhanced transconductance and reduced subthreshold

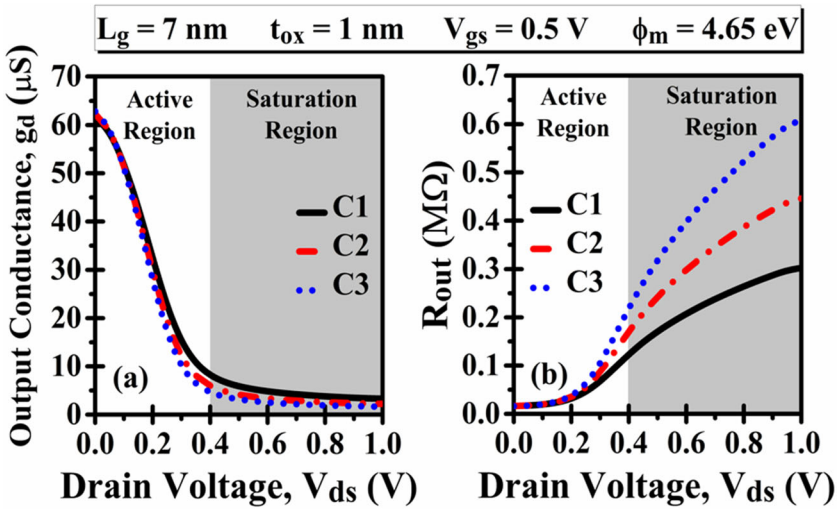


Fig. 9 Variation of (**a**) g_d (**b**) R_{out} against drain-source voltage for each simulated configuration

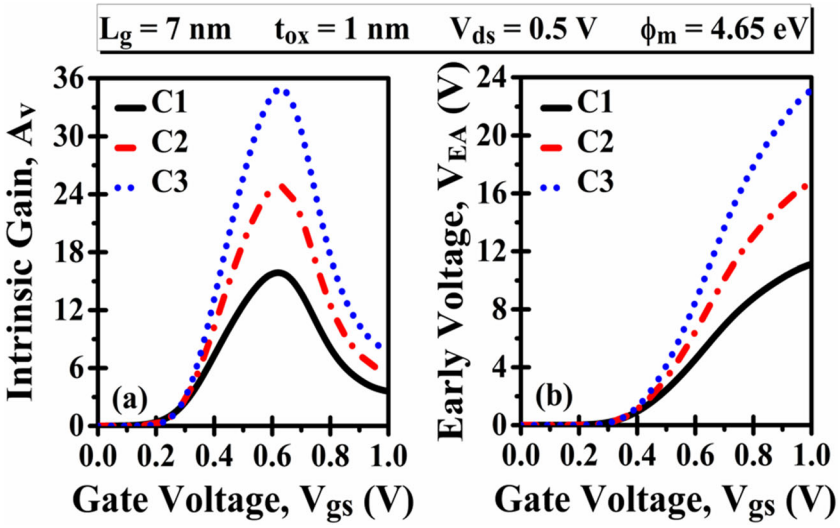


Fig. 10 Alteration of (**a**) A_v (**b**) V_{EA} against fin aspect ratio for all three configurations

swing. Consequently, device switching performance boosts with the increase in the fin AR.

The device's driving ability is estimated by output conductance ($g_d = \partial I_d/\partial V_{ds}$), and its value is determined through the region of device operation [43]. Figure 9(a) represents the output conductance variation against drain-source voltage for each simulated configuration. The g_d decreases with an increase in V_{ds} in the active region before maintaining a constant value in the saturation region. Further, g_d decreases in both active and saturation regions with increased fin AR, demonstrating the suppressed SCEs and improved gate controllability. The converse of output conductance is known as output resistance ($R_{out} = 1/g_d$), and it determines the device's available power gain. The deviation in R_{out} alongside fin AR is displayed in Fig. 9(b). The increment in R_{out} with the increase in fin AR is observed in both regions, with the rise in the saturation region is significantly higher than the active region. Compared to the C1 configuration, the C3 configuration acquires a 77.82 and 101.11 % increase in the active and saturation region.

$$A_V = \frac{g_m}{g_d}; V_{EA} = \frac{I_d}{g_d} \tag{6}$$

Equation (6) evaluates the other crucial analog parameters, i.e., intrinsic gain (A_v) and early voltage (V_{EA}) [43, 44]. To acquire superior analog performance, both these parameters must possess a high value. Figure 10(a) and (b) demonstrate the alteration in A_v and V_{EA} against fin AR for each configuration considered. An extensive enhancement in both the parameters is witnessed with the rise in fin AR. This increase with fin AR is due to the reduced g_d and enhanced g_m and I_d, as shown in Fig. 7. Therefore, superior device performance is attained for C3 configuration, i.e., analog parameters enhance noticeably with the increase in the fin AR, as indicated in Table 2.

5.3 Fin Aspect Ratio Optimization for Linearity Performance

Linearity is necessary to confirm the minimum harmonic distortion and intermodulation at the output [45]. The transistor's non-linearity is evaluated through higher-order derivatives of transconductance, i.e., g_{m2} and g_{m3}.

$$g_{m2} = \frac{\delta^2 I_d}{\delta V_{gs}^2}; g_{m3} = \frac{\delta^3 I_d}{\delta V_{gs}^3} \tag{7}$$

Table 2 Summary of different static (at the center of channel region) and analog parameters of each simulated configurations

Configuration	Parameters									
	Electric Field (V/cm)	Potential (V)	Electron Conc. (/cm^3)	I_{on} (μA)	g_m (S)	QF (μS-dec/mV)	g_d (μS)	Rout (MΩ)	A_v	V_{EA} (V)
C1	0.89×10^5	0.78	6.34×10^{18}	36.79	7.73×10^{-5}	0.95	3.30	0.30	15.91	11.12
C2	1.01×10^5	0.80	1.17×10^{19}	37.51	8.32×10^{-5}	1.13	2.24	0.44	24.97	16.74
C3	1.12×10^5	0.82	1.67×10^{19}	38.05	8.74×10^{-5}	1.26	1.64	0.61	35.27	23.14

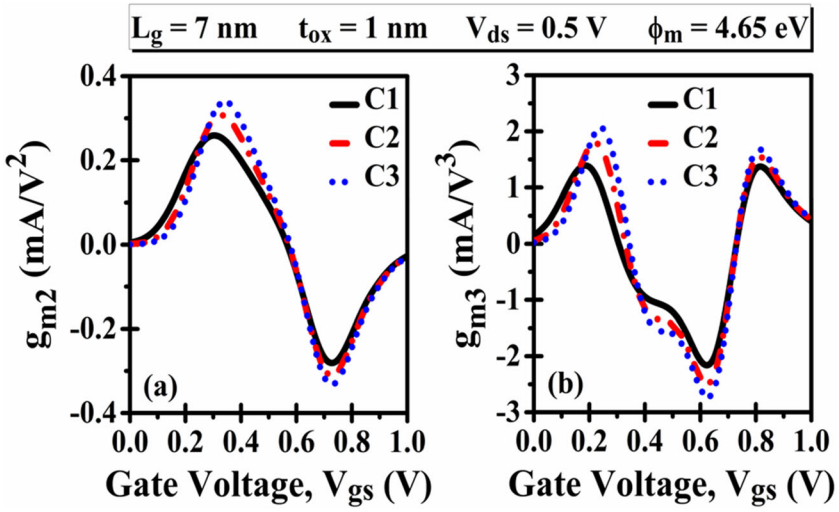


Fig. 11 Variation of (**a**) g_{m2} (**b**) g_{m3} against V_{gs} for each simulated configuration

Equation (7) defines g_{m2} and g_{m3}, and both parameters should be less for improved linearity operation [46]. Figure 11(a) and (b) depict the effect of fin AR on g_{m2} and g_{m3} against V_{gs}. The maximum value of g_{m3} and g_{m2} lessens with the reduction in the fin AR, signifying the enhancement in the linearity performance with fin AR reduction. Equations (8–11) calculates the different linearity parameters such as voltage intercept points (VIP2, VIP3), third-order input intercept point (IIP3), 1-dB compression point [47].

$$\mathrm{VIP2} = \frac{4 \times g_m}{g_{m2}} \tag{8}$$

$$\mathrm{VIP3} = \sqrt{\frac{(24 \times g_m)}{g_{m3}}} \tag{9}$$

$$\mathrm{IIP3} = \frac{2 \times g_m}{3 \times g_{m3} \times R_s} \tag{10}$$

$$1-\mathrm{dB\ compression\ point} = 0.22 \times \sqrt{\frac{g_m}{g_{m3}}} \tag{11}$$

where R_s is the source-channel resistance with a value of 50 Ω.

VIP2 is the extrapolated gate voltage at which second-order harmonic amplitude and fundamental tone amplitude are equal. Similarly, VIP3 is the extrapolated gate voltage at which third-order harmonic and fundamental tone amplitude is equal. The values of VIP2 and VIP3 should be high for better device linearity characteristics. Figure 12(a) and (b) portrays a plot of VIP2 and VIP3 against V_{gs} for each different configuration. The highest value of VIP2 and VIP3 is attained for the C1 configuration, showing improved linearity performance with decreased fin AR. The inverse proportionality

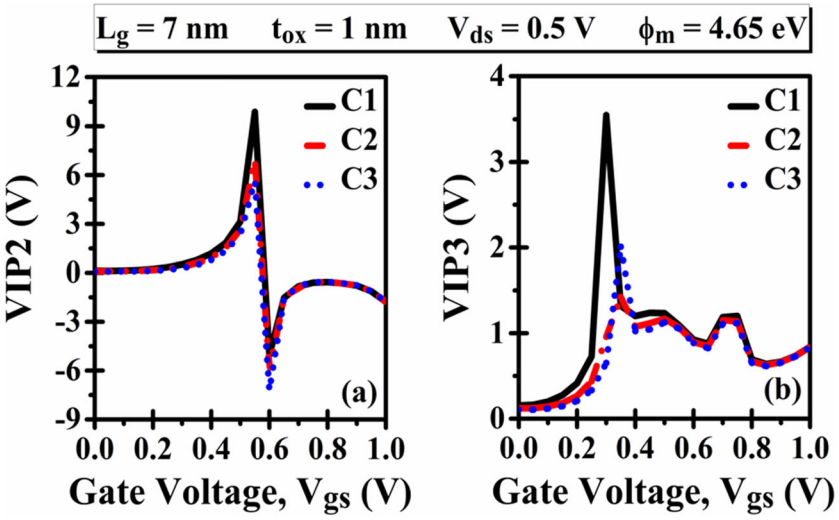


Fig. 12 Alteration in (**a**) VIP2 (**b**) VIP3 against V_{gs} for all three configurations

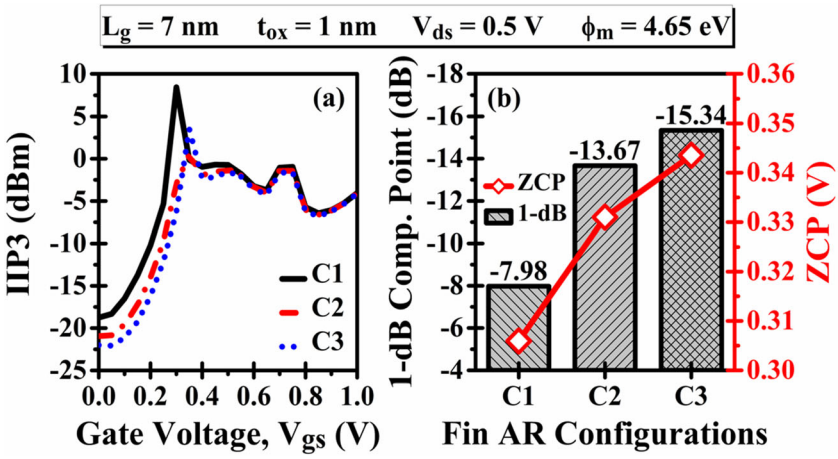


Fig. 13 Plot of (**a**) IIP3 (**b**) 1-dB compression point and ZCP for each configuration

relationship with g_{m2} and g_{m3} is the reason for this increase in VIP2 and VIP3, respectively. Further, the peak value of the C1 configuration is obtained at a lesser value of the V_{gs} that demonstrates superior linearity performance can be obtained with less power for the lower fin AR.

IIP3 is the extrapolated input power at which third and first-order harmonics powers are equivalent. Therefore, IIP3 should be large for better device linearity. Figure 13(a) reflects IIP3 variation against V_{gs} for each configuration considered. The IIP3 maximum value is obtained for the C1 configuration and acquired at a lower gate-source voltage, showing the enhanced linearity characteristics with decreased fin AR. The 1-dB compression point is the input power at which the gain falls by 1 dB from the normal gain. For highly linear applications and to attain maximum gain, 1-dB compression point should have a high value. Figure 13(b) represents the 1-dB compression point variation against V_{gs}. The 1-dB compression point attains the maximum value for C1 configuration, specifying superior linearity performance with decreased fin AR. The zero crossover point (ZCP) is the V_{gs} value for which higher-order derivatives of g_m (g_{m2} and g_{m3}) are zero [48]. The device optimum bias point is determined by ZCP, and it is desirable to be low for better linearity performance. Figure 13(b) represents the ZCP for each configuration, and it shows the distortion is more effectively suppressed in the device with low fin AR. Hence, linearity parameters improve considerably with the fin AR decrease, i.e., superior device performance is attained for the C1 configuration.

5.4 Fin Aspect Ratio Optimization for Distortion Performance

The rise in distortion due to the non-linear device performance in analog/RF applications is a significant concern. Distortion vitiates the strength of the signal by generating unwanted components whose frequency doesn't match the proper band of frequencies [49]. Thus, distortion at the output of a linear amplifier should be as minor as possible. The IFM (Integral Function Method) approach has been used instead of Fourier-based methods to extract distortion parameters as they acknowledge DC measurements in place of AC characterization

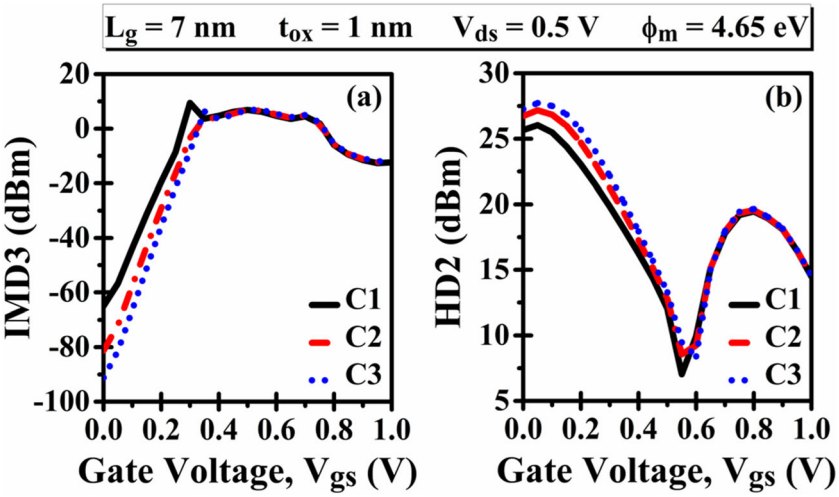


Fig. 14 Variation of (**a**) IMD3 (**b**) HD2 against V_{gs} for each simulated configuration

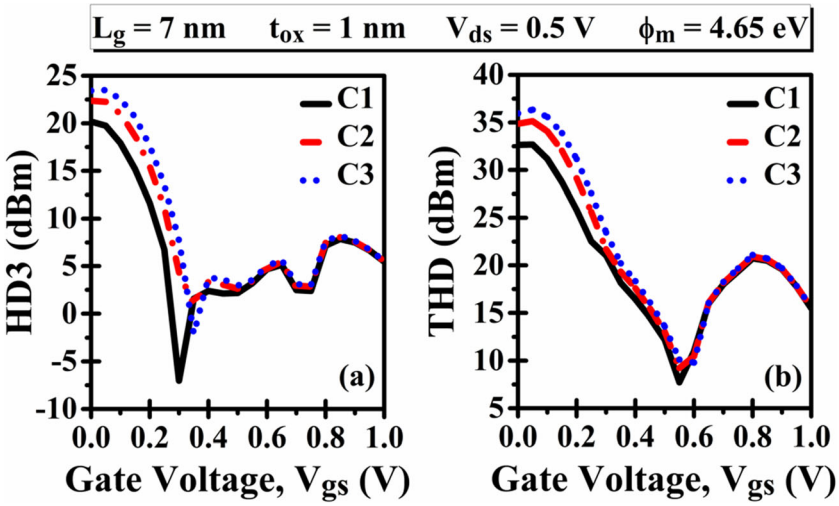


Fig. 15 Alteration of (**a**) HD3 (**b**) THD against V_{gs} for each configuration

[50, 51]. Equations (12–15) provides mathematical relation of several distortion parameters like third-order intermodulation distortion (IMD3), second and third-order harmonics (HD2, HD3), and total harmonic distortion (THD) [50–52].

$$\mathrm{IMD3} = \left[4.5 \times (\mathrm{VIP3})^2 \times g_{m3}\right]^2 \times R_s \tag{12}$$

$$\mathrm{HD2} = 0.5 V_a \frac{\left(\frac{dg_m}{dV_{gs}}\right)}{2g_m} \tag{13}$$

$$\mathrm{HD3} = 0.25 V_a^2 \frac{\left(\frac{d^2 g_m}{dV_{gs}^2}\right)}{6g_m} \tag{14}$$

$$\mathrm{THD} = \sqrt{(\mathrm{HD2})^2 + (\mathrm{HD3})^2 + \ldots} \tag{15}$$

where R_s is the source-channel resistance with a value of 50 Ω, and V_a is the small AC signal amplitude of about 50 mV.

IMD3 is the extrapolated intermodulation distortion power at which third and first-order intermodulation power are equal. For improved linear characteristics, IMD3 should have a low value [45]. Figure 14(a) reveals IMD3 variation against V_{gs} for each configuration considered. In the IMD3-V_{gs} plot, for lower values of V_{gs}, IMD3 increases with the increase in fin AR, whereas an opposite trend is observed at higher values of V_{gs}. This surge in IMD3 is because the increase in the VIP3 dictates the decrease in the g_{m3} at lower V_{gs} for the C1 configuration.

HD2, HD3, and THD are extracted using the approximate analytical expression to estimate the effect of fin AR on the device harmonic distortion characteristics. These harmonics parameters should have a low value to enhance the gain and linearity of the system and reduce the distortion. Figure 14(b) represents the variation of HD2, whereas the HD3 and THD alteration against gate-source voltage for all three simulated configurations is displayed in Fig. 15(a) and (b). It is witnessed from the simulated results that distortion parameters decrease with the decrease in fin AR, showing the distortion in the device decreases with the fin AR reduction. This decrease in the distortion parameters is because the reduction in higher-order transconductance dominates over the g_m reduction. Consequently, distortions parameters

Table 3 Summary of different linearity and intermodulation distortion parameters of each simulated configurations

Configuration	Parameters										
	g_{m2} (mA/V^2)	g_{m3} (mA/V^3)	VIP2 (V)	VIP3 (V)	IIP3 (dBm)	1-dB Compression Point (dB)	ZCP (V)	IMD3 (dBm)	HD2 (dBm)	HD3 (dBm)	THD (dBm)
C1	0.27	1.48	9.88	3.54	8.44	-7.97	0.30	1.90	25.66	20.18	32.64
C2	0.31	1.81	6.95	1.42	0.50	-13.66	0.33	1.99	26.73	22.37	34.86
C3	0.35	2.26	5.63	2.01	3.50	-15.33	0.34	2.25	27.26	23.44	35.95

boost substantially with the reduction in the fin AR, i.e., better device performance is achieved for the C1 configuration, as mentioned in Table 3.

6 Conclusion

This paper describes the numerical study of JAM-GS-GAA FinFET while optimizing the fin AR for upgraded static, analog, and distortion performance. The C3 configuration (fin AR = 5) exhibits the most improved static and analog performance compared to the two other configurations. Compared with the C1 configuration (fin AR = 1.25), an 18.12 and 33.18 % increase in g_m and QF is observed, while R_{out}, A_v, and V_{EA} are enhanced by more than two times in magnitude for the C3 configuration. On the other hand, lesser distortion and enhanced linearity are acquired for the device with low fin AR, i.e., the C1 configuration. The VIP2 and VIP3 increased by 75.40 and 76.61 % for the C1 configuration compared to the C3 configuration. Besides, for the C1 configuration, linearity parameters like 1-dB compression point and IIP3 get enhanced by 47.99 and 140.80 %. The device's optimum bias point determined by ZCP is reduced by 10.94 % for the C1 configuration. Thus, for RF applications, a device with a high fin AR is beneficial, and for designing less distortion and low-power linear applications, a device with a low fin AR can be considered an attractive solution. So, engineers can design the 3-D devices using the paper's findings, but we also must consider the fabrication limitations while creating such configurations.

Acknowledgements For providing all the research facilities, the authors are thankful to Microelectronics Research Laboratory, DTU.

Author Contributions All authors contributed to the study's conception and design.

Data Availability The authors mentioned above have all the relevant data associated with this research work and will be dedicated to sharing that they will be asked to do so in the future.

Declarations

The authors have seen all the Ethical Standards and will suppose to follow them in the future.

Conflict of Interests The authors declare that they have no known conflict of interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent to Participate & for Publication Since the concerned research paper is based on the 'non-life science journal.' So, 'Not Applicable' here. However, the authors have gone through all journal policies and consented to the authorities for further processing.

References

1. Moore GE (1998) Cramming more components onto integrated circuits. Proc IEEE 86:82–85. https://doi.org/10.1109/JPROC.1998.658762
2. Chaudhary A, Kumar MJ (2004) Controlling short-channel effects in deep-submicron SOI MOSFETs for improved reliability: A review. IEEE Trans Device Mater Reliab 4:99–109. https://doi.org/10.1109/TDMR.2004.824359
3. Iwai H (2009) Roadmap for 22 nm and beyond (Invited Paper). Microelectron Eng 86:1520–1528. https://doi.org/10.1016/j.mee.2009.03.129
4. Kumar A, Gupta N, Chaujar R (2016) TCAD RF performance investigation of transparent gate recessed channel MOSFET. Microelectron J 49:36–42. https://doi.org/10.1016/j.mejo.2015.12.007
5. Samal A, Pradhan KP, Mohapatra SK (2021) Improvising the switching ratio through low-k/High-k spacer and dielectric gate stack in 3D FinFET - a simulation perspective. Silicon 13:2655–2660. https://doi.org/10.1007/s12633-020-00618-8
6. Colinge JP, Lee CW, Afzalian A, Akhavan ND, Yan R, Ferain I, Razavi P, O'Neill B, Blake A, White M, Kelleher AM, McCarthy B, Murphy R (2010) Nanowire transistors without junctions. Nat Nanotechnol 5:225–229. https://doi.org/10.1038/nnano.2010.15
7. Biswas K, Sarkar A, Sarkar CK (2018) Fin shape influence on analog and RF performance of junctionless accumulation-mode bulk FinFETs. Microsyst Technol 24:2317–2324. https://doi.org/10.1007/s00542-018-3729-1
8. Kim TK, Kim DH, Yoon YG, Moon JM, Hwang BW, Moon DI, Lee GS, Lee DW, Yoo DE, Hwang HC, Kim JS, Choi YK, Cho BJ, Lee SH (2013) First demonstration of junctionless accumulation-mode bulk FinFETs with robust junction isolation. IEEE Electron Device Lett 34:1479–1481. https://doi.org/10.1109/LED.2013.2283291
9. Lee CW, Ferain I, Afzalian A, Yan R, Akhavan ND, Razavi P, Colinge JP (2010) Performance estimation of junctionless multigate transistors. Solid State Electron 54:97–103. https://doi.org/10.1016/j.sse.2009.12.003
10. Lo SH, Buchanan DA, Taur Y, Wang W (1997)Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFET's. IEEE Electron Device Lett 18:209–211. https://doi.org/10.1109/55.568766
11. Ribes G, Mitard J, Denais M, Bruyere S, Monsieur F, Parthasarathy C, Vincent E, Ghibaudo G (2005) Review on high-k dielectrics reliability issues. IEEE Trans Device Mater Reliab 5:5–19. https://doi.org/10.1109/TDMR.2005.845236
12. Kerber A, Cartier E, Pantisano L, Degraeve R, Kauerauf T, Kim Y, Hou A, Groeseneken G, Maes HE, Schwalke U (2003) Origin of the threshold voltage instability in SiO_2/HfO_2 dual layer gate dielectrics. IEEE Electron Device Lett 24:87–89. https://doi.org/10.1109/LED.2003.808844
13. Onishi K, Kang CS, Choi R, Cho HJ, Gopalan S, Nieh RE, Krishnan SA, Lee JC (2003) Improvement of surface carrier mobility of HfO_2 MOSFETs by high-temperature forming gas annealing. IEEE Trans Electron Devices 50:384–390. https://doi.org/10.1109/TED.2002.807447
14. Gupta N, Chaujar R (2016) Optimization of high-k and gate metal workfunction for improved analog and intermodulation performance of Gate Stack (GS)-GEWE-SiNW MOSFET. Superlattices Microstruct 97:630–641. https://doi.org/10.1016/j.spmi.2016.07.021
15. Huang YC, Chiang MH, Wang SJ, Fossum JG (2017) GAAFET versus pragmatic FinFET at the 5nm Si-based CMOS Technology node. IEEE J Electron Devices Soc 5:164–169. https://doi.org/10.1109/JEDS.2017.2689738

16. Kumar B, Kumar A, Chaujar R (2020) The effect of gate stack and high-κ spacer on device performance of a junctionless GAA FinFET. IEEE VLSI Device, Circuit Systems Conference, 159–163. https://doi.org/10.1109/VLSIDCS47293.2020.9179855
17. Mohapatra SK, Pradhan KP, Singh D, Sahu PK (2015) The role of geometry parameters and Fin aspect ratio of sub-20nm SOI-FinFET: An analysis towards analog and RF circuit design. IEEE Trans Nanotechnol 14:546–554. https://doi.org/10.1109/TNANO.2015.2415555
18. Nagy D, Indalecio G, Garcia-Loureiro AJ, Elmessary MA, Kalna K, Seoane N (2018) FinFET versus gate-all-around nanowire FET: Performance, scaling, and variability. IEEE J Electron Devices Soc 6:332–340. https://doi.org/10.1109/JEDS.2018.2804383
19. Boukortt N, Hadri B, Patanè S, Caddemi A, Crupi G (2017) Investigation on TG n-FinFET parameters by varying channel doping concentration and gate length. Silicon 9:885–893. https://doi.org/10.1007/s12633-016-9528-3
20. Coquand R, Jaud MA, Rozeau O, ElOudrhiri AI, Martinie S, Triozon F, Pons N, Barraud S, Monfray S, Boeuf F, Ghibaudo G, Faynot O (2013) Comparative simulation of TriGate and FinFET on SOI: Evaluating a multiple threshold voltage strategy on triple gate devices. IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference 13–14. https://doi.org/10.1109/S3S.2013.6716523
21. Kumar B, Chaujar R (2021) Fin aspect ratio optimization of novel junctionless Gate Stack Gate All Around (GS-GAA) FinFET for Analog/RF applications. Microelectronics, circuits and systems. Lect Notes Electr Eng 755:59–67. https://doi.org/10.1007/978-981-16-1570-2_6
22. Biswas K, Sarkar CK (2018) Optimizing Fin aspect ratio of junctionless bulk FinFET for application in Analog/RF circuit. IEEE Electron Devices Kolkata Conf, 591–595. https://doi.org/10.1109/edkcon.2018.8770515
23. Gupta N, Kumar A (2020) Assessment of high-k gate stack on sub-10 nm SOI-FinFET for high-performance analog and RF applications perspective. ECS J Solid State Sci Technol 9:123009. https://doi.org/10.1149/2162-8777/abcf14
24. Liu Y, Kijima S, Sugimata E, Masahara M, Endo K, Matsukawa T, Ishii K, Sakamoto K, Sekigawa T, Yamauchi H, Takanashi Y, Suzuki E (2006) Investigation of the TiN gate electrode with tunable work function and its application for FinFET fabrication. IEEE Trans Nanotechnol 5:723–728. https://doi.org/10.1109/TNANO.2006.885035
25. Sjöblom G (2006) Metal gate technology for advanced CMOS devices, Ph.D. dissertation Dept. Engg. Sci., Uppsala Univ., Sweden
26. Vitale SA, Kedzierski J, Healey P, Wyatt PW, Keast CL (2011)Work-function-tuned TiN metal gate FDSOI transistors for subthreshold operation. IEEE Trans Electron Devices 58:419–426. https://doi.org/10.1109/TED.2010.2092779
27. ATLAS User's Manual (2016) SILVACO International, Santa Clara
28. Shockley W, Read WT (1952) Statistics of the recombinations of holes and electrons. Phys Rev 87:835–842. https://doi.org/10.1103/PhysRev.87.835
29. Hall RN (1952)Electron-hole recombination in Germanium. Phys Rev 87:387. https://doi.org/10.1103/PhysRev.87.387
30. Arora ND, Hauser JR, Roulston DJ (1982) Electron and hole mobilities in silicon as a function of concentration and temperature. IEEE Trans Electron Devices 29:292–295. https://doi.org/10.1109/T-ED.1982.20698
31. Hurkx GAM, Klaassen DBM, Knuvers MPG (1992) A New Recombination Model for Device Simulation Including Tunneling. IEEE Trans Electron Devices 39:331–338. https://doi.org/10.1109/16.121690
32. Crowell CR, Sze SM (1966) Temperature dependence of avalanche multiplication in semiconductors. Appl Phys Lett 9:242–244. https://doi.org/10.1063/1.1754731
33. Dirac PAM (1926) On the theory of quantum mechanics. Proc R Soc London Ser A. Contain Pap a Math Phys Character 112:661–677. https://doi.org/10.1098/rspa.1926.0133
34. Lee H, Yu LE, Ryu SW, Han JW, Jeon K, Jang DY, Kim KH, Lee J, Kim JH, Jeon SC, Lee GS, Oh JS, Park YC, Bae WH, Lee HM, Yang JM, Yoo JJ, Kim SI, Choi YK (2006)Sub-5nm all-around gate FinFET for ultimate scaling. Dig Tech Pap - Symp VLSI Technol 58–59. https://doi.org/10.1109/vlsit.2006.1705215
35. Auth C, Aliyarukunju A, Asoro M, Bergstrom D et al (2017) A 10nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, Self-Aligned Quad Patterning, contact over active gate and cobalt local interconnects. IEEE International Electron Devices Meeting, 673–676. https://doi.org/10.1109/IEDM.2017.8268472
36. Choi SN, Moon SE, Yoon SM (2021) Impact of oxide gate electrode for ferroelectric field-effect transistors with metal-ferroelectric-metal-insulator-semiconductor gate stack using undoped HfO_2 thin films prepared by atomic layer deposition. Nanotechnology 32:085709. https://doi.org/10.1088/1361-6528/abc98c
37. Kumar A, Gupta N, Tripathi SK, Tripathi MM, Chaujar R (2020) Performance evaluation of linearity and intermodulation distortion of nanoscale GaN-SOI FinFET for RFIC design. AEU - Int J Electron Commun 115:153052. https://doi.org/10.1016/j.aeue.2019.153052
38. Long W, Ou H, Kuo JM, Chin KK (1999)Dual-Material Gate (DMG) field effect transistor. IEEE Trans Electron Devices 46:865–870. https://doi.org/10.1109/16.760391
39. Kumar A, Tripathi MM, Chaujar R (2018) Comprehensive analysis of sub-20 nm black phosphorus based junctionless-recessed channel MOSFET for analog/RF applications. Superlattices Microstruct 116:171–180. https://doi.org/10.1016/j.spmi.2018.02.018
40. Kumar B, Chaujar R (2021) TCAD temperature analysis of gate stack gate all around (GS-GAA) FinFET for improved RF and wireless performance. Silicon 1–13. https://doi.org/10.1007/s12633-021-01040-4
41. Doornbos G, Passlack M (2010) Benchmarking of III-V n-MOSFET maturity and feasibility for future CMOS. IEEE Electron Device Lett 31:1110–1112. https://doi.org/10.1109/LED.2010.2063012
42. Gupta N, Jain A, Kumar A (2021) 20 nm GAA-GaN/Al_2O_3 nanowire MOSFET for improved analog/linearity performance metrics and suppressed distortion. Appl Phys A Mater Sci Process 127:1–9. https://doi.org/10.1007/s00339-021-04673-9
43. Pradhan KP, Mohapatra SK, Sahu PK, Behera DK (2014) Impact of high-k gate dielectric on analog and RF performance of nanoscale DG-MOSFET. Microelectronics J 45:144–151. https://doi.org/10.1016/j.mejo.2013.11.016
44. Kumar B, Chaujar R (2021) Analog and RF performance evaluation of Junctionless Accumulation Mode (JAM) Gate Stack Gate All Around (GS-GAA) FinFET. Silicon 13:919–927. https://doi.org/10.1007/s12633-020-00910-7
45. Kumar SP, Agrawal A, Chaujar R, Gupta RS, Gupta M (2011) Device linearity and intermodulation distortion comparison of dual material gate and conventional AlGaN/GaN high electron mobility transistor. Microelectron Reliab 51:587–596. https://doi.org/10.1016/j.microrel.2010.09.033
46. Kumar A, Tripathi MM, Chaujar R (2018) Reliability issues of In_2O_5Sn gate electrode recessed channel MOSFET: Impact of interface trap charges and temperature. IEEE Trans Electron Devices 65:860–866. https://doi.org/10.1109/TED.2018.2793853
47. Sreenivasulu VB, Narendar V (2021) A comprehensive analysis of Junctionless Tri-Gate(TG) FinFET towards low-power and high-

frequency applications at 5-nm gate length. Silicon. https://doi.org/10.1007/s12633-021-00987-8

48. Gupta N, Kumar A, Chaujar R (2015) Effect of dielectric engineering on analog and linearity performance of gate electrode workfunction engineered (GEWE) silicon nanowire MOSFET. IEEE-NANO2015–15th Int Conf Nanotechnol, 928–931. https://doi.org/10.1109/NANO.2015.7388768
49. Dutta A, Koley K, Sarkar CK (2014) Analysis of Harmonic distortion in asymmetric underlap DG-MOSFET with high-k spacer. Microelectron Reliab 54:1125–1132. https://doi.org/10.1016/j.microrel.2013.12.001
50. Cerdeira A, Alemán MA, Estrada M, Flandre D (2004) Integral function method for determination of nonlinear harmonic distortion. Solid State Electron 48:2225–2234. https://doi.org/10.1016/j.sse.2004.06.001
51. Doria RT, Cerdeira A, Raskin JP, Flandre D, Pavanello MA (2008) Harmonic distortion analysis of double gate graded-channel MOSFETs operating in saturation. Microelectronics J 39:1663–1670. https://doi.org/10.1016/j.mejo.2008.02.006
52. Groenewold G, Lubbers WJ (1994) Systematic distortion analysis for MOSFET integrators with use of a New MOSFET model. IEEE Trans Circuits Syst II Analog Digit Signal Process 41:569–580. https://doi.org/10.1109/82.326583

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

2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)

Personalized and Dynamic top-k Recommendation System using Context Aware Deep Reinforcement Learning

Anubha Kabra §
Adobe Systems
anubhakabradtu@gmail.com

Anu Agarwal §
Samsung R&D
anuag7699@gmail.com

2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC) | 978-1-6654-2463-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/COMPSAC51774.2021.00042

***Abstract*—With the ever-increasing number of choices in just about every domain, ranging from e-commerce platforms to music and video streaming apps, it requires a highly personalized and best-in-class experience to make the users stay. Relevant, dynamic and smart item recommendations play a significant role in providing quality user experience. The majority of existing work either relies on explicit user feedback or gives static recommendations that poorly adapt to the changing user preferences. This paper proposes novel model architectures, namely, Sequential Feature Bandits(SFB) and Cross Feature Bandits(CFB) for providing online top-k context-aware recommendations. It leverages the item features, and uses implicit user feedback with a reinforcement learning backed dynamic algorithm. We use deep-q learning to maximize the cumulative estimated reward(CMR) and click-through rate(CTR) along with an experience replay strategy to prevent the proposed models from getting destabilized. Detailed experiments were conducted on a commercially available news article data set with implicit user-feedback and user-item context features. A thorough experimentation shows an average increase of 23.4% in CTR values and proves the effectiveness of the proposed SFB and CFB architectures for real-time recommendation.**

***Index Terms*—Context Aware Recommender Systems, Ubiquitous Computing , Reinforcement Learning, Intelligent and self-organizing computing, Deep Learning**

I. Introduction

A good recommender system is a new "must-have" for any product or service-based online venture today. It helps the users to systematically deal with a substantial number of choices available without going through the usual 'hit-and-trial'. Owing to the improved user experience and high potential profits for several E-commerce and Social Media platforms, significant effort has been made to capture the customer interest. However, this is not possible if adequate attention is not paid to the orientation of the market trends and customer behavior. This necessitates the recommendations to be real-time and personalized.

There are four main challenges that abstain the industry from providing customers pertinent item recommendations:

- Static Systems: Most recommendation systems are passive [11, 15, 28] in learning because they have to be retrained to handle a dynamic environment. This prevents the incorporation of changing trends and stochastic nature of the customer. For example, A customer generally likes buying fast food online. However, with time, she decides to eat healthy and starts ordering nutritious food. A static or user feature-based recommendation system will keep suggesting her fast food items long after she has changed her choices, unless retraining is done. This may lead to lower customer retention rates over time.
- Neglecting user, context and item features: Commercial collection of user, item and session context features (refers to session-specific values of parameters such as timestamp, location, session size, etc. which facilitate the recommender system to differentiate between transactions made in a different context) is common. Most state-of-the-art strategies either neglect all feature importance [20, 29], or take into account only user-context features or item features,but not all. This leaves a vast scope for policies to utilize all available features for a more customized recommendation.
- Cold Start Problem: This problem occurs when the amount of data available for recommendation is not sufficient and leads to the model giving close to random results. This happens generally, when a new user enters the system. The time taken by the system to habituate to the user, it may act as a disadvantage for competitive markets who rely on captivating users through item recommendations.
- Disregarding the recent trends: While many recommendation systems are dynamic in nature, they may neglect the recent trends.Generally experience replay [27] collects information over big batches and randomly chooses a mini batch to update the model. This compromises with the sequential nature of the data,

§Equal contribution — Work done while in Delhi Technological University, Delhi, India

978-1-6654-2463-9/21/$31.00 ©2021 IEEE
DOI 10.1109/COMPSAC51774.2021.00042

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

which evidently exists in the majority of the real-time databases.

Sequential Feature Bandits(SFB) and Cross Feature Bandits(CFB) have **tackled all the above mentioned challenges** by adopting the following measures, hence affirming to a novel style for addressing recommendations :

- Dynamic Learning: SFB and CFB adopt active astute learning, hence updating the model on each trial and over short successive periods to learn the continuously changing context.
- Leveraging all features: Proposed methods ensure that both user and item features, along with session context features for each trial are leveraged in the process of recommendation to personalize the user experience.
- Tackling the cold start problem: SFB and CFB use feature-based learning rather than relying on user-specific ids. Hence, this dynamic similarity between the user features can be used to match the new user entering the environment.
- Continuous Feedback: The Experience Replay in the proposed policies is done for a relatively smaller batch size. This guarantees that the sequential pattern in the data is kept intact and the current trends are well exploited by policies for a superior recommendation.

The rest of the paper is structured as follows: Section II defines the related work, Section III gives the background of the problem, Section IV focuses on the problem definition, and Section V describes our SFB and CFB techniques along with their architecture. Section VI gives an overview of the experiments conducted and evaluation of the results. Section VII concludes this paper and discusses the scope of future work in this domain.

II. Related Work

Recommendation Systems have attracted a lot of attention among researchers since the first attempt at recommenders [24] in 1990s. Traditional recommendation systems rely on either popularity-based or collaborative techniques. On the one hand where popularity based recommenders give no value whatsoever to user preferences, Collaborative Filtering (CF) utilizes the user-item interactions and rating values to model user relations and preferences. CF based techniques can be further subdivided into item-based, content-based [15, 19] and hybrid CF [11, 21, 25, 28, 29] where the recommendations are shown based on either existing users having similar choices or items similar to the items already rated by the user. Matrix Factorization [20] algorithm is yet another extension of CF. It expresses the user-item interaction matrix as a product of two lower dimensional rectangular matrices to extract and utilize the latent features present in the user-item interactions.

In Spite of performing remarkably well on several benchmark datasets such as MovieLens [13] having only user-id, item-id, and rating information available, none of the aforementioned techniques effectively differentiates between user actions performed in different settings or context. In many application domains, a context-independent representation may lose predictive power because potentially useful information from multiple contexts is aggregated [1]. Impact of including context on the performance of recommenders has been thoroughly investigated in [1, 16] and [30]. Attempts have been made to incorporate contextual information into the existing matrix factorization algorithm [5]. This is performed by modelling the interaction of the contextual factors with item ratings introducing additional model parameters. Since user specific and context-based information keeps changing over time, it is important for the recommender systems to be able to adapt to these changes and recommend items accordingly. A Reinforcement Learning based technique called Multi-Armed Bandit (MAB) [4, 6, 18] has recently drawn significant attention for its ability to generate dynamic recommendations. To better handle the exploration-exploitation dilemma involved in MAB, several extensions such as ϵ-greedy, softmax, Upper Confidence Bound (UCB) [3] and Thompson Sampling [2] have been proposed. A contextual bandit approach was proposed in [23] in which a learning algorithm sequentially selects items to serve users based on contextual information about the users and items. It adapts its item selection strategy simultaneously based on user-click feedback to maximize total user clicks. Linear pay-offs with contextual bandits were explored in [9] with an algorithm called linUCB. Attempts have been made to combine contextual bandits with traditional techniques such as matrix Factorization [31], Collaborative Filtering[12] and novel multi arm bandit policies [10]. Q-Learning [32] is another RL algorithm widely used for recommenders. It differs from the MAB approach in its action-value function. It tries to maximize the cumulative discounted future reward for a given state-action pair rather than just the immediate reward. Deep-Q Network (DQN)[26, 27] is another algorithm which uses an experience replay strategy to remove any collinearities present in the data obtained from the environment and a target function for stabilized learning. Our approach combines the best of deep-Q learning and contextual bandits along with a novel framework which performs considerably better than the existing approaches.

III. Background

A. Contextual Multi Arm Bandits (cMAB)

In reinforcement learning, the agent learns a model that can maximize long-term cumulative reward by interacting with the environment [23].The contextual bandits refer to the multi arm bandits problem in reinforcement learning, except it uses the context information for choosing the arm for a particular state. cMAB depends on the sequence of incoming states and for each discrete sequence information and gives a positive or negative reward according to the

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

decision it takes. For real time situations in recommendation systems, where the user may like more than one action in action space, multi arm bandits can be used to simulate partial feedback for only the actions shown to the users. cMAB ensures that the user state space and features are also used in this decision making process.
Formally, we define a fixed set of actions $A = \{a_1, a_2, ...a_N\}$,contextual state space S, a contextual vector, $s_t \in S$ and a reward vector r_t whose length is equal to the length of action space and each $r_{t_a} \in [0, 1]$. We also define a model $\pi = S \longmapsto A$. i.e. $a = \pi(s)$.
A contextual multi armed bandit describes step-by-step interaction of the agent with the environment. At each timestep t, the problem can be viewed as a series of following substeps:

1) The environment generates a context-reward tuple (s_t, r_t) from an unknown underlying distribution D. However, the reward vector is not revealed to the agent just yet.
2) The agent uses the model π to generate an action for the given contextual vector s_t and obtains a reward from the environment.
3) Based on the reward obtained in the previous step being positive or negative, the model π is updated.

The model parameters are updated in such a manner that it maximizes the cumulative reward. It is important to note here that only the reward for the chosen action is revealed. Ideally, the model should not be updated if we choose an action which is different from the true label since we do not have any reward signal available for that action.

B. Important Terms

The important terms are briefly described below for the ready reference of the readers.

- **State:** State refers to combination of user and context features. This consists of all the user's information, except for the explicit user_id (A unique id to refer to a particular user). The id is discarded to maintain user privacy.
- **Action:** The action refers to the variety of arms available for recommendation to the user. The top ranked k arms , according to their score values during prediction will be recommended for a particular state.
- **Transitions:** Transitions are used to define the movement from one state to another, with partial feedback for a particular state.
- **Experience Replay:** Experience replay refers to updating the model with the accumulated state, action, and reward tuples(s,a,r) in batches.

C. The exploration and exploitation trade off

The exploration-exploitation trade-off is a well-known problem that occurs in scenarios where a learning system has to repeatedly make a choice with uncertain pay-offs. In essence, the dilemma for a decision-making system that only has incomplete knowledge of the world is whether to repeat decisions that have worked well so far (exploit) or to make novel decisions, hoping to gain even greater rewards (explore). With our end goal as recommendations, the model has to strategically choose between exploration and exploitation. We use the decaying ϵ-Greedy exploration in our methods. The ϵ-Greedy strategy is alleviated with the use of Gaussian Distribution for selecting a value to decide for exploration or exploitation. The benefit of this distribution is that it relies on a mean parameter, which can be used to bias the value selection process. Hence, this parameter was set to slightly increase the selection that supports exploitation. The mean value was kept at 0.7, while the rest of the parameters were set as default. This experimentally worked better than uniform distribution.
The formula of Gaussian distribution (with parameters: mean μ and standard deviation σ) is as follows:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{1}{2}(\frac{x-\mu}{\sigma})^2 \quad (1)$$

Each time an exploration is performed, the ϵ value reduces the decay time (decay being a fixed value).

$$\epsilon = (\epsilon - \epsilon * decay) \quad (2)$$

IV. Problem Definition

We define our problem as follows:
Given a fixed set of N arms (items) , whenever a user with feature set u requests agent G for recommendations, the environment generates a set of context features c. The proposed work generates k most appropriate arms as recommendations for this request. The notations used throughout the paper are summarized in Table I.

TABLE I: Notations

Notation	Meaning
G	Agent
π	Model
R	Reward
S	State Space
s_i	State i
U	User set
u	User features
a	Arm features
c	Context features
N	No. of candidate items
ϵ	Epsilon
k	Number of recommended items
decay	Exploration Decay

Our work leverages :

1) **User features**: The common user features which define user behaviour are taken into consideration. These features include be 'age', 'region', 'profession', etc. Whenever a new user enters the system, this user's features are compared with existing users' feature embedding to create a contextual similarity.
2) **Arm(item) features**: The item features refer to various properties describing each item. These may include "category", 'word count" etc. It is a vector

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

space, used to learn similarity between different items. However, since the item actually selected by the user is not known beforehand, this feature vector cannot be treated as direct input to the architecture.

3) **Context features:** These features are specific to a particular user session. E.g. 'session_length', 'session_start_time', 'location' etc. Inclusion of these features helps the recommender in differentiating between user-item interactions made in different sessions.

4) **Sequential information from time stamp**: Reliance on timestamp for recommendation benefits from the latent sequential pattern in the data. Using timestamp with partial feedback ensures that changing user behaviour patterns are used effectively in the process of recommendation. To simulate a real-life scenario, we use a dataset where the old news articles go out of trend after a particular period of time (Figure 3). Capturing sequential patterns becomes a necessity in such cases.

V. Methodology

In this section, we elaborate on the proposed Feature Bandits's architecture and its variants. We further describe the model framework (Figure 2) in a step by step fashion.

A. Feature Bandits Architecture

The two variants of the feature bandit architecture are:

1) Sequential Feature Bandit : In the first architecture (Figure 1(a)), the item features are directly concatenated with the user and context features. This concatenated vector is then input into the model, i.e. $(u+c+a_i)$, where $i \in N$. The model decides the probability of the i_{t_h} item being selected by the user.

2) Cross Feature Bandit : Initially, the model tackles context + user features and item features separately (Figure 1(b)). The subsequent layer reduces them to a common length. These reduced vectors are then concatenated i.e. $((u+c)_h, a_{i_h})$, where $i \in N$ and h represent the hidden layer output. The rationale behind this alteration being that in several real-time datasets, there might be a skew in features. Either the user and context features or the structure features may dominate. To equalize their impact, both are reduced to a common length before proceeding further.

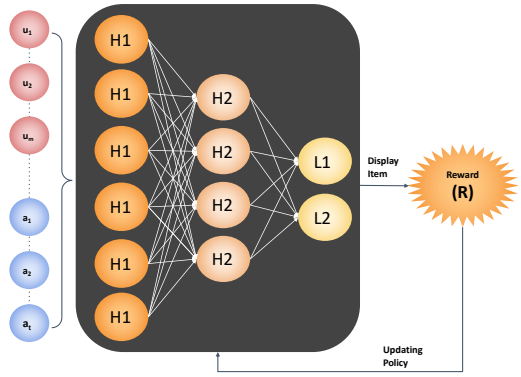
B. Model Framework

The environment generated set of covariates are passed as inputs into the *Feature bandits* model and a list of k recommended items is generated at each time step. Detailed pipeline of the framework has been given in Figure 2. The major steps involved in the proposed *Feature Bandits* framework are as follows:

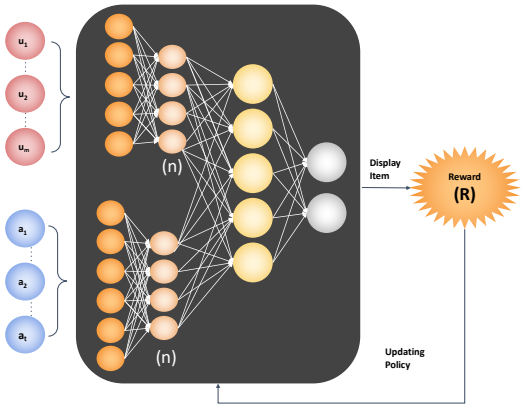
1. Explore vs Exploit : Since decaying ϵ-greedy exploration strategy is being used, the first step involves deciding between exploration and exploitation based on the current value of ϵ. In case of exploration, k random unexplored arms are added to the recommendation list. Else, for exploitation, the user-context features are merged one-by-one with features corresponding to each arm to form N state-action pairs. Now, the agent predicts the immediate rewards for each one of these state-action pairs and sends only k arms with the highest predicted rewards further to be added to the recommendation list.

2. Feedback : In this step, the recommendation list generated in the previous step consisting of k-items is pushed to the user to obtain feedback. As mentioned earlier, in the contextual bandit setting, only the reward corresponding to the action chosen by the user is revealed. Hence, if any item from the recommendation list is same as the chosen item, model is updated with reward 1. For rest of the selected items, the model is updated with reward 0.

3. Model Update : As soon as the user feedback for an item is obtained, the model π is updated with the state-action-reward tuple (s_i, a_i, r_i). Simultaneously, these



(a) Sequential Feature Bandits



(b) Cross Feature Bandits

Fig. 1: Feature Bandits network architecture
H : Hidden Layer, L: Labels , n: equal number of nodes for CFB user and item features before feeding them together.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

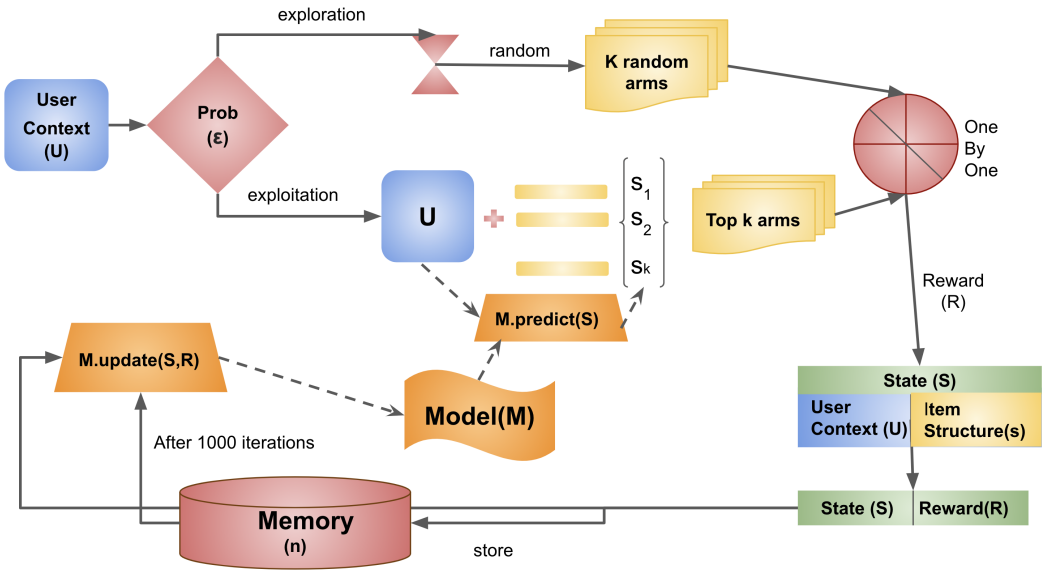


Fig. 2: Our *Feature Bandits* framework

TABLE II: Abbreviations for all the policies mentioned.

Abbreviation	Policy
LR	Logistic Regression
W&D	Wide and Deep
CB	Contextual Bandits
BUCB	Bootstrapped Upper Confidence Bound
BTS	Bootstrapped Thompson Sampling
AGT	Adaptive Greedy Threshold
AGP	Adaptive Greedy Percentile
AE	Active Explorer
AAG	Active Adaptive Greedy
SFB	Sequential Feature Bandits
CFB	Cross Feature Bandits

tuples are added to the memory to be used further in the *experience-replay* module and exploration parameters are also updated.

4. Experience Replay : Once enough experiences are accumulated in the memory i.e. the number of state-action-reward tuples stored in the memory becomes equal to a *threshold batch size* (default 1000, refer Figure 2), these experiences are replayed (similar to the replay strategy mentioned in [27]), for a fixed number of iterations. This step is performed to reinforce the more recent trends learned by the model without completely wiping out the knowledge it has gained previously. After each experience replay step, the memory buffer is reset to make way for the new experiences to be inculcated.

VI. Experimentation

A. Dataset

Since the problem necessitates the involvement of user, context and item features, we resorted to a dataset that fulfills these requirements. This dataset consists of user interaction logs from a news portal which was provided

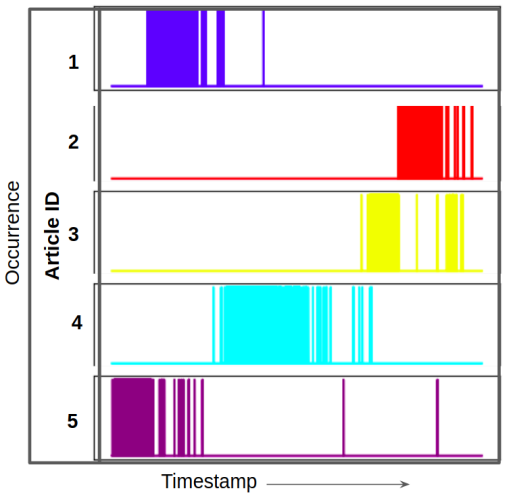


Fig. 3: Occurrence of different Articles with timestamp

by Globo.com [§], a popular news portal in Brazil. Here the users refer to the news reading population, whereas the items are news articles.The dataset has features corresponding to both users and items. With the fast-growing dependence on electronic media, click-log datasets are able to capture the changing user behavior and non-stationarity of the market with frequently changing news articles.To avoid class imbalance due to several hardly repetitive articles as well as the vastness of the data [17], we chose to select the most frequently appearing 50 articles. The dataset initially had a combination of categorical and

[§]https://www.kaggle.com/gspmoreira/news-portal-user-interactions-by-globocom

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

TABLE III: CTR values for the given baselines and proposed methods for k=1, k=3 and k=5. *Here, RL = Reinforcement Learning and k = No. of items shown.* Proposed methods are mentioned in **bold**. Best CTR values are underlined.

Type	**Policy**	**CTR**		
		k = 1	k = 3	k = 5
Non RL Based	LR	0.0022	0.0042	0.0043
	W&D	0.0032	0.0010	0.0005
RL Based	CB + BUCB	0.0243	0.0244	0.0245
	CB + BTS	0.0243	0.0244	0.0245
	CB + AGT	0.0252	0.0230	0.0256
	CB + AGP	0.0244	0.0243	0.0245
	CB + AE	0.2280	0.1023	0.0733
	CB + AAG	0.1864	0.0859	0.0627
	SFB	**0.2311**	**0.1386**	**0.1058**
	CFB	**0.2286**	**0.1201**	**0.0996**

continuous columns. The categorical features were one-hot encoded. All the continuous features were normalized. The final dataset has 1,82,494 trials, from which CTR and Relative CTR are evaluated for 30% test data to compare with the static recommendation baselines. CMR is calculated for all dynamic baselines. Over all the dataset has **38 user context features** such as "Country","Operating System","Click Environment" etc. and **62 item features** such as "Article Id","Word Count" etc.

B. Baselines

With multiple strategies for a recommendation, we decided to compare it with two types of techniques - static and dynamic. Both types consider user features, however, they lack the capacity to use the combination of user and item features. [10].

1) Static Techniques (Non RL Based):

- **LR** [22] : Logistic Regression is widely used in the industry as a common baseline method due to its easy implementation and high efficiency. It takes all contextual features of users as input.
- **W&D** : Wide & Deep [7, 8] is a widely used state-of-art deep learning model combining memorization (through logistic regression on wide combinations of categorical features) and generalization (through a deep neural network embedding of the raw features) to predict the click label.

2) Dynamic Techniques (RL Based):

- **Bootstrapped UCB**[10] : UCB exploration strategy involves calculating the confidence bound for each item. This bound depicts the confidence about the actual action-value of an item. Greediest action is selected to maximize UCB.
- **Bootstrapped TS**[10] : This strategy involves forming a prior distribution for each arm. A parameter theta is drawn from this distribution for each selected arm. The arm with the highest theta value is selected, and the distribution for each arm is updated.
- **Adaptive Greedy**[10] : This technique sets a certain threshold value for the probability scores generated by the base classifier. Now, based on the score obtained for a given state, this threshold value is used to decide between exploration and exploitation. However, the threshold value changes when the number of trials exceeds a chosen window size so as to adapt the model to give better predictions. Here, three variations of this technique are tried, namely Adaptive Greedy Threshold, Adaptive Greedy Percentile and Adaptive Active Greedy.
- **Active Explorer**[10] : This strategy uses active learning and tries to choose an item depending upon the amount of leverage it gets on the model if its label were known.

C. Evaluation

1) Simulated Online Environment: The evaluation for all the methodologies was done in an online simulation. The following conditions were adhered to :

- The data was sequential, timestamp sorted and not shuffled. This maintained the time-dependent pattern of the data to simulate how users would enter the system in a real-time scenario.
- Unlike separate training and testing in traditional methods, the proposed work employs continuous training and feedback.

The simulated online evaluation is performed as follows: For any given user request, a recommender system generates a list of k items. If any of these k items matches the actual label clicked by the user for that particular request, it is considered a success. Else, it is considered as a failure.

For example : If a user enters a news website, k news items will be recommended to her. She can decide to choose one or none. If any of the items are clicked, it counts for a success, else failure.

.

2) Evaluation Metrics: The following metrics were used for the evaluation of all techniques compared in this paper:

- **CTR** : Click Through Rate is an effective measure for assessing the effectiveness of recommender systems in understanding user click behaviour. It is calculated for each item as follows:

$$CTR_i = \frac{Total\,number\,of\,clicks}{Total\,number\,of\,impressions} \tag{3}$$

Furthermore, the average of CTR values of all items is calculated to obtain the final CTR value as follows:

$$CTR = \frac{\sum_{i=0}^{N} CTR_i}{N} \tag{4}$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

• **Relative CTR** : It is a relative measure of comparison between CTR scores given by different strategies. The Relative CTR for i_{t_h} strategy can be calculated as:

$$Relative\,CTR = \frac{CTR_i}{\sum CTR\,of\,other\,strategies} \quad (5)$$

• **CMR** : Cumulative Mean Reward is a useful measure to continuously analyze the performance of recommender systems over time.At any given point, if R is the number of rows evaluated so far, CMR is calculated as follows:

$$CMR_R = \frac{\sum_{i=0}^{R} Reward_i}{R} \quad (6)$$

D. Experimentation Setting

The experimentation settings for the proposed method uses a deep neural network with two hidden layers. The batch size is 1000, i.e., the model is updated after every trial, and the experience replay after the batch size is reached. The optimizer used is Adam and the loss is Binary Crossentropy.We experimented with three variations of k : 1,3 and 5

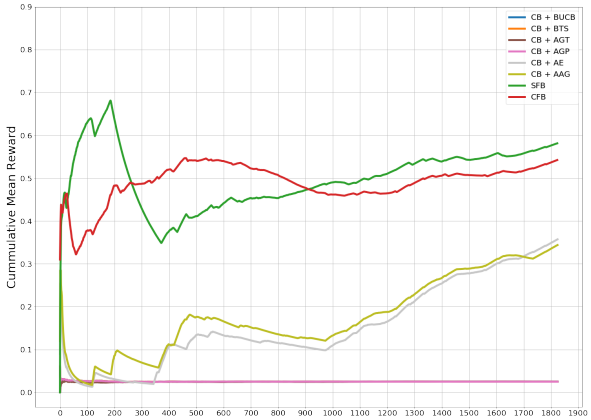
VII. Results and Analysis

The CTR scores for all the baseline models are summarized in Table III. RCTR and CMR trends can be observed from Figure 5 and Figure 4 respectively. SFB shows an average improvement of 23.4% in CTR values while CFB shows an average upgrade of 17.5%. Detailed analysis of the results are explained below. (All the abbreviations used below are mentioned in Table II.)

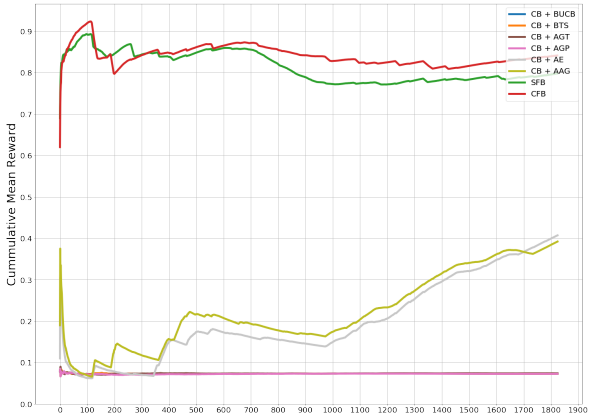
• *Adaptability*: It can be clearly observed from Figure 4 that our proposed methods are performing exceptionally well compared to the existing strategies for different values of k. The baselines like AAG and AE show some promise in terms of performance. However, since our techniques incorporate step-by-step update instead of batch update and reinforcement of recently learnt trends using experience replay, they are able to outperform AAG and AE by a significant margin.

• *Pattern grasping pace*:We can observe from Figure 4 that our techniques achieve their peak relatively early and maintain a consistent performance throughout whereas the baselines achieve their peak much later in the timeline. Thus, baselines learn over time rather than grasping the progression immediately. SFB and CFB learn patterns from the very beginning owing to partial feedback after every positive or negative reward for an action. Experience replay helps to stabilize the model which may incline towards randomness due to exploration.

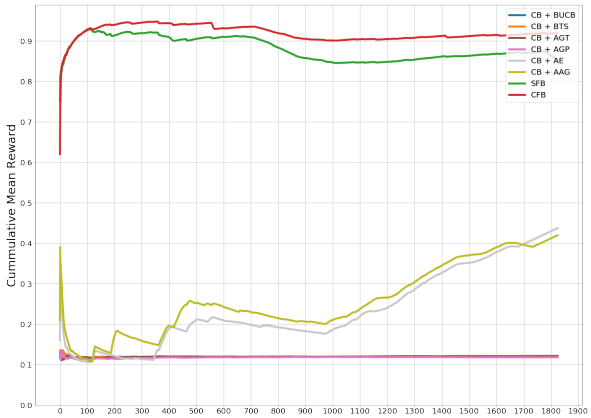
• *Static vs Dynamic Training*: It can be observed from Table III, that there is a stark difference between CTR values for RL and Non-RL based techniques. It points to the fact that any technique which relies on static model training, such as LR and W&D only learns the correlations present in the training interval which are bound to turn obsolete in most real time scenarios and thus inapplicable



(a) $k = 1$



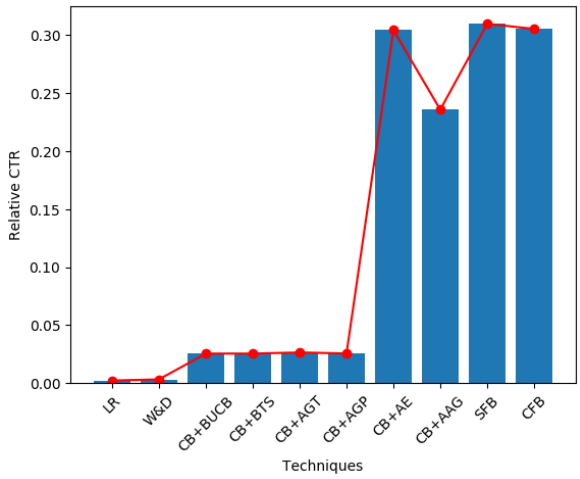
(b) $k = 3$



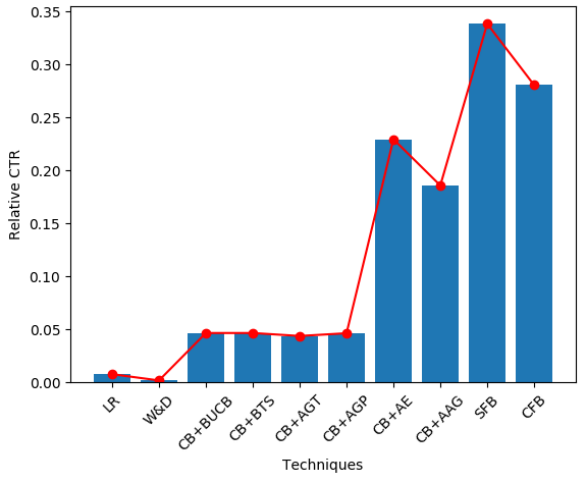
(c) $k = 5$

Fig. 4: CMR trends for all RL based policies for different values of k

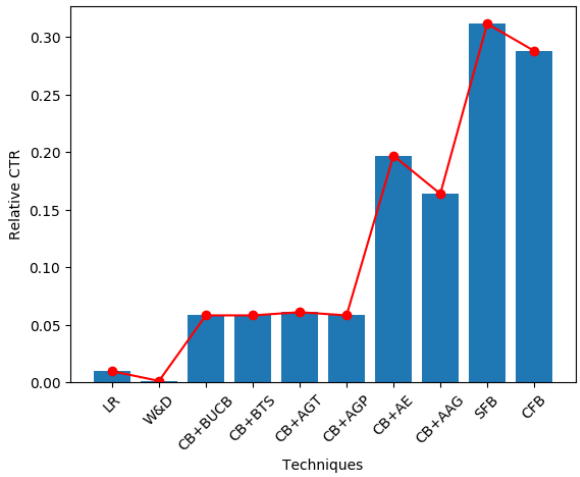
Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.



(a) $k = 1$



(b) $k = 3$



(c) $k = 5$

Fig. 5: Relative CTR trends for all RL-based policies for different values of k

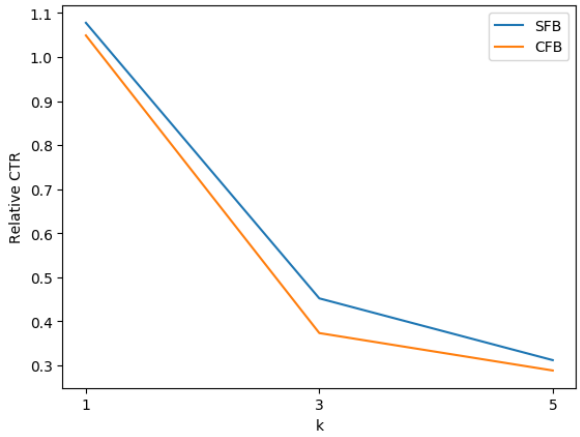


Fig. 6: The comparison of relative CTR of SFB and CFB with increasing k.

for the testing data. This explains the low relative CTR values for these methods as shown in Figures 5.

• *SFB vs CFB* : Although SFB and CFB both give higher CTR values than the baseline, SFB is performing better than CFB in terms of CTR (Figure 6). The performance of CFB may improve on data sets which have an evident relative skew between user features and item features, since this architecture provides a method of equalizing the weightage of user features and item features. Graphical representation of these trends in terms of RCTR is shown in Figure 5.

• *Handling explicit irregularities in user choices* : Another essential observation worth mentioning is the non-repeating dynamic trend present in time series datasets such as this one. Figure 3 shows the occurrence of 5 most popular article_ids over the complete timeline. It can be clearly observed that every article remains popular only for a certain interval of time and is hardly ever chosen again by the users after this peak interval has passed. This highlights two major strengths of our proposed framework. Firstly, any other technique which only relies on user context features might not be capable of recognizing which article would attain popularity in the future. Our technique however, takes the article structure into consideration, thus making it capable of understanding the future trends based on correlation between user context and arm structure. Secondly, since the occurrences in Figure 3 are not diffused but distinct, it can be inferred that the future trends are more likely to be related to the most recent trends rather than the older ones. Thus, replaying the recent accumulated experiences after certain fixed number of timesteps helps the agent to focus more on the relatively recent trends, hence giving more relevant recommendations.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

VIII. Conclusion

The proposed work provides two variants of feature bandit technique for real-time top-k recommendations. These are potent methods to overcome the existing issues of non-stationary environment and cold start problems, while embracing the user and item features together. After thorough experimentation, it is clear that these methods outperform the existing contextual bandit approaches. For future work, item clustering methods can be incorporated in our framework for performance optimization and scalability. Further, one can experiment with other exploration strategies such as dueling bandit gradient descent [14] that may positively impact the recommendations.

References

[1] G. Adomavicius and A. Tuzhilin. Context-aware recommender systems. In *Recommender systems handbook*, pages 217–253. Springer, 2011.

[2] S. Agrawal and N. Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on learning theory*, pages 39–1, 2012.

[3] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422, 2002.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[5] L. Baltrunas, B. Ludwig, and F. Ricci. Matrix factorization techniques for context aware recommendation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 301–304, 2011.

[6] D. A. Berry and B. Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5:71–87, 1985.

[7] Y. Cao and Q. Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10836–10846, 2019.

[8] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10, 2016.

[9] W. Chu, L. Li, L. Reyzin, and R. Schapire. Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214, 2011.

[10] D. Cortes. Adapting multi-armed bandits policies to contextual bandits scenarios. *arXiv preprint arXiv:1811.04383*, 2018.

[11] G. De Francisci Morales, A. Gionis, and C. Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 153–162, 2012.

[12] C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765, 2014.

[13] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[14] K. Hofmann, S. Whiteson, and M. De Rijke. Balancing exploration and exploitation in learning to rank online. In *European Conference on Information Retrieval*, pages 251–263. Springer, 2011.

[15] W. IJntema, F. Goossen, F. Frasincar, and F. Hogenboom. Ontology-based news recommendation. In *Proceedings of the 2010 EDBT/ICDT Workshops*, pages 1–6, 2010.

[16] A. Kabra, A. Agarwal, and A. S. Parihar. Cluster based deep contextual reinforcement learning for top-k recommendations. *arXiv preprint arXiv:2012.02291*, 2020.

[17] A. Kabra, A. Chopra, N. Puri, P. Badjatiya, S. Verma, P. Gupta, and B. Krishnamurthy. Mixboost: Synthetic oversampling using boosted mixup for handling extreme imbalance. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1082–1087. IEEE, 2020.

[18] M. N. Katehakis and A. F. Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

[19] M. Kompan and M. Bieliková. Content-based news recommendation. In *International conference on electronic commerce and web technologies*, pages 61–72. Springer, 2010.

[20] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[21] R. Kumar and P. K. Bala. Recommendation engine based on derived wisdom for more similar item neighbors. *Information Systems and e-Business Management*, 15(3):661–687, 2017.

[22] C. W. Lau. News recommendation system using logistic regression and naive bayes classifiers. 2011.

[23] L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[24] G. Lilien, P. Kotler, and K. ShridharMoorthy. Marketing models. prentice hall international, inc. 1992.

[25] P. Melville, R. J. Mooney, and R. Nagarajan. Content-boosted collaborative filtering for improved recommendations. *Aaai/iaai*, 23:187–192, 2002.

[26] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Play-

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

ing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[27] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[28] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.

[29] L. Terveen and W. Hill. Beyond recommender systems: Helping people help each other. *HCI in the New Millennium*, 1(2001):487–509, 2001.

[30] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval. Context-aware recommender systems for learning: a survey and future challenges. *IEEE Transactions on Learning Technologies*, 5(4):318–335, 2012.

[31] H. Wang, Q. Wu, and H. Wang. Factorization bandits for interactive recommendation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[32] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on September 22,2021 at 11:01:19 UTC from IEEE Xplore. Restrictions apply.

Electrical Engineering
https://doi.org/10.1007/s00202-021-01394-3

ORIGINAL PAPER

PV integrated grid synchronization technique using modified SOGI-FLL and zero-crossing detector

Hemant Saxena[1] · Alka Singh[1] · Jitendra Nath Rai[1] · Manoj Badoni[2]

Received: 15 September 2020 / Accepted: 25 August 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
The operation of a grid-connected photovoltaic (PV) system with improvement in power quality depends mostly on the estimation of synchronizing signals. A hybrid algorithm for the estimation of the synchronizing unit templates has been proposed in this paper. Modified second-order generalized integrator (MSOGI) has been combined with the frequency locked loop (FLL) and the traditional zero-crossing detector (ZCD) to extract the synchronizing signal. The synchronizing signal produced from the proposed technique has almost zero total harmonic distortion. Work on the proposed MSOGI-FLL-ZCD has been verified through simulation and experimental studies conducted under different grid voltage problems such as phase shift, frequency change, harmonic, DC-offset and noise. In addition, the proposed algorithm has been successfully applied to the three-phase grid-connected solar PV system for improving power quality.

Keywords Synchronization · Shunt active power filter · PV · Voltage source converter

1 Introduction

The efficient and reliable integration of renewable energy sources with the utility grid is based on power electronics-based converter technology. The accurate and reliable design of this system depends on the control technique employed under normal and abnormal grid conditions. The control structure of shunt active power filter (SAPF) requires the extraction of fundamental component of the load current and the estimation of synchronizing signals from the grid voltage [1–4]. A significant role in the control of grid-connected inverter is that of the synchronization unit, which achieves the function of extracting phase and frequency information from the grid voltage [5,6]. These are nonlinear controllers and can be designed in different reference frames viz. stationary and rotatory reference frames [7–9]. The use of FLL is relatively less common as compared to PLL for grid integration applications. However, developments in the field of small-signal modelling and filtering capability are enhancing the use of FLLs in power applications [10]. The design of PLL/FLL should be capable of working adequately under abnormal or distorted grid voltage conditions such as harmonics, sag, swell, DC offset, and frequency variations [11]. Several FLLs are proposed in the literature, among them the standard (SOGI-FLL) used to extract the fundamental component from the input signal and also its quadrature component [12]. Dual SOGI-FLL (DSOGI-FLL) can be formed by connecting two SOGI with FLL in parallel to enhance the transient response of the system in the presence of abnormal grid voltage [13]. It is observed that the presence of the DC component can significantly degrade the performance of the synchronization schemes. The sources for this component can be measurement errors, improper control of renewable sources, and signal conditioning circuits [14,15].

The presence of DC components can be eliminated using a mixed second- and third-order generalized integrator PLL (MSTOGI-PLL), which has been proposed for the three-phase system [16]. The modified SOGI-FLL has also been reported in the literature for estimating the fundamental load component [17]. A method of FLL based on a reduced-order resonant controller is reported in [18], which does not

✉ Alka Singh
alkasingh.eed@gmail.com

Hemant Saxena
hsaxena531@gmail.com

Jitendra Nath Rai
jnrai@dce.ac.in

Manoj Badoni
manojbadoni23@gmail.com

[1] Delhi Technological University, Delhi, India

[2] Thapar Institute of Engineering and Technology, Patiala, India

Published online: 24 September 2021

compute positive and negative sequence components and suffers from phase lock error in the presence of DC offset. A reduced-order generalized integrator (ROGI) is also utilized for grid synchronization, which can extract frequency, positive and negative sequence components from grid voltage in the presence of abnormalities, including DC component [19,20]. Thus, a few papers are available on new FLL-based synchronization techniques; however, their application for shunt compensation in PV-based systems is not discussed. SOGI-FLL estimates frequency more precisely than SOGI-PLL. The resonating frequency of SOGI-PLL is fixed at 50Hz while the resonating frequency of SOGI-FLL varies as the frequency of the input signal and it is decided by the FLL block.

Shunt compensation relies on the estimation of the reference currents for the generation of the switching signals in the grid-connected solar PV system. A suitable control unit is a prerequisite to generate accurate synchronizing signals. It is pertinent that the synchronizing signal should be distortion-free and in phase with the grid voltage, irrespective of the distortion of the input voltage. Zero crossing detection with the sine wave generator may be used for the generation of pure sinusoidal synchronizing signals [21,22]. However, the generation of synchronizing signals by the ZCD technique is not perfect; the results are badly affected in case of grid voltage contains noise or there is frequency variation. Thus, the ZCD technique cannot be used independently as an effective synchronization technique. In this paper, a new technique has been proposed, which uses the advantages of MSOGI-FLL for the filtered output signal and the accurate estimation of frequency. MSOGI-FLL has the capability of DC-offset elimination from the grid voltage signal. The filtered signal and the estimated frequency are input to the ZCD-based synchronizing signal generator. The hybrid combination of MSOGI-FLL-ZCD ensures accurate synchronization in different voltage quality issues, irrespective of the disturbance in the grid voltage signal. The important highlights of the paper are:

1. The proposed method estimates frequency precisely in different voltage quality problems such as phase change, frequency, noise, harmonic, DC-offset, and issues with combinations of different problems.
2. Perfect synchronization signals are ensured with less than 0.6% THD in the output signal under all voltage disturbances.
3. The proposed synchronization technique is applied for harmonic mitigation, power factor improvement and active power support in a three-phase grid-connected PV system.

The remaining sections of the paper are as follows: Sect. 2 describes the modelling of standard SOGI-FLL and its extended version MSOGI-FLL has been explained in Sect. 3 for the DC-offset rejection. Working of ZCD has been explained in Sect. 4 and the proposed MSOGI-FLL-ZCD is demonstrated in Sect. 5. The simulation results and analysis for the proposed technique is shown in Sect. 6. System configuration and its details have been explained in Sect. 7. The experimental result for the proposed technique has been shown in Sect. 8. Estimation of fundamental load current and generation of switching pulses for VSC explained in Sect. 9. Experimental results for SAPF operation and the final conclusion have been done in Sects. 10 and 11 respectively.

2 Modelling of standard SOGI-FLL

The conventional SOGI-PLL gives the filtered in-phase and quadrature-phase signal of the input sinusoidal signal [23]. The structure of the conventional second-order generalized integrator-phase locked loop (SOGI-PLL) and SOGI-FLL are shown in Fig. 1.

Consider the grid-voltage equation in a simpler form as-

$$v_g(t) = V_p \sin(\underbrace{\omega_0 t + \phi}_{\theta_{est}}) \tag{1}$$

In (1), v_g is the grid voltage, V_p is the peak of the fundamental, ω_0 is the standard grid frequency and θ is the phase of grid voltage. The filtered output of SOGI-FLL is obtained and expressed as in (2) and (3).

$$v_\alpha(t) = V_p \sin\theta_{est} \tag{2}$$

$$v_\beta(t) = V_p \cos\theta_{est} \tag{3}$$

In (2) and (3), the components v_α and v_β are the in-phase and the quadrature component of the grid voltage. V_p and θ_{est} are the estimated fundamental magnitude and the estimated phase angle of the grid voltage. The transfer functions of the in-phase and quadrature-phase output parameter of the SOGI are expressed as,

$$T_d(s) = \frac{v_\alpha(s)}{v_g(s)} = \frac{k_s\omega_{est}s}{s^2 + k_s\omega_{est}s + \omega_{est}^2} \tag{4}$$

$$T_q(s) = \frac{v_\beta(s)}{v_g(s)} = \frac{k_s\omega_{est}^2}{s^2 + k_s\omega_{est}s + \omega_{est}^2} \tag{5}$$

In (4) and (5), ω_{est} is the estimated grid frequency from the FLL block. The transfer function of the SOGI is similar to that of the standard second-order system. The value of gain (k_s) of SOGI-PLL is chosen as $k_s = \sqrt{2}$ for the optimum speed and the filtering capacity.

The SOGI-PLL gives perfect filtered orthogonal signal only if the grid frequency is almost constant and very close

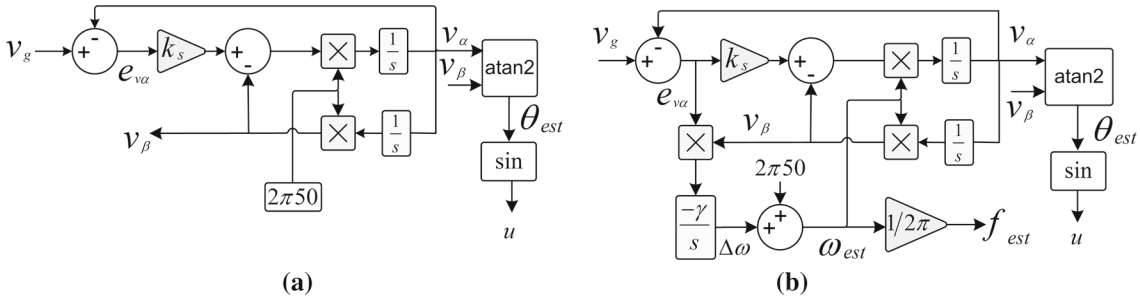


Fig. 1 Structure of **a** SOGI-PLL, **b** SOGI-FLL

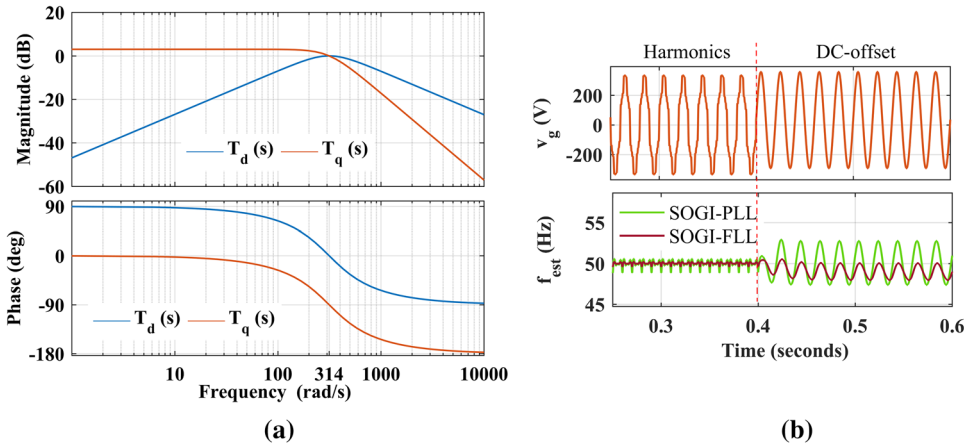


Fig. 2 **a** Bode-plot of SOGI, **b** estimated frequency by SOGI-PLL and SOGI-FLL

to the standard frequency. However, if there is a change in grid frequency, the output of the SOGI-PLL will not give perfect filtered orthogonal signals. An FLL is then required to estimate grid frequency, and it gives the perfect orthogonal signals, even when frequency deviates from the standard frequency [24]. The deviation in standard frequency is estimated by the FLL block and is expressed as,

$$\Delta\omega = \int \gamma e_{v\alpha} v_\beta(t)\mathrm{d}t \tag{6}$$

Frequency estimated by the FLL block is given as,

$$\omega_{\text{est}} = \Delta\omega + \omega_0 \tag{7}$$

The SOGI-FLL is advantageous over the SOGI-PLL, because the frequency estimation by SOGI-FLL is more precise and having less oscillations. Moreover, it does not require a PI controller to estimate phase and frequency information of the grid voltage as required by a PLL. The bode plot of the SOGI is shown in Fig. 2a. Bode plot of $T_d(s)$ is similar to the and band pass filter (BPF) and attenuates the harmonic and the DC-offset from the grid voltage while, bode plot of $T_q(s)$ shows the characteristics similar to the low pass filter (LPF), which attenuates the harmonics but is unable to attenuate the DC-offset.

Figure 2b shows the simulation analysis for the estimation of frequency by SOGI-PLL and SOGI-FLL under harmonic and DC-offset grid voltage conditions. Initially, only harmonics are present in the grid voltage, and the frequency estimated by the SOGI-FLL shows a better steady-state response than SOGI-PLL. At $t = 0.4$ s, the 10% of DC-offset is introduced in the grid voltage and harmonics removed. The results show that both the SOGI-PLL and SOGI-FLL are not capable of filtering the DC-offset. Hence, the SOGI-FLL requires a DC-offset rejection block to reject the DC-offset and estimate precise frequency.

3 MSOGI-FLL for DC-offset rejection

The structure of the modified SOGI-FLL has been shown in Fig. 3. Adding the DC-rejection block in the SOGI-FLL makes it a third-order system. The transfer functions of the MSOGI for the in-phase component, quadrature, and the offset component are expressed as,

$$H_d(s) = \frac{v_\alpha(s)}{v_g(s)} = \frac{k_{\text{dc}}\omega_{\text{est}}s^2}{s^3 + (k_s + k_{\text{dc}})\omega_{\text{est}}s^2 + \omega_{\text{est}}^2 s + k_{\text{dc}}\omega_{\text{est}}^3} \tag{8}$$

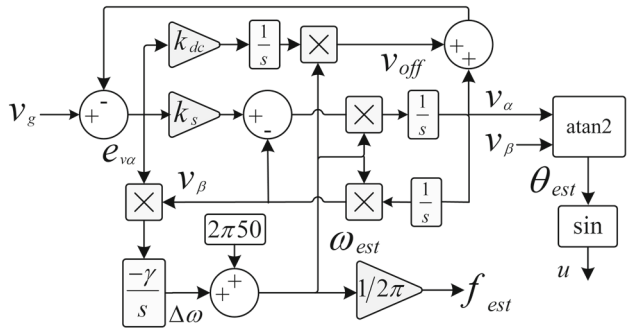


Fig. 3 Structure of MSOGI-FLL

$$H_q(s) = \frac{v_\beta(s)}{v_g(s)} = \frac{k_{\text{dc}}\omega_{\text{est}}^2 s}{s^3 + (k_s + k_{\text{dc}})\omega_{\text{est}}s^2 + \omega_{\text{est}}^2 s + k_{\text{dc}}\omega_{\text{est}}^3} \tag{9}$$

$$H_{\text{off}}(s) = \frac{v_{\text{off}}(s)}{v_g(s)} = \frac{k_{\text{dc}}\omega_{\text{est}}(s^2 + \omega_{\text{est}}^2)}{s^3 + (k_s + k_{\text{dc}})\omega_{\text{est}}s^2 + \omega_{\text{est}}^2 s + k_{\text{dc}}\omega_{\text{est}}^3} \tag{10}$$

The magnitude of the grid voltage (V_p) in single-phase can be estimated by,

$$V_p = \sqrt{v_\alpha^2 + v_\beta^2} \tag{11}$$

The bode plot of the MSOGI is shown in Fig. 4a. A reasonable value of k_{dc} (0.4) has been chosen for the optimum dynamic and steady-state response [16,25]. Bode plot $H_d(s)$ and $H_q(s)$ show similar characteristics of the band-pass filter, which attenuates the DC-offset of the grid voltage signal, and it estimates precise frequency from the grid voltage. The phase plot shows that the in-phase component is in the same phase with the grid voltage and the quadrature-phase is shifted with 90° from the grid voltage. The characteristics of $H_{\text{off}}(s)$ is similar to the low pass filter, which passes the DC-offset from the grid voltage.

Simulation response of the MSOGI-FLL is shown in Fig. 4b for the 10% of DC-offset in grid voltage. Initially, the grid voltage contains no DC-offset, and the performance for the frequency estimation by SOGI-FLL and MSOGI-FLL is almost the same. At $t = 0.4$ s, DC-offset of 10% is introduced in the grid voltage, and the frequency estimated by the MSOGI-FLL shows no oscillations while the SOGI-FLL shows oscillations and fails to track frequency accurately. The DC-offset (V_{off}) in the grid voltage is also estimated by the MSOGI-FLL and shown in Fig. 4b. Hence, the frequency estimated by the MSOGI-FLL shows superior performance over SOGI-PLL and SOGI-FLL under offset conditions. The estimation of accurate frequency depends upon the filtering capability of the algorithms. The performance of the MSOGI-FLL block can be further improved by integrating the ZCD block.

4 Zero-crossing detector (ZCD)

The working of the ZCD [22] is illustrated in Fig. 5.

Switch 1 detects the input signal and gives the time output with the help of a clock signal. Both the switches 1 and 2 give time signals which are complementary to each other. The output signals of switches 1 and 2 are given to the sinusoidal signal generator. The upper unit sine wave generator gives the positive half signal (u_+) with zero included and the lower unit sine wave generator with negative gain generates the sine wave with only negative half cycle (u_-) with zero excluded. For combining both the positive and negative

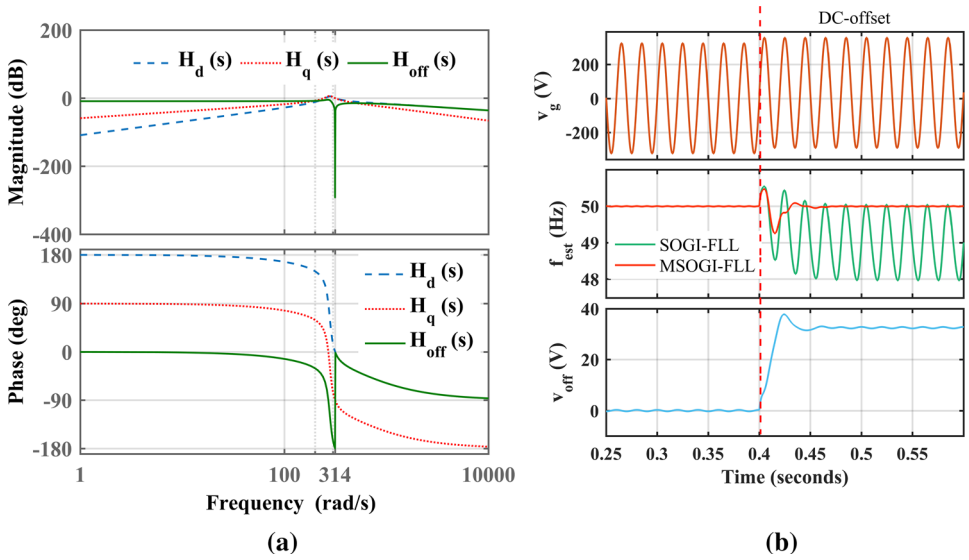


Fig. 4 **a** Bode-plot of MSOGI for $H_d(s)$, $H_q(s)$ and $H_{\text{off}}(s)$, **b** estimated frequency by MSOGI-FLL and SOGI-FLL for the grid voltage under DC-offset in grid voltage

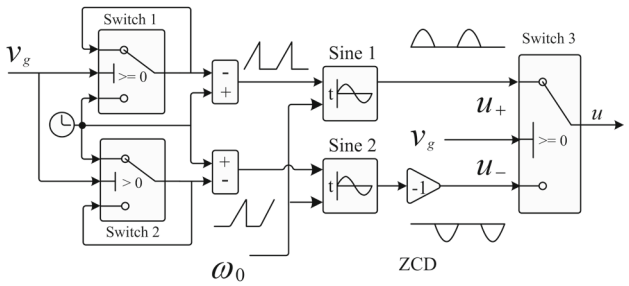


Fig. 5 Structure of ZCD

unit half-cycle switch 3 has been used. Combining u_+ and u_- gives the complete pure unit synchronizing signal. The switch 3 follows simple logic and behaves as (12).

$$u = \begin{cases} u_+ & \text{if } v_\alpha \geq 0 \\ u_- & \text{if } v_\alpha < 0 \end{cases} \tag{12}$$

The conventional ZCD generates the pure sine wave in case of phase change and harmonic conditions. However, ZCD used alone fails in case of frequency change of input signal and the presence of noise. In case of the noisy signal, there will be multiple zero-crossings at zero value of input signal. Therefore, conventional ZCD fails to perform well and produce errors at zero-crossings.

5 Proposed MSOGI-FLL-ZCD technique

The structure of the proposed MSOGI-FLL-ZCD algorithm is shown in Fig. 6.

The MSOGI-FLL gives the filtered output and the frequency of the grid voltage. The ZCD can be useful if the input of the ZCD is the filtered grid voltage with precise estimated frequency. The proposed algorithm uses the advantages of both the MSOGI-FLL and ZCD. It is proposed in this paper that the ZCD takes the filtered grid voltage and the estimated frequency and gives the pure sinusoidal synchronizing signal. The proposed method is expected to generate pure sinusoidal synchronizing signals under all grid voltage problems such as phase change, frequency change, DC-offset, harmonics, and noise. The application of ZCD is preferred after the MSOGI-FLL block as ZCD is an effortless technique and always ensures perfect sinusoidal wave (less than 0.6% THD) is generated. Since ZCD works on the zero crossings of the input signal, synchronized signals are almost perfectly sinusoidal and least distorted. The output of the ZCD block has the same frequency as the input signal, which itself is obtained from the MSOGI-FLL.

6 Simulation results and analysis of the MSOGI-FLL-ZCD technique

The simulation analysis of the proposed algorithm is shown in Fig. 7 for the different problems in grid voltage. Sampling time (T_s) for the simulation is 10 μs. Figure 7 shows the waveforms of grid voltage (v_g), system frequency, and generated unit sine wave (u). Several cases are represented to show the effectiveness of the proposed MSOGI-FLL-ZCD-based synchronizing technique, and these are given as follows:

Case I Phase shift: a phase shift of 30° is applied at $t = 0.4$ s (Fig. 7a)

Case II Frequency change: frequency change of +2 Hz at $t = 0.4$ s (Fig. 7b)

Case III Harmonic grid voltage: input grid voltage has harmonics 14.20% THD at $t = 0.4$ s (Fig. 7c)

Case IV Noisy grid voltage: the presence of noise in the input signal (4.65% THD) at $t = 0.4$ s (Fig. 7d)

Case V DC-offset; a 10% DC-offset is additionally present in the grid voltage at $t = 0.4$ s (Fig. 7e)

Case VI DC-offset with harmonics: worst-case scenario with 10% DC-offset and 14.20% THD containing 5th, 7th, 11th, 13th, and 19th order harmonic at $t = 0.4$ s (Fig. 7f).

In Fig. 7a, a 30° phase change occurs at time $t = 0.4$ s, due to which there is a sudden dip in frequency from 50 Hz to around 52 Hz, which recovers back within 8ms. Figure 7b shows system response in the event of sudden frequency change at $t = 0.4$ s, which also shows that the frequency is estimated to be 52 Hz within 8 ms. Figure 7c presents the case of harmonics included in the grid voltage at $t = 0.4$ s. Due to the presence of harmonics, there is distortion in the supply

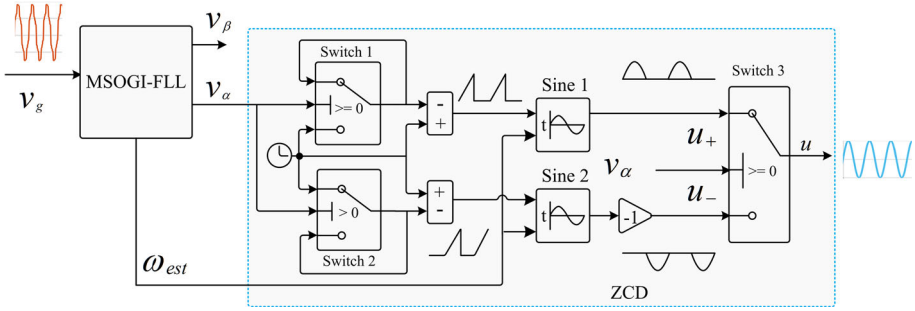


Fig. 6 Structure of proposed MSOGI-FLL-ZCD

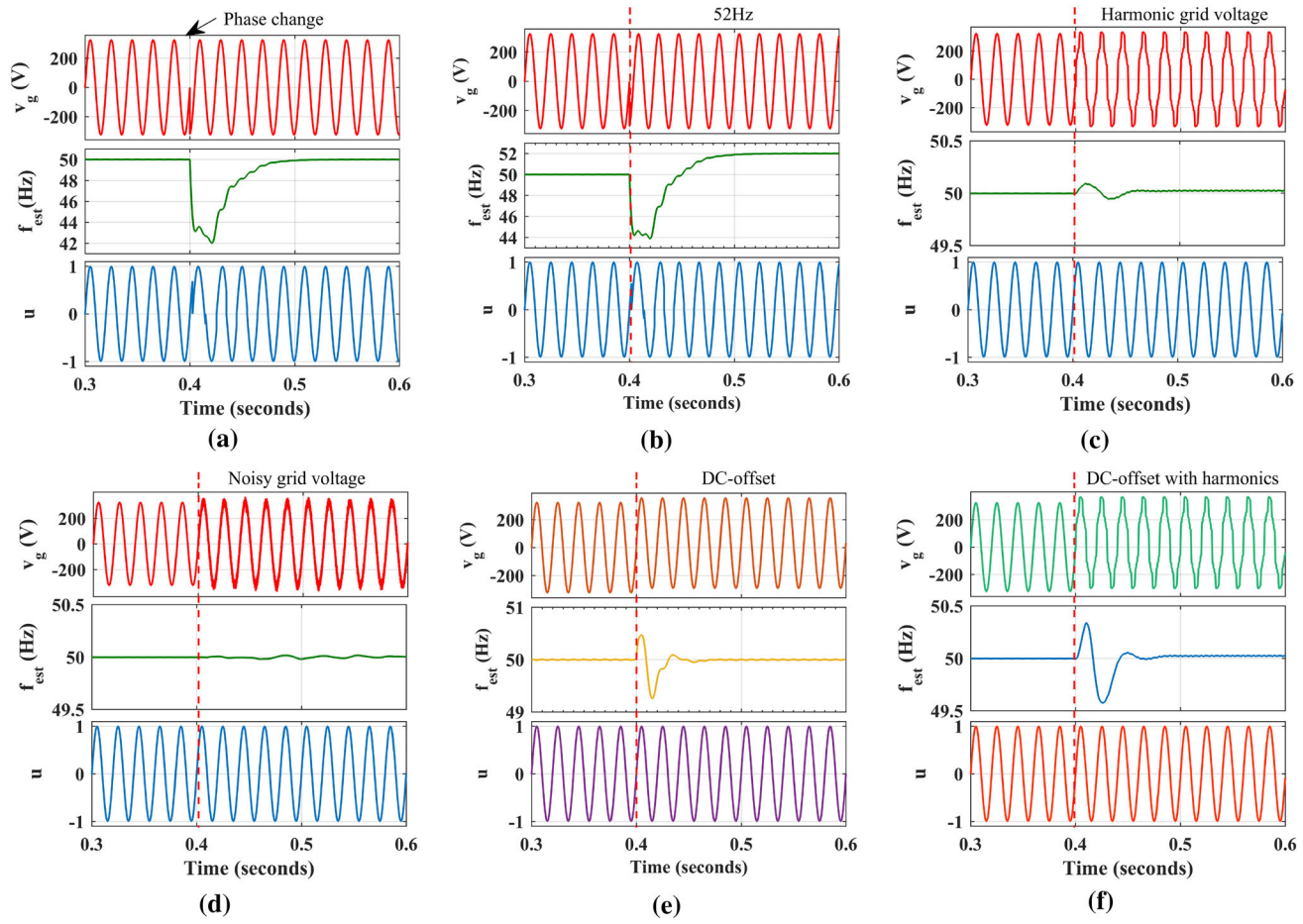


Fig. 7 Simulation performance analyses of different issues in grid voltage **a** Case I: phase change of 30°, **b** Case II: frequency change of +2 Hz, **c** Case III: harmonic grid voltage, **d** Case IV: noisy grid voltage, **e** Case V: DC-offset of 20%, **f** Case VI: DC-offset with harmonics

voltage, and THD is 14.20%. The frequency estimated by the proposed algorithm having negligible oscillations. Figure 7d illustrates the presence of noise in the grid voltage at $t = 0.4$ s, which introduces distortion in the grid voltage and hence a THD of 4.65%. The estimated frequency is free from oscillations, and due to the filtered input to ZCD, the estimated unit template is a pure sinusoid. Figure 7e shows the presence of a 10% DC-offset in the grid voltage at $t = 0.4$ s. The estimated frequency has a negligible steady-state error and settles within 2 cycles. Figure 7f shows the case when 10% of DC-offset with harmonics of 14.20% THD are added in the grid voltage at $t = 0.4$ s. The frequency obtained by the algorithm is free from oscillations, and the unit template has almost zero harmonic components. With the proposed MSOGI-FLL-ZCD, synchronizing signal (u) is obtained with very small transients in some cases (Ist and IInd) and with negligible transients in other cases (IIIrd to VIth). Also, the proposed algorithm does not depend much upon the filtering capability of SOGI-FLL. It requires precise frequency estimation and the detection of zero crossings. Therefore, the performance of the proposed technique is found to be suitable for unit vector estimation even after when various power quality disturbances are simultaneously present.

7 System configuration

Figure 8 illustrates the system configuration of a three-phase grid-connected solar PV system feeding the nonlinear load.

The grid supply has a small line impedance (R_s and L_s). A PV simulator is integrated with load/grid supply by using a three-phase VSC and interfacing inductor (L_d). Chroma made 62,050 H–600 S Programmable DC Power Supply Solar Array simulator has been used as PV and connected to the DC-link of VSC. The proposed algorithm has been applied to a three-phase system; hence for the voltage supply, a three-phase auto-transformer is used. The nonlinear load is modelled using a three-phase full-wave uncontrolled rectifier with resistive–inductive (R–L) series branch on the DC side. The uncontrolled rectifier is made of six 16A rated diode. dSpace 1102 has been used to process the algorithm and generate required pulses for the VSC. The rating of VSC

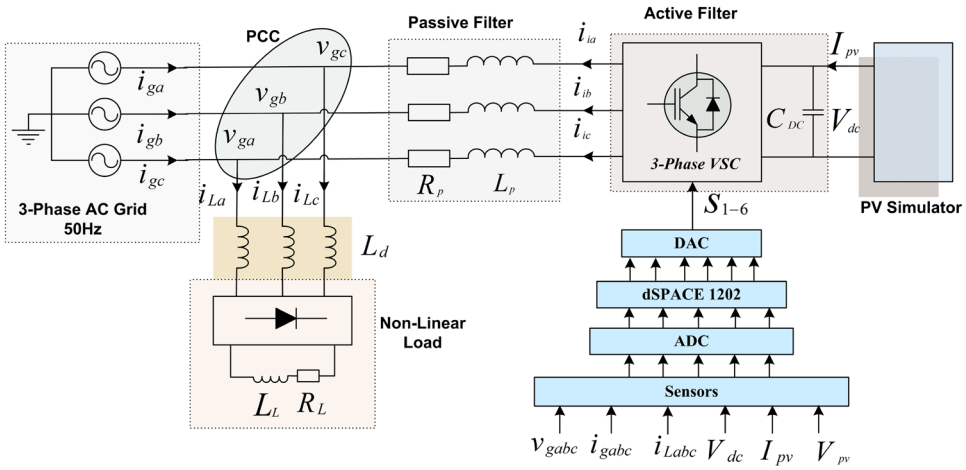


Fig. 8 System configuration for three-phase grid-connected PV system

is 25 kVA and is made by Semikron. The proposed algorithm has been tested experimentally for the estimation of synchronizing signals. Gwinstek make APS 1102 single-phase programmable AC/DC power supply that has been used to generate voltage quality problems. The proposed algorithm has been applied to the single-stage three-phase grid-tied PV system for power quality improvement. The proposed algorithm has also been compared with conventional ZCD and SOGI-FLL experimentally. The synchronizing signals generated from the proposed scheme have been used for power quality improvement in a grid-connected PV system.

8 Open-loop experimental results

The estimation of synchronizing signals and the THD of the obtained signal is presented for different grid voltage conditions. The purity of the synchronizing signal is reflected by the low THD content of the synchronizing signal.

Case I Harmonic grid voltage Figure 9 shows the THD of the grid voltage and the THD of the estimated synchronizing signals by conventional ZCD, MSOGI-FLL, and MSOGI-FLL-ZCD. The THD of the grid voltage is 13.797%, and THD of the synchronizing signal generated by the ZCD, MSOGI-FLL, and MSOGI-FLL-ZCD are 1.816%, 1.028%, and 0.403%, respectively. It is evident that the MSOGI-FLL-ZCD combination produces a near-perfect sine signal.

Case II Frequency change Figure 10 shows the THD of the grid voltage and the THD of the estimated synchronizing signals in the event of sudden frequency change to 52 Hz. The THD of the grid voltage is 0.729%, and THD of the synchronizing signal generated by the ZCD, MSOGI-FLL, and MSOGI-FLL-ZCD are 4.895%, 0.755%, and 0.346%, respectively.

Case III Noisy grid voltage Figure 11 shows the THD of the grid voltage and the THD of the estimated synchronizing signals in case of noisy grid conditions. The THD of the grid voltage is 6.632%, and THD of the synchronizing signal generated by the ZCD, MSOGI-FLL, and MSOGI-FLL-ZCD are 7.255%, 0.563%, and 0.451%, respectively. It is observed that the ZCD scheme is insufficient for generating synchronizing signals in the presence of noise.

Case IV Harmonic grid voltage with 20% DC-offset Figure 12 shows the THD of the grid voltage and the THD of the estimated synchronizing signals in case of harmonic grid with 20% DC-offset. The THD of the grid voltage is 13.732%, and THD of the synchronizing signal generated by the ZCD, MSOGI-FLL, and MSOGI-FLL-ZCD are 8.996%, 1.025%, and 0.560%, respectively. It is observed that the ZCD scheme is insufficient for generating synchronizing signals in the presence of DC-offset.

Based on the simulation and experimental performance and the characteristics of the algorithm, a comparison table has been drawn and shown in Table 1.

THD of the unit vector template or synchronizing signal obtained by the proposed algorithm is less than conventional ZCD and MSOGI-FLL. Equal sampling time of 50 μs is used for the fair comparison between the algorithms. Table 1 clearly shows the proposed algorithm has better performance in different voltage conditions.

9 Estimation of fundamental load component and generation of switching pulses for VSC

The application of the synchronizing scheme developed to the grid-connected solar PV system is considered in this section. Functions such as feeding active power to the load/grid and compensating power quality problems are primarily considered. These functions require the generation of synchronizing signals for generating the fundamental component of the load current. The fundamental load component is estimated by the SOGI algorithm.

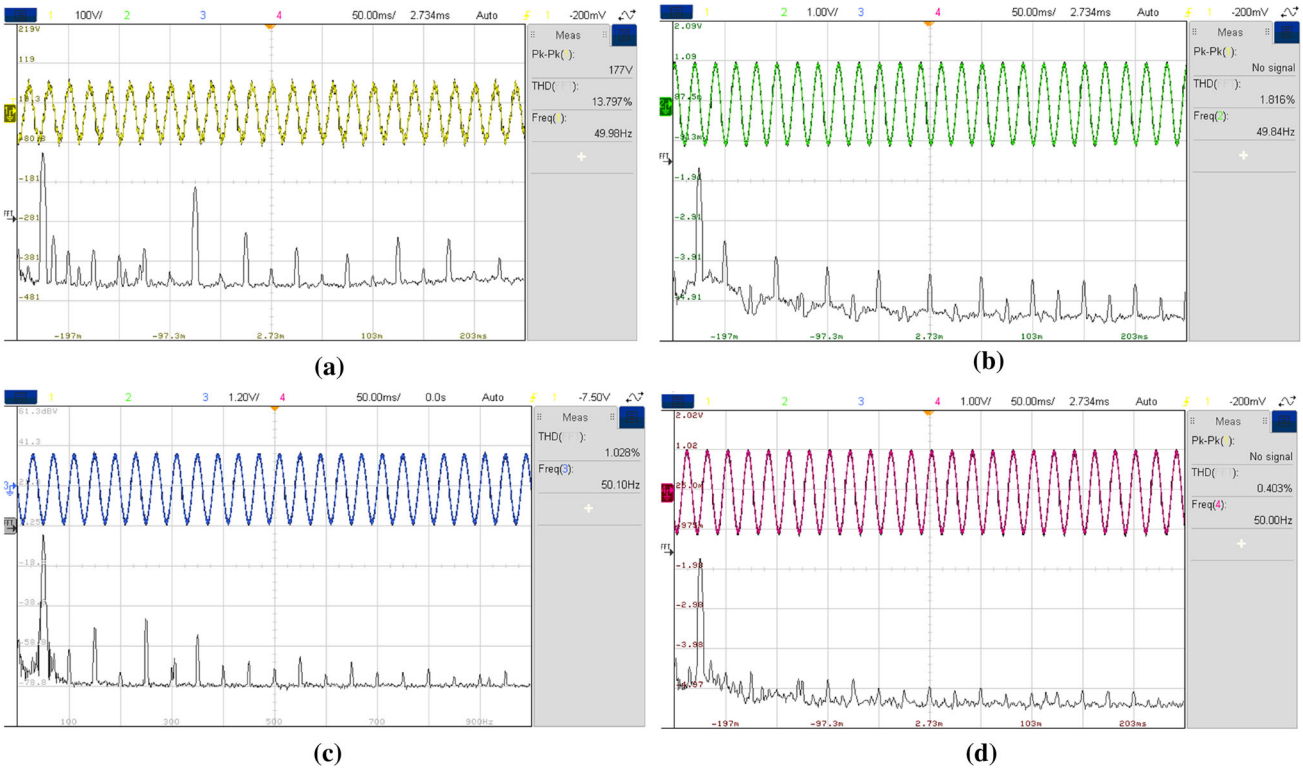


Fig. 9 THD analysis for Case I (**a**) grid voltage (**b**) unit synchronizing signal by ZCD (**c**) unit synchronizing signal by MSOGI-FLL (**d**) unit synchronizing signal by MSOGI-FLL-ZCD

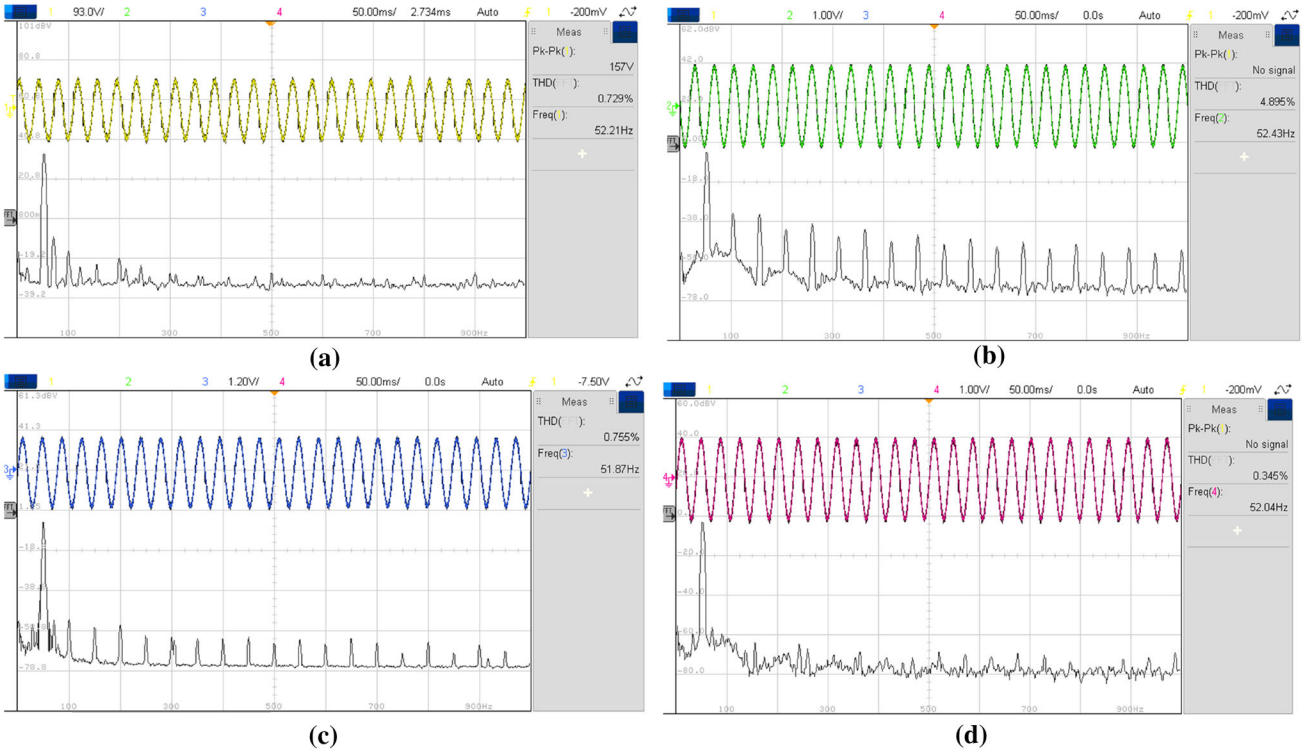


Fig. 10 THD analysis for Case II (**a**) grid voltage (**b**) unit synchronizing signal by ZCD (**c**) unit synchronizing signal by MSOGI-FLL (**d**) unit synchronizing signal by MSOGI-FLL-ZCD

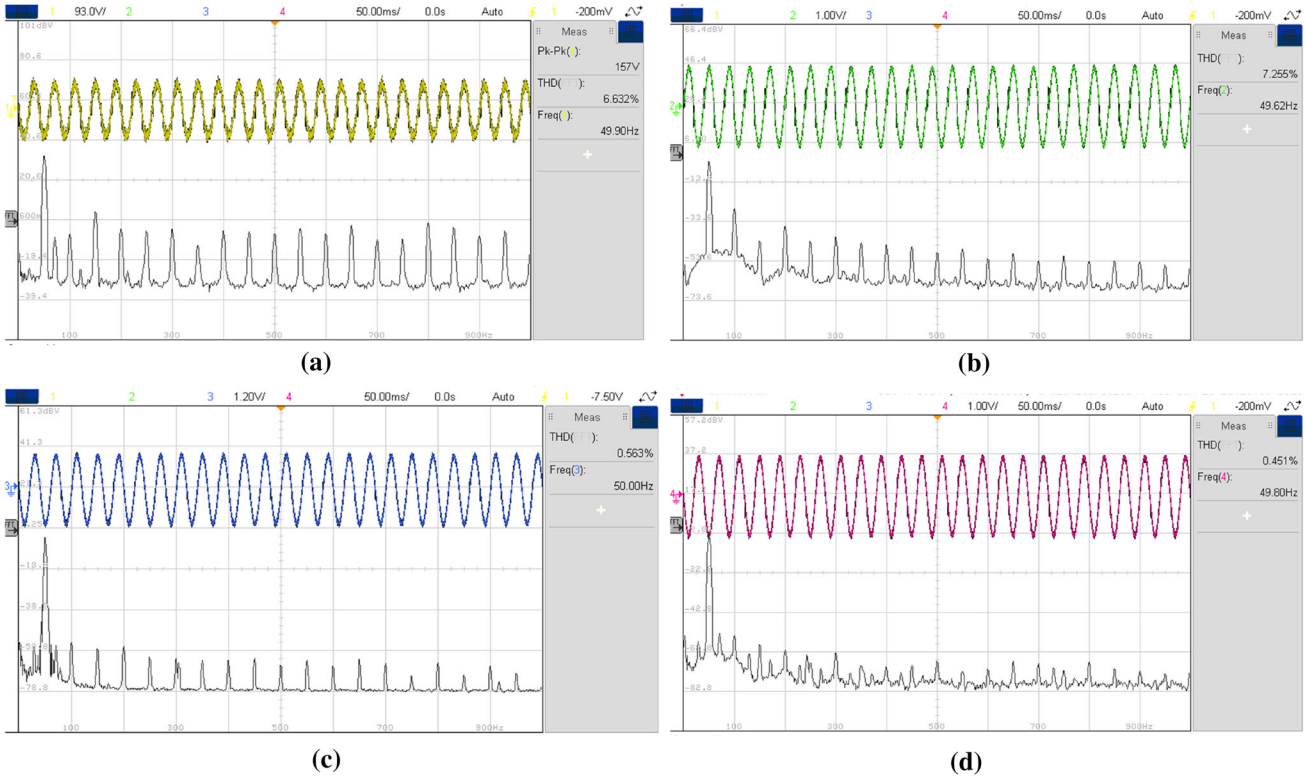


Fig. 11 THD analysis for Case III (**a**) grid voltage (**b**) unit synchronizing signal by ZCD (**c**) unit synchronizing signal by MSOGI-FLL (**d**) unit synchronizing signal by MSOGI-FLL-ZCD

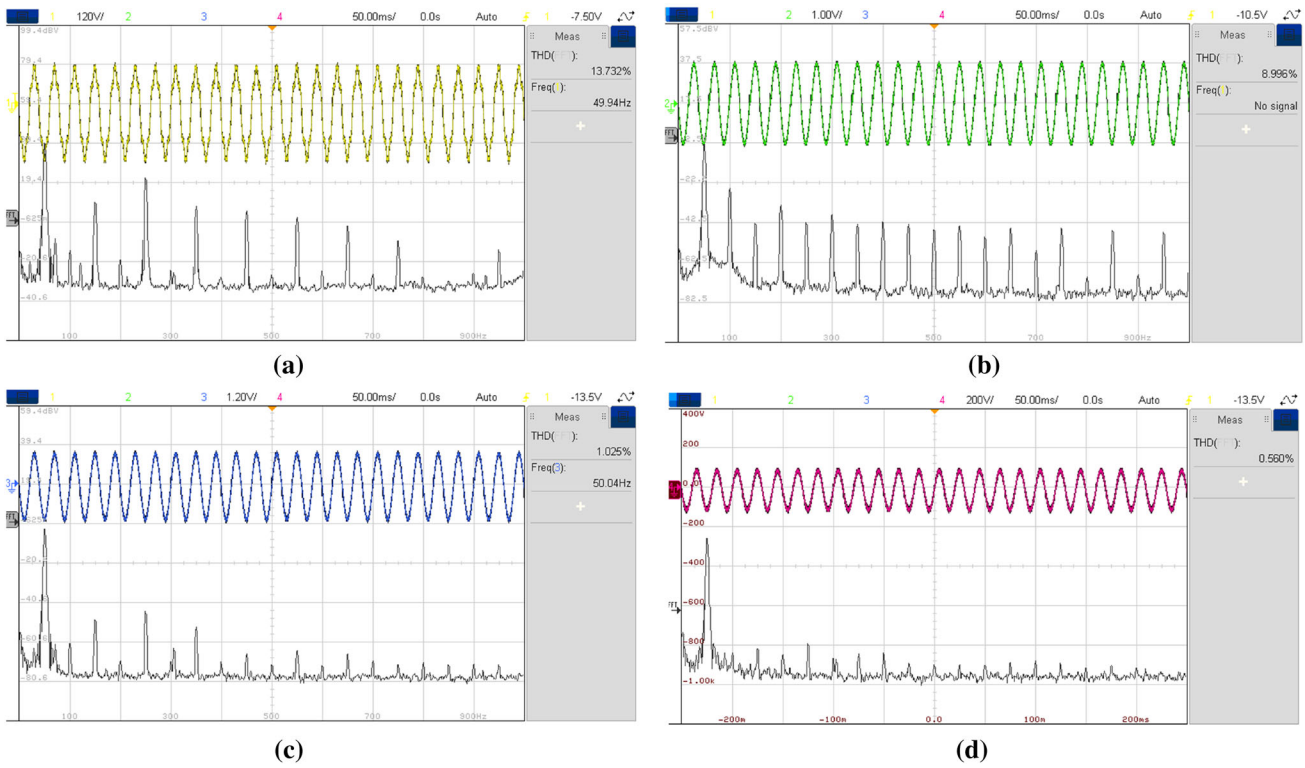


Fig. 12 THD analysis for Case IV (**a**) grid voltage (**b**) unit synchronizing signal by ZCD (**c**) unit synchronizing signal by MSOGI-FLL (**d**) unit synchronizing signal by MSOGI-FLL-ZCD

Table 1 Comparison between conventional ZCD, MSOGI-FLL and MSOGI-FLL-ZCD

Operations	Conventional ZCD	MSOGI-FLL	MSOGI-FLL-ZCD
Type	Non-adaptive	Adaptive	Adaptive
Frequency estimation	No	Yes	Yes
DC-offset rejection	No	No	Yes
Complexity	Low	Very Low	Low
Case I harmonic grid (THD = 13.797%)	1.816%	1.028%	0.403%
Case II frequency change (THD = 0.729%)	4.895%	0.755%	0.346%
Case III noisy grid voltage (THD = 6.632%)	6.632%	0.563%	0.451%
Case IV harmonic grid with 20% DC-offset (THD = 13.732%)	8.996%	1.025%	0.560%

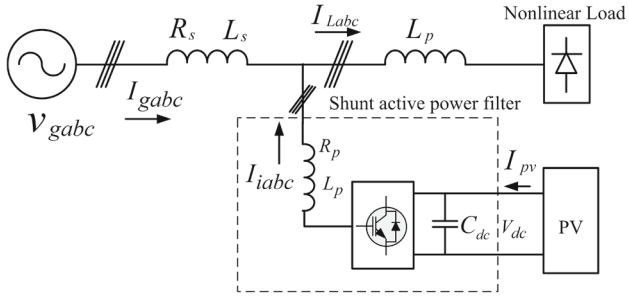


Fig. 13 Single-line schematic diagram for three-phase grid connected-PV system

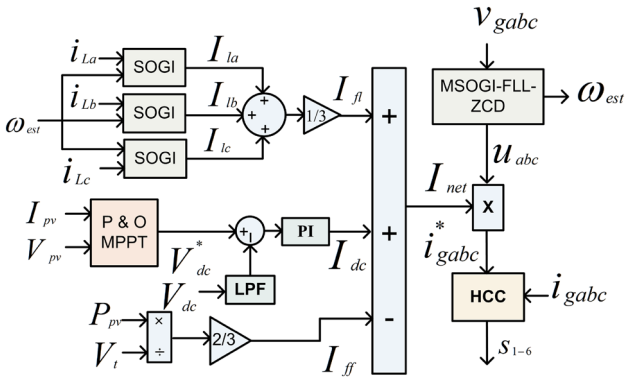


Fig. 14 Control structure with the proposed algorithm

The resonating frequency of the SOGI algorithm is estimated by the MSOGI-FLL applied to the grid voltage signal. The proposed algorithm has been integrated into the three-phase grid-tied PV system. The working of the SOGI is already explained in Sect. 2, and the fundamental load component of phase-a is estimated as,

$$I_{\mathrm{La}} = \sqrt{i_{a\alpha}^2 + i_{a\beta}^2} \tag{13}$$

Similarly, the fundamental component of phase-b and phase-c is given as

$$I_{\mathrm{Lb}} = \sqrt{i_{b\alpha}^2 + i_{b\beta}^2} \text{ and } I_{\mathrm{Lc}} = \sqrt{i_{c\alpha}^2 + i_{c\beta}^2} \tag{14}$$

The average fundamental component of three-phase load current is explained as

$$I_{\mathrm{fl}} = \frac{I_{\mathrm{La}} + I_{\mathrm{Lb}} + I_{\mathrm{Lc}}}{3} \tag{15}$$

The characteristics of PV are nonlinear in nature. Hence, an MPPT technique is required to achieve maximum power from the PV array. Perturb and Observe ($P\&O$) technique is used to generate reference voltage (V_{dc}^*), as this is a simple and efficient MPPT technique [26]. The DC bus voltage is regulated by using a proportional-integral (PI) controller. The reference DC bus voltage (V_{dc}^*) is compared with sensed DC bus voltage (V_{dc}) voltage error (V_{de}) is obtained. The error component (V_{de}) is given input to the PI, and thus, output (I_{dc}) is acquired at kth sampling instant as,

$$I_{\mathrm{dc}}(k) = I_{\mathrm{dc}}(k-1) + k_p\{V_{\mathrm{de}}(k) - V_{\mathrm{de}}(k-1)\} + k_i V_{\mathrm{de}}(k) \tag{16}$$

In the control algorithm, the contribution of PV current is termed as feed forward current (I_{ff}) and it is estimated by (17).

$$I_{\mathrm{ff}} = \frac{2P_{\mathrm{pv}}}{3V_t} \tag{17}$$

Here, P_{pv} is the power given by PV and V_t is the magnitude of three-phase voltage. The output of the PI controller (I_{dc}) is added to the fundamental load component (I_{fl}), and the feed-forward term (I_{ff}) is subtracted to obtain reference current component (I_{net}) as

$$I_{\mathrm{net}} = I_{\mathrm{fl}} + I_{\mathrm{dc}} - I_{\mathrm{ff}} \tag{18}$$

The reference current component (I_{net}) is multiplied with MSOGI-FLL-ZCD generated synchronizing signals (u_{abc}),

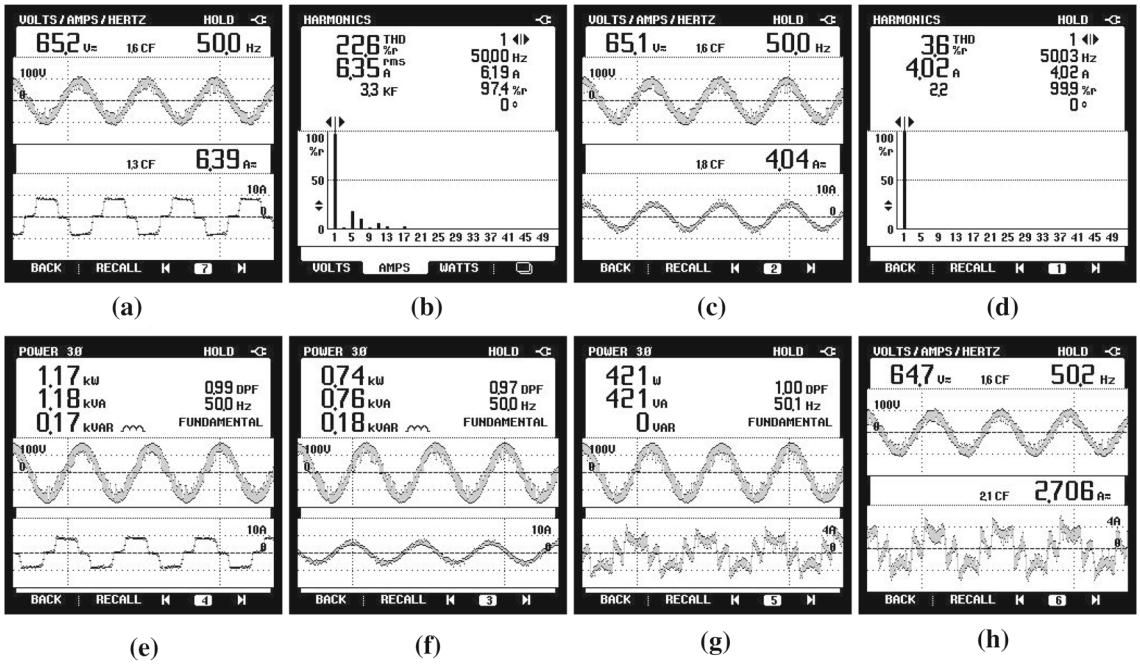


Fig. 15 Closed-loop steady state experimental results for three-phase grid connected system **a** v_{ga} and i_{La}, **b** THD of i_{La}, **c** v_{ga} and i_{ga}, **d** THD of i_{ga}, **e** load power demand, **f** grid power, **g** VSC power, **h** v_{ga} and i_{ia}

to obtain reference grid currents as

$$i^*_{gabc} = I_{net} u_{abc} \tag{19}$$

The reference grid currents (i^*_{gabc}) are compared with sensed grid currents (i_{gabc}) and the produced current error is passed through hysteresis current controller (HCC) for the generation of six switching pulses ($S_1 - S_6$) for VSC. Figures 13 and 14 show the single-line schematic and control structure for the three-phase single-stage grid-tied PV system under a nonlinear load, respectively.

10 Experimental results of the estimation of fundamental current applied to SAPF operation

The proposed synchronization algorithm has been implemented and tested for the SAPF operation. SAPF consists of the VSC, interfacing inductors and control algorithm to generate required pulses for the VSC. Figure 15 shows the closed-loop steady-state results for the SAPF operation in a grid-connected PV system.

The nonlinear load is connected to the grid, and the nonlinear current waveform is shown in Fig. 15a. THD of the nonlinear load current is 22.6% and shown in Fig. 15b. When no SAPF operation has been performed, the THD of the grid current is the same as the load current. After applying the SAPF operation, the THD of the grid current is observed to improve to 3.6%, which follows the IEEE-519. Grid current and its THD are shown in Fig. 15c and d, respectively. Load power demand, grid power, VSC power, and its current waveforms are shown in Fig. 15e–g. Figure 15h shows the PCC voltage and the VSC current. The active power demanded by the load is 1.17 kW, and the power delivered by PV via VSC is 421 W. The extra power demanded by the load is supplied by the grid (0.74 kW). The proposed algorithm has been successfully applied practically for harmonic mitigation in the three-phase distribution side. The power generated from the PV simulator is transferred via VSC to the load without affecting the quality of the grid current.

11 Conclusion

An improved MSOGI-FLL algorithm with a ZCD operator has been proposed in this paper for the estimation of synchronizing signals. The proposed algorithm has been compared with conventional ZCD and SOGI-FLL in terms of THD of the synchronizing signal. The THD of the unit sine template obtained from the proposed algorithm MSOGI-FLL-ZCD is observed to be the least under various voltage distortions. Further, the technique is applied to a three-phase grid-connected system for feeding real power as well as demonstrating shunt compensation. Both the functions are effectively performed, and the grid current is an almost pure sine wave. Thus, the feasibility of the proposed MSOGI-FLL-ZCD is effectively demonstrated for synchronization

purposes and for efficient controller design in mitigating power quality issues.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00202-021-01394-3.

Acknowledgements The authors are thankful to the Department of Science and Technology, Science and Engineering Research Board for the Sponsored Project (EMR/2016/001874).

Appendix

Experimental system parameters $V_{\mathrm{LL}}(RMS) = 110$ V, $f = 50$ Hz, $L_d = 1$ mH, $L_p = 3$ mH, Non-linear load; Full bridge rectifier with 2 kW rated resistive load for 240 V and $L_L = 80$ mH, $V_{\mathrm{dc}} = 200$ V, $k_p = 0.05$, $k_i = 0.006$, Sampling Time $T_s = 50$ µs, PV array parameter from simulator; maximum power point voltage $V_{\mathrm{mp}} = 200$ V, Maximum power point current $I_{\mathrm{mp}} = 2.5$ A, Maximum power $P_{\mathrm{mp}} = 0.5$ kW.

References

1. Repak M, Otcenasova A, Altus J, Regula M (2017) Grid-tied power converter for model of photovoltaic power plant. Electr Eng 99(4):1377–1391
2. Sedo J, Kascak S (2017) Design of output LCL filter and control of single-phase inverter for grid-connected system. Electr Eng 99(4):1217–1232
3. Sundaram E, Venugopal M (2016) On design and implementation of three phase three level shunt active power filter for harmonic reduction using synchronous reference frame theory. Int J Electr Power Energy Syst 81:40–47
4. Kumar R, Bansal HO (2018) Shunt active power filter: current status of control techniques and its integration to renewable energy sources. Sustain Cities Soc 42:574–592
5. Karimi-Ghartema M (2014) Enhanced phase-locked loop structures for power and energy applications. Wiley, Hoboken
6. Rodriguez P, Luna A, Munoz-Aguilar RS, Etxeberria-Otadui I, Teodorescu R, Blaabjerg F (2012) A stationary reference frame grid synchronization system for three-phase grid-connected power converters under adverse grid conditions. IEEE Trans Power Electron 1(27):99–112
7. Golestan S, Guerrero JM, Musavi F, Vasquez JC (2019) Single-phase frequency-locked loops: a comprehensive review. IEEE Trans Power Electron 34(12):11791–11812
8. Pal K, Kumar S, Singh B, Kandpal TC (2019) Improved phase-locked loop-based control for grid-integrated pv system. IET Renew Power Gener 14(5):705–712
9. Al-Gahtani SF, Nelms RM (2020) A frequency adaptive control scheme for a three-phase shunt active power filter. Electr Eng 103:595–606
10. Bei T-Z, Wang P (2016) Robust frequency-locked loop algorithm for grid synchronisation of single-phase applications under distorted grid conditions. IET Gener Trans Distrib 10(11):2593–2600
11. Xu J, Bian S, Liu M, Zhang Z, Xie S (2019) Stability analysis of grid-connected inverters during the transient of grid voltage fluctuations in weak grid cases. In: 2019 IEEE energy conversion congress and exposition (ECCE), pp 5185–5190. IEEE
12. Shah P, Hussain I, Singh B (2018) Single-stage secs interfaced with grid using ISOGI-FLL-based control algorithm. IEEE Trans Ind Appl 55(1):701–711
13. Yan Q, Zhao R, Yuan X, Ma W, He J (2018) A DSOGI-FLL-based dead-time elimination PWM for three-phase power converters. IEEE Trans Power Electron 34(3):2805–2818
14. Xie M, Wen H, Zhu C, Yang Y (2017) DC-offset rejection improvement in single-phase SOGI-PLL algorithms: methods review and experimental evaluation. IEEE Access 5:12810–12819
15. Kanjiya P, Khadkikar V, Moursi MSE (2018) Adaptive low-pass filter based DC offset removal technique for three-phase PLLS. IEEE Trans Industr Electron 65(11):9025–9029
16. Zhang C, Zhao X, Wang X, Chai X, Zhang Z, Guo X (2018) A grid synchronization PLL method based on mixed second-and third-order generalized integrator for DC offset elimination and frequency adaptability. IEEE J Emerg Select Top Power Electron 6(3):1517–1526
17. Patil K, Patel HH (2017) Modified SOGI based shunt active power filter to tackle various grid voltage abnormalities. Eng Sci Technol Int J 20(5):1466–1474
18. Dong S, Li Y, Wang A, Xi W (2015) Control of PMSG wind turbines based on reduced order resonant controllers under unbalanced grid voltage conditions. In: 2015 IEEE 11th international conference on power electronics and drive systems, pp 326–329. IEEE
19. Jorge SG, Busada CA, Solsona J (2013) Reduced order generalised integrator-based current controller applied to shunt active power filters. IET Power Electron 7(5):1083–1091
20. Xie C, Li K, Zhao X, Vasquez JC, Guerrero JM (2017) Reduced order generalized integrators with phase compensation for three-phase active power filter. In: 2017 IEEE applied power electronics conference and exposition (APEC), pp 2759–2766. IEEE
21. Konara KMSY, Kolhe ML, Sankalpa WGCA (2015) Grid synchronization of dc energy storage using voltage source inverter with ZCD and PLL techniques. In: 2015 IEEE 10th international conference on industrial and information systems (ICIIS), pp 458–462. IEEE
22. Azrik M, Ahmed KH, Finney SJ, Williams BW (2012) Voltage synchronization scheme based on zero crossing detection for parallel connected inverters in ac microgrids. In: 2012 IEEE international symposium on industrial electronics, pp 588–593. IEEE
23. Han Y, Luo M, Zhao X, Guerrero JM, Xu L (2016) Comparative performance evaluation of orthogonal-signal-generators-based single-phase PLL algorithms-a survey. IEEE Trans Power Electron 31(5):3932–3944
24. Yi H, Wang X, Blaabjerg F, Zhuo F (2017) Impedance analysis of SOGI-FLL-based grid synchronization. IEEE Trans Power Electron 32(10):7409–7413
25. Xie M, Wen H, Zhu C, Yang Y (2017) DC offset rejection improvement in single-phase SOGI-PLL algorithms: methods review and experimental evaluation. IEEE Access 5:12810–12819
26. Devassy S, Singh B (2017) Design and performance analysis of three-phase solar PV integrated UPQC. IEEE Trans Ind Appl 54(1):73–81

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

Relation between Schrödinger's Equation and Einstein's Equation

Aman Yadav[1]

ABSTRACT:

The relationship between Einstein's Field Equation and Schrodinger's Equation is examined in this work. I adjusted Schrodinger's Equation to offer the solution, and utilizing the wave equation, I came up with two cases: In case 1, I discovered the structure and dimension of the equations in a manner similar to Einstein's Field Equation, and in case 2, the Helmholtz equation replaces the modified Schrodinger's equation. Finally, the findings suggested that wave functions may have relevance beyond determining the position of a particle, and that they may be used to determine the structure of space-time at the quantum level.

1. INTRODUCTION:

For decades physicists are struggling to identify a solid relationship between general relativity and Quantum Mechanics but are facing defeat [1]. These two subjects forms the bedrock of the current understanding of physics yet general relativity is unable to explain the structure of space-time at a small scale. In quantum mechanics, events produced by the interaction of subatomic particles happen in quantum leaps, with probabilistic rather than definite outcomes which we perceive in General Relativity. The non-deterministic nature of quantum mechanics makes it highly inefficient to understand the quantum world; instead, we can determine the probability (density) of a wave-like particle to exist at a certain position. For example, If ψ is the wave function of a particle then |ψ|^2 is the probability (density) of that particle at a certain position, let's say "x". Yet, there is a simple approach through which we can modify Schrödinger's Equation in a manner that can be directly comparable with Einstein's Field Equation in conclusion giving us the physical significance of Wave Function ψ.

In the next section I have shown some calculations that directly describe the conversion of Schrödinger's equation to the required form, still, the left side of the resultant equation can be expanded and it requires more study to convert it into a more meaningful tensor notation.

2. Modification of Schrodinger's Equation:

Time independent Schrodinger's Equation (3 Dimensional, particle of mass m in a box of dimensions "$x \times y \times z$"):

$$\nabla^2\Psi\,(x,y,z) + \frac{8\pi^2 m(E-V)}{h^2}\Psi = 0$$

$$\nabla^2\Psi\,(x,y,z) = -\frac{8\pi^2 m(E-V)}{h^2}\Psi \qquad (1)$$

(Note that the terms in the brackets denote the derivatives taken with respect to the given variables.)

Using, wave-function $\Psi = Ae^{\left(\frac{i}{\hbar}(Px-Et)\right)}$, and taking its double partial derivative with respect to time coordinate.

[1] Undergraduate Student (Minor Under Engineering Physics) in Delhi Technological University, New Delhi
Email ID: amanyadav_2k19ce017@dtu.ac.in

$$\frac{\partial^2\Psi\,(t)}{\partial t^2} = \frac{(-i.E)^2}{\hbar^2}\Psi = -4\pi^2\frac{E^2}{\mathrm{h}^2}\,\Psi$$

Now, expanding one energy term in the equation in terms of kinetic energy ($\frac{1}{2}mv^2,\ where\ "v"\ is\ the\ speed\ of\ the\ particle$) and potential energy ($V$) and leaving another term intact, we will get.

$$\frac{\partial^2\Psi\,(t)}{\partial t^2} = -4\pi^2\frac{(\frac{1}{2}mv^2+V)\times E}{\mathrm{h}^2}\,\Psi = -\frac{8\pi^2}{\mathrm{h}^2}\left(\frac{1}{4}v^2\left(m+\frac{2V}{v^2}\right)E\right)\Psi$$

$$\frac{1}{v^2}\frac{\partial^2\Psi\,(t)}{\partial t^2} = -\frac{8\pi^2}{\mathrm{h}^2}\left(\frac{1}{4}\left(m+\frac{2V}{v^2}\right)E\right)\Psi \qquad (2)$$

Both eq(1) and eq(2) are dimensionally satisfying each other now, adding both of the equation,

$$\nabla^2\Psi\,(x,y,z) + \frac{1}{v^2}\frac{\partial^2\Psi\,(t)}{\partial t^2} = -\frac{8\pi^2}{\mathrm{h}^2}\left(m(E-V)+\frac{1}{4}\left(m+\frac{2V}{v^2}\right)E\right)\Psi$$

Considering right side of the equation first,

$$= -\frac{8\pi^2}{\mathrm{h}^2}\left(\left(\frac{Em}{4}+\frac{EV}{2v^2}\right)+Em-Vm\right)\Psi = -\frac{8\pi^2}{\mathrm{h}^2}\left(\frac{5Em}{4}+V\left(\frac{E}{2v^2}-m\right)\right)\Psi$$

$$= -\frac{8\pi^2}{\mathrm{h}^2}\left(\frac{5Em}{4}+\frac{V}{4}\left(\frac{2E}{v^2}-m-3m\right)\right)\Psi \qquad \because \frac{2E}{v^2}-m=\frac{2V}{v^2}$$

$$\Longrightarrow -\frac{8\pi^2}{\mathrm{h}^2}\left(\frac{5Em}{4}+\frac{V}{4}\left(\frac{2V}{v^2}\right)-\frac{3Vm}{4}\right)\Psi = -\frac{8\pi^2}{\mathrm{h}^2}\left(\frac{5Em}{4}+\frac{V^2}{2v^2}-\frac{3Vm}{4}\right)\Psi$$

$$= 8\pi\left(-\frac{\pi}{\mathrm{h}^2}\left(\frac{5Em}{4}+\frac{V^2}{2v^2}-\frac{3Vm}{4}\right)\right)\Psi \qquad (3)$$

Now, considering left side of the equation,

$$= \nabla^2\Psi\,(x,y,z) + \frac{1}{v^2}\frac{\partial^2\Psi\,(t)}{\partial t^2} \qquad (4)$$

And according to wave equation we can write second term of the above equation as:

$$\frac{1}{v^2}\frac{\partial^2\Psi}{\partial t^2} = \frac{\partial^2\Psi}{\partial x^2}+\frac{\partial^2\Psi}{\partial y^2}+\frac{\partial^2\Psi}{\partial z^2} = \nabla^2\Psi(x,y,z)$$

3. Result & Discussion:

Case I

So, the equation (4) can be written as: $= 2.\nabla^2\Psi$

$$2.\nabla^2\Psi = 8\pi\left(-\frac{\pi}{\mathrm{h}^2}\left(\frac{5Em}{4}+\frac{V^2}{2v^2}-\frac{3Vm}{4}\right)\right)\Psi \Longrightarrow \nabla^2\Psi = 8\pi\left(-\frac{\pi m}{2\mathrm{h}^2}.\frac{1}{4}\left(5E+\frac{2V^2}{mv^2}-3V\right)\right)\Psi \qquad (5)$$

Multiplying and dividing the volume "η" of the box in right side of the equation:

$$\nabla^2\Psi = 8\pi\left(-\frac{\pi m}{2\mathrm{h}^2}.\frac{\eta}{4}\left(\frac{5E}{\eta}+\frac{V^2}{\left(\frac{1}{2}mv^2\right)\eta}-\frac{3V}{\eta}\right)\right)\Psi$$

Where, E/η is the energy density which we can denote as T_{00} using Einstein's stress-energy Tensor and $\frac{V}{\eta}$ can be manipulated as pressure term and be denoted as T_{xx} where, $x \neq 0$ [2] [3].

$$\nabla^2\Psi = 8\pi\left(-\frac{\pi m}{2\mathrm{h}^2}.\frac{\eta}{4}\left(5T_{00}+\frac{V^2}{\left(\frac{1}{2}mv^2\right)\eta}-3T_{xx}\right)\right)\Psi$$

$$\frac{\nabla^2\Psi}{\Psi} = 8\pi\left(-\frac{\pi m}{2\mathrm{h}^2}.\frac{\eta}{4}\left(5T_{00}+\frac{V^2}{\left(\frac{1}{2}mv^2\right)\eta}-3T_{xx}\right)\right) \quad (6)$$

Right side of the above equation seems similar to the structure of Einstein's Field Equation i.e. $8\pi G T_{\mu\nu}/c^4$. If dimensionally we analyse the (6) with respect to right hand side of Einstein Field equation then $G/c^4 \equiv \frac{\pi m}{2\mathrm{h}^2}.\frac{\eta}{4}$ and at the left side equation of (6) it's still a scalar field. Using tensor calculus one can apply Laplacian Operator to the scalar field in terms of tensor for all the coordinate system. If somehow, we correlate $\frac{\nabla^2\Psi}{\Psi}$ to Curvature scalar or Curvature tensor then we can define the distortion in structure of the space-time due to the sub-atomic particles in quantum levels.

Case II

Now, using equation (5),

$$\nabla^2\Psi = 8\pi\left(-\frac{\pi m}{2\mathrm{h}^2}.\frac{1}{4}\left(5E+\frac{2V^2}{mv^2}-3V\right)\right)\Psi$$

$$\Longrightarrow \nabla^2\Psi + 4\pi^2\left(\frac{\pi m}{2\mathrm{h}^2}.\frac{1}{4}\left(5E+\frac{2V^2}{mv^2}-3V\right)\right)\Psi = 0 \quad (7)$$

It's coming out to be Helmholtz Equation i.e. $\nabla^2 A + k^2 A = 0$.

Where, $k^2 = 4\pi^2\left(\frac{\pi m}{2\mathrm{h}^2}.\frac{1}{4}\left(5E+\frac{2V^2}{mv^2}-3V\right)\right)$.

4. Conclusion:

we expanded the interpretation of the physical significance of the wave function. Mostly, it is used to identify probability of the position of a particle having at a certain time nevertheless using the Schrödinger's Equation we are able to deduce the equation similar to the Einstein's Field Equation which promises to explain the certain structure of space-time at quantum level. Furthermore, applying wave equation in the modified Schrödinger Equation we are able to perceive Helmholtz Equation out of it.

5. Reference:

[1] C. Kiefer, "Conceptual Problems in Quantum Gravity and Quantum Cosmology," *ISRN Math. Phys.*, vol. 2013, pp. 1–17, Aug. 2013, doi: 10.1155/2013/509316.

[2] "Classification of the Ricci tensor and the energy-momentum tensor," Cambridge University Press, 2003, pp. 57–67.

[3] S. Carlip, "The stress-energy tensor," Oxford University Press, 2019, pp. 59–65.

International Journal of Reconfigurable and Embedded Systems (IJRES)
Vol. 10, No. 2, July 2021, pp. 130~136
ISSN: 2089-4864, DOI: 10.11591/ijres.v10.i2.pp130-136

Restoration circuits for low power reduce swing of 6T and 8T SRAM cell with improved read and write margins

Ram Murti Rawat. Vinod Kumar
Department of Computer Science and Engineering, Delhi Technological University, Delhi, India

Article Info

Article history:

Received Mar 4, 2021
Revised May 12, 2021
Accepted Jun 4, 2021

Keywords:

Dual node voltage
Low power SRAM
Read noise margin
Swing restoration
Write noise margin

ABSTRACT

This article clarifies about the variables that influence the static noise margin (SNM) of a static random-access memory. Track down the improved stability of proposed 8T SRAM cell which is superior to conventional 6T SRAM cell utilizing Swing Restored circuit with voltages Q and QB bar are peruse and Compose activity. This SRAM cell strategy on the circuit or engineering level is needed to improve read static noise margin (RSNM), write static noise margin (WSNM) and hold static noise margin (HSNM). This article relative investigation of conventional 6T, standard 8T and proposed 8T SRAM cells with improved stability and static noise margin is finished for 180 nm CMOS innovation. This paper is coordinated as follows: Introduction in area 1, The 6T SRAM cell are portrayed in segment 2. In area 3, proposed 8T SRAM cell is portrayed. In area 4, standard 8T SRAM cell. Segment 5 incorporates the simulation and results which give examination of different boundaries of 6T and 8T SRAM cells and segment 6 conclusions.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Ram Murti Rawat
Department of Computer Science and Engineering
Delhi Technological University
Shahbad Daulatpur, Bawana Road, Delhi-110042
Email: rammurtirawat@dtu.ac.in

1. INTRODUCTION

The scaling of semiconductor process technologies has been continuing for more than five decades. Advancements in process technologies are the fuel that has been moving the semiconductor industry [1]. In response to growing customer demand for enhanced performance and functionality at reduced cost, a new process technology generation has been introduced by the semiconductor industry every two to three years during the past four decades [2]. Memory is used widely in all electrical systems: mainframes, microcomputers and cellular phones, etc [3]. The growing demand of portable battery-operated systems has made energy efficient processors a necessity [4]. The performance of these devices is limited by the size, weight and lifetime of batteries. Serious reliability problems, increased design costs and battery-operated applications prompted the IC design community to look more aggressively for new approaches and methodologies that produce more power-efficient designs, which means significant reductions in power consumption for the same level of performance [5]. Memory circuits form an integral part of every system design as dynamic RAMs, static RAMs, ferroelectric RAMs, ROMs or flash memories, significantly contributing to the system level power consumption [6]. Reducing the power dissipation in memories can significantly improve the system power- efficiency, performance, reliability and overall costs [7]. RAMs have experienced a very rapid development of low-power low-voltage memory design during recent years due to an increased demand for notebooks, laptops, hand-held communication devices and IC memory cards [8]. There are various approaches that are adopted to reduce power dissipation, like design of circuits with

Proceedings of the Second International Conference on Electronics and Sustainable Communication Systems (ICESC-2021)
IEEE Xplore Part Number: CFP21V66-ART; ISBN: 978-1-6654-2867-5

Review of Dynamic Comparators for ADCs

Yogesh Sachdeva
Department of Electrical Engineering
Delhi Technological University
Delhi, India
yogeshsachdeva27@gmail.com

Nalin Nehra
Department of Electrical Engineering
Delhi Technological University
Delhi, India
nehranalin99@gmail.com

Shikhar Bansal
Department of Electrical Engineering
Delhi Technological University
Delhi, India
shikhar1551998@gmail.com

Garima
Department of Electrical Engineering
Delhi Technological University
Delhi, India
garima.mait@gmail.com

2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC) | 978-1-6654-2867-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICESC51422.2021.9532865

***Abstract*— In this paper, we review different dynamic comparator circuits that have been inspired for various use cases and functions. Comparators are crucial analog blocks and find their use in most of the analog ICs and even in many digital ICs. A lot of work has happened to improve the designs and model better ones that are optimized for specific functions, such as low power, small area, high gain, low offset, high bit resolution, etc. Many comparator designs are used in Analog to Digital converter circuits, making it an even important analog block. There are multiple design styles such as dynamic comparators, regenerative comparators, continuous time comparators, etc. All these design styles have their own advantages and certain tradeoffs.**

Keywords—Dynamic Comparator, Low-Power, Low-Offset, Resolution, ADCs, CMR, ICMR, kickback-noise.

I. INTRODUCTION

One of the most crucial and decisive elements in today's analog circuitry are comparators. After Operational amplifiers, comparators are the second most widely used analog blocks. They play a key role in the performance of various analog circuits that prove essential for a variety of functions. One major application of Comparators is in Data converters. Data converters have various important parameters which are directly or indirectly affected by the comparators present at the heart of the circuitry. These parameters include slew rate, gain, resolution, speed of conversion, frequency of operation, power consumption, along with many others. Gain is one of the crucial parameters. It affects the resolution of the converter and hence directly affects the functionality of the circuit. High gain is an important characteristic of circuits that measure very small voltage differences.

Another major application of comparators is in portable battery operated communication circuits. These circuits require low power and high speed ADCs [1]. This has been made possible by shifting to smaller technology nodes. But, this introduces process variations and it increases as the process gets more finer and feature size decreases. In ADCs, the comparator is most affected by these mismatch and process variations. Mismatch also introduces offset [1]. Complexities such as gate-induced drain leakage, random channel effects, band tunnelling across junction, random doping requirements, high channel doping requirements [2] are introduced. Also, supply voltages also have to be scaled down as feature size decreases, introducing reliability issues [3].

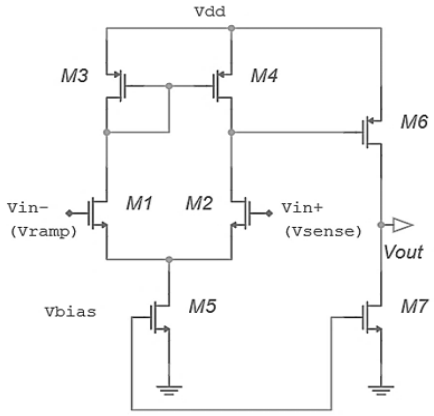


Figure-1 Conventional continuous time comparator [4-5]

Fig-1 illustrates the continuous time-domain comparator [4-5]. This conventional design is the first step towards all complex continuous-time comparator designs. It consists of two stages; the first one being a differential stage with active load followed by a pmos gain stage. This circuit closely resembles the two-stage operational amplifier. The only difference being is that this doesn't require compensation. This is a robust design, but it suffers from static power dissipation as I_{BIAS} is always from the supply rails. It also has the problem of offsets, noise, and common mode range sensitivity. Because of these problems, continuous time comparators are not very widely used for low power, low offset, and high-speed applications. Dynamic comparators are better candidates when it comes to such requirements.

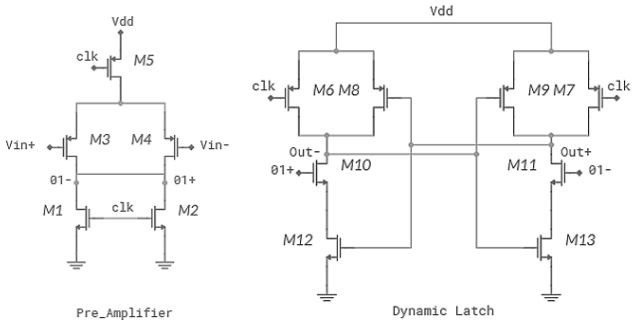


Figure-2 conventional dynamic comparator with its preamplifier and latch stages [4]

Fig-2 shows the conventiona1 2-stage dynamic comparator [4]. The first stage is the preamplifier stage, followed by a latch stage. Before the comparison operation begins, the comparator is in a reset state (Preset phase).

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

Output nodes of the 1st stage are discharged to zero potential and of the second stage to V_{DD}. Clk and Clkb goes to low and high, respectively, in the evaluation phase. The output voltages of the 1st stage (V_{O1+}; V_{O1-}) start to grow. As soon as the output voltages of first stage are high enough to turn on the transistors M10 and M11, second stage comes into picture. This positive feedback kicks in, increasing the differential output voltage till the latch reaches its final state.

II. Literature Review

A. Ultra-low energy dynamic comparator for CIS applications

Parameters such as area and power consumption become an important factor for designs that employ a large number of ADCs. One key example of such systems is CMOS Image Sensors (CIS) [6-10]. For massively parallel CISs, A/D conversion happens at the pixel level. To ensure that the power consumption and die/chip area remains within bounds, the die area and power dissipation of a single pixel should be low enough. This is necessary as both are proportional to the number of pixels. These factors are very crucial because even low-resolution CISs contain unto tens of thousands of pixels. This calls for the need to use single-slope ADC as it is simple and occupies relatively lesser area. This is an efficient use case for in-pixel implementation.

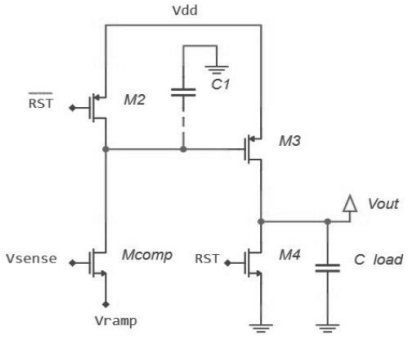


Figure-3 Two-stage design [11]

The conventional dynamic and and continuous time comparators discussed previously pose the problem of increased power dissipation, making them an unsuitable candidate. Continuous time comparators have static power dissipation. An N-bit dynamic comparators perform 2^N comparisons. Each comparison comes with energy dissipation. A power-down mechanism can help lower power consumption even further, by disabling the comparator once conversion is completed. For some designs, this reduction is still not enough.

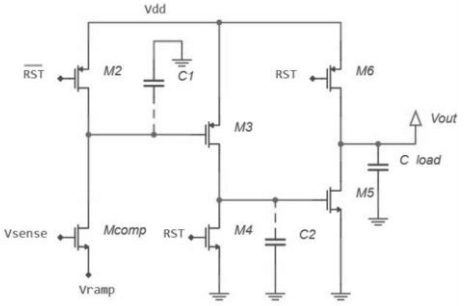


Figure-4 Three-stage design [11]

The design that we discuss here is a dynamic CMOS analog comparator [11]. This results in a power consumption that is about two to three orders lower than the two standard comparators designs. It works without a power down mechanism and the area size is smaller due to the in-pixel implementation.

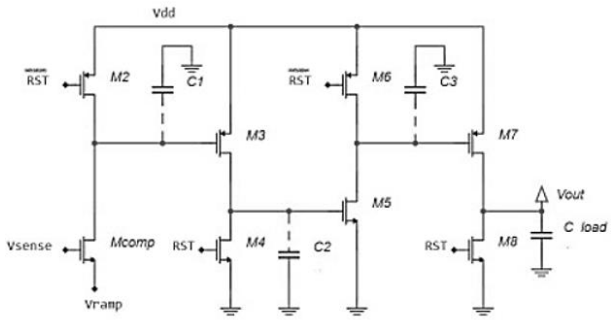


Figure-5 Four-stage design [11]

The Ultra-low voltage and its two-, three-, and four-stage designs are shown Fig 3-5. All these designs vary in power consumption per transition and the bit resolution. The transistor M_{COMP} is the device the carries out the comparison operation at the first stage. The PMOS M_2 precharges the capacitor C_1 to V_{DD}. The consecutive stages have C_1, C_2, C_3 as parasitic capacitances and C_{load} as the output load capacitance. The RST signal precharges the capacitances C_1 and C_3, and discharges C_2 and C_{load}, when it is high. Initially, V_{sense} has a value anywhere between 0-V_{DD}. Initial value of V_{ramp} is V_{DD}, and hence M_{COMP} is switched off. When V_{ramp} starts to go low from V_{DD} to 0V, M_{COMP} turns on at the following condition

$$V_{ramp}(t) \approx V_{sense} - V_{TH} \qquad (1)$$

Where V_{TH} is the threshold voltage of M_{COMP}. M_{COMP} starts to discharge C_1, and M_3 begins to charge up the output node, or output capacitance C_{load} at a high speed. Energy consumption can be given by the following simplistic expression of capacitance charging (or discharging),

$$E_{COMP} = (C_2 + C_2 + \ldots\ldots C_{load}) . V_{DD}^2 \qquad (2)$$

The comparator's state changes only once. Hence, the power consumption for the comparator is given by (2)
It can be seen that the comparator is performing voltage-to-time conversion, which results V_{sense} as a pulse of time duration T_{out}. This time T_{out} can be calculated as:

$$T_{out}(V_{sense}) = (T_{ramp}/V_{DD}) * (V_{DD} - V_{TH}) \qquad (3)$$

For $V_{TH} < V_{sense} < V_{DD}$ and $0 < V_{ramp} < V_{DD}$

The V_{TH} value may even go as down as 0V in advanced CMOS process (using depletion mode devices), but this introduces leakage, which is undesirable for this application. In pixel image photosensors voltages below 0.2-0.5V are not used, hence this range is sufficient for V_{SENSE} to be used in image sensors [12-14]

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

B. *Low-power, low-offset dynamic comparator for ADCs*

This design [15] is essentially a regenerative one, which employs voltage regeneration via a cross-coupled circuit. Conventional cross-coupled circuit designs suffer from a trade-off between speed, offset, power consumption of the comparator due to mismatch among transistors. The design discussed here is a low-offset, low-power comparator for ADCs. To keep the offset (which is due to mismatch between transistors) low, it has two operational phases.

To achieve low offset in comparators, transistors need to be sized up, which introduces increased power consumption and more area. This new design of a dynamic comparator aims to solve this tradeoff, by lowering offset, while keeping power consumption and area in bounds.

There is a trade-off between power consumption and input referred offset of the comparator, one can be increased at the expense of the other. The reason for this trade-off is that the comparison process and regeneration phase happen at the same time. As a result of using a double phase architecture, there's a cascade of amplifying stages along with a latch stage. Also, we see a lower offset voltage, without any tradeoff of power or area.

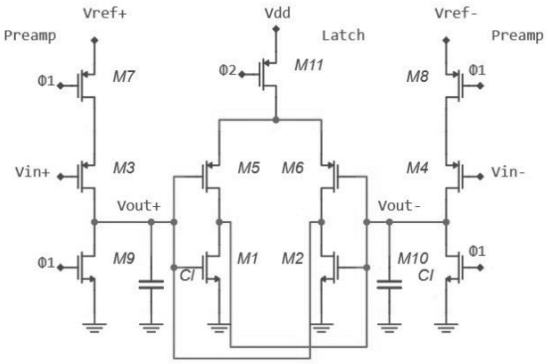


Figure-6 The low-offset design [15]

The design of the comparator is shown in Fig-6. The two latch and preamplification stages are present in the design. Clock pulses φ_1 and φ_2 signals control the two stages separately. This independent operation allows the input MOSFETS to nullify any mismatch inside the latch before this offset can affect the circuit operation. This helps reduce input referred offset. Moreover, the transistor count is lesser as differential input voltage is sensed using M_3 and M_4, which further reduces the offset from mismatch of these transistors which are no longer present in the design. The design employs a unique 3-phase signaling. The waveforms are shown in Fig-7. In the first phase both φ_1 and φ_2 are high, & hence the two cross coupled inverters are turned off. Output nodes are precharged to ground. In the second phase of signaling, amplification occurs when φ_1 is low and φ_2 is still high. Output nodes are disconnected from ground while the input cascade conduct as they are connected to V_{ref}. The voltage difference between input and reference is related by the current difference I_{IN+} - I_{IN}. In the 3rd phase, called the comparative phase, the latch is on and the regenerative loop of cross coupled inverters help boost the differential voltage.

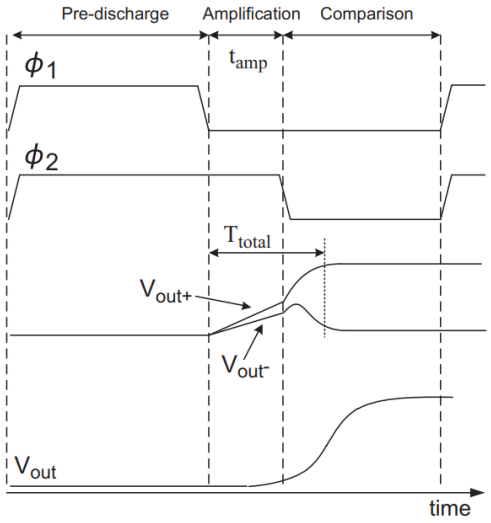


Figure-7 Conceptual Waveforms [15]

C. *High-Speed CMOS Latched Comparator With FVF Input Stage*

This comparator [16] is suitable for Successive Approximation type ADCs but can also be modified to be used with Flash type ADCs. It is a low power, high speed latched design that employs a 'Flipped voltage follower' (FVF) [17] as the first stage. This is another version of a dynamic latched comparator. This architecture tries to correct problems such as component mismatch, random noise, and increased power consumption. The advantages of using a flipped voltage follower as input stage over a class AB amplifier include enhanced ICMR, and the possibility of large voltage swings. The low output impedance of the follower enables it to source relatively higher currents. Improvement in slew rate and resolution is also observed in this design. The higher slew rate is a result of dynamic bias current setting in FVF. Higher swings lead to large resolutions. This design also sees a lower quiescent current.

The two stages of the design are: a class AB differential amplifier and a regenerative latch followed by an SR latch for output. Many variations of class AB amplifier are present in the literature [18-19]. In this circuit, transistors M_{P1}, M_{P2}, and the flipped voltage follower realise a pseudo differential class AB input stage. The FVF cell comprises of M_{P3}, M_{P4}, and the current source I_B. As mentioned before, this circuit is capable of achieving a high slew rate and resolution, in addition to a low quiescent current. Proper design via transistor sizing and careful layout helps in reducing inaccuracies like offset voltage that arise because of dissimilar transistors. As FVF is inherently a voltage follower, it sees a very low output impedance at the circuit node V_H. The impedance seen at the source of M_{P4} can be approximated by [20],

$$R_{VH} = 1/(g_{M4}\, g_{M3}\, r_{O3}) \qquad (4)$$

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

which is very low, usually in the range of 20-100ohms. The source-gate voltage of M_{P4} is constant as current through M_{P4} is I_B (which is constant if we neglect body-effect). This implies that any change in voltage at the gate terminal of M_{P4} is also reflected at the source. The shunt feedback though M_{P4} also adjusts V_{SG}, M_{P3} to ensure current requirements of the circuit are in place.

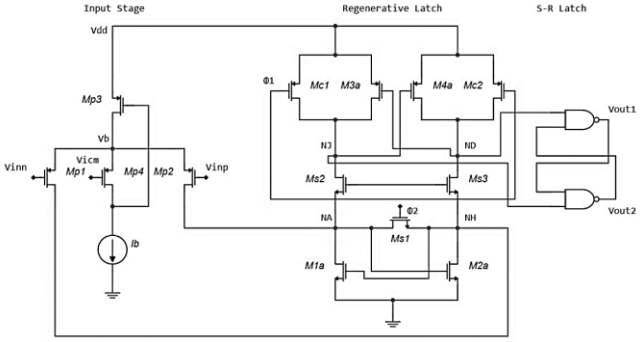


Figure-8 schematic of the dynamic comparator circuit [16]

The circuit comprises of two regenerative n-type MOS devices (M_{1A}, M_{2A}) with two n-channel transistors (M_{S2}, M_{S3}), an n-channel transistor (M_{S1}) for reset operation, two regenerative p-type devices (M_{3A}, M_{4A}), and two p-channel devices (M_{C1},M_{C2}). Clock signals f_1 and f_2 are two non-overlapping clocks.

During the positive feedback regeneration operation, f_1 goes high and f_2 goes low, and alongside that, a voltage proportional to the input voltage established among the drain terminals of M_{1A} and M_{2A}. All this while, M_{C1} and M_{C2} are disabled while M_{S2} and M_{S3} are closed. The n-channel and p-channel flip-flops uplift the voltage connecting the circuit points NA and NH, and the ones connecting NJ and ND.

D. High-Speed Low Power Comparator for ADCs

The following design [21], the varying voltage of 1st stage is reduced to attain a low-power, high-speed comparator featuring an acceptable offset. Along with improved ICMR, when compared to the conventional design, latch circuit is turned on stronger. The conventional circuits again suffer from the tradeoff between offset and power consumption. In the conventional design, the positive feedback effect of the latch amplifies the differential voltage until the latch is locked to a stable state. The voltage of one output node of latch reached Gnd, whereas the latter node settles close to VDD. Both V_{OUT} voltages of 1st stage settle close to VDD. The sizing of 1st stage devices (M_3, M_4) is done large enough so that it leads to lower offsets. This leads to large power dissipation. This tradeoff shows that the conventional circuit is unsuitable alternative for low-power large-resolution comparators and amplifies the need for better designs.

The new design is show in Fig-9. During the reset cycle, the voltages of 1st stage are half of the supply voltage, ensuring it will be high so as to not turn the 2nd stage off. M_{14} is placed to turn off the path from V_{DD} to Ground in reset phase to avoid static power dissipation.

Voltage range of the 1st stage limits to half of the supply voltage. This reduces the power dissipation of 1st stage by an order of 2, when compared to the conventional comparator. The drive currents of M_1 and M_2 are reduced because of the smaller gate-source voltage. Speed is not reduced as it only affects reset/precharge phase and has no effect on the evaluation phase. The latch speed improvement because of the large CM voltage at the 1st stage latch (>half of supply) improves the speed considerably.

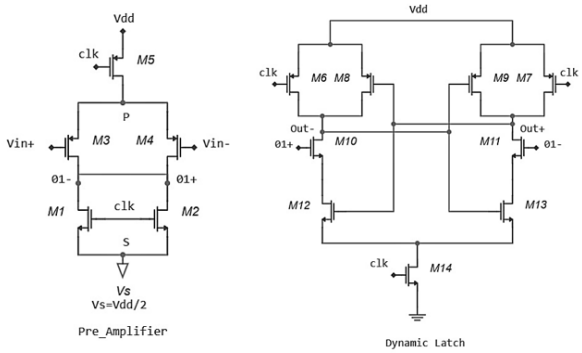


Figure-9 The new high-power low speed design [21]

The lower power consumption helps solve the offset issues of this design. To ensure that offset of circuit is within bounds, sizing of input devices of both phases (i.e., M_3, M_4, M_{10} and M_{11}) is done carefully. Fortunately, as this design greatly enhances speed and helps lower the power dissipation, the devices can be sized large so as to keep offset under limits. The comparator delay comprises of two terms,

$$T_{COMP} = T_{PRE\text{-}AMP} + T_{LATCH} \tag{5}$$

$T_{PRE\text{-}AMP}$ is the time that is taken to charge up the parasitic capacitances to the threshold voltage,

$$T_{pre-amp} = \frac{C_{01+} + C_{02-}}{I_{M5}} \times V_{thn} \tag{6}$$

T_{LATCH} is given by,

$$T_{latch} = \tau \times \ln\left(\frac{0.5(VDD - GND)}{\Delta V_{in}}\right) \tag{7}$$

Where τ is,

$$\tau = \frac{C_L}{G_{Mpmos} + G_{Mnmos}} \tag{8}$$

E. Low Kickback Fully-Differential Dynamic Comparator for Pipelined ADCs.

"Kickback noise" is a phenomenon that occurs due to voltage changes at regenerative nodes in comparator circuits. This design aims at lowering the value of this effect as an improvement on conventional circuits, along with a low-offset error. In a pipelined ADC with k-stages, the total noise is given by,

$$e_{Total}^2 = e_1^2 + \frac{e_2^2}{G_1} + \frac{e_3^2}{G_2} + \cdots + \frac{e_k^2}{\prod_{i=1}^{k-1} G_i} \tag{9}$$

Where e_i^2 represents the input stage noise power and G_i represents i^{th} stage interstage gain. Once the topology of the comparator has been checked for the 1st stage, the noise

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

components are added up for the remaining of the stages of the pipeline ADC. This error is given by,

$$V_{OS} = \pm \frac{r}{2^{B_i+r}} V_r \quad (10)$$

Where B_i is the digital output and r represents the redundancy stage. The smaller offset voltage does not enhance the general functioning of the pipelined Analog-to-Digital Converters. Dynamic comparators have precedence over other typologies, this topology does not require any extra preamplifier stage which would add to the quiescent power consumption. Also, no additional comparators are used in the dynamic comparator. The design also benefits from the low kickback noise across a large voltage range at input, while, simultaneously, the offset voltage accuracy remaining within bounds and the power being very low.

This new circuit [22] (Fig-10) proposes added transistors in the design to reduce the kickback effect by coupling the digital input and output of the design to the input analog voltage.

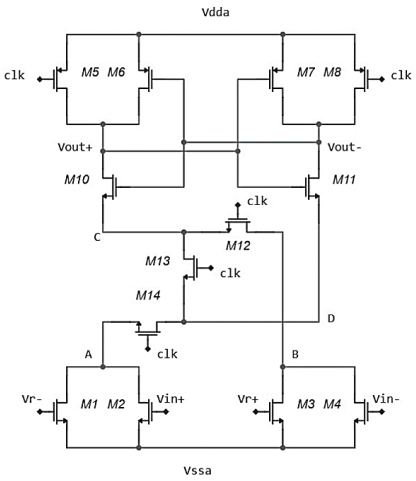


Figure-10 Low-offset low kickback noise dynamic comparator [22]

When clk goes high, the transistors turn off, disconnecting the input side from the cross-coupled latch. The transistor M13 is used for node C voltage to follow node D voltage, to help lower the offset error. Positive feedback mechanism is triggered as soon as the clk goes down, but the offset error is kept within bounds as drain of M13 and source of M13 are at the same voltage.

F. High-Speed, Low-Power Common-Mode Insensitive Comparator

This new design [23] features fast speed, low common-mode variation, low power consumption and input referred noise and offset that is within bounds when compared to conventional double tail comparators. Fig-11 presents the new design.

The latch which is working in current-mode receives the analog signal. M_{10} and M_{11} precharge the nodes D_P, D_N to V_{DD} and the cascaded inverters (M_6–M_9) begin the comparator operation after M_3 switches on with M_{10} and M_{11} switched off by clk1. clk2 is also the delayed signal of clk1 so that M_4 and M_5 are switched on later to accelerate the positive feedback effect.

The comparator features the high-speed comparison with lowest power dissipation when compared against some of the conventional double-tailed latch designs. The input-referred variation in voltage and offset of this design are within bounds, and even better than some of its conventional counterparts. This transistor is very well suited for data conversion applications and high-speed applications.

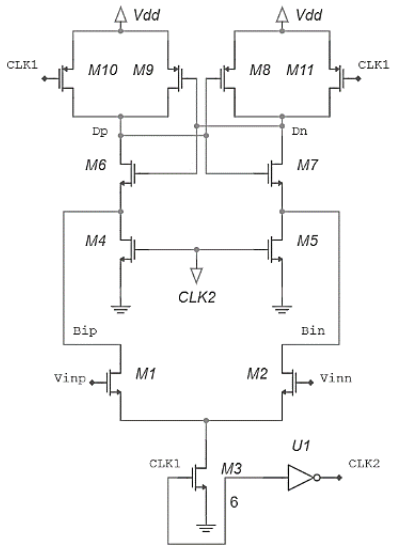


Figure-11 The common-mode insensitive design [23]

G. Double-Tail Latch-Type Voltage Sense Amplifier with 18ps Setup + Hold Time

The latch type sense amplifiers, or sense amplifier-based flip flops are considerably efficacious designs. They attain quick stable states because of regenerative feedback and low offset because of the differential input and large feature size. However, so many transistors stacked make the need for a large voltage headroom. Also, conventional designs have offsets and speed that show sensitivity towards common mode level at the input. This poses a problem for comparator designs that require wide common mode ranges.

The design discussed [24] here has less transistors stacked, enabling the possibility of a low headroom. The double tail enables both a large current in the latching stage for fast latching irrespective of V_{CM}. A small current in the 1st stage ensure low offset. Fig-12 presents the new design. During the precharge phase (c1k=0), M_7 and M_8 precharge the D_I nodes to V_{DD}, that further leads to M_{10} and M11 to discharge through ground node. In the evaluation stage (clk=V_{DD}), the tail transistors M_9 and M_{12} to turn ON. The common mode voltage at the D_i node drops at a decreasing rate defined by I_{M9}/C_{Di} and , and further, input controlled differential voltage ΔV_i develops. This voltage is passed on to cross-coupled inverter pair via M_{10} and M_{11}, and it also provides shielding between input and output, with low kickback noise. When the CM voltage falls at the D_i nodes, it is no longer adequate for M_{10} and M_{11} to clamp the output nodes to GND, and the inverter start to regenerate the voltage difference.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

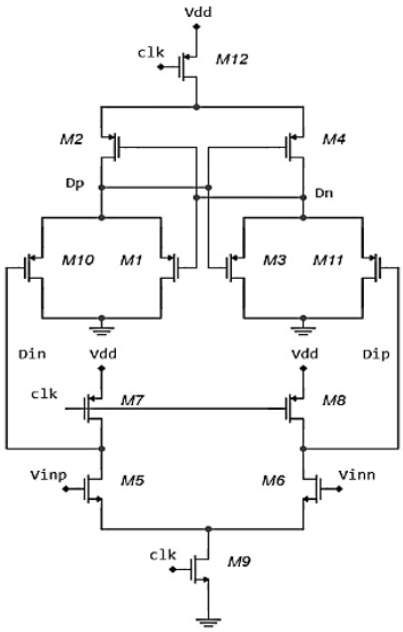


Figure-12 Double Tail Latch Sense Amplifier Design [24]

Outside an interval of 18ps, there were no bit errors (input jitter is a part of this interval), hence the required hold and setup time is less than 18ps. This setup + hold time can further be reduced by increasing the aspect ratio of the transistor M_9. However, this might come with a tradeoff of increased offset and noise as saturation operation time is decreased for transistors M_5 and M_6.

In conclusion, this double tail topology has lesser limitations when it comes to a balance among speed, power consumption, offset, common mode variation. This design isolates input and output well enough and can be operated at lower voltages.

H. Clocked Comparator for High-Speed Applications

A clocked comparator usually consists of 2-stages. The function of first input stage is to interface with the input signals. A pair of cross-coupled inverters are present in the 2nd stage and connected in the form of a latch topology. For low power designs, positive feedback operation might start its operation from supply level [25] or ground level [26]. For CMOS implementations, these stages consume very low power as wire to GND is controlled by MOS switches.

This design [27] is of a clocked, large speeds design for large frequencies low voltage range signal applications. The comparator is suitable for the applications where speed of operation is of the utmost importance and CM variation is limited, example: Checking of high frequency signals. Also, it features a low-cost technique for tracking the least possible clock period that can be applied to the circuit while ensuring proper working. M_{10} transistor is off and it ensures positive feedback stage does not operate.

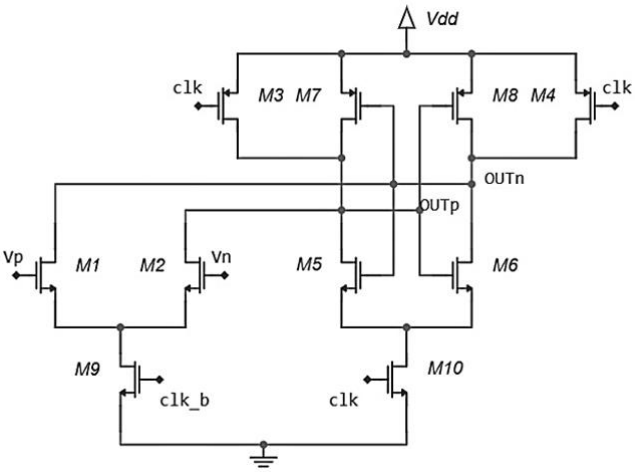


Figure-13 Clocked comparator for high-speed applications [27]

When the clock goes to high state, M_9 switches off disabling the operation of the amplification stage. On the other hand, M_{10} is switched on enabling the operation of the positive feedback stage. This part of circuit begins its working alongside V_{IN} being supplied through an amplified form of the differential voltages V_N and V_P. The charging time t_0 is given by,

$$t_0 = \frac{C_L V_{charge}}{I_{Eq}}$$

where C_L denotes load capacitance and I_{Eq} denotes resultant flow of M_3 less the current of M_{10} and V_{CHARGE} equals V_{THN}.

I. Low Noise High Speed Dynamic Analog Comparator

Multistage open loop comparators are characterized by high precision. Its speed is limited to 1GSPS due to bandwidth limitation of the amplifier. Comparatively large offset may be the only disadvantage of this preamplifier-based design. Dynamic comparators make the comparison only once every clock period and feature very less offset voltage. The main drawback is relatively large power dissipation. Attempts have been made to improve on this issue in this design [28].

With the main motives of preserving the low-offset, low-noise and high-speed, this design does not need a preamplifier stage for inputs. It shows effective behavior in terms of static and dynamic operation. The regeneration latch architecture is same as that present in a conventional design, as shown in Fig-14. No current flows in tail transistor M2 when clk is low, thereby reducing static power dissipation. When M6 is OFF the source and drain nodes of M6 are toggled depending on the input applied. When clk goes high, M6 starts to function, creating a low resistance path between those two points. Either of these nodes across M6 can be fed to a voltage buffer, producing the comparator output. The width of the clock pulse fed to this circuit is equal to half of the input pulse, controlling the operation of the comparator circuit. The transistors are sized in manner to achieve improved performance in terms of speed, delay, power, and area.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

Table-1 Comparison of various comparator designs

	Conventional Continuous Time	Conventional Dynamic	[A]	[B]	[C]	[D]	[E]	[F]	[G]	[H]	[I]	[J]
Technology (CMOS)	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm	180 nm
Offset Voltage (mV)	13	-	10.0	33.0	0.048	2.50	1.30	7.78	8.50	-	7.33	60
Power Consumption (μW)	140	190	1.6	51	63.5	420	70.12	600	113	461	50.57	40
Supply Voltage (V)	1.8	1.8	1.8	1.8	1	1.8	3	1.2	1.2	1.2	1	0.9
Output Slew Rate (V/μs)	1.25	85	2.05	-	-	-	-	-	-	-	64.11	-
Propagation Delay (ns)	-	-	<7	0.15	26	1.2	0.2	0.018	0.17	0.02	0.470	0.25

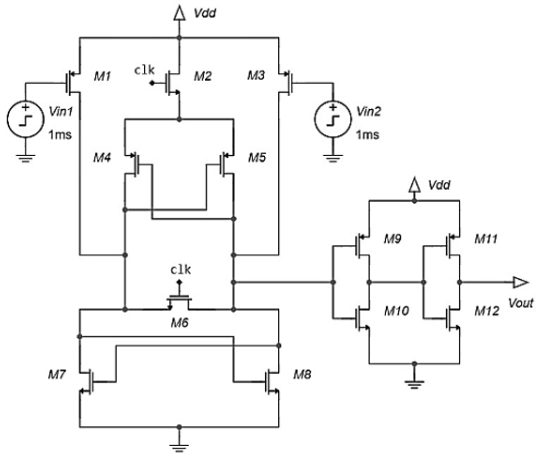


Figure-14 Low noise high speed design [30]

J. Fully Synthesizable, Rail-to-Rail Dynamic Voltage Comparator for Operation down to 0.3V

A rail-to-rail Common Mode Range (CMR) can be accomplished by noticing that using NOR gates would lead to dual circuit operation [29] (Fig-15). The comparator operates accurately only when common-mode input voltage is near ground, and it fails to operate correctly when common-mode input voltage goes above V_P (which is equal to $V_{DD}/2$). Comparator fails for common-mode voltages over V_P as OUT_1 and OUT_2 is unable to pull up properly.

The converse CMR limitations of NAND-3 and NOR-3 based Dynamic Voltage Comparator (DVC) leads to a DVC with rail-to-rail Common Mode Range can be achieved by combining a NAND-3 and a NOR-3 based first level. Also, NAND-3 logic can correctly work for the region where Common Mode Range is near Vdd, and NOR3 works for CMR near GND. This illustrates the "complementary differential pair concept". The input-stages of a NAND-3 based Dynamic Voltage Comparator and for a NOR-3 Dynamic Voltage Comparator are shown in Fig-16.

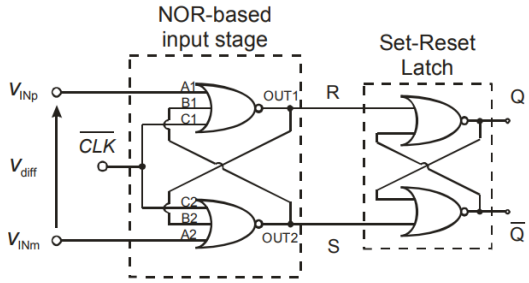


Figure-15 fully synthesizable dynamic voltage comparator with NOR based input stage and common mode range [29]

A dual-in Set Reset latch with two "set" inputs and two "reset" inputs are together passed through an OR operation, in the truth table in Fig-17. These inputs are driven from the outputs of the cross coupled NAND3 and NOR3 gates. The correct operation of just one input stage is enough to drive the SR latch accurately if the signals from other stages remain at their precharged levels.

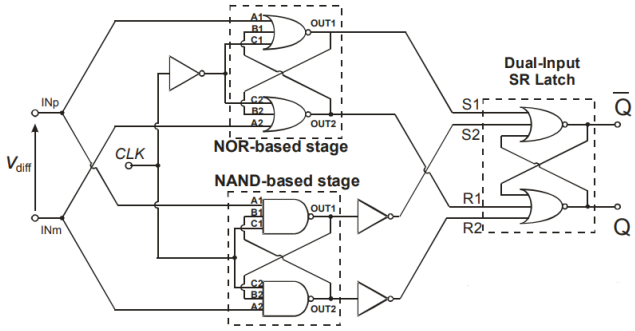


Figure-16 New Rail-to-rail DVC design with complementary NAND/NOR input stage and dual input SR latch [29]

The NAND3 gates operate correctly when common-mode input is closer to Vdd than ground potential, and generated expected outputs S1 and R1, whereas NOR3 gates do not function as their outputs remain at their precharged value (ie. ground). For the above given condition, the dual input SR-latch stores the value decided by the NAND3 gates, while being unchanged by the output

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

of NOR3 gates. This ensures correct operation in high CMR region (closer to Vdd). Opposite behavior entails when the CMR value is close to the ground than V_{DD}. In this case, NOR3 gates function properly, generating the expected SR latch S outputs. The output of the NAND3 gates stays at VDD.

S1	S2	R1	R2	Q	$\overline{Q}$	
1	-	0	0	1	0	SET
-	1	0	0	1	0	
0	0	1	-	0	1	RESET
0	0	-	1	0	1	
0	0	0	0	Memory		
Other				Not specified		

Figure-17 Truth table of Dual-Input SR latch [29]

III. Conclusion

In this paper, we presented an overview of a series of dynamic comparator designs that have been developed in the recent past. The trade-off challenges between offset voltage, power consumption, high speed operation, transistors' aspect ratio etc. all lead to very feature focused designs. All comparators tend to solve the issue of high offsets, in one way or the other. Some are suitable for low power operation, while others are suitable for low noise and high CMR operation.

While most design seek to achieve low power operation, there is varying range in the power metrics attributed to power supply and (Table 1). It is evident that there is a direct inverse relation that exists between power consumption and offset voltage. The simple reason for this is that to decrease offset, transistor sizes are increased, which leads to increased power consumption. We conclude that dynamic comparators are a lucrative choice for various ADC architectures.

IV. References

[1] B. Razavi, B. Wooley, Design techniques for high-speed, high-resolution comparators, IEEE J. Solid-State Circuits 27 (12) (1992) 1916–1926 (paper 2) 1

[2] Semiconductor Industry Association (SIA), International Technology Roadmap for Semiconductors (ITRS), 2008 2

[3] C. Fayomi, G.W. Roberts, M. Sawan, Low power/low voltage high speed CMOS differential track and latch comparator with rail-to-rail input, Proc. ISCAS 5 (2000) 653–656 3

[4] P.E. Allen, D.R. Holberg, CMOS Analog Circuits Design. Chapter 8: Comparators (Oxford University Press, USA, 2002) 4

[5] Gregorian, Roubrik, Introduction to CMOS OP-AMPs and Comparators (Wiley, 2008) 5

[6] J.B. Kim, S.K. Hong, O.K. Kwon, A low-power CMOS image sensor with area-efficient 14-bit twostep SA ADCs using pseudomultiple sampling method. IEEE Trans. Circuits Syst. II Express Briefs 62, 451–455 (2015) 6

[7] S. Okura et al., A 3.7 M-pixel 1300-fps CMOS image sensor with 5.0 G-Pixel/s high-speed readout circuit. IEEE J. Solid-State Circuits 50, 1016–1024 (2015) 7

[8] I. Takayanagi, CMOS Image Sensors, in Image Sensors and Signal Processing for Digital Still Cameras, ed. by J. Nakamura (Taylor & Francis Group, Boca Raton, 2006), pp. 143–178 8

[9] T. Toyama, K. Mishina, H. Tsuchiya, et al.: A 17.7Mpixel 120fps CMOS image sensor with 34.8Gb/s readout, in Digest of Technical Papers IEEE Int. Solid-State Circuits Conference (ISSCC) (2011), pp. 20–24 9

[10] S. Yoshihara, Y. Nitta, M. Kikuchi et al., A 1/1.8-inch 6.4 MPixel 60 frames/s CMOS image sensor with seamless mode change. IEEE J. Solid-State Circuits 41, 2998–3006 (2006) 10

[11] Jendernalik, W. An Ultra-Low-Energy Analog Comparator for A/D Converters in CMOS Image Sensors. *Circuits Syst Signal Process* 36**,** 4829–4843 (2017)

[12] W. Jendernalik, G. Blakiewicz, J. Jakusz, S. Szczepa´nski, R. Piotrowski, An analog sub-miliwatt CMOS image sensor with pixel level convolution processing. IEEE Trans. Circuits Syst. I Reg. Papers 60, 279–289 (2013) 12

[13] W. Jendernalik, J. Jakusz, G. Blakiewicz, R. Piotrowski, S. Szczepa ´nski, CMOS realisation of analogue processor for early vision processing. Bull. Pol. Acad. Sci. Tech. Sci. 59, 141–147 (2011) 13

[14] M. Kłosowski,W. Jendernalik, J. Jakusz, G. Blakiewicz, S. Szczepa ´nski, A CMOS pixel with embedded ADC, digital CDS and gain correction capability for massively parallel imaging array. IEEE Trans. Circuits Syst. I Reg. Papers 64, 38–49 (2017) 14

[15] M. Hassanpourghadi, M. Zamani, M. Sharifkhani, A low-power low-offset dynamic comparator for analog to digital converters, Microelectronics Journal, Volume 45, Issue 2, 2014, Pages 256-262,

[16] H.J. Achigui, C. Fayomi, D. Massicotte, M. Boukadoum, Low-voltage, high-speed CMOS analog latched voltage comparator using the "flipped voltage follower" as input stage, Microelectronics Journal, Volume 42, Issue 5, 2011, Pages 785-789

[17] R. Gonzalez, J. Ramirez-Angulo, A.J. Lopez-Martin, A. Torralba, J.A.G. Galan, A. Carlosena, F.M. Chavero, The flipped voltage follower: a useful cell for low-voltage low-power circuit design, IEEE Trans. Circuits Syst. I 52 (7) (2005) 1276–1291. 19

[18] C. Fayomi, G.W. Roberts, M. Sawan, Low power/low voltage high speed CMOS differential track and latch comparator with rail-to-rail input, Proc. ISCAS 5 (2000) 653–656. 20

[19] H.-C. Hong, G.-M. Lee, A 65-fJ/conversion-step 0.9-V 200-kS/s rail-to-rail 8-bit successive approximation ADC, IEEE J. Solid State Circuits 42 (10) (2007) 2161–2168. 21

[20] G.M. Yin, F. Op't Eynde, W. Sansen, A high-speed CMOS comparator with 8-b resolution, IEEE J. Solid State Circuits 27 (2) (1992) 208–211. 22

[21] A. Khorami, M. Sharifkhani, High-speed low-power comparator for analog to digital converters, AEU - International Journal of Electronics and Communications, Volume 70, Issue 7, 2016, Pages 886-894,

[22] Diaz-Madrid, J-A, Domenech-Asensi, G, Hauer, J, Mateu, L. A low kickback fully differential dynamic comparator for pipeline analog-to-digital converters. *Engineering Reports*. 2019; 1:e12055.

[23] Gao, J., Li, G. and Li, Q. (2015), High-speed low-power common-mode insensitive dynamic comparator. Electron. Lett., 51: 134-136

[24] D. Schinkel, E. Mensink, E. Klumperink, E. van Tuijl and B. Nauta, "A Double-Tail Latch-Type Voltage Sense Amplifier with 18ps Setup+Hold Time," 2007 IEEE International Solid-State Circuits Conference. Digest of Technical Papers, 2007, pp. 314-605

[25] B.Goll, et. al., "A 0.12m CMOS comparator requiring 0.5V at 600MHz and 1.5V at 6GHz ," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, San Francisco, CA, Feb. 11-15, 2007, pp. 316-317. 27

[26] D. Schinkel, et. al., "A Double-tail latch-type voltage sense amplifier with 18ps setup+hold time," in IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers, San Francisco, CA, Feb. 11-15, 2007, pp. 314-315 28

[27] M. Abbas, Y. Furukawa, S. Komatsu, J. Y. Takahiro and K. Asada, "Clocked comparator for high-speed applications in 65nm technology," 2010 IEEE Asian Solid-State Circuits Conference, 2010, pp. 1-4

[28] A. Majumder, M. Das, B. Nath, A. J. Mondal and B. K. Bhattacharyya, "Design of low noise high speed novel dynamic Analog Comparator in 65nm technology," *2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2016, pp. 115-120

[29] O. Aiello, P. Crovetti and M. Alioto, "Fully Synthesizable, Rail-to-Rail Dynamic Voltage Comparator for Operation down to 0.3 V," *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1-5

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:35:05 UTC from IEEE Xplore. Restrictions apply.

International Journal of Peptide Research and Therapeutics
https://doi.org/10.1007/s10989-021-10289-7

Role of Different Peptides for Cancer Immunotherapy

Ritika Luthra[2] · Shreeja Datta[2] · Arpita Roy[1]

Received: 4 June 2021 / Accepted: 13 September 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Due to the enormous benefits derived from host immune response, immunotherapy has been regarded among the most effective methods for defending against cancers. Nevertheless, developing robust solutions with strong biosecurity as well as the potential to modulate immunogenicity in an effective manner, continues to be a challenge. Peptides' structural similarities to protein molecules endow them with tremendous ability to solve such issues by either specifically inducing immune responses or enhancing therapeutic effect. Peptide-based substances have a lot of potential as immunotherapeutic agents for treating a variety of cancers. The development of cancer vaccines as well as delivery mechanisms based on peptides by imitating the biological functions of protein having extremely specialized immuno-regulatory roles has been the subject of significant research. Thus, this review gives a systemic comprehension of cancer immunotherapy, role of the immune system, and how peptides pose a huge potential in cancer immunotherapeutic. The role of peptides as vaccine antigens and adjuvants, checkpoint blockades, in targeted delivery and in combinatorial therapies has been addressed. Recent research on the same has been summarised, which demonstrates their enormous advantages, thus allowing their clinical application for cancer immunotherapy in near future.

Keywords Peptides · Cancer immunotherapy · Vaccine · Peptide hydrogels · Drug delivery

Abbreviations

PD-1	Programmed death-1.
CTLAT-4	Cytotoxic T-lymphocyte–associated antigen 4.
LAG-3	Lymphocyte-activation gene 3.
TIM-3	T cell immunoglobulin and mucin domain-containing protein 3.
TIGIT	T-cell immunoglobulin and ITIM domain.
CycMSH	Cyclized alpha-melanocyte-stimulating hormone peptide.
Tc-99m	Technetium-99 m.
ccRCC	Clear cell renal carcinoma.
DC	Dendritic cell.
TAM	Tumour- associated macrophages.
TGF-β	Transforming growth factor-β.
IL	Interleukin.
IFN-γ	Interferon gamma.
APC	Antigen presenting cell.
MHC	Major histocompatibility complex.
TLR	Toll-like receptors.
CTL	Cytotoxic T lymphocytes.
NK	Natural killer cell.
KIR2DS1	Killer cell immunoglobulin-like receptor 2DS1.
TIMP1	Tissue inhibitor matrix metalloproteinase 1.
MART-1	Melanoma-associated antigen recognized by T cells.
HGP100	Peptide containing amino acids 25 to 33 fragment of human melanoma antigen gp100.
TRP	Transient receptor potential.
TNF	Tumour necrosis factor.
MUC1	Mucin 1.
WES	Whole exome sequencing.
EPT	Epitope peptides.
HLA	Human leukocyte antigen.
IMP	Immunizing peptide.
ASP	Assay peptide.
AMP	Antimicrobial peptide.
ROS	Reactive oxygen species.
VP	Venom peptide.
PLA-2	Phospholipase A2.
MRD	Melittin-RADA32-DOX.

✉ Arpita Roy
arbt2014@gmail.com

[1] Department of Biotechnology, School of Engineering & Technology, Sharda University, Greater Noida, India

[2] Department of Biotechnology, Delhi Technological University, Delhi, India

Published online: 23 September 2021

DOX	Doxorubicin.
KN93	KN93 phosphate (It is a CaM kinase II inhibitor).
CAMKII	Calcium/calmodulin-dependent protein kinase II.
MRK	MR52-KN93-hydrogel.
CDN	Cyclic dinucleotide.
CDP	Cytidine diphosphate.
NP	Nanoparticle.
PDT	Photodynamic therapy.
HSA	Hepatocyte specific antigen.
GSH	Glutathione.
DMA	Dimethyl maleic anhydride.
EPR	Enhanced permeability and retention.

Introduction

Cancer is among the typical life-threatening illnesses, and despite advances in traditional approaches such as surgery, chemotherapy, as well as radiation treatments, adequate care remains a challenge (Littman et al. 2015; Del Paggi et al. 2018). Immunotherapy has gained a lot of acceptance as a treatment option in recent times owing to its excellent therapeutic potential and reduced adverse effects from off-targets of medication (Papanikolaou and Bosio 2021). Immunotherapy, rather than targeting cancerous cells specifically, induces the host's normal immune function, which destroys these malignant cells. Since the development of first cancer immunotherapy, a number of cancer vaccines have been established and used (Sahin et al. 2018). It has lately made significant progress due to newer techniques that activate immune response by influencing various immune cells. They affect innate or adaptive immune systems (Netea et al.2020). Cancer vaccines, checkpoint blockades, immune adjuvants, modified T lymphocytes as well as cytokines, along with many other developing groups, may be categorised based on the process of activation of immune cells or whole system (Park et al.2016; Byun et al.2017; Ribas et al.2018). The peptides, which are composed of analogues of protein molecules (or amino acids), are being widely used in pharmaceutical as well as biomaterial production for tissue repair and drug delivery (Zhang et al. 2017). Peptide sequences may have a proclivity to interact with infectious agents or cancerous cells and can be employed for peptide-based therapies. This is associated with their significant capabilities arising from protein molecules or imitating characteristics of the cellular proteins.

Vaccines made up of constructed peptides mimicking cytotoxic T-cell epitopes, is an example of simple and cost-effective melanoma treatment options (Zhang et al.2019). Peptide short chain therapies have some significant benefits over larger protein antibodies in terms of delivery as well as tumour aggregation. Peptides with shorter chains are easier to synthesise and could have a better chance of penetrating dense malignant tissues. Short peptide structures may also be fine-tuned to avoid allergic reactions or adverse effects caused by medication. Adjusting noncovalent interactions of peptide immunotherapies also enables the development of self-assembly of peptides into nano-medicines that typically have predictable pharmacokinetic profile (Abbas et al. 2017). Peptide assemblies that are sensitive to the microenvironment of cancerous cells, especially biomarkers are excellent anchors for cancer diagnostics (Miao et al.2018; Li et al.2019; Yin et al.2019) as well as targeted delivery (Zhang et al.2016), and thus are being widely used in treating cancers (Liu et al. 2015). Incorporating such benefits along with the membrane-spanning abilities, peptides have a lot of scope in cancer immunotherapy as therapeutic agents or scaffolds for delivering drugs (Riley et al.2019). Peptide nucleotide sequences generated via natural domains of proteins have the capability of interacting with receptors found in the immune system (innate or adaptive) and also on cancerous cells, allowing them to attenuate the human immune response, therefore serving as powerful vaccines (Pompano et al.2014). Mostly these peptide-based materials are made for targeting specific cells and can be employed for incorporating particular functions such as cell targeting, reactive cleavage sites, endocytic transportation, and therapeutic activities (Zhang et al.2015; Rodriguez-Cabello et al.2016). In the literature various reports are available on use of natural and synthetic peptide in cancer immunotherapy. Most of the reports have used one or two therapeutic peptides for treatment of cancer. This review provides a comprehensive overview on flexibility and variety of peptides and materials made from them possessing different biological activities along with their potential role in cancer immunotherapy.

Cancer Immunotherapy

Cancer immunotherapy is a major advancement for treating cancer, which is accompanied by radiotherapy, surgery and cytotoxic therapy (Lipson et al.2015). Therapies involved in cancer usually lead to tumour cell apoptosis as well as antigen release- presented via the dendritic cells of lymph nodes (tumour-draining). In the treatment of cancer, two of the key principles are there in cancer immunotherapy (Sanmamed and Chen 2018). First, is the checkpoint inhibitor's engagement of T lymphocytes having intrinsic capacity for adapting as well as memorising? Here, the mechanism is based on its durable reactions and observed prolonged survival. Second is how immunotherapy can treat immune cells by using either histology of tumours or associated driver mutation. Third is the reactions of checkpoint inhibitors are different as compared to that of chemotherapy as well

as targeted factors. Finally, efficiency of immunotherapy is enhanced when combined with existing strategies for treatment (Ascierto et al. 2014).

Since cancer immunotherapy treats or medicates immune cells, it is effectively non- adherent with respect to the histology of tumours or associated mutations. One of the studies showed that in the case of melanoma, ipilimumab (cancer immunotherapy agent) had a continuing impact for mucosal, cutaneous as well as ocular melanoma- all of them having distinct biology (Maio et al.2013). Ascierto et al.(2021) also concluded that no difference was found in the middle of BRAF- as well as NRAS-mutated melanomas with ipilimumab. As a result, it is plausible to attain prolonged life in some of the patients with clear-cut tumour classes, wherein various response estimates in tumours contemplate individual immunogenicity. Based on this mechanism for activation of the immune system, different curative agents targeting their respective immune responses can be grouped under cancer vaccines, engineered T-cells, immune adjuvants, checkpoint blockades and cytokines as emerging categories. At present, checkpoint blockade immunotherapy (Byun et al.2017; Ribas et al. 2018) and immunotherapy using engineered T-cells (Park et al.2016) are the most promising strategies exploited in the clinical trials.

Peptides Used in Cancer Immunotherapy

Peptides are usually defined structurally as shorter chains of amino acids (linear) which are < 50 amino acids in size and generally attain stability due to disulphide bonds (Hayashi et al.2012). They can be formed via conventional methods with higher specificity for binding as well as modulating protein interactions required. A peptide sequence is moderated easily because of their ease of synthesis via molecular biological techniques or chemical means (Marqus et al.2017). These peptides, also known as therapeutic peptides comprise many advantages over antibodies, due to their smaller size, easier synthesis and ability for penetrating the cell membrane. They even comprise higher affinity and specificity; minimum drug interactions; as well as biological and chemical diversities. One of the added advantages is that they do not assemble in particular organs (for instance, liver or even kidney), which helps in reducing their toxic aftereffects, if any. Infact, their role in cancer includes carrying radionuclides, cytotoxic drugs, hormones and vaccines (Yadav et al. 2020). Application of various peptides in cancer immunotherapy is shown in Fig. 1.

These are used in the therapeutics of different types of cancers like lung, colorectal, breast, pancreatic, prostate and gastric cancers from early diagnosis (Xiao et al.2015). In addition to this, they are also used for the treatment of osteosarcoma, skin and renal cancers and so on (Table 1). Besides, Wu et al.(2014) highlighted that properdistatin (peptide composed from properdin which is a plasma protein) inhibits angiogenesis in the human melanoma models (A-07). Liu et al. (2014) observed how CycMSH peptide conjoined by Tc-99 m displayed capability of targeting melanomas (particularly, metastatic ones), suggesting the efficiency to detect the metastatic melanoma using CycMSH. Gonz´alez et al.(2014) conducted a study, which showed that peptides obtained from melanocortin-1 receptor indices cytotoxic T- cell responses in order to destroy melanoma cells.

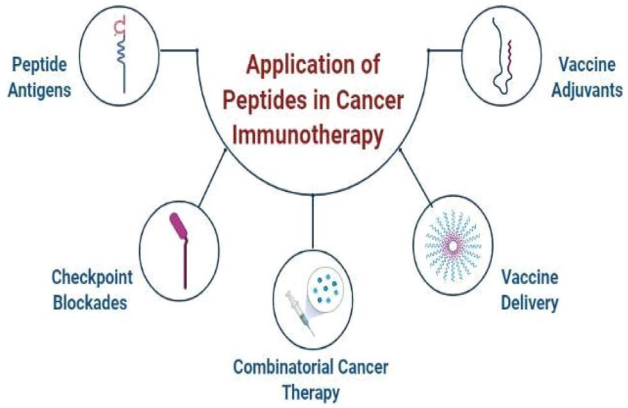


Fig. 1 Various applications of peptides that are utilized for cancer immunotherapy

Even though they have been successful in activation of immune cells, according to some preclinical trials, peptide therapeutics with respect to the immune system, still show low immunogenicity in clinical studies or clinical phase trials (Table 2). Reducing the sequence size of these proteins to partial epitopes notably decreases the peptide affinity to their respective targeting receptors, along with even enhancing the probability of enzymatic degradation of shorter peptides, as a result lowering the circulation lifetime as well as their aggregation around the lymph nodes. Besides this, heterogeneity of antigen presenting cells of patients poses a challenge in the case of efficient immunogenicity of peptide therapeutics. They even comprise lesser or no resistance at all, in-vivo, to cleavage via serum proteases, shorter half-life, lower bioavailability, as well as production and manufacturing challenges (Ali et al.2013). Thus, these considerations prove that further research of peptide functions for cancer immunotherapy is very crucial.

Table 1 List of different peptides (with their role) that are used as therapeutics in cancer immunotherapy

Name of peptide	Role/pathway affected	Ref.
AUNP-12	PD-1 Blockades	Chen et al. 2019
TPP-1	PD-L1 Blockades	Li et al. 2018
MAGE-A3	Peptide antigens activating CD8 + T-cells	Neek et al. 2018
OFA 2	Peptide antigens activating CD4 + and CD8 + T-cells	Tsoras et al. 2018
TRP2	Peptide antigens activating CD8 + T-cells	Wakabayashi et al. 2018
Peptide-57 and Peptide-71	PD-L1 Blockades	Magiera-Mularz et al. 2017
PDLong1	PD-L1 Blockades	Munir et al. 2016
$OVA_{323-339}$	Peptide antigens activating CD8 + T-cells	Qian et al. 2016
DPPA-1 and DPPA-2	PD-L1 Blockades	Chang et al. 2015
HGP100	Peptide antigens activating CD8 + T-cells	Guo et al. 2015
$E7_{49-57}$	Peptide antigens activating CD8 + T-cells	Liu et al. 2015
HAC-I and HAC-V	PD-L1 Blockades	Maute et al. 2015
$E7_{43-57}$	Peptide antigens activating CD4 + and CD8 + T-cells	Liu et al. 2013
$OVA_{257-264}$	Peptide antigens activating CD8 + T-cells	Molino et al. 2013
$OVA_{253-266}$	Peptide antigens activating CD8 + T-cells	Black et al. 2012

Role of Immune System

An individual needs a robust immune system, which can effectively detect and remove an elusive niche of cancers in order to attain successful results of immunotherapy. Innate as well as adaptive immune systems are two main immune systems of the body, in which antigen specificity serves as the primary difference among them. Innate immune system provides a non-specific defensive response in the presence of antigens or foreign particles without any delay. Adaptive immune system, also called acquired immunity, on the other hand, is much more complex owing to antigen-specific immune responses, which necessitate antigen screening and identification. When an antigen is identified, the body's adaptive immune response causes the immune cells to target it, and this reaction is long lasting since memory T-cells are long-lived and extremely selective in nature (Zhang et al.2019).

Innate immune system

The innate system consists of a broad range of cells, primarily macrophages, DCs, neutrophils, as well as natural killer cells, which serve an important purpose in immunity against cancers, altogether.

Tumour-associated macrophages (TAM) are a form of phagocytic cells, which, because of the potential to migrate outside of the vascular system, could destroy cancerous cells with fewer restrictions. They may encourage growth and migration of cancers throughout the tumour micro - environment by releasing cytokines such as chemokines, inflammatory markers, as well as cell proliferation agents (Gao et al.2019). Therefore, as a result, TAMs have two distinct morphologic activation phases: the anti-tumoral M1 and pro-tumoral M2. Clinical trials of several TAM targeting peptide-based materials have been reported (David et al. 2017. Qian et al. (2017) created an immunotherapy targeting two molecules, which consisted of an M2-TAM peptide (M2 Peptide: YEQDPWGVKWWY) with a α-peptide (SR-B1). This formulation directly inhibited the signalling pathway of M2-TAM in skin cancers, resulting in substantial M2-TAM loss inside the cancer.

Dendritic cells(DCs), which are amongst the highly crucial antigen-presenting cell types (APCs), are needed for the activation of both CD4 + as well as CD8 + T cells, which then cause specific cytotoxic effects in cancerous cells. MHCs (major histocompatibility complex), CD86, CD40, CD80 as well as TLRs (toll-like receptors) are all essential for introducing peptides to B and T lymphocytes, and DCs might be involved in their expression (Zhang et al.2019). Peptides mounted onto these cells cause antitumor immune response, laying the groundwork for the evolution of peptide-based cancer immunotherapy. A fused peptide for delivering to DCs has been discovered, which when combined with ovarian epitopes resulted in improving the Clec9a+'s capability to trigger CD8 + T-lymphocytes as well as reduce metastasis in lungs (Yan et al.2016). To cause immunity towards breast as well as liver cancers, a study conducted by Choi et al. (2018) used peptides obtained from stem-like cancerous cells in the form of antigens. They allowed effective stimulation for developing mature DCs, improved activation of T cells and boosted the volume of CTLs via circulating these cells using CD44 and EpCAM related peptides.

Neutrophils are key components in innate immunity and can be used for treating different cancers. They communicate

with macrophages, dendritic and natural-killer cells, mesenchymal stem cells as well as both lymphocytes for removing cancerous cells after they have been incorporated into these tissues. They have antitumor properties, produce cytotoxic effects, and are involved in releasing cytokines and chemokines for regulating various immune responses (Patel et al. 2018).

Because of the efficiency to identify stressful cells without requiring any antibodies or MHCs, natural killer (NK) cells react quickly to tumorigenesis. As several cancerous cells do not contain MHC-I markers, all immune cells are unable to identify and kill them, thereby making this characteristic of NK cells critical in cancer immunotherapeutic (Das et al. 2015; Guillerey et al. 2016). An HLA-C*06:02-exhibited peptide was involved in binding to NK inducing receptor KIR2DS1 as well as triggering primary KIR2DS1(+) NK cells, according to Chapel et al. (2017).

Adaptive Immune System

T-lymphocytes (which include CD4 + helper as well as CD8 + cytotoxic T-cells) and B-lymphocytes make up the adaptive immune system. To enable innate cells (including natural killer cells), B- lymphocytes as well as CD8 + killer T- lymphocytes to recognise and destroy tumours, CD4 + helper T-cells generate cytokines. Several existing cancer immunotherapeutics depend upon the amplitude of cytotoxic CD8 + T-lymphocytes to invade tumours as well as kill cancerous cells. These include checkpoint blockade antibodies and even adaptive T-cell transfer (Netea et al. 2020).

One of the latest treatment strategies to aid in improving capability of the immune system to identify as well as destroy secret cancerous cells is checkpoint blockade immunotherapy (Postow et al. 2015; Robert et al. 2015; Desrichard et al. 2016). Immune checkpoints are adverse controllers of immune response that contribute in maintaining homeostasis and inhibit autoimmune systems from targeting cells randomly (Buchbinder et al. 2016). Immune checkpoint systems can, nevertheless, be engaged in cancerous cells to block emerging antitumor immune responses, allowing them to evade and flourish (Sharma and Allison 2015; Gubin et al. 2014). Suppression of immunological checkpoints enables in restricting the cancer cells' immune escape as well as activation of body's immune cells like cytotoxic T-cells to safeguard from cancerous cells. This strategy has been hailed to be an attractive and promising technique in terms of cancer immunotherapy (Li et al. 2019). Principally, PD-1, CTLA-4, T-cell immunoglobulin 3 (TIM-3), lymphocyte-activation gene 3 (LAG-3) as well as T-cell immunoglobulin and ITIM domain (TIGIT) are the key inhibiting receptors secreted by active T-cells (Joller et al. 2014; Li et al. 2019). As a result, such receptors that serve as immune-checkpoint targets enable the generation of blockers for triggering immunological responses, such as PD-1, CTLA-4, etc., which are widely employed in present preclinical and clinical studies. CTLA-4 appears to block T-cell immunological responses via signalling or non-signalling channels, according to recent findings. CTLA-4 stimulates phosphatases, which dephosphorylates the T-cell receptor (TCR) signal, as per the signalling channels. CTLA-4. The non-signalling route suggests how CTLA-4 may utilise the trans-endocytosis mechanism for collecting and eliminating CD-80 and CD-86 proteins off the membranes of APCs, reducing CD28 expression (Li et al. 2019). APCs as well as the tumour cells produce PD-1 ligands, particularly PD-L1 (B7-H1 or CD 274) and PD-L2 (B7-DC or CD 273) (Li et al. 2019). When PD-1 binds to its ligand, T-cell kinases are inactivated, while TCR signals are dephosphorylated, lowering inflammatory-cytokine generation and controlling T-cell functioning (Gordan et al. 2017). Trying to block these PD-1/PD-L1 pathways preserves cancer-specific T-cell functioning and enables the immune cells to re-recognise and destroy tumour cells, avoiding immune escape by these cells. Previously, DPPA-1, DPPA-2 and HAC-I, HAC-V peptides have been utilised for blocking the PD-L1 pathway, thereby preventing the escape of tumor cells from the immune system (Chang et al. 2015; Maute et al. 2015). Li et al. (2018) reported a PD-L1 directed peptide showing great binding specificity. The tendency of same to prevent association of PD-1 with PD-L1 was also confirmed in in-vitro studies. This peptide was found to be more effective in comparison to PD-L1 antibody and could reduce tumour growth in the mice model, implying that it is as effective as therapeutics based on antibodies. Similarly, Magiera-Mularz et al.(2017) utilised macrocyclic-peptides (peptide-57 and peptide-71) as inhibitors for PD-1/PD-L1 pathway and they do so by directly binding with PD-L1, thereby restoring the functioning of T-cells. Li et al. (2018) and Munir et al. (2016) studied the effectiveness of PDLong1 and TPP-1 peptides as checkpoint blockades for PD-L1 pathways. Recently, AUNP-12, a precursor of CA-170, is a small molecule, which has also been reported to act as PD-L1 inhibitor (Chen et al. 2019).

B-lymphocytes are also crucial in vaccinations because they produce antibodies, which are involved in recognising particular antigens and also inactivating or labelling them for elimination. Almost 90% of cell types go through apoptosis when a tumour is removed, whereas a tiny proportion of specialized memory T-cells are preserved in order to bestow a quick reaction in case the same cancer cell reappears in the body (Cha et al. 2014). Also known as immunologic memory, it is the defining characteristic of adaptive systems. Cytotoxic and helper T-cells, as well as T-regulatory (T_{reg}) cells are the 3 different kinds of mature T-lymphocytes. Since CD8 + cytotoxic cells can recognise and even destroy tumour cells, which display antigens specific to them, they are essential for eliminating these cells.

Production of CD4 + regulatory cells, that limits the action of effector T cells, is linked towards the induction of the cytotoxic CD8 + cells (Flippe et al. 2019).

Directly delivering CTL-peptide activators to the tumour has a significant impact on tumorigenesis. Cheng et al. (2018) created a curative peptide that combines targeting of tumours as well as the on-demand secretion of both DPPA-1, a protein inhibitor of apoptosis-ligand 1, as well as NLG919, which inhibits indole-amine 2, 3-dioxygenase, into a single peptide. Apart from this, cancer vaccines constructed using B-cell peptide usually consist of an adjuvant along with an immunogenic protein having a B-cell epitope peptide which can cause these cells to produce antibodies. Such vaccines generate polyclonal formulations of antibodies, which identify a number of antigens, as opposed to monoclonal antibodies containing unique target areas. One more vaccine strategy would be using B-cell peptide imitates which can act synergistically to cellular receptors particular to tumours, preventing polymerization of the receptor. Besides that, antibodies as well as their peptide imitates exhibit anti-tumour effects, which is based on their ability to attach to tumour receptors and inhibit downstream signalling (Kaumaya et al. 2015).

Peptide Based Cancer Vaccines

Cancer vaccines provide a lot of promising results in immunotherapy because of antigen based immune reaction as well as long- lasting memory in terms of immunity (Melero et al. 2014). Because of versatility, ease of processing, and reliability, these vaccines contribute to being undoubtedly the most effective and realistic vaccination strategy (Malonis et al. 2019). Owing to its specific structure, small sequence peptides containing limited antigenic determinants for associating to the target receptor are established as an alternate solution (Skwarczynski et al. 2014). Antigens must have a strong potency for TLR or MHC (major histocompatibility complex) receptors to promote development of CD4 + or CD8 + T cells to further cause immune response. Since these smaller peptide antigens have restricted immunogenicity, using peptide adjuvants to aid the triggering of immune reaction of these antigens is a popular technique, as well as perhaps a number of peptide adjuvants are already produced (Melero et al. 2014). Designed peptide antigens correlated with MHC class I receptor for CD8 + T cell activation generally contain 8–10aa, while antigens armed by MHC class II receptor for CD4 + cells usually contain 13–18aa, although there is no hard rule on size of peptide sequence. This depends on the activating pathway (Melero et al. 2014). Immune resistance to peptide antigens is already identified in experimental studies based on differences of MHC receptors in patient populations. They are promising in cancer immunotherapy when targeted delivery is combined with the help of peptide adjuvants in immune responses (Zhu et al.2017). Skin cancer immunotherapy has made extensive use of peptide-based vaccines (Zhang et al.2019). A majority of skin cancer peptide antigens are obtained via proteins that cause melanocyte differentiation which include tyrosinase, MART-1 (Melan-A), and glycoprotein100 (Zhang et al.2019). These are primarily involved in promoting cytotoxic T lymphocyte formation for the immune system. Tyrosinase, for example, a rate-limiting enzyme in melanogenesis, comprises 2 peptides acting as immunogens, namely tyrosinase$_{1-9}$ and tyrosinase$_{368-376}$ (Zhang et al.2019). The 368–376 region of tyrosinase was found to have exceptional immunogenicity when asparagine residue was replaced with aspartic acid at the third position. Antigens derived from the MART-126–35 domain (Jäger et al.2002) were used in medical studies for treating people suffering from skin cancer. The response of the immune system towards cancerous cells was strengthened by coupling vaccine-restricted domains glycoprotein100$_{209-217}$ with immune activating agent IL-2 or interleukin-2. Guo et al. developed a nano-vaccine composition after combining poly (D, L-lactide-co-glycolide) nanostructures complexed with both antigen HGP100 as well as an adjuvant monophosphoryl lipid and mannose-coated liposome (Guo et al.2015). For producing vaccines, Shakuj et al.linked this antigen peptide HGP100 to immunogenic spherical nucleic acid (Skakuj et al.2018). Wakabayashi and his colleagues used a solid-oil nano-dispersion as carriers to deliver antigen TRP-2 peptide, which was altered by 3 lysine residues (KKKGSVYDFFVWL) along with adjuvant Resiquimod(R-848). In a tumour containing mice model, this method showed potential in suppressing skin cancer progression as well as combating lung metastasis (Wakabayashi et al.2018). HGP100 and TRP2 antigens have underwent co-encapsulation to form hollow nano-spheres of silica, that proved to be effective in stimulation and development of dendritic cells, as well as in secretion of IFN-γ or interferon gamma, TNF- or tumour necrosis factor, IL-12 or interleukin 12, and IL-4 for immune promotion (Xie et al.2017). Chemical glycosylation of MUC1 (Mucin 1) antigenic determinants can also be used for designing peptide antigens similar to MUC1. Numerous vaccines have been crafted as well as produced by coupling sequence (HGVTSAPDTRPAPGSTAPPA) to assembly domain Q11 using threonine residues that have undergone glycosylation at 9th or 16th position. Such medications incorporating B-cell epitopes exhibited strong responses from cytotoxic T-cells when stimulated by Helper T-cells (Type1). Covalent joining of antigens (a glycopeptide- TSAPDTRPAP) to assembly sequence (Nap-G^{D}F^{D}F^{D}Y^{D}K) also lead to formation of anti-tumour vaccine correlated to MUC1 (Zhao et al. 2017). Apart from these, newly formed peptide antigens including multifaceted ones, multi-antigens, hybrids and also neoantigens have garnered

considerable interest in medical studies and have shown to be very effective in cancer immunotherapy. Even after the use of several adjuvants for peptide based immunotherapeutics, such as salts of Al, presently offered adjuvants such as oil emulsions, virosome, and TLR ligands often struggle from structural variability, which makes process comprehension and administering challenges (Acar et al. 2017). Different steps that are involve in the mechanisms of anticancer peptides to inhibit cancer growth shown in Fig. 2. Steps that are involved in the mechanism of peptide-based vaccines for the treatment of cancer shown in Fig. 3.

Neoantigen Peptide Vaccines

One of the greatest examples in the area of personalized therapeutics is neoantigen vaccines in treating cancer, its manufacture being a challenge for individual patients (Hu et al. 2018; Sahin et al. 2018; Keskin et al. 2019). Such vaccines are prepared because of an assembly of mutations (primarily, the somatic ones) which could be formed into a tumour that encodes tumour-definite antigens, namely, neoantigens. Immune targeting of such neoantigens is anticipated to encourage particular activation of immunity in opposition to neoplastic cells and further avert immunological diversion. It has been reported that five clinical trials have tested these neoantigen vaccines (personalized) in patients suffering from melanoma as well as glioblastoma multiforme and observed that they generated immune responses in humans (Keskin et al. 2019; Ott et al. 2017; Hilf et al.2019; Carreno et al. 2015; Sahin et al. 2017). Four out of these five trials have used synthetic peptides in order to immunize the formulations (Keskin et al. 2019; Ott et al. 2017; Hilf et al. 2019; Carreno et al. 2015), whereby fifth clinical trial involved use of synthetic RNA (Sahin et al. 2017). All these trials, however, used around a dozen shorter peptides for each patient for conducting studies ex vivo with respect to monitoring of the immune system. Generally, studies on personalized neoantigen vaccines involve hundreds of peptides ranging around 8–30 aa in length. For designing these peptides, many trials have accepted the system as: post a biopsy of tumour, the mutated epitopes were detected by WES (whole-exome sequencing) of the tumour along with the normal cells in parallel; EPTs (epitope peptides) (8–10 aa) which can attach to the personal human leukocyte antigen allele (HLA) were further detected by class I predictive algorithms of binding (Hoof et al. 2009); around 20 class I epitope peptides were selected as the targets for neoantigen vaccine, and thus were incorporated inside the synthetic long IMPs (immunizing peptides) with size of 15–30 aa. Further, they were then fabricated via a commercial vendor for peptides, cleaved as well as refined in presence of proper manufacturing measure or GMP conditions, and finally delivered to the patient in the form of immunizing peptides. Besides this, IMPs were produced for vaccine delivery since similar peptides have displayed effective stimulation of CD4 + as well as CD8 + T cells (Kenter et al. 2009). On the other hand, shorter overlapping ASPs (or assay peptides) were generated for evaluation of immune responses.

Previously, studies have shown that the average lead time for generation of 20 IMPs was 18–20 weeks, which was a time-consuming as well as expensive synthesis process of clinical- division peptides (Keskin et al. 2019; Ott et al. 2017). Therefore, minimizing this time as well as cost is important for allowing medication of different cancers like metastatic. Truex et al. (2020) described that automated flow peptide synthesis has the ability to ease generation of neoantigen peptides for developing personalized vaccines for cancer. The study showed that this flow of peptide synthesis produced higher quality IMPs within 35 min or less than that, while other methods take numerous hours or even days with comparable equivalents of response. The results even illustrated that automated flow technology could produce rapid synthesis of peptides for neoantigen vaccines (personalized), which is convenient for wider contexts. Thus, arrival of rapid as well as affordable genome- sequencing technology can most probably allow classes of personalized therapeutics using peptides, for treating cancers.

Antimicrobial Peptides (AMPs)

Those peptides that prevent humans from different microbial infections are known to be antimicrobial peptides or AMPs. These are active compounds, which are generated as an immediate innate immunological reaction and even possess a defence machinery in humans via utilization of cytotoxicity on the invading pathogens. Their physicochemical as well as structural features help in maintaining their specificity in regard with the target cells. There are two techniques for an AMP to enter the cell. First, a translocation process across the membrane and second is because of the secondary structure of peptide- leading to pore formation on the membrane (Datta and Roy 2021). Because of their nominal immunogenicity, effective tissue perceptibility as well as easier fabrication and production, this class of peptides make an absolute choice for treating cancers (Yavari et al. 2018). Previously, studies have approved that AMPs display anticancer action along with protection of regular cells at the same time. Defensins is a class of plant derived AMPs, which has a potential of augmenting potency of immunotherapy with respect to cancer. Li et al. (2014) with the help of recombinant plasmid technology, found a plasmid, which expressed beta-defensin 2 and played a significant role in gene therapy as well immunotherapy of cancer. Further, results suggested that physiological deflections taking place in dendritic cells (immature) probably increased

the adaptive immunosurveillance (anticancer). Sang et al. (2017) highlighted that ABP-dHC-Cecropin A along with its counterpart ABP-dHC-Cecropin A-K showed cytotoxicity against leukemic cells. Cecropin A has also been known to cause induction of cell death in HL-60 (a promyelocytic cell line) via ROS mechanism (Cerón et al. 2010), and through caspase-3, and −8 as well as Fas and Fas-L expressions, in human carcinoma cells (hepatocellular ones) (Jin et al. 2010).

Farshid et al. (2020) showed the impact of AMPs in lung cancer treatment. However, chemotherapy in lung cancer is often associated with after reactions, toxic behaviour as well as resistance to multidrug. To overcome such fallouts, anticancer peptides extracted from AMPs pose a promising strategy for lung cancer treatment. They display anticancer activity via lysis of cell membranes (membranolytic-based mechanisms) or non-membranolytic mechanisms. The former includes toroidal, carpet, aggregate channel and barrel-stave models, which have developed with time on the basis of net charge, amino acid sequence, membrane composition, peptide concentration, structure, oligomerization and amphipathicity (Lei et al. 2019). Using barrel-stave as well as toroidal pore generation mechanisms, lung cancer cells undergo structural damage followed by cell death. Even though actual process of cell death of these cells is not studied, some mechanistic processes like lethal depolarization of the membrane, interactions with intracellular targets post internalization, leakage of cellular constituents caused via development of pores and disturbing functions of membrane via disorganizing its asymmetry, are some of the proven research (Divyashree et al. 2020). On the other hand, non-membranolytic anticancer activities are carried out via induction of antiangiogenic, anti-metastatic as well as anti-apoptotic processes (Aaghaz et al. 2019). Therefore, future applications associated with AMPs will possibly encounter a major enhancement because of the advances in designs of peptides along with capacity of large-scale synthesis provided by therapeutic companies.

Venom Peptides

Venom Peptides (VPs) tend to show higher selectivity as well as specificity against neoplastic cells as compared to other agents, along with impacts on angiogenesis, cell progression, migration and invasion, plus modulation of immune responses (Roy and Bharadvaja 2021). Studies on VPs help in the fabrication of detailed peptide-based drugs with respect to specified cancer therapy. They offer several advantages in comparison to monoclonal antibodies associated therapies. Firstly, VPs show high resistance towards protease degradation, pH and temperature. They show natural stability via post-translational modifications like amidation of C-terminus, glycosylation of threonines, tyrosine sulfation and tryptophan bromination, and further via chemical modifications namely D-amino acid incorporation. Secondly, they comprise phylogenetically conserved domains, providing an opportunity for identification of similar peptides or isoforms with increased specificity in the rest of venom species, which results in a bigger amount of therapeutically dynamic peptides. Ultimately, advancements in synthesis of peptide techniques at a huge scale can turn VP-derived medications into affordable ones for patients (Ma et al. 2017; Pérez-Peinado et al. 2020). Snake, bee, and wasp venom possess some properties which can be utilised for boosting innate as well as adaptive immune mechanisms. Innate immune cells including NK cells as well as APCs regulate immediate responses to invading pathogens or cancerous cells, which is accompanied with an adaptive immunity driven by B and T cells. In immunocompromised mice, VPs obtained via snakes and bees, like phospholipase A2 (PLA2) and also melittin, boost NK-cell function by boosting synthesis of IFN-γ (Orsolic et al. 2012; Mahadevappa et al. 2017). In certain cases, bee venom has proven to circumvent the host's innate responses as well as actively stimulate the adaptive responses mediated by T-cells. PLA2 and melittin can potentially increase T-cell response by causing monocytes to produce more IL-1 and TNF-a. T-lymphocytes are also activated by bee PLA2 because it causes the membranes to break down along with the formation of antigens like free fatty acids and lysophospholipids (Orsolic et al. 2012). Owing to their higher specificity, these peptides will set a benchmark in the drug discovery for cancer, however, at present, only a few of them have reached clinical trials. More applications of such peptide-linked drugs are however affected by extraction of newer peptides involving use of bioactivity-led approaches which is cumbersome. Thus, for developing VP libraries, approaches involving high-efficiency screening, which will combine proteomics as well as genomics, are presently ongoing.

Peptide Hydrogels

Hydrogels are cross-linked polymer-based frameworks swollen in water, which are made by combining one or more monomers via a simple process. Numerous forms of hydrogels are being studied for cancer immunotherapy owing to their high adsorption rates, ability to sustain drug release, as well as biocompatible nature (Li et al.2016). Therapeutic agents as well as immunologic compounds are typically inserted in the scaffold present in these hydrogels. Recently, peptide-based hydrogels are being employed extensively for immunotherapy in cancer. Wu et al. (2017) demonstrated interleukin-15 (IL-15) and cisplatin encapsulated in combination in a thermosensitive polypeptide hydrogel for an

integrated skin cancer chemo immunotherapy. Jin et al. (2018) showed that incorporating melittin (a cationic peptide procured from bee venom) into the scaffold of hydrogel significantly decreases adverse effects while also improving antitumor activity and immunological responses. MRD (melittin-$RADA_{32}$-DOX) hydrogel's concurrent distribution of melittin and DOX results in a profound and long-lasting antitumor activity, enhanced innate immune responses, decreased immunodeficiency and caused induction of effector T-cells. Furthermore, findings also indicated that MRD hydrogels promote cell necrosis, the primary reason for MRD hydrogel-induced *in-vitro* cell deaths (Jin et al. 2018). They possess promising therapeutic activity against subcutaneous and metastatic tumours owing to its regulated dosage forms and ability to control the innate immune system, reduce M2-like TAMs, and specific anticancer as well as immune modulation properties (Jin et al. 2018).

For effective immunotherapies towards melanoma as well as hepatoma ascites in mice, Dai et al. (2020) created melittin-(RADA)6 (MR52) hydrogel scaffolds can be equipped using KN93, a particular calcium ion/CAMKII inhibitor. This MR52-KN93-hydrogel or MRK hydrogel greatly enhanced the combatting properties as well as immunogenic apoptosis level on tumour cells in comparison with independent KN93. In these mice treated using this MRK hydrogel, macrophages expressed more apoptotic protein ligand-1 (PD-L1), pointing to a potential move towards combination anti-PD-1 treatments exhibiting survival rates of about 30%.

Wang et al. (2016) identified a self-assembly peptide hydrogel that released Ovalbumin (OVA) for cancer immunotherapeutics. Treatment with these hydrogels raised IgG levels dramatically, suggesting the potential to elicit robust immune function in-vivo, improve antigen absorption, as well as cause DC maturation. Subsequently, Leach et al. (2018) established STINGel, an implantable hydrogel, by fusion of multi-domains of peptide hydrogel and cyclic dinucleotides (CDNs) for releasing CDNs (group of STINGs or stimulator of the interferon genes, agonists involved in immunotherapy of cancer) as per the demands. This enhanced survival rates in rat models of oral cancer unlike with CDN infusion treatment, as well as activated a powerful memory response that suppressed recurrence of tumours. Yu et al. (2018) formed a thermo- and ROS- sensitive peptide hydrogel platform (Seq: P(Me-D-1MT)-PEG-P(Me-D-1MT) in order to deliver aPD-L1 as well as D-1MT or dextro-1-methyl tryptophan that acted in the form of a drug carrier for secretion of checkpoint inhibitors and also controlled redox ratios in tumour environment. Lifespan of mice bearing B1610 tumour was substantially enhanced by this hydrogel. Song et al. (2019) fabricated a hydrogel from PEG-b-poly (l-alanine) using TCLs, GM-CSF or granulocyte-macrophage colony stimulating factor, and CTLA-4 (cytotoxic T-lymphocyte-associated antigen) and PD-1 antibodies loaded on it via independent self-aggregation of polypeptides in water-based solution. This hydrogel allowed consistent antigen release, while the GM-CSF could acquire and stimulate DCs as well as increase T-cell numbers, allowing for synergic regulation of T-cell based immune response and suppression of immunocompromised tumour microenvironment when used in conjunction with double checkpoint inhibitor. In 4T1 tumour-bearing and B16F10 melanoma models, treatment with this hydrogel led to improved production of IgG, cytokine concentrations comprising IFN-, IL-4, as well as TNF-α, and also proportion of CD8 + cells, whereas the numbers of T_{reg} cells declined. Recently, Hou et al. (2020) created polypeptide hydrogels co-encapsulated with silver sulphide, DOX, as well as Bestatin (a protease inhibitor) for treating mammary cancers. It improved activity of T-cells, which boosted antitumor immune responses. Development of primary tumours along with the incidence of proximal lung metastatic nodules were both prevented by this multipurpose hydrogel-mediated therapy. More such trials have been carried out in the past, which highlight the potential uses of peptide hydrogels solely, as well as in combinatorial therapies. Peptide-based hydrogels thus provide a lot of promise to be employed for cancer immunotherapy.

Peptide Assemblies

Peptide assemblies into nanostructures of morphologies like nanoparticles, nanotubes, nanoribbons, nanofibers and other hierarchical networks, managed by noncovalent interconnections namely hydrophobic, π-π stacking and hydrogen bonding interactions, show huge potential for addressing issues in cancer immunotherapy (Hendricks et al. 2017). The resulting nanostructures are ideal for peptide delivery and therapeutics since they have unique structural characteristics like biocompatibility as well as biodegradability (Acar et al. 2017; De Santis et al. 2015). Incorporating peptide therapeutics into peptide nanostructures prolongs therapeutic circulation and increases their accord with attacking receptors emerging from multifunctional impact. Additionally, active and passive targeting ability of peptide-based nanostructures facilitate aggregation of curatives at regions of tumours. Such advantages result in the huge potency

of peptide-derived biomaterials like vaccine adjuvants for enhancing immune responses (antigen based) or delivery of immunotherapeutic operatives. Keeping aside the regular roles of peptide assemblies as delivery agents, modulation of their accumulating features also permits for formation of self-delivering systems known as drug amphiphiles (Hu et al. 2013). Drug amphiphile assemblies enhance antigen density on the exterior of the platform and even remove the useless constituents in drug formations, hence increasing the biological safety, lowering the production price, as well as easing the administration of drugs. In addition, longer intervals for immune reactions in comparison to immediate treating impact observed in standard therapeutics like chemotherapy as well as phototherapy further limits the clinical applications of cancer immunotherapeutics in the patients found in later stages of cancer. As a result, merging the standard treatment for tumours with immunotherapy increases the immune response and has collective impacts for curable metastatic cancer therapy (Sharma and Allison 2015; Mahoney et al. 2015). Thus, for effective extension of survival rates in patients, combinatorial treatment including conventional treatment options and immunotherapy have been recently evolved. Because of their inherent ability for taking up cargos, both covalently as well as non-covalently, peptide assemblies have been widely employed in combinatorial immunotherapies. In addition to this, many preclinical trials have found that they show remarkable properties for modulating immune responses and inhibiting tumour growth, therefore being a great alternative for future immunotherapy in clinical trials.

Peptide Based Nano Drugs

Being chemically diverse, a class of peptides are also exploited for construction of supramolecular-based peptide nanodrugs (Chen et al. 2015). Various such nanostructures having different morphologies play significant roles in delivery of tumour antigens, chemotherapeutic drugs or photosensitive drugs (Panda et al. 2014). Mainly, peptide-based nanocarriers have various advantages like higher content for drug loading, better accumulation of tumours, as well as increased efficiency with respect to treatment techniques. Peptide sequence is adapted in order to become functional in a way that it provides these nanocarriers with specified properties, which cannot be easily attained in other carriers of drug delivery, namely, targeting of tumour as well as activating antitumor immune reactions.

Apart from this, there is a class of peptide-based nanocarriers namely nanospheres, which have a function of encapsulating photodynamic drugs. Another kind is that of injectable dipeptide-based nanoparticle that is further fabricated for PDT (photodynamic therapy: an efficient technique for clinical treatment) via self-aggregation of GA-associated (or glutaraldehyde) CDP (cationic diphenylalanine) (Ma et al. 2016). Chang et al. (2019) along with his research colleagues used Hep-based (heparin polymer) CDP nanoparticles (NPs) for forming peptide-derived NPs ranging around 100 nm in size with a negative voltage of roughly − 25 mV. Thereafter, acquired peptide-associated nanocarriers displayed an enhanced permeability and retention (EPR) effect in the delivery of Ce6 (the model photosensitive drug). Thus, these observations suggest that the CDP NPs could be used as flexible nanocarriers for many drugs. Ce6 charged with peptide-based NPs showed a prolonged circulation lifetime as compared to free Ce6 in solution, as a result increasing photodynamic anticancer effectiveness (Sun et al. 2012) for designing a potential vehicle for drug delivery, Chang et al. (2019) developed a tumour-receptive nanocarrier for PDT. Further, they examined human serum albumin (HSA) (consisting of different amino acids), poly-l-lysine, as well as modulator polyethylene glycol in order to develop nanospheres for their role in multi triggered peptide-associated drug delivery against tumour responsiveness (Zhang et al. 2016). Apart from this, HSA accelerated the breaking down of NPs because of the protease overexpression in tumour cells, and its isoelectric point at pH 5.3 supplied the specified pH reaction, and further disulphide appearance led the nanospheres to counter reducing agent glutathione or GSH. Therefore, above-mentioned adjustable carriers responsiveness to proteinase, pH as well as GSH, may release loaded drugs like Ce6, verteporfin as well as protoporphyrin IX.

Based on their properties, these peptides can be classified as tumour-targeting peptides, allergen peptides as well as tumour microenvironment responsive peptides. Novel nanodrugs are designed for incorporating such peptides-having specific functions of peptides as well as therapeutic impact of drugs, and as a result serve a crucial role in tumour treatment development. Tumour-targeting peptides are widely used for treatment of cancer and even for targeting lesions. These target the receptors on the tumour cell's surface or tumour's microenvironment, like in the case of tumour vasculature. RGD (Arg-Gly-Asp) peptide,

for instance, is a tumour vascular targeting peptide, which particularly binds αvβ3 as well as αvβ5 integrins and thus, has been used in various cancer therapies. Another study showed how a bifunctional peptide comprising RGD, NGR (short for Asn-Gly-Arg), AVPIAQK (a peptide sequence for apoptosis induction), and HAV (a histidine-alanine-valine sequence) was utilized in the form of nanocarrier for delivering chemotherapeutic drug Docetaxel. This indicated that Docetaxel or peptide NPs impede the tumour development in comparison to the group without peptide, suggesting efficiency of the bifunctional peptide-associated vehicle (Fan et al. 2017). Another class being the tumour micro-environment, is more related with tumour development as well as growth. In order to unravel the treatment of tumours, specified surroundings of the tumour microenvironment must be known for providing guidance (Altorki et al. 2019). This mechanism has been considered as a newer target, wherein various nanodrugs have been further fabricated for responding towards this microenvironment like acidity as well as enzyme contents. Some of the research works have proved that certain peptides respond to tumour microenvironment, while nanodrugs based on such peptides undergo changes in microenvironment like structure as well as shape differences (Han et al. 2017). Thus, responsive peptides have recently been introduced in therapeutic applications of cancer.

Apart from this, Han et al. (2017) developed nanodrugs having altering geometrical shapes, which were caused by pH. Further, they designed a chimeric peptide consisting of alkylated PpIX along with dimethylmaleic anhydride (DMA)-based peptide with a sequence namely, PEAK-DMA. This sequence can self-aggregate into spherical NPs in the accompaniment of physiological conditions as the DMA groups present prevents electrostatic interactions amongst PEAK moieties. When they enter into tumour sites (which are acidic), DMA groups are parted, and this retrieval of ion complementarity leads to the production of tiny rod-like NPs. Therefore, post an intravenous injection, PEAK-DMA stored in the tumour regions because of EPR effects, further undergoes alterations in morphology in order to increase its incorporation into cells, and achieve good PDT potency. These results suggest that the use of pH- reactive chimeric peptides in PDT might be a good approach for drug delivery. These peptide-modulated nanodrugs provide an easy method for incorporation of various functional molecules as well as treatment strategies. Such nanodrugs have advantages of controlled structure as well as morphology, increased tumour accumulation, and most importantly, improved therapeutic potency. Hence, peptides that self-assemble into nanostructures are widely used in cancer drug delivery vehicles with various promising applications as well in cancer therapies.

Challenges and Future Prospects

A lot of progress has been made since last ten years, clinical trials of peptide-derived therapeutics for cancer immunotherapy are challenging. The main challenge which makes this therapy as less successful is the immunogenicity in most of them (Skwarczynski et al. 2016). Antibodies or vaccines, epitopes derived from standard peptides generally show lesser affinity with respect to a particular target. Such association might be further dropped via a targets or receptors phenotypic heterogeneity in individuals. Other challenges include their safe administration, based on the observations of significant side effects in preclinical trials. These usually arise as a result of incompatibility of peptide therapeutics or their delivery system, off-target therapeutic delivery as well as associated autoimmunity. Therefore, development of peptide therapeutics accompanied with higher immunogenicity as well as formulation with standard safety of administration will be the next step in the peptide-based cancer immunotherapy. Besides this, a design for multivalent

Table 2 List of clinical trials of peptides, their conditions, different peptides that are used and their phage of trails which is in cancer immunotherapy(https://www.clinicaltrials.gov/)

Title	Conditions	Interventions/ Peptide used	Phase	NCT number
Ability of a dendritic cell vaccine to immunize melanoma or epithelial cancer patients against defined mutated neoantigens expressed by the autologous cancer	Melanoma, Gastrointestinal Cancer, Breast Cancer, Ovarian Cancer, Pancreatic Cancer	Peptide loaded dendritic cell vaccine	Phase 2	NCT03300843
Cancer peptides plus GM-CSF and adjuvant in breast cancer	Breast Cancer	ESR1 peptide vaccine	Phase 1	NCT04270149
Pooled mutant KRAS-targeted long peptide vaccine combined with Nivolumab and Ipilimumab for patients with resected MMR-p colorectal and pancreatic cancer	Colorectal Cancer, Pancreatic Cancer	KRAS peptide vaccine, Nivolumab, Ipilimumab	Phase 1	NCT04117087
Nivolumab, Ipilimumab and OTSGC-A24 therapeutic peptide vaccine in Gastric Cancer - a Combination Immunotherapy Phase Ib Study.	Gastric Cancer	OTSGC-A24, Nivolumab, Ipilimumab	Phase 1	NCT03784040
Dendritic Cell Immunotherapy for the Patients With Pancreatic Cancer	Dendritic Cells, Pancreatic Neoplasms	Dendritic cells pulsed with MUC-1/WT-1 peptides	Phase 1 and 2	NCT03114631
Immunotherapy Vaccine and Herceptin in Breast Cancer	Breast cancer	GP2 peptide + GM-CSF vaccine plus trastuzumab	Phase 1	NCT03014076
Universal Cancer Peptide-based Vaccination in Metastatic NSCLC	Metastatic Non-small Cell Lung Cancer	UCPVax	Phase 1 and 2	NCT02818426
IDH1 Peptide Vaccine for Recurrent Grade II Glioma	Brain Cancer, Brain Neoplasm (Primary), Brain Neoplasms (Recurrent), Brain Tumour	PEPIDH1M vaccine	Phase 1	NCT02193347
Phase I Study of an Oncofetal Antigen Multi-Peptide Immunotherapy in Subjects With Hematologic Cancer	Acute Myelogenous Leukemia (AML), Multiple Myeloma (MM), Myelodysplastic Syndrome (MDS), Smoldering Multiple Myeloma (sMM)	Sargramostim, BB-MPI-03, Montanide	Phase 1	NCT02240537
Pilot Study of a Breast Cancer Vaccine Plus PolyI-CLC for Breast Cancer	Breast Cancer	Poly-ICLC, 9 Peptides from Her-2/ neu, CEA, & CTA, Peptide-tet	Phase 1	NCT01532960
A Phase I Safety Study of a Cancer Vaccine to Treat HLA-A2 Positive Advanced Stage Ovarian, Breast and Prostate Cancer	Ovarian Neoplasms, Breast Neoplasms, Prostatic Neoplasms	DPX-0907 (consisting of 7 tumour-specific HLA-A2-restricted peptides, a universal T Helper peptide, a polynucleotide adjuvant, a liposome and Montanide ISA51 VG)	Phase 1	NCT01095848
Radiation, Chemotherapy, Vaccine and AntiMART-1 and Anti-gp100 Cells for Patients With Metastatic Melanoma	Melanoma	MART-1: 26–35(27 L) Peptide	Phase 2	NCT00923195
PSMA and TARP Peptide Vaccine With Poly ICLC Adjuvant in HLA-A2 (+) Patients With Elevated PSA After Initial Definitive Treatment	Prostate Cancer	PSMA and TARP peptide vaccine with Poly IC-LC adjuvant	-	NCT00694551
Vaccine Therapy in treating patients with Metastatic, Progressive Prostate Cancer	Prostate Cancer	NY-ESO-1/LAGE-1 HLA class I/II peptide vaccine	Phase 1	NCT00616291
Stem cell transplant, Chemotherapy, and Biological Therapy in Treating Patients with HighRisk or Refractory Multiple Myeloma	Multiple Myeloma and Plasma Cell Neoplasm	CMV pp65 peptide, hTERT I540/R572Y/ D988Y multipeptide vaccine	Phase 1 and 2	NCT00499577
GV 1001 Immunotherapy in Patients with Non Small Cell Lung Cancer (NSCLC)	Carcinoma (Non-Small-Cell Lung)	GV 1001 Telomerase peptide	-	NCT00509457

Table 2 (continued)

Title	Conditions	Interventions/ Peptide used	Phase	NCT number
Proteinase 3 PR1 Peptide Mixed with Montanide ISA-51 VG Adjuvant and Administered with GMCSF and PEG-INTRON(R)	Leukemia	Peptide Vaccine (PR1 Peptide)	Phase 2	NCT00415857
Activity and safety of peptide-based immunotherapy in patients with Squamous cell carcinoma of the head and neck.	Oropharynx Squamous Cell Carcinoma, Larynx Squamous Cell Carcinoma, Hypopharynx Squamous Cell Carcinoma, Oral Cavity Squamous Cell Carcinoma	IO102 Peptide vaccine	Phase 2	NCT04445064

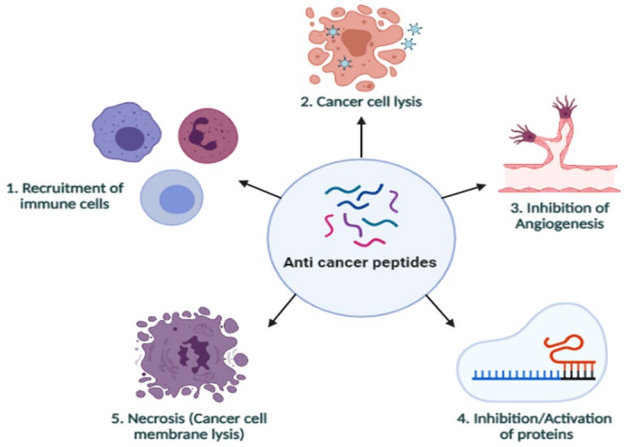


Fig. 2 Different steps that are involve in the mechanisms of anticancer peptides to inhibit cancer growth

peptide neoantigens, checkpoint blockades as well as antigens (Eskandari et al. 2017) is a multifaceted strategy for improvement in the affinity of peptide therapeutics with target substrates, thereafter efficiently resulting in higher immune response (Li et al. 2019). In addition to this, formation of a multi-biomarker-controlled drug release system and integration of numerous target-guiding epitopes into delivery systems might probably avert off-targeting release of curatives. Therefore, peptide-based cancer immunotherapy poses a hopeful and effective option for cancer treatment in the future.

Conclusions

Cancer immunotherapy shows a lot of promise for treating cancers owing to their potential to prevent tumour recurrence as well as metastatic spread by causing long-term memory and also triggering hosts' immunological responses to defend from localized cancerous cells. The present review outlines numerous approaches incorporating peptides in cancer immunotherapies, which presented therapeutic roles of peptides and peptide-based assemblies derived from different sources as well as their mode of action for attenuating the immunological function. Short peptides are being used as therapeutic agents like checkpoint blockades, antigens, or even vaccine adjuvants accounting to their remarkable biocompatibility and structural characteristics obtained from proteins. Peptide assembly, on the other hand, has demonstrated improved functionality in attacking and also co-delivery of drugs in a controlled way. Several clinical

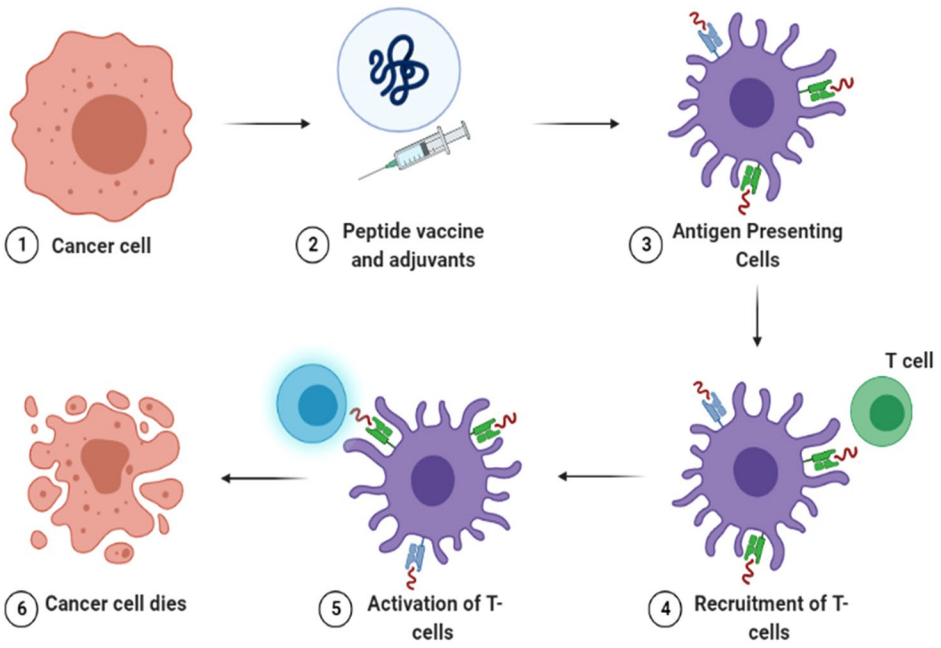


Fig. 3 Steps that are involved in the mechanism of peptide-based vaccines for the treatment of cancer

and preclinical studies mentioned before have proven that peptide-based therapeutic approaches have an impressive ability to modulate immunological responses and prevent tumour development, indicating their enormous potential in immunotherapeutics. Furthermore, clinical studies should be conducted on these peptide-based immunotherapies, which makes them a promising strategy for treating cancers in near future.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Aaghaz S, Gohel V, Kamal A (2019) Peptides as potential anticancer agents. Curr Top Med Chem 19(17):1491–1151

Abbas M, Zou Q, Li S et al (2017) Self-assembled peptide- and protein based nanomaterials for antitumor photodynamic and photothermal therapy. Adv Mater 29:1605021

Acar H, Srivastava S, Chung EJ et al (2017) Self-assembling peptide based building blocks in medical applications. Adv Drug Deliver Rev 110–111:65–79

Ali R, Rani R, Kumar S (2013) New peptide based therapeutic approaches. In: Ghulam Md A, Ishfaq Ahmed S (eds) Advances in protein chemistry. OMICS Group eBooks, Jeddah

Altorki NK, Markowitz GJ, Gao D, Port JL, Saxena A, Stiles B, McGraw T, Mittal V (2019) The lung microenvironment: an important regulator of tumour growth and metastasis. Nat Rev Cancer 19(1):9–31. https://doi.org/10.1038/s41568-018-0081-9

Ascierto PA, Marincola FM (2014) What have we learned from cancer immunotherapy in the last 3 years? J Transl Med 12:141

Black M, Trent A, Kostenko Y et al (2012) Self-assembled peptide amphiphile micelles containing a cytotoxic T-cell epitope promote a protective immune response in vivo. Adv Mater 24:3845–3849

Buchbinder EI, Desai A (2016) CTLA-4 and PD-1 pathways. Am J Clin Oncol 39:98–106

Byun DJ, Wolchok JD, Rosenberg LM et al (2017) Cancer immunotherapy—immune checkpoint blockade and associated endocrinopathies. Nat Rev Endocrinol 13:195–207

Carreno BM et al (2015) A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific t cells. Science 348:803–808. https://doi.org/10.1126/science.aaa3828

Cerón JM, Contreras-Moreno J, Puertollano E, De Cienfuegos GA, Puertollano MA, De Pablo MA (2010) The antimicrobial peptide cecropin A induces caspase-independent cell death in human promyelocytic leukemia cells. Peptides 31(8):1494–1503. https://doi.org/10.1016/j.peptides.2010.05.008

Cha E, Klinger M, Hou Y, Cummings C, Ribas A, Faham M et al (2014) Improved survival with T cell clonotype stability after anti-CTLA-4 treatment in cancer patients. Sci Transl Med 6:238ra70

Chang R, Zou Q, Xing R, Yan X (2019) Peptide-Based Supramolecular Nanodrugs as a New Generation of Therapeutic Tool Boxes against Cancer. Adv Therap 2:1900048. https://doi.org/10.1002/adtp.201900048

Chang HN, Liu BY, Qi YK et al (2015) Blocking of the PD-1/PD-L1 interaction by a D-peptide antagonist for cancer immunotherapy. Angew Chem Int Ed 54:11760–11764

Chapel A, Garcia-Beltran WF, Hölzemer A, Ziegler M, Lunemann S, Martrus G et al (2017) Peptide-specific engagement of the activating NK cell receptor KIR2DS1. Sci Rep 7:2414

Chen C, Liu K, Li J, Yan X (2015) Functional architectures based on self-assembly of bio-inspired dipeptides: structure modulation and its photoelectronic applications. Adv Colloid Interface Sci 225:177

Chen T, Li Q, Liu Z et al (2019) Peptide-based and small synthetic molecule inhibitors on PD-1/PD-L1 pathway: a new choice for immunotherapy? Eur J Medicinal Chem 161:378–398

Cheng K, Ding Y, Zhao Y, Ye S, Zhao X, Zhang Y et al (2018) Sequentially responsive therapeutic peptide assembling nanoparticles for dual-targeted cancer immunotherapy. Nano Lett 18:3250–3258

Choi YJ, Park S-J, Park Y-S, Park HS, Yang KM, Heo K (2018) EpCAM peptide-primed dendritic cell vaccination confers significant anti-tumour immunity in hepatocellular carcinoma cells. PLoS One 13:e0190638

Das J, Khakoo SI (2015) NK cells: tuned by peptide? Immunol Rev 267:214–227

Datta S, Roy A (2021) Antimicrobial peptides as potential therapeutic agents: a review. Int J Pept Res Ther 27:555–577. https://doi.org/10.1007/s10989-020-10110-x

David A (2017) Peptide ligand-modified nanomedicines for targeting cells at the tumour microenvironment. Adv Drug Delivery Rev 119:120–142

Dai X, Meng J, Deng S, Zhang L, Wan C, Lu L, Huang J, Hu Y, Zhang Z, Li Y, Lovell JF, Wu G, Yang K, Jin H (2020) Targeting CAMKII to reprogram tumour-associated macrophages and inhibit tumour cells for cancer immunotherapy with an injectable hybrid peptide hydrogel. Theranostics 10(7):3049–3063. https://doi.org/10.7150/thno.42385

De Santis E, Ryadnov MG (2015) Peptide self-assembly for nanomaterials: the old new kid on the block. Chem Soc Rev 44:8288–8300

Del Paggio JC (2018) Cancer immunotherapy and the value of cure. Nat Rev Clin Oncol 15:268–270

Desrichard A, Snyder A, Chan TA (2016) Cancer neoantigens and applications for immunotherapy. Clin Cancer Res 22:807–812

Divyashree M, Mani MK, Reddy D et al (2020) Clinical applications of antimicrobial peptides (AMPs): where do we stand now? Protein Pept Lett 27(2):120–134

Fan R, Mei L, Gao X et al (2017) Self-assembled bifunctional peptide as effective drug delivery vector with powerful antitumor activity. Adv Sci (Weinh) 4(4):1600285. https://doi.org/10.1002/advs.201600285

Zandsalimi F, Talaei S, Ahari MN, Aghamiri S, Raee P, Roshanzamiri S, Yarian F, Bandehpour M, Zadeh ZZ (2020) Antimicrobial peptides: a promising strategy for lung cancer drug discovery? Expert Opin Drug Discov 15(11):1343–1354. https://doi.org/10.1080/17460441.2020.1791080

Flippe L, Bézie S, Anegon I, Guillonneau C (2019) Future prospects for CD8 + regulatory T cells in immune tolerance. Immunol Rev 292(1):209–224

Gao S, Yang D, Fang Y, Lin X, Jin X, Wang Q et al (2019) Engineering nanoparticles for targeted remodeling of the tumour microenvironment to improve cancer immunotherapy. Theranostics 9:126–151

Gonzalez FE, Ram´ırez M, Allerbring EB et al (2014) Melanocortin 1 receptor- derived peptides are efficiently recognized by cytotoxic T lymphocytes from melanoma patients. Immunobiology 219(3):189–197

Gordon SR, Maute RL, Dulken BW et al (2017) PD-1 expression by tumour-associated macrophages inhibits phagocytosis and tumour immunity. Nature 545:495–499

Gubin MM, Zhang X, Schuster H et al (2014) Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. Nature 515:577–581

Guillerey C, Huntington ND, Smyth MJ (2016) Targeting natural killer cells in cancer immunotherapy. Nat Immunol 17:1025

Guo Y, Wang D, Song Q et al (2015) Erythrocyte membrane-enveloped polymeric nanoparticles as nanovaccine for induction of antitumor immunity against melanoma. ACS Nano 9:6918–6933

Han K, Zhang WY, Ma ZY, Wang SB, Xu LM, Liu J, Zhang XZ, Han HY (2017) Acidity-triggered tumour retention/internalization of chimeric peptide for enhanced photodynamic therapy and real-time monitoring of therapeutic effects. ACS Appl Mater Interfaces 9(19):16043–16053. https://doi.org/10.1021/acsami.7b04447

Hayashi MA, Ducancel F, Konno K (2012) Natural peptides with potential applications in drug development, diagnosis, and/or biotechnology. Int J Pept 2012:757838

Hendricks MP, Sato K, Palmer LC et al (2017) Supramolecular assembly of peptide amphiphiles. Acc Chem Res 50:2440–2448

Hilf N et al (2019) Actively personalized vaccination trial for newly diagnosed glioblastoma. Nature 565:240–245. https://doi.org/10.1038/s41586-018-0810-y

Hoof I et al (2009) Netmhcpan, a method for mhc class i binding prediction beyond humans. Immunogenetics 61:1–13. https://doi.org/10.1007/s00251-008-0341-z

Hou X, Dai X, Yang J, Zhang B, Zhao DH, Li CQ, Yin ZY, Liu B (2020) Injectable polypeptide-engineered hydrogel depot for amplifying the anti-tumour immune effect induced by chemo-photothermal therapy. J Mater Chem B Mater Biol Med 8(37):8623–8633

Hu X, Hu J, Tian J et al (2013) Polyprodrug amphiphiles: hierarchical assemblies for shape-regulated cellular internalization, trafficking, and drug delivery. J Am Chem Soc 135:17617–17629

Hu Z, Ott PA, Wu CJ (2018) Towards personalized, tumour-specific, therapeutic vaccines for cancer. Nat Rev Immunol 18:168–182. https://doi.org/10.1038/nri.2017.131

Jäger E, Höhn H, Necker A et al (2002) Peptide-specific CD8 + T-cell evolution in vivo: response to peptide vaccination with Melan-A/ MART-1. Int J Cancer 98:376–388

Joller N, Lozano E, Burkett PR et al (2014) Treg cells expressing the coinhibitory molecule TIGIT selectively inhibit proinflammatory Th1 and Th17 cell responses. Immunity 40:569–581

Jin X, Mei H, Li X, Ma Y, Zeng A-H, Wang Y et al (2010) Apoptosis-inducing activity of the antimicrobial peptide cecropin of Musca domestica in human hepatocellular carcinoma cell line BEL-7402 and the possible mechanism. Acta Biochim Biophys Sin (Shanghai) 42(4):259–265. https://doi.org/10.1093/abbs/gmq021

Ma K, Xing R, Jiao T, Shen G, Chen C et al (2016) Injectable self-assembled dipeptide-based nanocarriers for tumor delivery and effective in vivo photodynamic therapy. ACS Appl Mater Interfaces 8(45):30759–30767. https://doi.org/10.1021/acsami.6b10754

Kaumaya PT (2015) A paradigm shift: cancer therapy with peptide-based B-cell epitopes and peptide immunotherapeutics targeting multiple solid tumour types: emerging concepts and validation of combination immunotherapy. Hum Vaccines Immunother 11:1368–1386

Kenter GG et al (2009) Vaccination against hpv-16 oncoproteins for vulvar intraepithelial neoplasia. N Engl J Med 361:1838–1847. https://doi.org/10.1056/NEJMoa0810097

Keskin DB et al (2019) Neoantigen vaccine generates intratumoral t cell responses in phase ib glioblastoma trial. Nature 565:234–239. https://doi.org/10.1038/s41586-018-0792-9

Leach DG, Dharmaraj N, Piotrowski SL, Lopez-Silva TL, Lei YL, Sikora AG, Young S, Hartgerink JD (2018) STINGel: controlled release of a cyclic dinucleotide for enhanced cancer immunotherapy. Biomaterials 163:67–75

Lei J, Sun L, Huang S et al (2019) The antimicrobial peptides and their potential clinical applications. Am J Transl Res 11(7):3919–3931

Li C, Zhang N, Zhou J, Ding C, Jin Y, Cui X et al (2018) Peptide blocking of PD-1/PD-L1 interaction for cancer immunotherapy. Cancer Immunol Res 6:178–188
Li D, Wang W, Shi HS, Fu YJ, Chen X, Chen XC et al (2014) Gene therapy with beta-defensin 2 induces antitumor immunity and enhances local antitumor effects. Hum Gene Ther 25(1):63–72. https://doi.org/10.1089/hum.2013.161
Li J, Pu K (2019) Development of organic semiconducting materials for deep-tissue optical imaging, phototherapy and photoactivation. Chem Soc Rev 48:38–71
Li M, Zhao X, Dai J et al (2019) Peptide therapeutics and assemblies for cancer immunotherapy. Sci China Mater 62:1759–1781
Li Y, Fang M, Zhang J, Wang J, Song Y, Shi J, Li W, Wu G, Ren J, Wang Z, Zou W, Wang L (2016) Hydrogel dual delivered celecoxib and anti-PD-1 synergistically improve antitumor immunity. Oncoimmunology 5:e1074374
Lipson EJ, Forde PM, Hammers HJ, Emens LA, Taube JM, Topalian SL (2015) Antagonists of PD-1 and PD-L1 in cancer treatment. Semin Oncol 42:587e600
Littman DR (2015) Releasing the brakes on cancer immunotherapy. Cell 162:1186–1190
Liu LQ, Miao YB (2014) Metastatic melanoma targeting property of a novel Tc-99m-labeled HYNIC-conjugated lactam bridge-cyclized alpha-MSH peptide. Nucl Med Biol 41(7):622
Liu TY, Hussein WM, Jia Z et al (2013) Self-adjuvanting polymer–peptide conjugates as therapeutic vaccine candidates against cervical cancer. Biomacromol 14:2798–2806
Liu TY, Hussein WM, Giddam AK et al (2015) Polyacrylate-based delivery system for self-adjuvanting anticancer peptide vaccine. J Med Chem 58:888–896
Liu Y, Zhang D, Qiao ZY et al (2015) A peptide-network weaved nanoplatform with tumour microenvironment responsiveness and deep tissue penetration capability for cancer therapy. Adv Mater 27:5034–5042
Ma R, Mahadevappa R, Kwok HF (2017) Venom-based peptide therapy: insights into anti-cancer mechanism. Oncotarget 8(59):100908
Magiera-Mularz K, Skalniak L, Zak KM et al (2017) Bioactive macrocyclic inhibitors of the PD-1/PD-L1 immune checkpoint. Angew Chem Int Ed 56:13732–13735
Mahadevappa R, Ma R, Kwok HF (2017) Venom peptides: improving specificity in cancer therapy. Trends Cancer 3(9):611–614. https://doi.org/10.1016/j.trecan.2017.07.004
Mahoney KM, Rennert PD, Freeman GJ (2015) Combination cancer immunotherapy and new immunomodulatory targets. Nat Rev Drug Discov 14:561–584
Maio M, Danielli R, Chiarion-Sileni V, Pigozzo J, Parmiani G, Ridolfi R et al (2013) Efficacy and safety of ipilimumab in patients with pre-treated, uveal melanoma. Ann Oncol 24:2911e5
Malonis RJ, Jonathan RL, Olivia V (2019) Peptide-based vaccines: current progress and future challenges. Chemical Rev 120(6):3210–3229
Marqus S, Pirogova E, Piva TJ (2017) Evaluation of the use of therapeutic peptides for cancer treatment. J Biomed Sci 24:21. https://doi.org/10.1186/s12929-017-0328-x
Maute RL, Gordon SR, Mayer AT et al (2015) Engineering high-affinity PD-1 variants for optimized immunotherapy and immuno-pet imaging. Proc Natl Acad Sci USA 112:E6506–E6514
Melero I, Gaudernack G, Gerritsen W et al (2014) Therapeutic vaccines for cancer: an overview of clinical trials. Nat Rev Clin Oncol 11:509–524
Miao Q, Yeo DC, Wiraja C et al (2018) Near-infrared fluorescent molecular probe for sensitive imaging of keloid. Angew Chem Int Ed 57:1256–1260
Molino NM, Anderson AKL, Nelson EL et al (2013) Biomimetic protein nanoparticles facilitate enhanced dendritic cell activation and cross- presentation. ACS Nano 7:9743–9752
Munir Ahmad S, Martinenaite E, Hansen M et al (2016) PD-L1 peptide co- stimulation increases immunogenicity of a dendritic cell-based cancer vaccine. Oncoimmunology 5:e1202391
Neek M, Tucker JA, Kim TI et al (2018) Co-delivery of human cancer testis antigens with adjuvant in protein nanoparticles induces higher cell-mediated immune responses. Biomaterials 156:194–203
Netea MG et al (2020) Defining trained immunity and its role in health and disease. Nat Rev Immunol 20:375–388. https://doi.org/10.1038/s41577-020-0285-6
Orsolic N (2012) Bee venom in cancer therapy. Cancer Metastasis Rev 31:173–194
Ott PA et al (2017) An immunogenic personal neoantigen vaccine for patients with melanoma. Nature 547:217–221. https://doi.org/10.1038/nature22991
Panda JJ, Chauhan VS. Polym (2014) Short peptide based self-assembled nanostructures: implications in drug delivery and tissue engineering. Chem 5:4418
Papanikolaou E, Bosio A (2021) The promise and the hope of gene therapy. Front Genome Editing 3:4
Park JH, Geyer MB, Brentjens RJ (2016) CD19-targeted CAR T-cell therapeutics for hematologic malignancies: interpreting clinical outcomes to date. Blood 127:3312–3320
Patel S, Fu S, Mastio J, Dominguez GA, Purohit A, Kossenkov A et al (2018) Unique pattern of neutrophil migration and function during tumour progression. Nat Immunol 19:1236
Pérez-Peinado C, Defaus S, Andreu D (2020) Hitchhiking with nature: snake venom peptides to fight cancer and superbugs. Toxins 12(4):255
Pompano RR, Chen J, Verbus EA et al (2014) Titrating T-cell epitopes within self-assembled vaccines optimizes CD4 + helper T cell and antibody outputs. Adv Healthcare Mater 3:1898–1908
Postow MA, Chesney J, Pavlick AC, Robert C, Grossmann K, McDermott D et al (2015) Nivolumab and ipilimumab versus ipilimumab in untreated melanoma. N Engl J Med 372:2006–2017
Qian Y, Jin H, Qiao S et al (2016) Targeting dendritic cells in lymph nodes with an antigen peptide-based nanovaccine for cancer immunotherapy. Biomaterials 98:171–183
Qian Y, Qiao S, Dai Y, Xu G, Dai B, Lu L et al (2017) Molecular-targeted immunotherapeutic strategy for melanoma via dual-targeting nanoparticles delivering small interfering RNA to tumour-associated macrophages. ACS Nano 11:9536–9549
Ribas A, Wolchok JD (2018) Cancer immunotherapy using checkpoint blockade. Science 359:1350–1355
Riley RS, June CH, Langer R, Mitchell MJ (2019) Delivery technologies for cancer immunotherapy. Nat Rev Drug Discov 18(3):175–196
Robert C, Schachter J, Long GV, Arance A, Grob JJ, Mortier L et al (2015) Pembrolizumab versus ipilimumab in advanced melanoma. New Engl J Med 372:2521–2532
Rodriguez-Cabello JC, Arias FJ, Rodrigo MA, Girotti A (2016) Elastin-like polypeptides in drug delivery. Adv Drug Delivery Rev 97:85–100
Roy A, Bharadvaja N (2021) Venom-derived bioactive compounds as potential anticancer agents: a review. Int J Pept Res Ther 27:129–147. https://doi.org/10.1007/s10989-020-10073-z
Rudra JS, Sun T, Bird KC et al (2012) Modulating adaptive immune responses to peptide self-assemblies. ACS Nano 6:1557–1564

Sahin U et al (2017) Personalized rna mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. Nature 547:222–226. https://doi.org/10.1038/nature23003

Sahin U, Tureci O (2018) Personalized vaccines for cancer immunotherapy. Science 359:1355–1360. https://doi.org/10.1126/science.aar7112

Sang M, Zhang J, Zhuge Q (2017) Selective cytotoxicity of the antibacterial peptide ABP-dHC-Cecropin A and its analog towards leukemia cells. Eur J Pharmacol 803:138–147. https://doi.org/10.1016/j.ejphar.2017.03.054

Sanmamed MF, Chen L (2018) A paradigm shift in cancer immunotherapy: from enhancement to normalization. Cell 175(2):313–326

Sharma P, Allison JP (2015) The future of immune checkpoint therapy. Science 348:56–61

Skakuj K, Wang S, Qin L et al (2018) Conjugation chemistry-dependent T-cell activation with spherical nucleic acids. J Am Chem Soc 140:1227–1230

Skwarczynski M, Toth I (2014) Recent advances in peptide-based subunit nanovaccines. Nanomedicine 9:2657–2669

Skwarczynski M, Toth I (2016) Peptide-based synthetic vaccines. Chem Sci 7:842–854

Song H, Yang P, Huang P, Zhang C, Kong D, Wang W (2019) Injectable polypeptide hydrogel-based co-delivery of vaccine and immune checkpoint inhibitors improves tumour immunotherapy. Theranostics 9(8):2299

Truex NL, Holden RL, Wang BY et al (2020) Automated flow synthesis of tumour neoantigen peptides for personalized immunotherapy. Sci Rep 10:723. https://doi.org/10.1038/s41598-019-56943-5

Tsoras AN, Champion JA (2018) Cross-linked peptide nanoclusters for delivery of oncofetal antigen as a cancer vaccine. Bioconj Chem 29:776–785

Wakabayashi R, Sakuragi M, Kozaka S et al (2018) Solid-in-oil peptide nanocarriers for transcutaneous cancer vaccine delivery against melanoma. Mol Pharm 15:955–961

Wang H, Luo Z, Wang Y, He T, Yang C, Ren C, Ma L, Gong C, Li X, Yang Z (2016) Enzyme-catalyzed formation of supramolecular hydrogels as promising vaccine adjuvants. Adv Funct Mater 26(11):1822–1829

Wu DD, Gao YF, Qi YM, Chen LX, Ma YF, Li YZ (2014) Peptide-based cancer therapy: opportunity and challenge. Cancer Lett 351(1):13–22

Wu X, Wu Y, Ye H, Yu S, He C, Chen X (2017) Interleukin-15 and cisplatin co-encapsulated thermosensitive polypeptide hydrogels for combined immuno-chemotherapy. J Control Release 255:81–93

Xiao YF, Jie MM, Li BS, Hu CJ, Xie R, Tang B, Yang SM (2015) Peptide-based treatment: a promising cancer therapy. J Immunol Res 2015:761820. https://doi.org/10.1155/2015/761820

Xie J, Yang C, Liu Q et al (2017) Encapsulation of hydrophilic and hydrophobic peptides into hollow mesoporous silica nanoparticles for enhancement of antitumor immune response. Small 13:1701741

Yadav A, Mohite S (2020) Recent advances in protein and peptide drug delivery. *Res J Pharma* Dosage Forms Tech 12(3):205 212

Yan Z, Wu Y, Du J, Li G, Wang S, Cao W et al (2016) A novel peptide targeting Clec9a on dendritic cells for cancer immunotherapy. Oncotarget 7:40437

Yavari B, Mahjub R, Saidijam M et al (2018) The potential use of peptides in cancer treatment. Current Protein Pept Sci 19(8):759–770

Yin L, Sun H, Zhang H et al (2019) Quantitatively visualizing tumor-related protease activity in vivo using a ratiometric photoacoustic probe. J Am Chem Soc 141:3265–3273

Yu S, Wang C, Yu J, Wang J, Lu Y, Zhang Y, Zhang X, Hu Q, Sun W, He C (2018) Injectable bioresponsive gel depot for enhanced immune checkpoint blockade. Adv Mater 30(28):e1801527

Zhang S (2017) Discovery and design of self-assembling peptides. Interface Focus 7:20170028

Zhang D, Qi GB, Zhao YX, Qiao SL, Yang C, Wang H (2015) *In situ* formation of nanofibers from purpurin18-peptide conjugates and the assembly induced retention effect in tumour sites. Adv Mater 27:6125–6130

Zhang L, Huang Y, Lindstrom AR, Lin TY, Lam KS, Li Y (2019) Peptide-based materials for cancer immunotherapy. Theranostics 9(25):7807–7825. https://doi.org/10.7150/thno.37194

Zhang N, Zhao F, Zou Q, Li Y, Ma G, Yan X (2016 Nov) Multitriggered tumour-responsive drug delivery vehicles based on protein and polypeptide coassembly for enhanced photodynamic tumour ablation. Small 12(43):5936–5943. https://doi.org/10.1002/smll.201602339

Zhu G, Zhang F, Ni Q et al (2017) Efficient nanovaccine delivery in cancer immunotherapy. ACS Nano 11:2387–2392

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

ISSN No:-2456-2165

Soil and Water Assessment Tool (SWAT) Modelling: A Review

Noopur Awasthi[1], Munendra Kumar [2],
[1]Student, Department of Civil Engineering, Delhi Technological University, New Delhi, India
[2]Professor, Department of Civil Engineering, Delhi Technological University, New Delhi, India

Abstract:- In the recent years human activities and climate change affected the hydrological cycle under the shifting environmental conditions worldwide and dispersed hydrological modelling based on practical methods have become the key topic for hydrological research. There are so many hydrological tools are developed to predict the future patterns by analyzing the historical data. Soil & Water Assessment Tool (SWAT) model is one of the distinctive distributed hydrological models which has been extensively used in hydrological research and to predict the future streamflow of rivers and sediment flow and agricultural production. This paper aims to analyze the research progress based on Soil & Water Assessment Tool in recent years. This paper focused on some prevailing glitches, sought the future research areas, and anticipated model enhancement to acclimate to the local environmental variations. This study should be beneficial to beginners in the field of SWAT modelling and also should be valuable for future development of SWAT models. It will also provide reference for users of SWAT modelling in related subject.

Keywords:- *Climate Change, Hydrological Modelling, Soil & Water Assessment Tool (SWAT).*

I. INTRODUCTION

Hydrological modelling has been used globally as a commanding tool for water resources research and management. Many researchers have focused on hydrological modelling of the areas that have experienced an enlarged frequency of extreme events i.e. drought and flood. Over recent decades, hydrological modelling has facilitated in the enhancement of the conservation and sustainable use of water resources as well as other natural resources, exclusively through researches on impacts of land use/land cover changes, mitigation of climate change and water pollution [1].

The hydrological modelling principally includes empirical models, conceptual models and physically based models. Due to the boosted impact of human activities and climate change on the environment, the underlying surface changes significantly, more extreme precipitation and temperature events happened, and conceptual models could not reflect the change of runoff and nutrients caused by the spatial difference of the underlying surface. But distributed models divide the whole watershed into small units with almost identical properties, the simulation results of which are closer to authenticity [2,3].

The essential responsibilities of water resources research under changing environment are to objectively evaluate the periodicity and randomness of the water cycle and quantitatively recognize the aids and modes of natural and artificial factors that drive the evolution of water cycle, reproduce the historical process of the water cycle to predict the future, and serve water resources regulation and control decision-making. The distributed hydrological model can quantitatively pronounce the process of hydrological elements and is an effective technical way to identify the evolution law of water cycle. It is of great significance for the rational development, utilization and management of water resources [4].

The Soil and Water Assessment Tool (SWAT) was established to forecast the variation in the quantity and quality of water, sediment yield, chemical and pollution level of water in different sources of water due to different types of water management practices across the globe. Due to its prediction efficiency this model has been used by all over the world for water management and various hydrological research applications. After the successful development of a SWAT model, it is calibrated with actual data and validated for future prediction. A number of methods are available for calibration and validation of the SWAT model. The Sequential Uncertainty Fitting-Version 2 (SUFI-2) algorithm is mostly used technique for the calibration and validation analysis of a SWAT model [5]. The established results of the SUFI-2 indicate the efficiency of SWAT model and integrated SWAT model helps in simulation of streamflow and sediment yield.

II. OBJECTIVE OF THE STUDY

The primary objectives of this study are: (1) to summarize the structure and principle of SWAT model, and (2) to analyse the existing problems in the development of SWAT model based on previous published literatures and the summary of the application status of SWAT model both at home and abroad in recent years.

The results of this study will help the beginners of hydrological modelling to thoroughly understand the SWAT model and its data inputs, structure and applications. It will also provide reference for further enhancement of the SWAT model.

ISSN No:-2456-2165

III. SWAT MODEL DESCRIPTION

SWAT model is developed by the USDA Agriculture Research Service (USDA-ARS), and it is a semi-distributed model of watershed scale and continuous time [6]. SWAT officially launched the original SWAT 94.2 version in the early 1990s, and has been constantly updated and upgraded till now. The latest version is SWAT 2012 (https://swat.tamu.edu/). The source code of SWAT model is available online for free. Functioning of SWAT model is based on daily time step. The main objective to develop SWAT was to predict the impact of various activities i.e. land use/ land cover change, climate change over the years on stream flow of gauged and ungauged watersheds and water quality parameters. SWAT simulation process entirely can be divided into two major parts the land surface which includes runoff and slope confluence and the water surface which contains concentration of channel. The SWAT input contains the meteorological data (Temperature, rainfall, windspeed etc.) and LULC map, slope map of the watershed. The SWAT model simulates infiltration, percolation, plant growth, runoff, and nutrient loads to enumerate the effects of climate change, land use and water management. The major advantage of SWAT model is that it proficient of constant simulation over a long period as much as 100 years. SWAT processes all the input data and determines the movement of water, sediment and other substances from the river network to the basin and the calculate the load [7,8].

In the SWAT modelling the bigger watersheds are further divided into small catchment areas which are known as sub basins and sub basins are further divided into smaller unit which have identical features are known as the Hydrologic Response Units (HRU's). Modified SCS curve number and Green-Ampt infiltration method are two methods which SWAT use for simulating the streamflow or surface runoff. Penman-Monteith method, Hargreaves method and Priestley-Taylor method are the three methods which are used by SWAT to calculate the potential evapotranspiration. Lateral flow is simulated by the kinematic storage model [10]. Shallow aquifer is used for simulation of the return flow. Water balance equation of SWAT model which directs the hydrological balance is expressed as:

$$SW_t = SW_0 + \sum_{i=1} \left(R_{day} - Q_{surf} - E_a - W_{sweep} - W_{gw}\right)$$

Where:
- SW_t: Final water content of the soil
- SW_0: Initial water content of the soil on i^{th} day
- Q_{surf}: i^{th} day surface runoff
- R_{day}: Precipitation i^{th} day precipitation
- E_a: i^{th} day evapotranspiration
- W_{gw}: i^{th} day return flow
- W_{seep}: Seepage water on day i
- All the parameters are in mm

The SCS curve number is defined with the subsequent equation:

$$Q = \frac{(P - I_a)^2}{(P - I_a) + S}$$

Where:
- Q: Runoff depth in mm
- P: Effective precipitation mm)
- S: Maximum potential retention
- I_a: Initial abstraction of water in (mm)

I_a is the function of Maximum potential retention

Therefore,
$I_a = \lambda S$
Where:
λ=0.2. Therefore, I_a=0.2 S
Hence, by integrating both Eqs. we have;

$$Q = \frac{(P - 0.2S)^2}{(P + 0.8S)}$$

The runoff occurs when the value of P is greater than 0.5 S. Maximum retention potential of S is accompanying to the unitless parameter SCS curve number (CN) using the following equation.

$$S = 25.4\left(\frac{1000}{CN} - 10\right)$$

The SCS curve number (CN) is subjected with permeability of soil, land use, soil–water conditions and infiltration. The SCS curve number value can be described by the three conditions (wet, moist, and dry). The SWAT output develop and validate various parameters i.e., streamflow, surface runoff, evapotranspiration, reservoir water balance, deep aquifer, interception storage and infiltration [11].

The model simulation results are generally evaluated using the regression correlation coefficient (R^2), Nash-Sutcliffe modelling efficiency (NSE). Generally, if R^2 >0.6, NSE>0.5, the simulation accuracy of the model meets the requirements, that is, the SWAT model is suitable to simulate the watershed [12].

ISSN No:-2456-2165

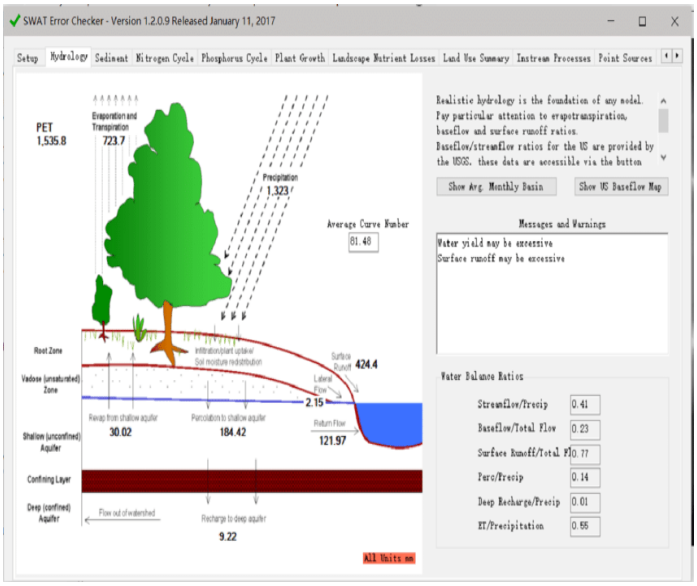


Figure 1. The hydrological cycle process of SWAT model.

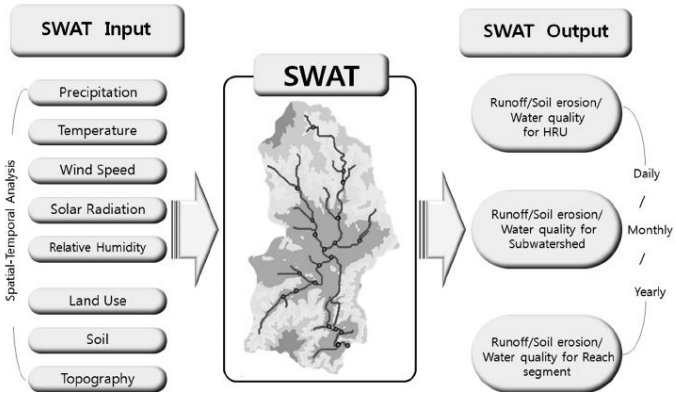


Figure 2: SWAT Model Input and Output

IV. RUNOFF SIMULATION

Runoff is the product of the combination of various natural and geographical factors such as climate and underlying surface in the basin. Runoff simulation is a significant part of hydrological simulation research and the basis for studying other hydrological problems [13]. The runoff simulation is mainly divided into two stages. The first part is to determine the parameters in the model based on historical data of a certain length, and the second one is to use the model to simulate the runoff of the verification period or predict the future runoff. The runoff simulation is mainly composed of surface runoff, evapotranspiration, soil water and groundwater [14]. The methods for calculating surface runoff are Soil Conservation Service Curve Number (SCS-CN) method and Green-Ampt infiltration method. Generally, the former is used more frequently. In the literature related to runoff simulation scholars applied SWAT to runoff simulation which are no longer just the simple application of the model in recent years. Foreign scholars are more concentrated on studies of influencing factors of simulation accuracy [15].

V. IMPACT OF CLIMATE CHANGE ON WATER RESOURCES

"Hydrological cycle research under changing environment (i.e. global climate change and human activity influence)" is the core scientific issue of Global Water Systems Programme (GWSP). Land Use and Land Cover (LULC) Change is an important manifestation of "changing environment" [16]. The combination of water cycle and the terrestrial carbon cycle is a new cross-direction. The scenario analysis research methods of LULC hydrological impacts usually include reference comparison method, historical inversion method, model prediction method, extreme land use method, and land use spatial allocation method. At present, many scholars have applied the SWAT model in different watersheds to analyse the impact of climate change and confirmed the adaptability and credibility of SWAT model [17].

VI. DISCUSSION

SWAT is widely used in many aspects of water cycle simulation, but it still has certain limitations. For the current era of big data, there are still some problems and improvements in the application of SWAT.

- Time scale problem

In the application process of SWAT model, most scholars used the year, season or month as the time scale for simulation and analysis, and rarely took day as the time step to simulate. The result is mostly consistent with that of previous study that SWAT can effectively simulate the hydrological process of long time series, but the simulation accuracy decreased with the shortening of time step, especially the simulation of daily runoff has systematic errors [18].

- Uncertainty analysis issues

SWAT was proposed by American scholars. The data used in the model, such as land use types and soil types, are all based on the localization of the United States. Chinese scholars need to establish attribute database according to the land use types and soil characteristics in the study area when applying the model, which is the main reason for the low simulation accuracy of the model for the research area with incomplete data. And the input of parameters also has some uncertainty on the simulation accuracy. At present, many scholars have done a lot of research on sensitivity analysis and uncertainty analysis of parameters [19].

- Model updating problem

Under changing environment, the relationship between rainfall and runoff is non-stationary, and real-time dynamic runoff simulation is required. The current SWAT model needs to use the latest data for calibration to make the prediction result more accurate when new data are added, which forms a mechanism of constraint and feedback between the latest data and future prediction data. But this method is time consuming and laborious. In the future research, an algorithm can be added on the basis of the SWAT model source code, so that the calibration process

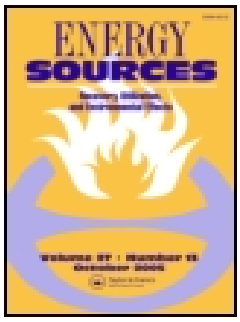
becomes a self-learning process of the model, and the model opens the self-learning mode while the latest data enters the database, which can be self-calibrated for more accurate simulation [20].

VII. CONCLUSIONS

The wide application of SWAT model at both domestic and abroad has confirmed that it is a powerful and comprehensive hydrological modelling tool. At present, domestic and foreign studies on this model mainly focus on model application, model accuracy and model coupling. The application field of the model are still mainly runoff simulation, climate change and water pollution. But the research content is no longer the simple application of the model. The main research contents of runoff simulation are the study on model accuracy, such as the influence of DEM resolution and other factors on the simulation accuracy, and the adaptability of the model in special areas. In order to adapt to the changing environment, SWAT model still has many directions for improvement, so as to better simulate the water cycle, serve water resources scheduling decisions, and provide support for water resources management.

REFERENCES

[1]. Abbaspour, K.C.; Faramarzi, M.; Ghasemi, S.S., and Yang, H., "Assessing the impact of climate change on water resources in Iran," *Water Resource Research*, 45(10),
[2]. Arnold, J.G.; Srinivasan, R.; Muttiah, R.S., and Williams, J.R., "Large-area hydrologic modeling and assessment: Part I model development", American Water Resources Association, 34, 73-89, 1998.
[3]. Gassman, Philip W., et al. "The soil and water assessment tool: historical development, applications, and future research directions." Transactions of the ASABE, 50.4, 1211-1250, 2007.
[4]. Alcamo, Joseph, Martina Flörke, and Michael Märker. "Future long-term changes in global water resources driven by socio-economic and climatic changes." *Hydrological Sciences Journal*, 52.2, 247-275, 2007.
[5]. Armitage, Peter, C. K. McPherson, and B. C. Rowe. "Repeated significance tests on accumulating data." *Journal of the Royal Statistical Society: Series A (General)*, 132.2, 235-244, 1969.
[6]. Baymani-Nezhad, M., and D. Han. "Hydrological modeling using effective rainfall routed by the Muskingum method (ERM)." *Journal of Hydro informatics*, 15.4, 1437-1455, 2013
[7]. Benham, Brian L., et al. "Modeling bacteria fate and transport in watersheds to support TMDLs." *Transactions of the ASABE*, 49.4, 987-1002, 2006.
[8]. Beven, Keith J. "*Rainfall-runoff modelling: the primer"*. John Wiley & Sons, 2011.
[9]. Beven, Keith. "Causal models as multiple working hypotheses about environmental processes." *Comptes rendus geoscience*, 344.2, 77-88, 2012.
[10]. Borah, D. K., et al. "Sediment and nutrient modeling for TMDL development and implementation." *Transactions of the ASABE*, 49.4, 967-986, 2006.
[11]. Castilla-Rho, Juan Carlos, et al. "Sustainable groundwater management: How long and what will it take?" *Global Environmental Change*, 58, 101972, 2019.
[12]. Chahine, Moustafa T. "The hydrological cycle and its influence on climate." *Nature*, 359.6394, 373-380, 1992.
[13]. Chambel-Leitão, P., et al. "Coupling SWAT and tempQsim Mohid River Network." *Geophysical Research Abstracts*, Vol. 8., 2006.
[14]. Hargreaves, George H., and Zohrab A. Samani. "Reference crop evapotranspiration from temperature." *Applied engineering in agriculture*, 1.2, 96-99, 1985.
[15]. Jain, Sharad K., and V. P. Singh. "Hydrological cycles, models and applications to forecasting." *Handbook of Hydrometeorological Ensemble Forecasting*, 2017.
[16]. Kumar, Rakesh, R. D. Singh, and K. D. Sharma. "Water resources of India." *Current science*, 794-811, 2005.
[17]. Mausbach, Maurice J., and Allen R. Dedrick. "The length we go measuring environmental benefits of conservation practices.", 2004.
[18]. Neitsch, Susan L. "Soil and water assessment tool." *User's Manual Version 2005*, 476, 2005.
[19]. Wani, Suhas P., and Kaushal K. Garg. "Watershed management concept and principles." 1-11, 2019.
[20]. Xu, Jianhua, et al. "Climate change and its effects on runoff of Kaidu River, Xinjiang, China: a multiple time-scale analysis." *Chinese Geographical Science*, 18.4, 331-339, 2008



Energy Sources, Part A: Recovery, Utilization, and Environmental Effects

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ueso20

Solution of optimal power flow problem using sine-cosine mutation based modified Jaya algorithm: a case study

Saket Gupta, Narendra Kumar & Laxmi Srivastava

To cite this article: Saket Gupta, Narendra Kumar & Laxmi Srivastava (2021): Solution of optimal power flow problem using sine-cosine mutation based modified Jaya algorithm: a case study, Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, DOI: 10.1080/15567036.2021.1957043

To link to this article: https://doi.org/10.1080/15567036.2021.1957043

Published online: 09 Sep 2021.

Submit your article to this journal

Article views: 15

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=ueso20

ENERGY SOURCES, PART A: RECOVERY, UTILIZATION, AND ENVIRONMENTAL EFFECTS
https://doi.org/10.1080/15567036.2021.1957043

Taylor & Francis
Taylor & Francis Group

Check for updates

Solution of optimal power flow problem using sine-cosine mutation based modified Jaya algorithm: a case study

Saket Gupta [a], Narendra Kumar [a], and Laxmi Srivastava [b]

[a]Department of Electrical Engineering, Delhi Technological University, Delhi, India; [b]Department of Electrical Engineering, Madhav Institute of Technology & Science, Gwalior, India

ABSTRACT

For planning the economic operation and control of an existing electricity grid and its future extension planning, the optimal power flow (OPF) results are essential. In this paper, a hybrid algorithm is proposed, which is based on a sine-cosine mutation operator and a modified Jaya (SCM-MJ) algorithm to solve the OPF problems. The efficacy of the SCM-MJ algorithm is primarily evaluated using thirteen (unimodal and multimodal) mathematical benchmark functions. Later, the SCM-MJ algorithm is applied on the Algerian 59-bus system and IEEE 118-bus test system to handle the OPF problems. The SCM-MJ has successfully achieved a minimum value of an objective function over several runs than other modern meta-heuristic optimization approaches in all the thirteen mathematical benchmark functions as well as in OPF case studies. For example, in the Algerian 59-bus system, the quadratic fuel cost obtained by SCM-MJ is 1688 $/hour, with a saving of 13.12% of the initial fuel cost. This benefit increases further with the size of the system. The SCM-MJ algorithm has provided high-quality solutions for mathematical benchmark functions and OPF problems quickly and efficiently. The comparison of numerical outcomes demonstrates that the suggested SCM-MJ algorithm dominates over other approaches for solving the OPF problem.

ARTICLE HISTORY
Received 14 May 2021
Revised 9 July 2021
Accepted 10 July 2021

KEYWORDS
Meta-heuristic; modified jaya algorithm; opf; SCM-MJ algorithm; sine-cosine mutation

Introduction

The incorporation of various types of loads such as Electric Vehicles and Renewable Energy Sources in the sub-transmission and distribution system have increased the uncertainty in load pattern in the modern restructured electrical grid. Therefore, the importance of solving the OPF problem has increased many folds. The OPF is an essential tool to find the control variables setting, which can minimize a pre-specified objective function while all the constraints are satisfied and power flows are within their operating limits. OPF is a complex constrained optimization problem, which includes a large number of complex constraints and decision variables (Abido 2002).

In the early decades, many classical methods (Momoh, Adapa, and El-Hawary 1999) like quadratic programming, interior point method, linear programming, non-linear programming and mixed-integer non-linear programming (Fortenbacher and Demiray 2019; Jiang et al. 2010), etc. have been applied to deal with the OPF problem, but these methods have certain limitations, for example, a requirement of initial or starting point (except for linear programming and convex optimization). Except for hill-climbing with random restart, almost all the conventional algorithms are deterministic. For the same starting point, the final solutions would be similar. In other words, the conventional optimization methods are less efficient, particularly, when the constraints and objective functions are non-linear, non-convex, and having multiple local optima (Srivastava and Singh 2015).

CONTACT Saket Gupta sguptamits@gmail.com Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India

© 2021 Taylor & Francis Group, LLC

To overcome the demerits of classical optimization methods and to handle such difficulties, various meta-heuristic algorithms came into being as alternatives to the classical optimization methods. As a result, meta-heuristic algorithms have recently got a lot of attention and are being applied to solve various real-world optimization problems in Science and Engineering fields. Some of these are Genetic-algorithm (Devaraj and Yegnanarayana 2005), Particle swarm optimization (PSO) (Abido 2002), Differential Evolution (DE) (Varadarajan and Swarup 2008), Gravitational search algorithm (GSA) (Bhattacharya and Roy 2012), Moth swarm algorithm (MSA) (Mohamed et al. 2017), black-hole-based optimization (BHBO) approach (Bouchekara 2014a), Salp swarm optimizer algorithm (SSA) (El-Fergany and Hasanien 2020), League championship algorithm (LCA) (Bouchekara et al. 2014b), teaching learning-based optimization (TLBO) (Farahani, Aghaei, and Rashidi 2018).

Adhvaryyu and Adhvaryyu (2021) presented the Krill Herd algorithm to solve static optimal load flow problem. Ajayi, Nwulu, and Damisa (2020) applied two recently developed optimization techniques namely, Ant Lion Optimizer and Exchange Market Algorithm to solve the DC-OPF problem. Kotb and El-Fergany (2020) used the Moth Swarm Optimizer to solve optimal power flow problem. Jamal, Men, and Khan (2020) solved the Optimal Reactive Power Dispatch problem using the gray wolf optimizer (GWO) algorithm.

When used to solve complex real-life engineering optimization problems, standard versions of some of the more common meta-heuristic approaches have been found to have some limitations. For example, premature convergence or local optima trapping is a common occurrence in GA and moth search optimization (MSO) algorithms. Similarly, the simulated annealing (SA) and PSO algorithms are relatively ineffective in searching optimal global solutions (Singh et al. 2020). In addition, poor communication in the TLBO algorithm during the second phase (Learning Phase) may result in insufficient knowledge sharing, therefore may get trapped into the local solution. Furthermore, the Jaya algorithm has a strong capacity to explore the search space for optimization, but occasionally it suffers from premature convergence (Elattar and ElSayed 2019). The Stud Krill Herd (SKH) algorithm (Pulluri, Naresh, and Sharma 2017a) was used to overcome the sluggish convergence of the Krill Herd (KH) algorithm and to find an optimum solution to solve the OPF problem. Various modifications and hybridization of meta-heuristic algorithms have been proposed in the literature to address the shortcomings of the poorly performing standard versions of meta-heuristic approaches.

To enhance the global search ability for solving the OPF problem, many improved variants have been explored in recent years. Some of these are, Quasi-oppositional TLBO (Mandal and Roy 2014), Novel improved social spider optimization algorithm (Nguyen 2019), Improved evolutionary programming (Ongsakul and Tantimaporn 2006), Enhanced genetic algorithm (Bakirtzis et al. 2002), Adaptive Biogeography based Predator–Prey Optimization (Christy and Raj 2014), Levy mutation TLBO algorithm (Ghasemi et al. 2015), G-best guided ABC algorithm (Jadhav and Bamane 2016), Improved group search optimization (Tan et al. 2015), Enhanced self-adaptive differential evolution (ESDE) (Pulluri, Naresh, and Sharma 2017b), Modified Sine Cosine Algorithm (MSCA) (Attia, Sehiemy, and Hasanien 2018), Particle swarm optimization with an aging leader and challengers algorithm (ALC-PSO) (Singh, Mukherjee, and Ghoshal 2016), Adaptive particle swarm optimization (Niknam et al. 2012), Adaptive real coded biogeography-based optimization (Kumar and Premalatha 2015), Improved moth-flame optimization (IMFO) algorithm (Taher et al. 2018).

A hybrid meta-heuristic algorithm is the most recent research trend to solve practical optimization problems. Recently, meta-heuristic algorithms hybridization has become more popular because of its improved ability to deal with complex optimization problems. In recent literature, a large number of hybrid meta-heuristics algorithms have been proposed to solve complex optimization problems successfully. Ongsakul and Bhasaputra (2002) used a hybrid tabu search and simulated annealing method to solve the OPF problem incorporating the FACTS device. Joorabian and Afzalan (2014) have applied a hybrid fuzzy-based particle swarm optimization and Nelder–Mead algorithm to solve the OPF problem.

Ghasemi et al. (2014) solved the OPF problem using a hybrid imperialist competitive algorithm (ICA) and teaching-learning algorithm (TLA). The OPF problem with uncertainty was solved using

a hybrid fuzzy dependent PSO technique presented by Liang et al. (2011). To check the efficacy of the algorithm, it has been applied on the IEEE 30-bus and IEEE 118-bus test systems to solve OPF problems including wind energy systems.

Niknam, Narimani, and Jabbari (2013) solved the dynamic optimal power flow problem using a hybrid particle swarm optimization and simulated annealing approach that considers the ramp rate constraints, valve-point loading effects, and prohibited zones. Mahdad and Srairi (2015) presented the first stage of the practical approach using Grey Wolf Optimizer integrated with the local search (pattern search) algorithm to solve the blackout risk prevention under the critical situation of the power system. Reddy (2019) proposed hybrid differential evolution and harmony search algorithm to solve the OPF problem. Radosavljevic et al. (2015) applied particle swarm optimization and GSA to find the solution to the OPF problem. The hybrid method utilized the social thinking of the PSO algorithm and the exploitation quality of GSA. Gacem and Benattous (2017) solved the OPF problem using hybrid GA and PSO algorithms. Khelifi, Bentouati, and Chettih (2019) proposed a hybrid Firefly Krill Herd method to solve the OPF problem. To improve the local search performance of the hybrid algorithm, a light intensity operator was employed in the Krill Herd algorithm. Azizivahed et al. (2020) suggested a dynamic economic dispatch problem using hybrid GWO and PSO algorithms. Naderi et al. (2021) presented a solution to the OPF problem using self-adaptive PSO and DE algorithms.

Niknam, Narimani, and Abarghooee (2012) presented a hybrid shuffled frog leaping algorithm incorporating simulated annealing (SA) to find the solution OPF problem. The proposed hybrid algorithm solved the OPF problem taking into account practical issues of generator units like prohibited operating zones and valve loading effects. The OPF problem was solved using an evolving ant direction hybrid differential evolution algorithm by Vaisakh and Srinivas (2011). Mahdad and Srairi (2014) presented differential evolution (DE), adaptive particle swarm optimization (APSO), and pattern search (PS) for the planning of power systems under normal and critical loading conditions. Kaur and Narang (2020) proposed a hybrid invasive weed optimization method to solve the OPF problem. The proposed hybrid method is a combination of Powell's pattern search and invasive weed optimization algorithm. Abbasi, Abbasi, and Ivatloo (2021) solved the OPF problem using a differential-based harmony search algorithm.

Despite their advantages, meta-heuristic algorithms have some drawbacks. To find the near-global optimal solution, they require parameters tuning. The parameters' tuning needs multiple trials and hence takes a long time to get the optimal solution. Moreover, the best solutions achieved by such algorithms cannot be replicated exactly thus several trials should be performed to ensure accuracy and meaningful statistical results. From that context, the Jaya algorithm (Rao 2016) is an algorithm-specific parameter-less optimization algorithm developed by Rao. The main advantage of the Jaya algorithm is that no algorithm-specific parameters are required in it and thus it is simple to implement this algorithm for solving various kinds of optimization problems. The Jaya algorithm has a strong capacity to explore search space for optimization, but sometimes it is suffering from premature convergence and it might be stuck at a sub-optimal solution. To alleviate this problem of the classical Jaya algorithm and make it more effective, the SCM-MJ algorithm which combines sine-cosine mutation strategy and modified Jaya algorithm is proposed in this paper.

The contribution of the paper can be summarized as follows:

- A novel hybrid SCM-MJ algorithm has been developed, which is a combination of the modified Jaya and sine-cosine mutation operator strategy. The proposed hybrid algorithm can address the limitations of the traditional Jaya algorithm such as sluggish convergence to the near-global optimum and to avoid premature convergence problems.
- The SCM-MJ method offers flexibility to balance the exploration and exploitation capability of the algorithm during the search process. The sine-cosine mutation operator is capable to maintain the diversity of the solutions throughout the search to avoid local optimal solutions, while modified Jaya can explore the search space very well.

- The novel hybrid SCM-MJ algorithm can deal with complex constrained optimization problems very well and it is successfully applied for the first time to solve the OPF problem.
- The SCM-MJ algorithm is formulated for solving the OPF problem on the Algerian 59-bus and IEEE 118-bus system. The Algerian 59-bus system is a practical power system, which is considered to demonstrate the validity of the SCM-MJ algorithm to solve a practical power system problem. In addition, the IEEE 118-bus system is considered in this paper to validate the scalability of the proposed algorithm i.e. capability to handle large dimensional OPF problems, as in this system the numbers of control variables are more than 130.
- It can be noted that in all the benchmark functions (Mirjalili 2015) as well as in all the case studies of OPF problems, the proposed SCM-MJ algorithm provided the lowest value of objective functions consistently. Also, SCM-MJ offered the least standard deviation over several runs in comparison with other efficient optimization algorithms presented in recent literature.
- Based on the OPF results and statistical analysis, the proposed SCM-MJ algorithm is found to be more efficient (in terms of solution quality) and robust over the modified Jaya algorithm and other meta-heuristic algorithms reported in recent literature.

The main objective of this research work is to solve the OPF problem considering technical and economical objective functions namely, fuel cost minimization, voltage profile improvement, and real power loss minimization. The effectiveness of the proposed SCM-MJ algorithm is examined by applying the SCM-MJ algorithm to solve various unimodal and multimodal mathematical benchmark functions as well as OPF problems. Based on computational analysis, the proposed SCM-MJ algorithm is found to offer better performance in comparison with other reported algorithms. The remaining of the paper is structured as follows. The objective functions for the OPF problem are described in Section-2. Section-3 provides details of the modified Jaya and sine-cosine mutation operator. Section-4 describes the SCM-MJ algorithm in detail. The numerical outcomes attained by the SCM-MJ algorithm are given in section-5, while sections 6 and 7 include the statistical analysis and discussion. The conclusion is given in the last section.

Objective functions

In this paper three objective functions are taken into account. This section represents details of these three objective functions.

Fuel cost minimization (FCM)

The minimization of total fuel cost is selected as main objective in this case. This objective function can be described as (Singh et al. 2020):

$$Z_{FCM}(u,x) = \sum_{i=1}^{NGN} f\left(P_{g_i}\right)(\$/h) = \sum_{i=1}^{NGN} A_i + B_i P_{g_i} + C_i P_{g_i}^2 (\$/h) \tag{1}$$

Eq. (2) and (3) represent control variables and dependent variables vectors:

$$u^T = \left[P_{g_2} \ldots P_{g_{NGN}}, V_{g_1} \ldots V_{g_{NGN}}, Q_{C_1} \ldots Q_{C_{NC}}, T_1 \ldots T_{NTR}\right] \tag{2}$$

$$x^T = \left[P_{g_{Slack}}, V_1 \ldots V_{NLB}, Q_{g1}, \ldots Q_{gNGN}, S_1, \ldots S_{Ntl}\right] \tag{3}$$

Voltage profile improvement (VPI)

The voltage magnitude at a load bus must not deviate much from 1.0 pu to improve the voltage profile. Thus, the main motive, in this case, is to minimize voltage variation from 1.0 pu at all the load buses. The objective function can be expressed in the present case as (Singh et al. 2020):

$$Z_{VPI}(u,x)=\sum_{i\epsilon NLB}|V_i-1| \tag{4}$$

The minimizations of fuel cost and voltage profile are simultaneously taken into account by using a weighted sum method as follows:

$$COF = Z_{VPI}(u,x) = Z_{FCM}(u,x) + W_{VPI} \times Z_{VPI}(u,x) \tag{5}$$

Where W_{VPI} is a weighting factor. In this case, W_{VPI} is set equal to 260.

Real power losses minimization (RPLM)

For a given electrical network, real power losses can be calculated using the following expression (Singh et al. 2020):

$$Z_{Loss}(u,x) = \sum_{i=1}^{NB} P_i = \sum_{i=1}^{NB} P_{gi} - \sum_{i=1}^{NB} P_{di} \tag{6}$$

The minimization of fuel cost and real power losses are simultaneously taken into account by using a weighted sum method as follows:

$$COF = Z_{VPI}(u,x) = Z_{FCM}(u,x) + W_{RPLM} \times Z_{Loss}(u,x) \tag{7}$$

In this case, W_{RPLM} is set equal to 10.

Constraints

Two kinds of constraints are part of the OPF problem: equality and inequality constraints. The equality constraints consist of power flow equations. The power flow equations are represented as follows:

$$\sum_{i=1}^{NB} P_{gi} = P_d + P_{Loss} \tag{8}$$

$$\sum_{i=1}^{NB} Q_{gi} = Q_d + Q_{Loss} \tag{9}$$

The inequality constraints are represented as the operating limits of various power system equipment, which are defined as follows:

$$P_{g_k}^{min} \le P_{g_k} \le P_{g_k}^{max} \; k = 1 \ldots .NGN \tag{10}$$

$$V_{g_k}^{min} \le V_{g_k} \le V_{g_k}^{max} \; k = 1 \ldots ..NGN \tag{11}$$

$$Q_{g_k}^{min} \le Q_{g_k} \le Q_{g_k}^{max} \; k = 1 \ldots .NGN \tag{12}$$

$$T_k^{min} \le T_k \le T_k^{max} \; k = 1 \ldots .NTR \tag{13}$$

$$Q_{C_k}^{min} \le Q_{C_k} \le Q_{C_k}^{max} k = 1 \ldots . NGN \tag{14}$$

$$V_{L_k}^{min} \le V_{L_k} \le V_{L_k}^{max} k = 1 \ldots . NLB \tag{15}$$

$$S_{l_k} \le S_{l_k}^{max} k = 1 \ldots \ldots ntl \tag{16}$$

Constraints handling

The inequalities are incorporated into the objective function to find a feasible solution and the augmented objective function can be defined as:

$$Z_{aug} = Z(.) + R_1 \left(P_{G_{Slack}} - P_{G_{Slack}}^{lim} \right)^2 + R_2 \sum_{i=1}^{NGN} \left(Q_{g_k} - Q_{g_k}^{lim} \right)^2 + R_3 \sum_{i=1}^{NLB} \left(V_{L_k} - V_{L_k}^{lim} \right)^2 + R_4 \sum_{i=1}^{Ntl} \left(S_{l_i} - S_{l_i}^{lim} \right)^2 \tag{17}$$

In this paper, R_1, R_2, R_3, and R_4 are all set equal to 10^5.

Modified Jaya Algorithm

Jaya algorithm is a comparatively new meta-heuristic optimization algorithm developed by Rao (Elattar and ElSayed 2019; Rao 2016). The working principle of the Jaya algorithm is that the numerical solution that has been obtained should go toward the best solution and should avoid the inferior solutions for a particular optimization problem. The key benefit of the Jaya algorithm is that algorithm-specific parameter tuning is not required at all in it and thus it is easy to implement this algorithm for solving various kinds of optimization problems. The modified Jaya algorithm has been derived by changing the solution update equation of the Jaya algorithm using (18).

$$J_{i+1,j,k} = J_{i,j,k} + \gamma_{i,j,1} \times \left(J_{i,j,B} - J_{i,j,k} \right) - \gamma_{i,j,2} \times S \times \left(J_{i,j,W} - J_{i,j,k} \right) \tag{18}$$

Where $J_{i,j,k}$ represents the value of the j^{th} variable for the k^{th} candidate during the i^{th} iteration. $J_{i,j,W}$ and $J_{i,j,B}$ are the worst candidate and best candidate value of the j^{th} variable respectively.$\gamma_{i,j,1}$and $\gamma_{i,j,2}$ (*rand2*) are the two random numbers for the *j*th variable during the *i*th iteration in the range [0, 1].

where *S* is updated in each iteration as follows;

$$S = \begin{Bmatrix} 1 & If\ r\ and\ 2 > 0.5 \\ -1 & else \end{Bmatrix} \tag{19}$$

Sine-Cosine Mutation Operator

Zhou et al. (2017) suggested a sine-cosine mutation operator that enhances the global search ability of meta-heuristic techniques. The key benefit of the sine-cosine mutation operator is that it avoids the loss of diversity of populations during the search process. The sine-cosine mutation operator typically constructs new offspring (solutions) similar to the parent candidate solutions. When the sine-cosine mutation operator is used during the search process, smaller steps are taken, allowing the candidate solutions to explore every corner of the solution space. The sine-cosine mutation operator is capable of diversifying the population and making global searches more effective so that the objective function does not converge prematurely with local optima. The sine-cosine mutation operator function is calculated as follows:

$$J_{new(i+1,j,k)} = \begin{cases} J_{i,j,k_{c1}} + A \times sin(B) \times \left| C.J_{i,j,k_{c2}} - J_{i,j,k_{c3}} \right| & rand < 1.0 \\ J_{i,j,k_{c1}} + A \times cos(B) \times \left| C.J_{i,j,k_{c2}} - J_{i,j,k_{c3}} \right| & otherwise \end{cases} \quad (20)$$

Eq. (20) is used to achieve the k^{th} candidate's new position value$J_{new(i+1,j,k)}$ during $(i+1)^{th}$ iteration. In Eq. (20), $J_{i,j,k_{c1}}$, $J_{i,j,k_{c2}}$ and $J_{i,j,k_{c3}}$ are three random solutions in the i^{th} iteration, which are different from one another. The parameter A is changed iteratively as determined by Eq. (21), whereas B and C are random numbers in the range [0, 2π] and [0, 2] respectively.

$$A = 2 \times (1 - iter/iter_{max})^2 \quad (21)$$

To maintain a constant population size in next generations, sine-cosine mutation uses a greedy selection procedure. The new position value of $J_{new(i+1,j,k)}$ replaces the old value $J_{(i,j,k)}$ if and only if $f(J_{new(i+1,j,k)}) < f(J_{(i,j,k)})$.

The proposed methodology

Jaya algorithm is an efficient optimization algorithm. This algorithm is good in the exploration of the search space but is slow in the exploitation part. Therefore, hybridizing Jaya with an algorithm having strong exploitation capability might balance the exploitation and exploration of the Jaya algorithm. Hence, to boost the operational efficacy of the Jaya algorithm, a sine-cosine mutation operator has been incorporated in it. The motivation to incorporate the sine-cosine mutation operator with the Jaya algorithm is to combine the benefits of both of them.

The SCM-MJ algorithm can maintain the population's diversity throughout the search space to avoid sub-optimal solutions, enhance the convergence speed, and find near-global solutions as well. The hybridization of the modified Jaya algorithm and sine-cosine mutation operator would lead to better results for real-world complex, constrained and high dimensional optimization problems. The flow chart of the proposed hybrid method is given in Figure 1.

The main computational step for solving OPF using the SCM-MJ algorithm can be summarized as follows:

Step 1: Initialize population 'P' having control variables, with the dimension of problem 'D' and set the stopping criteria.

Step 2: For each individual, run NRLF (Newton Raphson load flow) program and evaluate the value of the augmented objective function (18)

Step 3: Set iteration *Iter* =0.

Step 4: Identify the worst and best solution in the population based on the value of the augmented objective function (18).

Step 5: Modify the solution based on the worst and the best solutions using (19).

Step 6: For each individual, is the new fitness value better than the previous one? If yes, then replace the previous solution with a new solution. Otherwise, keep the previous solution.

Step 7: Apply sine-cosine mutation operator to update the position of all the members of the population and calculate the augmented objective function value (18).

Step 8: If a new solution achieved after the mutation operator is better than the previous solution, accept the new solution otherwise keep the previous solution.

Step 9: If stopping criteria is satisfied then go to step 10, otherwise go to step 4 and increase the iteration count by 1, i.e. *Iter* = *Iter* + 1.

Step 10: Stop and display optimal solutions.

Case studies

To check the efficacy of the proposed SCM-MJ algorithm, it has been applied to 13 standard mathematical benchmark functions. Table 1 represents the algorithm specific parameters of the SCM-

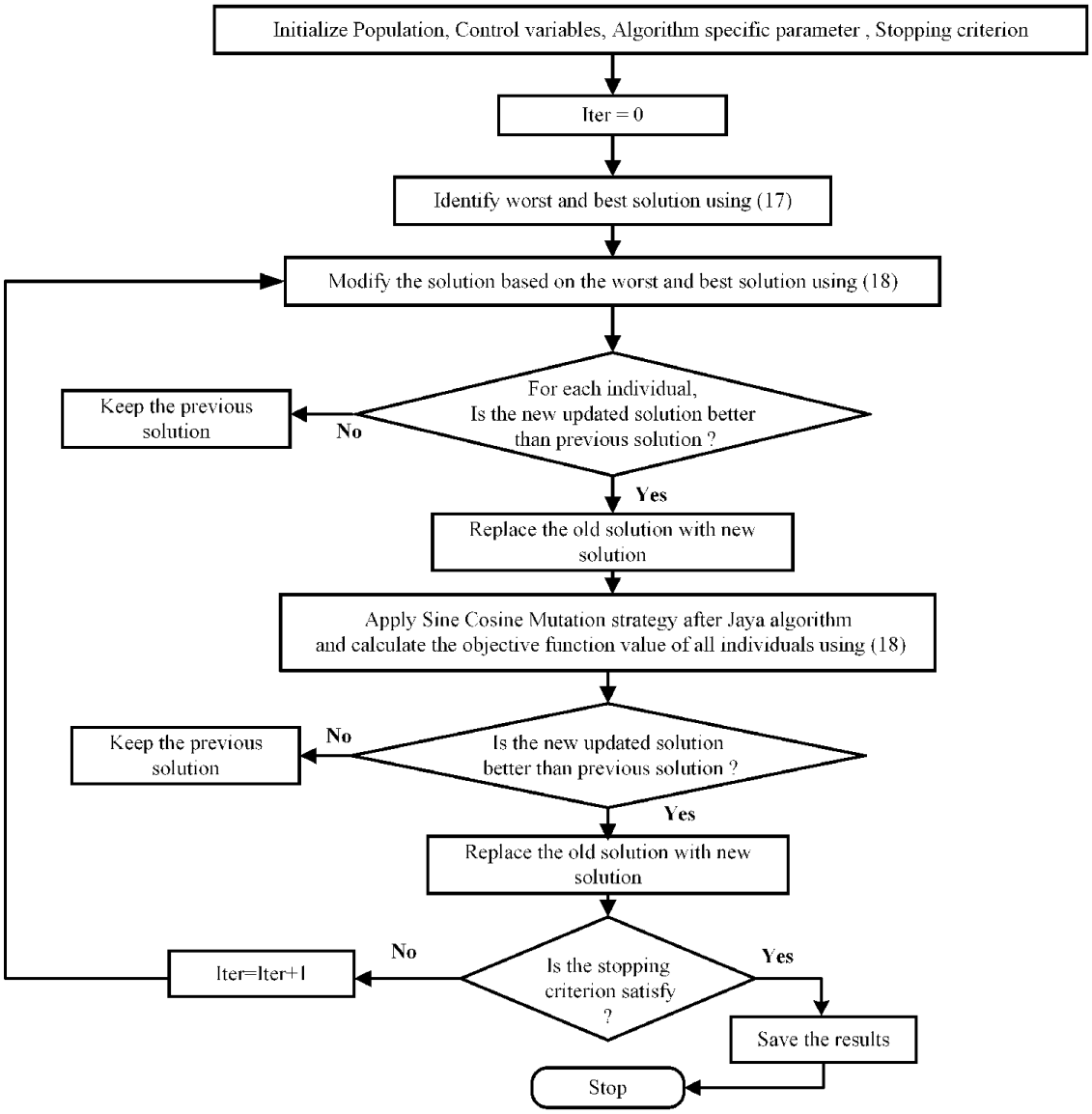


Figure 1. Flowchart of proposed SCM-MJ algorithm.

Table 1. Algorithm specific parameters setting of SCM-MJ algorithm.

S. No.	Parameter	Value
1	Population size	100 (for benchmark function), 30 (for OPF problem)
2	Maximum Iteration	2500 (for benchmark function), 150 (for OPF problem)
3	Number of function Evolution	(2,50,000 × M-Jaya+2,50,000× SCM) = 5,00,000 (for benchmark function) (45,00 × M-Jaya +45,00× SCM) = 9,000 (for OPF problem)

MJ algorithm. Details of these benchmark functions and simulation outcomes can be seen in Table 2 and Table 3 respectively.

The performance of the proposed SCM-MJ is compared with M-Jaya algorithm, ALO, BA, CS, FPA, FA, GA, PSO, and SMS algorithms (Mirjalili 2015). Out of various evolutionary computing (EC) based algorithms, the PSO and GA are the old but well-established and well-known optimization techniques. This is interesting to note that most of the recently developed swarm and evolutionary

Table 2. Mathematical benchmark functions (Mirjalili 2015).

S. no	Formulation	Dim	Search range	fmin
Unimodal benchmark functions				
1	$F_1(x) = \sum_{i=1}^{n} x_i^2$	200	[−100, 100]	0
2	$F_2(x) = \sum_{i=1}^{n} \lvert x_i \rvert + \prod_{i=1}^{n} \lvert x_i \rvert$	200	[−10, 10]	0
3	$F_3(x) = \sum_{i=1}^{n} \left(\sum_{j-1}^{i} x_j \right)^2$	200	[−100, 100]	0
4	$F_4(x) = max_i\{\lvert x_i \rvert, 1 \le i \le n\}$	200	[−100, 100]	0
5	$F_5(x) = \sum_{i=1}^{n-1} \left[100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2 \right]$	200	[−30, 30]	0
6	$F_6(x) = \sum_{i=1}^{n} (\lvert (x_i + 0.5) \rvert)^2$	200	[−100, 100]	0
7	$F_7(x) = \sum_{i=1}^{n} i x_i^4 + random[0, 1)$	200	[−1.28, 1.28]	0
Multimodal benchmark functions				
8	$F_8(x) = \sum_{i=1}^{n} -x_i \sin\left(\sqrt{\lvert x_i \rvert}\right)$	200	[−500, 500]	$-418.982 \times Dim$
9	$F_9(x) = \sum_{i=1}^{n} \left[x_i^2 - 10\cos(2\pi x_i) + 10 \right]$	200	[−5.12, 5.12]	0
10	$F_5(x) = -20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} \right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n} \cos(2\pi x_i) \right) + 20 + e$	200	[−32, 32]	0
11	$F_6(x) = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$	200	[−600, 600]	0
12	$F_7(x) = \frac{\pi}{n}\left\{ 10\sin(\pi y_1) + \sum_{i=1}^{n-1} (y_i - 1)^2\left[1 + 10\sin^2(\pi y_{i+1})\right] + (y_n - 1)^2 \right\} + \sum_{i=1}^{n} u(x_i, 10, 100, 4) y_i = 1 + \frac{x_i+1}{4}$ $u = (x_i, a, k, m) = \begin{cases} k(x_i - a)^m x_i > a \\ 0 - a < x_i < a \\ k(-x_i - a)^m x_i < a \end{cases}$	200	[−50, 50]	0
13	$F_8(x) = 0.1\left\{ \sin^2(3\pi x_1) + \sum_{i=1}^{n} (x_i - 1)^2[1 + \sin^2(3\pi x_i + 1)] + (x_n - 1)^2[1 + \sin^2(2\pi x_n)] \right\} + \sum_{i=1}^{n} u(x_i, 5, 100, 4)$	200	[−50, 50]	0

Table 3. Performance comparison of SCM-MJ algorithm with other algorithms.

	SCM-MJ		**M-Jaya**		**ALO** (Mirjalili 2015)		**PSO** (Mirjalili 2015)		**SMS** (Mirjalili 2015)	
Function	**ave**	**std**	**ave**	**std**	**ave**	**std**	**ave**	**std**	**ave**	**std**
F1	0.0000	0.0000	0.0088	0.0094	7.8e-07	1.1e-07	23.799	11.721	1039.2	0.4243
F2	0.0000	0.0000	7.6897	1.2159	530.82	222.67	237.87	22.432	1832.4	0.0122
F3	0.0000	0.0000	1.6e+05	1.4e+05	2331.4	507.18	4693.3	503.57	2034.8	0.3780
F4	0.0000	0.0000	30.013	1.5509	30.58	1.1446	40.111	0.5879	300.26	0.0023
F5	188.71	0.2105	627.31	12.737	167.04	49.746	911.23	95.245	3863.5	0.5329
F6	2.5 e-07	1.3 e-07	0.0455	0.0601	7.6e-07	7.3e-08	43.421	14.206	2494.4	0.0003
F7	1.8 e-05	1.0 e-05	0.3724	0.0689	0.0505	0.0144	17.321	4.0133	28.359	1.99e-05
F8	−50837	2395.9	−22549	14867	−44,426	1442.5	−18,136	4962.4	−35,969	0.8765
F9	0.0000	0.0000	1445.3	424.37	613.89	66.795	748.58	24.301	480.01	0.2365
F10	8.8e-16	0.0024	2.3669	1.2515	2.3058	0.2554	15.183	0.5762	17.293	0.0974
F11	0.0000	0.0000	0.0200	0.0391	0.0074	0.0065	3241.2	137.49	4801.5	0.8532
F12	0.7421	2.1e-02	6.4463	1.6378	5.3982	0.5959	4.0e+05	4.7e+05	1.0e+08	1.9e-05
F13	19.772	2.3e-04	154.35	21.675	0.1391	0.2219	1.2e+06	5.8e+05	1.0e+08	1.9e-05
	BA (Mirjalili 2015)		**FPA** (Mirjalili 2015)		**CS** (Mirjalili 2015)		**FA** (Mirjalili 2015)		**GA** (Mirjalili 2015)	
	ave	**std**	**ave**	**std**	**ave**	**std**	**ave**	**std**	**ave**	**std**
F1	1117.3	20731	55. 989	32. 678	3.8e-05	1.8e-05	76.128	1.5744	227.75	186.56
F2	3842.8	468.28	280.6	6.9384	400.10	0.8656	611.19	71.219	6322.6	1092.7
F3	1090.7	475.06	24219	8540	12,957	633.75	14852	6418.4	11,206	3986.1
F4	65.667	2.8293	37.689	2.4572	30.936	1.6877	2.736	0.5472	101.54	2.5321
F5	1410.8	591.07	3150.7	1490.6	332.67	159.88	1321.7	114.76	964.49	748.76
F6	51.205	12.005	166.99	41.109	8.1e-05	4.5e-05	78.42	2.3405	482.56	278.61
F7	2.4344	0.1275	4.8391	1.5354	0.4013	0.0087	0.0273	0.0041	116.56	60.161
F8	−25632	869.47	−45771	3097.8	−52600	156.04	−39,753	649.69	−28,660	1011.0
F9	723.38	100.96	702.95	69.653	541.58	41.889	475.45	28.058	1645.8	37.155
F10	18.159	0.0677	17.544	0.1668	17.654	2.9820	2.4297	0.0385	20.361	0.1425
F11	4937.0	268.42	180.74	36.084	0.0011	0.0011	1.7048	0.0143	3306.8	113.30
F12	1.6e+09	4.2e+08	4.3e+07	3.2e+07	1.0e+10	0.0045	23.426	0.5598	8.1e+09	9.5e+08
F13	2.2e+09	8.8e+08	9.8e+07	3.8e+07	1.0e+10	0.0568	2.8614	0.0568	1.3e+10	1.4e+09

computing-based algorithms are derived from the already established GA or PSO-based methods (Radosavljevic 2018).

To compare the performance of the proposed SCM-MJ with M-Jaya algorithm, ALO, BA, CS, FPA, FA, GA, PSO, and SMS algorithms (Mirjalili 2015), these algorithms were applied and run with the population size of 100 for the 5000 iterations on all the 13 benchmark functions. The number of functions evolution (NFE) for ALO, BA, CS, FPA, FA, GA, PSO, and SMS algorithms, proposed SCM-MJ and M-Jaya algorithm are also same which is equal to 5,00,000. Details of the implementation of hybrid SCM-MJ and M-Jaya algorithms are given in Table 1.

SCM = Sine-cosine mutation operator; M-Jaya = Modify Jaya algorithm

Ant Lion Optimizer (ALO): States of Matter Search (SMS): Bat Algorithm (BA): Flower Pollination Algorithm (FPA): FA = firefly algorithm: Cuckoo Search (CS) algorithm

The SCM-MJ algorithm is applied subsequently on the two systems (Algerian 59-bus and IEEE 118-bus) to solve the OPF problem. Several trials were taken, but the best results obtained and given in this paper are with population size 30 and the maximum number of iterations 150 for both the power systems. The simulation outcomes obtained by the SCM-MJ algorithm are compared to other meta-heuristic algorithms and it is observed that the SCM-MJ algorithm produces better results than its competitors for mathematical benchmark functions as well as for OPF problems. The computational work was carried out on a personal computer having a 1.7 GHz Intel Processor, 4GB RAM, Core i3, and 64-bit operating system using MATLAB-13a computing environment. Several case studies were carried out in this paper and are given in Table 4.

Algerian 59-bus system

To evaluate the effectiveness of the proposed SCM-MJ algorithm, it is tested for solving the OPF problem in the Algerian 59-bus system. The system data, maximum and minimum limits of the control variable are the same as provided in (Bouchekara et al. 2014b). It is worth mentioning that generator at bus No. 13 is not in service. For this system, 30 runs were taken using SCM-MJ and M-Jaya algorithms to solve the different objective functions of the OPF problem and the best results are given here.

***Case 1#* (FCM)**: The proposed SCM-MJ algorithm has been applied to solve the OPF problem in the Algerian 59-bus system to test its usefulness in solving practical power system problems. Table 5 shows the numerical results attained by the SCM-MJ algorithm and other methods mentioned in recent papers, while Table 6 shows the OPF results as well as the optimal control variable settings of the SCM-MJ algorithm. The numerical outcomes reveal that the proposed SCM-MJ algorithm produces better results as compared to other meta-heuristic algorithms for solving OPF problems. The convergence characteristic of Case 1 is represented in Figure 2.

Case 2# (FCM+W_{VPI}×VPI): The proposed SCM-MJ algorithm has been applied to solve the OPF problem which involves the minimization of fuel cost and improvement of voltage profile by minimization of total voltage deviation. In this case, the voltage profile and fuel cost obtained by the SCM-MJ algorithm are 1.8815pu and 1718.4893 $/h, respectively. The numerical outcomes of case 2 attained by the proposed method and modern optimization algorithms mentioned in the recent literature are compared in Table 5. The OPF results along with optimum control variables settings are given in Table 6. The voltage profiles of the base case, achieved by the SCM-MJ and M-Jaya algorithm presented in

Table 4. Various case studies of OPF problem investigated in this work.

Algerian 59 bus system	Case 1 : (FCM) Case 2 : (FCM+ W_{VPI} × VPI) Case 3 : (FCM + W_{RPLM} × RPLM)
IEEE-118 bus system	Case 4 : (FCM) Case 5 : (VPI) Case 6 : (RPLM)

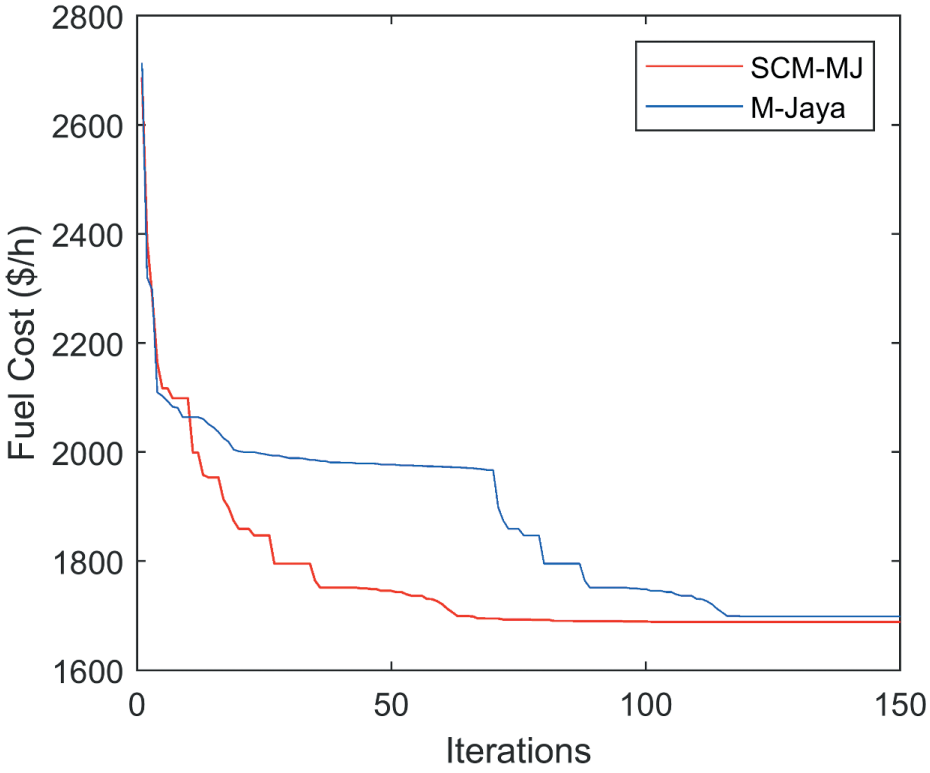


Figure 2. Convergence characteristics of fuel cost for Case 1.

Table 5. Comparison of SCM-MJ algorithm OPF results for Case 1–Case 3 of Algerian 59-bus system.

Algorithm	Fuel cost ($/h)	Voltage profile (p.u)	Real power loss (MW)	COF	Computation time (Sec)
Case – 1 (FCM)					
SCM-MJ	1688.5933	-	-	-	39.80
M-Jaya	1689.0281	-	-	-	37.23
BHBO (Bouchekara 2014a)	1710.0859	-	-	-	-
LCA (Bouchekara et al. 2014b)	1688.5122	-	-	-	-
SKH (Pulluri, Naresh, and Sharma 2017a)	1688.5742	-	-	-	47.32
ESDE-MC (Pulluri, Naresh, and Sharma 2017b)	1688.5586	-	-	-	42.34
ESDE-EC (Pulluri, Naresh, and Sharma 2017b)	1690.3171	-	-	-	45.93
ESDE (Pulluri, Naresh, and Sharma 2017b)	1692.0624	-	-	-	40.02
Case-2 (FCM + W_{VPI} × VPI)					
SCM-MJ	1718.48938	1.8815	-	2207.6711	39.78
M-Jaya	1719.8954	1.8842	-	2209.7874	36.09
LCA (Bouchekara et al. 2014b)	1755.5775	1.8404	-	-	-
Case-3 (FCM + W_{RPLM}× RPLM)					
SCM-MJ	1724.01823	-	15.4330	1878.3480	39.23
M-Jaya	1719.8684	-	15.9020	1878.8884	36.73
ESDE-MC (Pulluri, Naresh, and Sharma 2017b)	1719.9456	-	16.1682	-	127.34
ESDE-EC (Pulluri, Naresh, and Sharma 2017b)	1723.9823	-	16.5586	-	137.43
ESDE (Pulluri, Naresh, and Sharma 2017b)	1724.6272	-	16.7660	-	120.53

Table 6. Control variables settings of Case 1–Case 3 of Algerian 59-bus system using SCM-MJ algorithm.

S. No.	Control variable	Minimum value	Initial case	Case-1	Case-2	Case-3	Maximum value
Generator active power output							
1	P_{g2}	0.1	0.7	0.23372	0.20114	0.3073	0.7
2	P_{g3}	0.3	0.7	1.01775	1.0115	1.06134	5.1
3	P_{g4}	0.2	1.15	1.10507	1.20036	1.38107	4.0
4	P_{g13}	0.15	0	0	0	0	1.5
5	P_{g27}	0.1	0.4	0.25832	0.35014	0.39958	1.0
6	P_{g37}	0.1	0.3	0.50955	0.47721	0.43504	1.0
7	P_{g41}	0.15	1.1	0.97038	0.71526	0.61906	1.4
8	P_{g42}	0.18	0.7	1.40855	1.49762	1.29505	1.75
9	P_{g53}	0.3	2	1.03189	1.19992	1.11455	4.5
Generator voltage							
10	V_{g1}	0.94	1.06	1.1	1.06605	1.09999	1.1
11	V_{g2}	0.94	1.04	1.08796	1.07318	1.09973	1.1
12	V_{g3}	0.94	1.05	1.1	1.09944	1.09998	1.1
13	V_{g4}	0.94	1.0283	1.09998	1.03878	1.08747	1.1
14	V_{g13}	0.94	1	1.09878	0.99749	1.1	1.1
15	V_{g27}	0.94	1.0266	1.09932	1.03703	1.08661	1.1
16	V_{g37}	0.94	1.0273	1.1	1.02878	1.1	1.1
17	V_{g41}	0.94	1.0966	1.1	1.01045	1.09404	1.1
18	V_{g42}	0.94	1.034	1.1	1.1	1.1	1.1
19	V_{g53}	0.94	1	1.1	1.09991	1.09999	1.1
Fuel Cost ($\h)		-	1943.7010	**1688.5933**	**1718.4893**	**1724.0182**	-
Voltage Profile (p.u)		-	1.5757	2.8062	**1.8815**	**2.9395**	-
Real Power Loss (MW)		-	29.1409	27.6423	24.9567	**15.4330**	-
P_{g1}(Slack bus)		-	8.2409	58.2193	43.7417	38.2340	-
Combined Objective Function (COF)		-	-	-	2207.6711	1878.3480	-

Figure 3. This can be observed from Figure 3, that voltages at various load buses are within the specified limits.

Case 3 # **(FCM+** W_{RPLM} **×RPLM)**: The main motive of case 3 is to minimize the real power loss along with fuel cost. The multi-objective function is transformed into a single objective function by using the weighted sum method. In the combined objective function, the value of the weighting factor (W_{RPL}) assigned to real power loss was 10. The OPF results attained using the SCM-MJ and M-Jaya algorithm are compared with other reported results in Table 5. As can be observed from Table 5, the SCM-MJ algorithm provided the total fuel cost as 1724.01823 $/h and real power loss as 15.4330 MW, while the ESDE-MC algorithm provided the total fuel cost and real power loss as 1719.9456 $/h and 16.1682 MW, respectively.

In this case, the real power loss obtained by the proposed algorithm is less than the reported ESDE-MC results, while the fuel cost obtained by the proposed SCM-MJ algorithm is slightly higher than the reported ESDE-MC method (Pulluri, Naresh, and Sharma 2017b). Such type of situations will be there only because of the preference given to one objective as compared to the other objective and as discussed above, in this case, relative weightage given to real power loss is 10 as compared to fuel cost. The OPF results along with the control variable setting for this case also are shown in Table 6, while Figure 4 displays the real power loss convergence characteristics of the SCM-MJ and M-Jaya algorithms in this case.

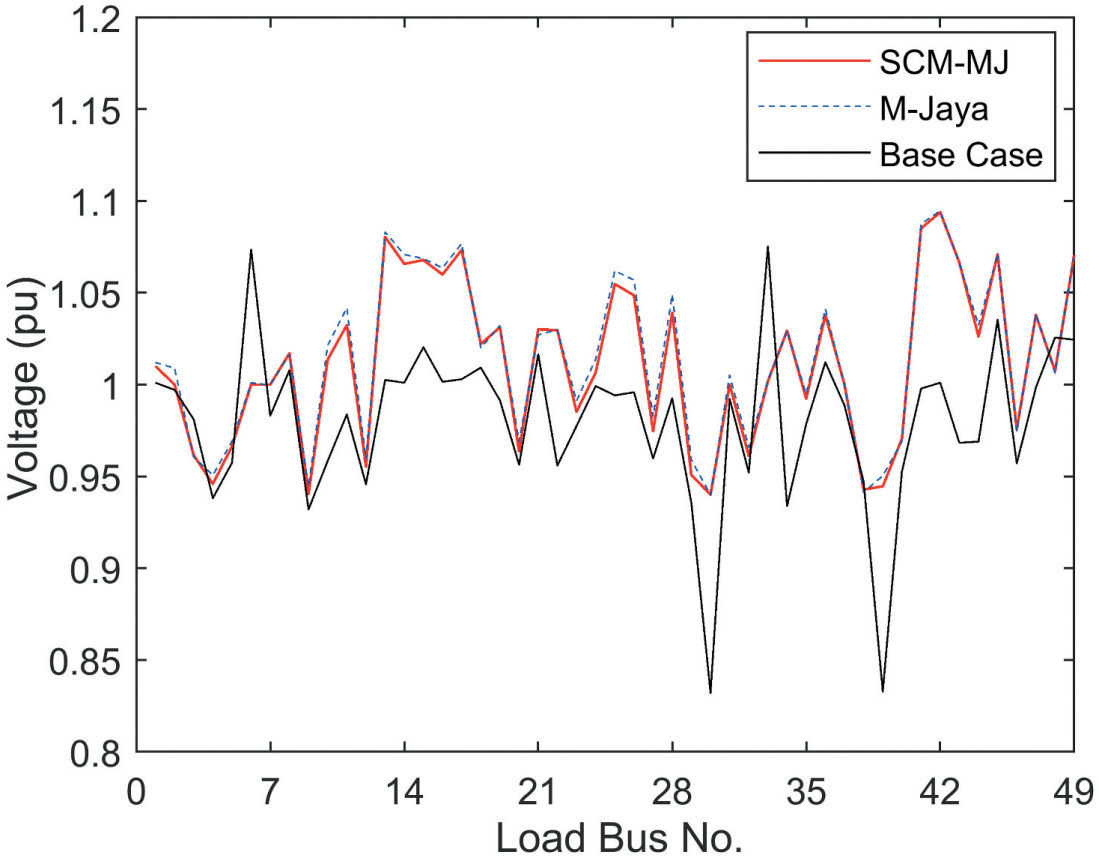


Figure 3. Load bus voltage profile for Case 2.

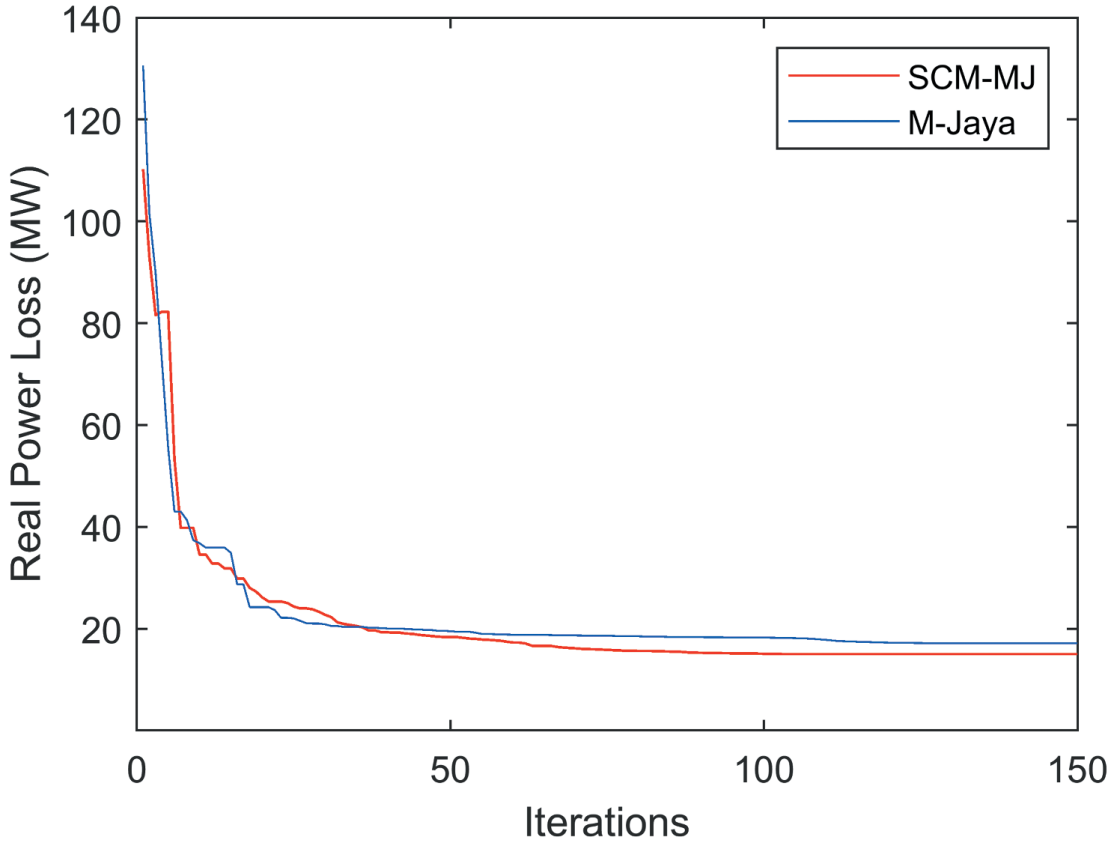


Figure 4. Convergence characteristics of real power loss for Case 3.

Table 7. Optimum values of control variables of Case 4–Case 6 for IEEE 118-bus system.

S. No.	Control variables	Initial	Case 4	Case 5	Case 6	S. No.	Control variables	Initial	Case 4	Case 5	Case 6
1	P_{g1}	0	0.0074	0.2476	0.7493	**68**	V_{g32}	0.963	1.0064	1.012	1.0013
2	P_{g4}	0	0.0352	0.0904	0.0879	**69**	V_{g34}	0.984	0.9818	1	1.004
3	P_{g6}	0	0.1114	0.1369	0.2065	**70**	V_{g36}	0.98	0.9569	0.9964	0.9947
4	P_{g8}	0	0.1151	0.4611	0.7209	**71**	V_{g40}	0.97	0.9917	1.0088	0.9877
5	P_{g10}	4.5	4.0965	3.0116	1.044	**72**	V_{g42}	0.985	1.0561	1.0033	1.0028
6	P_{g12}	0.85	0.8912	0.6187	1.7026	**73**	V_{g46}	1.005	0.9457	1.0037	1.0226
7	P_{g15}	0	0.0537	0.1774	0.887	**74**	V_{g49}	1.025	0.9765	1.0197	1.0049
8	P_{g18}	0	0.0778	0.1317	0.8205	**75**	V_{g54}	0.955	1.0265	1.0176	1.0166
9	P_{g19}	0	0.027	0.5426	0.9457	**76**	V_{g55}	0.952	1.0264	0.9804	1.0203
10	P_{g24}	0	0.2123	0.0001	0.1748	**77**	V_{g56}	0.954	1.0269	0.9938	1.0173
11	P_{g25}	2.2	1.8093	1.9346	0.1804	**78**	V_{g59}	0.985	0.9837	0.9547	1.015
12	P_{g26}	3.14	2.8809	2.1913	0.2349	**79**	V_{g61}	0.995	0.9705	0.9993	1.0161
13	P_{g27}	0	0.231	0.4348	0.9624	**80**	V_{g62}	0.998	0.9713	0.9699	1.0192
14	P_{g31}	0.07	0.0799	0.0224	0.9218	**81**	V_{g65}	1.005	0.9586	0.9486	1.0223
15	P_{g32}	0	0.0789	0.0647	0.3108	**82**	V_{g66}	1.05	0.9756	1.0436	1.0162
16	P_{g34}	0	0.5587	0.2384	1	**83**	V_{g69}	1.035	0.9869	1.0373	1.0126
17	P_{g36}	0	0.0755	0.5328	0.2322	**84**	V_{g70}	0.984	0.9542	0.9928	1.0159
18	P_{g40}	0	0.0019	0.7794	0.9251	**85**	V_{g72}	0.98	0.9439	0.9679	1.0153
19	P_{g42}	0	0.6441	0.1771	0.9828	**86**	V_{g73}	0.991	0.9895	0.9926	1.0463
20	P_{g46}	0.19	0.2165	0.2659	0.9008	**87**	V_{g74}	0.958	0.9713	0.9786	1.0228
21	P_{g49}	2.04	1.831	1.886	1.403	**88**	V_{g76}	0.943	0.9547	1.0195	1.0024
22	P_{g54}	0.48	0.4066	0.4281	1.1292	**89**	V_{g77}	1.006	0.9801	1.0011	1.0069
23	P_{g55}	0	0.3165	0.6398	0.9997	**90**	V_{g80}	1.04	1.0254	1.0384	1.0177
24	P_{g56}	0	0.0687	0.1811	0.8698	**91**	V_{g85}	0.985	0.9769	1.0239	0.9996
25	P_{g59}	1.55	1.4468	1.1159	2.3818	**92**	V_{g87}	1.015	0.9565	0.9738	1.0269
26	P_{g61}	1.6	1.4983	0.9604	0.9092	**93**	V_{g89}	1.005	1.0313	0.9937	0.9859
27	P_{g62}	0	0.01	0.1131	0.2029	**94**	V_{g90}	0.985	1.0071	0.9824	0.9881
28	P_{g65}	3.91	3.2906	3.3113	2.3234	**95**	V_{g91}	0.98	1.022	0.9535	0.9849
29	P_{g66}	3.92	3.0704	3.1084	1.8803	**96**	V_{g92}	0.99	1.0584	1.008	0.9934
30	P_{g70}	0	0.0411	0.1137	0.6109	**97**	V_{g99}	1.01	1.0509	1.0517	0.9888
31	P_{g72}	0	0.1232	0.0082	0.2463	**98**	V_{g100}	1.017	1.0525	1.007	1.0054
32	P_{g73}	0	0.1004	0.1719	0.1671	**99**	V_{g103}	1.01	1.0204	1.0391	1.0115
33	P_{g74}	0	0.568	0.6359	0.6442	**100**	V_{g104}	0.971	0.9664	1.0178	1.0177
34	P_{g76}	0	0.0944	0.9307	0.5954	**101**	V_{g105}	0.965	0.9723	1.0015	1.007
35	P_{g77}	0	0.066	0.1893	1	**102**	V_{g107}	0.952	0.9747	1.0339	0.9822
36	P_{g80}	4.77	4.0032	3.7289	3.3519	**103**	V_{g110}	0.973	1.0487	0.9994	1.041
37	P_{g85}	0	0.1829	0.0703	0.4759	**104**	V_{g111}	0.98	1.0114	1.0035	1.0599
38	P_{g87}	0.04	0.045	0.0337	0.3066	**105**	V_{g112}	0.975	1.0595	1.0273	1.0288
39	P_{g89}	6.07	5.0749	4.05	0.9493	**106**	V_{g113}	0.993	0.9989	0.9884	1.0184
40	P_{g90}	0	0.0029	0.3006	0.9232	**107**	V_{g116}	1.005	0.9881	1.0167	1.0238
41	P_{g91}	0	0.1777	0.0136	0.4551	**108**	T_{5-8}	0.985	0.9009	0.9016	0.989
42	P_{g92}	0	0.0083	0.1151	0.7886	**109**	T_{26-25}	0.96	0.9432	0.9919	0.9057
43	P_{g99}	0	0.094	0.6149	0.0599	**110**	T_{30-17}	0.96	1.0151	1.019	0.9779
44	P_{g100}	2.52	2.678	2.0898	1.697	**111**	T_{38-37}	0.935	1.0627	0.9422	0.9394
45	P_{g103}	0.4	0.402	0.0909	0.2765	**112**	T_{63-59}	0.96	1.0973	1.0673	1.0151
46	P_{g104}	0	0.2142	0.3885	0.121	**113**	T_{64-61}	0.985	1.0682	1.0303	1.0208
47	P_{g105}	0	0.0959	0.0661	0.8569	**114**	T_{65-66}	0.935	1.0956	1.0847	0.9631
48	P_{g107}	0	0.1201	0.3794	0.3457	**115**	T_{68-69}	0.935	0.9036	0.9	0.9158

(*Continued*)

Table 7. (Continued).

S. No.	Control variables	Initial	Case 4	Case 5	Case 6	S. No.	Control variables	Initial	Case 4	Case 5	Case 6
49	P_{g110}	0	0.1723	0.0102	0.6948	**116**	T_{81-80}	0.935	0.9001	0.9303	1.0363
50	P_{g111}	0.36	0.4485	0.361	0.416	**117**	Q_{C5}	0	0.136	0.0155	0.2081
51	P_{g112}	0	0.0194	0.4748	0.2227	**118**	Q_{C34}	0	0.0683	0.152	0.0029
52	P_{g113}	0	0.0271	0.1551	0.3846	**119**	Q_{C37}	0	0.1855	0.0305	0.0471
53	P_{g116}	0	0.0144	0.0216	0.4062	**120**	Q_{C44}	0	0.0824	0.042	0.1477
54	V_{g1}	0.995	1.0203	0.9991	1.0276	**121**	Q_{C45}	0	0.2582	0.1835	0.1186
55	V_{g4}	0.998	0.9948	0.9866	0.9998	**122**	Q_{C46}	0	0.0141	0.2967	0.0114
56	V_{g6}	0.99	1.0054	0.9914	0.9978	**123**	Q_{C48}	0	0.0013	0.0142	0.2974
57	V_{g8}	1.015	0.959	0.9617	0.9783	**124**	Q_{C74}	0	0.1131	0.2324	0.1259
58	V_{g10}	1.05	1.0122	0.9651	0.9909	**125**	Q_{C79}	0	0.2902	0.2698	0.0911
59	V_{g12}	0.99	1.0041	1.0126	1.015	**126**	Q_{C82}	0	0.082	0.0363	0.0599
60	V_{g15}	0.97	0.9876	1.0004	1.01	**127**	Q_{C83}	0	0.0982	0.0217	0.2881
61	V_{g18}	0.973	0.9813	0.9885	1.0165	**128**	Q_{C105}	0	0.2955	0.024	0.0371
62	V_{g19}	0.962	0.9754	1.0178	1.0164	**129**	Q_{C107}	0	0.0011	0.0268	0.0484
63	V_{g24}	0.992	1.0311	0.9905	0.9927	**130**	Q_{C110}	0	0.2277	0.0703	0.0032
64	V_{g25}	1.05	1.0273	0.9931	0.9975	**Fuel Cost ($/h)**		131220.02	**129171.96**	130879.62	163639.21
65	V_{g26}	1.015	1.0587	0.949	0.9846	**Voltage Profile (p.u)**		1.4389	1.2721	**0.4366**	0.6099
66	V_{g27}	0.968	1.0253	0.9971	1.0026	**Real Power Loss (MW)**		132.8101	114.5312	85.8153	**19.1525**
67	V_{g31}	0.967	0.9954	1.0021	1.0018	**P_{g69} (Slack Bus)**		513.8101	459.6014	439.9788	52.9582

IEEE 118-bus system

To evaluate the scalability of the SCM-MJ algorithm and to verify its efficacy to solve large scale problems, the SCM-MJ method was applied to solve the OPF problem in the IEEE 118-bus test system. The system data and maximum and minimum limits of the control variable are given (Power system test cases). For this system also, 30 runs were taken using SCM-MJ and M-Jaya algorithm to solve the different objective functions of the OPF problem and the best results are presented here.

In IEEE 118-bus system, in cases 4, 5 and 6, the three objective functions fuel cost, real power loss and voltage profile improvement were considered separately to evaluate the performance of the proposed SCM-MJ algorithm to solve single objective optimization (SOO) problems and to compare its performance with the reported results.

Case 4: Fuel Cost Minimization (FCM)

As in Case 1, FCM is chosen as the main objective function in Case 4 as defined in (1). Table 7 displays the OPF outcomes along with the optimal control variable settings of the SCM-MJ algorithm. The numerical results attained by the SCM-MJ technique are compared with the reported results available in recent literature in Table 8. The proposed SCM-MJ algorithm provides the minimum fuel cost as compared to other reported results with all the operating constraints within their pre-specified limits. This demonstrates the effectiveness of the SCM-MJ algorithm as compared to its other competitors. Figure 5 represents the fuel cost convergence characteristics for M-Jaya and SCM-MJ algorithms.

***Case 5#* (VPI)**: The proposed SCM-MJ method has been employed to solve the OPF problem for minimization of voltage deviation from 1.0pu at all the PQ (load) buses. The OPF results along with the optimum value of control variables are shown in Table 7. The voltage profile provided by the SCM-MJ algorithm is 0.4366pu. The OPF results of case 5 attained by the SCM-MJ method and other efficient methods reported in the recent literature are compared in Table 8. The voltage profile of the

Table 8. Comparison of SCM-MJ algorithm OPF results for Case 4–Case 6 of IEEE 118-bus system.

Algorithm	Fuel cost ($/h)	Voltage profile (p. u)	Real power loss (MW)	COF	Computation time (Sec)
Case – 4 (FCM)					
SCM-MJ	**129171.96**	-	-	-	63.85
M-Jaya	129248.10	-	-	-	59.79
NLP (Fortenbacher and Demiray 2019)	129700.00	-	-	-	0.80
QP (Fortenbacher and Demiray 2019)	129600.00	-	-	-	0.36
Interior point (Jiang et al. 2010)	129720.70	-	-	-	0.13
MSCA (Attia, Sehiemy, and Hasanien 2018)	129620.22	-	-	-	-
SCA (Attia, Sehiemy, and Hasanien 2018)	129622.65	-	-	-	-
ALC-PSO (Singh, Mukherjee, and Ghoshal 2016)	129546.08	-	-	-	-
IMFO (Taher et al. 2018)	131820.00	-	-	-	-
PSOGSA (Radosavljevic et al. 2015)	129733.58	-	-	-	-
FAHSPSO-DE (Naderi et al. 2021)	129519.38	-	-	-	64.93
Case- 5 (VPI)					
SCM-MJ	-	**0.4366**	-	-	62.98
M-Jaya	-	0.6771	-	-	60.08
MSCA (Attia, Sehiemy, and Hasanien 2018)	-	0.995	-	-	-
SCA (Attia, Sehiemy, and Hasanien 2018)	-	1.32	-	-	-
ALC-PSO (Singh, Mukherjee, and Ghoshal 2016)	-	0.4412	-	-	-
Case- 6 (RPLM)					
SCM-MJ	-	-	**19.1525**	-	63.19
M-Jaya	-	-	21.6419	-	59.07
QOTLBO (Mandal and Roy 2014)	-	-	35.3191	-	-
TLBO (Mandal and Roy 2014)	-	-	36.8482	-	-
MSCA (Attia, Sehiemy, and Hasanien 2018)	-	-	77.0873	-	-
SCA (Attia, Sehiemy, and Hasanien 2018)	-	-	77.1113	-	-
EGA (Reddy 2019)	-	-	31.3519	-	-
DE-HS (Reddy 2019)	-	-	30.78047	-	-

base case and this case is displayed in Figure 6. The voltage profile obtained by the proposed SCM-MJ and M-Jaya algorithm shows that all the PQ buses' voltages are within the operating limits.

***Case 6#* (RPLM)**: In this case, minimization of real power loss independently is selected as an objective function. The power loss is significantly improved in this case as compared to the base case. Table 7 shows the OPF results with the optimal control variable settings of the SCM-MJ algorithm and Table 8 compares the numerical outcomes attained by the SCM-MJ algorithm and other reported results. The convergence characteristics of the proposed algorithm are depicted in Figure 7. As can be seen from Figure 7, the proposed SCM-MJ method has offered smooth convergence characteristics as compared to the M-Jaya algorithm.

Statistical analysis

The lowest values of the best, worst, average, and standard deviation offered by the SCM-MJ algorithm demonstrate that the proposed SCM-MJ algorithm provides statistically significant results in all the cases. Thirty separate trials for the same population size and number of iterations were conducted for each case of both the power systems. Table 9 represents the results of the 30 trials which were used to measure the best, worst, average (mean), and standard deviation (SD). In all the cases, the lowest values of the best, worse, average, and standard deviation of the SCM-MJ algorithm are clear evidence

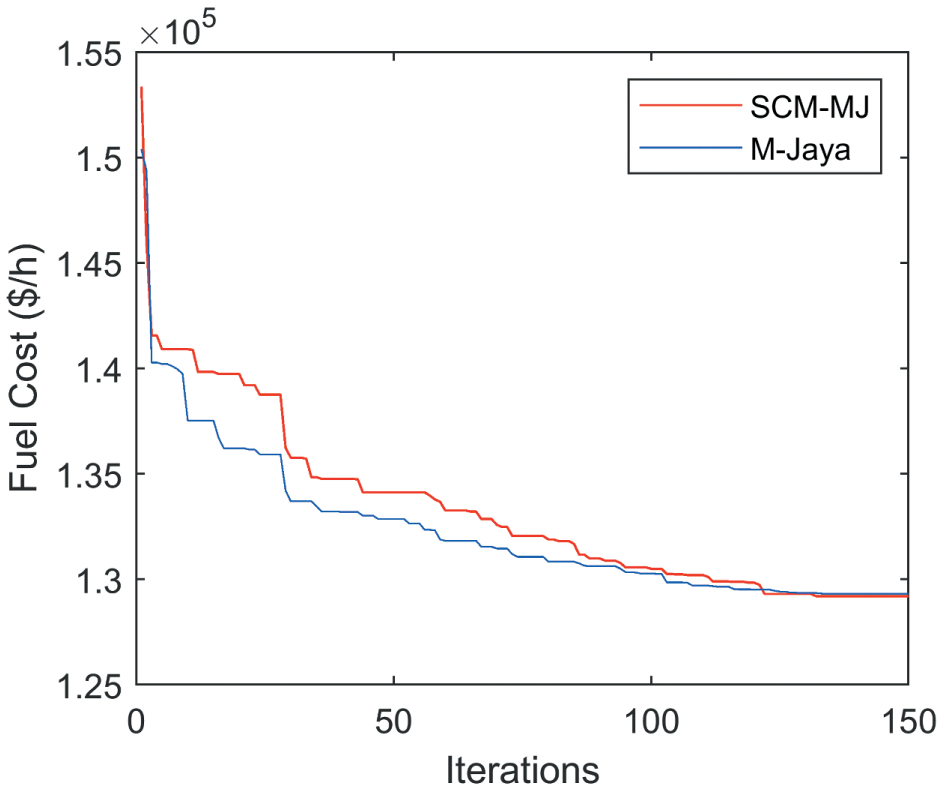


Figure 5. Convergence characteristics for fuel cost minimization (Case 4).

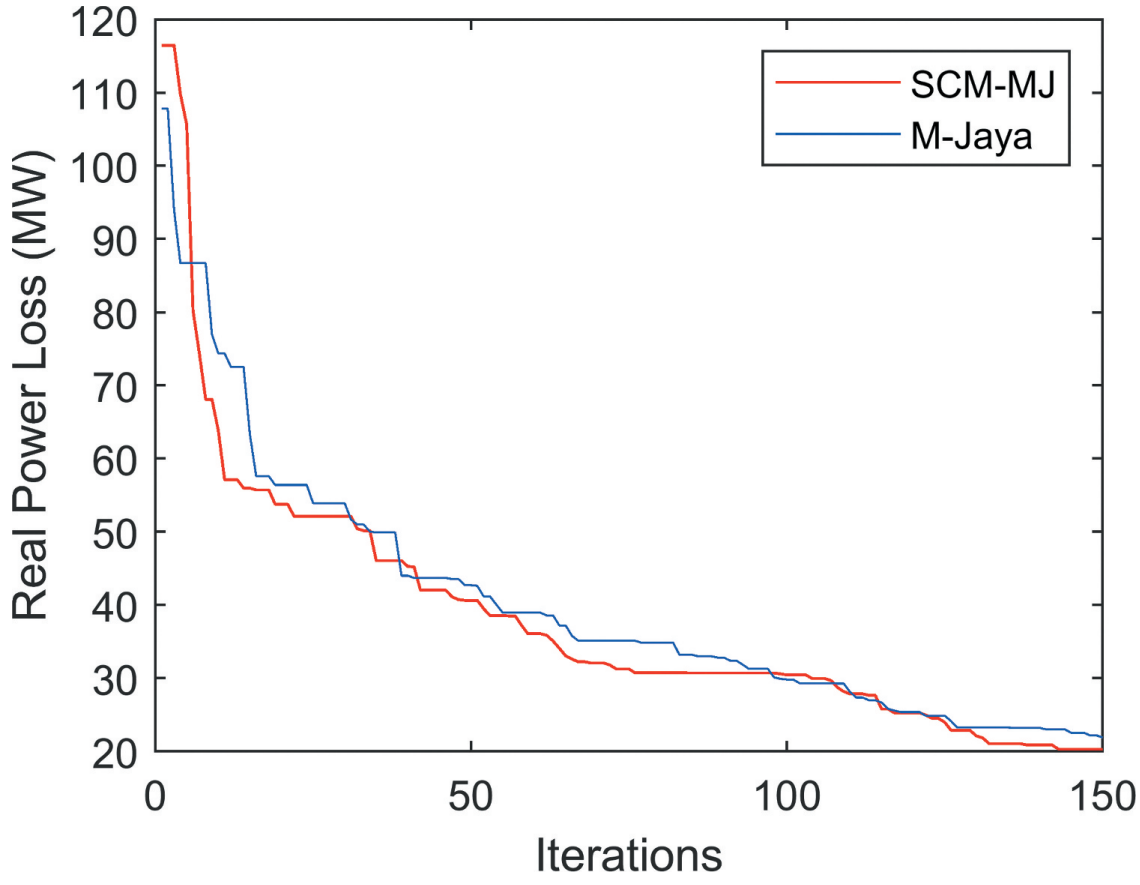


Figure 6. Load bus voltage profile for Case 5.

Table 9. Statistical analysis of the various cases using the SCM-Jaya and M-Jaya algorithms.

	Algerian 59-bus system				IEEE 118-bus system			
Algorithm	Best	Worst	Mean	Std. deviation	Best	Worst	Mean	Std. deviation
Case 1							**Case 4**	
SCM-MJ	1688.5933	1689.0236	1688.7502	0.0330	129171.96	129188.14	129181.62	9.190
M-Jaya	1689.0281	1690.2490	1689.9762	0.0372	129248.10	129264.24	129255.09	9.276
Case 2							**Case 5**	
SCM-MJ	2207.6711	2210.9225	2208.2497	0.0290	0.4366	0.4401	0.4386	0.029
M-Jaya	2210.0867	2214.7244	2212.9852	0.0341	0.6771	0.6779	0.6775	0.034
Case 3							**Case 6**	
SCM-MJ	1878.3480	1880.1783	1879.9610	0.0271	19.1525	20.7123	19.6571	0.028
M-Jaya	1862.9864	1864.7612	1864.0213	0.0421	21.6419	23.0154	22.1765	0.031

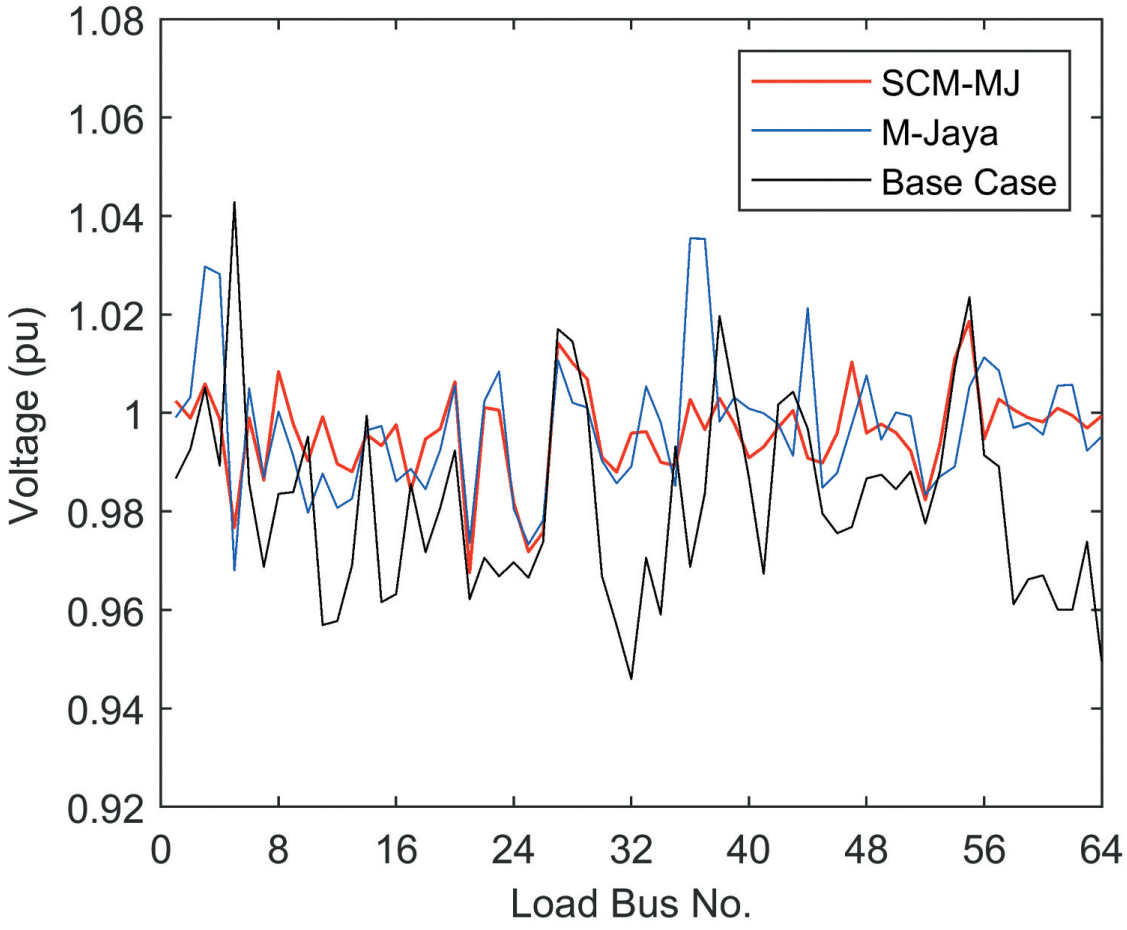


Figure 7. Convergence of SCM-MJ algorithm for Case 6.

of statistically significant outcomes from the proposed SCM-MJ method. This proves the strength of the proposed SCM-MJ algorithm.

Discussion

Like other population-based algorithms, the Jaya algorithm sometimes suffers from premature convergence. A versatile combination of two meta-heuristic algorithms may overcome their common weaknesses while taking advantage of the strengths of the two algorithms. The proposed SCM-MJ

algorithm aims to maintain the diversity of the solutions throughout the search to avoid sub-optimal solutions, and find near-global optimum solutions.

To show the effectiveness of the suggested SCM-MJ algorithm, this algorithm has been applied on 13 benchmarks (unimodal and multimodal) functions. The numerical results are listed in Table 3. Observations of the numerical results shown in Table 3 prove the dominance of the SCM-MJ algorithm over eight well-known optimization methods reported in the recent publications: ALO, BA, CS, FPA, FA, GA, M-Jaya, PSO, and SMS.

To check the efficacy of the SCM-MJ algorithm, it is applied for solving the OPF problem in the Algerian 59-bus system and standard IEEE 118-bus system. The Algerian 59-bus system is a practical power system used to demonstrate the validity of the SCM-MJ algorithm to solve the real-world OPF problem. The IEEE 118-bus system is considered for minimization of fuel cost, voltage profile, and real power loss using the SCM-MJ algorithm to validate its capability to handle large dimensional problems, as in this system the number of control variables is 130. The results given in Tables 5, 6, 7, and 8 shows that the numerical outcomes attained using the SCM-MJ algorithm are well aligned to those of other methods reported in recent literature.

The lowest values of the best, worst, average, and standard deviation are given by the SCM-MJ algorithm demonstrate that the proposed SCM-MJ solution provides statistically significant results in all the cases of OPF problem. This confirms the effectiveness of the SCM-MJ algorithm to solve the large-scale complex optimization problem.

Conclusions

The significance of OPF solution has increased many folds with the addition of various types of power electronic equipment and renewable energy sources in the interconnected electrical grid. In this paper, a sine-cosine mutation-based modified Jaya algorithm for solving the OPF problem has been proposed. The suggested SCM-MJ algorithm is found to be faster and immune to the local optima trapping as compared to the classical Jaya algorithm. To test the efficacy of the suggested SCM-MJ method, it is applied to calculate the mean value and standard deviations using 13 benchmark functions. The statistical analysis indicates that the SCM-MJ method is a reliable and robust optimization algorithm over other modern meta-heuristic optimization approaches proposed in recent literature. Later, the proposed algorithm was implemented in the Algerian 59-bus and IEEE 118-bus system for solving the OPF problem for minimization of fuel cost, voltage profile improvement, and real power loss minimization. Comparison of the optimization results acquired from the SCM-MJ algorithm with those of modern meta-heuristic optimization approaches published in recent literature demonstrates that the proposed SCM-MJ algorithm is highly efficient and robust over other recently developed popular algorithms.

In all the cases considered in this paper, the SCM-MJ algorithm consistently offered the lowest objective function value irrespective of the problem dimension and system size. For example, in the Algerian 59-bus system fuel cost saving is about 13.12% (from base case fuel cost of 1943$ to optimized fuel cost of 1688$), which is equivalent to a reduction of 2,234,746$ per year in fuel cost. For the IEEE 118-bus test system, the SCM-MJ algorithm provided a fuel cost reduction of 1.56% in comparison to the base case (from base case fuel cost of 131220$ to optimized fuel cost of 129171$), which is equivalent to a cost-saving of 17,940,830$ annually.

As the future scope of the work, the same hybrid algorithm can be applied to solve OPF problems under normal loading, severe loading, and contingency conditions in restructured power system. Dynamic models of FACTS devices can be developed and placed at optimal locations to examine their

impact on voltage profile and voltage stability of power networks at normal and severe loading conditions. The inclusion of discrete control variables like transformer taps-setting and shunt compensators in OPF problem formulation is another direction for future studies.

Nomenclature

x	Dependent variable	$V_{g_k}^{max}$ and $V_{g_k}^{min}$	Maximum and minimum voltage limit of the kth generator bus
u	Control variable	$Q_{g_k}^{max}$ and $Q_{g_k}^{min}$	Maximum and minimum limits of reactive power output of kth generator bus
$Z(u, x)$	Objective function	$P_{g_k}^{max}$and $P_{g_k}^{min}$	Maximum and minimum active power limit of the kth generating units
$P_{di}$and Q_{di}	Active and reactive power demand of ith load bus	$T_k^{min}$and T_k^{max}	Minimum and maximum tap setting of kth transformers
P_{gi} and Q_{gi}	Active and reactive power generations of ith generator units	$V_{L_k}^{min}$ and $V_{L_k}^{max}$	Minimum and maximum voltage limit of kth load bus
NB	Number of buses	$S_{l_k}^{max}$	Maximum MVA flow in kth transmission line
NLB	Number of load buses	C_1,C_2, C_3 and C_4	Penalty factors
Ntl	Number of transmission lines	A_i, B_i, and C_i	Fuel cost coefficients of the ith generator units
NGN	total numbers of generators units	$P_{g_{Slack}}$	Active power output at the slack bus.
NC	Number of VAR compensation units	$\alpha_i, \beta_i, \gamma_i, \xi_i, \lambda_i$	Emission coefficients of the ith generator units
NTR	Number of regulating transformers	$Iter_{max}$	Maximum number of iterations
P_{Loss} and Q_{Loss}	Real and reactive power loss of system		

Notes on contributors

Saket Gupta obtained his B.E. degree in Electrical Engineering from Samrat Ashok Technological Institute, Vidisha, India in 2012 and M. E. degree in Industrial Systems & Drives from Madhav Institute of Technology and Science Gwalior India, in 2015, respectively. He is currently working toward the Ph.D. degree at Delhi Technological University, New Delhi, India. His research interests include Nature Inspired algorithm and their applications in Power System.

Narendra Kumar received his B. Tech. degree in Electrical Engineering and M. Tech. degree in Power System and Drives from Aligarh Muslim University, Aligarh, India, in 1984 and 1987, respectively and Ph.D. in Electrical Engineering from University of Roorkee (Presently IIT Roorkee) India. He is currently working as a Professor at Delhi Technological University, New Delhi, India. Dr. Kumar has presented / published 230 research papers in reputed international/national journals and conference proceedings. He worked as an active reviewer of research papers of various international journals of repute like IEEE Transactions on Power System, IEEE Transactions on Smart Grid, IEEE Systems, IET Proceedings Generation, Transmission and Distribution, Electric Power Application, Power Electronics, Renewable Energy, Int. Journal of Computers & Electrical Engineering, Applied Soft Computing, Journal of Electric Power Components and Systems etc. His research interests include Power System Operation & Control, Soft Computing Applications to Electrical Power System, FACTS, AGC etc. He is Life Member Systems Society of India, Life Member Indian Society for Technical Education, Senior Member, IEEE, Fellow of Institution of Engineers India, Member Institution of Engineering and technology (IET, UK), Member IEEE USA.

Laxmi Srivastava obtained her B.E and Ph.D. in Electrical Engineering from University of Roorkee (Presently IIT Roorkee) India and M. Tech. in Electrical Engineering from IIT, Kanpur. She is a Professor in the Department of Electrical Engineering, M.I.T.S. Gwalior, India. Dr. Srivastava has published / presented more than 245 research papers in reputed international/national journals and conference proceedings. She is currently involved in research in AI applications in Power System Security Analysis, Optimization and Control. Dr. Srivastava has worked as reviewer of research papers of various international journals of repute. She is a Fellow of the Institution of Engineers, IE(I), Senior Member of IEEE, Life member of Indian Society for Technical Education (ISTE), Institution of Electronics & telecommunication Engineers (IETE), Computer Society of India (CSI) and Indian Women Scientist Association (IWSA).

ORCID

Saket Gupta http://orcid.org/0000-0003-4541-5798
Narendra Kumar http://orcid.org/0000-0003-1083-2129
Laxmi Srivastava http://orcid.org/0000-0002-5922-6835

References

Abbasi, M., E. Abbasi, and B. M. Ivatloo. 2021. Single and multi-objective optimal power flow using a new differential-based harmony search algorithm. *Journal of Ambient Intelligence and Humanized Computing* 12 (1):851–71. doi:10.1007/s12652-020-02089-6.

Abido, M. A. 2002. Optimal power flow using particle swarm optimization. *International Journal of Electrical Power & Energy Systems* 24 (7):563–71. doi:10.1016/S0142-0615(01)00067-9.

Adhvaryyu, P. K., and S. Adhvaryyu. 2021. Static optimal load flow of combined heat and power system with valve point effect and prohibited operating zones using Krill Herd algorithm. *Energy Systems* 12 (1):133–56. doi:10.1007/s12667-020-00378-9.

Ajayi, O., N. Nwulu, and U. Damisa. 2020. Application of metaheuristic algorithms in DC-optimal power flow. *African Journal of Science, Technology, Innovation and Development* 12 (7):867–72. doi:10.1080/20421338.2020.1726084.

Attia, A.-F., R. A. E. I. Sehiemy, and H. M. Hasanien. 2018. Optimal power flow solution in power systems using a novel Sine-Cosine algorithm. *International Journal of Electrical Power and Energy Systems* 99:331–43. doi:10.1016/j.ijepes.2018.01.024.

Azizivahed, A., A. Arefi, E. Naderi, H. Narimani, M. Fathi, and M. R. Narimani. 2020. An Efficient Hybrid Approach to Solve Bi-objective Multi-area Dynamic Economic Emission Dispatch Problem. *Electric Power Components and Systems* 48 (4–5):485–500. doi:10.1080/15325008.2020.1793830.

Bakirtzis, A. G., P. N. Biskas, C. E. Zoumas, and V. Petridis. 2002. Optimal power flow by enhanced genetic algorithm. *IEEE Transactions on Power Systems* 17 (2):229–36. doi:10.1109/TPWRS.2002.1007886.

Bhattacharya, A., and P. K. Roy. 2012. Solution of multi-objective optimal power flow using gravitational search algorithm. *IET Generation, Transmission & Distribution* 6 (8):751–63. doi:10.1049/iet-gtd.2011.0593.

Bouchekara, H. R. E. H. 2014a. Optimal power flow using black-hole-based optimization approach. *Applied Soft Computing* 24:879–88. doi:10.1016/j.asoc.2014.08.056.

Bouchekara, H. R. E. H., M. A. Abido, A. Chaib, and R. Mehasni. 2014b. Optimal power flow using the league championship algorithm: A case study of the Algerian power system. *Energy Conversion and Management* 87:58–70. doi:10.1016/j.enconman.2014.06.088.

Christy, A. A., and P. A. D. Raj. 2014. Adaptive biogeography based predator–prey optimization technique for optimal power flow. . *International Journal of Electrical Power & Energy Systems* 62:344–52. doi:10.1016/j.ijepes.2014.04.054.

Devaraj, D., and B. Yegnanarayana. 2005. Genetic-algorithm-based optimal power flow for security enhancement. *IEE Proceedings - Generation, Transmission and Distribution* 152 (6):899–905. doi:10.1049/ip-gtd:20045234.

Elattar, E. E., and S. K. ElSayed. 2019. Modified JAYA algorithm for optimal power flow incorporating renewable energy sources considering the cost, emission, power loss and voltage profile improvement. *Energy* 178:598–609. doi:10.1016/j.energy.2019.04.159.

El-Fergany, A. A., and H. M. Hasanien. 2020. Salp swarm optimizer to solve optimal power flow comprising voltage stability analysis. *Neural Computing and Applications* 32 (9):5267–83. doi:10.1007/s00521-019-04029-8.

Farahani, H. F., J. Aghaei, and F. Rashidi. 2018. Optimal power flow of HVDC system using teaching–learning-based optimization algorithm. *Neural Computing and Applications* 30 (12):3781–89. doi:10.1007/s00521-017-2962-3.

Fortenbacher, P., and T. Demiray. 2019. Linear/quadratic programming-based optimal power flow using linear power flow and absolute loss approximations. *Electrical Power and Energy Systems* 107:680–89. doi:10.1016/j.ijepes.2018.12.008.

Gacem, A., and D. Benattous. 2017. Hybrid genetic algorithm and particle swarm for optimal power flow with non-smooth fuel cost functions. *International Journal of System Assurance Engineering and Management* 8 (S1):146–53. doi:10.1007/s13198-014-0312-8.

Ghasemi, M., S. Ghavidel, M. Gitizadeh, and E. Akbari. 2015. An improved teaching–learning-based optimization algorithm using Lévy mutation strategy for non-smooth optimal power flow. . *International Journal of Electrical Power & Energy Systems* 65:375–84. doi:10.1016/j.ijepes.2014.10.027.

Ghasemi, M., S. Ghavidel, S. Rahmani, A. Roosta, and H. Falah. 2014. A novel hybrid algorithm of imperialist competitive algorithm and teaching learning algorithm for optimal power flow problem with non-smooth cost functions. *Engineering Applications of Artificial Intelligence* 29:54–69. doi:10.1016/j.engappai.2013.11.003.

Jadhav, H. T., and P. D. Bamane. 2016. Temperature dependent optimal power flow using g-best guided artificial bee colony algorithm. *International Journal of Electrical Power & Energy Systems* 77:77–90. doi:10.1016/j.ijepes.2015.11.026.

Jamal, R., B. Men, and N. H. Khan. 2020. A novel nature inspired meta-heuristic optimization approach of GWO optimizer for optimal reactive power dispatch problems. *IEEE Access* 8:202596–610. doi:10.1109/ACCESS.2020.3031640.

Jiang, Q., G. Geng, C. Guo, and Y. Cao. 2010. An efficient implementation of automatic differentiation in interior point optimal power flow. *IEEE Transactions on Power Systems* 25 (1):147–55. doi:10.1109/TPWRS.2009.2030286.

Joorabian, M., and E. Afzalan. 2014. Optimal power flow under both normal and contingent operation conditions using the hybrid fuzzy particle swarm optimisation and Nelder–Mead algorithm (HFPSO–NM). *Applied Soft Computing Journal* 14:623–33. doi:10.1016/j.asoc.2013.09.015.

Kaur, M., and N. Narang. 2020. An integrated optimization technique for optimal power flow solution. *Soft Computing* 24 (14):10865–82. doi:10.1007/s00500-019-04590-3.

Khelifi, A., B. Bentouati, and S. Chettih. 2019. Optimal Power Flow Problem Solution Based on Hybrid Firefly Krill Herd Method. *International Journal of Engineering Research in Africa* 44:213–28. www.scientific.net/JERA.44.213 .

Kotb, M. F., and A. A. El-Fergany. 2020. Optimal Power Flow Solution Using Moth Swarm Optimizer Considering Generating Units Prohibited Zones and Valve Ripples. *J. Electr. Eng. Technol* 15:179–92. doi:10.1007/s42835-019-00144-7.

Kumar, A. R., and L. Premalatha. 2015. Optimal power flow for a deregulated power system using adaptive real coded biogeography-based optimization. *Int J Electr Power & Energy Syst* 393:399–73. doi:10.1016/j.ijepes.2015.05.011.

Liang, R.-H., S.-R. Tsai, Y.-T. Chen, and W.-T. Tseng. 2011. Optimal power flow by a fuzzy based hybrid particle swarm optimization approach. *Electric Power Systems Research* 81 (7):1466–74. doi:10.1016/j.epsr.2011.02.011.

Mahdad, B., and K. Srairi. 2014. Multi objective large power system planning under sever loading condition using learning DE-APSO-PS strategy. *Energy Conversion and Management* 87:338–50. doi:10.1016/j.enconman.2014.06.090.

Mahdad, B., and K. Srairi. 2015. Blackout risk prevention in a smart grid based flexible optimal strategy using Grey Wolf-pattern search algorithms. *Energy Conversion and Management* 98:411–29. doi:10.1016/j.enconman.2015.04.005.

Mandal, B., and P. K. Roy. 2014. Multi-objective optimal power flow using quasi-oppositional teaching learning based optimization. . *Applied Soft Computing* 21:590–606. doi:10.1016/j.asoc.2014.04.010.

Mirjalili, S. 2015. The ant lion optimizer. *Advances in Engineering Software* 83:80–98. doi:10.1016/j.advengsoft.2015.01.010.

Mohamed, A.-A. A., Y. S. Mohamed, A. A. M. El-Gaafary, and A. M. Hemeida. 2017. Optimal power flow using moth swarm algorithm. *Electric Power Systems Research* 142:190–206. doi:10.1016/j.epsr.2016.09.025.

Momoh, J. A., R. Adapa, and M. E. El-Hawary. 1999. A review of selected optimal power flow literature to 1993. I. Nonlinear and quadratic programming approaches. *IEEE Transactions on Power Systems* 14 (1):96–104. doi:10.1109/59.744492.

Naderi, E., M. P. Kasmaei, M., . F. V. Cerna, and M. Lehtonen. 2021. A novel hybrid self-adaptive heuristic algorithm to handle single- and multi-objective optimal power flow problems. *International Journal of Electrical Power and Energy Systems* 125:106492. doi:10.1016/j.ijepes.2020.106492.

Nguyen, T. T. 2019. A high performance social spider optimization algorithm for optimal power flow solution with single objective optimization. *Energy* 171:218–40. doi:10.1016/j.energy.2019.01.021.

Niknam, T., M. R. Narimani, J. Aghaei, and R. A. Abarghooee. 2012. Improved particle swarm optimisation for multi-objective optimal power flow considering the cost, loss, emission and voltage stability index. *IET Gener, Transm Distrib* 515:527–6. doi:10.1049/iet-gtd.2011.0851.

Niknam, T., M. R. Narimani, and M. Jabbari. 2013. Dynamic Optimal power flow using hybrid particle swarm optimization and simulated annealing. *International Transactions on Electrical Energy Systems* 23 (7):975–1001. doi:10.1002/etep.1633.

Niknam, T., M. R. Narimani, and R. A. Abarghooee. 2012. A new hybrid algorithm for optimal power flow considering prohibited zones and valve point effect. *Energy Conversion and Management* 58:197–206. doi:10.1016/j.enconman.2012.01.017.

Ongsakul, W., and P. Bhasaputra. 2002. Optimal power flow with FACTS devices by hybrid TS/SA approach. *International Journal of Electrical Power & Energy Systems* 24 (10):851–57. doi:10.1016/S0142-0615(02)00006-6.

Ongsakul, W., and T. Tantimaporn. 2006. Optimal power flow by improved evolutionary programming. *Electric Power Components and Systems* 34 (1):79–95. doi:10.1080/15325000691001458.

Power system test cases [online]. Available: <http://www.ee.washington.edu/research/pstca >.

Pulluri, H., R. Naresh, and V. Sharma. 2017a. Application of stud krill herd algorithm for solution of optimal power flow problems. *International Transactions on Electrical Energy Systems* 27 (6):2316. doi:10.1080/15325008.2020.1793830.

Pulluri, H., R. Naresh, and V. Sharma. 2017b. An enhanced self-adaptive differential evolution based solution methodology for multiobjective optimal power flow. *Applied Soft Computing Journal* 54:229–45. doi:10.1016/j.asoc.2017.01.030.

Radosavljević, J., D. Klimenta, M. Jevtić, and N. Arsić. 2015. Optimal Power Flow Using a Hybrid Optimization Algorithm of Particle Swarm Optimization and Gravitational Search Algorithm. *Electric Power Components and Systems43* (17):1958–70. doi:10.1080/15325008.2015.1061620

Radosavljević, J., D. Klimenta, M. Jevtić, and N. Arsić. 2015. Optimal Power Flow Using a Hybrid Optimization Algorithm of Particle Swarm Optimization and Gravitational Search Algorithm. *Electric Power Components and Systems* 43 (17):1958–70. doi:10.1080/15325008.2015.1061620.

Radosavljevic, J., 2018. Metaheuristic optimization in power engineering. *IET Digital Library*. ISBN-13: 978-1-78561-546-7.

Rao, R. V. 2016. Jaya: A simple and new optimization algorithm for solving constrained and unconstrained optimization problems. Int. J. Ind. *Eng. Comput* 7 (1):19–34. doi:10.5267/j.ijiec.2015.8.004.

Reddy, S. S. 2019. Optimal power flow using hybrid differential evolution and harmony search algorithm. *International Journal of Machine Learning and Cybernetics* 10 (5):1077–91. doi:10.1007/s13042-018-0786-9.

Singh, P., N. K. Meena, J. Yang, E. Vega-Fuentes, and S. K. Bishnoi. 2020. Multi-criteria decision making monarch butterfly optimization for optimal distributed energy resources mix in distribution networks. *Applied Energy* 278:115723. doi:10.1016/j.apenergy.2020.115723.

Singh, R. P., V. Mukherjee, and S. P. Ghoshal. 2016. Particle swarm optimization with an aging leader and challengers algorithm for the solution of optimal power flow problem. *Applied Soft Computing Journal* 40:161–77. doi:10.1016/j.asoc.2015.11.027.

Srivastava, L., and H. Singh. 2015. Hybrid multi-swarm particle swarm optimisation based multi-objective reactive power dispatch. *IET Generation, Transmission & Distribution* 9 (8):727–39. doi:10.1049/iet-gtd.2014.0469.

Taher, M. A., S. Kamel, F. Jurado, and M. Ebeed. 2018. An improved moth-flame optimization algorithm for solving optimal power flow problem. *Int Trans Electr Energ Syst*. doi:10.1002/etep.2743.

Tan, Y., C. Li, Y. Cao, K. Y. Lee, L. Li, S. Tang, and L. Zhou. 2015. Improved group search optimization method for optimal power flow problem considering valve-point loading effects. *Neuro Computing* 148:229–39. doi:10.1016/j.neucom.2013.09.065.

Vaisakh, K., and L. R. Srinivas. 2011. Genetic evolving ant direction HDE for OPF with non-smooth cost functions and statistical analysis. *Expert Systems with Applications* 38 (3):2046–62. doi:10.1016/j.eswa.2010.07.142.

Varadarajan, M., and K. S. Swarup. 2008. Solving multi-objective optimal power flow using differential evolution. *IET Generation, Transmission & Distribution* 2 (5):720–30. doi:10.1049/iet-gtd:20070457.

Zhou, C., L. Chen, Z. Chen, X. Li, and G. Dai. 2017. A sine cosine mutation based differential evolution algorithm for solving node location problem. *International Journal of Wireless and Mobile Computing* 13 (3):3. doi:10.1504/IJWMC.2017.088531.

Environmental Earth Sciences (2021) 80:654
https://doi.org/10.1007/s12665-021-09954-3

ORIGINAL ARTICLE

Status of groundwater resource potential and its quality at Kabul, Afghanistan: a review

Ali Reza Noori[1,2] · S. K. Singh[1]

Received: 4 July 2020 / Accepted: 7 September 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Groundwater is an integral part of water resources. Groundwater has a vital role in water use in Kabul, Afghanistan. Nowadays, it is the only available source for water supply in Kabul city. The study aimed to comprehensively review the spatiotemporal groundwater distribution, physicochemical and bacteriological specifications of groundwater, and its consumption in Kabul, Afghanistan. Based on developed specific search terms to locate appropriate literature, the cited material is evaluated for relevance and reliability. Due to several causes, the mean annual decrease in groundwater levels between 2008 and 2016 is 1.7 m. About 33% of supply wells are not in operation due to declining groundwater tables. According to most reports, the physicochemical and bacteriological contents of groundwater in Kabul city exceeds World Health Organization (WHO) and National Drinking Water Quality Standard (NDWQS) limits. Individuals use water through four significant suppliers: Private wells, private suppliers, public water supply, and public wells. The study provides further insight into groundwater's current condition, and some recommendations are also presented to improve the sustainable development of groundwater.

Keywords Kabul basin · Aquifer · Groundwater · Water quality · Semi-arid climate · Kabul City

Introduction

Water is one of the main factors limiting the achievement of sustainable development, especially in cases wherein addition to quantitative limitations of water resources in terms of quality is also under serious threat. Quantitative constraints arise naturally due to population growth, while qualitative constraints arise due to overuse, inefficient use of water resources, and improper management of generated effluents. About one-quarter of the world's water needs are groundwater-dependent (Hao et al. 2018). As per a 2004 UNESCO report, groundwater supplies around 70 percent of the global populace, by 51 percent of nations allegedly extracting over 100 cubic meters for each capita yearly (Mengistu et al. 2019). The groundwater meets almost all of the water needs in many arid and semiarid areas. Just 3% of the total water source on Earth is known as freshwater. About 30% are accessible as groundwater, which is essential for human health, habitats, electricity, and other water-dependent topics (Kalhor et al. 2019). Recent decades have witnessed growing rates of decreasing water levels in the aquifer, wetland destruction, seawater intrusion, and a general deterioration of water quality (Niu et al. 2014). At the same time, the groundwater research considerably increased the analysis of a wide range of topics, such as groundwater quality assessment (Sadat-Noori et al. 2014; Varol and Davraz 2015; Selvakumar et al. 2017; Singaraja 2017; Kalhor et al. 2019; Saraswat et al. 2019; Verma et al. 2020), hydro geochemical evaluation (Krishna Kumar et al. 2015; Haritash et al. 2017, 2018; Srinivas et al. 2017; Chidambaram et al. 2018; Karami et al. 2018), vulnerability assessments (Naik et al. 2008; Awawdeh et al. 2015; Abdeslam et al. 2017; Oroji 2019), purification of groundwater (Uhl and Tahiri 2003; Ahmad et al. 2018; Dinh et al. 2020; Khan et al. 2020; Van Nguyen et al. 2020), groundwater quantity assessment and modeling, the relationship between groundwater, urbanization and land use (Malik et al. 2010; Kalhor and Emaminejad 2019; Zhang et al. 2019; Hall et al. 2020; Hou et al. 2020; Huang et al. 2020) and groundwater recharge (Prabhu and Venkateswaran

✉ Ali Reza Noori
a.noori@kpu.edu.af

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

[2] Department of Water Supply and Environmental Engineering, Faculty of Water Resources and Environmental Engineering, Kabul Polytechnic University, Kabul, Afghanistan

Published online: 18 September 2021

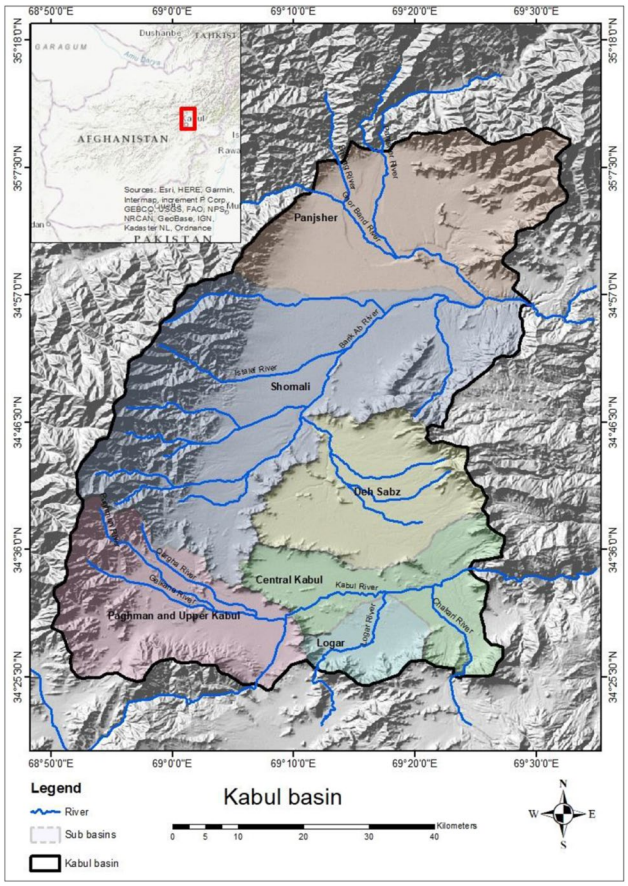


Fig. 1 The Kabul basin with groundwater sub-basins (reproduced from Mack et al. 2010 using ArcGIS and Degital Elevation Model = DEM)

2015; Senanayake et al. 2016; Yeh et al. 2016; Abijith et al. 2020; Achu et al. 2020; Adhikari et al. 2020; Arefin 2020; Lentswe and Molwalefhe 2020; Patel et al. 2020) are the most focused topics globally.

The Kabul basin, part of Kabul River (Indus) basin, a geologic valley, including Afghanistan's capital city, allocated north–south, surrounded via Pghman mountains westerly and the Saffi mountains easterly (Fig. 1) (Bohannon 2010). The basin was divided into six sub-basins of Central Kabul, including Kabul, Pagham, Upper Kabul, Logar, Deh Sabz, Shomali, and Panjsher (Mack et al. 2009). This investigation focuses specifically on the south area of the basin, in particular, the Central Kabul, Paghman, and upper Kabul and Logar sub-basins covering Kabul's city, which are of specific significance (Fig. 1).The Kabul basin is made up of sand, gravel conglomerates, and loess loams with sufficient filtration capacity. However, due to population growth, the town expanded widely in an unplanned way. And it created an impervious surface and prevented natural groundwater infiltration. Kabul metropolis has no wastewater collection and treatment. Drainage pit latrines, roadside ditches, sewerage, septic tank leakage, the Kabul River, and irrigation channels are significant pollution sources. Other pollution sources that contaminate the groundwater in the Kabul basin are cesspits, domestic waste, unregulated dispersion of garbage generated, over-exploitation, lack of water supply system, and low public awareness of issues related to water quality. The use of fertilizer for agriculture and landfill sites in highly permeable areas often causes groundwater contamination. Therefore, relatively low precipitation and slow groundwater recharge coupled with groundwater depletion result in negative stability amongst extractions and regeneration. It implies that cumulative groundwater exploitation is unaffordable and that the decreased amounts of groundwater cannot dilute rising groundwater contamination.

Kabul is one of those communities that served nearly all of its water needs by extracting groundwater. Kabul city having a population of about five million people, and almost all of their water demands are fulfilled through groundwater resources. On the other hand, in the present incarnation, groundwater is the dominant accessible source of potable water and all residential use by the capital inhabitants. Previous feasibility studies to expand the water supply infrastructure indicate a projected population of 4,089,000 in 2015, with water requirements of about 123.4 million m^3/yr (Zaryab et al. 2017). However, in Kabul city, the projected supply of groundwater is about 44 million m^3 / yr. The town is a community with limited water resources and less developed in water supply and sanitation. Just 1% of its residents have access to the sewerage network. Only 29% of them are connected to the public water supply system, and more than 70 private water supply companies provide water to the residents. Meanwhile, national distribution access to safe drinking water is 27%, but only about 20% of the rural population has access to clean drinking water (Saffi and Kohistani 2013; Paiman and Noori 2019; Arezoo 2020).

This study aims to audit existing literature on groundwater quality, quantity, status, and trends in Kabul basin. The evaluation makes use of several reports, articles, and published materials from various sources. In the last two decades, in addition to academic researches, many national and international organizations and institutions have worked on groundwater resources of Kabul basin in various scales. Therefore, it is necessary that these works be reviewed in general and its summary be presented in a very concise form. The current study presents details on groundwater status in a comprehensive manner with the goal of enhancing its factual and intellectual knowledge. The extent of future consequences can be deduced from the past status of groundwater. Understanding the cutting-edge scenario will help enhance future research topics and prevent similar and repetitive observations. It will also summarize the outcomes

of preceding studies and pave the way for better groundwater and alternative water resources management.

Climate and hydrology

An Afghan proverb is commonly said for Kabul.: "*May Kabul be without gold, but not without snow*" (Banks 2002). Kabul has an arid to semiarid climate (Qureshi 2002; JICA 2011; FAO 2012; Azimi et al. 2017; Qutbudin et al. 2019; Sayed Mir Agha et al. 2020; Shokory and Rabanizada 2020) with significant daytime and night-time temperature fluctuations. Historical hydro-meteorological records have been reachable from 1956 to 1977. During the common war, information was hindered until 2003 or later. As per chronicled information, the mean yearly precipitation is 330 mm/yr with a limit of 520 mm/yr in 1959 and at least 155 mm/yr in 1970. Normal rainfall (2004–2011) averaged around 312 mm/yr with a limit of 450 mm/yr in 2007 and about 210 mm/yr in 2004 and 2008. Based on recent records of Afghanistan meteorological department (2008–2018), the annual average precipitation measured at around 330 mm per year. Mostly precipitation is happening in the winter and spring. Months having the most noteworthy precipitation are March and April. Throughout the winter, it is snowing in Kabul; however, it is followed by spring rainfall. As per historical snowfall records, the most noteworthy snow was happening in the winter of 1963–1964 by just about a meter of snow. January and February possess the highest snowfall (Tünnermeier and Houben 2005; Houben et al. 2009a; Mack et al. 2009, 2010, 2014; Malyar 2017; Casale et al. 2020). The annual mean temperature in Kabul was increased from 10 °C to 13 °C from 1957 to 1977. The temperature ranges with a monthly mean of a maximum of 32.1 °C in July and a minimum of − 7.1 °C in January. From 1998 to 2004, many parts of Afghanistan, including Kabul, suffered drought. The annual evapotranspiration in the basin is approximately 1600 mm/yr which is a very high value compared to precipitation (Tünnermeier and Houben 2005; Malyar 2017; Zaryab et al. 2017; Mack 2018; Hashmi et al. 2020).

Three waterways have entered the city of Kabul. Paghman stream spilling out of the east to west. The Maidan River (Kabul River) entering into the city from the south and joins the Paghman river after a 21 km flow. The Logar River, which is an enormous tributary of the Kabul River, is flowing south–north and entering the Kabul River around 17 km downstream of the Paghman waterway's mouth (Fig. 1). The typical yearly streamflow of the Paghman waterway in the Hydrological Station of Qala-e-Malek somewhere in the range of 2009 and 2016 was 41 million cubic meters of water. Annually mean flow rate of the Logar river at Sang-e-Naweshta Hydrological Station in the same period was estimated at 231 million cubic meters. The average annual flow rate of the Maidan River at Tangi Saidan Hydrological Station measured 141 million cubic meters of water. According to Minister of Energy and Water (MoEW 2017) and (Zaryab et al. 2017), the aggregate of each of the three streams is around 413 million cubic meters of water. In recent years, the Kabul and Paghman rivers have some discharge only in late winter due to rainfall and also in the spring period due to snow melting and raining. During other seasons, the river beds are either dry or the wastewater from nearby houses and street ditches connected to the rivers flow into them. Comparison of hydrological figures between 1964 and 1974 and 2009–2016 shows that the average water flow in the Paghman, Maidan, and Logar rivers has dropped by almost more than 18 percent (MoEW 2017).

Geology and topography

The investigation region is a part of the Kabul basin that is encircled and underlain through mountain ranges made explicitly out of an assortment of metamorphic rocks. The Kabul basin structure is emerging because of plat developments during the past due to Paleocene time. The rocks are parts of Kabul block and are crossed by the fault of the Herat-Bamyan-Panjsher main fault at the west and north–west, the Sorobi fault at the east, and the Chaman fault framework in the Southeast. The encompassed mountain extents and bedrock are made regularly out of an assortment of transformative rocks, including gneiss, amphibolite, quartzite, slates, schist, marble, sandstone, mica of Precambrian age, and merged rocks (limestone, dolomite) of the Perm-Triassic age. The filling of the Kabul basin itself comprises of an aggregation of earthbound and lacustrine dregs, especially of Quaternary and Neogene ages unconsolidated and semi-combined lacustrine, fluvial, and Aeolian sedimentary rocks comprising of sand, rock, and residue (Saffi 2007; Lashkaripour and Hussaini 2008; Zaryab et al. 2017).

As indicated by an investigation by Japan International Cooperation Agency (JICA 2011) revealed, alluvial deposits comprised of unconsolidated silt, sand, and gravels such as clayey zone partly cover the Neogene in the examination territory with thickness ranging from 34.5 to 44.0 m. Underneath the Alluvial spread, exceptionally thick Neogene stores are dispensed extensively. Neogene deposits within the Kabul basin have been divided into upper Neogene (N_2) formations and the lower Neogene (N_1). The upper Neogene consisted mostly of unconsolidated clay or huge mudstone (consolidated clay). Sometimes it intercalates sandstone layers and includes gravels partially. Lower Neogene consisted of hard clay with gravels, unconsolidated sand and gravels, and consolidated sand and gravels (conglomerates). The upper Neogene thickness was from

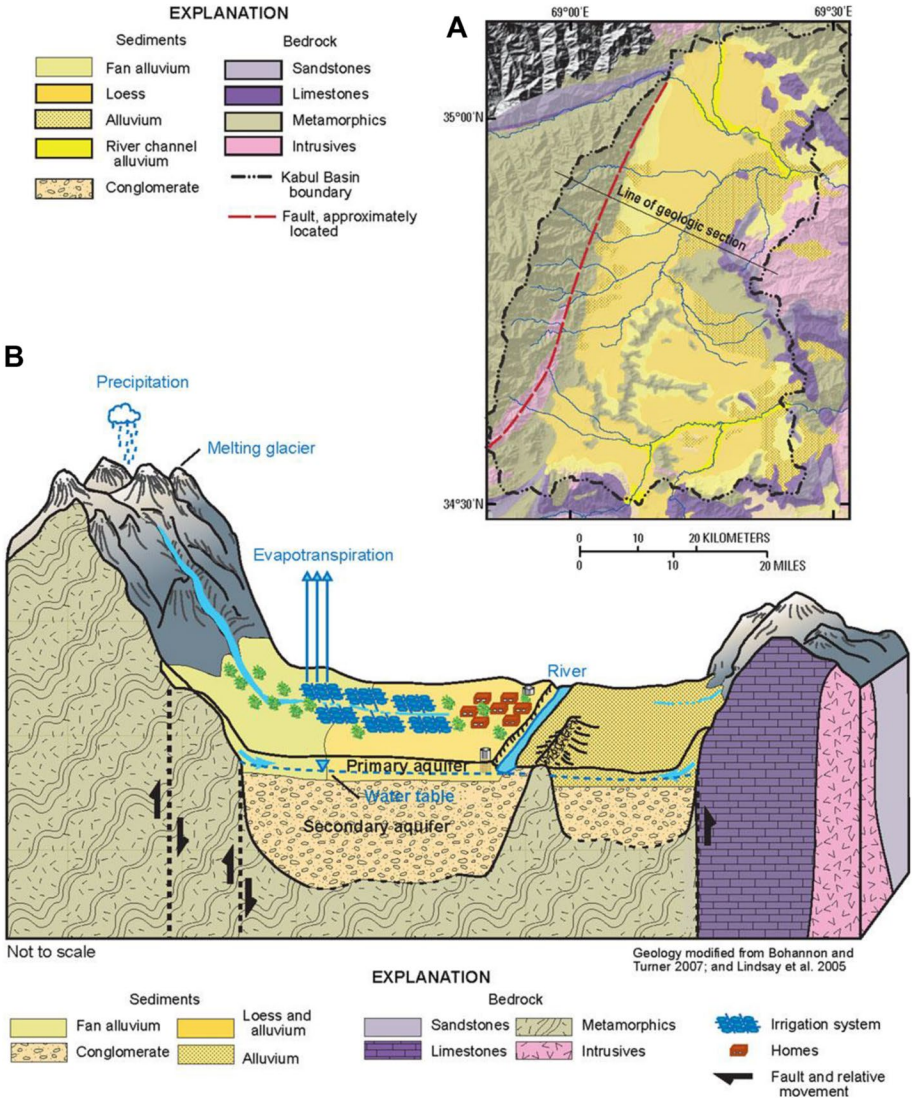


Fig. 2 Generalized surficial geology and cross section of the Kabul Basin, Afghanistan (Mack et al. 2013)

259 to 356 m, whilst that of the lower Neogene has been from more than 252 m to 159 m.

Mountains encircle the Kabul basin with a height of 4400 m at the west (Paghman mountains) and 3000 m at the east (Kohei Saffi mountains). For the most part, the interbasin edges rise around 200–500 m over the nearby valley floors. The sub basin's central parts are commonly plain, developing continuously to including bedrock outcrops (Fig. 2). The focal parts' heights range from about 1800 m within the central sub-basins of Logar and Kabul to 2200 m in the sub-basin of upper Kabul and Paghman. The valley-fill residue has been dismembered by endurant and intermittent stream channels (Fig. 2). The dynamic stream channels are usually narrow and shallow and do not regularly exceed 10 m in width and 5 m in depth (Tünnermeier and Houben 2005; Mack et al. 2009).

Methodology

This review article used data from a variety of sources. Google Scholar was mainly used to get a sense of the various types of data that were available. Additionally, Scopus, the United States Geological Survey (USGS) library, springer link, and science direct were used to find relevant and similar works. Some useful, valuable printed reports and information also were utilized. To create a list of research data, general search terms were primarily seen in online database searches. The procedure began by conducting a simple search on groundwater in Afghanistan. Some ways were figured out to use better alternatives to more refined phrases using online databases based on Google Scholar's data with that search basis.

The following search terms were selected for the current review: *Groundwater, Afghanistan, Kabul basin, Kabul city, groundwater quality, groundwater quantity, hydrogeology, condition, status, review on groundwater, aquifer, recharge, condition, assessment, analysis, study.* In order to achieve the most precisely specified and relevant documents, these phrases became combined in different ways with "AND" and "IN" directives. The majority of search phrases were generated using the findings of a preliminary search in Google Scholar and other databases. A number of papers and reports were found using the Snowball approach in addition to the database quest. In light of the intent of this review article, each of the keywords has been chosen for its suitability and significance.

In this study, the validity and reliability of all information are considered. The following measures were used to determine the resources: (a) Are the data up to date? (b) Is the source relevant to the study objectives? And (c) what are the author's qualifications for the source? The bulk of the references cited were from the last two decades. Those observations were searched in the publication's data that suggested correct analysis protocols had been followed. For groundwater's qualitative status, those samples were looked at which possess qualitative evaluations and highlighted the addressed qualitative issues in the area. On the other hand, for quantitative and groundwater tables, the collected data were scanned for finding the relevant research that addressed groundwater levels. Finally, in assessing qualitative and quantitative data, the collected data were summarized and synthesized to verify groundwater status in the Kabul basin, Afghanistan. The methodological flowchart of the review procedure used to come at this comprehensive review of groundwater status in Kabul, Afghanistan is shown in (Fig. 3).

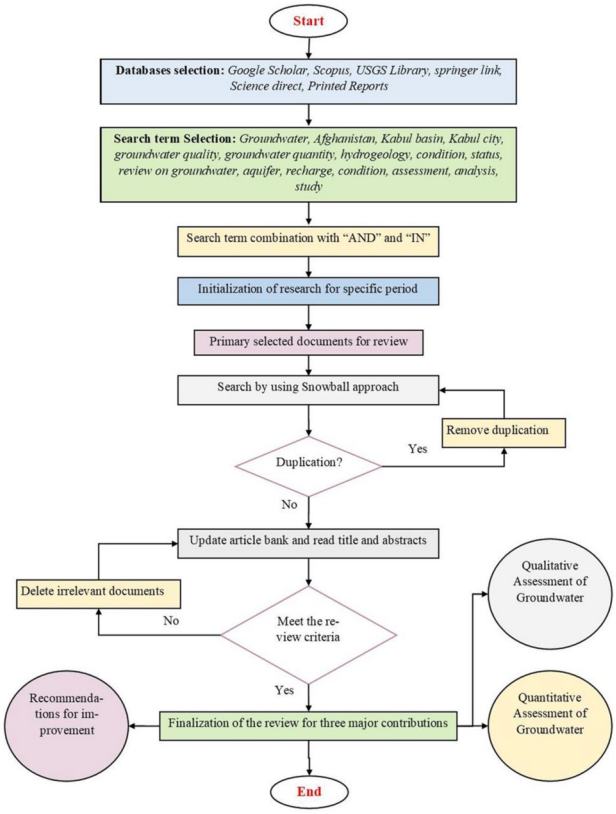


Fig. 3 The methodological flowchart of the review

Spatial and temporal variations of groundwater resources in Kabul city

Groundwater distribution

In the Kabul basin, groundwater flow is predominantly from soaked alluvium and other ground-fill residues. For most parts of the basin, the surface of the water table reflects topography, and generally, groundwater flows in the surface-water discharge direction (Broshears et al. 2005). The illustrated study area has four major interconnected Quaternary aquifers (Fig. 4). The lower Kabul (Sub-basins of Kabul-Logar) comprises two aquifers along the Logar river course and a lower part of the Kabul River. Upper Kabul (sub-basins of Darulaman-Paghman) also includes two aquifers lying along the Paghman river and Kabul river (Uhl and Tahiri 2003; Pell Frischmann 2012; Zaryab et al. 2017). The Logar aquifer is referred to as the Logar aquifer because it

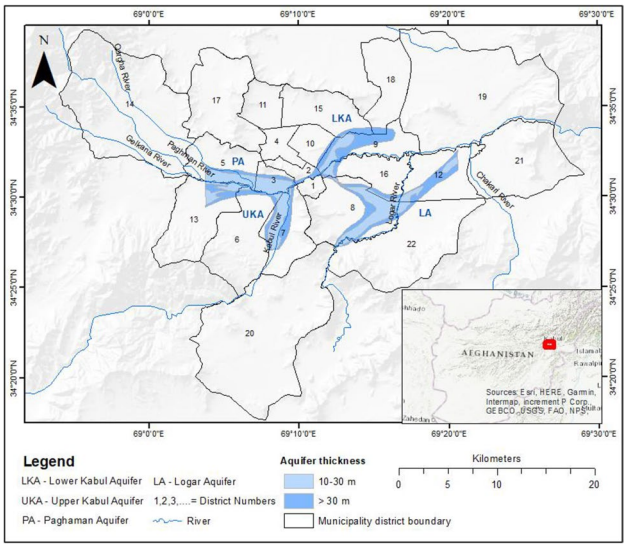


Fig. 4 Spatial distribution of aquifer in the Kabul city (modified from Landell Mills Ltd. 2018)

lies underneath and the Logar river's two sides. This aquifer has a length of approximately 10 km with a width of about 3 km. Its mean thickness is going from thirty to forty meters with a limit of seventy meters. As per pump test results, this

aquifer's water-driven conductivities are high yet dependent upon some heterogeneity (12–112 m/day). The Kabul aquifer is arranged in a coarse-grained deposit with a length of roughly 9 km and a width of around 2,5 km corresponding to the Kabul River. The thickness of the aquifer is ranging between 40 and 80 m. Permeability of the aquifer zone is ranging from 4.32 to 64.8 m/day. The gravel bed placed alongside the Paghman river is around 10 km protracted, 4 km wide, 45 m mean thickness, and close by thicknesses of 70 m. This aquifer is made exceptionally out of sand and gravel. Sandstones and less hard aggregates emerge in minimum amounts at specific regions. It's subsequently contemplated as the superb aquifer in the basin. The conductivity ranges from 1.73 to 25.92 m/day and is consequently penetrable. All through the aquifer, permeability is greater or less uniform (Houben et al. 2009a).

As indicated by Bundesanstalt für Geowissenschaften and Rohstoffe (BGR), the groundwater is recharging when the snowmelt is happening from direct infiltration of the waterways and lower region penetration at the edge of the basin (Niard 2005). Kabul's conventional continental climate, which possesses low precipitation and high evaporation levels, prevents groundwater regeneration explicitly from rainfall. In line with an observation by USGS and the Afghan Geological Survey from 2004 to 2013, water levels beneath the land surface ranged from less than 1.5 to 73.34 m, while static water levels ranged from 1.5 to 40 m. Seasonal water-level variances ranged from under 1 to 8 m from September 2005 to May 2006 (Houben et al. 2009a; JICA 2011; Taher et al. 2013).

The decrease in groundwater levels in Kabul has a long history because of different factors, for example, overexploitation, populace development, and diminishing of recharge zones, which has revealed in numerous reports (Banks 2002; Qureshi 2002; Uhl and Tahiri 2003; Tünnermeier and Houben 2005; Saffi 2007; Houben et al. 2009a; DACAAR 2011; Pell Frischmann 2012; Taher et al. 2013). According to Mack et al. (2013) the decline rate in the city's groundwater levels is higher for the period 2008–2012 (mean 1.5 m/yr) compared with 2004–2008 (mean 0–0.7 m/yr). In most recent years, an excessive decrease in groundwater level has been found. An evaluation from Zaryab et al. (2017) suggests that water levels have decreased by more than 10 m in the foothills and 5 to 6 m in the Kabul city area. The water levels in the upper Kabul basin regions decreased by more than 15 m from 2003 to 2016. From 1982 to 2003, the groundwater level decreased by a mean of 6 m, averaging 0.28 m per year, but from 2003 to 2016, there has been a mean of 15 m of groundwater level reduction, which averages 1.15 m per year. The average annual decrease in groundwater levels between 2008 and 2016 is 1.7 m. About 33% of supply wells are not in operation due to declining water levels. The leading cause for a decreased groundwater level might be the population growth, which has doubled since the 1990s and overexploitation of groundwater.

As indicated by JICA (2011), the aquifer structure of the Kabul basin incorporates three layers: shallow aquifer (alluvial aquifer), deep concludes (upper Neogene Aquifer), and deep aquifer (lower Neogene Aquifer). Outcomes of water quality examinations on the water test taken from all test wells demonstrate that water inside the profound aquifer must be considered "Fossil Water" isolated from the regular water circulation. Isotopic evaluation at the deep and upper deep test wells suggests that groundwater's age must be a glacial epoch. The synthetic and isotopic investigation of surface-water and groundwater tests led by Mack et al. (2010) shows that shallow groundwater (under 100 m subterranean surface) is regularly 20 to 30 years of age, while in more profound aquifers, groundwater is probably going to be thousands of years old.

Groundwater quality

The water quality examinations show a poor pattern and demonstrate that contamination of the groundwater has accelerated continuously. Results from the National Groundwater Monitoring Network of wells covering roughly 80 percent of Afghanistan's river basins for data management, evaluation, and mapping demonstrate that groundwater capacity in Afghanistan has been consistently declining, and water quality is progressively worried due to over-exploitation, low recharge, high evaporation and mismanagement (DACAAR 2011). The water quality examinations show a top-notch variety of groundwater quality explicitly in city environments. The essential qualitative groundwater issues in Kabul are nitrate, saltiness, boron, hardness, and coliform microorganism. Bacterial examinations indicate excessive and increasing degrees of fecal (coliform) bacterial in around 58–70% of the groundwater in Kabul's urban vicinity (Saffi 2011). Since the critical sources of drinking water are shallow groundwater and surface water, pollution-related diseases also affect the population. The observed intensity exceeds recommendations for drinking water by the World Health Organization (WHO) and National Drinking Water Quality Standard (NDWQS) (ANSA 2013).[1]

For nitrates, the evaluations show that one-third of the groundwater has a concentration of more than or equal to 45 mg/L, while the WHO has limited the concentration of nitrates for drinking water to 50 mg/L. An analysis from Danish Committee for Aid to Afghan Refugees (DACAAR 2019) for the entire province of Kabul indicates that the central part of the province, which mostly covers the city has nitrate contamination greater than WHO requirements

[1] Afghanistan National Standards Authority (ANSA).

Table 1 Contaminant concentration in Kabul, Afghanistan. Source: (Hayat and Baba 2017)

Contaminant	Statistics					# Of sample	Acceptable limits	
	Minimum	Maximum	Mean	SD	CoV		WHO	NDWQS
Arsenic	No detectable	0.690	0.003	0.044	14.136	241	0.01	0.05
Fluoride	No detectable	5.200	0.663	0.578	0.871	276	1.5	1.5
Boron	No detectable	4.000	0.784	0.660	0.882	276	2.4	2.4
Sulfate	18.0	1360.0	120.9	150.4	1.2	276	250	250
Nitrate	0.10	96.0	24.06	24.80	1.03	276	50	50
Fecal coliform	No detectable	820	112.32	64.69	5.25	276	0	0
E. Coli	No detectable	> 200				108	0	0

Note: The measuring unit for Arsenic, Flouride, Boron, Sulfate, Nitrate is mg/L, while for Fecal coliform and E. Coli is (Col/100 ml)

(WHO 2017; Brati et al. 2019). In reports, several levels of heavy metals showed comparatively high concentrations than WHO and NDWQS standard values. According to Barati et al. (2019), Mn, Ni, Zn are the elements exceeding the WHO standard level of drinking water. The elements of Cu, Fe, Me, NO_2, PO_4, COD, and TN are under the range of standard.

In line with Gesim and Okazaki (2018) that analyzed physical and chemical characteristics, including metallic elements, the outcomes indicate that the water quality index (WQI) ranges from 36.21 to 716.29. The result shows that only 0.2% of groundwater within the research area has outstanding quality, 19.69% appropriate quality, 62.21% poor quality, 13.65% very poor quality, and 4.25% are unsuitable for drinking. Some chemical elements like Arsenic, Fluoride, Boron, Sulfate, Nitrate, and bacteriological contaminations like Fecal coliform and E. coli are highlighted with higher concentrations than WHO and NDWQS limits in Kabul and Kabul Basin (Hayat and Baba 2017). The summary of the statistics illustrated in Table 1.

An evaluation by Frahmand (2011) also implies that the groundwater in Kabul metropolis and its environs are not of good quality. There is an excessive concentration of chemical elements such as lead and arsenic aligns with Coli and E-Coli. Physical parameters like specific conductivity, alkalinity, and bicarbonates also show higher values close to the city. As one moves far away from the city center, the concentration degree of essential constituents gets decreased, indicating decay postulation. Consequently, there is a high correlation between population concentration and the concentration of foremost constituents in groundwater in the city region.

Comprehensive investigations thru USGS imply that groundwater's essential ion chemistry became commonly of the bicarbonate of calcium–magnesium type. The quantities of numerous segments in evaluations assembled in central Kabul's sub-basin were higher than in assessments from the other sub-basins. The samples regularly reported anthropogenic contaminants' symptoms.

The situations in the aquifers did appear oxic with different levels of oxygen. Of the chemical elements, the amount of nitrate in the most noteworthy condensation was reported at 40.2 mg/lit in central Kabul. The reports explore that lead and uranium in three percent of the sample and arsenic, and selenium in one percent of tested samples exceeded WHO guidelines. The electrical conductivity is also reported at a higher level. Coliform organisms had been reported in almost all groundwater samples with concentrations higher than 100 colonies per100 mL have been recognized in six out of the seven sub-basins. E. coli got identified in 97 percent of groundwater quality assessments, but its concentration seemed to be scattered at random in the basin.

According to the particular CFC proportions surveyed, the groundwater young component's average age in the investigation region becomes roughly 21 yrs. Most of the water tested is reported to be young water. Averaging ages for groundwater in the sub-basin contrasted scarcely depending on the CFC assessment strategy utilized anyway had been around twenty years in the upper Kabul and Paghman sub-basin, 28 years at central Kabul and Deh Sabz sub-basin, and 15–19 years in the Shomali sub-basin. The records from CFC recommend that the age of groundwater expanded with depth. Basin-wide recharge rates have been approximated to be 0.4–0.8 m/yr for 30 percent expected porosity and 0.35–0.7 m/yr for 25 percent expected porosity based on the CFC ages interpretation and the depth at which samples were collected (Mack et al. 2009).

The water evaluation that is conducted via BGR in 2004 covered numerous parameters of water quality. Among physical characters, electrical conductivity was marked higher than WHO limits, and most tested samples were reported as “hard” and “very hard” water. A higher concentration of magnesium, sodium, and chloride are also distinguished. From the Redox potential perspective, the samples lie within the weakly oxidizing zone. In this report, however, nitrate and sulfate pollutions with more than WHO limits also have been reported. A striking element of the samples from the Kabul basin is that lone a couple of tests lie beneath the

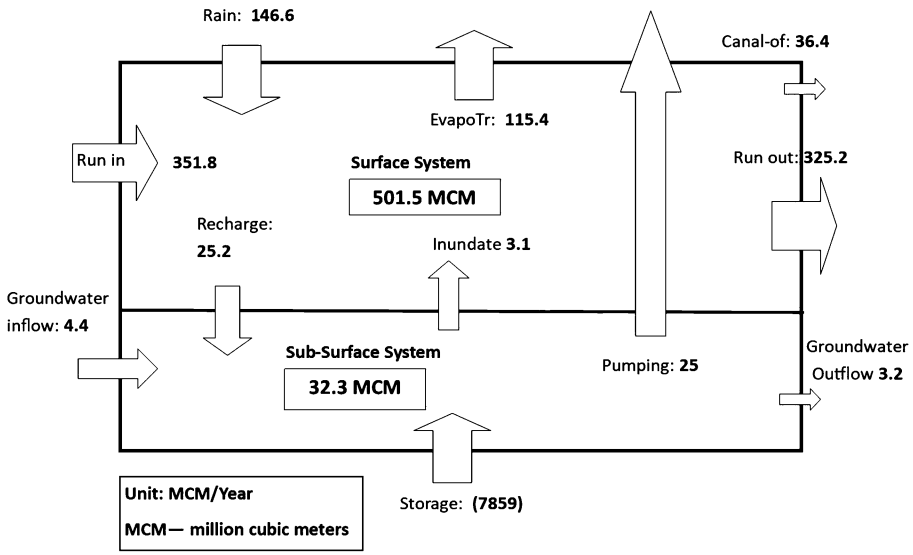


Fig. 5 Total water balance of Kabul basin (recent ten years average) source: (JICA 2011)

restriction of 0.5 mg/l B or two mg/l BO_2. Bacteriological examinations proved that most of the investigated wells had great bacterial content. A number of all explored wells had critical coliform microbes pollution and broadly surpassed the WHO limits of 0 CFU/100 ml (Houben and Tünnermeier 2005; Houben et al. 2009b).

However, an examination by DACAAR on drinking water wells chosen by UNHCR in various parts of the country, whose water is regularly utilized by returnees, additionally shows that the accompanying parameters are higher than WHO limits: physical parameters including EC, Hardness, Turbidity. Chemical variables such as Sulfate, Fluoride, Boron, Sodium, Arsenic, Chromium, and Fecal Coliform's bacteriological pollutants (Saffi and Jawid 2013).

A shallow aquifer quality evaluation by JICA (2011), which is rearranged as "Piper Diagram," implies that water characteristics of most parts of the study area had been placed within the zone of type-II: "Bicarbonate Calcium type," however roughly 30% were allotted in Type-I: "Non-Bicarbonate Calcium type" or in between the type II and I. The former is a typical water quality of a common shallow aquifer fed via rainwater. Still, the latter is an instead unique water quality of so-referred to as "Fossil Water," which is outdoor of the natural hydraulic cycle. It means that many observation wells have reached deep aquifer or have some combination of groundwater of shallow and deep aquifers. The wells labeled into type II water quality are usually freshwater; however, the wells categorized as type I or intermediate among II and I are as a substitute saline water.

To sum up the reviewed existing literature on groundwater quality, the essential qualitative issues of groundwater in Kabul are nitrate, boron, hardness, coliform microorganism such as fecal coliform and E. coli, sulfate, fluoride, heavy metals like Mn, Ni, Zn. as well as Arsenic, Selenium, Uranium, and lead are reported and indicate excessive and increasing degrees in the urban vicinity of Kabul. These are the critical qualitative parameters of groundwater that appear in the Kabul city that exceeded the WHO limits and the NDWQS limits.

The consumption of groundwater resources in Kabul

The total water balance evaluation for shallow aquifer by JICA (2011) indicates that groundwater recharge's primary resources are the surface system of 85.2%, accompanied by the subsurface influx of 14.7%. Groundwater extraction volume is calculated at approximately 25 million m^3/yr, which is about 79.8% and approximately 10.2% of groundwater is lost as outflow. The total storage quantity of groundwater is estimated at 7869 million cubic meters (Fig. 5). The total water balance for groundwater changed into in minus condition of 2.7 million m^3/yr. This minus portion of the groundwater system was supplied from the groundwater storage, and because of this, the groundwater level was drawdown overall.

The consumption of water is associated with the population and the per capita water demand. Previous studies notify that the people in Kabul metropolis had been 320,000 in 1962, 590,000 in 1972, and 1.8 million in 2001. The exact population is not clear yet. Currently, Kabul city's population is estimated at 4.3 million people based on Central Intelligence Agency and National Statistic and Information Authority (CIA 2020; NSIA 2020). On the other hand, the only source for water supply in the city is groundwater sources. Individuals use water through 4 significant suppliers: About 52% of families use their own wells, around 38% of households supplying by commercial water companies, public water supply covers 13% of the populace, and public

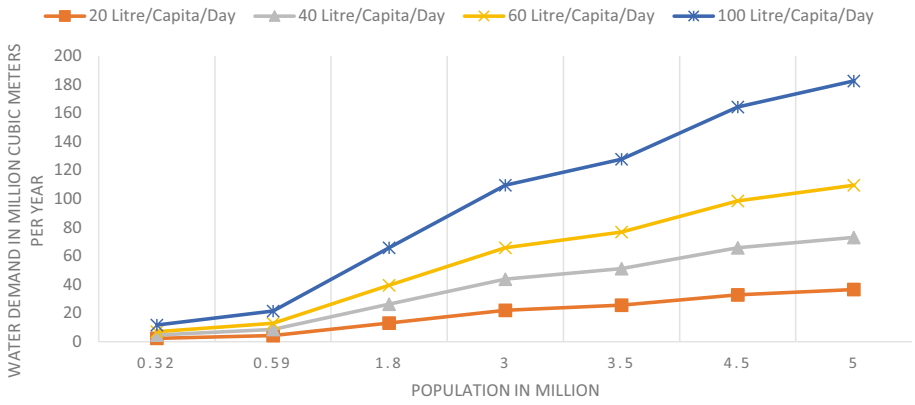


Fig. 6 Estimation of water requirement as a component of water utilization and populace. Produced from (Houben et al. 2009a)

wells 3% (Houben et al. 2009a; Zaryab et al. 2017; Brati et al. 2019). Six percent of the families used both public and private water supply companies. The genuine measurements on groundwater extraction are not accessible. Besides, the exact populace and per capita water utilization are likewise uncertain. Figure 6 shows the assessed amount of yearly water demand based on population growth and assumed water per capita rates.

The anticipated groundwater availability in Kabul is approximately 44×10^6 m^3/year, according to the Kreditanstalt für Wiederaufbau (KfW) study. In line with statistics, the Logar aquifer has the potential of about 24.64 million cubic meters yearly, followed by Allaudin and upper Kabul with a capability of approximately 12.48 million cubic meters annually. The Afshar and lower Kabul aquifers each have the potential of about 3.65 million cubic meters yearly. A study by Houben et al. (2009a) estimates that the entire volume of groundwater recharge in the wet period through all mechanisms might be anywhere between 15×10^6 and 40×10^6 m^3 /yr. In the maximum situation, this study estimates the entire storage volume of groundwater about 2405×10^6 m^3 with a depth of 50 m and effective porosity of 0.13.

Alternatives to drinking water in Kabul

Artificial groundwater recharge

To eliminate water scarcity and overcome excessive abstraction of groundwater, artificial recharge of Kabul Basin groundwater and rainwater harvesting as a source for this purpose is an urgent need for Kabul city (Rhimi and Murakmi 2017; Masoom 2018; Noori and Nasimi 2019). A pilot project on groundwater recharge named Kabul Managed Aquifer Recharge Project (KMARP) is being applied at up to four particular sites. It is an ongoing tentative project to verify groundwater recharge management's feasibility, which is one of the solutions to water scarcity in Kabul. It consists of hydrogeological studies, groundwater recharge experiments, and comparison with other alternatives. If these experiments were successful, the outcomes would lead to a comprehensive project to raise Kabul's groundwater levels and resources (KMARP 2017). The project is supporting financially by the Asian Development Bank (ADB) and the government of Afghanistan. The project start date is February 10, 2017, and the termination date is April 30, 2020.

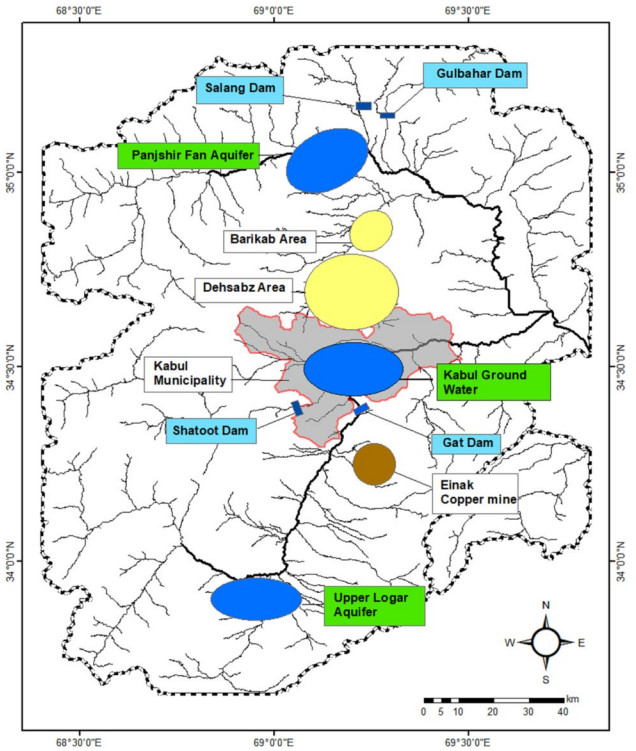


Fig. 7 Location of potential water resources (Reproduced from Zaryab et al. 2017 using Arc GIS and DEM)

Remote sources

The goodness for Kabul metropolis is that numerous surface water resources potentially neighboring it (Fig. 7). Feasibility studies of most dams above these water resources have already been done. But to construct each of these reservoirs, the government needs sufficient financial support. The feasibility study for the Shatoot dam was finished in 2010. The average amount of accessible water is estimated at 87.2×10^6 m^3/yr by constructing this reservoir.

Another under-plan reservoir called the Gulbahar dam is planned to be constructed on the river of Panjsher in the northern part of Kabul city. It is anticipated that the stored water behind it will be a key answer to the constant water shortage in the Kabul region by completing this multipurpose dam. The required feasibility study was led by the JICA and finished in Dec.2012. It is designed as an arch-type concrete dam. The proposed height is 140 m with a capacity of 240×10^6 m^3 storage.

In addition, a feasibility study for the Salang dam was done through JICA and completed in Dec. 2012. This dam is considered a rockfill dam type based on its construction material. The suggested height is 110 m and a capacity limit of 40×10^6 m^3, which can convey water for household and industrial purposes in the new city and guarantee water for ecological and farming utilization.

Outlook

As expressed, the groundwater level seriously declined in Kabul during the previous decades. However, the given information stated differences in statistics. This distinction may be referred to as the particular techniques implemented and the extraordinary observation duration in each case. Residents of Kabul, particularly in the west of the region, adapted to the decrease by depleting additional wells or switching their individual suppliers to a public or private agency. The residential water utilization per capita in Kabul was estimated at 61 L/capita/day (Brati et al. 2019). It appears that precipitation is not impressively changing to decrease the groundwater level. The fundamental reasons for groundwater decrease are the diminishing of its recharge zones and the expansion of water utilization. Inside the Kabul city, groundwater's net recharge is expected to be nearly zero from direct precipitation. The change in land cover from farmland to urban districts is likely to have caused a decline in groundwater revival because city territories have covered all penetration zones in the last two decades.

Managed aquifer recharge strategies must be investigated and implemented to strengthen sustainable groundwater stability, improve water supply safety, and stimulate the conservation and restoration of ecological and environmental systems. To compensate for this reduction, managed aquifer recharge project is a new and valuable step that the government has taken on a tentative basis. The development of this project with distinguishing of groundwater recharge zones and the creation of the appropriate zone maps, separation of artificial recharge methods in accordance with the variety of recharge areas, and having short, medium, and long-term plans is considered a must.

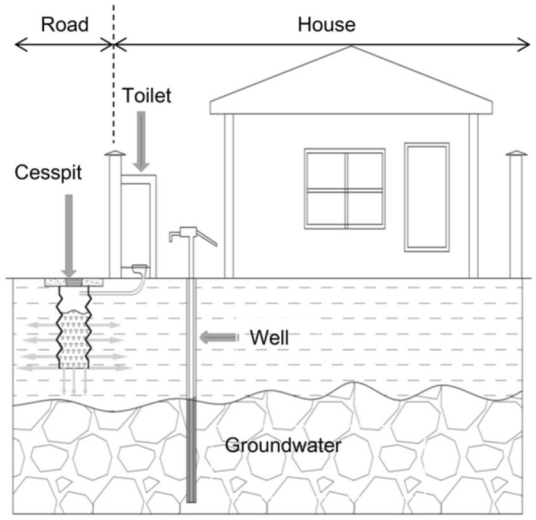


Fig. 8 Typical house with cesspit toilet (Brati et al. 2019)

Another extreme issue is water quality. The city of Kabul does not have a system for collecting and disposing of municipal wastewater. It is a considerable challenge. Only 1% of the households are connected to the sewage system, 49% of the family use holding tanks for disposing of the black-water, 37% use partially lined pits or cesspits (Fig. 8), and 10% dry latrines. Nevertheless, 84% dispose of their grey water directly into the open environment or to the river body, and only 11% use either the same or a different holding tank for the greywater. Five percent of the households additionally use a partially lined pit for the disposal of gray water. Also, 94% of the pits and septic tanks do not have an overflow outlet. Still, if such an outlet has existed there, it is either connected to the open environment (37%) or any other soak pit (63%) (Habib 2014; Paiman and Noori 2019; Rahmani and Anuar 2019). The domestic wastewater collection and disposal system in Kabul is considered a vital necessity. Creating an urban master plan for the sewage collection system should be considered in three short-term, medium-term, and long-term stages on a micro, mini and large scale. Municipality solid waste is counting a factor affecting the pollution of water resources. Solid waste management (collection, recycling, processing, and creating suitable landfill sites) must be systematic and standardized.

Based on integrated groundwater management tools, groundwater monitoring infrastructure must be strengthened and updated. The upgraded monitoring network must include all related organizations. A unit database hub addressing quantitative and qualitative real-time monitoring must be developed to provide timely, accurate, and extensive groundwater framework information to the public. Further investigations are required to identify and classify hydrogeological conditions in order to improve groundwater level and extraction control and resolve geological–environmental issues, including groundwater quality depletion, overexploitation, and land subsidence. Furthermore, additional research on shared groundwater surveillance and taxation, feasibility studies of remote surface sources for water supply, rainwater harvesting, and flush flood control should be conducted in water resource management.

Conclusion

The only source for domestic water uses in Kabul city is groundwater. This study was aimed to audit existing literature on groundwater quality, quantity, status, and trends in Kabul basin. The average annual decrease in groundwater levels between 2008 and 2016 is 1.7 m. Overexploitation, population growth, urban development, land cover changes, and diminishing recharge zones are the leading causes of groundwater level decline. Groundwater extraction volume is calculated at approximately 25 million m^3/yr. Studies in the maximum situation estimate groundwater's entire storage volume about 2405×10^6 m^3. The total water requirement for Kabul is anticipated ranging about 73–183 million cubic meters. The Kabul basin's aquifer structure incorporates three layers: Alluvial Aquifer, Upper Neogene Aquifer, and Lower Neogene Aquifer. Water inside the deep aquifer considered as "Fossil Water" isolated from the regular water circulation.

The physicochemical and bacteriological contents of groundwater in Kabul exceeds WHO and NDWQS limits. The result shows that only 0.2% of groundwater within the research area has outstanding quality, 19.69% appropriate quality, 62.21% poor quality, 13.65% very poor quality, and 4.25% are unsuitable for drinking. Individuals use water through four significant suppliers: Private wells (52%) of families, private suppliers (38%), public water supply (13%), and public wells (3%). To overcome the water scarcity in the Kabul city, two leading solutions have been studied. Artificial groundwater recharge will become more extensively employed if the pilot project "Managed Aquifer Recharge" succeeds. Use of surface water resources from the vicinity of Kabul city suggested as alternatives sources of water supply.

Funding No funds, grants, or other support was received.

Availability of data and material The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability Not applicable.

Declarations

Conflicts of interest Authors do not have any conflict of interest.

References

Abdeslam I, Fehdi C, Djabri L (2017) Application of drastic method for determining the vulnerability of an alluvial aquifer: Morsott - El Aouinet north east of Algeria: Using arcgis environment. Energy Procedia 119:308–317. https://doi.org/10.1016/j.egypro.2017.07.114

Abijith D, Saravanan S, Singh L et al (2020) GIS-based multi-criteria analysis for identification of potential groundwater recharge zones - a case study from Ponnaniyaru watershed, Tamil Nadu, India. HydroResearch 3:1–14. https://doi.org/10.1016/j.hydres.2020.02.002

Achu AL, Thomas J, Reghunath R (2020) Multi-criteria decision analysis for delineation of groundwater potential zones in a tropical river basin using remote sensing, GIS and analytical hierarchy process (AHP). Groundw Sustain Dev 10:100365. https://doi.org/10.1016/j.gsd.2020.100365

Adhikari RK, Mohanasundaram S, Shrestha S (2020) Impacts of land-use changes on the groundwater recharge in the Ho Chi Minh city, Vietnam. Environ Res 185:109440. https://doi.org/10.1016/j.envres.2020.109440

Ahmad A, Cornelissen E, van de Wetering S et al (2018) Arsenite removal in groundwater treatment plants by sequential Permanganate-Ferric treatment. J Water Process Eng 26:221–229. https://doi.org/10.1016/j.jwpe.2018.10.014

ANSA (2013) National Drinking Water Quality Standard. ANSA, Kabul, Afghanistan

Arefin R (2020) Groundwater potential zone identification at Plio-Pleistocene elevated tract, Bangladesh: AHP-GIS and remote sensing approach. Groundw Sustain Dev 10:100340. https://doi.org/10.1016/j.gsd.2020.100340

Arezoo TV (2020) Interview with Engineer Hamidullah Yalani, President of Afghanistan Urban Water Supply and Sewerage Corporation (AUWSSC)

Awawdeh M, Obeidat M, Zaiter G (2015) Groundwater vulnerability assessment in the vicinity of Ramtha wastewater treatment plant, North Jordan. Appl Water Sci 5:321–334. https://doi.org/10.1007/s13201-014-0194-6

Azimi AM, Mumtaz A, Saleh GA, Adam A (2017) Country environment profile for Afghanistan. Natioan Environmental Protection Agency

Banks D (2002) Guidelines for Sustainable Use of Groundwater in Afghanistan

Bohannon R (2010) Geologic and Topographic Maps of the Kabul South 30' × 60' Quadrangle, Afghanistan: U.S. Geological Survey Scientific Investigations Map 3137

Brati MQ, Ishihara MI, Higashi O (2019) Groundwater level reduction and pollution in relation to household water management in Kabul, Afghanistan. Sustain Water Resour Manag 5:1315–1325. https://doi.org/10.1007/s40899-019-00312-7

Broshears RE, Akbari MA, Chornack MP, et al (2005) Inventory of ground-water resources in the Kabul Basin, Afghanistan. USGS

Casale F, Bombelli GM, Monti R, Bocchiola D (2020) Hydropower potential in the Kabul River under climate change scenarios in the XXI century. Theor Appl Climatol 139:1415–1434. https://doi.org/10.1007/s00704-019-03052-y

Chidambaram S, Sarathidasan J, Srinivasamoorthy K et al (2018) Assessment of hydrogeochemical status of groundwater in a coastal region of Southeast coast of India. Appl Water Sci 8:1–14. https://doi.org/10.1007/s13201-018-0649-2

CIA (2020) The World Fact Book. https://www.cia.gov/library/publications/the-world-factbook/geos/af.html. Accessed 30 May 2020

DACAAR (2011) National Groundwater Monitoring Wells Network in Afghanistan. DACAAR, Kabul, Afghanistan

DACAAR (2019) Hydro-geological Booklet Kabul Province. DACAAR, Kabul, Afghanistan

Dinh M, Hakimabadi SG, Pham ALT (2020) Treatment of sulfolane in groundwater: A critical review. J Environ Manage 263:110385. https://doi.org/10.1016/j.jenvman.2020.110385

FAO (2012) Country profile-Afghanistan. Italy, Rome

Frahmand AS (2011) GIS based Ground Water Quality Analysis in Kabul City Region. Geospatial World Forum 1–13

Gesim NA, Okazaki T (2018) Groundwater Quality Mapping Using Water Quality Index in the Kabul City, Afghanistan. In: Proceedings of ISER 149th International Conference, Kyoto, Japan, 9th-10th August 2018. pp 29–34

Habib H (2014) Water Related Problems in Afghanistan. Int J Educ Stud 01:137–144

Hall B, Currell M, Webb J (2020) Using multiple lines of evidence to map groundwater recharge in a rapidly urbanising catchment: Implications for future land and water management. J Hydrol 580:124265. https://doi.org/10.1016/j.jhydrol.2019.124265

Hao A, Zhang Y, Zhang E et al (2018) Review: Groundwater resources and related environmental issues in China. Hydrogeol J 26:1325–1337. https://doi.org/10.1007/s10040-018-1787-1

Haritash AK, Mathur K, Singh P, Singh SK (2017) Hydrochemical characterization and suitability assessment of groundwater in Baga-Calangute stretch of Goa, India. Environ Earth Sci 76:1–10. https://doi.org/10.1007/s12665-017-6679-5

Haritash AK, Aggarwal A, Soni J et al (2018) Assessment of fluoride in groundwater and urine, and prevalence of fluorosis among school children in Haryana, India. Appl Water Sci 8:1–8. https://doi.org/10.1007/s13201-018-0691-0

Hashmi MZUR, Masood A, Mushtaq H et al (2020) Exploring climate change impacts during first half of the 21st century on flow regime of the transboundary kabul river in the hindukush region. J waJournal Water Clim Chang 11:1521–1538. https://doi.org/10.2166/wcc.2019.094

Hayat E, Baba A (2017) Quality of groundwater resources in Afghanistan. Environ Monit Assess. https://doi.org/10.1007/s10661-017-6032-1

Hou Q, Zhang Q, Huang G et al (2020) Elevated manganese concentrations in shallow groundwater of various aquifers in a rapidly urbanized delta, south China. Sci Total Environ 701:134777. https://doi.org/10.1016/j.scitotenv.2019.134777

Houben G, Niard N, Tünnermeier T, Himmelsbach T (2009a) Hydrogeology of the Kabul Basin (Afghanistan), part I: Aquifers and hydrology. Hydrogeol J 17:665–677. https://doi.org/10.1007/s10040-008-0377-z

Houben G, Tünnermeier T, Eqrar N, Himmelsbach T (2009b) Hydrogeology of the Kabul Basin (Afghanistan), part II: Groundwater geochemistry. Hydrogeol J 17:935–948. https://doi.org/10.1007/s10040-008-0375-1

Houben G, Tünnermeier T (2005) Hydrogeology of the Kabul Basin Part II : Groundwater geochemistry and microbiology. BGR

Huang G, Liu C, Li L et al (2020) Spatial distribution and origin of shallow groundwater iodide in a rapidly urbanized delta: A case study of the Pearl River Delta. J Hydrol 585:124860. https://doi.org/10.1016/j.jhydrol.2020.124860

JICA (2011) The study on groundwater resources potential in Kabul basin in the Islamic republic of Afghanistan final report. JICA

Kalhor K, Emaminejad N (2019) Sustainable development in cities: Studying the relationship between groundwater level and urbanization using remote sensing data. Groundw Sustain Dev 9:100243. https://doi.org/10.1016/j.gsd.2019.100243

Kalhor K, Ghasemizadeh R, Rajic L, Alshawabkeh A (2019) Assessment of groundwater quality and remediation in karst aquifers: A review. Groundw Sustain Dev 8:104–121. https://doi.org/10.1016/j.gsd.2018.10.004

Karami S, Madani H, Katibeh H, Fatehi Marj A (2018) Assessment and modeling of the groundwater hydrogeochemical quality parameters via geostatistical approaches. Appl Water Sci 8:1–13. https://doi.org/10.1007/s13201-018-0641-x

Khan MF, Yu L, Achari G (2020) Field evaluation of a pressurized ozone treatment system to degrade sulfolane in contaminated groundwaters. J Environ Chem Eng. https://doi.org/10.1016/j.jece.2020.104037

KMARP (2017) Kabul managed aquifer recharge project preparation. In: KMARP. https://kmarp.com/home. Accessed 2 May 2020

Krishna Kumar S, Logeshkumaran A, Magesh NS et al (2015) Hydrogeochemistry and application of water quality index (WQI) for groundwater quality assessment, Anna Nagar, part of Chennai City, Tamil Nadu, India. Appl Water Sci 5:335–343. https://doi.org/10.1007/s13201-014-0196-4

Landell Mills Ltd (2018) Environmental Due Diagence report on Proposed Pilot MAR Sites 3,4 and 6 Part 1 of 3

Lashkaripour GR, Hussaini SA (2008) Water resource management in Kabul river basin, eastern Afghanistan. Environmentalist 28:253–260. https://doi.org/10.1007/s10669-007-9136-2

Lentswe GB, Molwalefhe L (2020) Delineation of potential groundwater recharge zones using analytic hierarchy process-guided GIS in the semi-arid Motloutse watershed, eastern Botswana. J Hydrol Reg Stud 28:100674. https://doi.org/10.1016/j.ejrh.2020.100674

Mack TJ, Chornack MP, Taher MR (2013) Groundwater-level trends and implications for sustainable water use in the Kabul Basin, Afghanistan. Environ Syst Decis 33:457–467. https://doi.org/10.1007/s10669-013-9455-4

Mack TJ, Akbari MA, Ashoor MH, et al (2009) Conceptual Model of Water Resources in the Kabul Basin, Afghanistan. USGS

Mack TJ, Chornack MP, Coplen TB, et al (2010) Availability of Water in the Kabul Basin, Afghanistan. USGS

Mack TJ, Chornack MP, Verstraeten IM (2014) Chapter 11 Sustainability of Water Supply at Military Installations, Kabul Basin, Afghanistan. pp 199–226

Mack TJ (2018) Chapter 2 Groundwater Availability in the Kabul Basin, Afghanistan. In: Mukherjee A. (eds) Groundwater of South Asia. Singapore: Springer, Singapore. Springer Nature Singapore Pte Ltd., pp 247–255

Malik VK, Singh RK, Singh SK (2010) Impact of Urbanization on Groundwater of Gurgaon District, Haryana, India. Int J Rural Dev Manag Stud 5:45–57

Malyar I (2017) A Review of Current and Possible Future Relations in Kabul River Basin. Duran Publications, Kabul, Afghanistan

Masoom MF (2018) Artificial Recharge of Groundwater as a Management Tool for the Kabul Basin. Louisiana state University, Afghanistan

Mengistu HA, Demlie MB, Abiye TA (2019) Review: Groundwater resource potential and status of groundwater resource development in Ethiopia. Hydrogeol J 27:1051–1065. https://doi.org/10.1007/s10040-019-01928-x

MoEW (2017) The plan to identify surface and groundwater resources for sustainable drinking water supply for Kabul city (in persian language). MoEW
Naik PK, Tambe JA, Dehury BN, Tiwari AN (2008) Impact of urbanization on the groundwater regime in a fast growing city in central India. Environ Monit Assess 146:339–373. https://doi.org/10.1007/s10661-007-0084-6
Niard N (2005) Hydrogeology of the kabul basin part III: Modelling approach conceptual and numerical groundwater models. BGR
Niu B, Loáiciga HA, Wang Z et al (2014) Twenty years of global groundwater research: A Science Citation Index Expanded-based bibliometric survey (1993–2012). J Hydrol 519:966–975. https://doi.org/10.1016/j.jhydrol.2014.07.064
Noori KMA, Nasimi MN (2019) Kabul City Groundwater and Need for Artificial Recharge. Proc 4th Int Conf Civil, Struct Transp Eng https://doi.org/10.11159/iccste19.215
NSIA (2020) Estimation-Population-Final-Year 2020
Oroji B (2019) Groundwater vulnerability assessment with using GIS in Hamadan-Bahar plain. Iran Appl Water Sci 9:1–13. https://doi.org/10.1007/s13201-019-1082-x
Paiman Z, Noori AR (2019) Evaluation of Wastewater Collection and Disposal in Kabul City and Its Environmental Impacts. Mod Environ Sci Eng 5:451–458. https://doi.org/10.15341/mese(2333-2581)/05.05.2019/012
Patel PM, Saha D, Shah T (2020) Sustainability of groundwater through community-driven distributed recharge: An analysis of arguments for water scarce regions of semi-arid India. J Hydrol Reg Stud 29:100680. https://doi.org/10.1016/j.ejrh.2020.100680
Pell Frischmann (2012) Afghanistan Resource Corridor Development: Water Strategy Final Kabul River Basin Report Version 4.0
Prabhu MV, Venkateswaran S (2015) Delineation of Artificial Recharge Zones Using Geospatial Techniques in Sarabanga Sub Basin Cauvery River, Tamil Nadu. Aquat Procedia 4:1265–1274. https://doi.org/10.1016/j.aqpro.2015.02.165
Qureshi AS (2002) Water resources management in Afghanistan. International Water Management Institute
Qutbudin I, Shiru MS, Sharafati A, et al (2019) Seasonal Drought Pattern Changes Due to Climate Variability: Case Study in Afghanistan. Water
Rahmani H, Anuar AN (2019) Challenges and resolutions for sustainable domestic wastewater management in Kabul City, Afghanistan. Int J Eng Adv Technol 9:1394–1401. https://doi.org/10.35940/ijeat.A1222.109119
Rhimi O, Murakmi K (2017) Rooftop rainwater harvesting and its efficiency in Kabul new city. 土木学会論文集G (環境) 73:25–30
Sadat-Noori SM, Ebrahimi K, Liaghat AM (2014) Groundwater quality assessment using the Water Quality Index and GIS in Saveh-Nobaran aquifer. Iran Environ Earth Sci 71:3827–3843. https://doi.org/10.1007/s12665-013-2770-8
Saffi MH, Jawid A (2013) UNHCR public drinking water points analysed water quality data evaluation in DACAAR
Saffi MH, Kohistani AJ (2013) Water Resources Potential, Quality Problems, Challenges and Solutions in Afghanistan. DACAAR
Saffi MH (2007) Ground water resources at risk in Afghanistan. DACAAR
Saffi MH (2011) Groundwater natural resources and quality concern in Kabul Basin, Afghanistan Scientific Investigation Report in Afghanistan. DACAAR
Saraswat C, Kumar P, Dasgupta R et al (2019) Sustainability assessment of the groundwater quality in the Western India to achieve urban water security. Appl Water Sci 9:1–17. https://doi.org/10.1007/s13201-019-0956-2
Sayed Mir Agha M, Kamarul Azlan Mohd N, Mohammed Sanusi S et al (2020) Urban Flooding in the Northern Part of Kabul City : Causes and Mitigation. Earth Syst Environ. https://doi.org/10.1007/s41748-020-00165-7
Selvakumar S, Ramkumar K, Chandrasekar N et al (2017) Groundwater quality and its suitability for drinking and irrigational use in the Southern Tiruchirappalli district, Tamil Nadu, India. Appl Water Sci 7:411–420. https://doi.org/10.1007/s13201-014-0256-9
Senanayake IP, Dissanayake DMDOK, Mayadunna BB, Weerasekera WL (2016) An approach to delineate groundwater recharge potential sites in Ambalantota, Sri Lanka using GIS techniques. Geosci Front 7:115–124. https://doi.org/10.1016/j.gsf.2015.03.002
Shokory JAN, Rabanizada E (2020) Sustainable household water-saving and demand management options for Kabul City. IOP Conf Ser Earth Environ Sci. https://doi.org/10.1088/1755-1315/511/1/012003
Singaraja C (2017) Relevance of water quality index for groundwater quality evaluation: Thoothukudi District, Tamil Nadu, India. Appl Water Sci 7:2157–2173. https://doi.org/10.1007/s13201-017-0594-5
Srinivas Y, Aghil TB, Hudson Oliver D et al (2017) Hydrochemical characteristics and quality assessment of groundwater along the Manavalakurichi coast, Tamil Nadu, India. Appl Water Sci 7:1429–1438. https://doi.org/10.1007/s13201-015-0325-8
Taher MR, Chornack MP, Mack TJ (2013) Groundwater Levels in the Kabul Basin, Afghanistan, 2004–2013. USGS
Tünnermeier T, Houben DG (2005) Hydrogeology of the Kabul Basin Part I: Geology, aquifer characteristics, climate and hydrography. BGR
Uhl WV, Tahiri MQ (2003) An overview of groundwater resources and challenges. Uhl Uhl, Baron, Rana Associates, Inc., Washington Crossing, PA, USA, Vincent W
Van Nguyen D, Ho NM, Hoang KD et al (2020) An investigation on treatment of groundwater with cold plasma for domestic water supply. Groundw Sustain Dev 10:100309. https://doi.org/10.1016/j.gsd.2019.100309
Varol S, Davraz A (2015) Evaluation of the groundwater quality with WQI (Water Quality Index) and multivariate analysis: a case study of the Tefenni plain (Burdur/Turkey). Environ Earth Sci 73:1725–1744. https://doi.org/10.1007/s12665-014-3531-z
Verma P, Singh PK, Sinha RR, Tiwari AK (2020) Assessment of groundwater quality status by using water quality index (WQI) and geographic information system (GIS) approaches: a case study of the Bokaro district, India. Appl Water Sci 10:1–16. https://doi.org/10.1007/s13201-019-1088-4
WHO (2017) Guidelines for Drinking water Quality fourth edition incorporating the first addendum. Geneva
Yeh HF, Cheng YS, Lin HI, Lee CH (2016) Mapping groundwater recharge potential zone using a GIS approach in Hualian River. Taiwan Sustain Environ Res 26:33–43. https://doi.org/10.1016/j.serj.2015.09.005
Zaryab A, Noori AR, Wegerich K, Kløve B (2017) Assessment of water quality and quantity trends in Kabul aquifers with an outline for future drinking water supplies. Cent Asian J Water Res 3:3–11
Zhang F, Huang G, Hou Q et al (2019) Groundwater quality in the Pearl River Delta after the rapid expansion of industrialization and urbanization: Distributions, main impact indicators, and driving forces. J Hydrol 577:124004. https://doi.org/10.1016/j.jhydrol.2019.124004

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

Indian J Phys
https://doi.org/10.1007/s12648-021-02187-x

ORIGINAL PAPER

Study of contribution of doubly excited $3d^{10}$ configurations in excitation energies and SXR transition data of Fe-like ions

R Sharma[1] and A Goyal[2]*

[1]Department of Applied Physics, Delhi Technological University, Delhi 110042, India

[2]Department of Physics, Shyamlal College, University of Delhi, Delhi 110032, India

Received: 24 May 2021 / Accepted: 28 July 2021

Abstract: We present calculations of excitation energies, emitted photon wavelengths during the transition from upper level to lower level, transitions rates and other transition parameters for lowest 200 fine structure levels of Fe-like Xe, Cs and Ba ions by using multi-configuration Dirac–Fock (MCDF) technique. The significant contribution of the relativistic and quantum electrodynamic effects (QED) in addition to Breit corrections in energies of levels is observed and reported in this work. The calculated energies from both flexible atomic code (FAC) and General Purpose Atomic Structure Package (GRASP) are compared with the critically evaluated experimental NIST and other recent reported experimental and theoretical results. The computed data validate the experimentally available data. We have estimated the wavelengths of all soft X-ray (SXR) from the ground state in Fe-like isoelectronic sequence. Furthermore, sufficient new radiative data of Fe-like ions, which have not been published up to our knowledge, have been reported in present paper. Due to high abundance of iron ($Z = 26$) in cosmic environments, the present study can be beneficial in the field of modeling of fusion and astrophysical plasmas devices such as tokamaks, laser spectroscopy, electron beam ion trap (EBIT) cell biology and microscopic imaging.

Keywords: Soft X-ray; Quantum electrodynamics; Radiative data; MCDF

1. Introduction

The neutral and ionized xenon, cesium and barium have been used for the spectroscopic study of several plasmas. Their large atomic mass results in small Doppler width make them an attractive option for tokamak and other high-temperature plasma devices. The X-ray spectra of high Z elements such as xenon, cesium and barium are of special interest due to applications in charged plasma diagnostics, lithography and X-ray laser research. The ions of iron-group elements have complex spectra due to open 3d shells and contain thousands of lines. They are important in spectra of interstellar medium (ISM), hot and cool stars, galaxies, nebulae and quasi-stellar objects. Problems in modeling spectra in UV and X-ray range are mainly due to lack of precise atomic data of higher excited states of iron-group elements. Further, the urgent requirement of new energy sources has forced developed countries to collaborate or participate in the demonstration of scientific and technological achievability of magnetic fusion through ITER to produces ten times more power than required to heat the plasma [1] and this demands the atomic data of highly charged ion. Recently, tungsten is the most probable candidate but due to high radiative efficiency of tungsten ions, and the influx of tungsten to burning plasma is topic of diagnosis and require stringent control. Thus, the analysis and study of higher ionization states of inert gases such as xenon and other high Z ions can be used understand the heat dissipation flux in ITER divertor. Therefore, in the present manuscript, we have provided the atomic data of Fe-like Xe, Cs and Ba ions.

In the past few years, doubly excited states of positive ions have been observed in collision of slow multiply charged ions with atoms or molecules. These states are mainly formed due to electron capture or by transfer excitation process. But the major constraint in the study of these collision dynamics is the lack of understanding and

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12648-021-02187-x.

*Corresponding author, E-mail: arun.goyal.du@gmail.com

Published online: 28 August 2021

© 2021 IACS

availability of atomic and radiative data of the highly charged mid-Z elements with doubly excited states. High-precision atomic calculations are required to understand complex collision processes. Therefore, in the present work, we study the effect of doubly excited states on energy levels of other states for Fe-like ions.

2. Available data of Fe-like ions

In past few decades, only limited work is published in the literature. Carlson et al. [2] had presented ionization potentials for all ionization states of all elements up to atomic number Z = 103 including Fe-like ions. Wyart et al. [3] have observed spectra of Xenon ions including Fe-like Xe from TFR tokamak plasma. Further, Ekberg et al. [4] have observed the spectra of Fe-like ions from Ru to Gd from laser produced plasmas and grazing incidence spectrographs in the wavelength range from 2.6 nm to 7.8 nm. Doron et al. [5, 6] have studied the X-ray spectrum of xenon and barium from laser produced plasma in the wavelength range 8–13.5 Å and 16.5–19.5 Å, respectively, and identified spectral lines of Xe and Ba ions. Nagles et al. [7] have also observed spectra of highly ionized xenon from laser irradiated gas jet in the wavelength range 0.5 nm to 1 nm. Rodrigues et al. [8] have calculated the total atomic energies of ground state configurations for several heavier and highly charged ions in the Dirac–Fock approximation. Ding et al. [9] have studied X-ray spectra of highly charged xenon plasmas by using multi-configuration Dirac–Fock (MCDF) method. Osin et al. [10] have studied and analyzed spectra of highly ionized xenon ions with an electron beam ion trap (EBIT) in the wavelength range 4.5–19.5 nm. Since most of the work available in the literature is experimental therefore, our aim is to determine accurate and big theoretical atomic data of Fe-like ions.

3. Theoretical method

Several codes and online databases are available in the literature to predict and assess the accuracy of atomic structure and radiative data of ions that are of interest in various applications in atomic physics research, such as RELAC [11], Cowan code [12], HULLAC [13], GRASP [14] and GRASP2k [15, 16], MCDFGME [17, 18], RMBPT [19] and FAC [20]. Recently, few fully independent relativistic codes, such as FAC, GRASP, CI-MBPT [21], DIRAC [22], have become the most widely used codes. Comparison of GRASP and FAC results for same set of configurations helps to understand and analysis to know which theoretical contributions are included in the calculations and which code or method may potentially produce discrepancies and up to which factor and which one is taking more computational time. This also helps is in verification of results for same set of configurations from different codes and methods. In this case, the relative roles of the electron correlation contribution and the Breit and QED contributions to perform independent calculations are also important and therefore examined in the present work. To check our results especially for higher excited states from GRASP which are not verified from the literature, another independent flexible atomic code developed by Gu is used [20].

GRASP and FAC are based on multi-configuration Dirac–Fock method and modified multi-configuration Dirac–Hartree–Fock–Slater (DHFS) model, respectively. The notified difference between both is replacement of non-localized potential with localized potential in FAC. Further, DHFS method in FAC takes an approximate form of electron–electron interaction potential.

4. Computational procedure

The computational method we have adopted in our calculations involves the single excitations from outermost orbital 3d and within $n = 3$. For one electron excitations to outer orbitals, we have increased principal quantum number for outer orbitals one by one up to $n = 10$ and $l = 4$ to observe which configuration is contributing or not.

In the final step, we have taken into account the double electron excitations from 3s and 3p orbitals to 3d orbital to see the effect of $3d^{10}$ configurations.

5. Results and discussion

5.1. Excitation energies of Fe-like Xe

We have computed fine-structure energies of lowest 200 levels belonging to non-relativistic configurations $3s^23p^63d^8$, $3s^23p^53d^9$, $3s3p^63d^9$, $3s^23p^63d^74l$ ($l = 0, 1, 2$), $3s^23p^43d^{10}$ and $3s3p^53d^{10}$ using both GRASP and FAC codes which are listed in Table 1. The individual contributions of Dirac–Coulomb, relativistic corrections Breit and QED are listed separately in the table. It is evident from Table 1 that contribution of Breit correction is significant in comparison with QED for most of the levels.

Table 1 Excitation energies (in Ryd.) for lowest 200 levels for Fe-like Xe

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF5				FAC	NIST [23]	
		DC	Breit	QED	Total			
1	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3F_2$) $^3F_4^e$	0.0000	–	–	0.0000	0.0000	0.0000	–
2	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^1D_2$) $^1D_2^e$	4.1715E−01	1.6200E−02	− 2.8400E−05	4.3329E−01	4.3337E−01	4.2483E−01	0.0085
3	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3F_2$) $^3F_3^e$	9.4494E−01	− 2.2600E−02	2.3400E−06	9.2232E−01	9.2340E−01	9.2694E−01	0.0046
4	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3P_2$) $^3P_0^e$	1.1708E+00	2.3500E−02	− 2.4300E−05	1.1942E+00	1.1947E+00		
5	$3s1^23p^6$ ($^1S_0$) $3d^8$ ($^3P_2$) $^3P_2^e$	1.3570E+00	− 2.1200E−02	− 2.1900E−06	1.3358E+00	1.3372E+00	1.3287E+00	0.0071
6	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3P_2$) $^3P_1^e$	1.6000E+00	− 4.2200E−03	9.7700E−08	1.5957E+00	1.5975E+00	1.5823E+00	0.0134
7	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^1G_2$) $^1G_4^e$	1.7394E+00	− 3.5600E−02	2.7000E−06	1.7038E+00	1.7060E+00	1.6766E+00	0.0272
8	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3F_2$) $^3F_2^e$	2.2418E+00	− 4.9900E−02	− 1.8400E−05	2.1919E+00	2.1941E+00	2.1992E+00	0.0073
9	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^1S_0$) $^1S_0^e$	3.4693E+00	− 2.4000E−02	− 3.0000E−05	3.4453E+00	3.4472E+00		
10	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3F_4^o$	1.8122E+01	− 4.4700E−02	− 5.2800E−03	1.8072E+01	1.8012E+01	1.7838E+01	0.2340
11	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^1D_2^o$	1.8512E+01	− 3.9400E−02	− 5.2600E−03	1.8467E+01	1.8407E+01	1.8194E+01	0.2730
12	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3D_3^o$	1.9153E+01	− 5.2500E−02	− 5.2500E−03	1.9095E+01	1.9034E+01	1.8779E+01	0.3160
13	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3P_1^o$	1.9730E+01	− 3.8800E−02	− 5.3400E−03	1.9686E+01	1.9615E+01	1.9364E+01	0.3220
14	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3D_2^o$	1.9863E+01	− 6.6900E−02	− 5.2500E−03	1.9791E+01	1.9735E+01	1.9477E+01	0.3140
15	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3P_0^o$	2.0119E+01	− 5.5500E−02	− 5.3100E−03	2.0059E+01	2.0000E+01		
16	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3F_3^o$	2.0322E+01	− 6.9400E−02	− 4.7200E−03	2.0247E+01	2.0185E+01	1.9924E+01	0.3230
17	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3D_1^o$	2.0374E+01	− 5.7300E−02	− 5.3000E−03	2.0312E+01	2.0246E+01	1.9974E+01	0.3380
18	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3P_2^o$	2.3873E+01	− 1.0400E−01	1.1900E−03	2.3771E+01	2.3731E+01	2.3519E+01	0.2520
19	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3F_2^o$	2.4108E+01	− 1.4400E−01	1.2300E−03	2.3965E+01	2.3940E+01	2.3737E+01	0.2280
20	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^1F_3^o$	2.4533E+01	− 1.1400E−01	6.8100E−04	2.4420E+01	2.4369E+01	2.4050E+01	0.3700
21	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^1P_1^o$	2.5640E+01	− 1.3200E−01	9.5500E−04	2.5508E+01	2.5470E+01		
22	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^3D_3^e$	3.4132E+01	− 3.6100E−02	− 7.0400E−02	3.4026E+01	3.3938E+01		
23	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^3D_2^e$	3.4390E+01	− 3.6000E−02	− 6.8300E−02	3.4286E+01	3.4190E+01		
24	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^3D_1^e$	3.5172E+01	− 6.0600E−02	− 7.0400E−02	3.5041E+01	3.4963E+01		
25	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^1D_2^e$	3.5372E+01	− 6.7600E−02	− 6.4700E−02	3.5239E+01	3.5140E+01		
26	$3s^23p^4$ ($^3P_2$) $3d^{10}$ ($^1S_0$) $^3P_2^e$	3.8313E+01	− 8.5100E−02	− 1.5500E−02	3.8212E+01	3.7927E+01		
27	$3s^23p^4$ ($^1S_0$) $3d^{10}$ ($^1S_0$) $^1S_0^e$	3.8915E+01	− 7.5500E−02	− 1.0800E−02	3.8829E+01	3.8517E+01		
28	$3s^23p^4$ ($^3P_2$) $3d^{10}$ ($^1S_0$) $^3P_1^e$	4.2686E+01	− 1.5300E−01	− 3.9800E−03	4.2529E+01	4.2278E+01		
29	$3s^23p^4$ ($^1D_2$) $3d^{10}$ ($^1S_0$) $^1D_2^e$	4.3540E+01	− 1.5700E−01	− 6.7600E−03	4.3377E+01	4.3117E+01		
30	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_5^e$	4.7949E+01	− 2.3900E−02	1.8700E−02	4.7944E+01	4.7798E+01		
31	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^3F_4^e$	4.8133E+01	− 2.0600E−02	1.8700E−02	4.8131E+01	4.8008E+01		
32	$3s^23p^4$ ($^3P_2$) $3d^{10}$ ($^1S_0$) $^3P_0^e$	4.8230E+01	− 2.2100E−01	1.6900E−03	4.8011E+01	4.8207E+01		
33	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^3P_2^e$	4.8607E+01	3.4600E−03	1.8600E−02	4.8629E+01	4.8681E+01		
34	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^1P_1^e$	4.8658E+01	3.6800E−03	1.8600E−02	4.8680E+01	4.8736E+01		
35	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_3^e$	4.8748E+01	− 1.0300E−02	1.8600E−02	4.8756E+01	4.8811E+01		
36	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$)$^5F_2^e$	4.8842E+01	− 7.3500E−03	1.8600E−02	4.8853E+01	4.8913E+01		
37	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_4^e$	4.8897E+01	− 4.1400E−02	1.8700E−02	4.8874E+01	4.8951E+01		
38	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^3F_3^e$	4.9013E+01	− 3.4800E−02	1.8700E−02	4.8997E+01	4.9076E+01		
39	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_1^e$	4.9348E+01	− 2.6700E−02	1.8600E−02	4.9340E+01	4.9397E+01		
40	$3s^23p^63d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_5^e$	4.9369E+01	− 5.2800E−02	1.8700E−02	4.9334E+01	4.9407E+01		
41	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$) $^5P_3^e$	4.9398E+01	− 3.4100E−02	1.8600E−02	4.9383E+01	4.9450E+01		
42	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$)$^3F_2^e$	4.9399E+01	− 2.6000E−02	1.8600E−02	4.9391E+01	4.9464E+01		
43	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_4^e$	4.9438E+01	− 5.1300E−02	1.8700E−02	4.9406E+01	4.9475E+01		
44	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^3P_2^e$	4.9565E+01	− 3.1300E−02	1.8700E−02	4.9552E+01	4.9628E+01		
45	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^5P_1^e$	4.9608E+01	− 2.9100E−02	1.8600E−02	4.9598E+01	4.9661E+01		
46	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^3P_0^e$	4.9695E+01	− 2.8300E−02	1.8600E−02	4.9685E+01	4.9752E+01		
47	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_6^e$	4.9754E+01	− 7.3900E−02	1.8700E−02	4.9699E+01	4.9756E+01		
48	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^1H_5^e$	4.9830E+01	− 7.0700E−02	1.8700E−02	4.9778E+01	4.9841E+01		

Table 1 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF5				FAC	NIST [23]	
		DC	Breit	QED	Total			
49	$3s^2 3p^6 3d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_3^e$	4.9916E+01	− 5.2900E−02	1.8600E−02	4.9882E+01	4.9942E+01		
50	$3s^2 3p^6 3d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_4^e$	4.9943E+01	− 5.4000E−02	1.8600E−02	4.9907E+01	4.9969E+01		
51	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$)$^3D_3^e$	5.0337E+01	− 5.8200E−02	1.8600E−02	5.0297E+01	5.0366E+01		
52	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$)$^3D_2^e$	5.0356E+01	− 5.3800E−02	1.8600E−02	5.0321E+01	5.0387E+01		
53	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^3F_2^e$	5.0364E+01	− 2.8700E−02	1.8600E−02	5.0354E+01	5.0399E+01		
54	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^3F_3^e$	5.0394E+01	− 2.9800E−02	1.8600E−02	5.0382E+01	5.0425E+01		
55	$3s^2 3p^6 3d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^3P_1^e$	5.0420E+01	− 5.2600E−02	1.8600E−02	5.0386E+01	5.0457E+01		
56	$3s^2 3p^6 3d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^5P_2^e$	5.0442E+01	− 5.6900E−02	1.8600E−02	5.0404E+01	5.0480E+01		
57	$3s^2 3p^6 3d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^3P_1^e$	5.0690E+01	− 6.0600E−02	1.8600E−02	5.0648E+01	5.0716E+01		
58	$3s^2 3p^6 3d^7$ ($^2G_3$) $4s^1$ ($^2S_1$) $^1G_4^e$	5.0728E+01	− 9.0400E−02	1.8700E−02	5.0656E+01	5.0723E+01		
59	$3s^2 3p^6 3d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_5^e$	5.0744E+01	− 9.2900E−02	1.8700E−02	5.0670E+01	5.0738E+01		
60	$3s^2 3p^6 3d^7$ ($^4P_3$) $4s^1$ ($^2S_1$) $^3P_0^e$	5.0776E+01	− 6.2600E−02	1.8600E−02	5.0732E+01	5.0808E+01		
61	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$) $^3F_4^e$	5.1020E+01	− 7.6600E−02	1.8600E−02	5.0962E+01	5.1016E+01		
62	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^3D_1^e$	5.1056E+01	− 3.6300E−02	1.8600E−02	5.1038E+01	5.1082E+01		
63	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^1D_2^e$	5.1068E+01	− 3.8500E−02	1.8600E−02	5.1048E+01	5.1087E+01		
64	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^1F_3^e$	5.1084E+01	− 7.2600E−02	1.8600E−02	5.1030E+01	5.1092E+01		
65	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^5D_4^o$	5.1853E+01	4.6700E−03	− 4.4300E−04	5.1857E+01	5.1914E+01		
66	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3G_5^o$	5.1949E+01	− 3.3600E−03	− 4.7300E−04	5.1945E+01	5.2009E+01		
67	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_1^e$	5.2016E+01	− 6.8200E−02	1.8600E−02	5.1967E+01	5.2016E+01		
68	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_2^e$	5.2043E+01	− 6.8400E−02	1.8600E−02	5.1993E+01	5.2036E+01		
69	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_3^e$	5.2123E+01	− 5.2800E−02	1.8600E−02	5.2089E+01	5.2042E+01		
70	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$) $^1D_2^e$	5.2217E+01	− 5.1900E−02	1.8600E−02	5.2183E+01	5.2118E+01		
71	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$)$^3P_1^o$	5.2464E+01	2.8100E−02	− 4.8800E−04	5.2491E+01	5.2221E+01		
72	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$)$^3D_2^o$	5.2506E+01	2.5600E−02	− 1.5800E−03	5.2530E+01	5.2535E+01		
73	$3s^1 3p^5$ ($^2P_1$) $3d^{10}$ ($^1S_0$) $^3P_2^o$	5.2525E+01	− 8.1300E−02	− 7.4500E−02	5.2369E+01	5.2577E+01		
74	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5D_2^o$	5.2628E+01	1.9400E−02	− 6.0800E−04	5.2646E+01	5.2692E+01		
75	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_3^o$	5.2675E+01	1.3800E−02	− 4.0400E−04	5.2689E+01	5.2738E+01		
76	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_3^o$	5.2774E+01	− 1.4900E−02	− 4.0700E−04	5.2758E+01	5.2825E+01		
77	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_4^o$	5.2849E+01	− 2.1300E−02	− 4.5500E−04	5.2827E+01	5.2898E+01		
78	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_2^o$	5.3220E+01	− 7.9900E−03	− 3.8400E−04	5.3212E+01	5.3265E+01		
79	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_2^o$	5.3236E+01	− 6.5800E−03	− 4.0200E−04	5.3229E+01	5.3288E+01		
80	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_1^o$	5.3246E+01	− 3.2800E−03	− 4.3000E−04	5.3242E+01	5.3302E+01		
81	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	5.3270E+01	− 2.5800E−02	− 3.1400E−04	5.3244E+01	5.3302E+01		
82	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_5^o$	5.3280E+01	− 3.0400E−02	− 2.1700E−04	5.3249E+01	5.3306E+01		
83	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^5D_0^o$	5.3378E+01	− 4.8900E−03	− 3.3600E−04	5.3373E+01	5.3420E+01		
84	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.3409E+01	− 1.4500E−02	− 3.4500E−04	5.3394E+01	5.3463E+01		
85	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_6^o$	5.3438E+01	− 2.5600E−02	1.2400E−03	5.3413E+01	5.3474E+01		
86	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_5^o$	5.3456E+01	− 2.7100E−02	9.5600E−04	5.3430E+01	5.3493E+01		
87	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.3541E+01	− 8.1900E−03	− 4.4300E−04	5.3533E+01	5.3496E+01		
88	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	5.3651E+01	− 2.2700E−02	1.1200E−03	5.3629E+01	5.3592E+01		
89	$3s^2 3p^6 3d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_6^o$	5.3681E+01	− 5.3800E−02	− 4.6700E−04	5.3627E+01	5.3680E+01		
90	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.3690E+01	− 1.5500E−02	1.0000E−03	5.3675E+01	5.3711E+01		
91	$3s^2 3p^6 3d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3G_5^o$	5.3704E+01	− 4.6100E−02	− 3.7200E−04	5.3658E+01	5.3717E+01		
92	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_4^o$	5.3801E+01	− 3.3800E−02	− 4.5200E−04	5.3767E+01	5.3744E+01		
93	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1F_3^o$	5.3861E+01	− 2.6700E−02	− 3.1900E−04	5.3834E+01	5.3821E+01		
94	$3s^1 3p^5$ ($^2P_1$) $3d^{10}$ ($^1S_0$) $^3P_1^o$	5.3997E+01	− 8.3300E−02	− 7.5100E−02	5.3839E+01	5.3894E+01		
95	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^1S_0^o$	5.4023E+01	4.7300E−03	1.0500E−03	5.4029E+01	5.4075E+01		
96	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.4075E+01	− 7.1200E−03	1.0700E−03	5.4069E+01	5.4126E+01		

Table 1 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF5				FAC	NIST [23]	
		DC	Breit	QED	Total			
97	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3P_2^o$	5.4096E+01	− 1.0600E−02	8.2700E−04	5.4086E+01	5.4140E+01		
98	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$)$^5S_2^o$	5.4159E+01	− 1.8700E−02	1.1400E−04	5.4141E+01	5.4195E+01		
99	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$)$^3G_3^o$	5.4174E+01	− 1.1300E−02	5.9500E−05	5.4162E+01	5.4200E+01		
100	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5D_1^o$	5.4224E+01	− 1.7000E−02	5.4700E−04	5.4207E+01	5.4265E+01		
101	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.4245E+01	− 9.8800E−03	1.7800E−04	5.4235E+01	5.4280E+01		
102	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.4255E+01	− 3.1700E−02	− 1.1700E−04	5.4223E+01	5.4288E+01		
103	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.4265E+01	− 1.4100E−02	1.2400E−03	5.4253E+01	5.4310E+01		
104	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_1^o$	5.4283E+01	− 1.2300E−02	2.7500E−04	5.4271E+01	5.4342E+01		
105	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_4^o$	5.4326E+01	− 3.3700E−02	1.2500E−03	5.4293E+01	5.4359E+01		
106	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^5F_2^o$	5.4328E+01	− 1.2400E−02	2.7400E−04	5.4316E+01	5.4370E+01		
107	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^3G_3^o$	5.4367E+01	− 1.1000E−02	7.8700E−04	5.4357E+01	5.4400E+01		
108	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_5^o$	5.4369E+01	− 4.3100E−02	1.2600E−03	5.4327E+01	5.4417E+01		
109	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^1D_2^o$	5.4390E+01	− 2.2700E−02	4.0300E−06	5.4367E+01	5.4434E+01		
110	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3P_0^o$	5.4491E+01	− 3.9200E−02	− 4.7000E−04	5.4452E+01	5.4507E+01		
111	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^3D_1^o$	5.4506E+01	− 6.5100E−03	1.1300E−03	5.4501E+01	5.4571E+01		
112	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3F_3^o$	5.4554E+01	− 3.6800E−02	9.9300E−04	5.4518E+01	5.4589E+01		
113	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.4597E+01	− 3.4500E−02	9.8400E−04	5.4564E+01	5.4603E+01		
114	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_5^o$	5.4603E+01	− 7.1500E−02	− 3.8100E−04	5.4531E+01	5.4645E+01		
115	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$)$^1P_1^o$	5.4652E+01	− 3.5000E−02	− 1.6100E−04	5.4617E+01	5.4662E+01		
116	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.4666E+01	− 6.4400E−02	− 2.4200E−04	5.4602E+01	5.4681E+01		
117	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.4788E+01	− 2.5500E−02	7.3600E−04	5.4763E+01	5.4807E+01		
118	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_6^o$	5.4803E+01	− 5.5400E−02	1.2500E−03	5.4749E+01	5.4823E+01		
119	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.4833E+01	− 4.6600E−02	1.1100E−03	5.4787E+01	5.4847E+01		
120	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3F_3^o$	5.4853E+01	− 3.9200E−02	1.0200E−03	5.4815E+01	5.4877E+01		
121	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_4^o$	5.4863E+01	− 4.3600E−02	9.2200E−04	5.4820E+01	5.4880E+01		
122	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.4883E+01	− 2.1400E−02	1.6400E−04	5.4862E+01	5.4913E+01		
123	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3P_1^o$	5.4891E+01	− 2.3100E−02	5.7700E−04	5.4869E+01	5.4922E+01		
124	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$)$^3G_5^o$	5.4911E+01	− 5.3700E−02	1.1700E−03	5.4858E+01	5.4925E+01		
125	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^5D_3^o$	5.4930E+01	− 4.0800E−02	4.6500E−04	5.4890E+01	5.4955E+01		
126	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$)$^3D_2^o$	5.4933E+01	− 3.1000E−02	1.1500E−03	5.4903E+01	5.4970E+01		
127	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.4978E+01	− 5.5600E−02	4.2500E−05	5.4923E+01	5.4984E+01		
128	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5P_3^o$	5.5014E+01	− 3.4400E−02	1.1500E−03	5.4981E+01	5.5050E+01		
129	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.5022E+01	− 1.6900E−02	2.8800E−04	5.5005E+01	5.5062E+01		
130	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5D_0^o$	5.5045E+01	− 2.6500E−02	1.2400E−03	5.5019E+01	5.5094E+01		
131	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3S_1^o$	5.5058E+01	− 3.2100E−02	1.1900E−03	5.5027E+01	5.5108E+01		
132	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.5114E+01	− 4.3200E−02	6.6700E−04	5.5071E+01	5.5114E+01		
133	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_7^o$	5.5134E+01	− 7.7000E−02	1.2500E−03	5.5058E+01	5.5145E+01		
134	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3P_2^o$	5.5179E+01	− 2.9200E−02	1.2100E−03	5.5151E+01	5.5219E+01		
135	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.5231E+01	− 2.9500E−02	8.7600E−04	5.5202E+01	5.5272E+01		
136	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^1G_4^o$	5.5299E+01	− 6.2200E−02	1.1100E−03	5.5238E+01	5.5299E+01		
137	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1H_5^o$	5.5312E+01	− 6.2400E−02	1.2200E−03	5.5250E+01	5.5308E+01		
138	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3H_6^o$	5.5355E+01	− 7.2200E−02	1.2600E−03	5.5284E+01	5.5341E+01		
139	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	5.5411E+01	− 5.0000E−02	1.1900E−03	5.5362E+01	5.5422E+01		
140	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.5430E+01	− 4.9700E−02	1.0700E−03	5.5382E+01	5.5447E+01		
141	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1G_4^o$	5.5472E+01	− 5.4200E−02	1.1900E−03	5.5419E+01	5.5485E+01		
142	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^1H_5^o$	5.5483E+01	− 6.4200E−02	1.2500E−03	5.5420E+01	5.5490E+01		
143	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_3^o$	5.5677E+01	− 4.9900E−02	1.1900E−03	5.5628E+01	5.5686E+01		
144	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5P_1^o$	5.5734E+01	− 3.8500E−02	1.1000E−03	5.5697E+01	5.5742E+01		

Table 1 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF5				FAC	NIST [23]	
		DC	Breit	QED	Total			
145	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^3F_2^o$	5.5735E+01	− 4.3400E−02	7.9800E−04	5.5692E+01	5.5748E+01		
146	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3F_3^o$	5.5756E+01	− 4.7600E−02	1.0300E−03	5.5709E+01	5.5750E+01		
147	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3F_4^o$	5.5814E+01	− 5.8900E−02	1.2400E−03	5.5756E+01	5.5821E+01		
148	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^5P_2^o$	5.5861E+01	− 5.1100E−02	1.0200E−03	5.5811E+01	5.5872E+01		
149	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)^1G_4^o$	5.5900E+01	− 3.5200E−02	1.2100E−03	5.5866E+01	5.5910E+01		
150	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3P_1^o$	5.5926E+01	− 4.9100E−02	8.7800E−04	5.5877E+01	5.5942E+01		
151	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3P_2^o$	5.5956E+01	− 4.7100E−02	7.4700E−04	5.5910E+01	5.5972E+01		
152	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3P_0^o$	5.5967E+01	− 5.0300E−02	1.2000E−03	5.5918E+01	5.5989E+01		
153	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3D_1^o$	5.5971E+01	− 4.2500E−02	1.0700E−03	5.5929E+01	5.5994E+01		
154	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^1F_3^o$	5.5977E+01	− 5.0700E−02	1.1400E−03	5.5927E+01	5.5994E+01		
155	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3P_2^o$	5.6007E+01	− 4.0600E−02	9.9500E−05	5.5966E+01	5.6005E+01		
156	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3F_2^o$	5.6080E+01	− 3.7800E−02	5.2200E−04	5.6043E+01	5.6086E+01		
157	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)^1F_3^o$	5.6111E+01	− 4.8500E−02	5.5900E−04	5.6063E+01	5.6099E+01		
158	$3s^2 3p^6 3d^7\ (^4H_3)\ 4p^1\ (^2P_1)\ ^1I_6^o$	5.6120E+01	− 9.5800E−02	1.2500E−03	5.6025E+01	5.6121E+01		
159	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^1P_1^o$	5.6192E+01	− 5.3300E−02	3.0700E−04	5.6139E+01	5.6198E+01		
160	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3H_4^o$	5.6219E+01	− 8.3700E−02	1.2500E−03	5.6137E+01	5.6221E+01		
161	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)^1D_2^o$	5.6235E+01	− 5.8100E−02	1.0000E−03	5.6178E+01	5.6247E+01		
162	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)\ ^3G_4^o$	5.6245E+01	− 5.8300E−02	1.1800E−03	5.6188E+01	5.6258E+01		
163	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3S_1^o$	5.6270E+01	− 6.5900E−02	7.1800E−04	5.6205E+01	5.6264E+01		
164	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3H_5^o$	5.6296E+01	− 8.7500E−02	1.2500E−03	5.6210E+01	5.6274E+01		
165	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3P_0^o$	5.6315E+01	− 2.4800E−02	1.2700E−03	5.6291E+01	5.6321E+01		
166	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^1D_2^o$	5.6497E+01	− 6.4700E−02	1.1100E−03	5.6434E+01	5.6485E+01		
167	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3G_5^o$	5.6527E+01	− 8.0300E−02	1.2200E−03	5.6448E+01	5.6504E+01		
168	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3F_4^o$	5.6536E+01	− 7.3200E−02	1.2400E−03	5.6464E+01	5.6514E+01		
169	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3F_3^o$	5.6558E+01	− 4.4700E−02	1.2100E−03	5.6515E+01	5.6564E+01		
170	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3D_1^o$	5.6648E+01	− 3.1400E−02	1.1000E−03	5.6618E+01	5.6663E+01		
171	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3D_2^o$	5.6662E+01	− 4.8400E−02	1.1800E−03	5.6615E+01	5.6668E+01		
172	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^1F_3^o$	5.6709E+01	− 6.2100E−02	1.1100E−03	5.6648E+01	5.6705E+01		
173	$3s^1 3p^5\ (^2P_1)\ 3d^{10}\ (^1S_0)\ ^3P_0^o$	5.7414E+01	− 1.4700E−01	− 6.9000E−02	5.7198E+01	5.6916E+01		
174	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3P_1^o$	5.7418E+01	− 5.8000E−02	1.2700E−03	5.7361E+01	5.7397E+01		
175	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3D_2^o$	5.7462E+01	− 5.4000E−02	1.2500E−03	5.7409E+01	5.7448E+01		
176	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3F_4^o$	5.7550E+01	− 5.4600E−02	1.2500E−03	5.7497E+01	5.7524E+01		
177	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3F_3^o$	5.7561E+01	− 7.0300E−02	1.2400E−03	5.7492E+01	5.7543E+01		
178	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3P_0^o$	5.7584E+01	− 7.1700E−02	1.0100E−03	5.7513E+01	5.7567E+01		
179	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^1P_1^o$	5.7717E+01	− 5.9500E−02	1.2400E−03	5.7659E+01	5.7711E+01		
180	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^1D_2^o$	5.7726E+01	− 4.9800E−02	1.2100E−03	5.7677E+01	5.7715E+01		
181	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3D_3^o$	5.7869E+01	− 5.1100E−02	1.2400E−03	5.7819E+01	5.7857E+01		
182	$3s^1 3p^5\ (^2P_1)\ 3d^{10}\ (^1S_0)\ ^1P_1^o$	5.8754E+01	− 1.4600E−01	− 6.9600E−02	5.8539E+01	5.8228E+01		
183	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5F_4^e$	5.9530E+01	− 2.8200E−02	− 3.2100E−05	5.9501E+01	5.9537E+01		
184	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5F_5^e$	5.9579E+01	− 3.7600E−02	− 2.0700E−05	5.9541E+01	5.9579E+01		
185	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5P_3^e$	5.9590E+01	− 2.4200E−02	− 2.4000E−05	5.9565E+01	5.9604E+01		
186	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^3H_6^e$	5.9699E+01	− 4.0000E−02	9.4200E−07	5.9659E+01	5.9704E+01		
187	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^3D_3^e$	5.9908E+01	− 3.6200E−02	− 2.9200E−05	5.9872E+01	5.9903E+01		
188	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^3G_5^e$	5.9929E+01	− 4.5100E−02	− 8.4300E−06	5.9884E+01	5.9929E+01		
189	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5H_7^e$	5.9965E+01	− 5.0900E−02	4.9100E−06	5.9914E+01	5.9955E+01		
190	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5G_6^e$	5.9978E+01	− 4.6800E−02	3.9000E−07	5.9931E+01	5.9977E+01		
191	$3s^2 3p^6 3d^7\ (^2P_3)\ 4d^1\ (^2D_1)^3D_1^e$	6.0039E+01	− 5.0200E−03	− 7.9900E−05	6.0034E+01	6.0051E+01		
192	$3s^2 3p^6 3d^7\ (^2P_3)\ 4d^1\ (^2D_1)\ ^3D_2^e$	6.0093E+01	− 2.4200E−02	− 5.4500E−05	6.0069E+01	6.0097E+01		

Table 1 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF5				FAC	NIST [23]	
		DC	Breit	QED	Total			
193	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^3P_2^e$	6.0190E+01	− 2.0500E−02	− 9.3600E−06	6.0170E+01	6.0206E+01		
194	$3s^2 3p^6 3d^7$ ($^2P_3$) $4d^1$ ($^2D_1$) $^3F_3^e$	6.0193E+01	− 1.5500E−02	− 5.2900E−05	6.0178E+01	6.0209E+01		
195	$3s^2 3p^6 3d^7$ ($^2P_3$) $4d^1$ ($^2D_1$) $^3P_0^e$	6.0215E+01	− 6.0500E−03	− 8.8800E−05	6.0209E+01	6.0238E+01		
196	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5D_4^e$	6.0264E+01	− 3.8600E−02	− 1.2700E−05	6.0225E+01	6.0270E+01		
197	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5G_3^e$	6.0373E+01	− 2.5300E−02	− 2.3500E−05	6.0348E+01	6.0379E+01		
198	$3s^2 3p^6 3d^7$ ($^4P_3$) $4d^1$ ($^2D_1$) $^5P_1^e$	6.0434E+01	− 1.2200E−02	− 4.3700E−05	6.0422E+01	6.0438E+01		
199	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5G_4^e$	6.0435E+01	− 3.5800E−02	− 9.1300E−06	6.0399E+01	6.0445E+01		
200	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5F_2^e$	6.0456E+01	− 1.6300E−02	− 2.5400E−05	6.0439E+01	6.0462E+01		

Table 2 Comparison of excitation energies (in Ryd.) of five different configurations sets for first 24 excited states levels for Fe-like Xe

Level no	MCDF1	MCDF2	MCDF3	MCDF4	MCDF5	ΔE_1	ΔE_2	ΔE_3	ΔE_4
2	0.458	0.458	0.433	0.458	0.433	0.000	− 0.025	0.000	− 0.025
3	0.922	0.922	0.922	0.922	0.922	0.000	0.000	0.000	0.000
4	1.261	1.256	1.198	1.261	1.194	− 0.005	− 0.062	0.000	− 0.066
5	1.352	1.352	1.336	1.352	1.336	0.000	− 0.016	0.000	− 0.016
6	1.624	1.624	1.596	1.624	1.596	0.000	− 0.028	0.000	− 0.028
7	1.704	1.704	1.704	1.704	1.704	0.000	0.000	0.000	0.000
8	2.209	2.209	2.192	2.209	2.192	0.000	− 0.018	0.000	− 0.018
9	3.578	3.563	3.457	3.578	3.445	− 0.015	− 0.122	0.000	− 0.133
10	18.072	18.072	18.072	18.072	18.072	0.000	0.000	0.000	0.000
11	18.473	18.473	18.473	18.468	18.467	0.000	0.000	− 0.006	− 0.006
12	19.095	19.095	19.095	19.095	19.095	0.000	0.000	0.000	0.000
13	19.741	19.741	19.741	19.686	19.686	0.000	0.000	− 0.055	− 0.055
14	19.797	19.797	19.797	19.791	19.791	0.000	0.000	− 0.006	− 0.006
15	20.082	20.082	20.082	20.059	20.059	0.000	0.000	− 0.024	− 0.023
16	20.248	20.248	20.248	20.248	20.247	0.000	0.000	0.000	0.000
17	20.366	20.366	20.366	20.312	20.312	0.000	0.000	− 0.054	− 0.054
18	23.785	23.785	23.785	23.771	23.771	0.000	0.000	− 0.014	− 0.014
19	23.966	23.966	23.966	23.965	23.965	0.000	0.000	− 0.001	− 0.001
20	24.420	24.420	24.420	24.420	24.420	0.000	0.000	0.000	0.000
21	25.600	25.600	25.600	25.508	25.508	0.000	0.000	− 0.092	− 0.092
22	34.026	34.026	34.026	34.026	34.026	0.000	0.000	0.000	0.000
23	34.415	34.415	34.286	34.415	34.286	0.000	− 0.129	0.000	− 0.129
24	35.041	35.041	35.041	35.041	35.041	0.000	0.000	0.000	0.000
25	35.733	35.733	35.239	35.733	35.239	0.000	− 0.494	0.000	− 0.494

We have also compared our computed fine structure energies from GRASP with the experimentally available critically evaluated energies published by NIST [23]. The difference between the two ΔE is listed in Table 1.

In Table 2, we list the excitation energies of first 24 excited states for 5 different sets of configurations. In set MCDF1, we have considered only configurations obtained by single electron excitation from 3d orbitals to outer orbitals up to $n = 10$, 3s and 3p orbitals to $n = 4$ orbitals and within $n = 3$.

In set MCDF2, we have taken set MCDF1 plus $3p^6 3d^{10}$ configuration obtained by double electron excitation from 3s to 3d orbital.

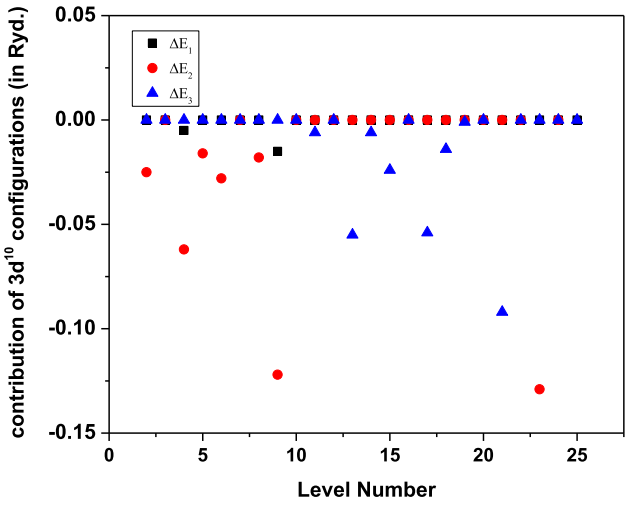


Fig. 1 Contribution of $3d^{10}$configurations in lowest 24 excited states of Fe-like Xe

In set MCDF3, we have taken set MCDF1 plus $3s^23p^43d^{10}$ configuration obtained by double electron excitation from 3p to 3d orbital.

In set MCDF4, we have taken set MCDF1 plus $3s3p^53d^{10}$ configuration obtained by double electron excitation from 3s and 3p to 3d orbital.

In final set MCDF5, we have taken set MCDF1 plus $3p^63d^{10}$, $3s^23p^43d^{10}$ and $3s3p^53d^{10}$ configuration obtained by double electron excitation to 3d orbital. The excitation energies from final set MCDF5 are given in Table 1.

In Table 2, difference of excitation energies of MCDF1 set with MCDF2, MCDF3, MCDF4 and MCDF5 is given under the columns ΔE_1, ΔE_2, ΔE_3 and ΔE_4, resp. From ΔE_1, we found that the effect of $3p^63d^{10}$ configuration is small and only in level number 4 and 9, while the effect of other two $3d^{10}$ configurations is considerable. We have also found that there is no change in excitation energies of some levels such as level number 7, 10, 12, 16, 20, 22 and 24 due to $3d^{10}$ configurations or we can say that the there is no or negligible contribution of $3d^{10}$ configurations to these levels. ΔE_4 shows the combined effect or contribution of $3d^{10}$ configurations on excitation energies.

In Fig. 1, we show the separate contribution of all $3d^{10}$ configurations in lowest 24 excited states for Fe-like Xe. From Fig. 1, we observe that the contribution of $3s3p^53d^{10}$ and $3s^23p^43d^{10}$ configurations is very small for most of the levels. From Table 1 and Fig. 1, we have also seen that $3d^{10}$ configurations are lowering the position of levels in energy spectrum. We have not found any change in excitation energies of levels greater than 25 due to inclusion of doubly excited $3d^{10}$ configurations and therefore, in Table 2 we only show up to lowest 25 levels.

5.2. Excitation energies of Fe-like Cs

For Fe-like Cs, fine structure energies of lowest 200 levels related to non-relativistic configurations $3s^23p^63d^8$, $3s^23p^53d^9$, $3s3p^63d^9$, $3s^23p^63d^74l$ (l = 0, 1, 2), $3s^23p^43d^{10}$ and $3s3p^53d^{10}$ from both GRASP and FAC are reported in Table 3. Separate contributions of Dirac–Coulomb and relativistic corrections Breit and QED in total energy have also presented.

We have also matched our computed fine structure energies with energies compiled by NIST [24]. The difference ΔE of our obtained fine structure energies from GRASP with energies compiled by NIST is listed in Table 3.

In Table 4, we list the excitation energies of first 24 excited states for 5 different sets of configurations. In Table 4, configuration sets MCDF1, MCDF2, MCDSF3, MCDf4 and MCDF5 are same as for Fe-like Xe in Table 2.

In Table 4, difference of excitation energies of MCDF1 set with MCDF2, MCDF3, MCDF4, MCDF5 is given under the columns ΔE_1, ΔE_2, ΔE_3 and ΔE_4, resp. From ΔE_1, we found that the effect of $3p^63d^{10}$ configuration is small and only in level number 4 and 9, while the effect of other two $3d^{10}$ configurations is considerable. We have also found that there is no change in excitation energies of some levels such as level number 7, 10, 12, 16, 20, 22 and 24 due to $3d^{10}$ configurations or we can say that the there is no or negligible contribution of $3d^{10}$ configurations to these levels.

In Fig. 2, we show the separate contribution of all $3d^{10}$ configurations in lowest 24 excited states for Fe-like Cs. From Fig. 2, we observe that the contribution of $3s3p^53d^{10}$ and $3s^23p^43d^{10}$ configurations is very small for most of the levels similar to Fe-like Xe. From Table 4 and Fig. 2, we have also seen that $3d^{10}$ configurations are lowering the position of levels in energy spectrum. We have not found any change in excitation energies of levels greater than 25 due to inclusion of doubly excited $3d^{10}$ configurations and therefore, in Table 4 we have only shown up to lowest 25 levels for Fe-like Cs.

5.3. Excitation energies of Fe-like Ba

For Fe-like Ba, fine-structure excitation energies of lowest 200 levels related to non-relativistic configurations $3s^23p^63d^8$, $3s^23p^53d^9$, $3s3p^63d^9$, $3s^23p^63d^74l$ (l = 0, 1, 2), $3s^23p^43d^{10}$ and $3s3p^53d^{10}$ from both GRASP and FAC are listed in Table 5 with separate contributions of Dirac–Coulomb and relativistic corrections Breit and QED in total energy.

Table 3 Excitation energies (in Ryd.) for lowest 200 levels for Fe-like Cs

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF				FAC	NIST[24]	
		DC	Breit	QED	Total			
1	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3F_2$) $^3F_4^e$	0.0000	–	–	0.0000	0.0000	0.0000	–
2	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^1D_2$) $^1D_2^e$	4.2798E−01	1.7400E−02	− 3.0100E−05	4.4538E−01	4.4543E−01	4.350E−01	0.0104
3	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3F_2$) $^3F_3^e$	1.0440E+00	− 2.4600E−02	2.5200E−06	1.0195E+00	1.0207E+00	1.029E+00	0.0095
4	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3P_2$) $^3P_0^e$	1.2154E+00	2.6200E−02	− 2.7400E−05	1.2415E+00	1.2419E+00	1.216E+00	0.0255
5	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3P_2$) $^3P_2^e$	1.4705E+00	− 2.2900E−02	− 2.8400E−06	1.4476E+00	1.4492E+00	1.443E+00	0.0046
6	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3P_2$) $^3P_1^e$	1.7173E+00	− 4.7100E−03	7.6900E−09	1.7126E+00	1.7145E+00	1.696E+00	0.0166
7	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^1G_2$) $^1G_4^e$	1.8566E+00	− 3.8700E−02	2.8900E−06	1.8179E+00	1.8201E+00	1.795E+00	0.0229
8	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^3F_2$) $^3F_2^e$	2.4478E+00	− 5.4300E−02	− 1.8900E−05	2.3935E+00	2.3960E+00	2.402E+00	0.0085
9	$3s^23p^6$ ($^1S_0$) $3d^8$ ($^1S_0$) $^1S_0^e$	3.7002E+00	− 2.7000E−02	− 3.1000E−05	3.6731E+00	3.6753E+00	3.660E+00	0.0131
10	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3F_4^o$	1.8697E+01	− 4.7700E−02	− 5.8600E−03	1.8644E+01	1.8581E+01	1.8408E+01	0.2360
11	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^1D_2^o$	1.9103E+01	− 4.1700E−02	− 5.8400E−03	1.9055E+01	1.8991E+01	1.8778E+01	0.2770
12	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3D_3^o$	1.9774E+01	− 5.5000E−02	− 5.8400E−03	1.9713E+01	1.9648E+01	1.9396E+01	0.3170
13	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3P_1^o$	2.0364E+01	− 3.9700E−02	− 5.9400E−03	2.0318E+01	2.0244E+01	1.9997E+01	0.3210
14	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3D_2^o$	2.0554E+01	− 7.1900E−02	− 5.8400E−03	2.0476E+01	2.0418E+01	2.0168E+01	0.3080
15	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3P_0^o$	2.0822E+01	− 5.9400E−02	− 5.8900E−03	2.0757E+01	2.0696E+01	2.0451E+01	0.3060
16	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3F_3^o$	2.1028E+01	− 7.5000E−02	− 5.3100E−03	2.0947E+01	2.0882E+01	2.0630E+01	0.3170
17	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3D_1^o$	2.1067E+01	− 6.2700E−02	− 5.8800E−03	2.0999E+01	2.0932E+01	2.0666E+01	0.3330
18	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3P_2^o$	2.4933E+01	− 1.1200E−01	1.2100E−03	2.4823E+01	2.4782E+01	2.4509E+01	0.3140
19	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^3F_2^o$	2.5252E+01	− 1.5500E−01	1.2600E−03	2.5098E+01	2.5072E+01	2.4879E+01	0.2190
20	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^1F_3^o$	2.5586E+01	− 1.2300E−01	7.2200E−04	2.5464E+01	2.5413E+01	2.5100E+01	0.3640
21	$3s^23p^5$ ($^2P_1$) $3d^9$ ($^2D_1$) $^1P_1^o$	2.6818E+01	− 1.4300E−01	9.7600E−04	2.6677E+01	2.6638E+01	2.6366E+01	0.3110
22	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^3D_3^e$	3.5503E+01	− 3.8700E−02	− 7.6200E−02	3.5389E+01	3.5297E+01		
23	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^3D_2^e$	3.5765E+01	− 3.8600E−02	− 7.3500E−02	3.5653E+01	3.5552E+01		
24	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^3D_1^e$	3.6647E+01	− 6.5200E−02	− 7.6200E−02	3.6506E+01	3.6425E+01		
25	$3s^13p^6$ ($^1S_0$) $3d^9$ ($^2D_1$) $^1D_2^e$	3.6820E+01	− 7.3300E−02	− 6.9200E−02	3.6678E+01	3.6572E+01		
26	$3s^23p^4$ ($^3P_2$) $3d^{10}$ ($^1S_0$) $^3P_2^e$	3.9548E+01	− 8.9900E−02	− 1.8300E−02	3.9439E+01	3.9152E+01		
27	$3s^23p^4$ ($^1S_0$) $3d^{10}$ ($^1S_0$) $^1S_0^e$	4.0134E+01	− 8.0000E−02	− 1.2000E−02	4.0042E+01	3.9724E+01		
28	$3s^23p^4$ ($^3P_2$) $3d^{10}$ ($^1S_0$) $^3P_1^e$	4.4336E+01	− 1.6400E−01	− 4.5400E−03	4.4168E+01	4.3913E+01		
29	$3s^23p^4$ ($^1D_2$) $3d^{10}$ ($^1S_0$) $^1D_2^e$	4.5205E+01	− 1.6800E−01	− 7.4800E−03	4.5030E+01	4.4767E+01		
30	$3s^23p^4$ ($^3P_2$) $3d^{10}$ ($^1S_0$) $^3P_0^e$	5.0358E+01	− 2.3800E−01	1.7100E−03	5.0122E+01	4.9907E+01		
31	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_5^e$	5.0959E+01	− 2.5900E−02	2.0400E−02	5.0953E+01	5.1015E+01		
32	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^3F_4^e$	5.1146E+01	− 2.2300E−02	2.0400E−02	5.1144E+01	5.1218E+01		
33	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^3P_2^e$	5.1632E+01	3.4100E−03	2.0400E−02	5.1656E+01	5.1706E+01		
34	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^1P_1^e$	5.1686E+01	3.8500E−03	2.0400E−02	5.1711E+01	5.1765E+01		
35	$3s^23p^63d^7$ ($^2D_1$) $4s^1$ ($^2S_1$) $^3D_3^e$	5.1798E+01	− 8.7900E−03	2.0400E−02	5.1810E+01	5.1861E+01		
36	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$)$^5F_2^e$	5.1894E+01	− 6.5400E−03	2.0400E−02	5.1908E+01	5.1966E+01		
37	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_4^e$	5.2004E+01	− 4.5200E−02	2.0400E−02	5.1980E+01	5.2055E+01		
38	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^3F_3^e$	5.2119E+01	− 3.9700E−02	2.0400E−02	5.2100E+01	5.2178E+01		
39	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$) $^5F_1^e$	5.2473E+01	− 2.9100E−02	2.0400E−02	5.2464E+01	5.2506E+01		
40	$3s^23p^63d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_5^e$	5.2481E+01	− 5.8000E−02	2.0400E−02	5.2444E+01	5.2530E+01		
41	$3s^23p^63d^7$ ($^4F_3$) $4s^1$ ($^2S_1$)$^3F_2^e$	5.2521E+01	− 2.8400E−02	2.0400E−02	5.2513E+01	5.2573E+01		
42	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^5P_3^e$	5.2525E+01	− 3.7300E−02	2.0400E−02	5.2508E+01	5.2584E+01		
43	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_4^e$	5.2555E+01	− 5.6200E−02	2.0400E−02	5.2519E+01	5.2587E+01		
44	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^3P_2^e$	5.2694E+01	− 3.4800E−02	2.0400E−02	5.2679E+01	5.2754E+01		
45	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^5P_1^e$	5.2734E+01	− 3.2000E−02	2.0400E−02	5.2722E+01	5.2785E+01		
46	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^3P_0^e$	5.2822E+01	− 3.1200E−02	2.0400E−02	5.2811E+01	5.2877E+01		
47	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_6^e$	5.2882E+01	− 8.0400E−02	2.0400E−02	5.2822E+01	5.2878E+01		
48	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^1H_5^e$	5.2960E+01	− 7.6900E−02	2.0400E−02	5.2904E+01	5.2966E+01		

Table 3 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF				FAC	NIST[24]	
		DC	Breit	QED	Total			
49	$3s^23p^63d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_3^e$	5.3064E+01	− 5.7000E−02	2.0400E−02	5.3027E+01	5.3087E+01		
50	$3s^23p^63d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_4^e$	5.3088E+01	− 5.8400E−02	2.0400E−02	5.3050E+01	5.3110E+01		
51	$3s^23p^6$ ($^1S_0$) $3d^7$ $4s^1$($^2S_1$) $^3F_2^e$	5.3503E+01	− 3.3100E−02	2.0400E−02	5.3491E+01	5.3532E+01		
52	$3s^23p^63d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^3F_3^e$	5.3529E+01	− 5.2100E−02	2.0400E−02	5.3498E+01	5.3554E+01		
53	$3s^23p^63d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^5F_3^e$	5.3550E+01	− 4.5200E−02	2.0400E−02	5.3525E+01	5.3579E+01		
54	$3s^23p^63d^7$ ($^2D_3$) $4s^1$ ($^2S_1$)$^3D_2^e$	5.3567E+01	− 5.8100E−02	2.0400E−02	5.3530E+01	5.3598E+01		
55	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^3P_1^e$	5.3635E+01	− 5.7500E−02	2.0400E−02	5.3598E+01	5.3668E+01		
56	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^5P_2^e$	5.3659E+01	− 6.2500E−02	2.0400E−02	5.3617E+01	5.3693E+01		
57	$3s^23p^63d^7$ ($^2P_3$) $4s^1$ ($^2S_1$)$^3P_1^e$	5.3920E+01	− 6.5900E−02	2.0400E−02	5.3874E+01	5.3941E+01		
58	$3s^23p^63d^7$ ($^2G_3$) $4s^1$ ($^2S_1$) $^1G_4^e$	5.3958E+01	− 9.8200E−02	2.0400E−02	5.3880E+01	5.3947E+01		
59	$3s^23p^63d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_5^e$	5.3974E+01	− 1.0100E−01	2.0400E−02	5.3893E+01	5.3961E+01		
60	$3s^23p^63d^7$ ($^4P_3$) $4s^1$ ($^2S_1$) $^3P_0^e$	5.4011E+01	− 6.8200E−02	2.0400E−02	5.3963E+01	5.3967E+01		
61	$3s^23p^63d^7$ ($^2F_3$) $4s^1$ ($^2S_1$) $^3F_4^e$	5.4246E+01	− 8.4100E−02	2.0400E−02	5.4182E+01	5.4038E+01		
62	$3s^23p^63d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^3D_1^e$	5.4249E+01	− 3.6700E−02	2.0400E−02	5.4233E+01	5.4235E+01		
63	$3s^23p^63d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^1D_2^e$	5.4262E+01	− 3.8400E−02	2.0400E−02	5.4244E+01	5.4273E+01		
64	$3s^23p^63d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^1F_3^e$	5.4307E+01	− 7.9900E−02	2.0400E−02	5.4247E+01	5.4284E+01		
65	$3s^13p^5$ ($^2P_1$) $3d^{10}$($^1S_0$)$^3P_2^o$	5.4474E+01	− 8.6600E−02	− 8.2100E−02	5.4306E+01	5.4305E+01		
66	$3s^23p^6$ $3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5D_4^o$	5.5001E+01	5.0900E−03	− 4.7600E−04	5.5006E+01	5.5061E+01		
67	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^3G_5^o$	5.5101E+01	− 3.5200E−03	− 5.0900E−04	5.5097E+01	5.5158E+01		
68	$3s^23p^63d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_1^e$	5.5327E+01	− 7.8300E−02	2.0400E−02	5.5269E+01	5.5322E+01		
69	$3s^23p^63d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_2^e$	5.5350E+01	− 7.7600E−02	2.0400E−02	5.5292E+01	5.5343E+01		
70	$3s^23p^63d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^3D_3^e$	5.5383E+01	− 5.8100E−02	2.0400E−02	5.5346E+01	5.5374E+01		
71	$3s^23p^63d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^1D_2^e$	5.5482E+01	− 5.8100E−02	2.0400E−02	5.5445E+01	5.5483E+01		
72	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$)$^3P_1^o$	5.5629E+01	3.0200E−02	− 5.3900E−04	5.5658E+01	5.5489E+01		
73	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.5672E+01	2.7100E−02	− 5.4100E−04	5.5699E+01	5.5699E+01		
74	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^5D_2^o$	5.5813E+01	2.2200E−02	− 4.5400E−04	5.5835E+01	5.5742E+01		
75	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_3^o$	5.5865E+01	1.6400E−02	− 4.4700E−04	5.5880E+01	5.5877E+01		
76	$3s^13p^5$ ($^2P_1$) $3d^{10}$($^1S_0$)$^3P_1^o$	5.6009E+01	− 8.8100E−02	− 8.1400E−02	5.5840E+01	5.5926E+01		
77	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_3^o$	5.6023E+01	− 1.6500E−02	− 4.3400E−04	5.6006E+01	5.6071E+01		
78	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_4^o$	5.6099E+01	− 2.3200E−02	− 4.9000E−04	5.6076E+01	5.6145E+01		
79	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_2^o$	5.6487E+01	− 8.8500E−03	− 4.2000E−04	5.6478E+01	5.6530E+01		
80	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_2^o$	5.6503E+01	− 7.0900E−03	− 4.3700E−04	5.6496E+01	5.6553E+01		
81	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_1^o$	5.6512E+01	− 3.5600E−03	− 4.6500E−04	5.6508E+01	5.6556E+01		
82	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	5.6530E+01	− 2.8600E−02	− 3.6400E−04	5.6501E+01	5.6559E+01		
83	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_5^o$	5.6539E+01	− 3.4400E−02	− 3.2400E−04	5.6504E+01	5.6565E+01		
84	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_0^o$	5.6645E+01	− 5.6700E−03	− 3.7800E−04	5.6639E+01	5.6685E+01		
85	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_3^o$	5.6685E+01	− 1.5800E−02	− 3.9400E−04	5.6668E+01	5.6735E+01		
86	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_6^o$	5.6748E+01	− 2.8000E−02	1.3900E−03	5.6722E+01	5.6779E+01		
87	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_5^o$	5.6763E+01	− 2.8900E−02	1.1700E−03	5.6735E+01	5.6795E+01		
88	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.6813E+01	− 8.9300E−03	− 4.8100E−04	5.6804E+01	5.6861E+01		
89	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_6^o$	5.6951E+01	− 5.8400E−02	− 4.9400E−04	5.6893E+01	5.6944E+01		
90	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	5.6967E+01	− 2.4500E−02	1.2900E−03	5.6943E+01	5.6986E+01		
91	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3G_5^o$	5.6978E+01	− 4.9800E−02	− 3.8400E−04	5.6928E+01	5.7021E+01		
92	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.6998E+01	− 1.7600E−02	1.1300E−03	5.6981E+01	5.7047E+01		
93	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_4^o$	5.7088E+01	− 3.6200E−02	− 4.8000E−04	5.7051E+01	5.7104E+01		
94	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1F_3^o$	5.7151E+01	− 2.8400E−02	− 3.0500E−04	5.7122E+01	5.7181E+01		
95	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^1S_0^o$	5.7344E+01	4.6900E−03	1.2100E−03	5.7350E+01	5.7393E+01		
96	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.7409E+01	− 5.8100E−03	1.2400E−03	5.7405E+01	5.7460E+01		

Table 3 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF				FAC	NIST[24]	
		DC	Breit	QED	Total			
97	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3P_2^o$	5.7429E+01	− 6.1600E−03	1.1000E−03	5.7424E+01	5.7476E+01		
98	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	5.7471E+01	− 1.2900E−02	− 1.1200E−04	5.7458E+01	5.7493E+01		
99	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5S_2^o$	5.7504E+01	− 2.1800E−02	6.5100E−06	5.7482E+01	5.7531E+01		
100	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.7553E+01	− 1.5000E−02	− 2.6100E−04	5.7538E+01	5.7581E+01		
101	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^1P_1^o$	5.7569E+01	− 1.2200E−02	9.2300E−04	5.7558E+01	5.7614E+01		
102	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^1F_3^o$	5.7606E+01	− 3.6300E−02	− 1.8400E−04	5.7569E+01	5.7632E+01		
103	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$)$^1G_4^o$	5.7616E+01	− 8.8300E−03	1.4000E−03	5.7609E+01	5.7660E+01		
104	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_1^o$	5.7633E+01	− 1.9600E−02	3.8100E−05	5.7613E+01	5.7682E+01		
105	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_2^o$	5.7665E+01	− 9.5200E−03	8.6300E−04	5.7656E+01	5.7710E+01		
106	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^5D_3^o$	5.7711E+01	− 1.0400E−02	1.0500E−03	5.7702E+01	5.7756E+01		
107	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_4^o$	5.7730E+01	− 4.2500E−02	1.4100E−03	5.7689E+01	5.7760E+01		
108	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^1D_2^o$	5.7746E+01	− 2.9100E−02	− 1.1900E−04	5.7717E+01	5.7782E+01		
109	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_5^o$	5.7779E+01	− 4.7100E−02	1.4200E−03	5.7734E+01	5.7804E+01		
110	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.7863E+01	− 4.2700E−02	− 5.0900E−04	5.7820E+01	5.7874E+01		
111	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3P_0^o$	5.7864E+01	− 5.5300E−03	1.2900E−03	5.7860E+01	5.7926E+01		
112	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3F_3^o$	5.7957E+01	− 4.0900E−02	1.1800E−03	5.7917E+01	5.7953E+01		
113	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_5^o$	5.7975E+01	− 7.8000E−02	− 4.2400E−04	5.7896E+01	5.7999E+01		
114	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.8000E+01	− 3.7100E−02	1.1000E−03	5.7964E+01	5.8032E+01		
115	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5P_1^o$	5.8030E+01	− 3.9300E−02	− 2.2400E−04	5.7991E+01	5.8041E+01		
116	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.8043E+01	− 7.0400E−02	− 2.9900E−04	5.7972E+01	5.8054E+01		
117	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.8167E+01	− 2.2500E−02	2.8200E−04	5.8145E+01	5.8189E+01		
118	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_6^o$	5.8218E+01	− 6.0800E−02	1.4100E−03	5.8158E+01	5.8215E+01		
119	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.8244E+01	− 5.5000E−02	1.0900E−03	5.8190E+01	5.8244E+01		
120	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	5.8265E+01	− 4.8100E−02	9.0900E−04	5.8218E+01	5.8278E+01		
121	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.8266E+01	− 2.6100E−02	7.6600E−04	5.8241E+01	5.8288E+01		
122	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_4^o$	5.8273E+01	− 4.6400E−02	8.3700E−04	5.8228E+01	5.8302E+01		
123	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3P_1^o$	5.8276E+01	− 2.0700E−02	3.0000E−04	5.8256E+01	5.8304E+01		
124	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3G_5^o$	5.8327E+01	− 5.8700E−02	1.3300E−03	5.8269E+01	5.8332E+01		
125	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.8332E+01	− 4.1500E−02	5.7700E−04	5.8291E+01	5.8355E+01		
126	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.8361E+01	− 3.5000E−02	1.3600E−03	5.8327E+01	5.8377E+01		
127	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1G_4^o$	5.8373E+01	− 5.7700E−02	4.3100E−04	5.8315E+01	5.8392E+01		
128	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3P_1^o$	5.8398E+01	− 2.0300E−02	6.4800E−04	5.8378E+01	5.8439E+01		
129	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.8435E+01	− 3.9400E−02	1.2800E−03	5.8397E+01	5.8465E+01		
130	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5D_0^o$	5.8484E+01	− 2.8500E−02	1.4100E−03	5.8457E+01	5.8528E+01		
131	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$)$^3S_1^o$	5.8488E+01	− 3.5300E−02	1.3600E−03	5.8454E+01	5.8529E+01		
132	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5P_3^o$	5.8527E+01	− 4.4800E−02	9.9900E−04	5.8483E+01	5.8539E+01		
133	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_7^o$	5.8562E+01	− 8.3800E−02	1.4100E−03	5.8479E+01	5.8556E+01		
134	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3P_2^o$	5.8610E+01	− 3.2300E−02	1.3600E−03	5.8579E+01	5.8646E+01		
135	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.8654E+01	− 3.2200E−02	1.1100E−03	5.8623E+01	5.8692E+01		
136	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^1G_4^o$	5.8728E+01	− 6.8300E−02	1.3000E−03	5.8660E+01	5.8721E+01		
137	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^1H_5^o$	5.8747E+01	− 6.8400E−02	1.3800E−03	5.8680E+01	5.8736E+01		
138	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3H_6^o$	5.8789E+01	− 7.8600E−02	1.4200E−03	5.8711E+01	5.8767E+01		
139	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	5.8857E+01	− 5.4200E−02	1.3600E−03	5.8805E+01	5.8862E+01		
140	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.8881E+01	− 5.3300E−02	1.2500E−03	5.8829E+01	5.8892E+01		
141	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.8915E+01	− 5.8400E−02	1.3600E−03	5.8858E+01	5.8926E+01		
142	$3s^23p^63d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1H_5^o$	5.8933E+01	− 6.8900E−02	1.4100E−03	5.8865E+01	5.8928E+01		
143	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3F_3^o$	5.9164E+01	− 4.8600E−02	1.1800E−03	5.9117E+01	5.9162E+01		
144	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.9204E+01	− 4.8100E−02	8.0600E−04	5.9157E+01	5.9203E+01		

Table 3 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF				FAC	NIST[24]	
		DC	Breit	QED	Total			
145	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.9207E+01	− 4.0900E−02	1.2600E−03	5.9167E+01	5.9214E+01		
146	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	5.9224E+01	− 5.5700E−02	1.2500E−03	5.9169E+01	5.9217E+01		
147	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	5.9324E+01	− 6.5300E−02	1.4000E−03	5.9261E+01	5.9310E+01		
148	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3P_2^o$	5.9340E+01	− 5.1100E−02	9.3000E−04	5.9290E+01	5.9323E+01		
149	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3G_4^o$	5.9346E+01	− 3.7600E−02	1.3700E−03	5.9310E+01	5.9342E+01		
150	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	5.9399E+01	− 5.0400E−02	7.4800E−04	5.9350E+01	5.9353E+01		
151	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3P_2^o$	5.9405E+01	− 4.0400E−02	2.9200E−04	5.9365E+01	5.9408E+01		
152	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5D_1^o$	5.9477E+01	− 5.3300E−02	1.3600E−03	5.9425E+01	5.9409E+01		
153	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5P_2^o$	5.9478E+01	− 5.5300E−02	6.3900E−04	5.9423E+01	5.9479E+01		
154	$3s^23p^63d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^3P_0^o$	5.9482E+01	− 5.5800E−02	1.3500E−03	5.9427E+01	5.9489E+01		
155	$3s^23p^63d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	5.9486E+01	− 5.7900E−02	1.2700E−03	5.9429E+01	5.9496E+01		
156	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^1F_3^o$	5.9550E+01	− 4.2200E−02	4.1300E−04	5.9508E+01	5.9502E+01		
157	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3F_2^o$	5.9553E+01	− 4.8300E−02	8.4600E−04	5.9506E+01	5.9559E+01		
158	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$)$^1I_6^o$	5.9651E+01	− 1.0400E−01	1.4100E−03	5.9548E+01	5.9570E+01		
159	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^1P_1^o$	5.9680E+01	− 5.9300E−02	4.1400E−04	5.9621E+01	5.9607E+01		
160	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3H_4^o$	5.9747E+01	− 9.2400E−02	1.4100E−03	5.9656E+01	5.9703E+01		
161	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$)$^1D_2^o$	5.9759E+01	− 6.4400E−02	1.2200E−03	5.9696E+01	5.9716E+01		
162	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	5.9778E+01	− 8.3700E−02	1.1400E−03	5.9695E+01	5.9761E+01		
163	$3s^23p^63d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3S_1^o$	5.9778E+01	− 6.4900E−02	1.3500E−03	5.9714E+01	5.9763E+01		
164	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3P_0^o$	5.9787E+01	− 2.3600E−02	− 8.0600E−04	5.9763E+01	5.9784E+01		
165	$3s^23p^63d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3H_5^o$	5.9824E+01	− 9.5300E−02	1.4100E−03	5.9731E+01	5.9790E+01		
166	$3s^13p^5$ ($^2P_1$) $3d^{10}$($^1S_0$)$^3P_0^o$	5.9832E+01	− 1.6000E−01	− 7.2700E−02	5.9599E+01	5.9793E+01		
167	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^1D_2^o$	6.0015E+01	− 7.2700E−02	1.3100E−03	5.9944E+01	5.9996E+01		
168	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3G_5^o$	6.0056E+01	− 8.8300E−02	1.3800E−03	5.9969E+01	6.0025E+01		
169	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3F_3^o$	6.0057E+01	− 4.4100E−02	1.3800E−03	6.0014E+01	6.0033E+01		
170	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	6.0063E+01	− 8.0500E−02	1.4000E−03	5.9984E+01	6.0058E+01		
171	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3D_1^o$	6.0134E+01	− 3.2000E−02	1.2800E−03	6.0104E+01	6.0145E+01		
172	$3s^23p^63d^7$ ($^2D_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	6.0179E+01	− 4.7300E−02	1.3400E−03	6.0133E+01	6.0181E+01		
173	$3s^23p^63d^7$ ($^2F_3$) $4p^1$ ($^2P_1$)$^1F_3^o$	6.0226E+01	− 7.0600E−02	1.2900E−03	6.0156E+01	6.0213E+01		
174	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3P_1^o$	6.0999E+01	− 6.4000E−02	1.4300E−03	6.0937E+01	6.0626E+01		
175	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3D_2^o$	6.1050E+01	− 6.5600E−02	1.4100E−03	6.0986E+01	6.0972E+01		
176	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3F_4^o$	6.1111E+01	− 6.0000E−02	1.4100E−03	6.1052E+01	6.1026E+01		
177	$3s^13p^5$ ($^2P_1$) $3d^{10}$($^1S_0$)$^1P_1^o$	6.1173E+01	− 1.5700E−01	− 7.5300E−02	6.0940E+01	6.1078E+01		
178	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$)$^3F_3^o$	6.1175E+01	− 8.0600E−02	1.3900E−03	6.1096E+01	6.1148E+01		
179	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3P_0^o$	6.1210E+01	− 8.0200E−02	1.3900E−03	6.1132E+01	6.1187E+01		
180	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^1D_2^o$	6.1288E+01	− 5.5000E−02	1.3800E−03	6.1234E+01	6.1267E+01		
181	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^1P_1^o$	6.1313E+01	− 6.8700E−02	1.3700E−03	6.1245E+01	6.1303E+01		
182	$3s^23p^63d^7$ ($^2D_1$) $4p^1$ ($^2P_1$) $^3D_3^o$	6.1438E+01	− 5.6200E−02	1.4000E−03	6.1383E+01	6.1419E+01		
183	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5F_4^e$	6.3047E+01	− 3.0300E−02	− 3.6200E−05	6.3016E+01	6.3048E+01		
184	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5F_5^e$	6.3102E+01	− 4.0200E−02	− 2.3100E−05	6.3062E+01	6.3097E+01		
185	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5P_3^e$	6.3107E+01	− 2.6000E−02	− 2.7600E−05	6.3081E+01	6.3116E+01		
186	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^3H_6^e$	6.3221E+01	− 4.3000E−02	4.9100E−07	6.3178E+01	6.3220E+01		
187	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^3D_3^e$	6.3456E+01	− 3.8700E−02	− 3.3900E−05	6.3417E+01	6.3445E+01		
188	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^3G_5^e$	6.3481E+01	− 4.9000E−02	− 1.0300E−05	6.3432E+01	6.3474E+01		
189	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5H_7^e$	6.3520E+01	− 5.5100E−02	4.7000E−06	6.3465E+01	6.3502E+01		
190	$3s^23p^63d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5G_6^e$	6.3533E+01	− 5.0700E−02	− 5.7500E−07	6.3482E+01	6.3524E+01		
191	$3s^23p^63d^7$ ($^2D_3$) $4d^1$ ($^2D_1$)$^3D_1^e$	6.3570E+01	− 4.9800E−03	− 8.7500E−05	6.3565E+01	6.3579E+01		
192	$3s^23p^63d^7$ ($^2P_3$) $4d^1$ ($^2D_1$) $^3D_2^e$	6.3640E+01	− 2.2400E−02	− 6.4000E−05	6.3618E+01	6.3640E+01		

Table 3 continued

Level no	Configurations	Energies (in Ryd.)						ΔE
		MCDF				FAC	NIST[24]	
		DC	Breit	QED	Total			
193	$3s^2 3p^6 3d^7$ ($^2F_3$) $4d^1$ ($^2D_1$)$^3F_3^e$	6.3733E+01	− 1.6800E−02	− 5.7200E−05	6.3716E+01	6.3744E+01		
194	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^3P_2^e$	6.3742E+01	− 2.5300E−02	− 6.7200E−06	6.3717E+01	6.3753E+01		
195	$3s^2 3p^6 3d^7$ ($^2P_3$) $4d^1$ ($^2D_1$) $^3P_0^e$	6.3750E+01	− 6.6100E−03	− 9.6500E−05	6.3744E+01	6.3769E+01		
196	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5D_4^e$	6.3834E+01	− 4.0200E−02	− 1.6000E−05	6.3794E+01	6.3835E+01		
197	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5G_3^e$	6.3938E+01	− 2.2400E−02	− 2.6000E−05	6.3916E+01	6.3941E+01		
198	$3s^2 3p^6 3d^7$ ($^4P_3$) $4d^1$ ($^2D_1$) $^5P_1^e$	6.3992E+01	− 1.2100E−02	− 4.7700E−05	6.3980E+01	6.4000E+01		
199	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5F_2^e$	6.4013E+01	− 1.4700E−02	− 2.7300E−05	6.3998E+01	6.4016E+01		
200	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5H_4^e$	6.4016E+01	− 3.5700E−02	− 8.5600E−06	6.3980E+01	6.4030E+01		

Table 4 Comparison of excitation energies (in Ryd.) of five different configurations sets for first 24 excited states levels for Fe-like Cs

Level no	MCDF1	MCDF2	MCDF3	MCDF4	MCDF5	ΔE_1	ΔE_2	ΔE_3	ΔE_4
2	0.471	0.471	0.445	0.471	0.445	0.000	− 0.026	0.000	− 0.026
3	1.020	1.020	1.020	1.020	1.020	0.000	0.000	0.000	0.000
4	1.311	1.306	1.246	1.311	1.242	− 0.005	− 0.065	0.000	− 0.070
5	1.464	1.464	1.448	1.464	1.448	0.000	− 0.016	0.000	− 0.016
6	1.741	1.741	1.713	1.741	1.713	0.000	− 0.028	0.000	− 0.028
7	1.818	1.818	1.818	1.818	1.818	0.000	0.000	0.000	0.000
8	2.411	2.411	2.394	2.411	2.394	0.000	− 0.017	0.000	− 0.017
9	3.805	3.790	3.685	3.805	3.673	− 0.015	− 0.121	0.000	− 0.132
10	18.644	18.644	18.644	18.644	18.644	0.000	0.000	0.000	0.000
11	19.062	19.062	19.062	19.056	19.055	0.000	0.000	− 0.006	− 0.007
12	19.713	19.713	19.713	19.713	19.713	0.000	0.000	0.000	0.000
13	20.378	20.378	20.378	20.319	20.318	0.000	0.000	− 0.060	− 0.060
14	20.481	20.481	20.481	20.476	20.476	0.000	0.000	− 0.005	− 0.005
15	20.781	20.781	20.781	20.757	20.757	0.000	0.000	− 0.024	− 0.024
16	20.947	20.947	20.947	20.947	20.947	0.000	0.000	0.000	0.000
17	21.051	21.051	21.051	20.999	20.999	0.000	0.000	− 0.052	− 0.052
18	24.837	24.837	24.837	24.823	24.823	0.000	0.000	− 0.014	− 0.014
19	25.099	25.099	25.099	25.099	25.098	0.000	0.000	− 0.001	− 0.001
20	25.464	25.464	25.464	25.464	25.464	0.000	0.000	0.000	0.000
21	26.768	26.768	26.768	26.677	26.677	0.000	0.000	− 0.091	− 0.091
22	35.389	35.389	35.389	35.389	35.389	0.000	0.000	0.000	0.000
23	35.802	35.802	35.653	35.802	35.653	0.000	− 0.149	0.000	− 0.149
24	36.506	36.506	36.506	36.506	36.506	0.000	0.000	0.000	0.000
25	37.199	37.199	36.678	37.198	36.678	0.000	− 0.521	0.000	− 0.521

We have compared our calculated fine structure energies with energies compiled by NIST [25]. The difference ΔE of our calculated fine structure energies from GRASP with energies compiled by NIST is listed in Table 5.

In Table 6, we list the excitation energies of first 24 excited states for 5 different sets of configurations. In Table 6, configuration sets MCDF1, MCDF2, MCDSF3, MCDf4 and MCDF5 are same as for Fe-like Xe in Table 2.

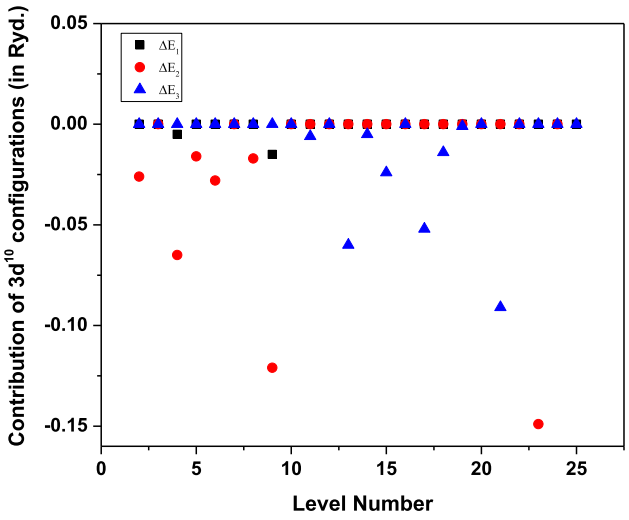


Fig. 2 Contribution of $3d^{10}$ configurations in lowest 24 excited states of Fe-like Cs

In Table 6, difference of excitation energies of MCDF1 set with MCDF2, MCDF3, MCDF4, MCDF5 is given under the columns ΔE_1, ΔE_2, ΔE_3 and ΔE_4, resp. From ΔE_1, we found that the effect of $3p^6 3d^{10}$ configuration is small and only in level number 4 and 9, while the effect of other two $3d^{10}$ configurations is considerable. We have also found that there is no change in excitation energies of some levels such as level number 7, 10, 12, 16, 20, 22 and 24 due to $3d^{10}$ configurations or we can say that the there is no or negligible contribution of $3d^{10}$ configurations to these levels.

In Fig. 3, we show the separate contribution of all $3d^{10}$ configurations in lowest 24 excited states for Fe-like Ba. From Fig. 3, we observe that the contribution of $3s3p^5 3d^{10}$ and 3 $os^2 3p^4 3d^{10}$ configurations is very small for most of the levels similar to Fe-like Ba. From Table 6 and Fig. 2, we have also seen that $3d^{10}$ configurations are lowering the position of levels in energy spectrum. We have not found any change in excitation energies of levels greater than 25 due to inclusion of doubly excited $3d^{10}$ configurations and therefore, in Table 4, we only show up to lowest 25 levels for Fe-like Ba.

So overall, for Fe-like ions, we found that the contribution of $3d^{10}$ configurations to some excited states is considerable and this affects the position of levels in energy spectrum and hence will also affect the calculations of collisional excitation cross sections and photoionization cross sections.

Further, it is also found that doubly excited $3s^2 3p^4 3d^{10}$ configurations have less energies as compared to $n = 4$ configurations ,while $3s3p^5 3d^{10}$ configurations position lies in between $n = 4$ configurations positions in energy spectrum for all Fe-like ions. So, inclusion of these doubly excited states is important for the study of collisional excitation and photoionization process of Fe-like ions.

5.4. Radiative data

The transition parameters for multipole radiative transitions can be calculated from reduced matrix element by following relation

$$\left\langle \psi_\alpha(PJM) \middle\| Q_k^{(\lambda)} \middle\| \psi_{\alpha'}(P'J'M') \right\rangle \tag{1}$$

where $Q_k^{(\lambda)}$ is the corresponding transition operator in length or velocity gauge and $\psi_\alpha(PJM)$ and $\psi_\alpha(P'J'M')$ are atomic state functions for initial and final states. Here, P represents parity, while J and M are angular quantum numbers.

For electric and magnetic dipole and quadrupole transitions, oscillator strength f_{ij} (is dimensionless) and transition rates A_{ij} (in s^{-1}) are related to the line strength S_{ij} (in a.u.) from following relations:

(i) for electric dipole transitions (E1)

$$A_{ji} = \frac{2.0261 \times 10^{18}}{\omega_j \lambda_{ji}^3} S_{ij} \quad \text{and} \quad f_{ij} = \frac{303.75}{\lambda_{ji}\omega_i} S_{ij} \tag{2}$$

(ii) for magnetic dipole transitions (M1)

$$A_{ji} = \frac{2.6974 \times 10^{13}}{\omega_j \lambda_{ji}^3} S_{ij} \quad \text{and} \quad f_{ij} = \frac{4.044 \times 10^{-3}}{\lambda_{ji}\omega_i} S_{ij} \tag{3}$$

(iii) for electric quadrupole transition (E2)

$$A_{ji} = \frac{1.1199 \times 10^{18}}{\omega_j \lambda_{ji}^5} S_{ij} \quad \text{and} \quad f_{ij} = \frac{167.89}{\lambda_{ji}^3\omega_i} S_{ij} \tag{4}$$

(iii) for magnetic quadrupole transition (M2)

$$A_{ji} = \frac{1.4910 \times 10^{13}}{\omega_j \lambda_{ji}^5} S_{ij} \quad \text{and} \quad f_{ij} = \frac{2.236 \times 10^{-3}}{\lambda_{ji}^3\omega_i} S_{ij} \tag{5}$$

Oscillator strength and transition rates are also related to each other by following expression,

$$f_{ij} = \frac{1.49 \times 10^{-16} \lambda_{ji}^2 \omega_j}{\omega_i} A_{ji} \tag{6}$$

In Eqs. (2)–(6), ω_i and ω_j are statistical weights for lower level (i) and upper level (j) and λ_{ji} denotes transition wavelength of radiative transition from upper level (j) to lower level (i).

We have listed transition data of electric and magnetic dipole and quadrupole transitions within lowest 200 levels of Fe-like Xe, Cs and Ba ions as supplementary material.

Table 5 Excitation energies (in Ryd.) for lowest 200 levels for Fe-like Ba

Level No	Configurations	Energies (in Ryd.)						ΔE
		MCDF				FAC	NIST[25]	
		DC	Breit	QED	Total			
1	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^3F_2)\ {}^3F_4^e$	0.0000	–	–	0.0000	0.0000	0.0000	–
2	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^1D_2)\ {}^1D_2^e$	4.3859E−01	1.8700E−02	− 3.1900E−05	4.5729E−01	4.5733E−01	4.470E−01	0.0103
3	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^3F_2)\ {}^3F_3^e$	1.1508E+00	− 2.6600E−02	2.7100E−06	1.1241E+00	1.1255E+00	1.135E+00	0.0109
4	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^3P_2)\ {}^3P_0^e$	1.2597E+00	2.9100E−02	− 3.0700E−05	1.2887E+00	1.2890E+00	1.263E+00	0.0257
5	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^3P_2)\ {}^3P_2^e$	1.5915E+00	− 2.4600E−02	− 3.5400E−06	1.5669E+00	1.5686E+00	1.564E+00	0.0029
6	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^3P_2)\ {}^3P_1^e$	1.8424E+00	− 5.2400E−03	− 9.0600E−08	1.8371E+00	1.8392E+00	1.820E+00	0.0171
7	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^1G_2)\ {}^1G_4^e$	1.9814E+00	− 4.2000E−02	3.0800E−06	1.9393E+00	1.9417E+00	1.922E+00	0.0173
8	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^3F_2)\ {}^3F_2^e$	2.6693E+00	− 5.8900E−02	− 1.9300E−05	2.6104E+00	2.6132E+00	2.620E+00	0.0096
9	$3s^2 3p^6\ (^1S_0)\ 3d^8\ (^1S_0)\ {}^1S_0^e$	3.9466E+00	− 3.0200E−02	− 3.2100E−05	3.9164E+00	3.9188E+00	3.903E+00	0.0134
10	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3F_4^o$	1.9281E+01	− 5.0800E−02	− 6.4900E−03	1.9224E+01	1.9158E+01	1.8989E+01	0.2350
11	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^1D_2^o$	1.9702E+01	− 4.4000E−02	− 6.4800E−03	1.9652E+01	1.9585E+01	1.9373E+01	0.2790
12	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3D_3^o$	2.0403E+01	− 5.7600E−02	− 6.4800E−03	2.0338E+01	2.0270E+01	2.0018E+01	0.3200
13	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3P_1^o$	2.1006E+01	− 4.0600E−02	− 6.5800E−03	2.0959E+01	2.0881E+01	2.0636E+01	0.3230
14	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3D_2^o$	2.1260E+01	− 7.7100E−02	− 6.4700E−03	2.1177E+01	2.1116E+01	2.0870E+01	0.3070
15	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3P_0^o$	2.1542E+01	− 6.3500E−02	− 6.5200E−03	2.1472E+01	2.1408E+01	2.1165E+01	0.3070
16	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3F_3^o$	2.1750E+01	− 8.0900E−02	− 5.9500E−03	2.1663E+01	2.1595E+01	2.1349E+01	0.3140
17	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3D_1^o$	2.1778E+01	− 6.8300E−02	− 6.5200E−03	2.1703E+01	2.1634E+01	2.1371E+01	0.3320
18	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3P_2^o$	2.6036E+01	− 1.2000E−01	1.2200E−03	2.5917E+01	2.5874E+01	2.5598E+01	0.3190
19	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^3F_2^o$	2.6447E+01	− 1.6700E−01	1.2800E−03	2.6281E+01	2.6254E+01	2.6063E+01	0.2180
20	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^1F_3^o$	2.6683E+01	− 1.3300E−01	7.5200E−04	2.6551E+01	2.6498E+01	2.6192E+01	0.3590
21	$3s^2 3p^5\ (^2P_1)\ 3d^9\ (^2D_1)\ {}^1P_1^o$	2.8048E+01	− 1.5400E−01	9.8500E−04	2.7895E+01	2.7855E+01	2.7578E+01	0.3170
22	$3s^1 3p^6\ (^1S_0)\ 3d^9\ (^2D_1)\ {}^3D_3^e$	3.6920E+01	− 4.1600E−02	− 8.2300E−02	3.6796E+01	3.6700E+01		
23	$3s^1 3p^6\ (^1S_0)\ 3d^9\ (^2D_1)\ {}^3D_2^e$	3.7182E+01	− 4.1400E−02	− 7.8800E−02	3.7062E+01	3.6954E+01		
24	$3s^1 3p^6\ (^1S_0)\ 3d^9\ (^2D_1)\ {}^3D_1^e$	3.8176E+01	− 7.0100E−02	− 8.2300E−02	3.8024E+01	3.7940E+01		
25	$3s^1 3p^6\ (^1S_0)\ 3d^9\ (^2D_1)\ {}^1D_2^e$	3.8313E+01	− 7.9300E−02	− 7.3600E−02	3.8160E+01	3.8047E+01		
26	$3s^2 3p^4\ (^3P_2)\ 3d^{10}\ (^1S_0)\ {}^3P_2^e$	4.0812E+01	− 9.4800E−02	− 2.1900E−02	4.0695E+01	4.0408E+01		
27	$3s^2 3p^4\ (^1S_0)\ 3d^{10}\ (\ ^1S_0)\ {}^1S_0^e$	4.1370E+01	− 8.4700E−02	− 1.3300E−02	4.1272E+01	4.0947E+01		
28	$3s^2 3p^4\ (^3P_2)\ 3d^{10}\ (^1S_0)\ {}^3P_1^e$	4.6038E+01	− 1.7600E−01	− 5.1600E−03	4.5857E+01	4.5598E+01		
29	$3s^2 3p^4\ (^1D_2)\ 3d^{10}\ (^1S_0)\ {}^1D_2^e$	4.6922E+01	− 1.8000E−01	− 8.2700E−03	4.6733E+01	4.6466E+01		
30	$3s^2 3p^4\ (^3P_2)\ 3d^{10}\ (^1S_0)\ {}^3P_0^e$	5.2571E+01	− 2.5600E−01	1.7000E−03	5.2317E+01	5.2101E+01		
31	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)\ {}^5F_5^e$	5.4047E+01	− 2.8000E−02	2.2300E−02	5.4041E+01	5.4101E+01		
32	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)\ {}^3F_4^e$	5.4238E+01	− 2.4200E−02	2.2300E−02	5.4236E+01	5.4308E+01		
33	$3s^2 3p^6 3d^7\ (^2P_3)\ 4s^1\ (^2S_1)\ {}^3P_2^e$	5.4736E+01	3.3000E−03	2.2200E−02	5.4761E+01	5.4809E+01		
34	$3s^2 3p^6 3d^7\ (^2P_3)\ 4s^1\ (^2S_1)\ {}^1P_1^e$	5.4793E+01	3.9400E−03	2.2200E−02	5.4819E+01	5.4872E+01		
35	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)\ {}^5F_3^e$	5.4925E+01	− 7.6300E−03	2.2300E−02	5.4940E+01	5.4989E+01		
36	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)^5F_2^e$	5.5024E+01	− 5.7500E−03	2.2300E−02	5.5041E+01	5.5096E+01		
37	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)\ {}^5F_4^e$	5.5199E+01	− 4.9200E−02	2.2300E−02	5.5172E+01	5.5246E+01		
38	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)\ {}^3F_3^e$	5.5313E+01	− 4.4300E−02	2.2300E−02	5.5291E+01	5.5369E+01		
39	$3s^2 3p^6 3d^7\ (^2G_3)\ 4s^1\ (^2S_1)\ {}^3G_5^e$	5.5681E+01	− 6.3500E−02	2.2300E−02	5.5640E+01	5.5701E+01		
40	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)^5F_1^e$	5.5685E+01	− 3.1600E−02	2.2300E−02	5.5675E+01	5.5740E+01		
41	$3s^2 3p^6 3d^7\ (^4F_3)\ 4s^1\ (^2S_1)^3F_2^e$	5.5731E+01	− 3.1100E−02	2.2300E−02	5.5722E+01	5.5784E+01		
42	$3s^2 3p^6 3d^7\ (^4P_3)\ 4s^1\ (^2S_1)^5P_3^e$	5.5738E+01	− 4.0600E−02	2.2300E−02	5.5720E+01	5.5786E+01		
43	$3s^2 3p^6 3d^7\ (^2H_3)\ 4s^1\ (^2S_1)\ {}^3H_4^e$	5.5758E+01	− 6.1400E−02	2.2300E−02	5.5719E+01	5.5791E+01		
44	$3s^2 3p^6 3d^7\ (^4P_3)\ 4s^1\ (^2S_1)^3P_2^e$	5.5911E+01	− 3.8300E−02	2.2300E−02	5.5895E+01	5.5952E+01		
45	$3s^2 3p^6 3d^7\ (^4P_3)\ 4s^1\ (^2S_1)^5P_1^e$	5.5947E+01	− 3.5000E−02	2.2300E−02	5.5934E+01	5.5968E+01		
46	$3s^2 3p^6 3d^7\ (^2P_3)\ 4s^1\ (^2S_1)\ {}^3P_0^e$	5.6036E+01	− 3.4200E−02	2.2200E−02	5.6024E+01	5.5996E+01		
47	$3s^2 3p^6 3d^7\ (^2H_3)\ 4s^1\ (^2S_1)\ {}^3H_6^e$	5.6097E+01	− 8.7100E−02	2.2300E−02	5.6032E+01	5.6087E+01		
48	$3s^2 3p^6 3d^7\ (^2H_3)\ 4s^1\ (^2S_1)\ {}^1H_5^e$	5.6178E+01	− 8.3400E−02	2.2300E−02	5.6117E+01	5.6088E+01		

Table 5 continued

Level No	Configurations	Energies (in Ryd.)					ΔE
		MCDF				FAC	NIST[25]
49	$3s^2 3p^6 3d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_3^e$	5.6299E+01	− 6.1300E−02	2.2300E−02	5.6260E+01	5.6177E+01	
50	$3s^2 3p^6 3d^7$ ($^2G_3$) $4s^1$ ($^2S_1$)$^3G_4^e$	5.6320E+01	− 6.2900E−02	2.2300E−02	5.6280E+01	5.6318E+01	
51	$3s^1 3p^5$ ($^2P_1$) $3d^{10}$ ($^1S_0$) $^3P_2^o$	5.6478E+01	− 9.2400E−02	− 8.8800E−02	5.6297E+01	5.6338E+01	
52	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^3F_2^e$	5.6730E+01	− 3.6100E−02	2.2300E−02	5.6716E+01	5.6756E+01	
53	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^3F_3^e$	5.6757E+01	− 3.9000E−02	2.2300E−02	5.6741E+01	5.6782E+01	
54	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$)$^3D_3^e$	5.6852E+01	− 6.8100E−02	2.2300E−02	5.6806E+01	5.6874E+01	
55	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$)$^3D_2^e$	5.6872E+01	− 6.4400E−02	2.2300E−02	5.6829E+01	5.6897E+01	
56	$3s^2 3p^6 3d^7$ ($^4P_3$) $4s^1$ ($^2S_1$)$^3P_1^e$	5.6944E+01	− 6.2600E−02	2.2300E−02	5.6903E+01	5.6972E+01	
57	$3s^2 3p^6 3d^7$ ($^4P_3$) $4s^1$ ($^2S_1$) $^5P_2^e$	5.6971E+01	− 6.8200E−02	2.2300E−02	5.6925E+01	5.7001E+01	
58	$3s^2 3p^6 3d^7$ ($^2P_3$) $4s^1$ ($^2S_1$) $^3P_1^e$	5.7243E+01	− 7.1300E−02	2.2200E−02	5.7194E+01	5.7259E+01	
59	$3s^2 3p^6 3d^7$ ($^2G_3$) $4s^1$ ($^2S_1$) $^1G_4^e$	5.7283E+01	− 1.0600E−01	2.2300E−02	5.7199E+01	5.7265E+01	
60	$3s^2 3p^6 3d^7$ ($^2H_3$) $4s^1$ ($^2S_1$) $^3H_5^e$	5.7298E+01	− 1.1000E−01	2.2300E−02	5.7211E+01	5.7278E+01	
61	$3s^2 3p^6 3d^7$ ($^4P_3$) $4s^1$ ($^2S_1$) $^3P_0^e$	5.7339E+01	− 7.4100E−02	2.2300E−02	5.7288E+01	5.7362E+01	
62	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^3D_1^e$	5.7528E+01	− 3.7800E−02	2.2300E−02	5.7512E+01	5.7535E+01	
63	$3s^2 3p^6 3d^7$ ($^2D_3$) $4s^1$ ($^2S_1$) $^1D_2^e$	5.7539E+01	− 3.8700E−02	2.2300E−02	5.7522E+01	5.7549E+01	
64	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$) $^3F_4^e$	5.7565E+01	− 9.2000E−02	2.2300E−02	5.7495E+01	5.7550E+01	
65	$3s^2 3p^6 3d^7$ ($^2F_3$) $4s^1$ ($^2S_1$)$^1F_3^e$	5.7625E+01	− 8.7600E−02	2.2300E−02	5.7560E+01	5.7559E+01	
66	$3s^1 3p^5$ ($^2P_1$) $3d^{10}$ ($^1S_0$) $^3P_1^o$	5.8074E+01	− 9.3200E−02	− 8.8200E−02	5.7893E+01	5.7616E+01	
67	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^3D_4^o$	5.8230E+01	5.5300E−03	− 5.0800E−04	5.8235E+01	5.8288E+01	
68	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$)$^3G_5^o$	5.8333E+01	− 3.6700E−03	− 5.4400E−04	5.8329E+01	5.8388E+01	
69	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_3^e$	5.8738E+01	− 6.3600E−02	2.2300E−02	5.8697E+01	5.8725E+01	
70	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$) $^3D_1^e$	5.8744E+01	− 8.8600E−02	2.2300E−02	5.8678E+01	5.8732E+01	
71	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^3D_2^e$	5.8756E+01	− 8.4400E−02	2.2300E−02	5.8694E+01	5.8743E+01	
72	$3s^2 3p^6 3d^7$ ($^2D_1$) $4s^1$ ($^2S_1$)$^1D_2^e$	5.8849E+01	− 6.7400E−02	2.2300E−02	5.8804E+01	5.8846E+01	
73	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3P_1^o$	5.8874E+01	3.2400E−02	− 5.6400E−04	5.8906E+01	5.8944E+01	
74	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_2^o$	5.8918E+01	2.9000E−02	− 5.7800E−04	5.8947E+01	5.8987E+01	
75	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_2^o$	5.9079E+01	2.4900E−02	− 4.8600E−04	5.9103E+01	5.9142E+01	
76	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_3^o$	5.9133E+01	1.8800E−02	− 4.8700E−04	5.9151E+01	5.9194E+01	
77	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_3^o$	5.9361E+01	− 1.8100E−02	− 4.6300E−04	5.9343E+01	5.9405E+01	
78	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_4^o$	5.9438E+01	− 2.5300E−02	− 5.2400E−04	5.9412E+01	5.9479E+01	
79	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5S_2^o$	5.9842E+01	− 9.6700E−03	− 4.5600E−04	5.9832E+01	5.9882E+01	
80	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_2^o$	5.9859E+01	− 7.6700E−03	− 4.7100E−04	5.9851E+01	5.9899E+01	
81	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_1^o$	5.9865E+01	− 3.8600E−03	− 4.9900E−04	5.9861E+01	5.9900E+01	
82	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	5.9876E+01	− 3.1600E−02	− 4.0900E−04	5.9844E+01	5.9905E+01	
83	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_5^o$	5.9885E+01	− 3.8300E−02	− 4.0300E−04	5.9846E+01	5.9916E+01	
84	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_0^o$	6.0001E+01	− 6.5000E−03	− 4.2000E−04	5.9994E+01	6.0037E+01	
85	$3s^2 3p^6 3d^7$ ($^4P_3$) $4p^1$ ($^2P_1$) $^5D_3^o$	6.0048E+01	− 1.7100E−02	− 4.4000E−04	6.0030E+01	6.0094E+01	
86	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5G_6^o$	6.0151E+01	− 3.0700E−02	1.5500E−03	6.0122E+01	6.0177E+01	
87	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^5F_5^o$	6.0163E+01	− 3.1400E−02	1.3500E−03	6.0133E+01	6.0189E+01	
88	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_1^o$	6.0173E+01	− 9.7500E−03	− 5.1900E−04	6.0162E+01	6.0218E+01	
89	$3s^2 3p^6 3d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3I_6^o$	6.0310E+01	− 6.2900E−02	− 5.0700E−04	6.0246E+01	6.0296E+01	
90	$3s^2 3p^6 3d^7$ ($^2H_3$) $4p^1$ ($^2P_1$) $^3G_5^o$	6.0340E+01	− 5.3300E−02	− 3.7400E−04	6.0287E+01	6.0343E+01	
91	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3F_4^o$	6.0375E+01	− 2.6600E−02	1.4600E−03	6.0350E+01	6.0423E+01	
92	$3s^2 3p^6 3d^7$ ($^4F_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	6.0398E+01	− 2.0100E−02	1.2200E−03	6.0379E+01	6.0441E+01	
93	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^3H_4^o$	6.0464E+01	− 3.8600E−02	− 4.9600E−04	6.0425E+01	6.0477E+01	
94	$3s^2 3p^6 3d^7$ ($^2G_3$) $4p^1$ ($^2P_1$) $^1F_3^o$	6.0530E+01	− 3.0000E−02	− 2.4500E−04	6.0500E+01	6.0558E+01	
95	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^1S_0^o$	6.0758E+01	4.6100E−03	1.3900E−03	6.0764E+01	6.0805E+01	
96	$3s^2 3p^6 3d^7$ ($^2P_3$) $4p^1$ ($^2P_1$) $^3D_3^o$	6.0835E+01	− 5.3200E−03	1.4300E−03	6.0831E+01	6.0867E+01	
97	$3s^2 3p^6 3d^7$ ($^2F_3$) $4p^1$ ($^2P_1$) $^3G_3^o$	6.0852E+01	− 1.4000E−02	− 2.3000E−04	6.0838E+01	6.0886E+01	

Table 5 continued

Level No	Configurations	Energies (in Ryd.)					ΔE
		MCDF				FAC NIST[25]	
98	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)\ ^3P_2^o$	6.0852E+01	− 3.0900E−03	1.3400E−03	6.0851E+01	6.0900E+01	
99	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^1D_2^o$	6.0911E+01	− 1.0600E−02	− 2.3200E−04	6.0900E+01	6.0935E+01	
100	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5G_2^o$	6.0973E+01	− 3.3800E−02	− 3.1700E−04	6.0939E+01	6.0993E+01	
101	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)\ ^1P_1^o$	6.1003E+01	− 6.4800E−03	1.2700E−03	6.0998E+01	6.1054E+01	
102	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^1F_3^o$	6.1053E+01	− 4.1300E−02	− 2.5300E−04	6.1011E+01	6.1073E+01	
103	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^3G_4^o$	6.1057E+01	− 7.7600E−03	1.5700E−03	6.1051E+01	6.1098E+01	
104	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^5D_1^o$	6.1082E+01	− 2.8200E−02	− 1.6200E−04	6.1053E+01	6.1119E+01	
105	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3F_2^o$	6.1102E+01	− 7.5200E−03	1.2400E−03	6.1095E+01	6.1146E+01	
106	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3D_3^o$	6.1152E+01	− 9.8800E−03	1.2900E−03	6.1143E+01	6.1199E+01	
107	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^5P_2^o$	6.1199E+01	− 3.4300E−02	− 2.2400E−04	6.1165E+01	6.1229E+01	
108	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5D_4^o$	6.1237E+01	− 4.7500E−02	1.5800E−03	6.1191E+01	6.1256E+01	
109	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5G_5^o$	6.1290E+01	− 5.1300E−02	1.5900E−03	6.1240E+01	6.1309E+01	
110	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3D_1^o$	6.1314E+01	− 4.6800E−03	1.4500E−03	6.1311E+01	6.1336E+01	
111	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)\ ^3P_0^o$	6.1331E+01	− 4.6400E−02	− 5.4800E−04	6.1284E+01	6.1372E+01	
112	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3I_5^o$	6.1442E+01	− 8.4800E−02	− 4.6500E−04	6.1357E+01	6.1412E+01	
113	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^3F_3^o$	6.1462E+01	− 4.5200E−02	1.3600E−03	6.1418E+01	6.1496E+01	
114	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^3D_2^o$	6.1496E+01	− 3.7300E−02	1.0500E−03	6.1460E+01	6.1497E+01	
115	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^5P_1^o$	6.1504E+01	− 4.3400E−02	− 2.8700E−04	6.1461E+01	6.1522E+01	
116	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3H_4^o$	6.1515E+01	− 7.6600E−02	− 3.4400E−04	6.1438E+01	6.1526E+01	
117	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3F_2^o$	6.1618E+01	− 2.2700E−02	1.6600E−04	6.1595E+01	6.1638E+01	
118	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3P_1^o$	6.1730E+01	− 1.6000E−02	− 5.2200E−05	6.1714E+01	6.1723E+01	
119	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^3H_6^o$	6.1733E+01	− 6.6600E−02	1.5800E−03	6.1668E+01	6.1729E+01	
120	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3G_4^o$	6.1742E+01	− 6.4800E−02	5.7000E−04	6.1678E+01	6.1752E+01	
121	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^3F_3^o$	6.1765E+01	− 5.7700E−02	5.3200E−04	6.1708E+01	6.1766E+01	
122	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^3F_2^o$	6.1773E+01	− 3.1100E−02	1.2600E−03	6.1743E+01	6.1791E+01	
123	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^5D_4^o$	6.1782E+01	− 5.0400E−02	1.1200E−03	6.1733E+01	6.1792E+01	
124	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5D_3^o$	6.1838E+01	− 4.2300E−02	9.8000E−04	6.1797E+01	6.1810E+01	
125	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^3G_5^o$	6.1843E+01	− 6.4000E−02	1.5200E−03	6.1781E+01	6.1842E+01	
126	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3G_4^o$	6.1878E+01	− 5.8300E−02	9.9400E−04	6.1821E+01	6.1859E+01	
127	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3P_1^o$	6.1890E+01	− 3.8500E−02	1.5400E−03	6.1853E+01	6.1882E+01	
128	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3D_2^o$	6.1891E+01	− 2.6600E−02	1.1300E−03	6.1865E+01	6.1916E+01	
129	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3D_3^o$	6.1955E+01	− 4.4100E−02	1.4600E−03	6.1912E+01	6.1929E+01	
130	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3S_1^o$	6.2020E+01	− 3.8500E−02	1.5400E−03	6.1983E+01	6.1980E+01	
131	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5D_0^o$	6.2023E+01	− 3.0600E−02	1.5800E−03	6.1994E+01	6.2047E+01	
132	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^5P_3^o$	6.2048E+01	− 4.6700E−02	1.3200E−03	6.2002E+01	6.2064E+01	
133	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3I_7^o$	6.2090E+01	− 9.1000E−02	1.5900E−03	6.2001E+01	6.2065E+01	
134	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3P_2^o$	6.2142E+01	− 3.5600E−02	1.5400E−03	6.2108E+01	6.2074E+01	
135	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3D_1^o$	6.2180E+01	− 3.5200E−02	1.3300E−03	6.2146E+01	6.2172E+01	
136	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^1G_4^o$	6.2257E+01	− 7.4700E−02	1.5000E−03	6.2184E+01	6.2213E+01	
137	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^1H_5^o$	6.2282E+01	− 7.4700E−02	1.5600E−03	6.2209E+01	6.2243E+01	
138	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3H_6^o$	6.2323E+01	− 8.5300E−02	1.6000E−03	6.2239E+01	6.2263E+01	
139	$3s^1 3p^5\ (^2P_1)\ 3d^{10}\ (^1S_0)\ ^3P_0^o$	6.2334E+01	− 1.7000E−01	− 8.1100E−02	6.2083E+01	6.2293E+01	
140	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^3G_3^o$	6.2404E+01	− 5.8500E−02	1.5400E−03	6.2347E+01	6.2402E+01	
141	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3D_2^o$	6.2431E+01	− 5.7000E−02	1.4500E−03	6.2376E+01	6.2437E+01	
142	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^1G_4^o$	6.2460E+01	− 6.2900E−02	1.5400E−03	6.2398E+01	6.2463E+01	
143	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^1H_5^o$	6.2483E+01	− 7.3700E−02	1.5800E−03	6.2411E+01	6.2471E+01	
144	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3F_3^o$	6.2723E+01	− 4.5700E−02	1.0700E−03	6.2679E+01	6.2710E+01	
145	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3F_2^o$	6.2772E+01	− 5.2800E−02	8.7400E−04	6.2720E+01	6.2764E+01	
146	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3D_1^o$	6.2778E+01	− 4.2700E−02	1.4400E−03	6.2736E+01	6.2779E+01	

Table 5 continued

Level No	Configurations	Energies (in Ryd.)					ΔE
		MCDF				FAC NIST[25]	
147	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3D_3^o$	6.2821E+01	− 6.6400E−02	1.5500E−03	6.2756E+01	6.2813E+01	
148	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^1G_4^o$	6.2890E+01	− 4.4700E−02	5.9500E−05	6.2846E+01	6.2880E+01	
149	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3P_2^o$	6.2891E+01	− 4.3200E−02	1.5400E−03	6.2849E+01	6.2890E+01	
150	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3F_4^o$	6.2945E+01	− 6.9500E−02	1.5800E−03	6.2877E+01	6.2939E+01	
151	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3P_2^o$	6.2948E+01	− 4.7900E−02	9.4000E−04	6.2901E+01	6.2947E+01	
152	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^3D_1^o$	6.2971E+01	− 5.6700E−02	6.7900E−04	6.2915E+01	6.2972E+01	
153	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3F_3^o$	6.3071E+01	− 4.7800E−02	3.0100E−04	6.3024E+01	6.3073E+01	
154	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^1D_2^o$	6.3076E+01	− 6.4800E−02	7.7600E−04	6.3012E+01	6.3074E+01	
155	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5D_1^o$	6.3099E+01	− 6.0500E−02	1.5600E−03	6.3040E+01	6.3103E+01	
156	$3s^2 3p^6 3d^7\ (^4P_3)\ 4p^1\ (^2P_1)\ ^3P_0^o$	6.3106E+01	− 6.1500E−02	1.5400E−03	6.3046E+01	6.3113E+01	
157	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^3G_3^o$	6.3111E+01	− 5.9100E−02	1.5500E−03	6.3053E+01	6.3116E+01	
158	$3s^2 3p^6 3d^7\ (^4F_3)\ 4p^1\ (^2P_1)\ ^5D_2^o$	6.3152E+01	− 5.8500E−02	1.2700E−03	6.3094E+01	6.3119E+01	
159	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^1P_1^o$	6.3277E+01	− 6.6100E−02	5.6700E−04	6.3211E+01	6.3164E+01	
160	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^1I_6^o$	6.3290E+01	− 1.1300E−01	1.5900E−03	6.3178E+01	6.3236E+01	
161	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3P_0^o$	6.3361E+01	− 2.6600E−02	1.6100E−03	6.3336E+01	6.3294E+01	
162	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3G_4^o$	6.3383E+01	− 1.0100E−01	1.5900E−03	6.3284E+01	6.3343E+01	
163	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)\ ^1D_2^o$	6.3392E+01	− 7.0700E−02	1.4200E−03	6.3323E+01	6.3358E+01	
164	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3G_3^o$	6.3405E+01	− 9.6100E−02	1.4400E−03	6.3311E+01	6.3379E+01	
165	$3s^2 3p^6 3d^7\ (^2P_3)\ 4p^1\ (^2P_1)\ ^3S_1^o$	6.3420E+01	− 7.1100E−02	1.5200E−03	6.3350E+01	6.3389E+01	
166	$3s^2 3p^6 3d^7\ (^2H_3)\ 4p^1\ (^2P_1)\ ^3H_5^o$	6.3461E+01	− 1.0300E−01	1.5900E−03	6.3359E+01	6.3419E+01	
167	$3s^2 3p^6 3d^7\ (^2G_3)\ 4p^1\ (^2P_1)\ ^3F_2^o$	6.3644E+01	− 8.0700E−02	1.5100E−03	6.3564E+01	6.3419E+01	
168	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3F_3^o$	6.3652E+01	− 4.3500E−02	1.5500E−03	6.3610E+01	6.3616E+01	
169	$3s^1 3p^5\ (^2P_1)\ 3d^{10}\ (^1S_0)\ ^1P_1^o$	6.3680E+01	− 1.6900E−01	− 7.9500E−02	6.3431E+01	6.3649E+01	
170	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ ^3G_5^o$	6.3694E+01	− 9.6900E−02	1.5600E−03	6.3599E+01	6.3654E+01	
171	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^3F_4^o$	6.3699E+01	− 8.8200E−02	1.5700E−03	6.3612E+01	6.3660E+01	
172	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3D_1^o$	6.3721E+01	− 3.3800E−02	− 3.3300E−04	6.3687E+01	6.3726E+01	
173	$3s^2 3p^6 3d^7\ (^2D_3)\ 4p^1\ (^2P_1)\ ^3D_2^o$	6.3790E+01	− 4.6300E−02	1.5200E−03	6.3745E+01	6.3789E+01	
174	$3s^2 3p^6 3d^7\ (^2F_3)\ 4p^1\ (^2P_1)\ ^1F_3^o$	6.3852E+01	− 7.9900E−02	1.4800E−03	6.3773E+01	6.3830E+01	
175	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3P_1^o$	6.4688E+01	− 6.9200E−02	1.6200E−03	6.4621E+01	6.4655E+01	
176	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3D_2^o$	6.4761E+01	− 7.7600E−02	1.5900E−03	6.4685E+01	6.4727E+01	
177	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3F_4^o$	6.4780E+01	− 6.5800E−02	1.5900E−03	6.4716E+01	6.4740E+01	
178	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^1F_3^o$	6.4908E+01	− 9.1200E−02	1.5700E−03	6.4819E+01	6.4873E+01	
179	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3P_0^o$	6.4955E+01	− 8.9100E−02	1.5700E−03	6.4868E+01	6.4924E+01	
180	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^1D_2^o$	6.4959E+01	− 6.0600E−02	1.5600E−03	6.4900E+01	6.4931E+01	
181	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^1P_1^o$	6.5027E+01	− 7.9100E−02	1.5700E−03	6.4950E+01	6.5009E+01	
182	$3s^2 3p^6 3d^7\ (^2D_1)\ 4p^1\ (^2P_1)\ ^3D_3^o$	6.5115E+01	− 6.1700E−02	1.5800E−03	6.5055E+01	6.5090E+01	
183	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5F_4^e$	6.6656E+01	− 3.2500E−02	− 4.0600E−05	6.6624E+01	6.6653E+01	
184	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5D_3^e$	6.6717E+01	− 2.7900E−02	− 3.1700E−05	6.6689E+01	6.6708E+01	
185	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5F_5^e$	6.6718E+01	− 4.2900E−02	− 2.5700E−05	6.6675E+01	6.6720E+01	
186	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^3H_6^e$	6.6837E+01	− 4.6200E−02	1.7900E−09	6.6790E+01	6.6829E+01	
187	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5P_3^e$	6.7098E+01	− 4.1200E−02	− 3.9200E−05	6.7057E+01	6.7081E+01	
188	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^3G_5^e$	6.7130E+01	− 5.3100E−02	− 1.2300E−05	6.7077E+01	6.7114E+01	
189	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5H_7^e$	6.7170E+01	− 5.9500E−02	4.4600E−06	6.7111E+01	6.7145E+01	
190	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^5G_6^e$	6.7184E+01	− 5.4800E−02	− 1.6000E−06	6.7129E+01	6.7168E+01	
191	$3s^2 3p^6 3d^7\ (^2D_3)\ 4d^1\ (^2D_1)^3D_1^e$	6.7194E+01	− 5.0300E−03	− 9.5400E−05	6.7188E+01	6.7199E+01	
192	$3s^2 3p^6 3d^7\ (^2P_3)\ 4d^1\ (^2D_1)\ ^3D_2^e$	6.7277E+01	− 2.0100E−02	− 7.2100E−05	6.7257E+01	6.7274E+01	
193	$3s^2 3p^6 3d^7\ (^2P_3)\ 4d^1\ (^2D_1)^3F_3^e$	6.7365E+01	− 1.8500E−02	− 6.1100E−05	6.7347E+01	6.7372E+01	
194	$3s^2 3p^6 3d^7\ (^2P_3)\ 4d^1\ (^2D_1)\ ^3P_0^e$	6.7378E+01	− 7.2400E−03	− 1.0400E−04	6.7371E+01	6.7393E+01	
195	$3s^2 3p^6 3d^7\ (^4F_3)\ 4d^1\ (^2D_1)\ ^3P_2^e$	6.7392E+01	− 3.0700E−02	− 5.7100E−06	6.7361E+01	6.7396E+01	

Table 5 continued

Level No	Configurations	Energies (in Ryd.)					ΔE
		MCDF				FAC NIST[25]	
196	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5D_4^e$	6.7499E+01	− 4.1500E−02	− 1.9300E−05	6.7457E+01	6.7493E+01	
197	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5G_3^e$	6.7594E+01	− 2.2000E−02	− 2.8200E−05	6.7572E+01	6.7594E+01	
198	$3s^2 3p^6 3d^7$ ($^4P_3$) $4d^1$ ($^2D_1$) $^5D_1^e$	6.7643E+01	− 1.1900E−02	− 5.2000E−05	6.7631E+01	6.7648E+01	
199	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5F_2^e$	6.7662E+01	− 1.4700E−02	− 2.9800E−05	6.7647E+01	6.7675E+01	
200	$3s^2 3p^6 3d^7$ ($^4F_3$) $4d^1$ ($^2D_1$) $^5H_4^e$	6.7688E+01	− 3.5600E−02	− 8.8000E−06	6.7652E+01	6.7685E+01	

Table 6 Comparison of excitation energies (in Ryd.) of five different configurations sets for first 24 excited states levels for Fe-like Ba

Level no	MCDF1	MCDF2	MCDF3	MCDF4	MCDF5	ΔE_1	ΔE_2	ΔE_3	ΔE_4
2	0.483	0.483	0.457	0.483	0.457	0.000	− 0.026	0.000	− 0.026
3	1.124	1.124	1.124	1.124	1.124	0.000	0.000	0.000	0.000
4	1.362	1.362	1.293	1.362	1.289	0.000	− 0.068	0.000	− 0.073
5	1.584	1.584	1.567	1.584	1.567	0.000	− 0.017	0.000	− 0.017
6	1.865	1.865	1.837	1.865	1.837	0.000	− 0.028	0.000	− 0.028
7	1.939	1.939	1.939	1.939	1.939	0.000	0.000	0.000	0.000
8	2.627	2.627	2.610	2.627	2.610	0.000	− 0.017	0.000	− 0.017
9	4.047	4.047	3.928	4.047	3.916	0.000	− 0.120	0.000	− 0.131
10	19.224	19.224	19.224	19.224	19.224	0.000	0.000	0.000	0.000
11	19.658	19.652	19.658	19.652	19.652	− 0.006	0.000	− 0.006	− 0.006
12	20.339	20.339	20.339	20.339	20.338	0.000	0.000	0.000	0.000
13	21.023	20.959	21.023	20.959	20.959	− 0.064	0.000	− 0.064	− 0.064
14	21.182	21.177	21.182	21.177	21.177	− 0.005	0.000	− 0.005	− 0.005
15	21.495	21.472	21.495	21.472	21.472	− 0.024	0.000	− 0.024	− 0.024
16	21.663	21.663	21.663	21.663	21.663	0.000	0.000	0.000	0.000
17	21.752	21.703	21.752	21.703	21.703	− 0.049	0.000	− 0.049	− 0.049
18	25.931	25.917	25.931	25.917	25.917	− 0.014	0.000	− 0.014	− 0.014
19	26.282	26.282	26.282	26.282	26.281	− 0.001	0.000	− 0.001	− 0.001
20	26.551	26.551	26.551	26.551	26.551	0.000	0.000	0.000	0.000
21	27.986	27.895	27.986	27.895	27.895	− 0.091	0.000	− 0.091	− 0.091
22	36.796	36.796	36.796	36.796	36.796	0.000	0.000	0.000	0.000
23	37.234	37.234	37.062	37.234	37.062	0.000	− 0.172	0.000	− 0.172
24	38.024	38.024	38.024	38.024	38.024	0.000	0.000	0.000	0.000
25	38.716	38.716	38.161	38.716	38.160	0.000	− 0.556	0.000	− 0.556

We have preferred only Babushkin gauge due to the effect of negative energy states is least on Babushkin gauge as compared to coulomb gauge. We have predicted all transitions to ground state in Fe-like ions lie in soft X-ray (SXR) region 1–50 Å and hence can be useful in soft X-ray lasers. We have also compared our calculated transition wavelengths with wavelengths compiled by NIST for Fe-like Xe and Ba as given in Table 7, and maximum difference is 0.908Å for 5–12 transition and 0.825Å for 5–17 transition in Fe-like Xe and Ba, respectively. This difference in transition wavelengths with NIST is due to difference in our excitation energies of levels with NIST, and reason behind is that the results available in NIST are obtained by compilation of available theoretical and experimental results.

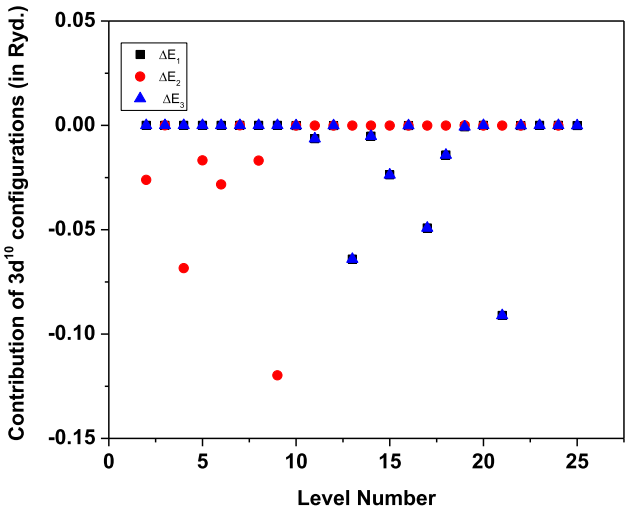


Fig. 3 Contribution of $3d^{10}$configurations in lowest 24 excited states of Fe-like Ba

6. Conclusions

Better understanding of two different codes and methods on the basis of computational time, cost of computation and accuracy of results with minimum discrepancy for same set of configurations is prerequisite. Therefore, in the present work, comparison of lowest 200 excited state energies from FAC and GRASP codes has been studied for Fe-like ions. Matching of results and subsequent analysis of discrepancies both from GRASP and FAC codes ensures that our results are trustworthy and the deviations are reasoned for adequately. We have also predicted that all transitions to ground state lie in SXR region and also provided transition data within lowest 200 levels. We understand that our reported data will be useful in fusion plasma as well as in cell biology and pave way for future comparison of experimental results of higher excited states.

The large amount of atomic data reported in this paper is examination and detailed study of reliability, authenticity and completeness atomic parameters which could be useful in the modeling and characterization of hot dense burning plasmas. Further, the reported big data and details may be useful in the tracing of SXR spectral lines from spectra of fusion and astrophysical devices for analysis, microscopic imaging and other applications of these spectral lines.

Table 7 Comparison of transition wavelengths for Fe-like Xe and Ba

Fe-like Xe					Fe-like Ba				
Transitions		λ_{MCDF} (in Å)	λ_{NIST} (in Å)	$\Delta\lambda$ (in Å)	Transitions		λ_{MCDF} (in Å)	λ_{NIST} (in Å)	$\Delta\lambda$ (in Å)
Lower level	Upper level				Lower level	Upper level			
1	10	50.4240	51.085	0.661	1	10	47.4030	47.990	0.587
1	12	47.7230	48.525	0.802	1	12	44.8050	45.523	0.718
1	16	45.0070	45.730	0.723	1	16	42.0660	42.684	0.618
2	11	50.5300	51.285	0.755	2	11	47.5240	48.147	0.623
2	13	47.3330	48.115	0.782	2	12	45.8960	46.574	0.678
2	17	45.8430	46.615	0.772	2	13	44.3670	45.137	0.77
3	10	53.1360	53.885	0.749	3	14	45.4320	46.173	0.741
3	14	48.2950	49.125	0.83	3	16	44.3680	45.083	0.715
3	16	47.1550	47.970	0.815	3	18	36.7340	37.258	0.524
3	18	39.8830	40.335	0.452	3	20	35.8390	36.368	0.529
3	20	38.7820	39.410	0.628	5	14	46.4970	47.204	0.707
5	12	51.3120	52.220	0.908	5	17	45.1830	46.008	0.825
5	17	48.0230	48.875	0.852	5	18	37.4270	37.928	0.501
6	18	41.0940	41.540	0.446	5	20	36.4980	36.998	0.5
7	16	49.1420	49.940	0.798	6	17	45.8230	46.574	0.751
7	20	40.1160	40.731	0.615	6	18	37.8650	38.324	0.459
8	19	41.8530	42.310	0.457	7	16	46.2020	46.904	0.702

Acknowledgements We are grateful to DTU and Shyamlal College, DU, for providing infrastructure and support for other facilities.

References

[1] M Shimada et al *Nucl. Fusion* **47** S1 (2007)
[2] T A Carlson, C W Nestor, J R N Wesserman and J D McDowell *ADNDT* **2** 63 (1970)
[3] J F Wyart, C Bauche-Arnoult and E Luc-Koenig *Phys. Scr.* **32** 103 (1985)
[4] J O Ekberg and U Feldman *J. Opt. Soc. Am. B* **5** 1275 (1988)
[5] R Doron, M Fraenkel, P Mandelbaum, A Zigler and J L Schwob *Phys. Scr.* **58** 19 (1998)
[6] R Doron, E Behar, P Mandelbaum, J L Schwob, H Fiedorowicz, A Bartnik, R Jarocki, M Szczurek and T Wilhein *Phys. Rev. A* **59** 188 (1999)
[7] V Nagels, C Chenais-Popovics, V Malka, J-C Gauthier, A Bachelier and J-F Wyart *Phys. Rev. A* **68** 233 (2003)
[8] G C Rodrigues, P Indelicato, J P Santos, P Patte and F Parente *ADNDT* **86** 117 (2004)
[9] X B Ding, C Z Dong, F Koike and T Kato *Phys. Lett. A* **371** 124 (2007)
[10] D Osin, J Reader, J D Gillaspy and Y Ralchenko *J. Phys. B* **45** 21 (2012)
[11] M Klapisch, J L Schwob, B S Fraenkel and J Oreg *J. Opt. Soc. Am.* **67** 148 (1977)
[12] R D Cowan, *The Theory of Atomic Structure and Spectra* (*University of California Press*, 1981)
[13] A Bar-Shalom, M Klapisch and J Oreg *J. Quant. Spectros. Radiative Transf.* **71** 169 (2001)
[14] K G Dyall, I P Grant, C Johnson, F A Parpia and E Plummer *Comput. Phys. Commun.* **55** 425 (1989)
[15] P Jönsson, X He, C Froese Fischer and I P Grant *Comput. Phys. Commun.* **177** 597 (2007)
[16] P Jönsson, G Gaigalas, J Bieron, C Froese Fischer and I P Grant *Comput. Phys. Commun.* **184** 2197 (2013)
[17] O Gorceix, P Indelicato and J-P Desclaux *J. Phys. B Atomic Mol. Phys.* **20** 639 (1987)
[18] P Indelicato, O Gorceix and J P Desclaux *J. Phys.B Atomic Mol. Phys.* **20** 651 (1987)
[19] M S Safronova, W R Johnson and U I Safronova *Phys. Rev. A* **53** 4036 (1996)
[20] M F Gu *Can. J. Phys.* **86** 675 (2008)
[21] M G Kozlov, S G Porsev, M S Safronova and I I Tupitsyn *Comput. Phys. Commun.* **195** 199 (2015)
[22] T Saue, R Bast, A S P Gomes, H J Aa Jensen, L Visscher, I A Aucar, et al., *J. Chem. Phys.* **152** 204104 (2020)
[23] E B Saloman *J. Phys. Chem. Ref. Data* **33** 765 (2004)
[24] J E Sansonetti *J. Phys. Chem. Ref. Data* **38** 761 (2009)
[25] J E Sansonetti and J J Curry, *J. Phys. Chem. Ref. Data* **39** 043103 (2010)

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

Study on parameter's optimization for the minimizing of the residual stress in powder mixed near-dry electric discharge machining

S. Sundriyal[1], Vipin[2], R.S. Walia[3*]

[1]Mechanical Engineering, Delhi Technological University, Delhi, 110089, India.
[2] Mechanical Engineering, Delhi Technological University, Delhi, 110089, India.
[3] Production and Industrial Engineering, Punjab Engineering College, Punjab, 160012, India
*Corresponding author/ *E-mail address*: - waliaravinder@yahoo.com (+91-9717325233)

Running Title: Experimental studies on Near Dry Electric Discharge Machining

Abstract There are some undesired effects left in the machined products by electric discharge machining (EDM) and the most prominent among those is the residual stress. These induced residual stresses influence fatigue behavior, dimensional stability, and stress corrosion which is responsible for the failure of the machined product. It was observed that the products obtained by powder mixed near dry- EDM or (PMND-EDM) have a considerably low value of residual stress (RS). The objective of this research is to minimize the residual stresses induced in the machined workpiece (EN-31) by optimizing the process parameters which affects the machining characteristics significantly. Selected parameters were tool diameter, mist flow rate, metallic powder concentration, and mist pressure while the workpiece selected was EN-31 due to its desirable mechanical properties and immense applications in the manufacturing industry. The minimum value of residual stress at optimized values of process parameters was found to be 106.32 MPa.

Keywords: Residual stress**,** EDM**,** Powder, Near dry, X-ray method.

Nomenclature:

EDM: Electric discharge machining

DEDM: Dry Electric discharge machining

NDEDM : Near dry Electric discharge machining

PM-EDM : Powder mixed Electric discharge machining

PMND-EDM : Powder mixed near dry Electric discharge machining

MRR: Material removal rate

TWR: Tool wear rate

Ra: Surface finish

RS: Residual stress

FWHM: Full width half maximum

ANOVA: Analysis of variance

CI_{ce} *: Confidence interval of confirmation experiments:*

CI_{pop}*: Confidence interval of population*

V_e*: Error variance*

MH: Micro-Hardness

HV: Vickers Hardness number

1 Introduction

Electric discharge machining (EDM) was developed by Russian scientist Lazarenko in 1943. EDM is one of the non-conventional machining methods which utilize thermal energy for material removal from the workpiece. EDM nowadays is widely used in global manufacturing engineering and this technology made it possible to cut intricate shapes in very hard materials. Further study of EDM by the researchers has led to the development of various hybrid EDM machining methods such as Dry EDM (DEDM), near dry EDM (NDEDM), powder mixed EDM (PM-EDM) and powder mixed near dry EDM (PMND-EDM). These developed methods were more prominent in terms of machining efficiency. NDEDM utilizes the mist of gas and minute amount of dielectric oil as a dielectric medium, therefore this machining method also eliminates the bulk amount of EDM oil required unlike traditional EDM [1]. The feasibility of the new EDM process which was known as PMND-EDM came into existence [2]. Gases such as nitrogen or air as a dielectric medium instead of dielectric oil were used in Dry EDM which makes processing eco-friendly as this new methodology produces less harmful fumes and at the same time also eliminates the need for EDM oil [3].

Previously research has been done in powder mixed EDM (PM-EDM) for performance enhancements such as improved material removal rate (MRR), tool wear rate (TWR), surface finish (Ra), and morphology. A study on machining efficiency was performed for powder mixed EDM and Taguchi L_9

(OA) was utilized for optimization. It was illustrated that PMND-EDM leads to reduced tool wear rate and residual stress due to stable arc conditions at the inter-electrode gap [4]. Rotary tool electrode was used as in Near dry EDM and there was enhanced MRR with improved surface finish with reduced micro-cracks and debris accumulation [5]. Similary, oxygen gas was utilized as dielectric medium in Near dry EDM and it was reported that more MRR was observed with the increased phenomenon of melting by rapid oxidation [6]. The experimental result and analysis revealed that PMND-EDM was better machining method than ND-EDM as in the former technique, the M RR increased by 45.04 % while surface and TWR reduced by 45.33 % and 60.60% respectively [7]. It was also noticed that PMND-EDM resulted in producing machined surface with better surface quality due to presence of metallic foreign particles in the dielectric medium which results in producing stable arcing at the machining gap [8].

The study has also been performed in the field of surface morphology such as improved microhardness in PMEDM. The presence of hard tungsten carbide (W_2C) and cementite (Fe_3C) increased the MH (~150%) on the machined surface [9]. Generation of the carbides on the machined surface, increased the micro-hardness as high as 912 HV, without sacrificing the quality of the machined surface [10]. Taguchi's philosophy was adapted to determine the suitable combination of EDM parameters to improve machining performances in terms of material removal and micro-hardness [11].

The manufactured products by different EDM machining methods have several undesirable effects on components and residual stress generated on the work surface is a major one. Metallurgical transformations due to sharp temperature gradients are responsible for the generation of residual stress in a mechanical component [12]. The intensity of this residual stress was found to be more than the yield point of the material which causes severe cleavage, twinning, and slip in the crystal structure [13]. The nature of residual stress was investigated and it was revealed that high residual stresses were found at the immediate surface of the machined sample and then gradually falls to a low value before giving a small way to compressive residual stresses [14]. There was a decrease in residual stress by 56.09 % over the machined surfaces in PMND-EDM as compared to the ND-EDM process [15]. A layer of hard zinc carbide hard was deposited over the surface of the machined sample which leads to a higher value of micro-hardness [16]. A peak value of residual stress in EDM was investigated on the machined surface of the workpiece and it was concluded that residual stress with a peak value of 590 MPa was found at 40μm beneath the top surface of the workpiece and was reduced to zero value at 200μm depth [17]. EDM results in the generation of residual stresses of high value which increase from the surface and reach the

maximum value. This maximum value is around the ultimate tensile strength of the material [18]. Optimization of the residual stress using X-ray cos α method in vibration- assisted hybrid EDM process was performed over high carbon high chromium D2 tool steel workpiece. It was revealed that discontinuous vibrations enhanced the material removal rate and reduced the residual stresses induced in the machined workpiece [19]. Finite element method was used for modeling of residual stresses in the workpiece machined by EDM. It was revealed by the study that the value of residual stresses at the subsurface of the workpiece was more than the top surface due to high surface roughness over the top surface and low discharge energy of the sparks was attributed to lower values of the residual stresses [20]. Finite element analysis of residual stresses in EDM was performed and techniques like nano-indentation technique and Raman spectroscopy were adopted to measure the residual stress in the AISI H13 tool steel machined by EDM process and it was revealed that microcracks which were responsible for relieving some part of residual stress [21]. A parametric study for residual stresses was performed in wire EDM and the Taguchi method of optimization was utilized to minimize the residual stresses in the aluminum workpiece [22].

There are many factors that are responsible for the residual stress such as heating, cooling, re-solidification, melting, and phase transformations with volume change. This induced residual stress results in a decrease in reliability of the product and at the same time leads to uncertain failure of the manufactured products. Therefore it was necessary to study the residual stress so that these stresses can be minimized and product reliability can be increased. These residual stresses (RS) are generated in the specimens due to a heating phenomenon and re-solidification over the surface of the workpiece. Thus these residual stress plays a vital role in determining product quality and its reliability. Although studies on performance enhancements and optimization techniques have been applied in PM-EDM but still investigations related to residual stress and optimization of critical parameters in PMND-EDM are still lacking therefore in this study contributions and efforts have been made to explore the important mechanical property (RS) of the machined products. This research deals with further in-depth investigation in the field of residual stress optimization and continuation of research conducted in the field of powder mixed near-dry EDM [16].

2 Experimentation

2.1 Experimental procedure

The process parameters for experimentation were classified into six categories which were dielectric, metallic powder, tool, electrical parameters, workpiece, and other parameters (Air pressure, oil pressure, and mist flow rate).

Series of experiments were performed as per Taguchi L_9 (OA) for residual stress (RS). Each series of experiments was run for three trials and the results obtained are shown in Table 1.

The experimentation methodology was derived from previous work in the field of EDM [15-16]. Each series of experiments was performed in three runs and a total of 27 experiments each were performed for analysis of RS by considering S/N (signal to noise ratio) of lower the better principle given in Equation 1, [16]:

$$S/N = -log_{10}\left(\frac{1}{n}\sum_{i=1}^{n} y_{ij}^2\right), \quad (1)$$

Where:

n = Replications;

y_{ij}= Response value.

The standard deviation error bar was plotted between the mean value of the obtained residual stress for different parametric experimental conditions (total twenty-seven experiments :three repeatabilities for each set of parametric condition) and the error was within +/-5% range as shown in Fig.1.

Measuring techniques of residual stress can be classified into different categories such as destructive testing, semi-destructive testing, and non-destructive testing (NDT). Non-destructive testing has several advantages over semi-destructive and destructive testing. Non-destructive testing can be applied to a wide variety of materials and the accuracy of measurement is high. There are various NDT methods of measuring residual stress such as the X-ray diffraction technique, Berkhausen noise method, neutron diffraction method, and the ultrasonic method. X-ray diffraction method has several advantages over other methods because the X-ray diffraction technique can be applied for measurements of residual stress over a wide variety of materials. Other advantage of the X-ray diffraction technique is that micro and macro residual stress can be analyzed and at the same time this technique is handheld which makes it possible to take the reading at any orientations.

The residual stress measurements were taken by μX-360 pulstec machine (Japanese built) as shown in Fig.2. It consists of mainly three parts which are computer, sensor unit and power unit. The setup consists of the X- ray tube (30 KV) and 2-D X- ray sensor for visual analysis along with a power supply unit to generate the X-ray.

The X-ray takes 90 seconds to take one reading of residual stress. The machine has a special feature of air cooling for efficient performance. The X- ray tube current was 1mA and the sample distance from the X-ray focusing lens was 38mm. The incident angle for the X- ray was set at 35 degrees while the information regarding specimen to be analyzed were lattice constant, interplanar spacing and diffraction planes (h, k, l), crystal structure, Poisson's ratio, and Young's modulus. The output responses measured were residual stress, shear stress, and full width half maximum value. The full width at half maximum (FWHM) is an expression of the extent of a function (profile peak diffraction) whereas maximum and minimum values of peak diffraction were given as α-max and α-min respectively. The procedure of residual stress measurement is divided into four stages. The first stage includes sample positioning and capture of the sample image. The second stage includes the incidence of the X- ray and its detection by a 2-D sensor. The third stage includes the calculation of residual stress and the final stage includes output data results of residual stress and value of full width half maximum (FWHM). The output reading also calculates the standard deviation in the reading obtained to maintain the repeatability of reading. The working principle of residual stress is based on X-ray diffraction scattering as per Bragg's law given by Equation 2, [23]:

$$n\lambda = 2dsin\theta \quad (2)$$

Where; n= is the order of reflection;

d = interplanar spacing of the crystal;

λ = wavelength of the incident X-ray;

and θ is the angle of incidence that equals the angle of scattering (diffraction angle) as shown in Fig. 3. X-ray diffraction occurs from various orientations of the grain crystals that satisfy Bragg's law as shown in Fig. 4. The diffracted X-ray form a cone around the incident X-ray axis because of the variation in crystal's orientation which also results in the formation of a ring of residual stress known as Debye-Scherrer ring as shown in Fig. 5. This Debye ring is formed by the single short-duration X-ray exposure.

Determination of the residual stress achieved by accurately measuring the position of the Debye- Scherrer rings and their positions are a direct measure of strain.

The magnitude of the strain is determined from the detected position of the debye-Scherrer ring was calculated by using the formula given in Eqs. (3) - (4) while residual stress and shear stress are calculated by Equations 5 - 6.

$$\epsilon_{\alpha 1} = \frac{1}{2}\{(\epsilon_{\alpha} - \epsilon_{\pi+\alpha}) + (\epsilon_{-\alpha} - \epsilon_{\pi-\alpha})\} \quad (3)$$

$$\epsilon_{\alpha 2} = \frac{1}{2}\{(\epsilon_{\alpha} - \epsilon_{\pi+\alpha}) - (\epsilon_{-\alpha} - \epsilon_{\pi-\alpha})\} \quad (4)$$

$$\sigma_x = -\frac{E}{1+\vartheta} * \frac{1}{\sin 2\eta} * \frac{1}{sin2\varphi^o} * \frac{\partial \varepsilon\alpha 1}{\partial\, cos\alpha} \quad (5)$$

$$\tau_{xy} = -\frac{E}{2\,(1+\vartheta)} * \frac{1}{\sin 2\eta} * \frac{1}{sin\varphi^o} * \frac{\partial \varepsilon\alpha 2}{\partial\, sin\alpha} \quad (6)$$

Where α = Azimuth angle of Debye-Sherrer ring; ε α_1 = Strain measured in the vertical direction; ε $_{\pi+\alpha}$ = Strain measured in the direction of π+α; σ_x = Residual stress; τ_{xy} = Residual shear stress; ν = Poisson's ratio, ε α_2 = Strain measured in the horizontal direction, E is Young's modulus; η = Angle between Debye –Sherrer ring axis and the sample diffraction detector X-ray, υ = Poisson's ratio.

3. Method and Developed Setup.

The schematic diagram for the working principle of PMND-EDM is shown in Fig. 6. A newly designed setup of PMND-EDM was made indigenously at precision engineering lab (DTU, Delhi) as shown in Fig. 7 [4]. The developed setup for PM-NDEDM included various technical features and parts which were designed and fabricated according to the required experimental conditions to achieve the desired output results. The setup body was made of stainless steel which comprised of EDM oil pressure regulator, air pressure regulator along with a mist flow rate controller.

The pressure gauges were analogous in nature with limit up to to 0.6 MPa. The mist flow regulator was mounted on the setup along with transparent scale with marking up to 20 ml min^{-1}. Manually controlled regulators were used to set up the desired mist flow rate and its pressure. A separate air compressor (Nu-Air built) of 2 H.P (horsepower), pressure limit of 0.8 MPa and 24-litre capacity was integrated with the setup. A mixing chamber was provided with the setup which was utilized to obtain 3-Ø (phase) mixture

of metallic powder, glycerol (a stabilizing agent) and EDM oil along with compressed air supplied separately from the air compressor (glycerol 5% of dielectric oil, metallic powder 2% dielectric oil). The dielectric mist generated was supplied at the EDM machining gap between the tool electrode and the workpiece with the help of flexible plastic tubes. The experiments were performed on Sparkonix EDM machine (Sparkonix limited Pune, 35 A series). Customized copper tool electrodes of different dimensions with special design features were fabricated and developed at the precision engineering lab (DTU, Delhi) as shown in Fig. 8. The concept of design of tool electrodes was taken from a previous study in the field of near dry EDM [24]. The detailed design features of the hollow tool electrode is shown in Fig. 9. A flexible tube of 6 mm inner diameter was inserted at the entry end of a tool for delivering a pressurized mist of dielectric medium. The different tool electrodes dimensions for experimentation are shown in Table 2. The experimental conditions set for PMND-EDM are shown in Table 3, while the chemical and physical properties of the tool electrode are shown in Table 4.

The workpiece chosen for machining was EN-31 grade due to its desirable physical properties such as high strength and hardness**.** The chemical composition and physical properties of EN-31 workpiece is shown in Table 5.

4 Working Principle of PMND-EDM

Electric Discharge Machining works on the basic principle of conversion of electric energy into thermal energy. The plasma channel of high heat energy was generated at a working gap between tool and workpiece or inter electrode-gap when a current was supplied to electrodes. The ionization effect takes place due to the interaction of ions and molecules at this gap which results in generating a spark that jumps from tool to workpiece in case of straight polarity and vice- versa. This repetitive cycle generates thousand of sparks which results in the formation of a plasma of very high temperature. This plasma channel of high heat energy is responsible for the melting and removal of material from the workpiece due to which erosion process starts. The flushing of eroded debris was done with the help of dielectric oil supplied at the working gap. The dielectric medium in PMND-EDM consists of metallic powder, compressed air and EDM oil in very minute amount along with glycerol which works as a stabilizing agent. The metallic powder works as an additive and helps in creating a chain-like structure of powder

grain particles in the plasma channel which reduces the insulating strength of dielectric medium and helps in rapid erosion. The compressed air supplied from the compressor helps in the formation of dielectric mist and also helps in stabilizing arc at inter-electrode gap due to the increased pressure of supplied mist. The glycerol liquid which has high viscosity was added to prevent the powder particles from settling down. This mixture is fed into the mixing chamber from the inlet of the tank and this heterogeneous mist is supplied at IEG (Inter electrode gap or gap between the tool and the workpiece) 25µm at high pressure with the help of tubes from the mixing chamber. The current is supplied to the electrodes as soon as the experimental conditions are obtained. The sparking phenomenon starts and erosion cycle takes place. The increased working gap due to the presence of metallic powder results in obtaining more uniformly energized plasma. The reason for an increase in working gap and energized plasma is attributed to the chain-like formation of powder particles due to interlocking of grains of powder particles at IEG which results in rapid sparking with high frequency.

The 3- phase dielectric was supplied at the final stage at the machining area through hollow copper electrode tool. Schematic diagram of the setup for PMND-EDM is shown in Fig. 10. The compressed air is suplied to the mixing chamber. The compressed air is mixed with dielectric EDM oil along with metallic powder additives. Afterwards, with pressure control regulators dielectric mixture is supplied to the hollow tool electrode through the plastic tube. The dielctric gets ejected out from the other end of the electrode and the initiation of spark generation takes place upon supply of power to the tool electrodes. This methodology of injecting heterogeneous mist at inter-electrode gap or working gap was efficient in terms of machining and at the same time also led to the reduced requirement of EDM oil. The consumption rate of EDM oil in this methodology was very minute unlike the traditional EDM method where the requirement of EDM oil was in bulk quantity. Therefore there was a requirement of a new setup with customized specifications as per the specified design which could fulfill the research objective and thus expected results can be obtained.

5 Results and Discussion.

5.1 Analysis of residual stress

The workpiece material (EN-31) chosen for residual stress analysis was B.C.C (Body centered cubic) crystal structure with interplanar spacing (d) of 1.170 $\dot{A}$, Poisson's ratio of 0.280, diffraction plane (h, k, l) of 2, 1, 1 and with diffraction angle of the X-ray at 156.39 degrees. Figs. 10(a-i) shows the residual stress analysis performed by the X-ray diffraction machine for nine series of experiments as per Taguchi

L_9 (OA). Fig. 11 shows display diffraction ring, residual stress, peak profile and FWHM along with their values for nine sets of experiments as per Taguchi L_9 (OA). The minimum value of residual stress was found to be 145 MPa among experimental values obtained from nine series of experiments. This obtained value of RS was much lower as compared to the high RS achieved in previous work [17]. Residual stress decreased due to induced microcracks over the machined surface. These induced microcracks as shown in Fig.12, were responsible for relieving some of the residual stress, another reason can be attributed to the uniform distribution of heat over the machined surface and improved flushing condition which ultimately relieves some part of the residual stresses over the machined surfaces [22].

The plasma at the machining gap is partially energized by the metallic powder concentration, which gives a surface with lesser induced micro-cracks as shown in Fig. 12. Improved ionization takes place by adding zinc powder in the dielectric fluid, as added powder enlarges and widens the discharge gap between the electrodes. This results in the easy removal of debris, which leads to an improvement in surface quality. It was revealed that by the addition of metallic powder, the discharging energy dispersion improves because the insulating strength of the dielectric fluid reduces with the addition of metallic powder. These favourable conditions improves the forming effect of the machined surface with improved surface characteristics in absence of large cracks [15]. This proves that a significant role was played by the metallic powder to modify the channel of the plasma. This resulted in improved de-ionization effect and discharging conditions. All these factors improve discharge frequency with improved flushing conditions. A solid and liquid mixture, such as mist assisted with metallic powder, changes the electric field strength of the dielectric medium and thus facilitates the discharge initiation. Enlarged discharge gap results in sufficient heat dissipation, due to which more molten materials were ejected by the explosive force by the gasification of solid and liquid phases. This phenomenon also restrain the excessive expansion of the discharge channel in its radius. Thus optimal parameters achieve a surface with reduced residual stress.

Table 6 shows the output results of residual stress analysis for nine sets of experiments where σ is residual stress, σ_1 is the standard deviation of residual stress, τ_{xy} is shear stress, σ_2 is the standard deviation of shear stress, FWHM is full width half maximum angle, α_{max} is maximum azimuth angle and α_{min} is minimum azimuth angle of debye scherrer ring.

Radial stress analysis over the machined surface was conducted to study the effect of variation of residual stresses with respect to the radial distance from the machined area. The radial residual stress analysis was performed at different radial distances from the center of the machined area.

It was observed that the nature of variation of stress was transforming from tensile to compressive value Similar trend of RS was also observed in research published in previous literature [12-14, 25]. There was a sudden reduction of temperature with respect to an increase in radial distance due to low heat flux in the workpieces. The phenomenon of phase transformation was negligible at a more radial distance from the center of the machined area due to which the grains were again restrained back to compressive nature.

The testing was performed at different sections over the machined samples with the help of residual stress measuring machined depicted in Fig.1. The thermal effect was reduced to a very low value with the further increase of radial distance and finally becomes negligible at a distance of 1.7 mm as shown in Fig.13. The average values and main effects values of residual stress and their S/N ratios were calculated at different levels which were L1, L2 and L3 as tabulated in Table 7 and Table 8.

The average values of residual stresses is shown in Fig. 14. The parameters A, B, C, and D at levels 2, 2, 1, 2 respectively were most effective in reducing RS as shown in Fig.14. The same trend for S/N ratios is shown in Fig. 14. The parameter A at 2^{nd} level, parameter B at 2^{nd} level, parameter C at 1^{st} level and parameter D at the 2^{nd} level were most effective in reducing residual stress.

The residual stress induced in the workpiece initially decreases with an increase in the internal diameter of tool-tip at 2^{nd} level (L_2). Better flushing takes place over the machined surface which provides a better cooling condition for dissipation of heat over the machined surface and therefore less heat was generated over the machined surface which results in reducing residual stress to the lowest value at L_2. At level 3 there was a phenomenon of rapid solidification. This rapid solidification was responsible for an increase in the value of RS [25]. Rapid solidification led to the solidification of sputtered melted liquid over the workpiece surface as shown in Fig. 15 (Scanning electron microscope image). It can be clearly seen from Fig. 15 that there is non-homogenous solidified material over the machined surface which was responsible for the change in a specific volume of the molten material and therefore the value of residual stresses was high.

The RS value initially decreases with an increase in flow rate values due to better heat dissipation at L_2 but with a further increase of flow rate, there was a phenomenon of stress corrosion due to which the RS value increases to its peak value at L_3. Extreme temperature differences are developed on the surface of the machined samples due to high thermal sparking at the inter-electrode gap in EDM. Therefore rapid cooling leads to high residual stresses in the EDM'ed surfaces [15].

The residual stress initially increases with increases in powder concentration but afterwards decreases. Metallic powder addition resulted in a decrease of tensile residual stresses due to spark gap enlargement by low discharge energy density. This addition of graphite powder to the dielectric reduces the tensile stresses which in return also increases the fatigue strength of the machined components [26]. Due to the presence of metallic powder particles the plasma channel becomes more energized and more heat was induced over the workpiece due to which high residual stress was found over the machined surface at L_2. The heat density or heat flux at the machining gap was increased many times by the increase of metallic powder concentration. The bridging effect between the metallic powder grains gets increased and maintains favorable conditions for sparking and thus the temperature over the machined area gets very high which was responsible for higher values of residual stress. Further increase of powder concentration leads to improper spark discharges due to which low heat energy gets imparted to the workpiece. Over clustering of metallic powder particles at the inter-electrode gap resulted in the formation of small heat plasma and therefore the heat intensity at the machining gap gets reduced. The lower value of temperature at the inter-electrode gap was responsible for low metallurgical transformation at the machined surface (Fig. 16 , Scanning electron microscope image) and this results in reducing residual stress at L_3.

RS value decreases with the increase of mist pressure but afterwards, it increases with further increase of pressure. The molecules density initially gets increased with increase of mist pressure which improved deionization effect. These conditions lead to better heat dissipation which reduces RS value at L_2. The relaxation time for residual stress gets decreased with an increase of mist pressure due to which there was a phenomenon of localized in-homogenous plastic deformation over the machined part as shown in Fig. 17 (Scanning electron microscope image). These conditions were responsible for the increase in value of residual stress at L_3. Sudden pressure change of mist at L_3 was very much significant in changing the grains structure of the surface layer of the machined part.

5.1.1 Residual Stress performance characteristics estimation

Response characteristic is determined by using Equation 7, [16] as:

$$RS = \bar{A}_2 + \bar{B}_2 + \bar{C}_1 + \bar{D}_2 - 3\,\overline{RS} \quad (7)$$

The confidence interval for confirmation experiments is given by Equation 8, [16]:

$$CI_{CE} = \sqrt{F_\alpha\,(1, f_e)V_e[\frac{1}{n_{eff}} + \frac{1}{R}]}\,, \quad (8)$$

The Confidence interval of population is given by Equation 9, [16]:

$$CI_{POP} = \sqrt{\frac{F_\alpha\,(1, f_e)V_e}{n_{eff}}}, \quad (9)$$

Where, F_α (1, fe) = the F ratio at the confidence level of (1 –α) against DOF 1.

$F_{0.05}$ (1,18) =3.5546 (Tabulated)

$$n_{eff} = \frac{N}{1 + [\text{DOF associated in the estimate of mean response}]} = 9$$

Treatment = 9, repetition = 3; N (total number of experiments) = 27; V_e(error variance) = 1816.33 (as per Table 8), f_e (error DOF) = 18 (as per Table 9), R (confirmation experiments for sample size) = 3

where;

$$F = \frac{Variation\ between\ the\ samples}{Variation\ within\ the\ samples}$$

This *F* value is compared with the *F* limit for respective DOF. If *F* value is equal to or greater than the *F*-limit value (refer table 8), it can be said that significant differences exist between the sample means.

So, RS=106.32 MPa, CI_{CE}= ± 53.56, CI_{POP}= ± 26.78

The predicted optimal range of CI_{CE} is:

Mean RS − CI_{CE}< RS (MPa) < mean RS + CI_{CE}

i.e 52.76 MPa < RS (MPa) < 159.88 MPa

Mean RS − CI_{POP} < RS (MPa) < mean RS + CI_{POP}

i.e 79.54 MPa < RS (MPa) < 133.10 MPa

5.1.2 Confirmation experiments for residual stress.

Three repetitions of confirmation experiments for RS were performed at optimized levels of process parameters. The confirmation test for residual stress was performed at the experimental condition of A_2, B_2, C_1 and D_2. The experimental residual stresses for three runs at optimized conditions were found to be 90.32 MPa, 110.39 MPa and 118.49 MPa respectively. The mean experimental residual stress calculated at the optimized process of parameters was 106.40 MPa which was within the confidence interval of predicted RS. Thus, it is proven by the range that the predicted value of RS lies in the optimum range as shown in Table 9.

6 Conclusion

PMND-EDM process is a hybrid method of machining which was efficient in machining rate with improved quality characteristic such as reduced residual stress in the machined product. Optimization technique was utilized in PMND-EDM to study the effect on residual. Confirmation experiments were performed at the optimized experimental condition. The following conclusions can be drawn for the developed setup:

1. The RS was found to be least when experimentation was performed at optimized values of process parameters. Tool diameter at level 2 (3mm), mist flow rate at level 2 (10 ml min^{-1}), metallic powder concentration at level 1 (2g l^{-1}) and dielectric mist pressure at level 2 (0.5 MPa) were the most influential values which resulted in lowest RS value of 106.32 MPa in the machined product.
2. The predicted optimal range of residual stress was 52.76 MPa < RS (MPa) < 159.88 MPa. Predicted mean of residual stress at 95% confidence interval was 79.54 MPa < RS (MPa) < 133.10 MPa. This validates that the mean residual stress were within the predicted optimal range of residual stress.
3. Residual stress analysis was performed at different radial distance from the machined area. It was observed that the residual stresses were tensile in nature at the machined area but the nature of residual stress became compressive as the radial distance was increasing. The residual stress becomes negligible at radial distance of 1.7 mm.

Refrences

[1] Kao C.C., Tao J., and Shih A .J. “Near dry electrical discharge machining”, Int. J. Mach. Tool. Manuf. **47**, pp. 2273–2281 (2007).

[2] Gao Q., Zhang Q.H., and Zhang J.H. “Experimental study of powder-mixed near dry electrical discharge machining”, Chin. J. Mech. Engg. **45**, pp. 169–175 (2009).

[3] Shen Y., Liu H.Y., and Zhang Y. “High-speed dry electrical discharge machining”, Int. J. Mach. Tool. Manuf. **93**, pp. 19–25 (2015).

[4] Sundriyal S., Walia R.S., and Vipin. “Powder Mixed Near Dry Electric Discharge Machining Parameter Optimization for Tool Wear Rate”, Advances in Unconventional Machining and Composites. Lecture Notes on Multidisciplinary Industrial Engineering. Springer, Singapore (2020).

[5] Yadav, V.K., Kumar P., and Dvivedi A. “Performance enhancement of rotary tool near-dry EDM of HSS by supplying oxygen gas in the dielectric medium”, Material. Manuf. Process. (2019)

[6] Yadav V.K., Kumar P., and Dvivedi A. “Effect of tool rotation in near-dry EDM process on machining characteristics of HSS”, Material. Manuf. Process. (2019).

[7] Sundriyal S., Walia R.S., and Vipin. “Near Dry and Powder Mixed Near Dry Electric Discharge Machining”, Int. J. Engg. Adv. Tech. **9**, pp. 2021-2025 (2019)

[8] Sundriyal S., Walia R.S., Vipin., and Tyagi M. “Investigation on surface finish in powder mixed near dry electric discharge machining method”, Materials Today: Proceedings. **25**, pp. 808-809 (2019).

[9] Gill A.S., and Kumar S. “Investigation of Micro-Hardness in Electrical Discharge Alloying of En31 Tool Steel with Cu–W Powder Metallurgy Electrode“, Arab. J. Sci. Engg. **43**, pp. 1499-1510 (2018).

[10] Chundru V.R., Koona R., and Pujari S.R. “Surface Modification of Ti6Al4V Alloy Using EDMed Electrode Made with Nano- and Micron-Sized TiC/Cu Powder Particles”, Arab. J. Sci. Engg. **44**, pp. 1425–1436 (2019).

[11] Rahul., Datta S., and Biswal B.B. “A Novel Satisfaction Function and Distance-Based Approach for Machining Performance Optimization During Electro-Discharge Machining on Super Alloy Inconel 718”, Arab. J. Sci. Engg. **42**, pp. 1999–2020 (2017).

[12] Lloyd H.K., and Warren R.H. "Metallurgy of Spark-Machined Surfaces", J. Iron. Steel Inst. **203**, pp. 238–247 (1965).

[13] Rebelo J.C., Dias A.M., and Kremer D. "Influence of Pulse Energy on the Surface Integrity of Martensitic Steels", J. Mat. Process. Technol. **84**, pp. 90–96 (1998).

[14] Crookall J.R., and Khor B.C. "Residual Stresses and Surface Effects in Electro Discharge Machining", Proceedings 13th MTDR Conf, Birmingham, PP. 16-27 (1972).

[15] Sundriyal S., Vipin., and Walia R.S. "Study on the Influence of Metallic Powder in Near-Dry Electric Discharge Machining", Strojniški vestnik – J. Mech. Engg. **66**, pp. 243-253 (2020).

[16] Sundriyal S., Vipin., and Walia R.S. "Experimental Investigation of the Micro-hardness of EN-31 Die Steel in a Powder-Mixed Near-Dry Electric Discharge Machining Method", Strojniški vestnik - J. Mech. Engg. **66**: 184-192 (2020).

[17] Ekmekci B., Elkoca O., and Tekkaya A.E. "Residual Stress State and Hardness Depth in Electric Discharge Machining: De-Ionized Water as Dielectric Liquid", Mach. Sci. Technology. **9**, pp. 39-61 (2007).

[18] Ekmekci B., Tekkaya A.E., and Erden A.D. "A semi-empirical approach for residual stresses in electric discharge machining (EDM) ", Int. J. Mach. Tool. Manuf. **46**, pp. 858–868 (2006).

[19] Kumar S., Grover S., and Walia R.S. "Effect of hybrid wire EDM conditions on generation of residual stresses in machining of HCHCr D2 tool steel under ultrasonic vibration", Int. J. Interact. Des. Manuf. **12**, pp. 1119-1137 (2018)

[20] Liu J.F., and Guo Y.B. "Residual Stress Modeling in Electric Discharge Machining (EDM) by Incorporating Massive Random Discharges", Procedia CIRP. **45**, pp. 299 – 302 (2016).

[21] Shabgard M., Seydi S., and Seyedzavvar M. "Novel approach towards finite element analysis of residual stresses in electrical discharge machining process", Int. J. Adv. Manuf. Technol. **82**, pp.1805–1814 (2016).

[22] Rao P.S., Ramji K., and Satyanarayana B. "Effect of wire EDM conditions on generation of residual stresses in machining of aluminum 2014 T6 alloy", Alex. Engg. J. **55**, pp. 1077–1084 (2016).

[23] Aruja E. Displacement of X-Ray Reflexions. Nature 154**,** 53 (1944)

[24] Shen Y., Liu Y., and Sun W. "High-speed near dry electrical discharge machining. J. Mat. Process. Technol", **233**, pp. 9-18 (2016).

[25] Kruth J.P., and Bleys P. "Measuring residual stresses caused by wire EDM of tool steel", Int. J. Elect. Mach. **5**, pp. 23–28 (2000).

[26] Talla G., Gangopadhyay S., Biswas C.K. "Influence of Graphite Powder Mixed EDM on the Surface Integrity Characteristics of Inconel 625", J. Part. Sci. Technol. **35**, pp. 219-226 (2016).

Table 1 Residual stress experimental results as per Taguchi L_9 OA

Table 2 Tool dimensions for experiments

Table 3 Condition for experimental machining

Table 4 Chemical and physical properties of tool

Table 5 Chemical composition and physical properties of the workpiece (EN-31**)**

Table 6 Results obtained by X-ray diffraction testing as per Taguchi L_9 (OA)

Table 7 Average values and Main effects: Residual stress, RS (in MPa)

Table 8 ANOVA pooled data for Residual stress

Table 9 Confirmation experiments for RS

Fig. 1 Standard deviation error bar for residual stress for different a set of experiments

Fig.2 Portable XRD - residual stress analyzer in operation

Fig. 3 Diffraction orientation of incident X-ray from the sample

Fig. 4 Formation of Debye- Scherrer ring

Fig. 5 Scattering phenomenon of diffracted X-ray

Fig. 6 Schematic diagram of the working principle of PMND-EDM

Fig. 7 Indigenously developed setup for PMND-EDM

Fig. 8 Developed copper tool electrodes for experiments (All dimensions are in mm)

Fig.9 Design feature of developed tool

Fig. 10 Schematic diagram of setup for PMND-EDM

Fig. 11 Residual stress analysis as per Taguchi L_9 (OA)

Fig. 12 Induced micro-cracks in PMND-EDM

Fig. 13 Residual stress (MPa) Vs radial distance (mm)

Fig. 14 Effect of process parameters on RS (MPa)

Fig. 15 Rapid solidification over the surface (SEM image)

Fig. 16 SEM image of low metallurgical transformation (SEM image)

Fig. 17 Localized in-homogenous plastic deformation (SEM image)

Table 1 Residual stress experimental results as per Taguchi L_9 OA

Experiment No	Parameteric conditions				Residual stress (RS) (MPa)			S/N (decibel)
	A	B	C	D	Run1	Run2	Run3	
1	2	5	2	0.4	565	558	530	-54.82
2	2	10	5	0.5	621	551	550	-55.19
3	2	15	8	0.6	510	659	678	-55.85
4	3	5	5	0.6	370	361	340	-51.05
5	3	10	8	0.4	145	169	186	-44.48
6	3	15	2	0.5	239	279	280	-48.51
7	4	5	8	0.5	251	214	311	-48.35
8	4	10	2	0.6	230	209	191	-46.46
9	4	15	5	0.4	556	501	587	-54.79
Total					3478	3501	3653	
Mean					386.44	389	405.88	
					Overall mean RS ($\overline{RS}$)= 394.11 MPa			

Table 2 Tool dimensions for experiments

Tool 1	Tool 2	Tool 3
$d_i = 2$; $d_o=3$	$d_i = 3$; $d_o= 4$	$d_i = 4$; $d_o=5$
d_i= inner dia; d_o = outer dia; All dimensions are in mm		

Table 3 Condition for experimental machining

Parameter	**Description**
Workpiece/ tool	EN-31 (30mmx15mmx15mm)/ Copper
Dielectric medium	Metallic powder + Air + oil (LL-221) +
Current	12A
Gap voltage	25V
Pulse On/ Off	500µs/ 75µs
Polarity	+ve
Tool diameter (mm)	2, 3, 4
Concentration of metallic powder	2, 5, 8 (g l^{-1})
Flow rate of dielectric medium	5, 10, 15 (ml min^{-1})
Pressure of dielectric	0.4, 0.5, 0.6 (MPa)
Metallic powder	Zinc
Grain size of powder	15µm
Stabilizing agent	Glycerol (5%)
Machining time	10 mins

Table 4 Chemical and physical properties of tool

Properties	Values
Bulk modulus (Gpa)	140
Tensile strength (MPa)	33.3
Hardness (Mohs)	2.5-3
Thermal conductivity ($Wm^{-1} K^{-1}$)	385
Boiling point (^{O}C)	2567
Melting point (^{O}C)	1083
Density(Kg m^{-3})	8960
Atomic number	29
Atomic weight	63.546

Table 5 Chemical composition and physical properties of the workpiece (EN-31**)**

Chemical properties		Physical properties	
ELEMENT	%	Thermal conductivity (W $m^{-1}K^{-1}$)	44.5
Carbon	0.90-1.20	Hardness(HRC)	63
Silicon	0.10-0.35	Yield stress (MPa)	450
Manganese	0.30-0.75	Melting point (^{o}C)	1540
Sulphur	0.050	Tensile strength (MPa)	750
Phosphorus	0.050	Density (kg m^{-3})	7850

Table 6 Results obtained by X-ray diffraction testing as per Taguchi L_9 (OA)

Exp. No	RS (σ) MPa	σ_1 (MPa)	(τ_{xy}) MPa	σ_2 (MPa)	FWHM (Degree)	α- max (Degree)	α- min (Degree)
1	530	20	-79	11	3.49	314.60	204.48
2	551	43	-61	26	3.80	20.88	241.20
3	659	49	13	33	3.90	83.52	200.16
4	370	66	23	53	4.42	34.56	216.72
5	145	64	-199	60	4.31	103.68	282.24
6	280	19	-133	23	3.95	39.60	216
7	251	51	-63	54	4.19	277.92	259.92
8	191	81	-115	44	4.21	161.28	110.88
9	280	19	-133	23	3.95	39.60	216

Table 7 Average values and Main effects: Residual stress, RS (in MPa)

Process parameter	Level	Tool dia (A)		Flow rate (B)		Powder conc (C)		Pressure (D)	
Data type		S/N	Data Raw	S/N	Data Raw	S/N	Data Raw	S/N	Data Raw
Average values (% RS)	L1	-55.29	580.22	-51.41	388.88	-49.93	342.33	-51.36	421.88
	L2	-48.01	263.22	-48.71	316.88	-53.68	493.00	-50.69	366.22
	L3	-49.87	338.88	-53.05	476.55	-49.56	347.00	-51.12	394.22
Main effects (% RS)	L2-L1	7.27	-317.00	2.69	-72.00	-3.74	150.66	0.67	-55.66
	L3-L2	-1.85	75.66	-4.34	159.66	4.11	-146.00	-0.43	28
Differences (L3-L2)- (L2-L1)		-9.12	392.66	-7.04	231.66	7.86	-296.66	-1.11	83.66

Table 8 ANOVA pooled data for Residual stress

Source	SS Raw	SS S/N	SS' raw	SS' S/N	D.O.F raw	D.O.F S/N	Variance raw	Variance S/N	F-ratio raw	F-ratio S/N	P% raw	P% S/N
Tool Diameter	493368.66	85.63	489736	84.92	2	2	246684.30	42.81	135.81	121.16	62.67	58.54
Flow rate	115088.66	28.82	111456	28.11	2	2	57544.33	14.41	31.68	40.78	14.61	19.70
Powder conc.	132114.66	31.11	128482	30.40	2	2	66057.33	15.55	36.36	44.02	16.78	21.27
Mist Press-ure	13944.66	0.70	*	-	2	--	6972.33	*	3.83	*	1.77	*
Error	32694	0.70	57536.67	2.82	18	2	1816.33	0.35	-	-	4.15	0.48
Total	787210.66	146.27	787210.70	146.27	26	8	*	-	-	-	100	100

*Significant at 95% confidence level, F- critical=3.55 (Tabular value for raw data), SS- Sum of squares, DOF- Degree of freedom, V- variance , F- crirtical=19 (Tabular value for S/N data), SS'= Pure sum of squares

Table 9 Confirmation experiments for RS

Tool Diameter (mm)	Flow Rate (ml min $^{-1}$)	Metallic Powder concentration (g l^{-1})	Mist pressure (MPa)	Experimental Results RS (MPa)	Average RS (MPa)
3	10	2	0.5	90.32 110.39 118.49	106.40

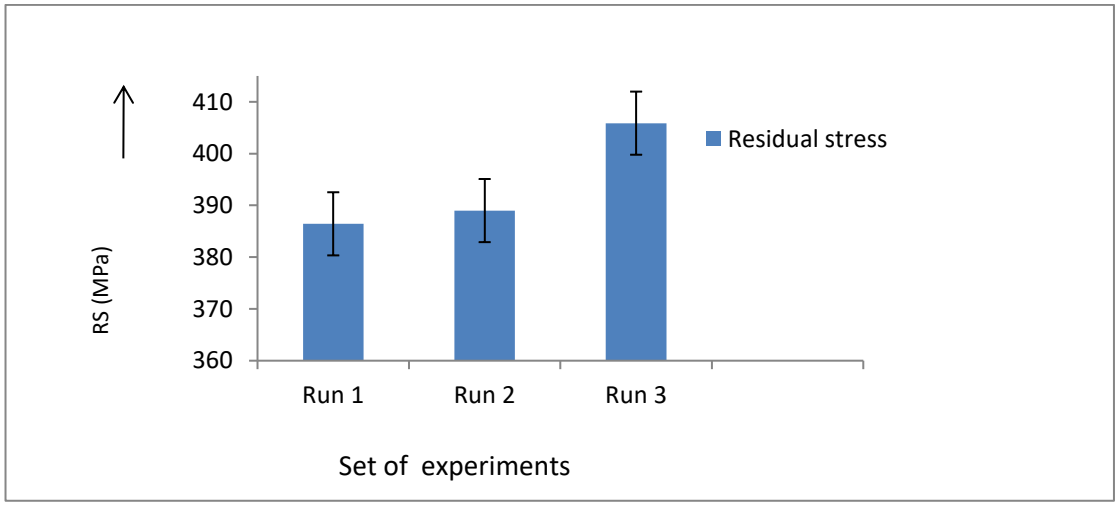


Fig. 1 Standard deviation error bar for residual stress for different a set of experiments

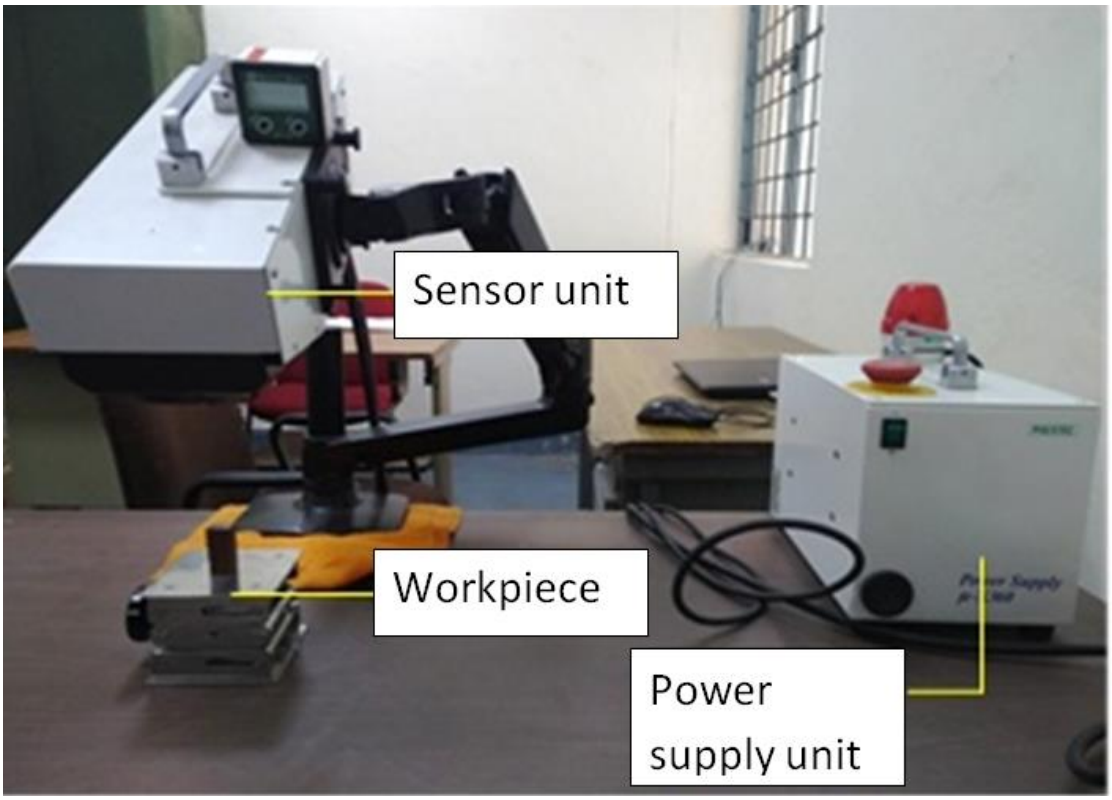


Fig.2 Portable XRD - residual stress analyzer in operation

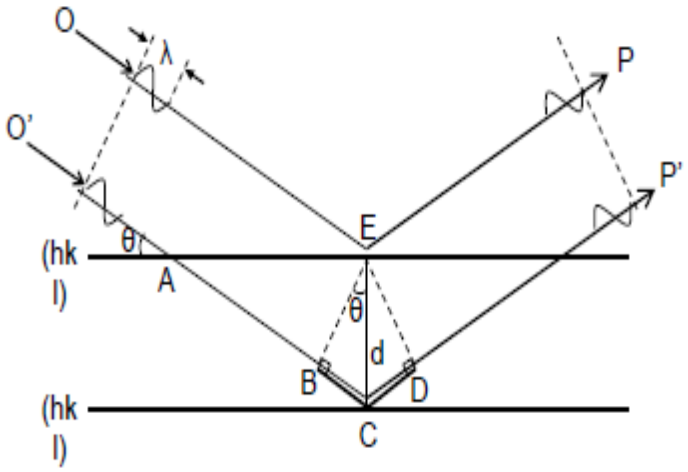


Fig. 3 Diffraction orientation of incident X-ray from the sample

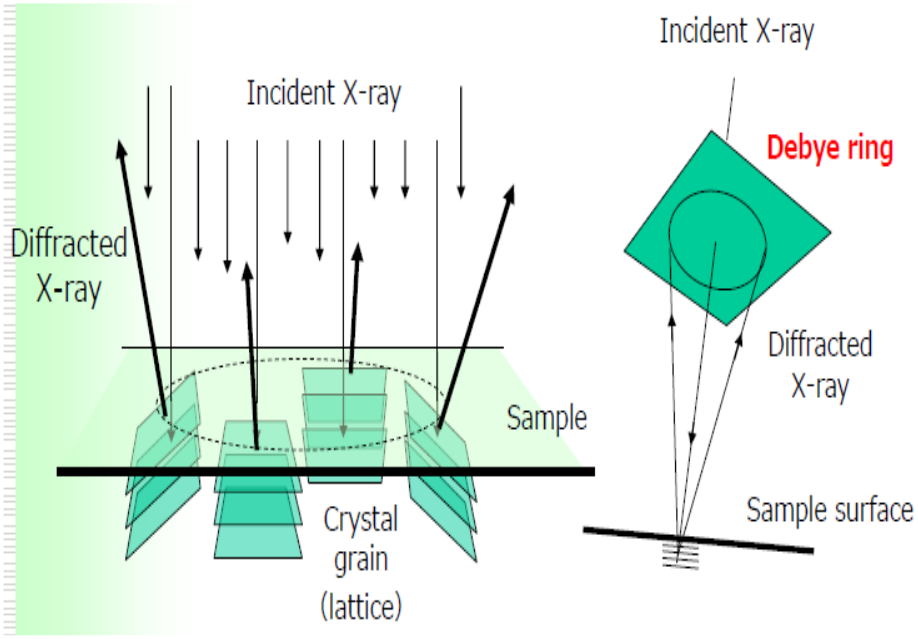


Fig. 4 Formation of Debye- Scherrer ring

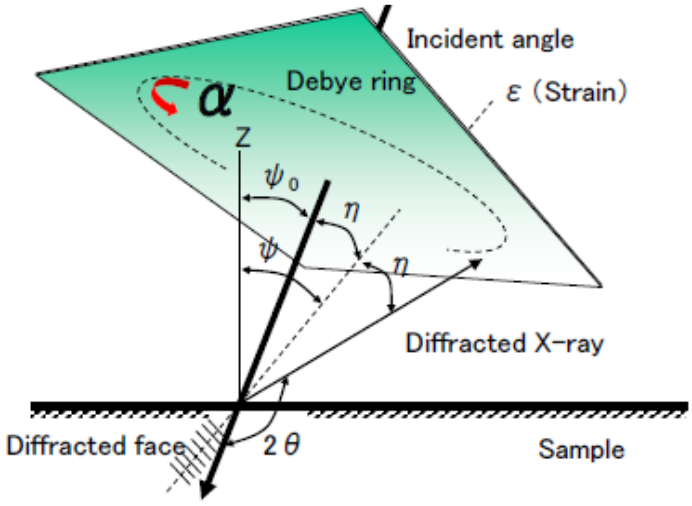


Fig. 5 Scattering phenomenon of diffracted X-ray

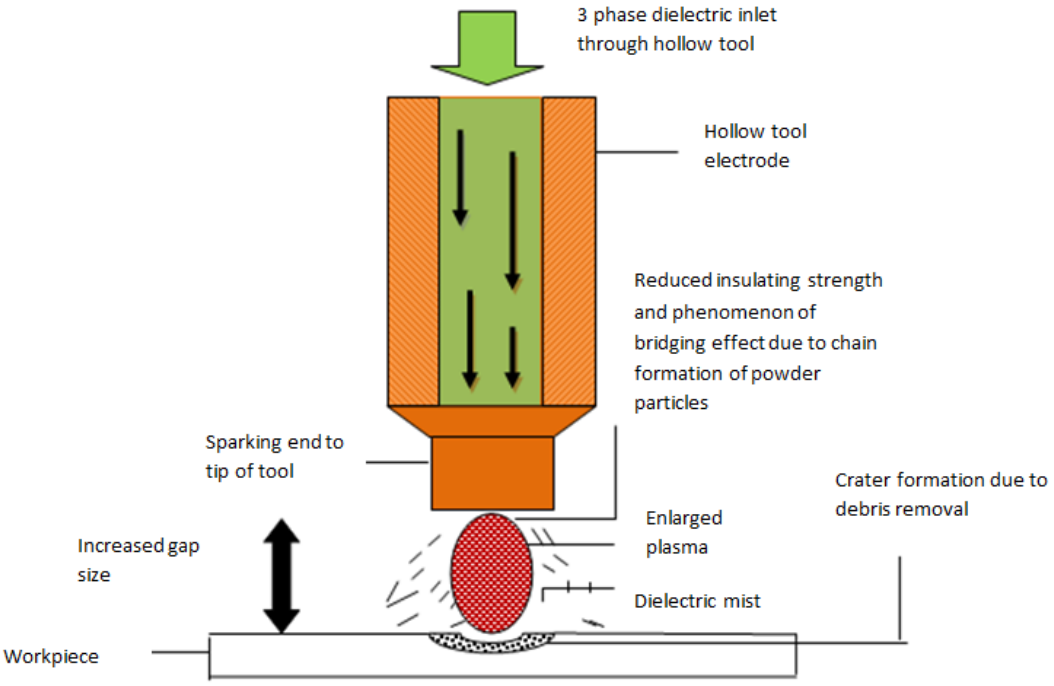


Fig. 6 Schematic diagram of the working principle of PMND-EDM

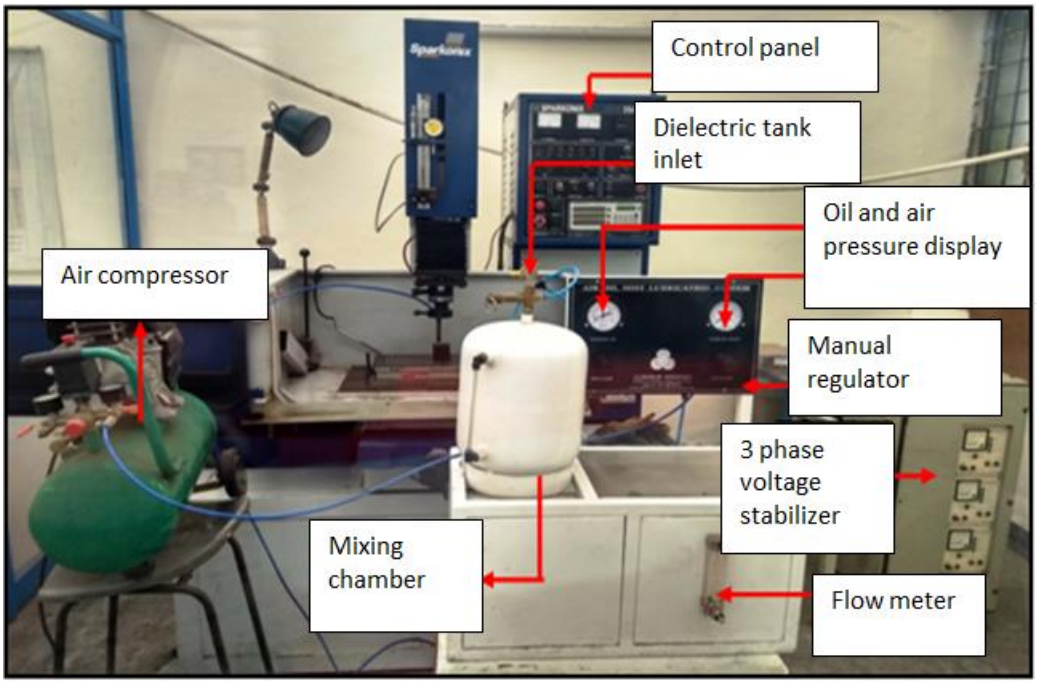


Fig. 7 Indigenously developed setup for PMND-EDM

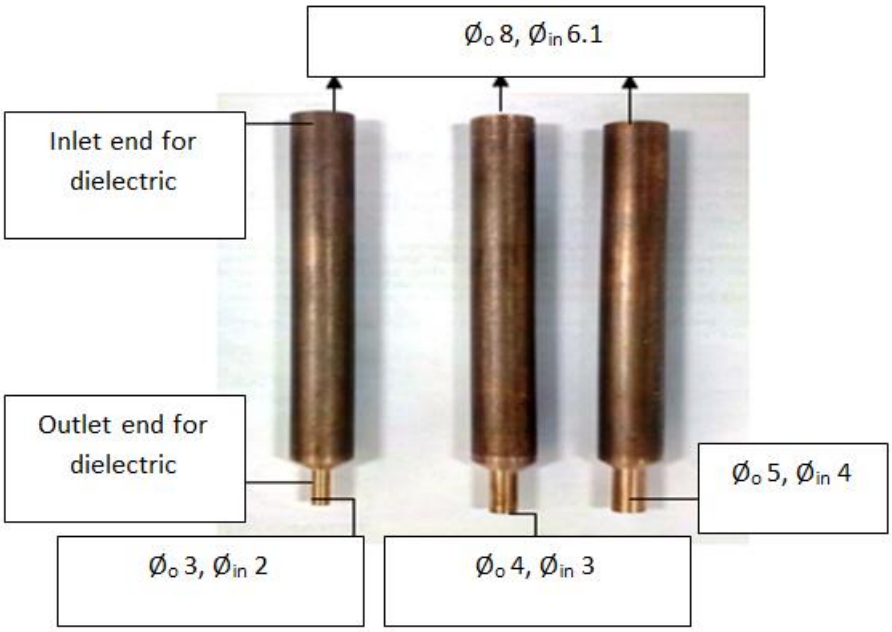


Fig. 8 Developed copper tool electrodes for experiments (All dimensions are in mm)

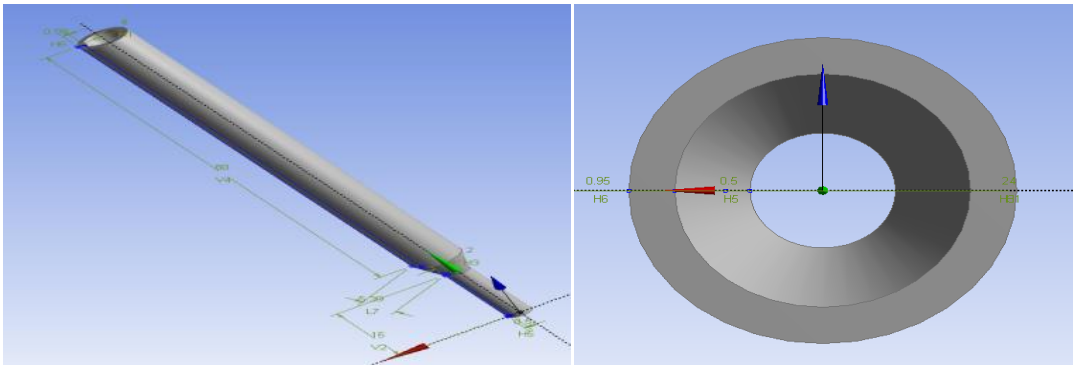


Fig.9 Design feature of developed tool

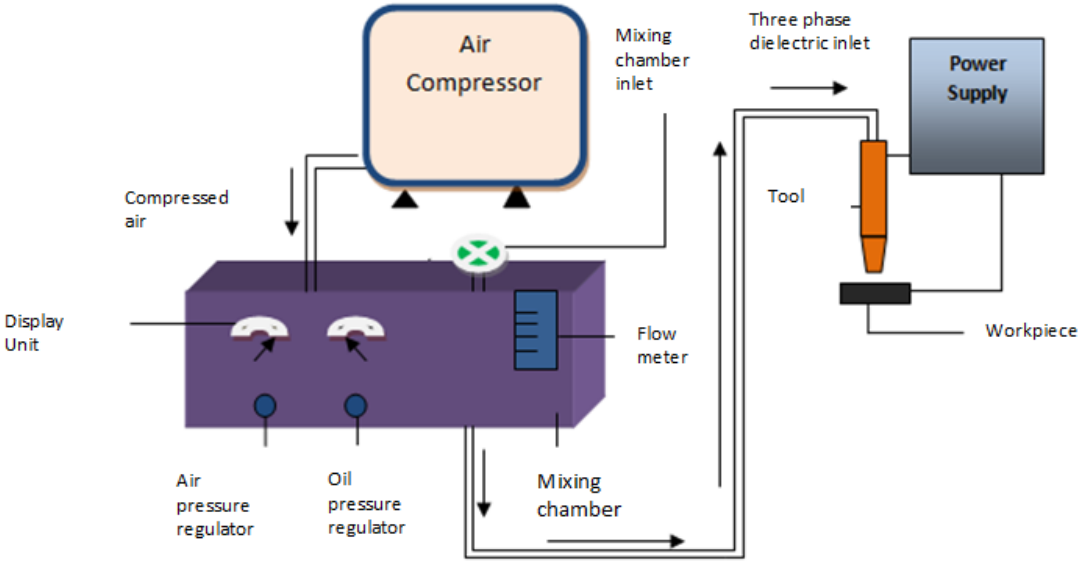


Fig. 10 Schematic diagram of setup for PMND-EDM

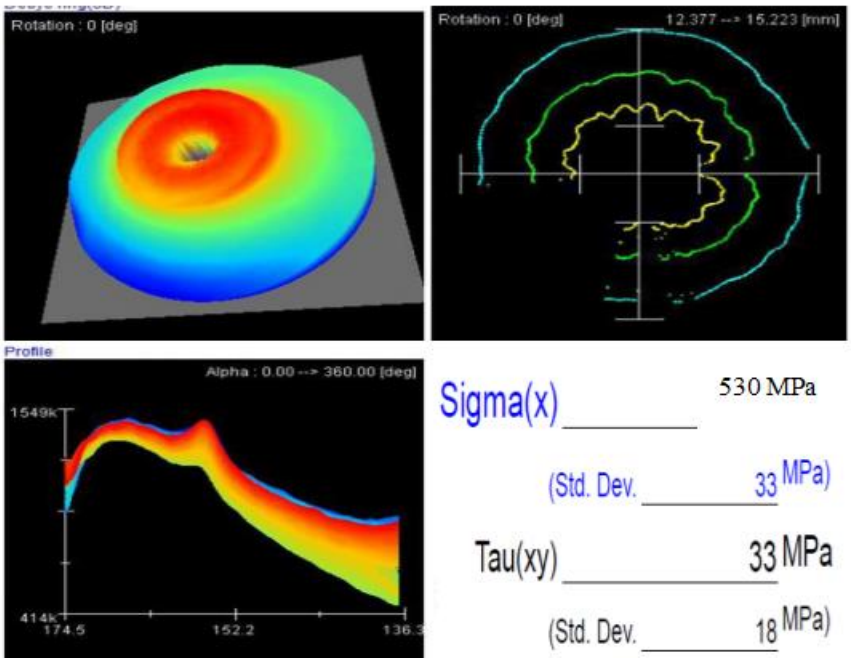


Fig 11 (a) Residual stress for 1st series experiment

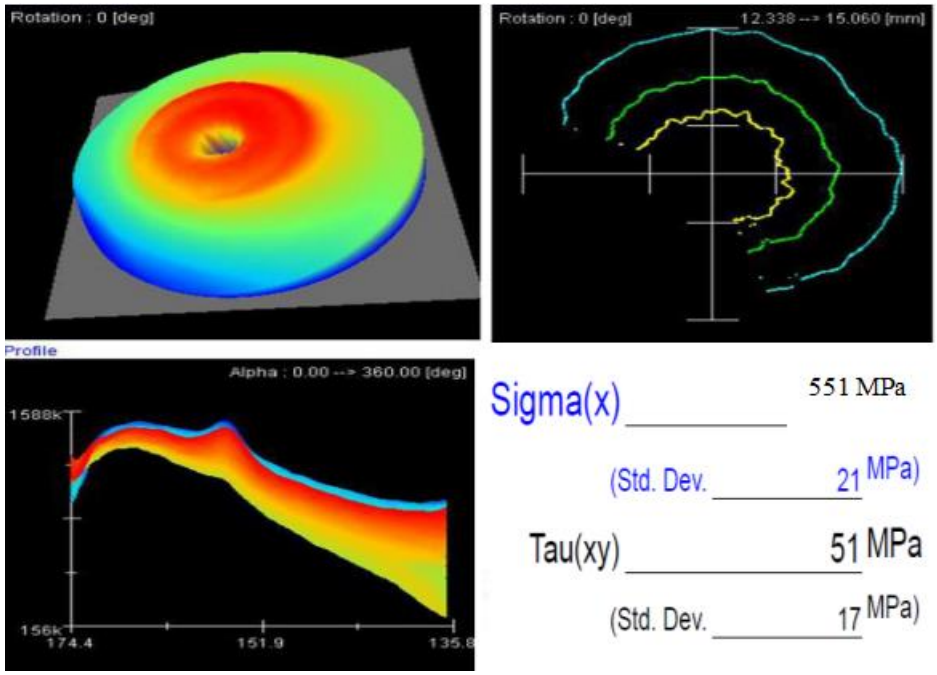


Fig 11 (b) Residual stress for 2nd series of experiment

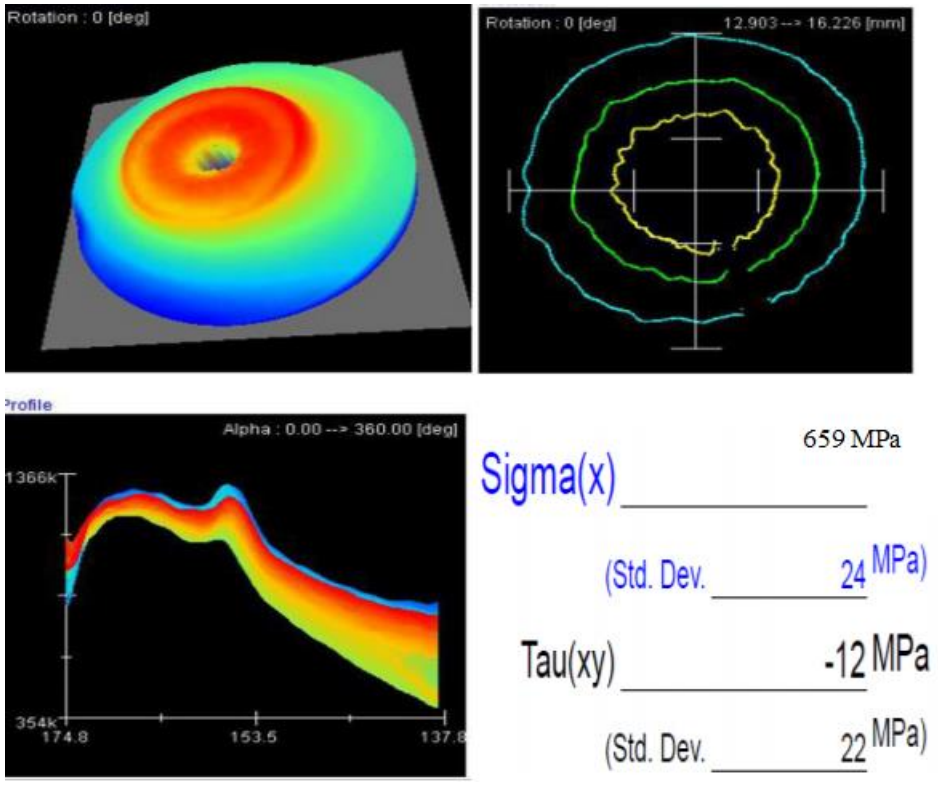


Fig 11 (c) Residual stress for 3rd series of experiment

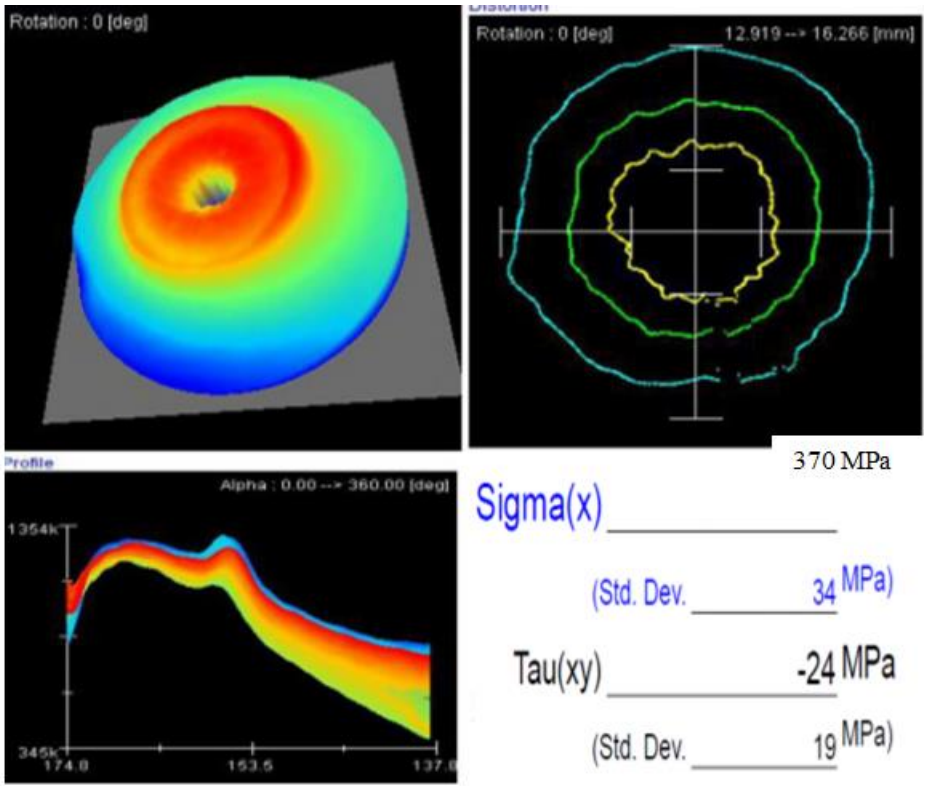


Fig 11 (d) Residual stress for 4th series of experiment

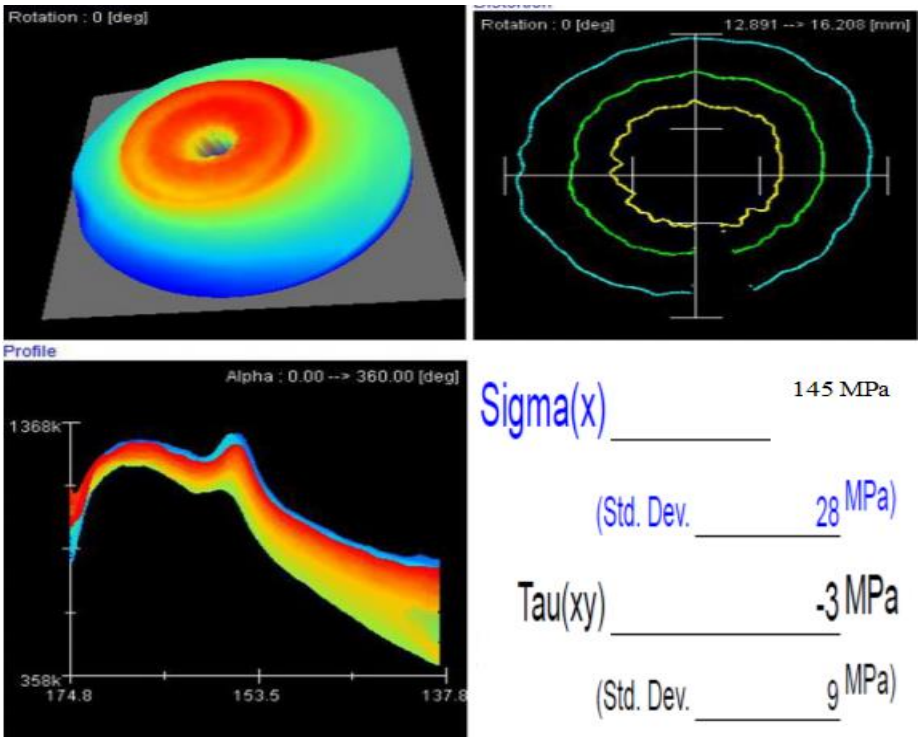


Fig 11 (e) Residual stress for 5th series of experiment

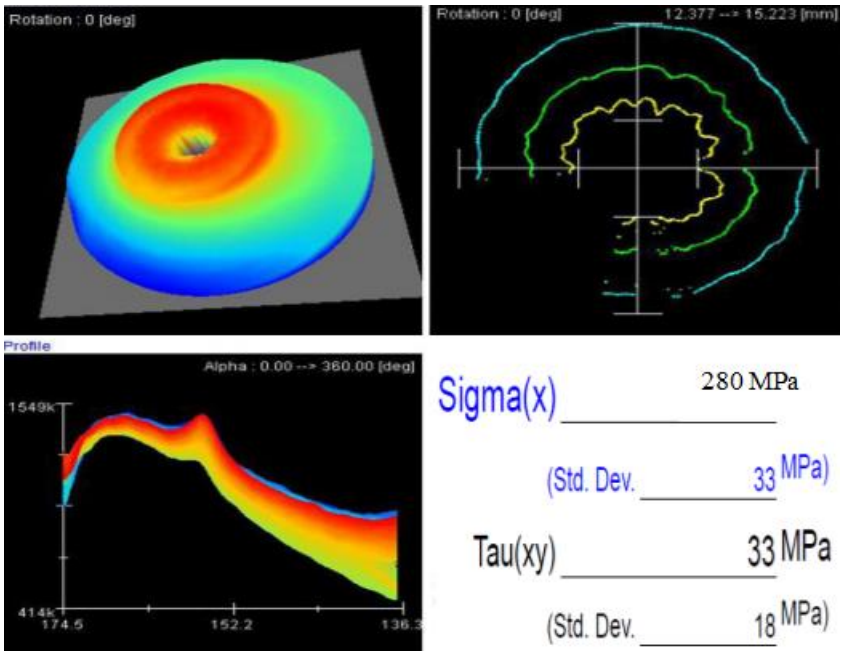


Fig 11 (f) Residual stress for 6th series of experiment

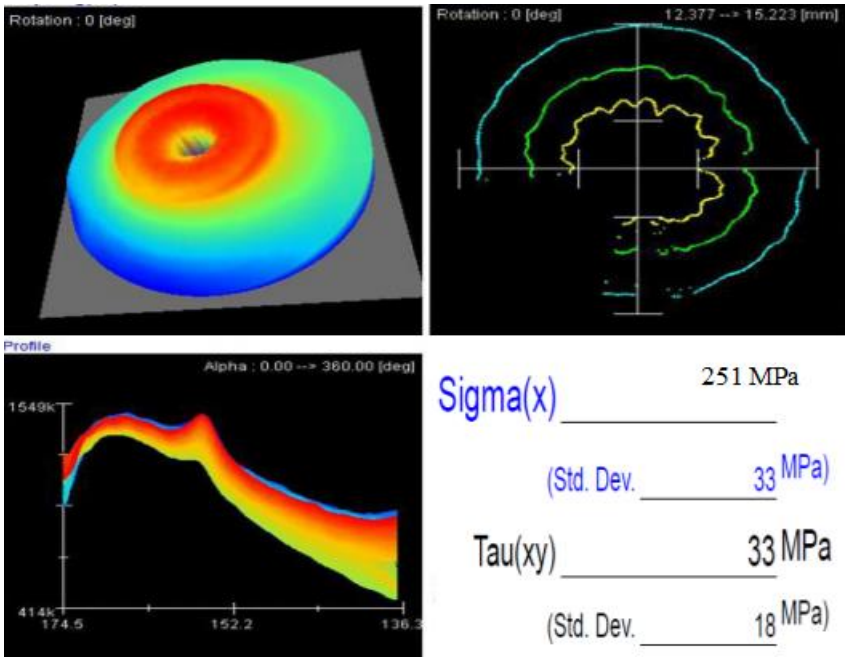


Fig 11 (g) Residual stress for 7th series of experiment

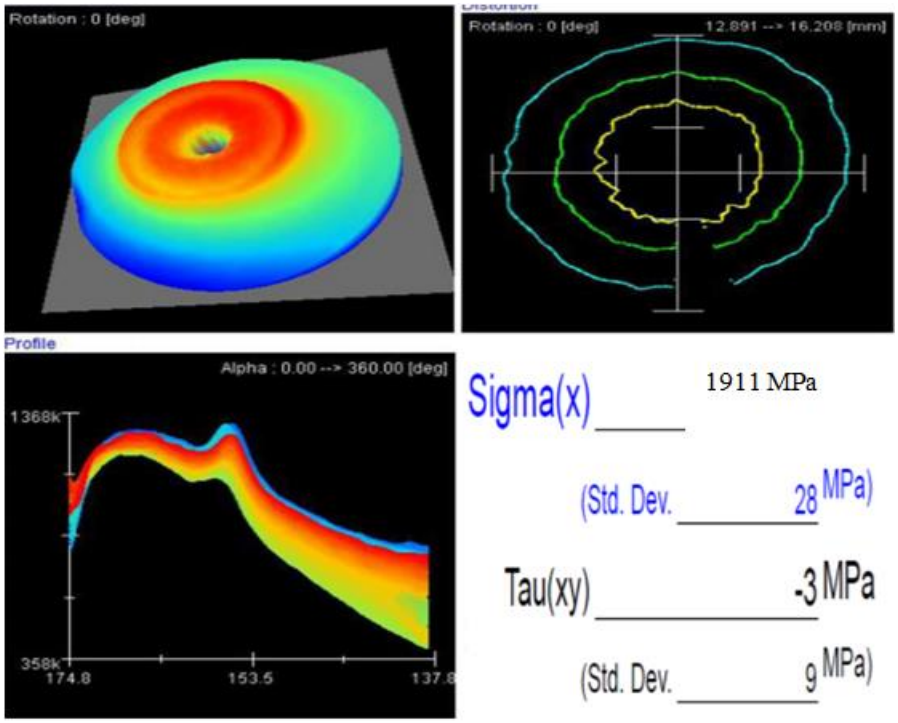


Fig 11 (h) Residual stress for 8th series of experiment

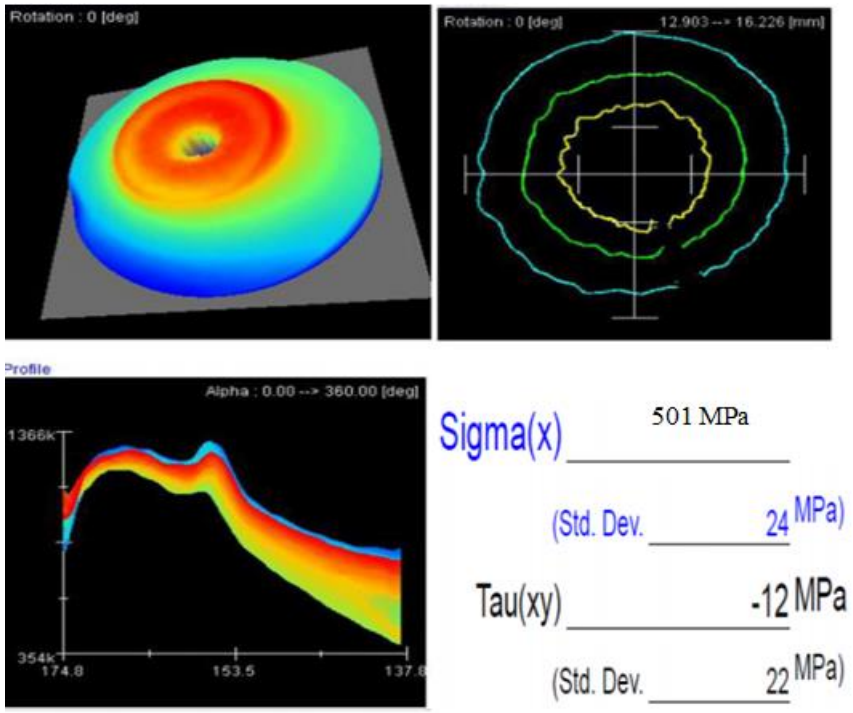


Fig 11 (i) Residual stress for 9th series of experiment

Fig. 11 Residual stress analysis as per Taguchi L_9 (OA) (a-i)

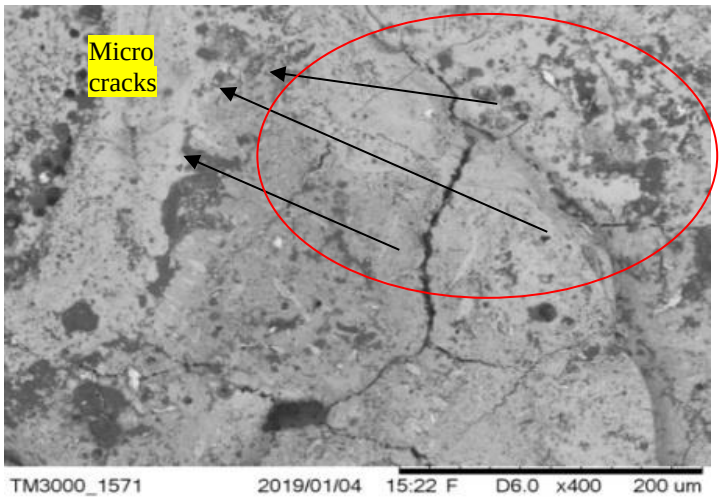


Fig. 12 Induced micro-cracks in PMND-EDM

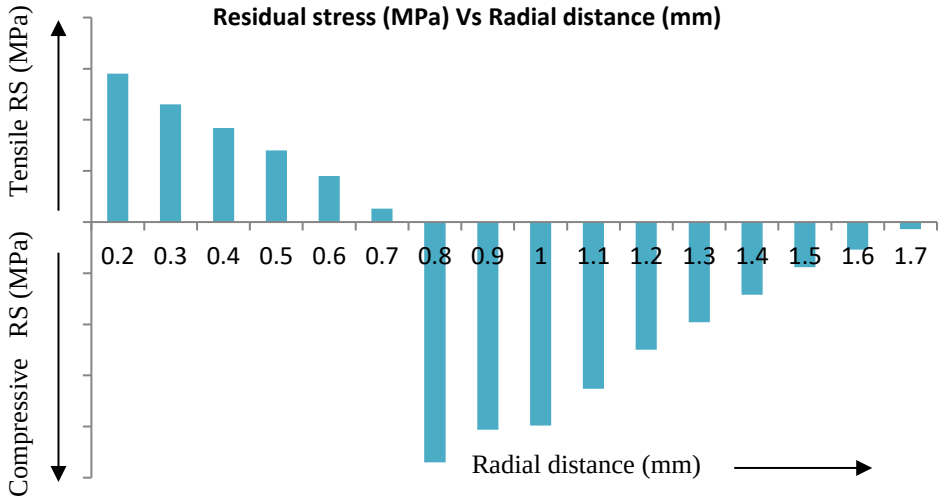


Fig. 13 Residual stress (MPa) Vs radial distance (mm)

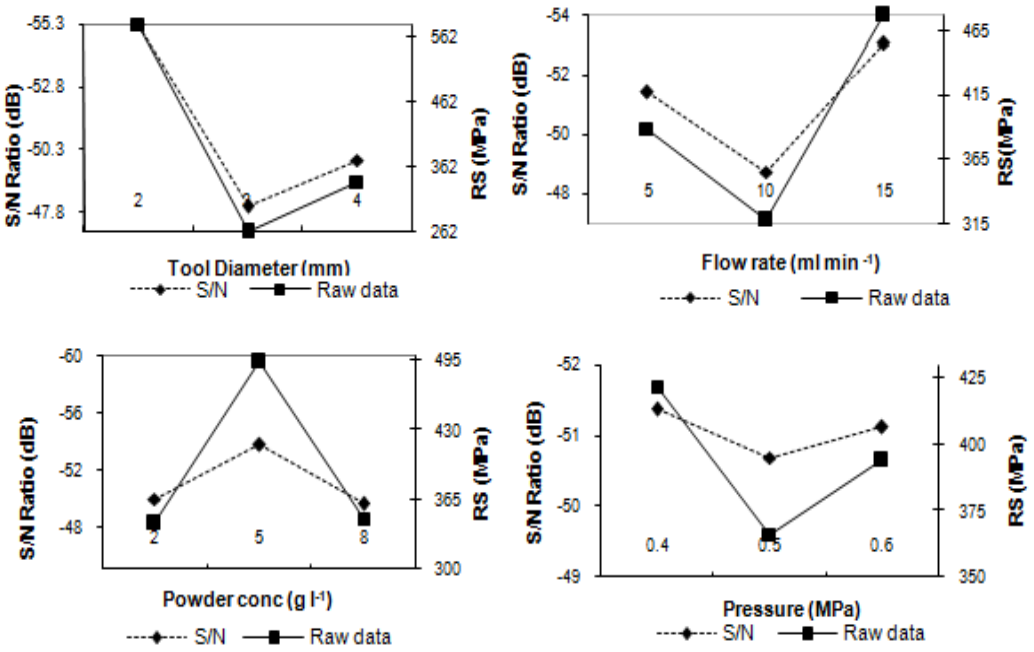


Fig. 14 Effect of process parameters on RS (MPa)

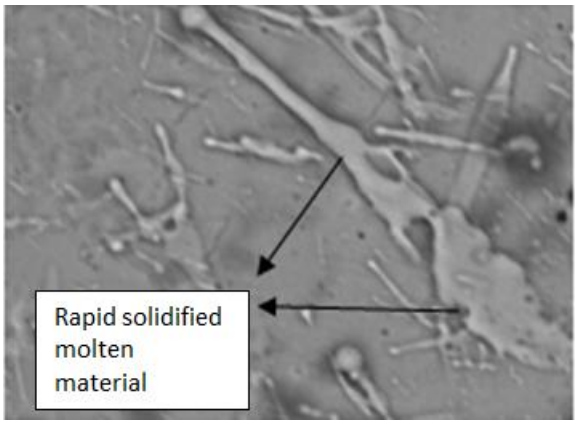


Fig. 15 Rapid solidification over the surface (SEM image)

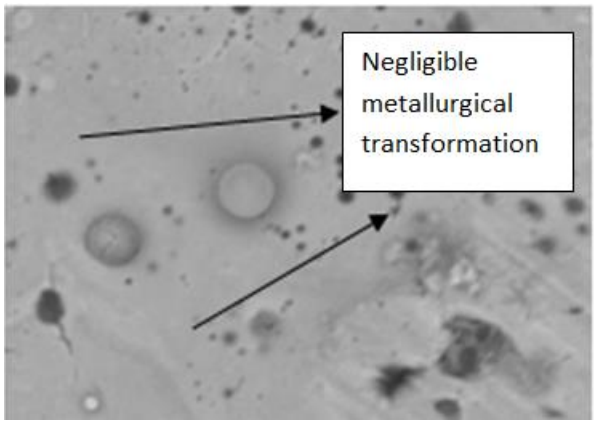


Fig. 16 SEM image of low metallurgical transformation (SEM image)

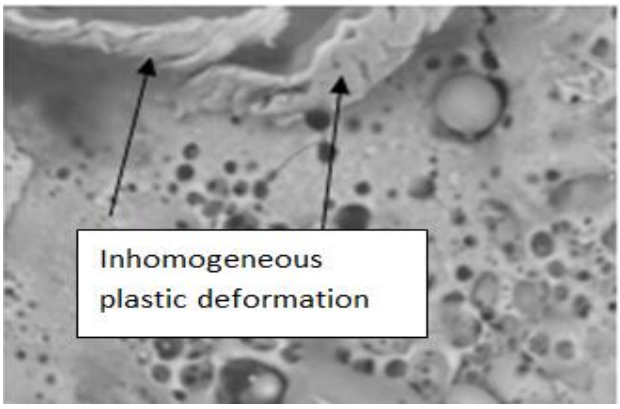


Fig. 17 Localized in-homogenous plastic deformation (SEM image)

Sanjay Sundriyal is a PhD student at Delhi Technological University of India. He graduated from College of Engineering Roorkee and received his masters from National Institute of Technology Warangal. His research interests are advanced manufacturing process and optimization.
Dr. Ravinderjit Singh Walia is a professor in Punjab Engineering College. He holds his PhD degree from Indian Institute of Technology (Roorkee). His area of interest includes advanced manufacturing process, optimization, modeling and industrial engineering. He is author of more than 100 papers, book chapters, conferences and referred journals.
Dr. Vipin is a professor in Delhi Technological University of India. He holds his PhD degree from Delhi Technological University. His area of interest includes advanced manufacturing process, optimization and modeling. He has authored many papers, book chapters, conferences and referred journals.

ACADEMIA
Accelerating the world's research.

Thermodynamic performances of ejector refrigeration systems (ERS) using low GWP ecofriendly HFO refrigerants

Editor IJREI

Related papers

Download a PDF Pack of the best related papers



Performance improvement of vapour compression refrigeration systems
Editor IJREI

Methods for improving thermodynamic performances of vapour compression refrigeration system
Editor IJREI

Thermal performance of multi cascaded vapour compression NH 3 H 2 O absorption systems for ultr...
Editor IJREI

International Journal of Research in Engineering and Innovation (IJREI)
journal home page: http://www.ijrei.com
ISSN (Online): 2456-6934



Thermodynamic performances of ejector refrigeration systems (ERS) using low GWP ecofriendly HFO refrigerants

R. S. Mishra

Department of Mechanical Engineering, Delhi Technological University Delhi, India

Abstract

Currently, the refrigeration and air conditioning is looking for energy efficient technologies which will reduce the electric power consumption from causing the damage to the environment. The ejector refrigeration system is found to be energy efficient system which will be able to provide the cooling by utilizing the environment friendly refrigerants. In this paper HFO refrigerants are used in ejector refrigeration system and comparison were made with other HFC and HCFC refrigerants. ©2020 ijrei.com. All rights reserved

Keywords: Energy-Exergy analysis, Ejector refrigeration system, Thermodynamic performances.

1. Introduction

In the context of recent developments in the field of energy, the aspect related to energy consumption is of excessive importance for experts. Many industries trust on refrigeration technologies; a great challenge being stated in energy savings in this sector. In this respect, efforts oriented towards efficient industrial refrigeration systems have revealed the necessity of a proper design. The most commonly used method of cooling is based on vapor compression cycles.

Ejector refrigeration technology is the one which can have utilized the waste heat to compress the refrigerant instead of compressor. Therefore, electrical energy was saved which can have employed to meet the other demand. When ejector is integrated with vapour compression refrigeration system, then system will use less compressor power to provide the natural cooling of products.

2. Ejector Refrigeration system

In the vapour compression refrigeration system, throttle device is used to expand the refrigerant from high pressure prevailing in the condenser to low pressure evaporator. Capillary tube and thermostatic expansion valve, used in the throttling devices in the vapour compression refrigeration system has highly irreversible process which causes larger amount of wastage of the kinetic energy of the refrigerant. Vapour compression refrigeration systems uses CFCs, HFCs and HCFC refrigerants and their leakage from the system causes emissions in the environment hence degrading its air quality. This emission can be direct emission or indirect or both. Therefore, energy efficient technologies which reduced the electric power consumption apart from the causing the damage to the environment. Ejector refrigeration system is an energy efficient device which is able to compress the refrigerant in the system and providing the cooling by utilizing the environment friendly refrigerant

Chen, J, [1] studied theoretically, the application of different nine working fluids in the ejector vapour refrigeration system and found theR600 gives higher first law efficiency (COP) of ejector refrigeration system. Chen, J, [2] had performed exergy analysis of an ejector refrigeration system using ecofriendly R134a refrigerant and found that by reducing the exergy destruction in the parts improves the quality of exergy analysis in terms of exergetic efficiency and concluded the maximum exergy destruction in ejector followed by generator and evaporator and tried for reducing ejector exergy destruction by improving the ejector efficiencies. Similarly, generator (boiler) exergy destruction can be reduced by improving design of other components of the system. Memet and A Preda [3] carried out theoretical analysis of ejector refrigeration system using R134a

Corresponding author: R.S. Mishra
Email Address: hod.mechanical.rsm@dtu.ac.in
https://doi.org/10.36037/IJREI.2020.4501.

and found that when by increasing the generating temperature, the coefficient of performance(COP) is increasing together and the best COP value being 0.178. also the generating temperature increase leads to the increase of the work input to the pump. Li, H. et.al. [4] had compared thermodynamic performance characteristics of ejector expansion refrigeration cycle using R1234yf and R134a as a refrigerants and found that R1234yf gives better performance than R134a at 40°C of condenser temperature and 5°C of evaporator temperature. Trung Kien Nguyen & Chi Hiep Le [5] carried out the first and second law of thermodynamics, an ejector vapour compressor cascade refrigeration system using R134a and R410A refrigerant for the ejector and compressor sub-cycle, respectively. Also an intercooler was inserted for setting temperature at 26.5 °C for minimizing total exergy loss in the system. The effects of generator temperature, evaporator temperature, condenser temperature, intercooler temperature and intercooler temperature difference are studied and found that a combined cycle can improve 30.8% of COP in single compressor cycle and 122% when compared to the single ejector cycle. M Yari, M Sirousazar [6] carried out the first and second law (exergy) analysis of the ejector-vapour compression refrigeration cycle using internal heat exchanger and intercooler for enhancing the performance of the cycle and investigated the effects of the evaporative and condenser temperatures on the coefficient of performance (COP), second law efficiency, exergy destruction rate and entrainment ratio. and found that the COP and second law efficiency values of the new ejector-vapour compression refrigeration cycle are on average 8.6% and 8.15 % higher than that of the conventional ejector-vapour compression refrigeration cycle using R125. It was also observed that the COP of the new ejector-vapour compression cycle is 21 per cent higher than that of the conventional vapour compression cycle.

Yapici and Ersoy [7] had compared the thermodynamic performance of constant pressure and constant area mixing ejector using HR123 refrigerant and found the optimum coefficient of performance and area ratio determined by using the constant area flow model are greater than those of constant pressure model. Yin-Hai Zhu & Peixue Jiang [8] developed refrigeration system which combines a basic vapour compression refrigeration cycle with an ejector cooling cycle. The ejector cooling cycle is driven by the waste heat from the condenser in the vapor compression refrigeration cycle. The additional cooling capacity from the ejector cycle is directly input into the evaporator of the vapor compression refrigeration cycle. The governing equations are derived based on energy and mass conservation in each component including the compressor, ejector, generator, booster and heat exchangers. The system performance was analyzed for the design conditions and found that the COP is improved by 9.1% for R22 system. Chaiwongsa and Wongwishes [9] have carried out experimental investigation on two phase ejector as expansion device using R134a and R12 refrigerants and observed the COP improvement at lower sink temperature because motive mass flow rate is highly dependent on heat sink temperature. Chaiwongsa and Wongwishes [10] have carried out experimental investigation on the effect of throat diameter of the nozzle on the performance on the refrigeration system using ecofriendly R134a refrigerant in the two phase ejector as expansion device and found maximum cooling capacity and COP at 0.8mm throat diameter. by varying the dimension of nozzle diameter with cooling capacity. Da-Wen Sun [11] described a new integrated refrigeration cycle based on the combination of an ejector cycle with a vapour compression cycle for maximizing performance of the conventional ejector cycles and providing high COP for refrigeration and observed that this cycle has a significant increase in system performance over the conventional systems and its COP values are competitive to the absorption machines. In this modified system, the waste heat was used and, higher COP value was observed. The system performance can be further improved if dual refrigerants are used. Lucas and Kochler [12] found experimentally the improvement in COP of the vapour compression refrigeration cycle using R744 (CO_2) as refrigerant and found the entrainment ratio and ejector efficiency increases with increase in high side pressure and also concluded that ejector efficiency is maximum decrement with decreasing evaporator pressure. Kornhauser A.A [13] analyzed thermodynamic performance of the ejector-expansion refrigeration cycle on constant mixing pressure using R12 and found 21% improvement in COP. Nehdi [14] carried out theoretical investigation on the performance of ejector-expansion refrigeration cycle with ecofriendly synthetic refrigerants and found maximum performance of 22% using R141b. Also investigated maximum first law performance (COP) using ejector as expansion device occurred at 9.9 area ratio and found best performance for using R141b and R408A for a given operating condition. Bilir and Ersoy [15] investigated theoretical improvement of ejector refrigeration cycle using two phase ejector using R134a refrigerant and observed that COP increases with decreasing evaporator temperature and also by increasing condenser temperature. Saleh B., [16] had numerical investigated parameter analysis of ejector refrigeration cycle with different refrigerants using ejector refrigeration cycle and found best thermodynamic performance using R245fa. Sarkar [17] carried out optimization of geometric parameters of the ejector expansion refrigeration cycle using natural refrigerant (R717) and hydrocarbon (R600a and R290) refrigerants and observed 21.55% maximum COP improvement using isobutane (R600a), 18% using propane(R290).12% using ammonia(R717).

Lot of work have been done by several investigators using CFCs, HFCs, natural and hydrocarbon refrigerants in the ejector refrigeration systems and still very less work has been done using HFO refrigerant. In this paper the utility of HFO refrigerants in ejector refrigeration system are highlighted for improving thermodynamic performances.

3. Results and Discussion

Following Numerical values have been chosen for numerical computation

Cooling Load(Qeva)	4.75 KW
Boiler temperature(T_boiler)	333 K
Evaporator temperature (T_eva)	273 K
Condenser temperature (T_cond)	303 K
Ambient temperature (To)	300 K
T_ref	T_eva+5

Ejector geometric input

(Length / Diameter} ratio of constant area mixing chamber(L/D)	10
Diameter of primary nozzle throat (D_throat)	0.5/1000
Diameter of mixing chamber(d_m)	1.4/1000
Exit diameter of primary nozzle (d_p)	0.8/1000
Diffuser angle (theta)	3
Diffuser Length (L_d)	112/1000
Refrigerant	R1234zf
Area Ratio	7.84

Table- 1(a): Effect of ecofriendly refrigerants on the variation of thermal performance parameters of vapour compression refrigeration system using ejector

Refrigerant	COP	EDR	2nd law Efficiency
R-1234ze(Z)	6.483	4.007	0.1997
R-1234ze(E)	6.213	3.56	0.2193
HFO-1336 mzz(Z)	4.723	3.573	0.2187
R-1243zf	6.016	3.754	0.2103
R1234yf	5.788	3.298	0.2327
R-1233zd(E)	5.799	3.833	0.2069
R-1224yd(Z)	5.163	3.468	0.2238
R-1225ye(Z)	5.734	3.212	0.2374
R-124	5.333	3.256	0.2349
R123	5.141	3.545	0.220

Table-1(a) shows the variation of first law efficiency (COP) of vapour compression refrigeration system using ejector using ecofriendly refrigerants and it was observed that maximum COP was observed by using R1234ze(Z) and minimum by using R123. Similarly, maximum second law efficiency was observed by using R-1225ye(Z)

Table-1(b) Effect of ecofriendly refrigerants on the variation of thermal performance parameters of vapour compression refrigeration system using ejector

Refrigerant	Entrainment Ratio(shy)	Compression Ratio (Mu_1)	Mu_2
R-1234ze(Z)	1.454	1.942	7.089
R-1234ze(E)	1.383	1.970	6.804
HFO-1336 mzz(Z)	1.535	2.347	5.328
R-1243zf	1.359	1.807	6.543
R1234yf	1.353	1.987	6.313
R-1233zd(E)	1.477	2.07	6.403
R-1224yd(Z)	1.467	2.229	5.714
R-1225ye(Z)	1.386	2.151	5.858
R-124	1.396	2.162	5.827
R123	1.496	2.254	5.648
R125	1.326	2.155	5.32

Table-1(b) shows the variation of Entrainment Ratio (shy) of vapour compression refrigeration system using ejector using ecofriendly refrigerants and it was observed that maximum Entrainment Ratio was observed by using HFO-1336mzz(Z) and minimum by using R125. Similarly, maximum compression Ratio was observed by using R-1225ye(Z).

Table-1(c) Effect of ecofriendly refrigerants on the variation of thermal performance parameters of vapour compression refrigeration system using ejector

Refrigerant	Effective ness	a	b
R-1234ze(Z)	0.07336	2.072	1.02
R-1234ze(E)	0.08521	1.963	1.03
HFO-1336 mzz(Z)	0.1063	2.134	0.9652
R-1243zf	0.08476	1.904	1.035
R1234yf	0.09896	1.891	1.033
R-1233zd(E)	3000.08411	2.097	1.022
R-1224yd(Z)	0.1034	2.063	0.9902
R-1225ye(Z)	0.1094	1.933	1.016
R-124	0.1085	1.959	1.011
R123	0.1030	2.092	0.9832

Table-1(c) shows the variation of Effectiveness of vapour compression refrigeration system using ejector using ecofriendly refrigerants and it was observed that maximum Effectiveness was observed by using R-1225ye(Z).

Table- 2(a): Variation with Generator (Boiler) temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_Boiler (K)	Exergetic efficiency	EDR	shy
333	0.2103	3.754	1.359
338	0.2.091	3.782	1.359
343	0.2080	3.807	1.359
348	0.2071	3.829	1.359
353	0.2062	3.848	1.359
358	0.2056	3.863	1.359
363	0.2053	3.871	1.359

Table-2(a) shows the variation of Exergetic efficiency of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of boiler temperature and it was observed that boiler temperature is increasing, second law efficiency (Exergetic efficiency) is decreasing.

Table- 2(b): Variation with Generator (Boiler) temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_Boiler (K)	a	b	mu_1
333	1.398	0.6952	1.665
338	1.449	0.7046	1.627
343	1.497	0.7132	1.593
348	1.542	0.7218	1.563
353	1.582	0.7296	1.537
358	1.619	0.7373	1.514
363	1.648	0.7444	1.495

Table-2(b) shows the variation of Compression Ratio of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of boiler temperature and it was observed that boiler temperature is increasing, Compression Ratio of ejector coupled vapour compression refrigeration system is decreasing.

Table- 3(a): Variation with condenser temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_Cond(K)	Exergetic efficiency	EDR	shy
300	0.2127	3.701	1.663
303	0.2138	3.676	1.807
306	0.2149	3.651	1.961
308	0.2158	3.634	2.069
309	0.2162	3.625	2.124

Table-3(a) shows the variation of Exergetic efficiency of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of condenser temperature and it was observed that condenser temperature is increasing, second law efficiency (Exergetic efficiency) is increasing. Similarly, entainment ratio is also increasing while exergy destruction ratio is decreasing

Table- 3(b): Variation with condenser temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_Cond(K)	COP	mu1	Mu2
300	1.363	1.807	1.486
303	0.9115	1.807	0.9951
306	0.6283	1.807	0.6867
308	0.4595	1.807	0.5376
309	0.4337	1.807	0.4746

Table-3(b) shows the variation of first law efficiency (COP) of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of condenser temperature and it was observed that condenser temperature is increasing, first law efficiency (COP) is decreasing.

Table-3(c) shows the variation of first law efficiency (COP) of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of condenser temperature and it was observed that condenser temperature is increasing, first law efficiency (COP) is decreasing.

Table- 3(c): Variation with condenser temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_Cond(K)	COP	a	b
300	1.363	1.582	0.8595
303	0.9115	1.489	0.8092
306	0.6283	1.413	0.7676
308	0.4595	1.368	0.7436
309	0.4337	1.348	0.7326

Table-4(a): Variation with evaporator temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_EVA[K]	Exergetic efficiency	EDR	shy
268	0.3826	1.613	3.002
273	0.3347	1.988	2.517
278	0.2791	2.583	2.126
283	0.2138	3.676	1.807

Table-4(a) shows the variation of second law efficiency (exergetic efficiency) of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of evaporator temperature and it was observed that evaporator temperature is increasing, second law efficiency (exergetic efficiency) is decreasing.

Table-4(b): Variation with evaporator temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_EVA[K]	COP	mu1	Mu2
268	0.3692	1.602	0.4163
273	0.4888	1.665	0.5451
278	0.6556	1.732	0.7233
283	0.9115	1.807	0.9950

Table-4(b) shows the variation of first law efficiency (COP) of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of evaporator temperature and it was observed that evaporator temperature is increasing,
first law efficiency (COP) is decreasing and entrainment ratio is also increasing.

Table- 4(c): Variation with evaporator temperature of ejector fitted vapour compression refrigeration system using R1243zf

T_EVA[K]	COP	a	b
268	0.3692	1.359	0.6478
273	0.4888	1.398	0.6952
278	0.6556	1.441	0.7483
283	0.9115	1.489	0.8092

Table-4(c) shows the variation of first law efficiency (COP) of ejector coupled vapour compression refrigeration system using HFO-1243zf with variation of evaporator temperature and it was observed that evaporator temperature is increasing, first law efficiency (COP) is decreasing and entrainment ratio is also increasing.

4. Conclusions

Following conclusions were drawn from this investigation.

- Maximum first law efficiency of ejector coupled vapour compression refrigeration system using HFO-1234ze(Z) while maximum second law efficiency was observed by using R-1225ye(Z)
- maximum Entrainment Ratio of ejector coupled vapour compression refrigeration system was observed by using HFO-1336mzz(Z) and minimum by using R125.
- Maximum Effectiveness of ejector coupled vapour compression refrigeration system was observed by using R-1225ye(Z)
- When boiler temperature of ejector coupled vapour compression refrigeration system is increasing, second law efficiency (Exergetic efficiency) is decreasing.
- When boiler temperature of ejector coupled vapour compression refrigeration system is increasing, compression Ratio of ejector coupled vapour compression refrigeration system is decreasing.
- When condenser temperature of ejector coupled vapour compression refrigeration system is increasing, second law efficiency (exergetic efficiency) is increasing. Similarly, entrainment ratio is also increasing while exergy destruction ratio is decreasing
- condenser temperature of ejector coupled vapour compression refrigeration system using R1243zf is increasing, first law efficiency (COP) is decreasing.

- When evaporator temperature of ejector coupled vapour compression refrigeration system using HFO-1243zf is increasing, first law efficiency (COP) is decreasing and entrainment ratio is also increasing.
- When evaporator temperature of ejector coupled vapour compression refrigeration system using HFO-1243zf is increasing, second law efficiency (exergetic efficiency) is decreasing.

References

[1] Chen J, et.al.{2014] screening of working fluids for the ejector refrigeration system, International journal of refrigeration, Vol-47, page-1-14

[2] Chen J, et.al. [2015] conventional and advanced exergy analysis of an ejector refrigeration system, Applied energy, Vol-144, page-139-151

[3] F Memet[1] and A Preda, [2015] An analysis of the performance of an ejector refrigeration cycle working with R134a IOP Conference Series: Materials Science and Engineering, Volume 95, Modern Technologies in Industrial Engineering (Mod Tech-2015), page-17–20 June 2015, Mammalia, Romania.

[4] Li, H. et.al. [2014] performance characteristics of R1234yf ejector expansion refrigeration cycle, Applied Energy, Vol-121, page-96-103

[5] Trung Kien Nguyen & Chi Hiep Le [2020] Thermodynamic analysis of an ejector–vapour compressor cascade refrigeration system Journal of Thermal Analysis and Calorimetry (2020), Published: 09 April 2020.

[6] M Yari, M Sirousazar[2007]Performance analysis of the ejector-vapour compression refrigeration cycle, Institute of Mechanical Engineers, https://doi.org/10.1243/09576509JPE484

[7] Yapici, R and Ersoy H.K. [2005] performance characteristics of the ejector expansion device, Energy conversion and Management, Vol-46, issue-18-19, page-3117-3135.

[8] Yin-Hai Zhu & Peixue Jiang [2012] Hybrid vapor compression refrigeration system with an integrated ejector cooling cycle International Journal of Refrigeration 35(1) · January 2012.

[9] Disawas S and Wongwishes S .[2004] experimental investigation on the performance of the refrigeration cycle using two phase ejector as expansion device, international journal of refrigeration, Vol-27, issue-6, page-587-594

[10] Chaiwongsa,P and Wongwishes S .[2007] Effect of throat diameters of the ejector on the performance of the refrigeration cycle using two phase ejector as an expansion device, International journal of refrigeration, Vol-30, issue-4, page-601-608

[11] Da-Wen Sun[1998] Evaluation of a combined ejector–vapour-compression refrigeration system, International Journal of energy research Volume22, Issue4, Pages 333-342,

[12] Lucus C and Kochler J [2020] experimental investigation of the COP improvement of a refrigeration cycle by using of an ejector, international journal of refrigeration, Vol-35, issue-6, page-1595-1603.

[13] Kornhauser A.A [1990], the use of an ejector as a refrigerant expander, proceedings of the USN/ IIR- Purdue refrigeration conference, page-10-19.

[14] Nehdi, E et.al., [2007], performance analysis of the vapour compression cycle using ejector as an expander, International journal of Energy Research, Vol-31, No.4, page-364-375

[15] Bilr N, and Ersoy H.K., [2009] Performance improvement of the vapour compression refrigeration cycle by a two phase constant area ejector, International Journal of Energy Research, Vol-33, No.5, page-469-480.

[16] Saleh B., [2016] performance analysis and working fluid selection for ejector refrigeration cycle, Applied Thermal Engineering, Vol-107, page-114-124.

[17] Sarkar J, et.al. [2010] geometric parameter optimization of ejector-expansion refrigeration cycle with natural refrigerants, International Journal of Energy Research, Vol-34, No.1, page-84-94.

Cite this article as: R.S. Mishra, Thermodynamic performances of ejector refrigeration systems (ERS) using low GWP ecofriendly HFO refrigerants, International Journal of Research in Engineering and Innovation Vol-4, Issue-5 (2020), 236-240. *https://doi.org/10.36037/IJREI.2020.4501*.

Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Two Different Approaches for Consistency of Intuitionistic Multiplicative Preference Relation using Directed Graph

Mamata Sahu
Delhi Technological University
Anjana Gupta (anjanagupta@dce.ac.in)
Delhi Technological University

Research Article

Keywords: Isomorphism, consistency, Directed graph, cycle, Intuitionistic multiplicative preference relation, asymmetric multiplicative preference relation

Posted Date: September 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-847781/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Noname manuscript No.
(will be inserted by the editor)

Two different Approaches for consistency of Intuitionistic Multiplicative Preference Relation using Directed Graph

Mamata Sahu[1] · **Anjana Gupta**[2,*]

Received: date / Accepted: date

Abstract Consistency is an important issue that causes wide public concern of decision-makers in the decision-making process. The lack of consistency in preference relations results in a vague solution. The main goal of this paper is to achieve the consistent intuitionistic multiplicative preference relation using a graphical approach. We have proposed two different characterizations of the consistency for intuitionistic multiplicative preference relation(IMPR). In the first approaches, we propose an algorithm to achieve the consistency of IMPR by using the cycles of various length in a directed graph. The second approach proves isomorphism between the set of IMPRs and the set of asymmetric multiplicative preference relations. That result is explored to use the methodologies developed for asymmetric multiplicative preference relations to the case of IMPRs and achieve the consistency of asymmetric multiplicative preference relation using a directed graph. Sometimes the decision maker may not be able to provide the complete relation. So the above-said method is applied for an incomplete IMPR also, here consistency plays an important role. The examples are provided to illustrate both the methods in all cases.

Keywords Isomorphism, consistency, Directed graph, cycle, Intuitionistic multiplicative preference relation, asymmetric multiplicative preference relation.

1 Introduction

In decision-making problems, preference relation is one of the most useful tools used by decision makers(DM's), as it can express their views over the alternatives/criteria. In the past decades, plenty of research has been done about preference relations. Among these efforts, the fuzzy preference relations (Orlovsky, 1978) and the multiplicative preference relations (Saaty, 1980) are two types of preference relations which are widely studied by many researchers. Additionally, the multiplicative preference relations have been deeply studied and successfully applied in the analytic hierarchy process (AHP)(Saaty, 1980). In fuzzy preference relation(FPR), expert provides the preference information using the 0-1 scale, and a multiplicative preference relation (MPR) (Saaty, 1980) in which the (1/9)-9 ratio scale is used to measure the intensity of the pairwise comparison of alternatives.

In these preference relations, the essential elements provide the degree that an alternative is prior to the other. However, sometimes DM's may also need to give the degree that an alternative is not prior to the other in some practical problems. In these cases, DM not sure about the preference information namely, hesitation degree (uncertainty degree). Intuitionistic fuzzy preference relations (IFPRs) (Xu 2007b) and intuitionistic multiplicative preference relations (IMPRs) (Xia et al., 2013) are proposed to indicate the

[1] Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India
E-mail: munmun.sahu@gmail.com ·
[2,*] Corresponding author, Department of Applied Mathematics, Delhi Technological University, Delhi 110042, India
E-mail: anjanagupta@dce.ac.in

positive information x_i is preferred to x_j, the degree of negative details x_i is not preferred to x_j and simultaneously the degree that cannot determine by the DM.

The consistency, a fundamental property of preference relations, in the decision-making process has drawn much attention in many decision-making fields. The level of agreement among the preference data given by the individual DMs' measures the consistency (Xia, Xu and Chen 2013). Based on the multiplicative consistency property, Xia et al.(Xia, Xu and Chen 2013) discussed the consistency of FPRs and developed an algorithm to improve the consistency level of FPRs. Herrera-Viedma et al. (2004) developed a method for constructing consistent FPRs and using the additive transitivity property of the FPRs they proposed a characterization of the consistency property. Chiclana et al. (2009) put forwarded a functional equation to demonstrate the "cardinal consistency in the strength of preferences" of reciprocal preference relations (RPRs).

Different type of reciprocal preference relations surveyed by Li et al. (2019) using the additive consistency. Furthermore, they research some fusion methods and consistency-driven decision making methods.

Capuano et al. (2018) proposed a new influence-guided GDM model and also the effects of social influence are generally relevant to cases when the presences of expertŠs knowledge are not known. Then use social influence network (SIN) to calculate the missing preferences and develop them by simulating the effect of the expert's interpersonal influence before aggregating them to select the best alternative. Ureña et al.(2015) reviewed and analyzed the state-of-the-art research efforts on group decision making from the perspective of the estimation of missing preferences using different types of preference relations. These estimation techniques mainly use the properties of additive or multiplicative consistency for the missing options among the known options, as well as to increase the overall consistency. In many cases it is the consensus of experts.

To improve the consistency of a MPR Xu and Wei (1999) proposed an algorithm to derive a definite reciprocal matrix with acceptable consistency. Ergu et al. (2011) developed a method to measure the consistency for MPRs.

Xu et al. (2013) first proposed the ordinal consistency of a FPR, then the authors proposed how to find the cycles in the direct graphs. In addition, they also proposed the process of calculating the ordinal consistency index and positioning each cycle, as well as the process of finding inconsistent judgments in the preference relationship. To repair the inconsistency in FPR, they developed an algorithm to find and eliminate the 3 cycles on the graph. Xu, Ma and Herrera (2019) first proposed the ordinal consistency of incomplete fuzzy linguistic preference relations. Also, they developed one novel algorithm to judge whether an incomplete FLPR is ordinaly consistent. To rectify the inconsistency of an incomplete FLPR they developed an another algorithm and improve it for ordinally consistent. Using the proposed algorithms they identify and rectify the ordinal inconsistency of the complete FLPRs. Xu et al. (2018) studied both the ordinal and multiplicative consistencies for FPRs. To measure the degree of ordinal consistency for FPRs, they introduced a new ordinal consistency index, and developed a new multiplicative inconsistency identification and modification method for FPR. Xu et al. (2021) developed an algorithms to rectify the multiplicative and ordinal inconsistencies for FPRs. Xu and Herrera (2019) proposed a graphical method to visualize and rectify different inconsistencies for FPRs. Using, ordinal and cardinal consistency, on a two-dimensional plan, they draw Gower plots to visualize the ordinal and cardinal inconsistencies and they formulated some optimization models to rectify different types of inconsistences, respectively. Xu, Herrera and Wang (2016) proposed a distance-based methodology is proposed to deal with ordinal and additive inconsistencies for FPR.

A large number of researchers are committed to using IFPR to make decisions in uncertain environments (Xu and Liao, 2015). The definitions of feasible-region-based consistency and priority derivation methods were proposed for IFPRs(Behret 2014, Gong, et al., 2011, Gong et al. 2009).

Jiang et al. (2015) discussed the consistency and acceptably consistent of an IMPR. Based on it, two approaches were developed to complement all missing elements of incomplete IMPRs. Ren et al.

(2016) defined IMWGA operator and proposed an iterative process to adjust the inconsistent IMPRs into an acceptably consistent one and repair the maximum in conformity gradually, which makes fewer changes to the DMs' original opinions. Also, they provide an adjustment process to restore and improves the consistency of inconsistent IMPR. Zhang and Guo (2017) developed a linear programming-based algorithm to check and adjust the flexibility of an IMPR. As well, Zhang and Guo (2017) discuss the relationships between an IMPR and a normalized intuitionistic multiplicative weight vector and developed two group decision methods based on complete and incomplete IMPRs, respectively. Zhang and Pedrycz (2018) manage the unity and consensus of IMPR, and establish a consistent and consensus-based method through various proposed goal programming models to address group decision making (GDM) with IMPR. In this paper, our work focuses on only IMPRs.

1.1 Focus of the present study

In 1995, Nishizawa (Nishizawa 1995) proposed an algorithm checking for the consistency of MPRs by using the cycle of a directed graph. After that Nishizawa (1996), proposed two algorithms to find the various cycles of odd and even length in the incomplete directed graph. Taking inspiration from the Nishizawa (1996), in this paper, we have developed two approaches to improve the consistency of IMPR. In the first approach, we have formed an ordered-pair binary matrix from the IMPR and which is split into two binary arrays and using Nishizawa (1996) algorithm, we have checked the consistency of IMPR.

The subsequent methodology, of this paper, is to demonstrate the isomorphism between the set of IMPR and the set of asymmetric multiplicative preference relations. This outcome can accordingly be abused to utilize techniques created for MPRs to the instance of IMPR and, eventually, to expand the utilization of IMPR in decision making and to overcome the computation of above mentioned complexity. In other words, this result will allow taking advantage of mature and well-defined methodologies developed for MPRs while controlling the flexibility of IMPR to model vagueness/uncertainty.

As shown in work mentioned above, the preference relations are generally with complete information. In the decision-making process, DM required $\frac{n(n-1)}{2}$ judgement at each level to present an entire preference relation, and when n is large, it becomes a tedious task. Here and there, DM may not yet have a good understanding of a particular question, and thus he/she is unable to make a direct comparison between every two objects. Therefore it is sometimes necessary to allow the DM to skip some dubious comparisons flexibly. Thus, due to lack of time and busy schedule of the DM sometimes incomplete preference relations are obtained. In this case, the whole process may slow down. Therefore in this paper, both the approaches are applied in incomplete IMPR scenario.

Indeed, an issue that can be addressed using the mentioned equivalence is the presence of incomplete IMPR in the decision-making process. Using directed graph (Nishizawa, 1996) we check the consistency of asymmetric multiplicative preference relation. A new section has been added that compare the two approaches and improve the consistency of IMPR.

1.2 Organization of the paper

The rest of the paper is arranged as follows. Section 2 represent some basic concepts of MPRs and IMPRs. We have developed two approaches to check the consistency of IMPR given in section 3 and 4. Section 3 demonstrates the first approach in which, we have developed an algorithm that improves the consistency of IMPRs and is discussed with some examples. This section includes incomplete IMPRs also. The second approach in section 4, the set of IMPR and the set of asymmetric multiplicative preference relations has proved mathematically isomorphic. Numerical examples are also given in this section for both complete or incomplete intuitionistic scenario. In this section, we have given a comparison between

the two approaches and improve the consistency of IMPR. Section 5 gives a comparative analysis of the proposed methodology relevant to the context. Concluding remarks are given in the last section.

2 Preliminaries

Before introducing the basic concept of MPR and IMPR, first we have discussed the analytical hierarchy process.

Generally, in the decision-making problems, Analytical hierarchy process (AHP) is applied with a finite number of alternatives. There are limited arrangement choices $X = \{x_1, \cdots, x_n\}$ and there is one objective in the decision process, from which the DM, is typically approached to choose the best one. Generally, in the decision-making process, DM is closest to assigning a score to each of the other options and then selecting the option with the most extreme value. That is, for a given arrangement of choices, $X = \{x_1, \cdots, x_n\}$, the DM ought to give a weight vector of

$$w = (w_1, \cdots, w_n)^T$$

where, w_i is a worth that correctly determine the score of option x_i. The weight vector w_i is more prominent, that better the i-th alternative. As demonstrated by the theory of Fishburn (Fishburn 1979), the rule is that if and only if $w_i > w_j$, the alternative x_i is preferred to the alternative x_j . The components w_i of the weight vectors are called priorities, or weights, of the alternatives x_i, and on the other hand weight vectors are used for ranking of alternatives. In particular, $w = (0.1, 0.4, 0.2, 0.3)^T$ infers $x_2 \succ x_4 \succ x_3 \succ x_1$ where $x_i \succ x_j$ implies that option x_i is preferred to x_j. Along these lines, the decisions process appears to be simple, but when complexity increases, it turns into a hard errand. As we will see, complexity increases affectionately intertwined with the quantity of choices furthermore, criteria. A DM could continue running into bothers when mentioned to exhibit a rating as a numerical vector for an enormous number of options. Some times, it happens that we can not settle on a few other options. Far and away more terrible, we can not choose what's more, in the end, understand that it was not the best choice. There is an effective way to use pairwise comparisons to overcome this problem. The explanation for this is that it allows DMs to consider two options at the same time. Afterwards, the technique is to decompose the original problem into many smaller sub-problems and manage the latter. Formally, from the set of pairwise comparisons, the pairwise comparison matrix $A = (a_{ij})_{n\times n}$, can be structured as follows

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \tag{1}$$

with $a_{ij} > 0$ communicating the level of inclination of x_i to x_j. More precisely, according to Saaty's theory, each entry must approximate the ratio between the two weights

$$a_{ij} \approx \frac{w_i}{w_j}$$

This implies, on the off chance that the entries precisely speak to proportions between weights, at that point the comparison matrix A can be communicated in the accompanying structure,

$$A = \left(\frac{w_i}{w_j}\right)_{n\times n} = \begin{pmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \cdots & \frac{w_1}{w_n} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \cdots & \frac{w_2}{w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{w_n}{w_1} & \frac{w_n}{w_2} & \cdots & \frac{w_n}{w_n} \end{pmatrix} \tag{2}$$

Table 1 Random Index

n	1	2	3	4	5	6	7	8	9	10
R.I	0	0	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

Note that, when we represent 1 and consider the 2, one condition of multiplicative reciprocity $a_{ij} = \frac{1}{a_{ji}} \forall\, i, j$ holds. The matrix A can be improved and rewritten as,

$$A = \begin{pmatrix} 1 & a_{12} & \cdots & a_{1n} \\ \frac{1}{a_{12}} & 1 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{1n}} & \frac{1}{a_{2n}} & \cdots & 1 \end{pmatrix}$$

In other words, the rearrangement structure of this form of pairwise comparison matrix comes from assumptions. Namely, when the option x_1 is twice as good as x_2, then we can infer that x_2 is $1/2$ of x_1. We are currently continuing this example, and imagine that the pairwise comparison matrix of a city has X set to features and x_1, x_2, x_3 are appended to the rows and columns of the matrix.

$$A = \begin{pmatrix} 1 & 3 & 6 \\ 1/3 & 1 & 2 \\ 1/6 & 1/2 & 1 \end{pmatrix}$$

In particular, the element a_{12} of this matrix A says that the alternative x_1 is considered 3 times better than x_2, i.e. $a_{12} = 3$.

There are many ways to find the priority vector w from the pairwise comparison matrix. In this example, it can be checked that the condition $a_{ij} = w_i/w_j, \forall i, j$ is satisfied, for instance, the sum of the following vector and its components is 1 i.e. $w = (6/9, 2/9, 1/9)^T$.

Several basic concepts related to MPR and IMPR are introduced, and accordingly the concept of incomplete IMPR is defined.

Definition 1 (Saaty, 1980) MPR $R = (r_{ij})_{n\times n}$ is called multiplicative consistent if it satisfied the multiplicative transitivity property

$$r_{ij} = r_{ik} r_{kj},\ \forall\, i, j, k = 1, 2, \cdots, n. \tag{3}$$

In this case, there exit a positive vector $w = (w_1, w_2, \cdots, w_n)^T$ such that $r_{ij} = \frac{w_i}{w_j}$ satisfying the conditions $\sum_{i=1}^{n} w_i = 1,\ w_i \geq 0,\ i, j \in N$, where w_i is the priority weight.

In general cases, equation 3 does not always hold, therefore the MPRs are inconsistent. Saaty (1980) gave a method to measure the consistency degree of an MPR for an inconsistent MPR R, by the concept of consistency index (CI) and consistency ratio (CR).

$$CI = \frac{\lambda_{max} - n}{n - 1},\ \ CR = \frac{CI}{RI_n} \tag{4}$$

where λ_{max} and n are the largest eigen value and the order of the MPR R respectively. RI_n is the random index that dependents on the orders of the preference relation matrices given in the table 1.

Definition 2 An MPR R is said to be acceptably consistent if $CR \leq 0.1$, else R is announced not acceptably consistent (or simply inconsistent), and returned to an expert to have a re-look at the preferences evaluations.

Expressly, if the value of the consistency ratio (CR) < 0.1, the comparison matrix is acceptable. There is no standard for matrix consistency. According to some experiments and experiences, the degree of inconsistency is 10% of the degree of inconsistency still acceptable.

Nishizawa, in 1995 proposed an algorithm for the consistency of MPRs by using the cycle of a directed graph. Later Nishizawa (1996) proposed two algorithms to find the cycles of various odd and even length using incomplete directed graph. For finding cycles, vertex matrix that is denoted by V of order $n \times n$ is needed, among them, n is the number of vertices, corresponding to the element (i, j), which is a directed graph determined by $V(i, j)$. If one points i is connected to another point j by an arrow, say "$i \to j$", then $V(i, j) = 1$ otherwise 0. Pairwise comparison data is represented by θ or $1/\theta$, where θ is the parameter whose binary AHP value is greater than 1 (Saaty, 1977). Nishizawa proposed two algorithms (Nishizawa, 1996) for even and odd length cycle in the incomplete directed graph. From the result Nishizawa (1996) one can easily judge the consistency of comparison matrix. For the sake of convenience, following the two algorithms are given by Nishizawa (1996) for finding the cycle of odd and even length.

Algorithm 1 (Nishizawa, 1996)(Finding even length cycle) Nishizawa (1996) discussed the cycle of even length $2m$ $(m = 2, 3, ..)$, of n elements be the form of $(i_0\ i_1\ \cdots\ i_{m-1}\ j_0\ j_1; \cdots\ j_{m-1})$.

1. Find the vertex matrix V and V^m.
2. Find pair of numbers i_0 and j_0 with $V^m(i_0, j_0) > 0$ and $V^m(j_0, i_0) > 0$, where $i_0 < j_0$. Without these pairs, there are no cycles that are $2m$ in length.
3. Find a pairs of numbers $i_\alpha (> i_0)$ and $j_\alpha (> i_0)$ under the conditions $V(i_{\alpha-1}, i_\alpha) = 1,\ V^m(i_\alpha, j_\alpha) > 0$, $V^m(j_\alpha, i_\alpha) > 0$, and $V(j_{\alpha-1}, j_\alpha) = 1$ for each $\alpha = 1 \sim m - 1$.
4. Confirm $V(i_{m-1}, j_0) = 1$ and $V(j_{m-1}, i_0) = 1$.
5. Repeat step 2 to 4 for all pairs of i_0, j_0 where $i_0 < j_0$.

Algorithm 2 (Nishizawa, 1996)(Finding odd length cycle) Nishizawa (1996) discussed the cycle of odd length $2m-1$ $(m = 2, 3, ..)$, of n elements be the form of $(i_0\ i_2\ \cdots\ i_{m-3}\ j_{m-2}\ \cdots\ j_3\ j_1\ k\ i_1\ i_3 \cdots\ i_{m-2}\ j_{m-3}$ $\cdots j_2\ j_0)$ if m is odd, and $(i_0\ i_2\ \cdots\ i_{m-2}\ j_{m-3}\ \cdots\ j_3\ j_1\ k\ i_1\ i_3 \cdots\ i_{m-3}\ j_{m-2} \cdots j_2\ j_0)$ if m is even.

1. Find the vertex matric V, V^{m-1}, and V^m.
2. Find pair of numbers i_0, k and j_0 with $V^{m-1}(i_0, k) > 0$ and $V^m(k, i_0) > 0$, $V^{m-1}(k, j_0) > 0$, $V^m(j_0, k) > 0$, and $V(j_0, i_0) = 1$, where $i_0 < k,\ i_0 < j_0$. Without these pairs, there are no cycles of length $2m - 1$.
3. Find pair of numbers $i_1 (> i_0)$ and $j_1 (> i_0)$ with $V^m(i_0, i_1) > 0$, $V^{m-1}(i_1, i_0) > 0$, $V(k, i_1) = 1$, $V^{m-1}(j_0, j_1) > 0$, $V^m(j_1, j_0) > 0$, and $V(j_1, k) = 1$.
4. Find a pairs of numbers $i_\alpha (> i_0)$ and $j_\alpha (> i_0)$ under the conditions $V^m(i_{\alpha-1}, i_\alpha) > 0,\ V^{m-1}(i_\alpha, i_{\alpha-1}) > 0$, $V(i_{\alpha-2}, i_\alpha) = 1$, $V^{m-1}(j_{\alpha-1}, j_\alpha) > 0$, $V^m(j_\alpha, j_{\alpha-1}) > 0$, and $V(j_\alpha, j_{\alpha-2}) = 1$ for each $\alpha = 2 \sim m - 2$.
5. Confirm $V(i_{m-2}, j_{m-3}) = 1$ and $V(i_{m-3}, j_{m-2}) = 1$.
6. Repeat step 2 to 4 for all pairs of i_0, k, j_0 where $i_0 < k,\ i_0 < j_0$.

If no cycles have found in the directed graph, then the comparison matrix is consistent. The comparison matrix is inconsistent if at least one cycle is found in the incomplete directed graph. In case of inconsistency, they find the minimum covering sets (Nishizawa, 1995) among the cycles, and then they eliminate the path of cycles such as the comparison matrix is consistent.

In this paper, we have discussed two approaches for checking the consistency of IMPRs. Here we are extending the above method of Nishizawa (1996) in IMPR scenario. For this, we recall some basic concept of IMPRs, incomplete IMPRs, and the consistency property. Xia et al. (2013) extended MPR into IMPR.

Definition 3 (Xia et al. 2013)Let $\tilde{R} = [\tilde{r}_{ij}(x_i, x_j)]_{n\times n}$ be an intuitionistic multiplicative preference relation (IMPR), where $\tilde{r}_{ij}(x_i, x_j) = (\wp(x_i, x_j), \Im(x_i, x_j)),\ i, j \in N$, is an intuitionistic multiplicative

number (IMN), and $\wp(x_i, x_j)$ indicates certainty degree to which x_i is preferred to x_j and $\Im(x_i, x_j)$ is the certainty degree to which x_i is not preferred to x_j, and it satisfy the following characteristics:

$$1/9 \leq \wp(x_i, x_j), \quad \Im(x_i, x_j) \leq 9, \wp(x_i, x_j) = \Im(x_j, x_i), \Im(x_i, x_j) = \wp(x_j, x_i),$$
$$\wp(x_i, x_i) = \Im(x_i, x_i) = 1, \qquad 0 < \wp(x_i, x_j)\Im(x_i, x_j) \leq 1, \quad \forall\, i, j \in N.$$

If $\wp(x_i, x_j).\Im(x_i, x_j) = 1,\ \forall i, j \in N$, the IMPR is equivalent to an MPR. The value of $\wp(x_i, x_j)$, $\Im(x_i, x_j)$ are not arbitrary. For easy understanding $\wp(x_i, x_j)$ and $\Im(x_i, x_j)$ are denoted by $\wp_{ij}$ and $\Im_{ij}$ respectively.

Xu (2013) has given the consistent property of IMPR $\tilde{R} = (\tilde{r}_{ij})_{n\times n} = (\wp_{ij}, \Im_{ij})_{n\times n}$ based on the transitive property:

$$(\wp_{ij}, \Im_{ij}) = (\wp_{ik}\wp_{kj}, \Im_{ik}\Im_{kj}), \text{ for all } i, j, k \in N \text{ and } \ i \leq k \leq j. \tag{5}$$

It is to note that the equation (5) is restricted for the condition $i \leq k \leq j$, while the transitive property of a MPR is unconstrained which satisfies for all $i, j, k \in N$. If the equation (5) is utilized to check the consistency of an IMPR for all $i, j, k \in N$, the transitivity and consistency properties some times do not hold. This is because that when 'k' comes from the row of the lower triangular matrix, the equation does not hold. For example:

$$\begin{pmatrix} (1,1) & (1/2,1) & (1,1/2) \\ (1,1/2) & (1,1) & (2,1/2) \\ (1/2,1) & (1/2,2) & (1,1) \end{pmatrix}_{3\times 3}$$

is a consistent IMPR given by Xu, 2013. Jiang et al. (2015) relax the condition $i \leq k \leq j$, it is follow that $a_{23} = (\wp_{21}\wp_{13}, \Im_{21}\Im_{13}) = (1, 1/4)$. But $a_{23} = (2, 1/2) \neq (1, 1/4)$. To over come this type of transitivity limitation, Jiang et al. (2015) proposed a more general consistency property of an IMPR split into two MPRs by using the formula

$$c_{ij} = \begin{cases} \wp_{ij} & i < j \\ 1 & i = j \\ 1/\Im_{ij} & i > j \end{cases} \text{ and } d_{ij} = \begin{cases} \Im_{ij} & i < j \\ 1 & i = j \\ 1/\wp_{ij} & i > j\,. \end{cases} \tag{6}$$

where the MPRs $C = (c_{ij})_{n\times n}$ and $D = (d_{ij})_{n\times n}$ are preferred and non-preferred information matrix given by the DM with respect to the alternative x_i over x_j. Based on the above concept, Jiang et al. (2015) defined the consistent IMPR.

Jiang et al. (2015) relax the condition $i \leq k \leq j$ and proposed a more general consistency property of IMPR to overcome this type of transitivity limitation.

Definition 4 (Jiang et al. 2015) An IMPR $\tilde{R} = (\tilde{r}_{ij})_{n\times n}$ is said to be consistent if both MPRs C and D obtained by splitting the IMPR $\tilde{R}$ are consistent such that

$$c_{ij} = c_{ik}c_{kj},\ d_{ij} = d_{ik}d_{kj} \ \ \forall i, j, k \in N \tag{7}$$

In this paper we have developed two different approaches to check the consistency of IMPR and incomplete IMPRs which are discussed in section 3 and 4 respectively.

3 Consistency for IMPR and Incomplete IMPR

The absence of consistency in decision making with preference relations is a big challenge to bring about conflicting conclusions. Numerous strategies on consistency measure and improvement of preference relations with various structures have been exhibited progressively. This section checks the consistency of both IMPR and incomplete IMPR.

Let $\tilde{R}$ be the IMPR,

$$\tilde{R} = \begin{pmatrix} (\wp_{11}, \Im_{11}) & (\wp_{12}, \Im_{12}) & \cdots & (\wp_{1i}, \Im_{1i}) & \cdots & (\wp_{1n}, \Im_{1n}) \\ (\wp_{21}, \Im_{21}) & (\wp_{22}, \Im_{22}) & \cdots & (\wp_{2i}, \Im_{2i}) & \cdots & (\wp_{2n}, \Im_{2n}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\wp_{i1}, \Im_{i1}) & (\wp_{i2}, \Im_{i2}) & \cdots & (\wp_{ii}, \Im_{ii}) & \cdots & (\wp_{in}, \Im_{in}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\wp_{n1}, \Im_{n1}) & (\wp_{n2}, \Im_{n2}) & \cdots & (\wp_{ni}, \Im_{ni}) & \cdots & (\wp_{nn}, \Im_{nn}) \end{pmatrix}$$

We define a ordered pair vertices matrix $V = \{v_{ij}\} = \{(a, b)\}$ as follows, where v_{ij} is an ordered pair (a, b) such that

$$\begin{cases} \text{For } \ i < j \begin{cases} if \ \wp_{ij} > 1 & \text{then } a = 1, \text{ otherwise } a = 0 \\ if \ \Im_{ij} > 1 & \text{then } b = 1, \text{ otherwise } b = 0 \end{cases} \\ \text{For } \ i > j \begin{cases} if \ \frac{1}{\Im_{ij}} > 1 & \text{then } a = 1, \text{ otherwise } a = 0 \\ if \ \frac{1}{\wp_{ij}} > 1 & \text{then } b = 1, \text{ otherwise } b = 0 \end{cases} \\ \text{For } i = j \text{ both } a = 0 \text{ and } b = 0. \end{cases} \tag{8}$$

To illustrate this consider the following IMPR

$$\begin{pmatrix} (1,1) & (1/2, 2/3) & (1/5, 5) \\ (2/3, 1/2) & (1,1) & (4/5, 3/4) \\ (5, 1/5) & (3/4, 4/5) & (1,1) \end{pmatrix}$$

Using the binary order paired vertex matrix V is given below

$$V = \begin{pmatrix} (0,0) & (0,0) & (0,1) \\ (1,1) & (0,0) & (0,0) \\ (1,0) & (1,1) & (0,0) \end{pmatrix}$$

Here, the binary order paired vertex matrix is split into two vertex matrix i.e lower vertex matrix V_L containing the lower element of order pair and upper vertex matrix V_U containing the upper element of order pair.

$$V_L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad V_U = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Using the concept of order pair, we have defined a new definition of consistent IMPR.

Definition 5 The IMPR is consistent if both the corresponding lower and upper vertex matrices are consistent.

In this section, we have applied the graphical approach of Nishizawa (1996) to check the consistency of both complete IMPRs and incomplete IMPRs scenario. We have developed an algorithm to illustrate the above method.

Algorithm 3 Step1: Let us consider an IMPR.
Step2: Order pair vertex matrix is obtained using equation 8. In case of incomplete IMPRs, missing elements are treated as zero in the vertex matrix.
Step3: The ordered pair vertex matrix is split into the lower vertex matrix and upper vertex matrix.
Step4: Then apply the algorithm 1 and 2 of section 2 for finding the cycle of even and odd length on both the vertex matrix.
Step5: If any cycle found then the vertex matrix is inconsistent, otherwise consistent.
Step6: IMPR is consistent if both the vertex lower and the upper matrix is consistent; otherwise matrix

is inconsistent.

Step7: If the IMPR matrix is inconsistent, then remove the minimum number of the path that covers the cycles.

An example illustrates the above-said method.

Example 1 Let us consider a decision making problem with four sets of alternatives $x_i, (i = 1, 2, 3, 4)$. The decision maker judge these four alternatives by pairwise comparison and provides his/her judgement as follows: $\tilde{r}_{12} = (1/2, 1/4)$, $\tilde{r}_{13} = (2, 1/8)$, $\tilde{r}_{14} = (2/3, 1/4)$, $\tilde{r}_{23} = (5, 1/7)$, $\tilde{r}_{24} = (7/5, 2/3)$, $\tilde{r}_{34} = (6, 1/7)$. The matrix representation is given by

$$\tilde{R}_1 = \begin{pmatrix} (1,1) & \left(\frac{1}{2},\frac{1}{4}\right) & \left(2,\frac{1}{8}\right) & \left(\frac{2}{3},\frac{1}{4}\right) \\ \left(\frac{1}{4},\frac{1}{2}\right) & (1,1) & \left(5,\frac{1}{7}\right) & \left(\frac{7}{5},\frac{2}{3}\right) \\ \left(\frac{1}{8},2\right) & \left(\frac{1}{7},5\right) & (1,1) & \left(6,\frac{1}{7}\right) \\ \left(\frac{1}{4},\frac{2}{3}\right) & \left(\frac{2}{3},\frac{7}{5}\right) & \left(\frac{1}{7},6\right) & (1,1) \end{pmatrix}$$

By using equation 8, the binary order pair vertex matrix is

$$V = \begin{pmatrix} (0,0) & (0,0) & (1,0) & (0,0) \\ (1,1) & (0,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,0) & (1,0) \\ (1,1) & (0,1) & (0,1) & (0,0) \end{pmatrix}$$

The lower and upper vertex matrices are

$$V_L = \begin{pmatrix} 0\,0\,1\,0 \\ 1\,0\,1\,1 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \end{pmatrix},\ V_U = \begin{pmatrix} 0\,0\,0\,0 \\ 1\,0\,0\,0 \\ 1\,1\,0\,0 \\ 1\,1\,1\,0 \end{pmatrix}$$

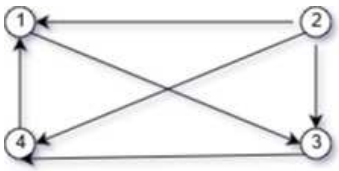


Fig. 1 Directed graph of V_L of example 1

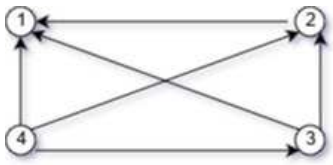


Fig. 2 Directed graph of V_U of example 1

There is only one cycle of length 3 is present in the lower vertex matrix V_L. Directly we can found the cycle from figure 1. To find cycles of length 3, we use step (2) to step (5) of odd-length algorithm (see algorithm 2), where $m = 2$. The format of a cycle of length 3 is $(i_0 - k - j_0)$.

To find the elements i_0, k, j_0, we need V_L and V_L^2. For this V_L^2 as follows:

$$V_L^2 = \begin{pmatrix} 0\,0\,0\,1 \\ 1\,0\,1\,1 \\ 1\,0\,0\,0 \\ 0\,0\,1\,0 \end{pmatrix}$$

At the starting we have $i_0 = 1$, $k = 3$, $j_0 = 4$, satisfying the step (2) of algorithm 2 of odd length cycle i.e $V_L(1,3) = 1$, $V_L^2(3,1) = 1$, $V_L(3,4) = 1$, $V_L^2(4,3) = 1$, and $V_L(4,1) = 1$. Then the cycle of length 3 is $(1 - 3 - 4)$. There is no other cycle available in V_L. Similarly there is no cycle of any length in figure 2. The lower vertex matrix V_L is inconsistent and the upper vertex matrix V_U is consistent. Therefore the

Table 2

Cycle	(1,3)	(1,4)	(3,4)
$(1-3-4)$	1	1	1

complete IMPR is inconsistent. Apply an extinguishing cycles algorithm based on minimum covering sets to provide the cause of inconsistency (see Nishizawa, 1995). The cycle-arc incidence matrix in Table 2. From table 2, we get the several pairs of edges which cover the cycle. we have to chose the pair eliminate the cycle. In this example if we will change the pair of vertex from any one of them from $(3,4)$ to $(4,3)$ in the original IMPR $\tilde{R}_1$, then the both the lower and upper vertex matrix are consistent. Then IMPR $\tilde{R}_1$ is also consistent.

Example 2 Let us consider a decision-making problem with seven sets of alternatives $x_i, (i = 1, 2, \cdots, 7)$. The DM judges these seven alternatives by pairwise comparison and provides his/her judgement. The matrix representation of the decision maker judgement is given by

$$\tilde{R}_2 = \begin{pmatrix} (1,1) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(7,\frac{1}{9}\right) & \left(3,\frac{1}{7}\right) & \left(\frac{5}{3},\frac{1}{7}\right) & \left(1,\frac{3}{5}\right) & \left(\frac{1}{4},\frac{5}{3}\right) \\ \left(\frac{1}{4},\frac{5}{3}\right) & (1,1) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(\frac{3}{5},1\right) & \left(\frac{1}{3},1\right) & \left(\frac{1}{4},3\right) \\ \left(\frac{1}{9},7\right) & \left(\frac{1}{4},\frac{5}{3}\right) & (1,1) & \left(1,\frac{3}{5}\right) & \left(\frac{1}{4},\frac{5}{3}\right) & \left(\frac{1}{3},3\right) & \left(\frac{1}{9},7\right) \\ \left(\frac{1}{7},3\right) & \left(\frac{1}{4},\frac{5}{3}\right) & \left(\frac{3}{5},1\right) & (1,1) & \left(\frac{3}{5},\frac{3}{5}\right) & \left(\frac{1}{3},3\right) & \left(\frac{1}{9},7\right) \\ \left(\frac{1}{7},\frac{5}{3}\right) & \left(1,\frac{3}{5}\right) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(\frac{3}{5},\frac{3}{5}\right) & (1,1) & \left(\frac{1}{3},3\right) & \left(\frac{1}{7},3\right) \\ \left(\frac{3}{5},1\right) & \left(1,\frac{1}{3}\right) & \left(3,\frac{1}{3}\right) & \left(3,\frac{1}{3}\right) & \left(3,\frac{1}{3}\right) & (1,1) & \left(\frac{1}{7},3\right) \\ \left(\frac{5}{3},\frac{1}{4}\right) & \left(3,\frac{1}{4}\right) & \left(7,\frac{1}{9}\right) & \left(7,\frac{1}{9}\right) & \left(3,\frac{1}{7}\right) & \left(3,\frac{1}{7}\right) & (1,1) \end{pmatrix}$$

Using 8, the binary order pair vertex matrix is given by

$$V = \begin{pmatrix} (0,0) & (1,0) & (1,0) & (1,0) & (1,0) & (0,0) & (0,1) \\ (0,1) & (0,0) & (1,0) & (1,0) & (0,0) & (0,0) & (0,1) \\ (0,1) & (0,1) & (0,0) & (0,0) & (0,1) & (0,1) & (0,1) \\ (0,1) & (0,1) & (1,1) & (0,0) & (0,0) & (0,1) & (0,1) \\ (0,1) & (1,0) & (1,0) & (1,1) & (0,0) & (0,1) & (0,1) \\ (0,1) & (1,0) & (1,0) & (1,0) & (1,0) & (0,0) & (0,1) \\ (1,0) & (1,0) & (1,0) & (1,0) & (1,0) & (1,0) & (0,0) \end{pmatrix}$$

The above vertex matrix split into two binary array.

$$V_L = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}, \quad V_U = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

By applying algorithm 1 and 2, there is no cycle found in V_L, and one cycle of length three, i.e., $(3-5-4)$ is located in V_U. In order to find the cycle of length 3 in V_U we use step (2) to step (5) of algorithm 2, where $m = 2$. The form of the cycle of length 3 is $(i_0 - k - j_0)$. To locate these element we

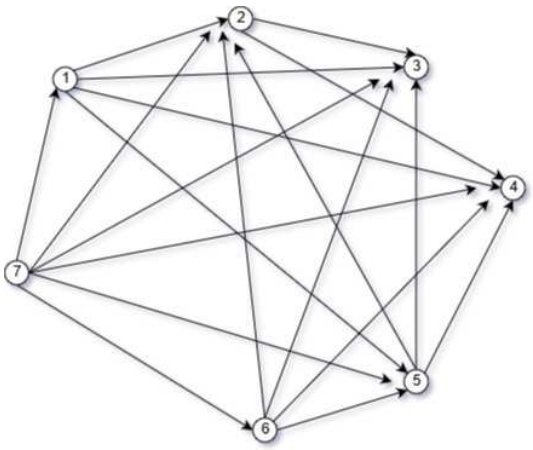


Fig. 3 Directed graph of V_L of example 2

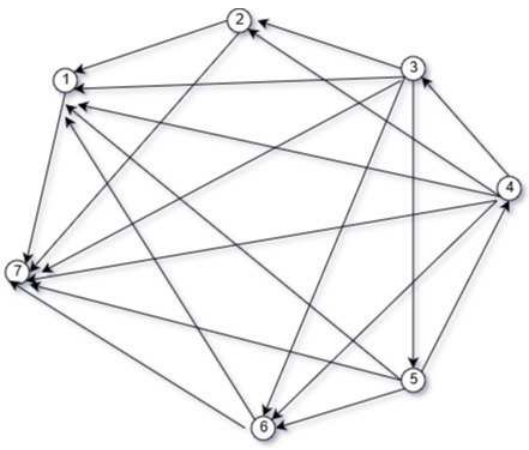


Fig. 4 Directed graph of V_U of example 2

Table 3

Cycle	(3,4)	(3,5)	(4,5)
$(3-5-4)$	1	1	1

need V_U and V_U^2. For this V_U^2 as follows:

$$V_U^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 1 & 0 & 1 & 4 \\ 3 & 1 & 0 & 0 & 1 & 1 & 4 \\ 2 & 1 & 1 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

At the starting we have $i_0 = 3$, $k = 5$, $j_0 = 4$, satisfying $V_U(3,5) = 1$, $V_U^2(5,3) = 1$, $V_U(5,4) = 1$, $V_U^2(4,5) = 1$, and $V_U(4,3) = 1$. Then the cycle of length 3 is $(3-5-4)$. Similarly there is no cycle of any length in figure 4. The lower vertex matrix V_L is consistent and the upper vertex matrix V_U is inconsistent. Therefore the complete IMPR is inconsistent.

We apply an extinguishing cycles algorithm based on minimum covering sets to provide the cause of inconsistency (see Nishizawa, 1995). The cycle-arc incidence matrix in Table 3. From table 3, we get the several pairs of edges which cover the cycle. we have to chose the pair eliminate the cycle. In this example if we will change the pair $(3,4)$ to $(4,3)$ in the original IMPR $\tilde{R}_2$, then both the lower and upper vertex matrix are consistent. Then IMPR $\tilde{R}_2$ is consistent.

The above said method is also applied in incomplete IMPRs scenario which is given in the following subsection.

3.1 **Consistency of incomplete IMPR**

In decision-making problem, the reality of the situation may prove that decision maker might not have a decent comprehension on a specific inquiry, and along these lines, he/she can not make an immediate correlation between two choices or criteria. Therefore, it is more fitting and adaptable to avoid a few similarities, and in those cases, the decision maker may like to express their judgments with incomplete preference relation. As indicated by the already talked about, the decision makers may not give exactly $\frac{n(n-1)}{2}$ judgments in practical decision making and the incomplete IMPR will be introduced. So, it is essential to investigate the incomplete preference relations as a useful tool in the decision-making

problem, and many research results have been developed. Herrera-Viedma et al. (2007) proposed the meaning of incomplete preference relation. The idea of IMPRs is reached out to the circumstances where the preference data given by DM is incomplete.

Definition 6 (Jiang et al. 2015) An IMPR $\tilde{r}_{ij} = (\wp_{ij}, \Im_{ij})_{n\times n}$ is called an incomplete IMPR if some elements in it are missing, and all available elements satisfy the characteristics of IMPR stated in Definition 3.

For an incomplete preference relation, it is very much essential for known elements to satisfy consistency. We are applying the same graphical approach in incomplete IMPRs scenario. For an incomplete comparison case, which includes unknown pairwise comparisons, we use the measure of inconsistency by the number of the cycle in the graph corresponding to known elements of the IMPR matrix. For the incomplete preference relation, if one wants to find the inconsistency, one condition is that the incomplete preference relation is acceptable, that is, in its directed graph, there is no isolated point.

Example 3 Let us consider a decision making problem with six sets of alternatives $x_i, (i = 1, 2, 3, 4, 5, 6)$. The decision maker judge these six alternatives by pairwise comparison and provides his/her judgement as follows: $\tilde{r}_{12} = (1/5, 2)$, $\tilde{r}_{13} = (5, 1/6)$, $\tilde{r}_{15} = (1/9, 8)$, $\tilde{r}_{16} = (2, 1/6)$, $\tilde{r}_{24} = (1/4, 1/5)$, $\tilde{r}_{25} = (1/2, 1/3)$, $\tilde{r}_{26} = (3, 1/5)$, $\tilde{r}_{35} = (6, 1/7)$, $\tilde{r}_{45} = (1/5, 4)$. The matrix representation is given by

$$\tilde{R}_3 = \begin{pmatrix} (1,1) & \left(\frac{1}{5},2\right) & \left(5,\frac{1}{6}\right) & (*,*) & \left(\frac{1}{9},8\right) & \left(2,\frac{1}{6}\right) \\ \left(2,\frac{1}{5}\right) & (1,1) & (*,*) & \left(\frac{1}{4},\frac{1}{5}\right) & \left(\frac{1}{2},\frac{1}{3}\right) & \left(3,\frac{1}{5}\right) \\ \left(\frac{1}{6},5\right) & (*,*) & (1,1) & (*,*) & \left(6,\frac{1}{7}\right) & (*,*) \\ (*,*) & \left(\frac{1}{5},\frac{1}{4}\right) & (*,*) & (1,1) & \left(\frac{1}{5},4\right) & (*,*) \\ \left(8,\frac{1}{9}\right) & \left(\frac{1}{3},\frac{1}{2}\right) & \left(\frac{1}{7},6\right) & \left(4,\frac{1}{5}\right) & (1,1) & (*,*) \\ \left(\frac{1}{6},2\right) & \left(\frac{1}{5},3\right) & (*,*) & (*,*) & (*,*) & (1,1) \end{pmatrix}$$

The order pair vertex matrix is obtained by using equation 8

$$V = \begin{pmatrix} (0,0) & (0,1) & (1,0) & (0,0) & (0,1) & (1,0) \\ (1,0) & (0,0) & (0,0) & (0,0) & (0,0) & (1,0) \\ (0,1) & (0,0) & (0,0) & (0,0) & (1,0) & (0,0) \\ (0,0) & (1,1) & (0,0) & (0,0) & (0,1) & (0,0) \\ (1,0) & (1,1) & (0,1) & (1,0) & (0,0) & (0,0) \\ (0,1) & (0,1) & (0,0) & (0,0) & (0,0) & (0,0) \end{pmatrix}$$

The lower and upper vertex matrices are given below

$$V_L = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \; V_U = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

To find the cycle of length 3 in the lower vertex matrix V_L, we need V_L and V_L^2. For this V_L^2 as follows:

$$V_L^2 = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

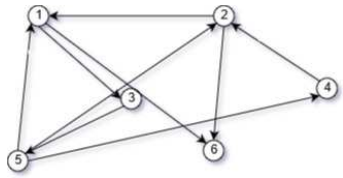


Fig. 5 Directed graph of V_L of example 3

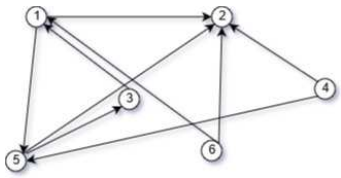


Fig. 6 Directed graph of V_U of example 3

Table 4

Cycle	(1,5)	(1,3)	(3,5)
$(1-3-5)$	1	1	1
$(1-5-3)$	1	1	1

At the starting we have $i_0 = 1$, $k = 3$, $j_0 = 5$, satisfying the step (2) of algorithm 2 i.e $V_L(1,3) = 1$, $V_L^2(3,1) = 1$, $V_L(3,5) = 1$, $V_L^2(5,3) = 1$, and $V_L(5,1) = 1$. Then the cycle of length 3 is $(1-3-5)$.

Similarly using algorithm 2, there is also one cycle of length 3 in V_U. Here also, for finding the cycle of length 3 in V_U, we need V_U and V_U^2. For this V_U^2 as follows:

$$V_U^2 = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Also we have $i_0 = 1$, $k = 5$, $j_0 = 3$, satisfying $V_U(1,5) = 1$, $V_U^2(5,1) = 1$, $V_U(5,3) = 1$, $V_U^2(3,5) = 1$, and $V_U(3,1) = 1$. Then the cycle of length 3 is $(1-5-3)$.

Both the lower and upper vertex matrix V_L, and V_U is inconsistent. Therefore the incomplete IMPR is inconsistent. To suggest the cause of inconsistency, we apply an algorithm for extinguishing cycles based on minimum covering sets. The cycle-arc incidence matrix in Table 4.

From table 4, we get several pairs of the edges which cover the cycle. We have to chose the couple to eliminate the cycle. In this example, if we will change the pair $(3,5)$ to $(5,3)$ in the original IMPR $\tilde{R}_4$, then both the lower and upper vertex matrix are consistent. Then incomplete IMPR $\tilde{R}_4$ is also consistent.

Example 4 Let us take an example of incomplete IMPRs of order 10×10 given as below

$$\tilde{R}_4 = \begin{bmatrix} (1,1) & * & (\frac{1}{2},\frac{1}{5}) & (\frac{1}{3},2) & (3,\frac{1}{6}) & * & * & (\frac{8}{3},\frac{1}{8}) & (\frac{1}{7},4) & (5,\frac{1}{5}) \\ * & (1,1) & (\frac{2}{3},\frac{6}{10}) & (\frac{1}{8},\frac{1}{2}) & (\frac{1}{4},1) & * & (\frac{1}{5},5) & * & (\frac{1}{3},3) & * \\ (\frac{1}{5},\frac{1}{2}) & (\frac{6}{10},\frac{2}{3}) & (1,1) & (\frac{3}{5},\frac{2}{3}) & * & (2,\frac{1}{2}) & (4,\frac{1}{8}) & (7,\frac{1}{8}) & (\frac{1}{8},5) & * \\ (2,\frac{1}{3}) & (\frac{1}{2},\frac{1}{8}) & (\frac{2}{3},\frac{3}{5}) & (1,1) & * & (3,\frac{1}{7}) & (5,\frac{1}{6}) & * & (2,\frac{1}{2}) & * \\ (\frac{1}{6},3) & (1,\frac{1}{4}) & * & * & (1,1) & (\frac{1}{8},7) & (8,\frac{1}{9}) & (\frac{1}{6},3) & * & (4,\frac{1}{5}) \\ * & * & (\frac{1}{2},2) & (\frac{1}{7},3) & (7,\frac{1}{8}) & (1,1) & * & * & (\frac{1}{5},4) & (2,\frac{1}{3}) \\ * & (5,\frac{1}{5}) & (\frac{1}{8},4) & (\frac{1}{6},5) & (\frac{1}{9},8) & * & (1,1) & (\frac{1}{9},6) & (\frac{1}{7},2) & (4,\frac{1}{8}) \\ (\frac{1}{8},\frac{8}{3}) & * & (\frac{1}{8},7) & * & (3,\frac{1}{6}) & * & (6,\frac{1}{9}) & (1,1) & * & (6,\frac{1}{6}) \\ (4,\frac{1}{7}) & (3,\frac{1}{3}) & (5,\frac{1}{8}) & (\frac{1}{2},2) & * & (4,\frac{1}{5}) & (2,\frac{1}{7}) & * & (1,1) & (\frac{1}{3},2) \\ (\frac{1}{5},5) & * & * & * & (\frac{1}{5},4) & (\frac{1}{3},2) & (\frac{1}{8},4) & (\frac{1}{6},6) & (2,\frac{1}{3}) & (1,1) \end{bmatrix}$$

The binary order pair vertex matrix is given by

$$V = \begin{bmatrix} (0,0) & (0,0) & (0,0) & (0,1) & (1,0) & (0,0) & (0,0) & (1,0) & (0,1) & (1,0) \\ (0,0) & (0,0) & (0,0) & (0,0) & (0,0) & (0,0) & (0,1) & (0,0) & (0,1) & (0,0) \\ (1,1) & (1,1) & (0,0) & (0,0) & (0,0) & (1,0) & (1,0) & (1,0) & (0,1) & (0,0) \\ (1,0) & (1,1) & (1,1) & (0,0) & (0,0) & (1,0) & (1,0) & (0,0) & (1,0) & (0,0) \\ (0,1) & (1,0) & (0,0) & (0,0) & (0,0) & (0,1) & (1,0) & (0,1) & (0,0) & (1,0) \\ (0,0) & (0,0) & (0,1) & (0,1) & (1,0) & (0,0) & (0,0) & (0,0) & (0,1) & (1,0) \\ (0,0) & (1,0) & (0,1) & (0,1) & (0,1) & (0,0) & (0,0) & (0,1) & (0,1) & (1,0) \\ (0,1) & (0,0) & (0,1) & (0,0) & (1,0) & (0,0) & (1,0) & (0,0) & (0,0) & (1,0) \\ (1,0) & (1,0) & (1,0) & (0,1) & (0,0) & (1,0) & (1,0) & (0,0) & (0,0) & (0,1) \\ (0,1) & (0,0) & (0,0) & (0,0) & (0,1) & (0,1) & (0,1) & (0,1) & (1,0) & (0,0) \end{bmatrix}$$

The lower and upper vertex matrices are given below:

$$V_L = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}, V_U = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

There are three cycles of length 3 is found in the lower vertex matrix i.e. $(1-10-9)$, $(6-10-9)$ and $(7-10-9)$. In order to find the cycle of length 3 in V_L we use step (2) to step (5) of the algorithm 2,

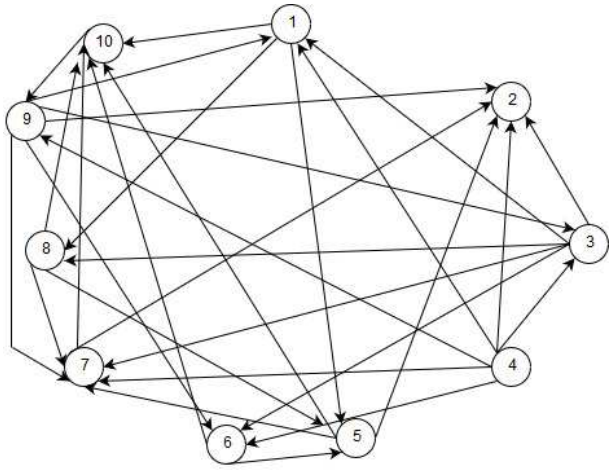


Fig. 7 Directed graph of V_L of example 4

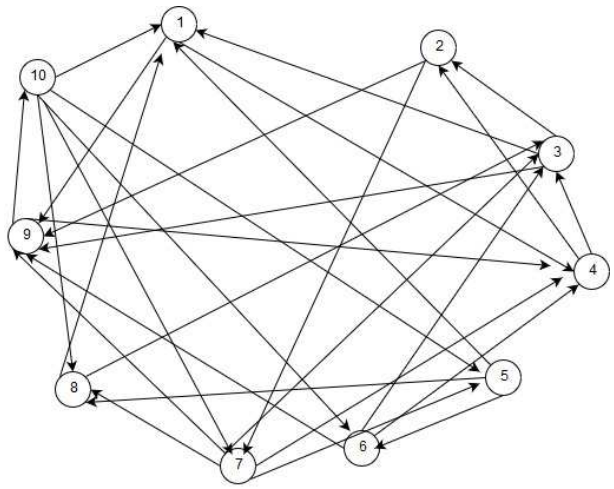


Fig. 8 Directed graph of V_U of example 4

where $m = 2$. The form of the cycle of length 3 is $(i_0 - k - j_0)$. To find the element of the cycle, we need V_L^2 as follows:

$$V_L^2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 2 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 3 & 0 & 1 & 1 & 0 & 4 \\ 2 & 3 & 1 & 0 & 2 & 2 & 2 & 2 & 0 & 3 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 2 \\ 1 & 2 & 0 & 0 & 2 & 1 & 1 & 2 & 0 & 3 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

For the cycle $(1 - 10 - 9)$, we have $i_0 = 1$, $k = 10$, $j_0 = 9$, satisfying $V_L(1, 10) = 1$, $V_L^2(10, 1) > 0$, $V_L(10, 9) > 0$, $V_L^2(9, 10) > 0$, and $V_L(9, 1) = 1$. Then the cycle of length 3 is $(1-10-9)$. Similarly for the cycle $(6 - 10 - 9)$, we have $i_0 = 6$, $k = 10$, $j_0 = 9$, such that $V_L(6, 10) > 0$, $V_L^2(10, 6) > 0$, $V_L(10, 9) > 0$, $V_L^2(9, 10) > 0$, and $V_L(9, 6) = 1$. For the cycle $(7 - 10 - 9)$, we have $i_0 = 7$, $k = 10$, $j_0 = 9$, such that $V(7, 10) > 0$, $V_L^2(10, 7) > 0$, $V_L(10, 9) > 0$, $V_L^2(9, 10) > 0$, and $V_L(9, 7) = 1$.

Next we try to find out the cycle of length 4. The form of the cycle of length 4 is $(i_0 - i_1 - j_0 - j_1)$. By using the algorithm 1, there are five cycle of length 4 that is $(5 - 10 - 9 - 6)$, $(1 - 10 - 9 - 3)$, $(3 - 6 - 10 - 9)$, $(3 - 7 - 10 - 9)$,and $(3 - 8 - 10 - 9)$. To find cycle of length 4, we need V_L^2. For the cycle $(5 - 10 - 9 - 6)$, we have $i_0 = 5$, $j_0 = 9$, satisfying step (2) of algorithm 2, since $V_L^2(5, 9) > 0$ and $V_L^2(9, 5) > 0$, and we have $i_1 = 10$ and $j_1 = 6$ satisfying step (3) of even length cycle algorithm 2 for $\alpha = 1$. Since $V_L(5, 10) = 1$, $V_L^2(10, 6) > 0$, $V_L^2(6, 10) > 0$ and $V_L(9, 6) = 1$. After that it is confirm $V_L(10, 9) = 1$ and $V_L(6, 5) = 1$. Then we have cycle of length 4 is $(5 - 10 - 9 - 6)$. Similarly cycle of length 4 is $(1 - 10 - 9 - 3)$, $(3 - 6 - 10 - 9)$, $(3 - 7 - 10 - 9)$,and $(3 - 8 - 10 - 9)$. To find the cycle of

length 5, using the algorithm for odd cycles, we need V_L, V_L^2 and V_L^3. For this V_L^3 as follows:

$$V_L^3 = \begin{bmatrix} 1 & 4 & 1 & 0 & 0 & 1 & 2 & 0 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 1 & 0 & 4 & 0 & 4 & 5 \\ 1 & 5 & 0 & 0 & 6 & 1 & 5 & 3 & 3 & 10 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 & 1 & 1 & 0 & 2 & 1 \\ 0 & 3 & 0 & 0 & 4 & 0 & 4 & 1 & 3 & 7 \\ 1 & 2 & 0 & 0 & 2 & 1 & 1 & 2 & 0 & 3 \end{bmatrix}$$

The form of the cycle of length 5 is $(i_0 - j_1 - k - i_1 - j_0)$. We have $i_0 = 3$, $k = 7$, $j_0 = 9$, satisfying $V_L^2(3,7) > 0$, $V_L^3(7,3) > 0$, $V_L^2(7,9) > 0$, $V_L^3(9,7) > 0$ and $V_L(9,3) = 1$. For $\alpha = 1$, we have to find out i_1 and j_1 where $i_1 = 10$ and $j_1 = 8$ with $V_L^3(3,10) > 0$, $V_L^2(10,3) > 0$, $V_L(7,10) = 1$, $V_L^2(9,8) > 0$, $V_L^3(8,9) > 0$ and $V_L(8,7) = 1$. After that it is confirm $V_L(10,9) = 1$ and $V_L(3,8) = 1$. Then we have cycle of length 5 is $(3-8-7-10-9)$. Similarly cycle of length 5 are obtained i.e $(3-6-5-10-9)$, $(1-8-10-9-3)$, $(1-5-10-9-3)$,and $(5-7-10-9-6)$.

Next we will try to find the cycle of length 6. The form of the cycle of length 6 is $(i_0-i_1-i_2-j_0-j_1-j_2)$. We have $i_0 = 1$ $j_0 = 10$ with $V_L^3(1,10) > 0$, $V_L^3(10,1) > 0$. For $\alpha = 1$ we have to find the vertices $i_1 = 5$ and $j_1 = 9$ satisfying step (3) of algorithm 1 satisfying $V_L(1,5) = 1$, $V_L^3(5,9) > 0$, $V_L^3(9,5) > 0$, $V_L(10,9) = 1$. Similarly, for $\alpha = 2$ we have $i_2 = 7$ and $j_2 = 3$ satisfying step (4) of algorithm 2 i.e. $V_L(5,7) = 1$, $V_L^3(7,3) > 0$, $V_L^3(3,7) > 0$, $V_L(9,3) = 1$. After that it is confirm $V_L(7,10) = 1$ and $V_L(3,1) = 1$. Then we have cycle of length 6 is $(1-5-7-10-9-3)$. Other cycle of length 6 are given by $(1-8-7-10-9-3)$, $(3-6-5-7-10-9)$, $(3-8-5-7-10-9)$,and $(1-8-5-10-9-3)$. Similarly, one cycle of length 7 is found that is $(1-8-5-7-10-9-3)$. There is no cycle of length 8, 9 and 10 is detected. For easy understanding, all the cycles of V_L are presented in Table 5.

Table 5 All possible cycle of V_L

Length	*Cycles*
3	(1 10 9), (6 10 9), (7 10 9)
4	(5 10 9 6), (1 10 9 3), (3 6 10 9), (3 7 10 9), (3 8 10 9)
5	(3 8 7 10 9), (3 6 5 10 9), (1 8 10 9 3),(1 5 10 9 3) , (5 7 10 9 6)
6	(1 5 7 10 9 3), (1 8 7 10 9 3), (3 6 5 7 10 9), (3 8 5 7 10 9), (1 8 5 10 9 3)
7	(1 8 5 7 10 9 3)
$8 \sim 10$	nothing

Similarly the possible cycles of the upper vertex matrix V_U are given in table 6.

Since both the lower and upper matrix V_L and V_U are inconsistent; therefore IMPR $\tilde{R}_5$ is also inconsistent.

To suggest the cause of inconsistency, we apply an algorithm for extinguishing cycles based on minimum covering sets (Nishizawa, 1995). In this example, from table 5 and 6, the path $2-7$, $3-4$, $9-10$ which cover all the cycle. If we will change the pair $(3,4)$ to $(4,3)$, $(2,7)$ to $(7,2)$, $(9,10)$ to $(10,9)$, in the original IMPR $\tilde{R}_5$, then the lower vertex matrix V_L and the upper vertex matrix V_U are both consistent. Then IMPR $\tilde{R}_5$ is also consistent.

Table 6 All possible cycle of V_U

Length	*Cycles*
3	(1 4 3), (2 7 4), (2 7 3), (3 9 4), (1 9 10), (6 9 10),
4	(2 7 4 3), (2 7 8 3), (1 9 4 3), (2 7 9 4), (3 9 10 6), (3 9 10 8), (1 9 10 8), (3 9 10 7), (1 9 10 5), (5 6 9 10), (2 9 4 3)
5	(2 7 5 6 3), (2 7 5 6 4), (1 4 2 7 5),(1 4 2 7 3) , (1 4 2 7 8), (2 7 5 8 3), (2 7 3 9 4),(1 4 2 9 10), (3 9 10 7 8), (1 9 10 8 3), (1 9 10 7 3), (3 9 10 7 4), (1 9 10 6 3),(3 9 10 6 4), (1 9 10 5 8), (2 9 10 8 3), (2 9 10 7 3), (2 9 10 6 3), (2 9 10 6 4), (2 9 10 7 4),
6	(2 7 5 6 4 3), (1 4 3 2 7 5), (1 4 2 7 5 8), (1 4 3 2 7 8), (1 4 2 7 8 3), (1 9 4 2 7 3),(2 9 10 7 4 3) (1 9 4 2 7 8), (2 7 8 3 9 4), (1 9 4 2 7 5), (2 7 5 6 9 4), (2 7 9 10 8 3), (1 4 2 9 10 5), (2 9 10 5 6 4), (1 4 3 9 10 8), (1 9 10 7 4 3), (1 9 10 6 4 3), (3 9 10 5 6 4)
7	(1 4 2 7 5 6 3), (1 4 2 7 5 8 3), (1 9 4 2 7 8 3), (2 7 5 6 3 9 4), (1 9 4 2 7 5 8), (1 4 2 9 10 6 3), (2 7 5 8 3 9 4), (1 9 4 3 2 7 5), (2 7 5 6 9 4 3), (1 4 3 9 10 5 8), (1 9 10 5 6 4 3), (1 4 3 2 9 10 8), (2 9 10 5 6 4 3), (1 4 2 9 10 5 8)
8	(1 9 4 2 7 5 6 3), (1 9 4 2 7 5 8 3), (1 9 4 3 2 7 5 8), (1 9 4 3 2 7 8 10) (1 4 3 2 9 10 5 8) , (1 4 3 2 7 9 10 8)
9	(1 4 2 7 8 3 9 10 5), (1 9 10 5 6 4 2 7 8)
10	(1 9 10 5 6 4 3 2 7 8)

4 Isomorphism between Intuitionistic multiplicative preference relations(IMPR) and asymmetric multiplicative preference relation

This section discusses the equivalence between the set of IMPR and the set of asymmetric multiplicative preference relation that leads to derive an asymmetric multiplicative preference relation from a given IMPR.

From the asymmetric multiplicative preference relation, we obtained the vertex matrix and checked for the consistency by using a directed graph in IMPR scenario.

Consider an IMPR

$$\tilde{R} = \begin{pmatrix} (\wp_{11}, \Im_{11}) & (\wp_{12}, \Im_{12}) & \cdots & (\wp_{1i}, \Im_{1i}) & \cdots & (\wp_{1n}, \Im_{1n}) \\ (\wp_{21}, \Im_{21}) & (\wp_{22}, \Im_{22}) & \cdots & (\wp_{2i}, \Im_{2i}) & \cdots & (\wp_{2n}, \Im_{2n}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\wp_{i1}, \Im_{i1}) & (\wp_{i2}, \Im_{i2}) & \cdots & (\wp_{ii}, \Im_{ii}) & \cdots & (\wp_{in}, \Im_{in}) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ (\wp_{n1}, \Im_{n1}) & (\wp_{n2}, \Im_{n2}) & \cdots & (\wp_{ni}, \Im_{ni}) & \cdots & (\wp_{nn}, \Im_{nn}) \end{pmatrix}$$

The above relation can completely characterised using just its upper triangular part, because the intuitionistic multiplicative element $(\wp_{ij}, \Im_{ij})$ is the mirror image of $(\wp_{ji}, \Im_{ji})$.

$$U\tilde{R} = \begin{pmatrix} (\wp_{11}, \Im_{11}) & (\wp_{12}, \Im_{12}) & \cdots & (\wp_{1i}, \Im_{1i}) & \cdots & (\wp_{1n}, \Im_{1n}) \\ & (\wp_{22}, \Im_{22}) & \cdots & (\wp_{2i}, \Im_{2i}) & \cdots & (\wp_{2n}, \Im_{2n}) \\ & & \ddots & \vdots & \ddots & \vdots \\ & & & (\wp_{ii}, \Im_{ii}) & \cdots & (\wp_{in}, \Im_{in}) \\ & & & & \ddots & \vdots \\ & & & & & (\wp_{nn}, \Im_{nn}) \end{pmatrix}$$

and this can be represented equivalently as the following MPR

$$R = \begin{pmatrix} \wp_{11} & \wp_{12} & \cdots & \wp_{1i} & \cdots & \wp_{1n} \\ \Im_{12} & \wp_{22} & \cdots & \wp_{2i} & \cdots & \wp_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \Im_{1i} & \Im_{2i} & \cdots & \wp_{ii} & \cdots & \wp_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \Im_{1n} & \Im_{2n} & \cdots & \Im_{in} & \cdots & \wp_{nn} \end{pmatrix}$$

Since $\wp_{ij} = \Im_{ji}$ and $\Im_{ij} = \wp_{ji}$ then the above MPR becomes

$$R = \begin{pmatrix} \wp_{11} & \wp_{12} & \cdots & \wp_{1i} & \cdots & \wp_{1n} \\ \wp_{21} & \wp_{22} & \cdots & \wp_{2i} & \cdots & \wp_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \wp_{i1} & \wp_{i2} & \cdots & \wp_{ii} & \cdots & \wp_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \wp_{n1} & \wp_{n2} & \cdots & \wp_{ni} & \cdots & \wp_{nn} \end{pmatrix}$$

Let $\widetilde{Q}$ denote the set of IMPRs, where

$$\widetilde{Q} = \left\{ \tilde{R} = (\tilde{r}_{ij}) | \forall i, j : \tilde{r}_{ij} = (\wp_{ij}, \Im_{ij}), \wp_{ij}, \Im_{ij} \in [1/9, 9], \wp_{ii} = \Im_{ii} = 1, \wp_{ij} = \Im_{ji}, \wp_{ji} = \Im_{ij}, 0 \leq \wp_{ij}\Im_{ij} \leq 1 \right\}$$

and $\Re$ be the set of MPRs $\Re = \{R = (r_{ij}) | \forall i, j : r_{ij} \in [1/9, 9]\}$. Define a mapping $f : [1/9, 9] \times [1/9, 9] \to [1/9, 9]$ by the function $f(x_1, x_2) = x_1$. We can define the following mapping, $F : \widetilde{Q} \to \Re$ between the set of IMPRs $\widetilde{Q}$ and the set of MPRs, $\Re$

$$\{f(\tilde{r}_{ij})\} = \{\wp_{ij}\} \; i.e \; R = F(\tilde{R})$$

The following properties can be proved

Proposition 1 *Function F is well defined, i.e. For given $\tilde{R} \in \widetilde{Q} \Rightarrow f(\tilde{R}) \in \Re$.*

Proof Let $\tilde{R} = (\tilde{r}_{ij}) \in \widetilde{Q}$.
Here $\tilde{r}_{ij} = (\wp_{ij}, \Im_{ij}) \Rightarrow f(\tilde{r}_{ij}) = f(\wp_{ij}, \Im_{ij}) = \wp_{ij} \in R$

Proposition 2 *Function F is one-one.*

Proof Let $\tilde{R}_1 = (\tilde{r}^1_{ij})$ and $\tilde{R}_2 = (\tilde{r}^2_{ij})$ are IMPR such that $F(\tilde{R}_1) = F(\tilde{R}_2)$. Then we have that

$$f(\tilde{r}^1_{ij}) = f(\tilde{r}^2_{ij}) \; \forall i, j \Leftrightarrow \wp^1_{ij} = \wp^2_{ij} \; \forall i, j.$$

Because of the conditions of $\wp^1_{ij} = \Im^1_{ji}$ and $\wp^2_{ij} = \Im^2_{ji}$, then it is obvious that $\Im^1_{ij} = \Im^2_{ij}, \forall i, j$ Therefore, we have that

$$(\wp^1_{ij}, \Im^1_{ij}) = (\wp^2_{ij}, \Im^2_{ij}) \Leftrightarrow \tilde{R}_1 = \tilde{R}_2 \; \forall i, j$$

For the function to be onto, the following conditions to be verified:

$$\forall R \in \Re \; \exists \; \tilde{R} \in \widetilde{Q} : F(\tilde{R}) = R.$$

By the definition of F and $\widetilde{Q}$, we have that $R = (r_{ij}) = (\wp_{ij})$ satisfies:

$$0 \leq r_{ij} r_{ji} = \wp_{ij}\wp_{ji} \leq 1.$$

Thus R is asymmetric multiplicative preference relation that proves the range of the function F is the subset of MPRs which are asymmetric.

Theorem 1 *The set of intuitionistic multiplicative preference relations is isomorphic to set of asymmetric multiplicative preference relations.*

Proof We know that when $\tilde{R} \in \widetilde{Q}$ has hesitancy degree always zero, then we have that:

$$\wp_{ij}\Im_{ij} = 1, \forall i, j \tag{9}$$

In this case, $F(\tilde{R}) = R$ is also reciprocal, i.e. $r_{ij}r_{ji} = 1\forall i, j$. The proof of this is quite simple as we have the following:

$$\forall i, j : r_{ij} = f(\tilde{r}_{ij}) = \wp_{ij} \wedge r_{ji} = f(\tilde{r}_{ji}) = \wp_{ji}.$$

Since $\tilde{R} \in \widetilde{Q}$ then we have that $\wp_{ji} = \Im_{ij}\forall i, j$ and by using equation 9 it is $r_{ij}r_{ji} = \wp_{ij}\wp_{ji} = \wp_{ij}\Im_{ij} = 1\forall i, j$.

Example 5 Isomorphic asymmetric multiplicative preference relation of IMPR of example 2 is

$$I_{\tilde{R}_2} = \begin{pmatrix} 1 & \frac{5}{3} & 7 & 3 & \frac{5}{3} & 1 & \frac{1}{4} \\ \frac{1}{4} & 1 & \frac{5}{3} & \frac{5}{3} & \frac{3}{5} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{9} & \frac{1}{4} & 1 & 1 & \frac{1}{4} & \frac{1}{3} & \frac{1}{9} \\ \frac{1}{7} & \frac{1}{4} & \frac{3}{5} & 1 & \frac{3}{5} & \frac{1}{3} & \frac{1}{9} \\ \frac{1}{7} & 1 & \frac{5}{3} & \frac{3}{5} & 1 & \frac{1}{3} & \frac{1}{7} \\ \frac{3}{5} & 1 & 3 & 3 & 3 & 1 & \frac{1}{7} \\ \frac{5}{3} & 3 & 7 & 7 & 3 & 3 & 1 \end{pmatrix}$$

The vertex matrix of above asymmetric MPR is

$$V = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}$$

By using the algorithm 1 and 2 of odd and even length cycle ,there is no cycle is present in the vertex matrix V. Hence $\tilde{R}_2$ is consistent. Also we can directly see from the figure 9.

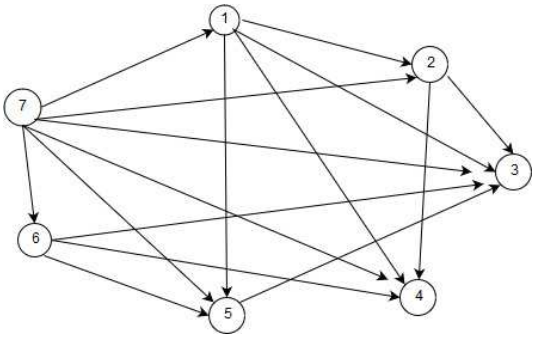


Fig. 9 Directed graph of example 5

Also same isomorphism approach is applied in incomplete IMPRs scenario.

Example 6 Isomorphic asymmetric MPR of of example 3 is

$$I_{\tilde{R}_3} = \begin{pmatrix} 1 & \frac{1}{5} & 5 & * & \frac{1}{9} & 2 \\ 2 & 1 & * & \frac{1}{4} & \frac{1}{2} & 3 \\ \frac{1}{6} & * & 1 & * & 6 & * \\ * & \frac{1}{5} & * & 1 & \frac{1}{5} & * \\ 8 & \frac{1}{3} & \frac{1}{7} & 4 & 1 & * \\ \frac{1}{6} & \frac{1}{5} & * & * & * & 1 \end{pmatrix}$$

Vertex matrix of the asymmetric MPR, $I_{\tilde{R}_3}$ is

$$V = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Using the algorithm 1 and 2, the is one cycle of length 3 i.e. $(1-3-5)$ is present. Therefore, $\tilde{R}_3$ is inconsistent. In this example, if we will change the pair $(3,5)$ to $(5,3)$ in the original incomplete IMPR

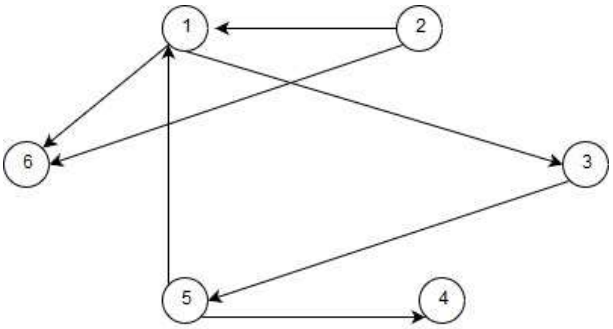


Fig. 10 Directed graph of 6

$\tilde{R}_3$, then it is consistent.

Example 7 Isomorphic asymmetric MPR of example 4 is

$$I_{\tilde{R}_4} = \begin{pmatrix} 1 & * & \frac{1}{2} & \frac{1}{3} & 3 & * & * & \frac{8}{3} & \frac{1}{7} & 5 \\ * & 1 & \frac{2}{3} & \frac{1}{8} & \frac{1}{4} & * & \frac{1}{5} & * & \frac{1}{3} & * \\ \frac{1}{5} & \frac{6}{10} & 1 & \frac{3}{5} & * & 2 & 4 & 7 & \frac{1}{8} & * \\ 2 & \frac{1}{2} & \frac{2}{3} & 1 & * & 3 & 5 & * & 2 & * \\ \frac{1}{6} & 1 & * & * & 1 & \frac{1}{8} & 8 & \frac{1}{6} & * & 4 \\ * & * & \frac{1}{2} & \frac{1}{7} & 7 & 1 & * & * & \frac{1}{5} & 2 \\ * & 5 & \frac{1}{8} & \frac{1}{6} & \frac{1}{9} & * & 1 & \frac{1}{9} & \frac{1}{7} & 4 \\ \frac{1}{8} & * & \frac{1}{8} & * & 3 & * & 6 & 1 & * & 6 \\ 4 & 3 & 5 & \frac{1}{2} & * & 4 & 2 & * & 1 & \frac{1}{3} \\ \frac{1}{5} & * & * & * & \frac{1}{5} & \frac{1}{3} & \frac{1}{8} & \frac{1}{6} & 2 & 1 \end{pmatrix}$$

vertex matrix of $I_{\tilde{R}_4}$ is

$$V = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

By using the algorithm of odd and even length cycle 1 and 2, the corresponding cycles are given in table 6.

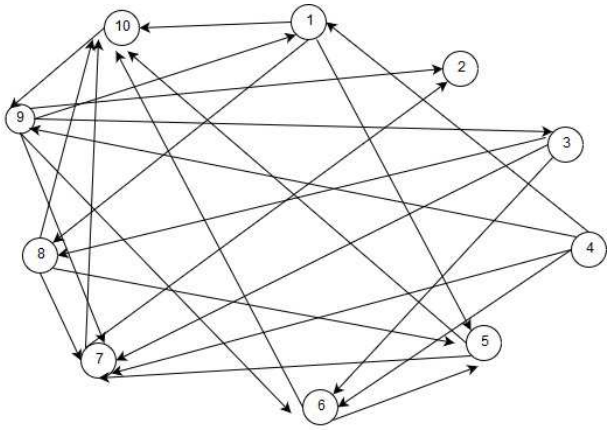


Fig. 11 Directed graph of Example 7

Table 7 All possible cycle of V

Length	*Cycles*
3	(1 10 9), (6 10 9), (7 10 9)
4	(5 10 9 6), (1 8 10 9), (1 5 10 9), (3 7 10 9), (3 8 10 9), (3 6 10 9)
5	(3 8 7 10 9), (3 6 5 10 9), (1 8 7 10 9),(1 8 5 10 9) , (5 7 10 9 6), (3 8 5 10 9), (1 5 7 10 9)
6	(1 8 5 7 10 9), (5 7 10 9 3 6), (3 6 5 7 10 9), (3 8 5 7 10 9), (5 7 10 9 3 8)
$7 \sim 10$	nothing

Therefore, $\tilde{R}_5$ is inconsistent. In table 7 it is conclude the path $9-10$ that cover all the cycle. If we will change the pair $(9, 10)$ to $(10, 9)$, in the original IMPR $\tilde{R}_5$, then $\tilde{R}_5$ is consistent.

4.1 Comparison between two different Approaches:

In example 2, $\tilde{R}_2$ is inconsistent according to approach 1 has no cycle found in the lower vertex matrix and one cycle of length 3 $(3-5-4)$ found in the upper vertex matrix. But it is consistent according to approach 2. To suggest the cause of inconsistency in approach 1, we will change the path $(3, 4)$ to $(4, 3)$

in the original IMPR $\tilde{R}_2$; then both the lower and upper vertex is consistent. Therefore $\tilde{R}_2$ is consistent. From the two approaches, we have concluded that the second approach is better than approach 1.

In example 3, $\tilde{R}_3$ is inconsistent according to approach 1 has one cycle found i.e. $(1-3-5)$ in the lower vertex matrix and one cycle of length 3, i.e. $(1-5-3)$ is found in the upper vertex matrix. Also according to approach 2 $\tilde{R}_3$ is inconsistent where one cycle is found in isometric matrix i.e. $(1-3-5)$. In both the approach the path $3-5$ which cover the cycle. We have to choose the couple to eliminate the cycle. In this case, if we will change the pair $(3,5)$ to $(5,3)$ in the original IMPR $\tilde{R}_4$, then both the lower and upper vertex is consistent. Then IMPR $\tilde{R}_4$ is also consistent. Therefore approach 2 is better than approach 1 because less no of the cycle is found in approach 2.

In example 4, According to approach 1, 19 cycles are found in the lower vertex matrix V_L and 78 cycles are found in V_U that are given in table 5 and 6. Similarly, according to approach 2, 21 cycles are found that are given in table 7. In approach 1, we conclude that if we change the path $9-10$, $4-3$, and $2-7$, then the IMPR $\tilde{R}_4$ is consistent. But according to approach 2, if we will change the pair $(9,10)$ to $(10,9)$, in the original IMPR $\tilde{R}_5$, then $\tilde{R}_5$ is consistent. Therefore, approach 2 is better than approach 1.

5 Comparative analysis with the existing methods

Lot of research has been developed in literature in the area of MPR, incomplete MPR, IMPR, incomplete IMPRs, and their consistency. In this section we have compare our methodology with these exsting techniques in literature.

Comparison with Xu et al. (2013)

Chiclana et al. (2001) studied the transformation function between reciprocal MPR with values in the interval scale (1/9, 9), and reciprocal FPRs with values in [0,1].

Proposition 3 *(Chiclana et al. (2001)) Suppose that $X = \{x_1, x_2, \cdots, x_n\}$, and associated to it a reciprocal multiplicative preference relation $A = (a_{ij})_{n\times n}$ with $a_{ij} \in [1/9, 9]$, then the corresponding reciprocal fuzzy preference relation, $P = (p_{ij})_{n\times n}$ with $p_{ij} \in [0,1]$, associated with A is given as follows* $p_{ij} = \frac{1}{2}(1 + log_9 a_{ij})$

After spliting the IMPR into two MPRs, the two MPRs is converted into two FPRs using the proposition 3. Xu et al. (2013) proposed the ordinal consistency index to measure the degree of ordinal consistency of a FPR, which is to count the unreasonable 3-cycles in a directed graph that represents the FPR. Using this concept we check the consistency of IMPR. Let us take example 1

$$\tilde{R}_1 = \begin{pmatrix} (1,1) & \left(\frac{1}{2},\frac{1}{4}\right) & \left(2,\frac{1}{8}\right) & \left(\frac{2}{3},\frac{1}{4}\right) \\ \left(\frac{1}{4},\frac{1}{2}\right) & (1,1) & \left(5,\frac{1}{7}\right) & \left(\frac{7}{5},\frac{2}{3}\right) \\ \left(\frac{1}{8},2\right) & \left(\frac{1}{7},5\right) & (1,1) & \left(6,\frac{1}{7}\right) \\ \left(\frac{1}{4},\frac{2}{3}\right) & \left(\frac{2}{3},\frac{7}{5}\right) & \left(\frac{1}{7},6\right) & (1,1) \end{pmatrix}$$

This IMPR $\tilde{R}_1$ is split into two MPRs using equation 6 that is given below

$$A_1 = \begin{bmatrix} 1 & \frac{1}{2} & 2 & \frac{2}{3} \\ 2 & 1 & 5 & \frac{7}{5} \\ \frac{1}{2} & \frac{1}{5} & 1 & 6 \\ \frac{3}{2} & \frac{5}{7} & \frac{1}{6} & 1 \end{bmatrix} \quad A_2 = \begin{bmatrix} 1 & \frac{1}{4} & \frac{1}{8} & \frac{1}{4} \\ 4 & 1 & \frac{1}{7} & \frac{2}{3} \\ 8 & 7 & 1 & \frac{1}{7} \\ 4 & \frac{3}{2} & 7 & 1 \end{bmatrix}$$

Using the proposition 3 the above two MPRs converted into two FPRs i.e P_1 and P_2 given by

$$P_1 = \begin{bmatrix} 0.5 & 0.3423 & 0.6577 & 0.4077 \\ 0.6577 & 0.5 & 0.8662 & 0.5766 \\ 0.3423 & 0.1338 & 0.5 & 0.9077 \\ 0.5923 & 0.4234 & 0.0923 & 0.5 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0.5 & 0.1845 & 0.0268 & 0.1845 \\ 0.8155 & 0.5 & 0.0572 & 0.4077 \\ 0.9732 & 0.9428 & 0.5 & 0.0572 \\ 0.8155 & 0.5923 & 0.9428 & 0.5 \end{bmatrix}$$

According to Xu et al. (2013) the adjacency matrix of P_1 is

$$E = \begin{bmatrix} 0\,0\,1\,0 \\ 1\,0\,1\,1 \\ 0\,0\,0\,1 \\ 1\,0\,0\,0 \end{bmatrix} \quad E^3 = \begin{bmatrix} 1\,0\,0\,0 \\ 1\,0\,1\,1 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}$$

The Hadamard product of E^2 and E^T is denoted by B that is given below

$$B = \begin{bmatrix} 1\,0\,0\,0 \\ 1\,0\,1\,1 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}$$

Thus $\frac{trac(E^3)}{3} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} b_{ij}}{3} = 1$. Then there is one 3-cycle in the FPR P_1. To identify the cycle we will collect all the 3×3 pricipal sub-matrices of B are

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ 1 \\ 2 \\ 4 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ 2 \\ 3 \\ 4 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array}, \quad \begin{array}{c} \\ 1 \\ 3 \\ 4 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} 1 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array}$$

There does not exist a zero row or column in the principal sub matrix $B[1,3,4]$. Therefore, there is one 3-cycle in diagraph of P_1 i.e. $V_1 \to V_4 \to V_3 \to V_1$ (see figure 12). The inconsistent entries are $r_{31}^{(0)}, r_{14}^{(0)}, r_{43}^{(0)}$. In our method also we got same inconsistent entries. Thus P_1 is order inconsistent and each entries appear once time in the cycle. So, $r_{31}^{(0)} = 0.3423, r_{14}^{(0)} = 0.4077, r_{43}^{(0)} = 0.0923$. According to Xu et al.(2013), we choose the entry that is closer to 0.5. From these above three value $r_{14}^{(0)} = 0.4077$ is near to 0.5. Therefore, take $r_{14}^{(1)} = 1 - 0.4077 = 0.5923$, and $r_{41}^{(1)} = 0.4077$ in P_1 matrix then apply same procedure it becomes consistent. Similarly, the adjacency matrix corresponding to P_2 is given by

$$E = \begin{bmatrix} 0\,0\,0\,0 \\ 1\,0\,0\,0 \\ 1\,1\,0\,0 \\ 1\,1\,1\,0 \end{bmatrix}$$

The Hadamard product of E^2 and E^T is denoted by B that is given below

$$B = \begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}$$

Here, $\frac{\sum_{i=1}^{4}\sum_{j=1}^{4} b_{ij}}{3} = 0$ then there is no 3-cycles are found in P_2 matrix. In our method also, the MPR A_2 has no cycles. Therefore P_2 is consistent. Since, both the FPRs P_1, P_2 are consistent, then the IMPR is consistent.

Let us take another example 2

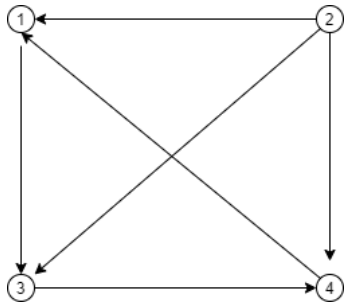


Fig. 12 Diagraph of FPR P_1 of example 1

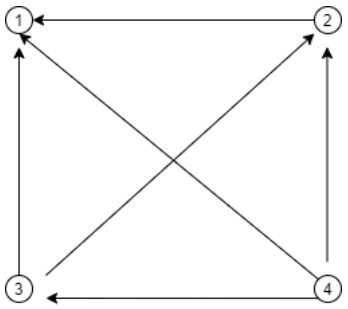


Fig. 13 Diagraph of FPR P_2 of example 1

$$\tilde{R}_2 = \begin{pmatrix} (1,1) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(7,\frac{1}{9}\right) & \left(3,\frac{1}{7}\right) & \left(\frac{5}{3},\frac{1}{7}\right) & \left(1,\frac{3}{5}\right) & \left(\frac{1}{4},\frac{5}{3}\right) \\ \left(\frac{1}{4},\frac{5}{3}\right) & (1,1) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(\frac{3}{5},1\right) & \left(\frac{1}{3},1\right) & \left(\frac{1}{4},3\right) \\ \left(\frac{1}{9},7\right) & \left(\frac{1}{4},\frac{5}{3}\right) & (1,1) & \left(1,\frac{3}{5}\right) & \left(\frac{1}{4},\frac{5}{3}\right) & \left(\frac{1}{3},3\right) & \left(\frac{1}{9},7\right) \\ \left(\frac{1}{7},3\right) & \left(\frac{1}{4},\frac{5}{3}\right) & \left(\frac{3}{5},1\right) & (1,1) & \left(\frac{3}{5},\frac{3}{5}\right) & \left(\frac{1}{3},3\right) & \left(\frac{1}{9},7\right) \\ \left(\frac{1}{7},\frac{5}{3}\right) & \left(1,\frac{3}{5}\right) & \left(\frac{5}{3},\frac{1}{4}\right) & \left(\frac{3}{5},\frac{3}{5}\right) & (1,1) & \left(\frac{1}{3},3\right) & \left(\frac{1}{7},3\right) \\ \left(\frac{3}{5},1\right) & \left(1,\frac{1}{3}\right) & \left(3,\frac{1}{3}\right) & \left(3,\frac{1}{3}\right) & \left(3,\frac{1}{3}\right) & (1,1) & \left(\frac{1}{7},3\right) \\ \left(\frac{5}{3},\frac{1}{4}\right) & \left(3,\frac{1}{4}\right) & \left(7,\frac{1}{9}\right) & \left(7,\frac{1}{9}\right) & \left(3,\frac{1}{7}\right) & \left(3,\frac{1}{7}\right) & (1,1) \end{pmatrix}$$

This IMPR R_2 is split into two MPRs using equation 6 that is given below

$$A_1 = \begin{pmatrix} 1 & \frac{5}{3} & 7 & 3 & \frac{5}{3} & 1 & \frac{1}{4} \\ \frac{3}{5} & 1 & \frac{5}{3} & \frac{5}{3} & \frac{3}{5} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{7} & \frac{3}{5} & 1 & 1 & \frac{1}{4} & \frac{1}{3} & \frac{1}{9} \\ \frac{1}{3} & \frac{3}{5} & 1 & 1 & \frac{3}{5} & \frac{1}{3} & \frac{1}{7} \\ \frac{3}{5} & \frac{5}{3} & 4 & \frac{5}{3} & 1 & \frac{1}{3} & \frac{1}{7} \\ 1 & 3 & 3 & 3 & 3 & 1 & \frac{1}{7} \\ 4 & 4 & 9 & 9 & 7 & 7 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{9} & \frac{1}{7} & \frac{1}{7} & \frac{3}{5} & \frac{5}{3} \\ 4 & 1 & \frac{1}{4} & \frac{1}{4} & 1 & 1 & 3 \\ 9 & 4 & 1 & \frac{3}{5} & \frac{5}{3} & 3 & 7 \\ 7 & 4 & \frac{5}{3} & 1 & \frac{3}{5} & 3 & 7 \\ 7 & 1 & \frac{3}{5} & \frac{5}{3} & 1 & 3 & 3 \\ \frac{5}{3} & 1 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 1 & 3 \\ \frac{3}{5} & \frac{1}{3} & \frac{1}{7} & \frac{1}{7} & \frac{1}{3} & \frac{1}{3} & 1 \end{pmatrix}$$

The MPRs converted into FPRs that is P_1 and P_2 is given by

$$P_1 = \begin{bmatrix} 0.5 & 0.6162 & 0.9428 & 0.75 & 0.6162 & 0.5 & 0.1845 \\ 0.3838 & 0.5 & 0.6162 & 0.6162 & 0.3838 & 0.25 & 0.1845 \\ 0.0572 & 0.3838 & 0.5 & 0.5 & 0.1845 & 0.25 & 0 \\ 0.25 & 0.3838 & 0.5 & 0.5 & 0.3838 & 0.25 & 0 \\ 0.3838 & 0.6162 & 0.8155 & 0.6162 & 0 & 0.25 & 0.0572 \\ 0.5 & 0.75 & 0.75 & 0.75 & 0.75 & 0.5 & 0.0572 \\ 0.8155 & 0.8155 & 1 & 1 & 0.9428 & 0.9428 & 0.5 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0.5 & 0.1845 & 0 & 0.0572 & 0.0572 & 0.3838 & 0.6162 \\ 0.8155 & 0.5 & 0.1845 & 0.1845 & 0.5 & 0.5 & 0.75 \\ 1 & 0.8155 & 0.5 & 0.3838 & 0.6162 & 0.75 & 0.9428 \\ 0.9428 & 0.8155 & 0.6162 & 0.5 & 0.3838 & 0.75 & 0.9428 \\ 0.9428 & 0.5 & 0.3838 & 0.6162 & 0.5 & 0.75 & 0.75 \\ 0.6162 & 0.5 & 0.25 & 0.25 & 0.25 & 0.5 & 0.75 \\ 0.3838 & 0.25 & 0.0572 & 0.0572 & 0.25 & 0.25 & 0.5 \end{bmatrix}$$

The adjacency matrix of the FPR P_1 is

$$E = \begin{bmatrix} 0\,1\,1\,1\,1\,0\,0 \\ 0\,0\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,1\,1\,1\,0\,0\,0 \\ 0\,1\,1\,1\,1\,0\,0 \\ 1\,1\,1\,1\,1\,1\,0 \end{bmatrix}, \quad E^3 = \begin{bmatrix} 0\,0\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,1\,1\,0\,0\,0 \\ 0\,2\,5\,5\,0\,0\,0 \end{bmatrix}$$

Thus $\frac{trac(E^3)}{3} = 0$. There is no 3-cycles. Similarly, the adjacency matrix of P_2 matrix is

$$E = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,1 \\ 1\,0\,0\,0\,0\,0\,1 \\ 1\,1\,0\,0\,1\,1\,1 \\ 1\,1\,1\,0\,0\,1\,1 \\ 1\,0\,0\,1\,0\,1\,1 \\ 1\,0\,0\,0\,0\,0\,1 \\ 0\,0\,0\,0\,0\,0\,0 \end{bmatrix}, \quad E^3 = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 2\,1\,1\,0\,0\,1\,5 \\ 3\,0\,0\,1\,0\,1\,6 \\ 3\,1\,0\,0\,1\,1\,5 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \end{bmatrix}$$

The Hadamard product of E^2 and E^T is denoted by B that is given below

$$B = \begin{bmatrix} 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,0\,0 \\ 0\,0\,1\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0 \end{bmatrix}$$

Thus $\frac{trac(E^3)}{3} = \frac{\sum_{i=1}^{4}\sum_{j=1}^{4} b_{ij}}{3} = 1$. Then there is one 3-cycle. To identify the cycle we have 35, 3×3 principal sub-matrices of B. Out of these, the principal sub matrix $B[3,4,5]$ has no zero row and column. Therefore, there are one 3-cycles in diagraph of adjacency matrix of P_2, i.e. $V_1 \to V_4 \to V_3 \to V_1$. The inconsistent entries are $r_{34}^{(0)}, r_{45}^{(0)}, r_{53}^{(0)}$. In our method also we got same inconsistent entries. Thus P_1 is order inconsistent and each entries appear once time in the cycle. So, $r_{34}^{(0)} = 0.3838, r_{45}^{(0)} = 0.3838, r_{53}^{(0)} =$ 0.0.3838. According to Xu et al. (2013), we choose the entry that is closer to 0.5. From these above three value take any one of them. Take $r_{34}^{(1)} = 1 - 0.3838 = 0.6162$ and $r_{43}^{(1)} = 0.3838$ in P_1 matrix then apply same procedure it becomes consistent.

Both the methods are equivalent. we are getting the same cycles in both the methods. In our method, we can take any order of matrix and check the consistency. But Xu et al. (2013) methods is complicated to find the inconsistency entries from the principal sub-matrix as for n-order matrix the number of principal sub-matrices of order 3 are n_{C_3}.

Comparison with Jiang et al. (2015)

Jiang et al. (2015) proposed the consistency property of an IMPR. They proposed two approaches to find the missing elements of incomplete IMPRs, i.e. “estimating step" and “adjusting step." To calculate the initial values of the missing element, a geometric mean method is used in the estimating step. To improve the initial values, another two different approaches are developed that is local optimizations model and iterative method. Jiang et al.(2015) method once getting the missing value, the consistency of IMPR will be checked by splitting the IMPRs into two MPRs. In this paper consistency of an arbitrary IMPRs are checked using the cyclic length of a directed graph. We have taken one example of Jiang et al. (2015) paper. Let us consider the example 2 of Jiang et al.(2015)

$$A = \begin{bmatrix} (1,1) & (3/2,*) & (1/3,1) & (1/2,1) \\ (*,3/2) & (1,1) & (1/4,1/2) & (*,2) \\ (1,1/3) & (2,1/4) & (1,1) & (1/3,1) \\ (1,1/2) & (2,*) & (1,1/3) & (1,1) \end{bmatrix}$$

The order pair vertex matrix of IMPR A is given

$$V_A = \begin{bmatrix} (0,0) & (1,0) & (0,0) & (0,0) \\ (0,0) & (0,0) & (0,0) & (0,1) \\ (1,0) & (1,0) & (0,0) & (0,0) \\ (1,0) & (0,0) & (1,0) & (0,0) \end{bmatrix}$$

The lower and upper vertex matrix are

$$V_L = \begin{bmatrix} 0\,1\,0\,0 \\ 0\,0\,0\,0 \\ 1\,1\,0\,0 \\ 1\,0\,0\,0 \end{bmatrix},\ V_U = \begin{bmatrix} 0\,0\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}$$

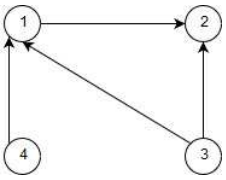


Fig. 14 Directed graph of V_L of IMPR A

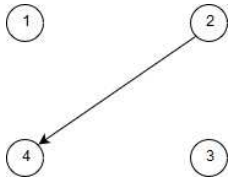


Fig. 15 Directed graph of V_U of IMPR A

From figure 14 and 15, we conclude that no cycle is found in both the lower and upper vertex matrix. Therefore, the incomplete IMPR A is consistent same as result of Jiang et al. (2015). Both the methods i.e our graphical methods and the Jiang et al. (2015) method are equivalent. But in graphical approach we can directly check the consistency of IMPR without finding the missing elements from the incomplete IMPR.

Comparison with Zhang, Z. and Pedrycz,W.,(2017)

Zhang, Z. and Pedrycz,W.,(2017), defined a new definition of consistency IMPR and proved that Xu's (Xu, Z. S., (2013)) definition of consistency and Jiang et al. (Jiang et al. (2015)) definition of consistency, are the special case of consistency defined by Zhang, Z. and Pedrycz,W.,(2017). According to Zhang, Z. and Pedrycz,W.,(2017) consistency is independent of alternative labels. For an multi-criteria decision making(MCDM) problem with n decision alternatives, the DMs pairwise comparison information can be structured by differently labeling the $n-$ alternatives and yields $n!$ IMPRs.

We take example 1 from the article by Zhang, Z. and Pedrycz, W.,(2017).

Consider three alternatives A, B, C in an MCDM problems. A DM employs the pairwise comparison method to elicit his/her judgements. There are six possible cases for labeling the three alternatives. If the three alternatives A, B, C are label by x_1, x_2 and x_3, then the DM's pairwise judgements are given by the following IMPR

$$\begin{array}{c} \\ x_1 : A \\ x_2 : B \\ x_3 : C \end{array} \begin{array}{c} \begin{array}{ccc} x_1 : A & x_2 : B & x_3 : C \end{array} \\ \begin{pmatrix} (1,1) & (\frac{2}{5},\frac{3}{2}) & (\frac{2}{5},\frac{3}{4}) \\ (\frac{3}{2},\frac{2}{5}) & (1,1) & (1,\frac{1}{2}) \\ (\frac{3}{4},\frac{2}{5}) & (\frac{1}{2},1) & (1,1) \end{pmatrix} \end{array}$$

By Jinag's (Jiang et al. (2015)) method it is consistent. According to our method, the order pair vertex matrix

$$V = \begin{bmatrix} (0,0) & (0,1) & (0,0) \\ (1,0) & (0,0) & (0,0) \\ (1,1) & (0,1) & (0,0) \end{bmatrix}$$

The lower and upper vertex matrix is

$$V_L = \begin{bmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 1\,0\,0 \end{bmatrix}, \quad V_U = \begin{bmatrix} 0\,1\,0 \\ 0\,0\,0 \\ 1\,1\,0 \end{bmatrix}$$

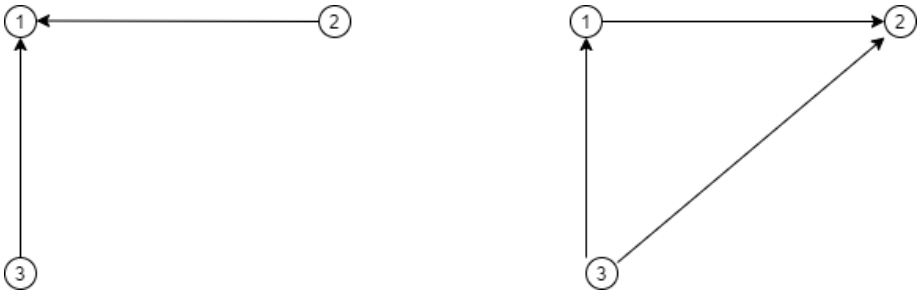


Fig. 16 Directed graph of V_L and V_U of IMPR of example 1 of Zhang, Z. and Pedrycz,W.,(2017)

From the directed graph 16, Both V_L and V_U are consistent because there are no circle was found. So, IMPR is consistent.

If the alternative A, B, C are labeled by x_2, x_3, x_1, then the DM's pairwise judgements should be restructured by the following IMPR i.e.

$$\begin{array}{c} \\ x_1 : C \\ x_2 : A \\ x_3 : B \end{array} \begin{array}{c} \begin{array}{ccc} x_1 : C & x_2 : A & x_3 : B \end{array} \\ \begin{pmatrix} (1,1) & (\frac{3}{4},\frac{2}{5}) & (\frac{1}{2},1) \\ (\frac{2}{5},\frac{3}{4}) & (1,1) & (\frac{2}{5},\frac{3}{2}) \\ (1,\frac{1}{2}) & (\frac{3}{2},\frac{2}{5}) & (1,1) \end{pmatrix} \end{array}$$

This re-structured IMPR is inconsistent according to Jiang et al. (2015). But it is consistent, according to Zhang, Z. and Pedrycz, W.,(2017). Now same re-structured IMPR is consistent according to our method. The ordered pair vertex matrix is

$$V = \begin{bmatrix} (0,0) & (0,0) & (0,0) \\ (1,1) & (0,0) & (0,1) \\ (1,0) & (1,0) & (0,0) \end{bmatrix}$$

The lower and upper vertex matrix is

$$V_L = \begin{bmatrix} 0\,0\,0 \\ 1\,0\,0 \\ 1\,1\,0 \end{bmatrix}, \quad V_U = \begin{bmatrix} 0\,0\,0 \\ 1\,0\,1 \\ 0\,0\,0 \end{bmatrix}$$

From the directed graph 17, Both V_L and V_U are consistent because there are no circle was found. So, re-structured IMPR is consistent. Therefore, the restructured IMPR, is consistent according to Zhang, Z. and Pedrycz,W.,(2017) as well as our method.

Comparison with Jin et al. (2018)



Fig. 17 Directed graph of V_L and V_U of re-structured IMPR of example 1 of Zhang, Z. and Pedrycz,W.,(2017)

Jin et al. (2018) proposed a method to derive normalized intuitionistic multiplicative weights on the basis of order consistent IMPR for decision making. Also, Jin et al. (2018) constructed an optimization model to generate the normalized intuitionistic multiplicative weights of IMPR, and compute the optimal deviation values to improve the consistency of the given IMPR, such that the repaired IMPR is constructed. To perform the comparison, we have taken the example 2 of (Jin et al. (2018)) the IMPR $P = (p_{ij})$ i.e.

$$P = \begin{pmatrix} (1,1) & (\frac{5}{3},\frac{1}{4}) & (7,\frac{1}{9}) & (3,\frac{1}{7}) \\ (\frac{1}{4},\frac{5}{3}) & (1,1) & (\frac{5}{3},\frac{1}{4}) & (\frac{5}{3},\frac{1}{6}) \\ (\frac{1}{9},7) & (\frac{1}{4},\frac{5}{3}) & (1,1) & (1,\frac{3}{5}) \\ (\frac{1}{7},3) & (\frac{1}{6},\frac{5}{3}) & (\frac{3}{5},1) & (1,1) \end{pmatrix}$$

According to Jin et al. (2018), the IMPR P is non-order consistent. Then, they obtained the optimal normalized intuitionistic multiplicative weights, using goal programming model (M-2) of Jin et al. (2018). To construct the repaired consist IMPR, they utilize the optimal deviation values. Therefore, they get the repaired IMPR is consistent. Here we solve the same IMPR by using our method. The binary order pair vertex matrix of IMPR P is

$$V = \begin{pmatrix} (0,0) & (1,0) & (1,0) & (1,0) \\ (0,1) & (0,0) & (1,0) & (1,0) \\ (0,1) & (0,1) & (0,0) & (0,0) \\ (0,1) & (0,1) & (0,1) & (0,0) \end{pmatrix}$$

The lower and upper vertex matrices are,

$$V_L = \begin{pmatrix} 0\,1\,1\,1 \\ 0\,0\,1\,1 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{pmatrix} \; ; \; V_U = \begin{pmatrix} 0\,0\,0\,0 \\ 1\,0\,0\,0 \\ 1\,1\,0\,0 \\ 1\,1\,1\,0 \end{pmatrix}$$

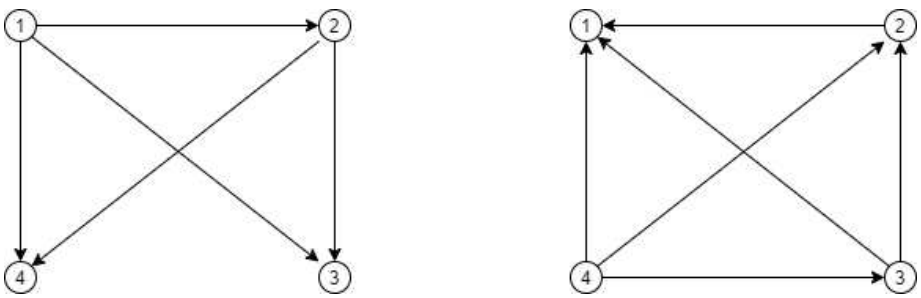


Fig. 18 Directed graph of V_L and V_U of IMPR of Jin et al. (2018)

From figure 18, it is clear that there is no cycle is found from the lower and upper vertex matrix. Both the V_L and V_U are consistent then the IMPR is also consistent. The IMPR P is consistent by our

method. Therefore both the methods gave the same result that proves the validity of our result and also in our method there is no need to solve any optimization problem.

6 Conclusion

In this paper, we have proposed two approaches to get the characterization of the consistency for intuitionistic multiplicative preference relation(IMPR). In the first approaches, we propose an algorithm to check the consistency of IMPR by using the cycles of various length in a directed graph and the same procedure apply for incomplete IMPRs also. The second approach proved isomorphic between the set of IMPRs and the set of asymmetric multiplicative preference relations. Also, consistency property is then checked of asymmetric preference relation using a directed graph that is used to get the consistency of IMPR.

Our future research will study the consistency of hesitant fuzzy preference relation HFPR using the cyclic length of a directed graph with the best additive consistency index, the worst additive consistency index and average additive consistency index of an HFPR (Zhang, Kou, Dong, 2018).

Zhang et al. (Zhang, Kou, Yu, Guo 2018) proposed an approach to deriving a priority weight vector from an incomplete HFPR using logarithmic least square method. Consistency index of an incomplete HFPR is defined. For an unacceptably consistent incomplete HFPR, an automatic algorithm is developed to improve the consistency. Also, we can check the consistency of incomplete HFPR using the cyclic approach that will result are then extended to propose a new procedure for the group analytic hierarchy process to deal with multi-criteria group decision-making problem.

Nowadays, linguistic large scale group decision-making problems are a handy tool in the decision-making process. Zhang et al. (2016) developed a new linguistic computational model with multi-granular linguistic distribution assessments for its application to large scale multi-attributes group decision-making problems with linguistic information. Our future research will study the consistency of IMPR with multi-granular linguistic information based on the developed model. It will be more interesting to analyze the consistency of IMPR with multi-granular linguistic information using the cyclic length of the directed graph of IMPR. Also, our studies can be extended to the social network group decision-making framework (Ureña et al. 2019).

Compliance with ethical standards

Conflict of interest: The authors declare that they have no conflict of interest regarding the publication of this manuscript.

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

References

Behret, H., (2014) Group decision making with intuitionistic fuzzy preference relations, Knowledge-Based Systems. 70, 33–43.

Capuano, N., Chiclana, F., Fujita, H., Herrera-Viedma, E., Loia, V., (2018) Fuzzy group decision making with incomplete information guided by social influence. IEEE Trans. On Fuzzy Systems 26(3) 1704–1718.

Chiclana, F., Herrera-Viedma, E., Alonso, S., Herrera. F., (2009) Cardinal Consistency of Reciprocal Preference Relations: A Characterization of Multiplicative Transitivity, IEEE Transactions on Fuzzy Systems

17(1)14–23.

Chiclana, F., Herrera, F. ,Herrera-Viedma, E., (2001) Integrating multiplicative preference relations in a multipurpose decision-making model based on fuzzy preference relations, Fuzzy Sets and Systems 122 277Ű-291.

Ergu, D., Kou, G., Peng, Y., and Shi, Y.,(2011) A simple method to improve the consistency ratio of the pair-wise comparison matrix in ANP. Eur. J. Oper. Res., vol. 213, pp. 246Ű-259.

Fishburn, P. C., Utility Theory for Decision making. R.E. Krieger pub.Co.(1979).

Gong, Z.W., Li, F.J., Forrest, J., & Zhao, Y., (2011) The optimal priority models of the intuitionistic fuzzy preference relation and application in selecting industries with higher meteorological sensitivity, Expert Systems with Applications. 38, 4394–4402.

Gong, Z. W., Li, L. S., Zhou, F.X., & Yao, T. X., (2009) Goal programming approaches to obtain the priority vectors from the intuitionistic fuzzy preference relations, Computers & Industrial Engineering. 57, 1187–1193.

Herrera-Viedma, E., Chiclana, F., Herrera, F., Alonso, S.,(2007) Group decision-making model with incomplete fuzzy preference relations based on additive consistency. IEEE Trans. Syst. Man Cybern. Ű Part B 37: 176Ű-189.

Herrera-Viedma, E., Herrera, F., Chiclana, F., and Luque, M.,(2004) Some issues on consistency of reciprocal relations, Eur. J. Oper. Res., vol. 154, 98Ű-109.

Jiang, Y., Xu, Z., Yu, X.,(2015) Group decision making based on incomplete intuitionistic multiplicative preference relations. Information Sciences 295: 33–52.

Jin, F., Ni, Z., Pei, L., Chen, H., Li, Y., (2018) Goal programming approach to derive intuitionistic multiplicative weights based on intuitionistic multiplicative preference relations. Int. J. Mach. Learn. & Cyber. 9, 641Ű-650.

Li, C-C, Dong, Y., Xu, Y., Chiclana, F., Herrera-Viedma, E., Herrera, F., (2019) An overview on managing additive consistency of reciprocal preference relations for consistency-driven decision making and Fusion: Taxonomy and future directions. Information Fusion, 52, 143–156.

Nishizawa, K., (1995)A consistency improving method in binary AHP. Journal of Operations Research Society of Japan, vol. 38, No. 1, 21–33.

Nishizawa, K., (1996) A method to find elements of cycles in an incomplete directed graph and its Applications- Binary AHP and Petri Nets. Computers and Mathematics with Application. 33, 33–46.

Orlovsky, S. A, (1978) Decision making with a fuzzy preference relation. Fuzzy Sets Syst. 1: 155–167.

Ren, P., Xu, Z., Liao, H., (2016) Intuitionistic multiplicative analytic hierarchy process in group decision making. Computers & Industrial Engineering, 101 513–524.

Saaty, T. L, (1977) A scaling method for priorities in hierarchy structures. J. Math. Psychol. 15: 234–281.

Saaty, T. L, (1980) The Analytic Hierarchy process. McGraw-Hill, New York.

Ureña, M. R., Chiclana, F., Morente-Molinera, J. A., Herrera-Viedma E., (2015) Managing Incomplete Preference Relations in Decision Making: A Review and Future Trends. Information Sciences 302(1)14–32.

Ureña, R., Kou, G., Dong, Y., Chiclana, F., Herrera-Viedma E., (2019) A review on trust propagation and opinion dynamics in social networks and group decision making frameworks. Information Sciences, Volume 478, 461–475.

Xia, M. M., Xu, Z.S., Chen, J., (2013) Algorithms for improving consistency or consensus of reciprocal [0, 1]-valued preference relations, Fuzzy Sets Syst. 216 108–133.

Xia, M.M., Xu, Z.S., Liao, H. C., (2013) Preference relations based on intuitionistic multiplicative information. IEEE Trans. Fuzzy Syst. 21, 113-Ű133.

Xu, Y., Herrera, F., (2019) Visualizing and rectifying different inconsistencies for fuzzy reciprocal preference relations, Fuzzy Sets and Systems, 362, 85–109.

Xu,Y., Herrera, F., Wang, H., (2016) A distance-based framework to deal with ordinal and additive inconsistencies for fuzzy reciprocal preference relations, Information Sciences, 328, 189–205.

Xu ,Y., Li,M., Cabrerizo ,F. J., Chiclana, F., and Herrera-Viedma, E., (2021) Algorithms to detect and rectify multiplicative and ordinal inconsistencies of fuzzy preference relations, IEEE Transactions on Systems, Man, and Cybernetics: Systems, 51(6), 3498–3511.

Xu,Y., Ma,F., Herrera, F.,(2019) Revisiting inconsistent judgments for incomplete fuzzy linguistic preference relations: Algorithms to identify and rectify ordinal inconsistencies, Knowledge-Based Systems, 163, 305–319.

Xu, Y., Patnayakuni, R., Wang, H. (2013) The ordinal consistency of a fuzzy preference relation, Information Sciences, 224, 152–164.

Xu, Y., Wang,Q., Cabrerizo,F. J., Herrera-Viedma, E., (2018) Methods to improve the ordinal and multiplicative consisetncy for reciprocal preference relation, Applied Soft Computing, 67, 479–493.

Xu, Z. S., (2007b) Intuitionistic preference relations and their application in group decision making. Information Sciences 177 (11) 2363–2379.

Xu, Z. S., (2013) Priority weight intervals derived from intuitionistic multiplicative preference relations. IEEE Trans. Fuzzy Syst. 21: 642–654.

Xu, Z., & Liao, H., (2015) A survey of approaches to decision making with intuitionistic fuzzy preference relations, Knowledge-Based Systems, 80, 131–142.

Xu, Z.S., Wei, C.P., (1999) A consistency improving method in the analytic hierarchy process, Eur. J. Oper. Res. 116, 443–449.

Zhang, Z. and Guo, C., (2017) Deriving priority weights from intuitionistic multiplicative preference relations under group decision-making settings, Journal of the Operational Research Society, 68, (12) 1582–1599.

Zhang, Z., Guo, C., and Martnez, L., (2016) Managing Multigranular Linguistic Distribution Assessments in Large-Scale Multiattribute Group Decision Making, IEEE Transactions on Fuzzy Systems, Man, and Cybernetics: Systems, 47 (11), 3063–3076.

Zhang, Z., Kou, X., Dong, Q., (2018) Additive consistency analysis and improvement for hesitant fuzzy preference relations, Expert Systems With Applications, 98, 118–128.

Zhang, Z., Kou, X., Yu, W., Guo, C., (2018) On priority weights and consistency for incomplete hesitant fuzzy preference relations, Knowledge-Based Systems, 143, 115–126.

Zhang, Z. and Pedrycz, W.,(2018) Goal Programming Approaches to Managing Consistency and Consensus for Intuitionistic Multiplicative Preference Relations in Group Decision-Making, IEEE Transactions on Fuzzy Systems, DOI 10.1109/TFUZZ.2018.2818074.

Zhang, Z. and Pedrycz,W.,(2017) Models of Mathematical Programming for Intuitionistic Multiplicative Preference Relations, IEEE Transactions on Fuzzy Systems, 25 (4) 945–957.

Waste Cooking Oil as Fuel in CI Engine – A Review

Adhirath Mandal and HaengMuk Cho
Department of Mechanical and Automotive Engineering, Kongju National University
Email: adhirath.mandal@gmail.com, hmcho@kongju.ac

***Abstract*— There has been an increasing demand for clean energy in the recent days. Fuels used now days are non-renewable fuels, a major reason for environment pollution. Increasing urbanization, increase the demand and also increase the emission from the conventional fuels which created a demand for efficient and alternative energy source. 1/3rd fuel produced in the world is consumed by the transportation sector and a major contributor to greenhouse gases. Production process of energy should be developed which can cater the demand of environmentally safe, renewable and sustainable. Alternate fuels are being researched because of the increasing demand of the convention no-renewable fuel. There have been many renewable energy, but the immediate replacement is biodiesel. Biodiesel is a replacement of the diesel, which is the major fuel in the transportation sector and consists of alkyl esters. Biodiesel are majorly produced from food crops, a major disadvantage and at the same time it increase the production cost. Recent studies show that biodiesel could be produced from waste cooking oil, this does not pose any threat to the food crops. Use of waste cooking biodiesel in the CI engine reduces the HRR and in-cylinder pressure. Many researches also showed that CO, HC, NO_x emission reduces with increase in blend ratio. This paper gives a review of the use and characteristics of waste cooking oil biodiesel used in compression ignition engine.**

***Index Terms*— waste cooking oil, biodiesel, NOx, CO, HC**

I. Introduction

Requirement for energy has been increasing day by and also the need for clean fuel. As we are progressing the need for clean fuel is becoming crucial and it is a concern. The fuels that are being used now are nonrenewable and harmful to the environment. Because of the conventional fuels the pollution have been on a rise and it has a damaging effect on the environment and on the health of humans. Increasing urbanization, increase the demand and also increase the emission from the conventional fuels which created a demand for efficient and alternative energy source. Production process of energy should be developed which can cater the demand of environmentally safe, renewable and sustainable. It is very important to have a cheap and cost competitive alternate fuels which can replace the arising demand of the conventional energy. It has been clearly seen that the transportation sector plays a curtail role in the use of consumption of crude oil. It was observed that 1/3rd crude oil is being consumed by the same and is the largest contributor of greenhouse gasses emission. In the period 2010 to 2040 it has been predicted by the many researchers that the demand for fossil fuels would increase by 40%, which would be because of the transportation sector. Alternate fuels are being researched because of the increasing demand of the convention no-renewable fuel. There have been many renewable energy, but the immediate replacement is biodiesel. biodiesel is a replacement of the diesel, which is the major fuel in the transportation sector and consists of alkyl esters [1].

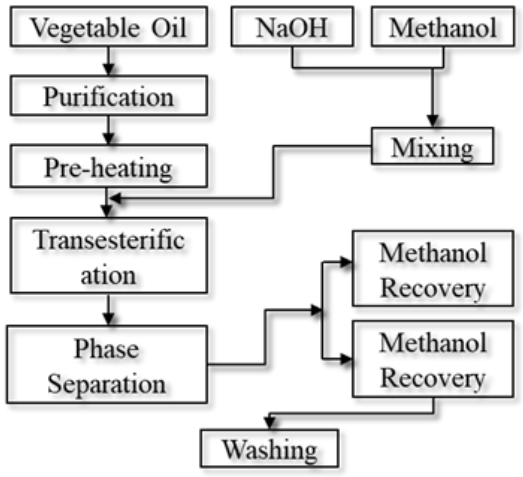


Figure 1. Biodiesel Production process [2]

Biodiesel is also known as methyl/ethyl esters which consists of long chain fatty acids. In the alternative fuels, biodiesel is a popular fuel because of being biodegradable, renewable, nontoxic and free from carbon. Because of similar physiochemical property of biodiesel, it can be replaced by diesel in a diesel engine with small modification or no modification. Biodiesel has a lot of advantages but production of biodiesel from crops is not wise. The process of making biodiesel is shown in Fig. 1. Use of waste cooking oil for production of biodiesel would be economical and a good research area. Gives an review on the use of waste cooking oil biodiesel and its advantages [2][3].

Manuscript received November 4, 2020; revised March 11, 2021.

© 2021 Int. J. Mech. Eng. Rob. Res

doi: 10.18178/ijmerr.10.9.485-492

II. Literature Review

A. New Oil as Biodiesel

Peng Geng et al. studied on the use of biodiesel in large ocean going ships for production of electricity. For the production of electricity with the use of diesel produces a lot of harmful emission. Alternate fuel are being suggested to reduce the emission in the marine diesel engine. Emissions of a 6-cyclinder turbocharged inter-cooling direct-injection marine auxiliary diesel engine, emission and combustion characteristics was studied by using clean waste cooking oil biodiesel, B70 (diesel containing 70 vol.% of biodiesel), B90 and Ultralow Sulfur Diesel. Study on NOx emission was studied with high biodiesel to diesel ratio at different loads of 25, 50 and 75% engine load at 1050to1500rpm speed situations. It was observed that with increase in biodiesel content the in-cylinder pressure decreased with reduction in ignition delay, advance in ignition and longer combustion duration. Maximum heat release rate was reduced, when the engine was operated at low load. The reduction was 14.3% and maximum HRR was 21.3% when the load was high. With increase in engine load, increase in peak heat release rate and cylinder pressure was observed. With increase in biodiesel content, temperature in the exhaust manifold reduces. Use of high substitution ratio of biodiesel, significantly reduces the NOx emission because of cylinder temperature reduction in the diffusion combustion mode. Increase in biodiesel content and engine load, NO2/No emission decreases [4]. **Mag ń Lapuerta et al.** studied the two different alcohol driven biodiesel, methyl ester and ethyl ester. Two different biodiesels and conventional diesel were examined on 2.2L CRDI engine in 100%, 70% and 30% volume basis. Pure biodiesel resulted in increase in fuel consumption compared to reference fuel. With use of biodiesel NOx emission was reduced slightly, but in case of smoke opacity, total hydrocarbon emission, particle emission there was sharp reduction, although volatile organic fraction of the particulate matter increased. Pure biodiesel resulted in increase in fuel consumption compared to reference fuel. With use of biodiesel NOx emission was reduced slightly, but in case of smoke opacity, total hydrocarbon emission, particle emission there was sharp reduction, although volatile organic fraction of the particulate matter increased [5]. **Lijiang Wei et al.** investigated on the emission reduction in a marine diesel engine by the use of biodiesel. Compression ratio, NOx emission and combustion process were studied by the experimental investigation on the use of waste cooking oil biodiesel. Four different types of bled were used in the experimentation, and was observed that increasing the blend ration would decrease in peak of heat release rate. At 1050rpm and 1500rpm speed, PHRR reduction was 11.04% and 19.86%. with increase in biodiesel blend brake thermal efficiency changed little but brake specific fuel consumption increased. With 1050rpm and 1500rpm at B50, different engine load the max increment in brake specific fuel consumption were 7-8.3% and 5.1-6.1%. Rise in biodiesel blend, NOx emission decreased. NOx raised with decline in in-cylinder. With increase in load and engine speed, significant reduction in NO2/NO for in-cylinder cooler regions. Waste cooking oil biodiesel could be considered for auxiliary diesel engine in marine [6]. **X.J. Mana et al.** studied naturally aspirated direct injection engine and focused on the emission , which were fueled with B10, B20, B30, B100 (Biodiesel 100%) & neat diesel. 13 mode test cycle was used to conduct the experiment for the diesel engine so that regulated and unregulated emission influenced by engine speed and engine loads could be identified. For the experimental observations it was seen that HC, particulate matter mass concentration and CO decreased whereas NOx increased with increase in biodiesel blend ratio. Increase in the biodiesel blend, xylene and toluene emission reduced as well as benzene emission also increased. As benzene increased was balanced by decrease in toluene and xylene, BTX emission was not changed. Emission were affected by the load and engine speed [7]. **Cheng Tung Chong et al.** experimented by using pool flames using extinction calibrated laser induced-incandescence (LII) and pre-vaporised diffusion jet flames under total carbon flow to compare and quantify the soot volume fraction of waste cooking oil biodiesel and its bled. Soot is formed highest at the flame and convected downstream for diesel fuel, but in the condition of biodiesel, soot volume fraction is the center of the flame was evenly distributed. Soot tendency reduced by increasing the biodiesel fraction which was visible in vapour flame and pool flame. SVF profiles for pool flame was noticed when radial profile of SVF was compared along the centerline, residence time for soot diffusion and growth compared to vapour flame was reflected, this showed that the pool burner had low mass flux. At higher temperature the diesel rich fuel flame soot was found which convert quickly upwards. 62.9nm (pool) for biodiesel and 51.6 for diesel. Soot produced was higher in the vapour flame than pool flame by a factor of two, mass consumption rate for same fuel. Soot volume fraction for waste cooking oil biodiesel was lowest irrespective of the flame type and owing to fuel chemistry and aromatic compound absent. The diameter of soot primary particle size was 1.5 lower mean diameter for waste cooking oil biodiesel than diesel produced soot. Large diameter particles of 22% was produced for diesel by pool flame and 8% for biodiesel [8].

Y. Zhang et al. studied the production of biodiesel from fresh and waste cooking vegetable oil, 4 different production ways. The four unlike methods were implemented to assess the technological gains and limitations. It was observed that fresh vegetable cooking oil required less equipment for the process but the raw material cost was high, alkali-catalysed process. Conversion of waste cooking oil by acid catalysed process was comparatively easy compared to alkali catalysed, at the same time use of waste cooking oil was cost effective and reduces raw material cost. Which makes waste cooking oil biodiesel produced by alkali catalysis, a good alternate fuel [9]. **Emilio A. Viornery-Portillo et al.** compared the environmental impact on the use of ultra low sulphur diesel and B25 biodiesel in a power generator of 33kW. It was seen from the result that

© 2021 Int. J. Mech. Eng. Rob. Res

the use of waste cooking oil biodiesel (B25) showed better performance than ultra-low sulphur diesel. Abiotic depletion by 39.48%, acidification potential by 38.73%, eutrophication potential by 39.24%, human toxicity by 39.44% and global warming by 35.77% were the major reduction. 41.54% NOx emission, CO by 52% reduced, compared to ultra-low Sulphur diesel. Harmful emission would reduce by the use of Biodiesel [10]. **Jehad Yamin et al.** Researched on the engine performance and mapping. He used diesel and used cooking oil biodiesel fuel. For the test 1500cc CI engine testbed was used and performance parameter were measured. Load varied from no load to full load, speed of crankshaft was 35 & 60rps and constant temperature of the coolant at 70oC. With the help of the test results the simulation model was adjusted. With biodiesel fuel there was slight improvement in the operation of the engine. It was also observed that when using biodiesel combustion was in the mixing phase. Use of potassium hydroxide as catalyst in equivalent to sodium hydroxide, yield was more by 45% which resulted in lower viscosity by 10% and calorific value by 5%. Result from mapping showed that thermal efficiency was lower than diesel. Engine when fueled with diesel, fuel consumption was low of about 250-300 g/kWh at mean effective pressure value and higher power but for biodiesel the fuel consumption was 350 g/kWh at low engine speed, mean cylinder pressure and power. Biodiesel has higher cetane number so it burns earlier than diesel [11]. **Mohammed S. Edam.** Researched on the use of caster methyl ester with diesel in different proportions and examined the performance, combustion characteristics and emission of a single cylinder engine. Diesel-rk simulation program was used to study numerically about the biodiesel. With increasing percentage of CME, near the top dead center peak pressure was closer. With increased blending the, brake specific fuel consumption also increased. NOx emission was higher for all the biodiesel blends of CME and diesel. Smoke level also decreased with increasing the biodiesel blend ratio. For B10, B20, B30, B50, B70 and B100, smoke level were reduction were, 15.25%, 35.3%, 40.7%, 45.71%, 49.43% and 52.73% respectively. The best blend was B20, have reduced CO emission and slight variation in performance in comparison to conventional diesel [12]. **Alireza Shirneshana. [13]** Experimented on the use of alternative fuel (biodiesel) in a 4 cylinder direct injection diesel engine at different engine load, speed of 1800 rev/min with different. Four different fuel blends (B20, B40, B60 and B80) were used for the test in an engine. The biodiesel was prepared from waste fry oil. Emission was lower (HC and CO) & NOx, CO2 increased when used biodiesel at higher load. Contrary the CO and HC were higher at lower load. Study showed that biodiesel compared to diesel showed lower exhaust emission. Combustion improves with addition of biodiesel as it increases the oxygen level of the fuel [13].

B. Waste Cooking Oil as Biodiesel

Sahar et al. reduced the cost of producing biodiesel by utilizing waste cooking oil, but as free fatty acids are high transesterification process is difficult. Waste cooking oil had acid value was 5.5 mg KOH/g, this also showed that FFA was high. By the use of HCL, H3PO4, & H2SO4 (all acidic catalyst), esterification of waste cooking oil was carried out, this method reduced the FFA by 88.8% at 600C with methanol to oil molar ratio by 1:2.5. 90% efficiency was achieved by the presence of KOH and fatty acid methyl ester transesterification at 500C at 1% catalyst. 1:3 methanol to oil archived 94% FAME yield at 60 oC reaction temp and 1% catalyst. The main component to produce biodiesel were stearic acid, oleic acid, palmitic acid and linoleic acid. Pretreatment of waste cooking oil biodiesel by acid followed by catalyzed reaction was a good method for treatment of waste cooking oil [14]. **L. Wei et al.** investigated the particulate emission, gaseous emission and combustion of a diesel engine which was fuelled by used cooking oil biodiesel. Different test cycles were used on a DI diesel engine with 20%, 50% and 75% biodiesel. With the use of biodiesel, in-cylinder pressure increases, maximum heat release rate is reduced and combustion & ignition delay in shortened. With the use of biodiesel the break specific fuel consumption also increased. There was no substantial change in break thermal efficiency. 1,3-butadiene, benzene, formaldehyde, ethane, propene and acetaldehyde increased the weighted brake specific emission but toluene and xylene reduced the weighted brake specific emission. The increase in weighted brake specific was because of the accelerated pyrolysis of long chain molecules in higher combustion temperature and in-cylinder high pressure. The decrease was reasoned as toluene and xylene have oxygen which encourages oxidation reaction. Ozone formation was not observed for any significant changes for the unregulated gaseous emissions for the test fuels. With use of biodiesel it was observed that weighted particle mass concentration and GMD was reduced but weighted total number concentration. Linear regression was conducted for particulate matter and weighted unregulated gases, results showed that correlation coefficient more than 0.9 was obtained which proven that biodiesel content was proportional to PM emission and unregulated gaseous [15]. **Xianbao Shen et al.** measured emission of on-road diesel vehicle which was fuelled with waste cooking oil biodiesel, with the love of portable emission measurement system. 2 LDDT truck and 2 HDDT truck was filled with 4 mixed fuels with different blend ratio. The ratio of blend was 0%, 5%, 20% and 100%. With the use of biodiesel + conventional diesel it was observed that there was no decrease in fuel consumption. Using waste cooking oil biodiesel and mixing with diesel helps in reducing emission, gives energy security and food security as well. With increasing biodiesel content in the mixture it was observed that HC, NOx, CO and PM2.5 decreases. In all the blends, emission was lowest in highway condition than non-highway. Biodiesel blend which were below 12.5% showed better emission characteristics (especially B5) than neat diesel in non-highway conditions. CO, NOx, HC and PM2.5 were reduced by increasing blend ratio [16]**. Mangesh G. Kulkarni et al.** discussed on the use of biodiesel form

waste cooking oil. Fatty acid methyl ester biodiesel is an alternative fuel which is biodegradable and nontoxic and produced form renewable source. The major difference in the commercialization of biodiesel is that the manufacturing cost and raw material cost is more than conventional diesel. The only low cost biodiesel production technique is by used cooking oil. As the cooking oil are being used for frying, which forms polymerized triglycerides and free fatty acid which affects the property of the biodiesel and transesterification reaction also. Transesterification process must be selected in the free fatty acid and water content. If FFA and water are <1 wt% and <0.5 wt% then alkaline catalyst is suitable, if FFA is more than 1 wt% than acidic catalyst should be selected. These catalyst were not recommended for transesterification because of high molar concentration, corrosion problem and high catalyst concentration, for waste cooking oil. Biodiesel processing would become costly if two-step (acid catalyzed and alkaline catalyzed step) process is adopted, which is also not friable. For chemical catalyzed reaction, enzyme-catalyzed transesterification is a good option which must be developed for commercialization. For the production of biodiesel from waste cooking oil, potential in catalyst-free supercritical methanol method was observed. But high molar ratio, pressure and temperature make it difficult for industrial scale. Emission are less and performance of the engine is better when commercial engine was run on waste cooking oil biodiesel, except NOx and carbon residue [17].

Yuan-Chung Lin et al. discussed on the cost of commercialization of biodiesel. The costliest part of the biodiesel is the raw material. To reduce the cost of biodiesel, raw material cost has to be reduced, which could be reduced by the use of waste cooking oil. Another problem is solve, i.e waste disposal by the use of waste cooking oil. The investigation was carried out on a heavy duty diesel engine to calculate brake specific fuel consumption, polycyclic aromatic hydrocarbons & carcinogenic potencies and regulated matter, under US-HDD transient cycle. The investigation was carried out on 5 different test fuels, 1. ultra-low sulfur diesel (ULSD), 2. WCOB5, 3. WCOB10, 4. WCOB20, and WCOB30. Results showed that blend of waste cooking oil biodiesel and ULSD showed reduced CO, PM and HC emission with respect to ULSD. CO lowered by 3.33%-13.1%, PAHs decreased by 7.53%-37.5%, HC reduced by 10.5%-36.0% and PM reduced by 5.29%-8.32% by use of ULSD/WCOB blend compared to ULSD.

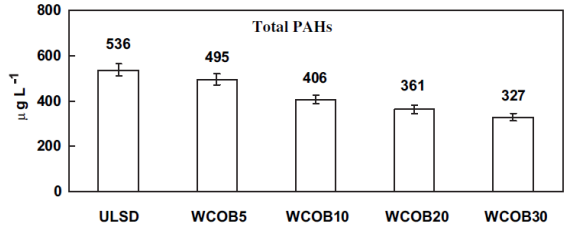


Figure 2. PAH emission for different blends [18]

With the use of waste cooking oil CO, HC and PM reduced and at the same time combustion efficiency increased. PAH in WCOB is not present because of which the biodiesel blend with high fraction resulted in lower PAH emission as shown in figure 2. Volumetric consumption was smaller between biodiesel a diesel because volumetric injection was compensated by high density biodiesel [18]. **Joonsik Hwang et al.** investigated on the use of waste cooking oil biodiesel in a CRDI single cylinder diesel engine to study the effect of injection timing and injection pressure on the emission and combustion characteristics and compared it with commercial diesel. Fuel property were compared to the conventional fuel including fatty acid composition.80 MPa and 160Mpa was the selected injection pressure in the test engine and 25 and 0 crank angle were selected for injection timing after top dead center at 2 different engine load. In casse of biodiesel the indicated specific fuel consumption in respect to injection timing was higher than diesel in all cases.

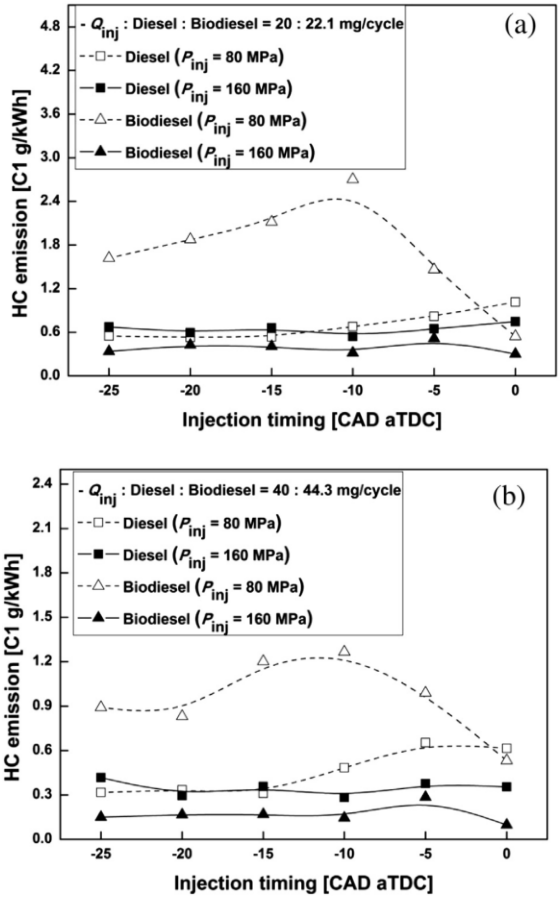


Figure 3. HC emission with injection timing at different load (a) low load, (b) high load [19].

For biodiesel, lower peak heat release rate and peak cylinder pressure was observed but longer ignition delay was seen in all conditions. HC emission comparison of different fuel with different load condition could be seen in Fig. 3. Hydrocarbon, carbon monoxide emission and smoke as reduced with the use of biodiesel especially at high injection pressure. NOx was comparatively higher for biodiesel compared to diesel [19]. **Ali M.A. Attia et al.** waste cooking oil methyl ester was produced by

© 2021 Int. J. Mech. Eng. Rob. Res

transesterification process. Different waste cooking oil biodiesel blends were tested on the engine performance. Soot tendency and viscosity analysis was carried out, by the use of waste cooking oil and its blend. In-cylinder peak pressure was determined by the engine load and blend ratio of biodiesel and diesel. Use of B20 showed best value of brake thermal efficiency and brake specific energy consumption. With blend ratio from B20 to B50, all were best for the engine environmentally. BSEC increased by 8% on use of 100% biodiesel, smoke opacity increased by 15%, NOx reduced by 10%, HC reduced by 15% and CO reduced marginally. 30% to 50% blending ratio was preferred for improvement in emission and engine performance [20]**. S. Senthur Prabu et al.** investigated on the use of hydroxyl-toluene and n-butanol additive with waste cooking oil. Exhaust emission and engine performance was evaluated. Sodium hydroxide, sulphuric acid and methanol was used as catalyst reaction for transesterification process of waste cooking oil to biodiesel. B0, B20, B30, B40 & B100 biodiesel blends were used to test the engine performance but the best blend was B30. Performance of B30 was meliorated than other blends because of the addition of n-butanol 20% & antioxidants of BHT 2000. BTE was 4.6% low and BSFC was 7.3% higher for B30 +butylated hydroxytoluene than diesel. Heat release were same as diesel for B20,B40 and B30+BHT blend but B30+butanol at higher load showed higher heat release rate than diesel. B30+n-butanol showed higher EGT (2.3%) & NOx (9%) compared to diesel fuel but Co emission was less by 37.5%. Compared to diesel, 100% biodiesel had 51% lesser exhaust gas temperature. Performance of the engine on BTE and BSFC reduces steadily as the biodiesel fraction increase in the blend as shown I Fig. 4 & Fig. 5. B30 showed better emission characteristics than diesel fuel.

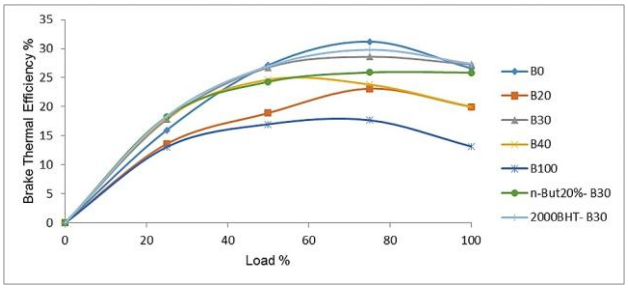


Figure 4. Brake thermal efficiencies of different blends at different loads [21].

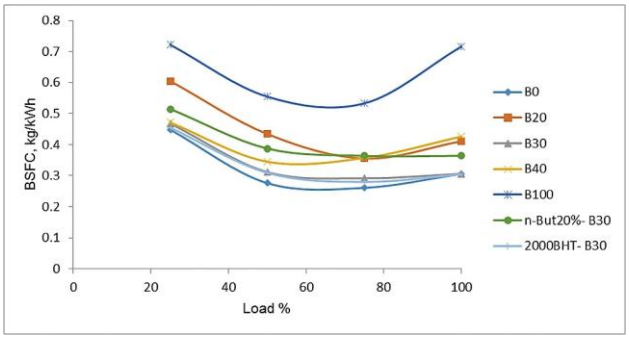


Figure 5. Brake specific fuel consumption of different blends at different loads [21]

A. Rajesh et al. discussed in the importance of alternative fuel, generated from waste and also help in reducing emission and environmental hazards. Investigation on waste cooking oil biodiesel was carried out to find the effect of anisole fraction. Compared result of biodiesel and baseline diesel with the waste cooking oil biodiesel with 30% vol. anisole blending. Ignition delay period is extended by increasing the anisole content in the biodiesel and also the heat release rate and in-cylinder pressure increase with increase in anisole fraction. Waste cooking oil biodiesel 90% and anisole 10% test fuel showed better brake thermal efficiency than biodiesel. At peak load NOx emission was lowest when the engine was fueled with W90A10 biodiesel, but with Diesel and biodiesel it decreased by 17% and 11%. Biodiesel had highest smoke opacity, soot was suppressed with the addition of anisole. Anisole also helped in reduction of CO and HC emission. Waste cooking oil methyl ester blend with anisole 10% was efficient [22].

M.S. Gad et al. researched on the use of waste cooking oil in conventional diesel engine. It was observed that, use of waste cooking oil biodiesel generated a lot of NOx emission. The biodiesel also had low volatility, high viscosity and high pour point. The impact of using gasoline with waste cooking oil biodiesel as an additive and its effect on combustion, emission, and exergy characteristics of a diesel engine run under various loads and a constant speed of 1500 rpm.

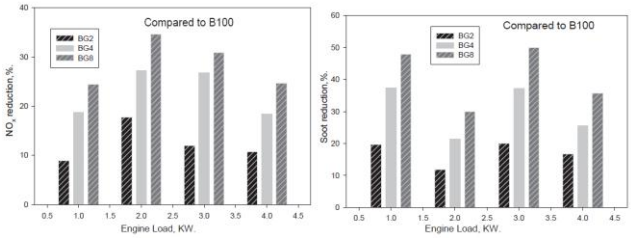


Figure 6. NOx emission at different engine load [23]

With the help of ultrasonic and mechanical dispersion, transesterification was carried out to produce waste cooking oil biodiesel and by applying GC-MS and FTIR the biodiesel was characterized. For BG2, BG4 and BG8 the viscosity was reduced by 5%, 11% and 21% respectively. Addition of gasoline with WCO, enlarges the Cylinder pressure and HRR. Reduction in emission was observed, CO by 25%, UHC by 30%, NOx by 20% and smoke opacity by 30% compared to pure waste cooking oil biodiesel as shown in Fig. 6. It was observed that by blending biodiesel with gasoline in small dose, improves the physical [23].

R. Selvaraj et al. Produced waste cooking oil biodiesel by transesterification with methanol. To increase the fatty acid methyl esters production, waste cooking oil was pre-treated and catalyst potassium methoxide 1% was used. Activated charcoal 2% was used to adsorb the free fatty acid. RSM method was used to optimize the process variables and Fame yield was predicted by ANN. Under the optimum condition the maximum conversion was 95%, time 60s, alcohol to oil

ratio 6:1. Concentration of catalyst 1% and 75oC temperature. ANN predicted 0.99 and RSM 0.98 yield of biodiesel shown in Fig. 7. Proton nuclear magnetic resonance, Fourier transformation infrared spectroscopy, carbon nuclear magnetic resonance and gas chromatography-mass spectrometry was use to examine FAME. 13CNMR for FAME from WCO, agreed to the presence of methyl esters in biodiesel at 174.2 ppm and 51.4ppm [24].

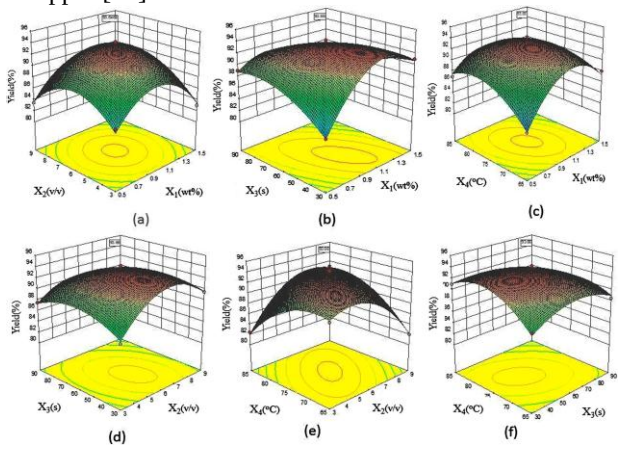


Figure 7. Response surface plots showing the effects on yield of FAME (%) (X1: Catalyst (wt %), X2: Alcohol: Oil ratio (v/v), X3: Time (s) and X4: Temperature (℃)) [24].

Mohamed F. Al-Dawody et al. investigated on the use of methyl ester of Waste cooking oil collected from different restaurants. Different blend ratio of B10, B20 and B100 on volume basis was used for testing in a diesel engine at constant speed. Diesel-rk simulation software was used to verify the experimental data and was found that both the data agree with each other. It was observed that with increasing the ration of methyl ester of waste cooking oil, maximum pressure reduced due to reduction in heating value of the fuel. NOx emission was also reduced with the use of methyl ester waste cooking oil. B20 was best ratio in terms of performance of the engine as show in Fig. 8 & Fig. 9 [25].

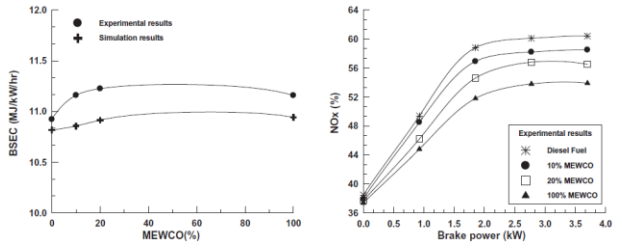


Figure 8. (a) MEWCO% v/s BSEC, (b) Load v/s NOx emissions [25]

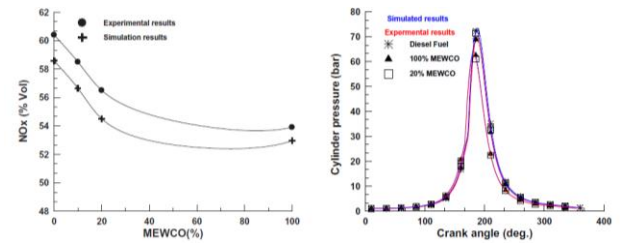


Figure 9. (a) MEWCO% v/s NOx emission, (b) Crank angle v/s cylinder pressure [25]

Mohd Affandi Mohd Ali et al. Used waste cooking oil as biodiesel which was produced by transesterification by microwave irradiation and activated limestone was used as catalyst. The screening was carried out by using response surface methodology and two level factorial design. Wet-impregnation was used to prepare the catalyst, which was characterized for porosity, surface element, surface area and surface morphology. Continuous microwave assisted reactor was used to execute the reaction and conversion was measured by gas chromatography. Waste cooking oil conversion was affected by methanol to oil molar ratio, catalyst loading and reaction time. At 5.47 wt% catalyst loading, methanol to oil molar ratio was 12.21:1 with reaction time of 55.26min, oil to biodiesel conversion was 96.65%. Time for transesterification was reduced by 77% by the use of CMAR. The biodiesel produced showed that particulate matter and NOx was reduced compared to diesel [26]. **Mohamed Shameer et al.** Investigated on the combustion, emission and performance characteristics of a diesel which was fueled by waste cooking oil biodiesel, animal fat biodiesel and camphor oil. All the alternate fuel were mixed in the ratio of This research paper aims at investigating the performance, emission and combustion characters of diesel engine with 20% volume concentrations of animal fat based biodiesel (AFO20), waste cooking oil biodiesel (WCO20), camphor oil (CMO20) and also including pure diesel fuel (D100). The peak cylinder pressure and heat release rate of biodiesel was about 4.82% higher and 13.49% lower than those of diesel fuel on average respectively. Start of combustion of alternate blends happened at earlier crank angles compared to base fuel. Combustion duration of all alternate fuel blends is higher than those of diesel at all load conditions. While fueling CMO20, AFO20 and WCO20, the NOx concentration in the emission shows 7.52%, 10.352% and 16.405% increment respectively with the biodiesel addition to diesel. However, significant reduction in NOx of about 43.8% was observed for camphor oil - diesel blend when compared to other biodiesel blends. The correlation between NOx emission level and in-cylinder temperature was premeditated by a novel procedure using thermal imager. The result shows that the increase in incylinder temperature contributed to the augmentation in NOx concentration. The main aim of this paper is to present the performance, emission and combustion characteristics of fuel blends obtained from camphor oil and bio-waste resources like animal fat residue oil and waste cooking oil. Also a novel procedure of usage of thermal imager to validate the impact of in-cylinder temperature on NOx formation in the compression ignition direct injection diesel engines fuelled with various blends has been insisted. Performance analysis resulted in lower BSFC of CMO20 among the other biodiesel blends. BTE scale for all biodiesel blends shows declining trend than those of diesel. AFO20, WCO20, CMO20 recorded 14.63%, 23.8% and 7.953% lower BTE when compared to diesel. On the whole camphor oil blend displays better performance than other blends, due to its lower viscous

property. Combustion analysis depicted that all alternate fuel blends exhibit the start of combustion at earlier crank angles. Ignition delay period decreases with increase in load for all blends. Camphor oil blends resulted in 3.61% lower PCP, 19.32% lower HRR and 22.22% lower ID than those of diesel fuel at full load condition. Combustion duration of all fuel blends is higher than the duration period of diesel, among which camphor oil presents the lowest CD due to its better atomization attributed to its lower viscosity. According to the emission results, NOx increased with increasing biodiesel concentration in all fuel blends. Camphor oil shows the lowest NOx values among the alternate fuels; meanwhile biodiesel obtained from waste cooking oil contributed higher NOx emission due to its higher oxygen content. On an average, NOx concentration for AFO20, WCO20 and CMO20 are 12.08%, 18.24% and 7.52% higher than those of diesel fuel respectively. On an average, the cylinder head temperature (CHT) for AFO20,WCO20 and CMO20 are 3.75%, 7.16% and 1.08% higher than those of diesel fuel respectively. Thus the increment in CHT communicates with higher NOx. Hence it could be concluded that Camphor oil blend shows more positive results in all performance, combustion and emission aspects than the other biodiesels derived from animal fat residue oil and waste cooking oil. Camphor oil blend at 20% volume concentration is appropriate for diesel engine without any engine modification. A novel procedure of emission investigation has been proposed by incorporating NOx emission concentration with cylinder head temperature measured by thermal imager. Future scope of this work is calibration and optimization of cylinder head temperatures from thermal imager for measuring NOx emission variation of different biodiesel [27].

III. Conclusion

Clean environment is a need of every one and is the major concern. The energy source used nowadays are non-renewable. Increasing urbanization, increase the demand and also increase the emission from the conventional fuels which created a demand for efficient and alternative energy source. In the Alternate fuels, Biodiesel is extremely popular because of non-toxic, biodegradable, carbon neutrality and renewable and same properties like diesel, it could be replace conventional fuel little modification or no modification. There are a lot of advantages of using biodiesel, but from sustainability point of view it was not advisable to produce biodiesel from food crops. Generation of biodiesel from the food crops would not be a good idea and at the same time it would not be economical. Therefore use of waste cooking oil to generate biodiesel would be economical and at the same time reduce the waste from the environment. Following could be concluded on the use of waste cooking oil biodiesel in compression ignition engine:

- Transesterification process must be selected in the free fatty acid and water content. If FFA and water are <1 wt% and <0.5 wt% then alkaline catalyst is suitable, if FFA is more than 1 wt% than acidic catalyst should be selected.
- Biodiesel blend increase, in-cylinder pressure decreased, reduced ignition delay, advance in ignition and longer combustion duration.
- Maximum heat release rate reduced, at low loads. The reduction in maximum HRR was 21.3% at high load. Engine load increases the peak heat release rate and cylinder pressure.
- The CO, HC, NO_X and $PM_{2.5}$ emissions for all vehicles reduced with increasing biodiesel blend.
- CO reduced by 13.1%, PAHs decreased by 37.5%, HC reduced by 36.0% and PM reduced by 8.32% by use WCOB blend compared to ULSD.
- BSEC increased by 8% when 100% biodiesel fuel was used, at different load.
- Engine smoke opacity increased by 15% but NOx reduced by 10%, unburnt HC reduced by 15% and CO reduced marginally.
- Addition of oxygenates to Diesel 20% - WCO biodiesel 20% reduces the smoke opacity, 75% smoke emission lower with octanol blend. CO and HC emission also declined.

It was clearly visible from the survey that the use of waste cooking oil biodiesel blend helps in reducing different emission but in some conditions increase the BSEC. Maximum heat release rate also increases at high loads. It would be recommended to develop biodiesel from waste cooking oil with higher fatty acids. Use of Waste cooking oil as biodiesel has to be researched in detail. Though the use of waste cooking oil reduces the emission but was to reduce the engine load and BSEC. In many conditions NO_x emissions increases. There for use of additives with waste cooking oil biodiesel would be an interesting topic and is an unexplored area. Impact of additives on the emission and engine performance has to be studied. hence, this review paper provides relevant knowledge on the waste cooking biodiesel to be used in CI engine.

Acknowledgement

This work was supported by The National Research Foundation of Korea (NRF) grant funded by the Korea Government (MIST) (NRF-2019R1A2C1010557)

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Alex Tangy, Indra Neel Pulidindi, Nina Perkas, Aharon Gedanken, "Continuous flow through a microwave oven for the large-scale production of biodiesel from waste cooking oil", *Bioresource Technology*, Vol. 224, pp. 333–341, October 2016.

[2] Arvind Kumar Madheshiya, Ajitanshu Vedrtnam, "Energy-exergy analysis of biodiesel fuels produced from waste cooking oil and mustard oil," *Fuel*, vol. 214, pp. 386–408, November 2017.

[3] Mortaza Aghbashlo, Meisam Tabatabaei, Soleiman Hosseinpour, "On the exergoeconomic and exergoenvironmental evaluation and optimization of biodiesel synthesis from waste cooking oil (WCO) using a low power, high frequency ultrasonic reactor," *Energy Conversion and Management*, Vol. 164, pp. 385–398, March 2018.

[4] Peng Geng, Hongjun Mao, Yanjie Zhang, Lijiang Wei, Kun You, Ji Ju, Tingkai Chen, "Combustion characteristics and NOx

emissions of a waste cooking oil biodiesel blend in a marine auxiliary diesel engine," *Applied Thermal Engineering*, vol. 115, pp. 947–954, January 2017.
[5] Magín Lapuerta, José M. Herreros, Lisbeth L. Lyons, Reyes García-Contreras, Yolanda Briceño, "Effect of the alcohol type used in the production of waste cooking oil biodiesel on diesel performance and emissions," *Fuel*, vol. 87, pp. 3161–3169, June 2008.
[6] Lijiang Wei, Rupeng Cheng, Hongjun Mao, Peng Geng, Yanjie Zhang, Kun You, "Combustion process and NOx emissions of a marine auxiliary diesel engine fuelled with waste cooking oil biodiesel blends," *Energy*, vol. 144, pp. 73-80, December 2017.
[7] X.J. Mana, C.S. Cheung, Z. Ning, L. Wei, Z.H. Huang, "Influence of engine load and speed on regulated and unregulated emissions of a diesel engine fueled with diesel fuel blended with waste cooking oil biodiesel," *Fuel*, pp. 41–49, April 2016.
[8] Cheng Tung Chong, Bo Tian, Jo-Han Ng, Luming Fan, Shiyao Ni, Kang Yao Wong, Simone Hochgreb, "Quantification of carbon particulates produced under open liquid pool and prevaporised flame conditions: Waste cooking oil biodiesel and diesel blends," *Fuel*, vol. 270. pp.1-10. March 2020.
[9] Y. Zhang, M.A. Dub, D.D. McLean, M. Kates, "Biodiesel production from waste cooking oil: 1. Process design and technological assessment," *Bioresource Technology*, vol. 89, pp. 1–16, January 2003.
[10] Emilio A. Viornery-Portillo, Brenda Bravo-Díaz, Violeta Y. Mena-Cervantes, "Life cycle assessment and emission analysis of waste cooking oil biodiesel blend and fossil diesel used in a power generator," *Fuel*, vol. 281, pp. 1-10, March 2020.
[11] Jehad Yamin, Ahmad Jehad Abu Mushref, "Performance and mapping of direct injection diesel engine using waste cooking oil biodiesel fuel," *Advances in Mechanical Engineering*, Vol. 11(5), pp 1–11, September 2020
[12] Mohammed S. Edam, Mohamed F. Al-Dawody, "Numerical Simulation for the Effect of Biodiesel Addition on the Combustion, Performance and Emissions Parameters of Single Cylinder Diesel Engine," *Al-Qadisiyah Journal For Engineering Sciences*, Vol. 12, pp 072–078, June 2019
[13] Alireza Shirneshana, "HC, CO, CO2 and NOx Emission evaluation of a diesel engine fueled with waste frying oil methyl ester," *in Proceedings 2nd International Conference on Leadership, Technology and Innovation Management*, Vol. 75, pp. 292 – 297. Istanbul, Turkey, October 2013.
[14] Sahar, Sana Sadaf, Javed Iqbal, Inam Ullah, Haq Nawaz Bhatti, Shazia Nouren, Habib-ur-Rehman, Jan Nisar, Munawar Iqbal, "Biodiesel production from waste cooking oil: An efficient technique to convert waste into biodiesel," *Sustainable Cities and Society*, vol. 41, pp. 220–226, May 2018.
[15] L. Wei, C.S. Cheung, Z. Ning, "Influence of waste cooking oil biodiesel on combustion, unregulated gaseous emissions and particulate emissions of a direct-injection diesel engine," *Energy*, vol. 127, pp. 175-185, March 2017.
[16] Xianbao Shen, Jiacheng Shi, Xinyue Cao, Xin Zhang, Wei Zhang, Hui Wu, Zhiliang Yao, "Real-world exhaust emissions and fuel consumption for diesel vehicles fueled by waste cooking oil biodiesel blends," *Atmospheric Environment*, vol. 191, pp. 249–257, August 2018.
[17] Mangesh G. Kulkarni, Ajay K. Dalai, "Waste Cooking Oils-An Economical Source for Biodiesel: A Review," *Ind Eng Chem Res*, vol. 45, pp. 2901-2913, March 2006.
[18] Yuan-Chung Lin, Kuo-Hsiang Hsu, Chung-Bang Chen, "Experimental investigation of the performance and emissions of a heavy-duty diesel engine fueled with waste cooking oil biodiesel/ultra-low sulfur diesel blends," *Energy*, vol. 36, pp. 241-248, November 2010.
[19] Joonsik Hwang, Donghui Qi, Yongjin Jung, Choongsik Bae, "Effect of injection parameters on the combustion and emission characteristics in a common-rail direct injection diesel engine fueled with waste cooking oil biodiesel," *Renewable Energy*, vol. 63, pp. 9-17, March 2014.
[20] Ali M.A. Attia, Ahmad E. Hassaneen, "Influence of diesel fuel blended with biodiesel produced from waste cooking oil on diesel engine performance," *Fuel*, vol. 167, pp. 316–328, November 2015.
[21] S. Senthur Prabu, M.A. Asokan, Rahul Roy, Steff Francis & M.K. Sreelekh, "Performance, combustion and emission characteristics of diesel engine fuelled with waste cooking oil bio-diesel/diesel blends with additives," *Energy*, vol. 122, pp. 638-648, January 2017.
[22] A.Rajesh, K. Gopal, De Poures Melvin Victor, B. Rajesh Kumar, A.P. Sathiyagnanam & D. Damodharan, "Effect of anisole addition to waste cooking oil methyl ester on combustion, emission and performance characteristics of a DI diesel engine without any modifications," *Fuel*, vol. 278, pp. 1-10, June 2020.
[23] M.S. Gad, Ahmed I. EL-Seesy, Ali Radwan, Zhixia He, "Enhancing the combustion and emission parameters of a diesel engine fueled by waste cooking oil biodiesel and gasoline additives," *Fuel*, Vol. 269, pp. 1-13, March 2020
[24] R. Selvaraj, I. Ganesh Moorthy, R. Vinoth Kumar, V. Sivasubramanian, "Microwave mediated production of FAME from waste cooking oil: Modelling and optimization of process parameters by RSM and ANN approach," *Fuel*, Vol. 237, pp. 40–49, October 2018
[25] Mohamed F. Al-Dawody, Ali A. Jazie, Hassan Abdulkadhim Abbas, "Experimental and simulation study for the effect of waste cooking oil methyl ester blended with diesel fuel on the performance and emissions of diesel engine," *Alexandria Engineering Journal*, vol. 58, pp. 9–17. January 2019
[26] Mohd Affandi Mohd Ali, Jolius Gimbun, Kun Lu Lau, Chin Kui Cheng, Dai-Viet N. Vo, Su Shiung Lam, Rosli Mohd Yunus, "Biodiesel synthesized from waste cooking oil in a continuous microwave assisted reactor reduced PM and NOx emissions," *Environmental Research*, vol. 185, pp 1-10, April 2020
[27] P. Mohamed Shameer, K. Ramesh, "Experimental evaluation on performance, combustion behavior and influence of in-cylinder temperature on NOx emission in a D.I diesel engine using thermal imager for various alternate fuel blends," *Energy*, vol. 118, pp. 1334-1344, November 2016

Copyright © 2021 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC-ND 4.0), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.



Adhirath Mandal has completed his bachelors in Mechanical & Automation Engineering from Guru Gobind Singh Indraprastha University, Delhi, India. After his graduation he has worked in a water RO Company for 1 year. He completed his M.Tech (masters) in production engineering from Delhi Technological University, Delhi, India. He has worked as Assistant Professor in engineering colleges in Delhi NCR. He has teaching experience of 3.5 year. He is pursuing his doctoral degree from Kongju National University, Cheonan, Korea in mechanical engineering. His area of interest is Biodiesel and electric vehicles. He has published papers in national and international journals. The papers are indexed in Scopus.



Professor HaengMuk Cho is a renowned professor in South Korea. He is working as professor at Kongju National University in South Korea. He did Post Doctorate Fellowship from Department of Automotive Engineering of Loughborough University in UK and North Texas University in USA. He earned his PhD from Hanyang University, South Korea in 1997 and completed his M.D. in Environment Engineering from Yonsei University in S Korea. He is working on national and international projects in the field of Bio-fuels for IC engines and renewable energy. Professor HaengMuk Cho is working in the field of Renewable Energy and Biofuels. He is also a member of many societies, like Korea Society of Automotive Engineers, Korean Society of Energy (Korea), Korea Society of Marine Engineering (Korea) & Korea Society of Environment Engineer (Korea). He has published many national and internal research papers. Most of his papers are indexed in Scopus and SCI.

Zero-Shot Learning with Missing Attributes using Semantic Correlations

Ali Braytee[†], Mohamad Naji[†], Ali Anaissi[∓], Kunal Chaturvedi[‡], Mukesh Prasad[†]
[†] School of Computer Science, University of Technology Sydney, Australia
[∓] School of Computer Science, The University of Sydney, Australia
[‡] Department of Information Technology, Delhi Technological University, India
Email: Ali.Braytee@uts.edu.au

2021 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-3900-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/IJCNN52387.2021.9533591

***Abstract*—Zero-shot learning (ZSL) aims to recognize instances belonging to unseen categories which are not available at training time. Previous ZSL models learn a projection function from the visual feature space to a semantic space which contains a description of the categories. The semantic attributes are often correlated with each other at the semantic space and it is not appropriate to learn them independently. Existing ZSL methods are designed to work on complete descriptions of the semantic attributes. However, because these attributes are human-designed values, they might be incomplete or contains noisy values which may affect the recognition performance of many existing ZSL models. This paper proposes a novel zero-shot learning approach (ZSL-MSA) to handle missing and noisy semantic attributes during the training process. Significantly, the proposed method learns a supplementary attribute matrix by exploiting the attribute correlation. The proposed method also learns the relevant feature coefficients in the projection matrix to identify the correlated attribute space. Th proposed method also adopts l_1 regularization norm to select the relevant sparse features. A constrained optimization function is formulated and solved using the accelerated proximal gradient method. Extensive experiments on three benchmark datasets using ZSL and generalized ZSL demonstrate the effectiveness of the proposed method.**

***Index Terms*—Zero-shot learning, semantic attributes, semantic correlation**

I. Introduction

The traditional supervised machine learning models require hundreds of labeled samples about categories at training time to be able to classify them at testing time. In visual recognition, this approach is not desirable due to the limited number of labeled images and the cost of the manual labelling process. Inspired by the ability of humans to identify objects without seeing examples, machine learning researchers introduce the concept of zero-shot learning (ZSL) as a process to recognize objects that they have never seen before. For example, it is an easy task for a child to recognize a snake in a photo if he/she has seen reptiles before and knows that a snake is like a reptile but it is longer, legless, lacks movable eyelids, etc.

Zero-shot learning introduces a similar process to human recognition by identifying new instances using the detailed descriptions of the categories which are known as semantic attributes and can be represented using continuous or binary values (i.e., 0/1) [1]. In ZSL, at training time, the only information available to create the classification model is: the labeled images which are known as seen data and the semantic attributes of seen and unseen classes. Most of the existing works on ZSL [2]–[4] learn a projection function from the visual feature space of the seen training data to a semantic space. To recognize unseen objects, the projection function is used to map the unseen visual feature images into the same semantic space, then the models use a simple nearest neighbour algorithm to classify the new instance to the closest unseen category/class.

Of the various attribute-based ZSL approaches [5], [6], they explore a certain amount of correlation exists between the semantic attributes, thus, it is not desirable to learn each attribute independently. Few works have exploited the correlation across the semantic attributes to improve the performance of zero-shot classification [6], [7]. Specifically, in the work proposed in [6], the similarity space is connected to the attribute space to make the latent attributes discriminative.

However, several challenges still exist which require further attention. The semantic attributes vector is generated manually by humans either by extracting the descriptions from Wordnet or crafting the attribute values [8], however it may be difficult to obtain complete attribute information for each instance, especially in the case of a large number of classes. For example, the ImageNet dataset contains 21,814 classes with 14 million images. Moreover, it is common for an attribute vector to contain flawed values due to human error. In Fig. 1, the original semantic attributes vector A_s for zebras contains a missing value in the '*white*' attribute and some errors in the attributes '*black*' and '*stripes*'. The similarity between the semantic attributes can be used to augment the missing and flawed values in the original attribute matrix. Thus, the important attributes to recognize zebras such as '*black*', '*white*',and '*stripes*' have larger weights/values in the supplementary attribute vector $\hat{A}_s$. Further, as mentioned by Rezaei et al. [9], currently the world is facing the COVID-19 pandemic with limited annotated chest X-rays. Obviously, traditional supervised methods may not be effective due to sparse images. Therefore, if clinicians provide information about the attributes even if they are incomplete due to a lack of sufficient information regarding COVID-19, the augmented ZSL method based on pre-trained medical models which can handle missing attributes will be able to recognize the new concepts. As a result, the existing ZSL methods may not accurately estimate the relation between visual features and the semantic attributes due to missing and noisy attributes.

978-0-7381-3366-9/21/$31.00 ©2021 IEEE

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.

Moreover, in most of the attribute-based ZSL models, the learned feature coefficients in the projection matrix are related to individual attributes, however in the case of correlation, they must be related to the correlated attribute space.

This paper proposes a zero-shot learning approach (ZSL-MSA) which first explores the negative impact of missing attributes on zero-shot classification. Then, it handles the missing attribute problem by learning a supplementary semantic attribute matrix along with capturing the new attribute correlations within an optimization process while training the ZSL model. Furthermore, the proposed method adopts the l_1 regularization norm to impose sparsity on the learned features which is used to map the visual features to the semantic space. The proposed method is formulated as a constrained optimization function which solved using the accelerated proximal gradient method with convergence guarantees.

To the best of our knowledge, this is the first zero-shot learning method that is able to handle the missing and noisy semantic attribute problem. In summary the contributions of the proposed method are:

- It handles missing and noisy semantic attributes while learning the projection function in the ZSL model.
- It learns a new supplementary attribute matrix.
- It solves the proposed constrained optimization problem using the accelerated proximal gradient approach.
- It evaluates the compared methods by conducting experiment analysis using three public datasets.

II. Related Work

With the growth of newly emerging objects which are rarely seen and the difficulty of annotating images, zero-shot learning becoming increasingly important to recognize new instances that have never been seen before using the category descriptions. Attributes are the textual descriptions of the image categories. Most existing zero-shot learning methods learn the mathematical projection to map the labeled images to the semantic space at training time. At testing time, the unseen images are projected using the mapping function to predict the attributes. Then, the unseen class will be identified with the most similar attributes to the predicted attributes [2]–[4], [8], [9]. Another unseen-class recognition learning named as Open-Set Recognition (OSR) [10] which aims to build a classifier for seen-classes and rejects unseen classes.

Three types of ZSL frameworks are proposed to learn the projection function: (1) visual $\Rightarrow$ semantic, (2) semantic $\Rightarrow$ visual, and (3) visual $\Rightarrow$ label and semantic $\Rightarrow$ label. First framework is the most popular framework which aims to learn a projection function to map the visual features to semantic space [2], [3], [11]. Second framework aims to learn the mapping function from a semantic to a visual feature space which can alleviate from the hubness problem [12]. Third framework is a new ZSL approach to learn two mappings by activating the image labels: visual space to label space using one-hot encodings and from semantic space to label space [4].

Earlier works on ZSL learn the attributes independently [1]. However, few works explore the difficulty of accurately learning the attributes independently due to the fact that many attributes are correlated with one another [5], [6]. A random forest approach was proposed by Jayaraman et al. [5] to predict accurate predictions for unseen classes by considering the correlation between attributes. Several methods take into account the relationship between attributes and categories to improve the prediction results [13]. Yu et al. [14] proposes to exploit the relationship between seen and unseen classes. Jiang et al. [6] proposes to learn discriminative latent attribute space in a dictionary learning framework with attribute space and similarity space.

III. Problem Formulation

This section presents the notations and formulation of the proposed ZSL-MSA method to identify the discriminative latent features in the case of missing and unreliable semantic attributes in zero-shot learning.

A. Preliminaries

In zero-shot learning, there are given two sets of data for seen and unseen classes $S = \left\{X_s \in \mathbb{R}^{N_s \times d}, Y_s, A_s \in \mathbb{R}^{c \times k}\right\}$ and $U = \left\{X_u \in \mathbb{R}^{N_u \times d}, Y_u, A_u \in \mathbb{R}^{c \times k}\right\}$ where X_s and X_u are the visual d-dimensional data matrices which contain N_s and N_u samples respectively. The class label vectors Y_s and Y_u are associated with the instances of X_s and X_u matrices. The set of semantic encodings (A_s and A_u) describe the class labels of seen and unseen classes using semantic attributes, for example, a class label 'cat' has a tail and is furry, therefore tail and furry attributes assigned weights. At training time, the unseen labelled images are not available, therefore, there is no overlap between the class labels Y_s and Y_u ($Y_s \cap Y_u = \phi$). The goal of classical ZSL is to predict only unseen classes Y_u based on ZSL model which trained on labeled data X_s and attributes $(A_s \cup A_u)$, however GZSL is more challenging which aims to predict both seen and unseen classes.

B. Missing and unreliable semantic attributes

The recognition capability of the ZSL classifier depends on the semantic properties of the classes. ZSL classifier are generated on the fly to recognize the new classes based on the visual attributes. For example, the model is able to create a zebras without labelling any images of zebras in the training data using an attribute signature such as '*black and white, striped, furry hair, etc.*'. However, if any of the important attribute values in the zebra signature are noisy or missing, it will degrade the classification accuracy of the ZSL model. Therefore, the key challenge is the possibility of missing and noisy attributes in the semantic encodings A_s. The semantic attributes are either human-designed [15], [16] or discovered [14] where an error can occurr in designing these attributes which may generate some noisy or missing attributes values. To tackle this problem, the proposed method learns a new semantic attribute matrix $\hat{A}_s$ which exploits the semantic attribute correlation in the Euclidean space to complete the missing information. Mathematically, the new predicted matrix

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.

$\hat{A}_s$ is initialized using the prior information of the attribute correlation $Z \in \mathbb{R}^{k \times k}$ as follows:

$$\hat{A}_s = A_s \times Z \tag{1}$$

where the element $\hat{A}_s^{i,j}$ is initialized using all the semantic attributes and then updated during the optimization process. Fig. 1 shows the importance of exploiting the similarities between semantic attributes to reconcile the missing and noisy values in the original semantic attribute matrix. In this simple example, we assume that the zebra class contains a missing value for the *white* attribute and two small weights for *black and stripes* due to noise. Using the similarity matrix, the newly predicted semantic attribute matrix initialized the *black, white and stripes* elements with high values due to the strong correlation between these attributes in the similarity matrix. Further, it can seen that the zebra class has a better description for its features *black, white and stripes*. Therefore, the similarity matrix can capture the real relations across the semantic attributes because the semantic attributes are naturally correlated with each other and it is not desirable to learn the attributes independently [6], [17].

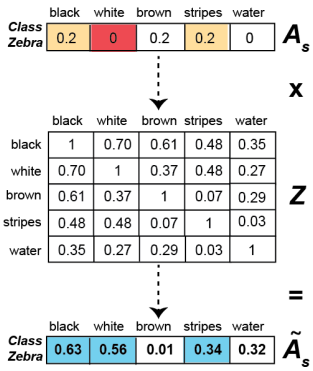


Fig. 1: A new prediction semantic attributes ($\hat{A}_s$) is identified using the similarity matrix of the attributes and the original semantic attributes. The red, yellow, and blue are the missing, noisy and predicted values respectively.

Previous studies [2], [3] project the visual features to the semantic attributes using a ridge classifier as follows:

$$\min_{W} \left\| X_s - A_s W^T \right\|_F^2 \tag{2}$$

where $W \in \mathbb{R}^{d \times k}$ is the projection matrix and $\|\|_F$ is the Frobenius norm. Based on the correlation assumption above, we propose the ZSL-MSA method by adopting the new semantic attribute matrix in the standard ZSL projection function. The newly predicted attribute matrix learns the accurate values of the missing attribute information in the original semantic matrix based on the similarity matrix Z as follows:

$$\min_{W} \left\| X_s W - \hat{A}_s \right\|_F^2 \quad \text{s.t. } \hat{A}_s = A_s Z \tag{3}$$

where the original attribute matrix A_s is replaced by the new attribute matrix $\hat{A}_s$. Further, to retain consistency between the new and original attribute matrices, we add a constraint $\hat{A}_s = A_s Z$ to the objective function.

C. Attribute Correlation

Previous studies ignored the correlation across semantic attributes during zero-shot learning [18]. The proposed method explicitly add the attribute correlation factor to the objective function to identify the discriminative features which are learned from the new semantic attribute matrix $\hat{A}_s$. This factor is significant during the learning process of the projection matrix W due to identifying the shared features space for the strongly correlated semantic attributes. The objective function is formulated as follows:

$$\min_{W} \left\| X_s W - \hat{A}_s \right\|_F^2 + \alpha D W^T W + \gamma \left\| W \right\|_1 \quad \text{s.t. } \hat{A}_s = A_s Z \tag{4}$$

where $D_{ij} = 1 - S_{ij}$, and D_{ij} and $S_{ij} \in \mathbb{R}^{k \times k}$ represents the distance and correlation between the predicted semantic attributes a_i and a_j respectively, $\alpha > 0$ and $\gamma > 0$ are the parameters. In the objective function (Eq. 4), if the distance between two semantic attributes in large, it will reduce the similarity between the coefficient vectors W_i and W_j. We also impose l_1 regularization norm on the project matrix W to induce sparsity on the attribute-features coefficients.

D. Optimization

The objective function in Eq. 4 is formulated as follows:

$$\begin{aligned} \min_{W,Z} F(W,Z) = &\left\| X_s W - \hat{A}_s \right\|_F^2 + \alpha D W^T W \\ &+ \beta \left\| \hat{A}_s - A_s Z \right\|_F^2 + \gamma \left\| W \right\|_1 \end{aligned} \tag{5}$$

where $\beta > 0$ and $\left\| \hat{A}_s - A_s Z \right\|^2$ factor to regularize the difference between A_s and $\hat{A}_s$. The objective function in Eq. 5 is a minimization problem of variables W and Z. This problem is partially convex where it minimizes one variable at a times and fix the other variable at the most recent values. Further, Eq. 5 is non-smooth due to adopting the l_1 sparsity regularization norm. Due to these reasons, accelerated proximal gradient (APG) method is used to solve the objective function. To solve Eq. 5, we update W and Z in each iteration as follows:

$$W^{t+1} = W^t - \frac{1}{L_{W^t}} \nabla_W F(W^t, Z^t) + g(W)$$

$$Z^{t+1} = Z^t - \frac{1}{L_{Z^t}} \nabla_Z F(W^{t+1}, Z^t)$$

where $L_{W^t} = \sqrt{XX^T + \alpha D}$ and $L_{Z^t} = \sqrt{A_s^T A_s}$ are Lipschitz constants of gradient $\nabla_W F(W, Z^t)$ and $\nabla_Z F(W^{t+1}, Z)$ by F as a function w.r.t W and Z respectively [19] and $g(W) = \gamma \left\| W \right\|_1$. To speed up the optimization process, Nesterov-type acceleration technique was used. It defines a positive weight $\Theta_t = \frac{r_{t-1}-1}{r_t}$, where $r_t = \frac{(1+\sqrt{1+4r_{t-1}^2})}{2}$. The idea of this technique is that if the current iteration is

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.

closer to the optimal point then Θ will be close to the current iteration [20]. The setting of problems W and Z is modified as follows:

$$P_W^t = W^t + \Theta_t(W^t - W^{t-1})$$

$$P_Z^t = Z^t + \Theta_t(Z^t - Z^{t-1})$$

Therefore, the coefficients of W and Z can be optimized by

$$W^{t+1} = P_W^t - \frac{1}{L_{W^t}} \nabla_W F(P_W^t, Z^t) + \gamma \|W\|_1 \quad (6)$$

$$Z^{t+1} = P_Z^t - \frac{1}{L_{Z^t}} \nabla_Z F(W^{t+1}, P_Z^t) \quad (7)$$

where $\nabla_W F(P_W^t, Z^t) = X_s X_s^T W - X_s^T \hat{A}_s + \alpha W D$ and $\nabla_Z F(W^{t+1}, P_Z^t) = -2\hat{A}_s^T A_s + A_s^T A_s Z$ by taking the partial derivative w.r.t W and Z where $\hat{A}_s = A_s Z$.

In every iteration, $g(W)$ is the l_1 norm regularization term in Eq.6 which is solved using the soft-thresholding operator S_γ as follows:

$$w_{ij}^{t+1} = S_\gamma[w_{ij}] = \begin{cases} w_{ij} - \gamma & \text{if } w_{ij} > \gamma \\ w_{ij} + \gamma & \text{if } w_{ij} < -\gamma \\ 0 & \text{otherwise} \end{cases}$$

where $\gamma > 0$ and w_{ij} is the element in the W matrix. The procedure of our method is described in Algorithm 1.

ALGORITHM 1: Zero-Shot Learning with Missing Semantic Attributes (ZSL-MSA)

Input: Training seen data $X_s \in \mathbb{R}^{N_s \times d}$
Semantic attributes $A \in \mathbb{R}^{c \times k}$
Parameters: α, β, and λ
Initialization:
$Z \in \mathbb{R}^{k \times k}$
$\hat{A}_s = A_s \times Z$
$Obj = \left\| X_s W - \hat{A}_s \right\|_F^2 + \alpha D W^T W + \beta \left\| \hat{A}_s - A_s Z \right\|_F^2 + \gamma \|W\|_1$
while $Obj < \varepsilon$ **do**
  Update W by solving the Eq. 6;
  Update Z by the solving the Eq. 7;
  Calculate Obj;
end
Output: Projection matrix W, Z

IV. EXPERIMENTS

The proposed ZSL-MSA method is validated on three datasets (AwA, CUB, and SUN) based on different levels of missing attributes that vary from moderate (20%) to extreme (80%). The results demonstrate the capability of the proposed method to handle missing and noisy attributes in zero-shot learning.

A. Datasets

The experiments are conducted on three challenging image datasets: Animal with Attributes (AwA) [21], CUB-200-2011 Birds (CUB) [22], and SUN Attributes (SUN) [23]. The details of the datasets are summarized in Table I. In the experiments, GoogleNet visual features [24] are extracted from the final pooling layer of the deep convolutional neural network.

TABLE I: Dataset characteristics

Dataset	#samples	seen/unseen classes	Semantic attributes
AwA	30475	40/10	85
CUB	11788	150/50	312
SUN	14340	645/72	102

B. Experimental settings

In the experiments, we randomly removed different proportions of semantic attributes which vary from moderate to extreme missing attributes: 10%, 20%,..., to 80%. We followed two settings for the missing attribute value that will replace the actual attribute value in the seen training data. First, to represent the missing attribute value, we normalize the semantic attributes between [0,1] and randomly replace the actual attribute value by -1. The other setting is used to represent the noisy data. After normalization, we generate the noisy value as a random number between [-1,1] to represent human or machine error. Our method comprises three parameters: α, β, and γ. The optimal values of the parameters are tuned using the grid-search strategy using five-fold cross-validation. To evaluate the performance of the methods in the standard ZSL setting, we use multi-way accuracy as in previous studies [2] to measure the classification accuracy of the unseen classes. For generalized zero-shot learning (GZSL), using five-fold cross-validation, we hold 20% of the seen training data and mix it with the unseen testing data. We report the average accuracy based on five-fold cross-validation for the seen testing classes (acc_s), unseen testing classes (acc_u) and the harmonic mean (H). The harmonic mean of seen and unseen classification accuracies [3] is defined as follows:

$$H = \frac{2 \times acc_s \times acc_u}{acc_s + acc_u}$$

The proposed method is compared with the most popular zero-shot learning methods SAE [2] and the recent ZSL method LAF [4]. To the best of our knowledge, the proposed method is the first in the literature to handle the missing semantic attributes.

C. Experiment Results

1) ZSL Results: Figs. 2 and 3 report the classification results of the algorithms with different proportions of missing and noisy attributes. The proposed ZSL-MSA method performs better than the previous methods. This indicates the importance of handling the missing and noisy semantic attributes to improve the classification performance of ZSL. Specifically, the classification accuracies of our proposed method in both

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.

TABLE II: Loss in classification accuracy for the three datasets.

Datasets	AwA			CUB			SUN		
	SAE	LAF	**ZSL-MSA**	SAE	LAF	**ZSL-MSA**	SAE	LAF	**ZSL-MSA**
Fig. 2	44.3%	42.5%	**19.4%**	61.8%	61.9%	**27.9%**	78.1%	72.3%	**44.0%**
Fig. 3	32.8%	30.0%	**16.6%**	72.7%	68.8%	**42.5%**	86.2%	68.9%	**57.5%**

Figs. 2 and 3 drop slightly to above 50% of the missing and noisy attributes. However, the classification performance of SAE and LAF quickly drops after 20% of missing attributes in most datasets. Furthermore, the proposed ZSL-MSA still achieves good classification results compared to the state-of-the-art in the case of extreme missing and noisy attributes. The reason for this is due to learning the discriminative features for the correlated attributes. Interestingly, we also report in Table II the loss in classification accuracy of the compared methods when the missing attribute proportions increase from 10% to 80%, as shown in Figs. 2 and 3. The results clearly demonstrate the capability of the proposed method to handle extreme proportions of the missing and noisy attributes. For example, in the AwA dataset, the loss in classification accuracy of ZSL-MSA for missing and noisy settings is only 19.4% and 16.6% respectively, however LAF suffers a loss of 42.5% and 30% which indicates that the LAF and SAE classification results are negatively affected by the missing and noisy attributes and the results become worse by increasing the proportions.

2) GZSL Results: The generalized zero-shot classification results are reported in Tables III, IV, and V for the three datasets using three evaluation metrics: acc_s, acc_u, and H. The accuracies in these tables are the average values of each compared algorithm using five-fold cross-validation. In each fold, we hold 20% of the training seen data and add it to the testing unseen data for evaluation. We specifically compared the performance of the methods based on their capability to predict the seen and unseen data at different proportions of the missing attributes. The proposed ZSL-MSA method achieves better results than the compared methods in all datasets as reported in Tables III, IV, and V. Similar to the results of standard ZSL, ZSL-MSA is able to retain a good classification performance compared to the other methods in the presence of extreme missing attribute proportions. For example, in Table III, for the extreme case of where 80% of the semantic attributes are missing, ZSL-MSA achieves 61%, 73%, and 66% in acc_s, acc_u, and H respectively, however LAF which is the second best method only achieves 44%, 55%, and 48% respectively. Importantly, ZSL-MSA achieves a remarkable improvement in predicting the seen testing data compared to SAE and LAF. Finally, it is clear that the classification performance decreases as the missing proportions increase, but we observe in the GZSL results that ZSL-MSA does not quickly drop its performance in the presence of large and extreme missing proportions as seen in SAE and LAF.

To perform an appropriate statistical comparative analysis, we conduct t-test on the experimental results. As reported in Tables III, IV, and V, our method's results are statistically different to the compared methods, the p-values are less than 0.05 in all comparisons which rejects the null hypothesis that all the methods have an equal performance. As a result, ZSL-MSA is superior across different datasets in ZSL and GZSL. The results indicate the importance of handling the missing semantic attributes in ZSL models to improve classification accuracy.

TABLE III: GZSL accuracy on AwA with different missing semantic attribute proportions. The value is the average of 5-fold cross-validation

Dataset	AwA								
Missing attribute	SAE			LAF			**ZSL-MSA**		
proportion ↓	acc_s	acc_u	H	acc_s	acc_u	H	acc_s	acc_u	H
20%	0.68	0.79	0.73	0.68	0.78	0.73	0.72	0.82	0.77
40%	0.58	0.76	0.66	0.58	0.75	0.66	0.68	0.81	0.74
60%	0.53	0.72	0.61	0.54	0.72	0.61	0.65	0.78	0.71
80%	0.43	0.55	0.48	0.44	0.55	0.48	0.61	0.73	0.66
t-test	*0.01*	*0.04*	*0.04*	*0.01*	*0.03*	*0.02*	base	base	base

TABLE IV: GZSL accuracy on CUB with different missing semantic attribute proportions. The value is the average of 5-fold cross-validation

Dataset	CUB								
Missing attribute	SAE			LAF			**ZSL-MSA**		
proportion ↓	acc_s	acc_u	H	acc_s	acc_u	H	acc_s	acc_u	H
20%	0.37	0.45	0.41	0.38	0.46	0.41	0.43	0.57	0.49
40%	0.29	0.36	0.32	0.30	0.37	0.33	0.41	0.45	0.43
60%	0.22	0.25	0.23	0.23	0.26	0.25	0.38	0.36	0.37
80%	0.15	0.14	0.14	0.16	0.15	0.15	0.32	0.29	0.30
t-test	*0.005*	*0.001*	*0.002*	*0.007*	*0.001*	*0.002*	base	base	base

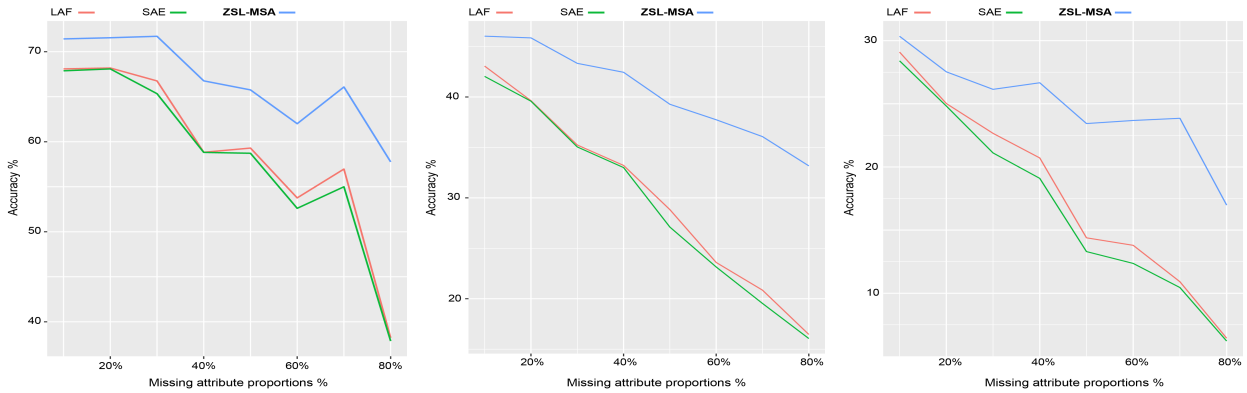
D. Parameter Sensitivity

This section examines the sensitivity of the parameters α, β, and γ in ZSL-MSA. We tune α and β using a grid-search from 0.1 to 1 with step-size 0.1 and $gamma$ is tuned from 50 to 300 with step-size 50. Due to space limitations, we only report the results for the AwA dataset using the H measure in GZSL. As shown in Fig. 4, we evaluate all the combinations of the grid search line for the three parameters, however to enable visualization, we fix one parameter value and visualize the other two parameters using H accuracy. We observe that generally the proposed method is not sensitive to the parameters in wide ranges. As shown in Figs. 4b and 4c, It is noted that when the value of the α parameter is increased, it achieves better accuracy, which indicates the importance of

TABLE V: GZSL accuracy on SUN with different missing semantic attribute proportions. The value is the average of 5-fold cross-validation

Dataset	SUN								
Missing attribute	SAE			LAF			**ZSL-MSA**		
proportion ↓	acc_s	acc_u	H	acc_s	acc_u	H	acc_s	acc_u	H
20%	0.23	0.13	0.17	0.24	0.14	0.18	0.28	0.26	0.27
40%	0.16	0.09	0.11	0.17	0.09	0.13	0.26	0.25	0.26
60%	0.11	0.05	0.07	0.12	0.06	0.08	0.24	0.15	0.18
80%	0.06	0.03	0.04	0.07	0.03	0.05	0.17	0.12	0.14
t-test	*0.005*	*0.002*	*<0.001*	*0.01*	*0.003*	*<0.001*	base	base	base

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.



(a) AwA (b) CUB (c) SUN

Fig. 2: Zero-Shot Learning (ZSL) accuracy results on three datasets with different missing semantic attribute proportions (10% to 80%).



(a) AwA (b) CUB (c) SUN

Fig. 3: Zero-Shot Learning (ZSL) accuracy results on three datasets with different noisy semantic attribute proportions (10% to 80%).

the contribution of the complete semantic attribute correlation to estimate the projection matrix W. Similarly, in Figs. 4a and 4b, the GZSL accuracy is improved by increasing the values of the γ parameter. This can be interpreted as a result of imposing sparsity over projection matrix W.

V. Conclusion

This paper proposes the first zero-shot learning approach to handle the missing and flawed attribute values in the semantic space while simultaneously training the ZSL model. To address the issue of missing or noisy semantic attributes, we incorporate a new supplementary attribute matrix by exploiting attribute correlation. Further, it estimates the relevant feature coefficients in the projection matrix to identify the specific correlated attribute space. Extensive experiments are conducted on three benchmark datasets which clearly demonstrate the superiority of our proposed algorithm over the compared algorithms in ZSL and GZSL.

References

[1] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1778–1785.

[2] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3174–3183.

[3] Y. Liu, Q. Gao, J. Li, J. Han, and L. Shao, "Zero shot learning via low-rank embedded semantic autoencoder." in *IJCAI*, 2018, pp. 2490–2496.

[4] Y. Liu, X. Gao, Q. Gao, J. Han, and L. Shao, "Label-activating framework for zero-shot learning," *Neural Networks*, vol. 121, pp. 1–9, 2020.

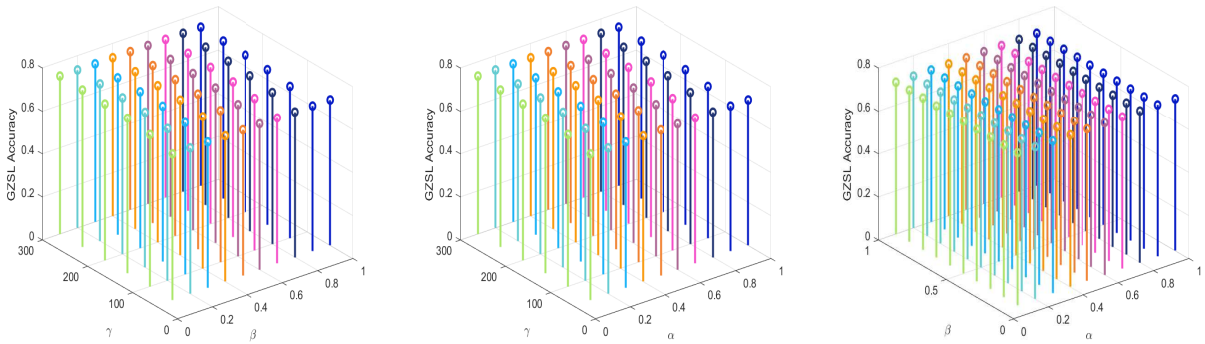
[5] D. Jayaraman and K. Grauman, "Zero shot recognition with unreliable attributes," *arXiv preprint arXiv:1409.4327*, 2014.

[6] H. Jiang, R. Wang, S. Shan, Y. Yang, and X. Chen, "Learning discriminative latent attributes for zero-shot classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4223–4232.

[7] M. Liu, D. Zhang, and S. Chen, "Attribute relation learning for zero-shot classification," *Neurocomputing*, vol. 139, pp. 34–46, 2014.

[8] M. Matsuki, P. Lago, and S. Inoue, "Characterizing word embeddings for zero-shot sensor-based human activity recognition," *Sensors*, vol. 19, no. 22, p. 5043, 2019.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.



(a) Optimize β and γ (b) Optimize α and γ (c) Optimize α and β

Fig. 4: GZSL accuracy results of ZSL-MSA on the AwA dataset w.r.t different parameter values. The missing semantic attribute proportions is 60%.

[9] M. Rezaei and M. Shahidi, "Zero-shot learning and its applications from autonomous vehicles to covid-19 diagnosis: A review," *Intelligence-based medicine*, p. 100005, 2020.

[10] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[11] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, and L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1993–2007, 2018.

[12] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto, "Ridge regression, hubness, and zero-shot learning," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2015, pp. 135–151.

[13] X. Wang and Q. Ji, "A unified probabilistic approach modeling relationships between attributes and objects," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2120–2127.

[14] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang, "Designing category-level attributes for discriminative visual recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 771–778.

[15] N. Turakhia and D. Parikh, "Attribute dominance: What pops out?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1225–1232.

[16] D. Parikh and K. Grauman, "Relative attributes," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 503–510.

[17] A. Braytee, W. Liu, A. Anaissi, and P. J. Kennedy, "Correlated multi-label classification with incomplete label space and class imbalance," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 5, pp. 1–26, 2019.

[18] C. Chai, Y. Lou, and S. Zhang, "Prior knowledge about attributes: Learning a more effective potential space for zero-shot recognition," *arXiv preprint arXiv:2009.06226*, 2020.

[19] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical programming*, vol. 103, no. 1, pp. 127–152, 2005.

[20] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, "Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix," *Coordinated Science Laboratory Report no. UILU-ENG-09-2214, DC-246*, 2009.

[21] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[22] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[23] G. Patterson, C. Xu, H. Su, and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," *International Journal of Computer Vision*, vol. 108, no. 1-2, pp. 59–81, 2014.

[24] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2927–2936.

Authorized licensed use limited to: DELHI TECHNICAL UNIV. Downloaded on October 01,2021 at 06:37:56 UTC from IEEE Xplore. Restrictions apply.