

SCHOLARLY PUBLICATIONS

*A CURRENT AWARENESS BULLETIN
OF RESEARCH OUTPUT*

@DTU

(120th Edition)

DECEMBER 2022

BY: CENTRAL LIBRARY

DELHI TECHNOLOGICAL UNIVERSITY

(FORMERLY *DELHI COLLEGE OF ENGINEERING*)

GOVT. OF N.C.T. OF DELHI

SHAHBAD DAULATPUR, MAIN BAWANA ROAD

DELHI 110042

PREFACE

This is the **One Hundred Twentieth** 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 **December, 2022**. The arrangement of the contents is alphabetical. The full text of the article which is either subscribed by the university or available in the web is provided in this bulletin.

Central Library

CONTENTS

1. A noise robust kernel fuzzy clustering based on picture fuzzy sets and KL divergence measure for MRI image segmentation, **8.Inder Khatri**, **3.Dhirendra Kumar** and **8.Aaryan Gupta**, Applied Mathematics, DTU
2. A Review on Fabrication of Universal Drilling Machine, **8.ROHAN KUMAR**, **8.SAATVIK SAGAR** and **3.AK MADAN**, Mechanical, DTU
3. A Switching NMOS Based Single Ended Sense Amplifier for High Density SRAM Applications, **6.Bhawna Rawat** and **3.Poornima Mittal**, Electronics, DTU
4. A2B corroles: fluorescent signalling system for Hg²⁺ ion, **6.ATUL VARSHNEY** and **3.ANIL KUMAR**, Applied Chemistry, DTU
5. An Embodied Conversational Agent to Minimize the Effects of Social Isolation During Hospitalization, Jemma Smith, **6.Aashish Bhandari**, Berkan Yuksel and A. Baki Kocaballi, CSE, DTU
6. Analyze the SATCON Algorithm's Capability to Predict Tropical Storm Intensity across the West Pacific Basin, **6.Monu Yadav** and **3.Laxminarayan Das**, Applied Mathematics, DTU
7. An Emotion-guided Approach to Domain Adaptive Fake News Detection using Adversarial Learning (Student Abstract), **8.Arkajyoti Chakraborty**, **8.Inder Khatri**, **8.Arjun Choudhry**, **8.Pankaj Gupta**, **3.Dinesh Kumar Vishwakarma** and Mukesh Prasad, IT, DTU
8. ANN prediction approach analysis for performance and emission of antioxidant-treated waste cooking oil biodiesel, **6.N. Kumar**, **6.K. Yadav** and **3.R. Chaudhary**, Mechanical, DTU
9. Assessment of water surface profile in nonprismatic compound channels using machine learning techniques, **6.Vijay Kaushik** and **3.Munendra Kumar**, Civil, DTU

10. Augmented thermoelectric performance of LiCaX (X = As, Sb) Half Heusler compounds via carrier concentration optimization, **6.Sangeeta** and **3.Mukhtiyar Singh**, Applied Physics, DTU
11. Capital Structure Study: A Systematic Review and Bibliometric Analysis, **6.Anjali Sisodia** and **3.G. C. Maheshwari**, DSM, DTU
12. Cervical Cancer Screening on Multi-class Imbalanced Cervigram Dataset using Transfer Learning, **6.Manisha Saini** and **3.Seba Susan**, CSE and IT, DTU
13. Characterization, utility, and interrelationship of household organic waste generation in academic campus for the production of biogas and compost: a case study, **6.Pradeep Kumar Meena**, **3.Amit Pal** and Samsher, Mechanical, DTU
14. Circuit Complexity in Z2 EEFT, Kiran Adhikari, Sayantan Choudhury, Sourabh Kumar, Saptarshi Mandal, **6.Nilesh Pandey**, Abhishek Roy, Soumya Sarkar, Partha Sarker and Saadat Salman Shariff, Applied Physics, DTU
15. CKS: A Community-based K-shell Decomposition Approach using Community Bridge Nodes for Influence Maximization (Student Abstract), **8.Inder Khatri**, **8.Aaryan Gupta**, **8.Arjun Choudhry**, **8.Aryan Tyagi**, **3.Dinesh Kumar Vishwakarma** and Mukesh Prasad, IT, DTU
16. Comparative Performance of DVR and STATCOM for Voltage Regulation in Radial Microgrid with High Penetration of RES, **6.Ritika Gour** and **3.Vishal Verma**, Electrical, DTU
17. Current Limiting Reactors based Time-Domain Fault Location for High Voltage DC Systems with Hybrid Transmission Corridors, **8.1.Vaibhav Nougain** and Sukumar Mishra, Electrical, DTU
18. Design and Implementation of a High-Performance 4-bit Vedic Multiplier Using a Novel 5-bit Adder in 90nm Technology, **6.Hemanshi Chugh** and **3.Sonal Singh**, Electronics, DTU

19. Development of Novel Model for the Assessment of Dust Accumulation on Solar PV Modules, **6.1.Astitva Kumar**, Muhannad Alaraj, **3.Mohammad Rizwan**, Ibrahim Alsaidan and Majid Jamil, Electrical, DTU
20. Development of water quality management strategies for an urban river reach: A case study of the river Yamuna, Delhi, India, **6.Nibedita Verma**, **3.Geeta Singh** and **6.Naved Ahsan**, Environmental, DTU
21. EMOTION-GUIDED CROSS-DOMAIN FAKE NEWS DETECTION USING ADVERSARIAL DOMAIN ADAPTATION, **8.Arjun Choudhry**, **8.Inder Khatri**, **8.Arkajyoti Chakraborty** and **3.Dinesh Kumar Vishwakarma**, IT, DTU
22. EV Control in G2V and V2G modes using SOGI Controller, **6.Sudhanshu Mittal**, **3.Alka Singh** and **3.Prakash Chittora**, Electrical, DTU
23. Experimental Simulation of Hydraulic Jump for the Study of Sequent Depth Using an Obstruction, **7.Daisy Singh**, **7.Abhishek Prakash Paswan**, **3.S. Anbukumar** and **6.Rahul Kumar Meena**, Civil, DTU
24. Exploiting Linguistic Information from Nepali Transcripts for Early Detection of Alzheimer's Disease using the State-of-the-art Techniques of Machine Learning and Natural Language Processing, **8.Surabhi Adhikari**, **8.Surendrabikram Thapa**, Usman Naseem, Priyanka Singh, Angela Huo, Gnana Bharathy & Mukesh Prasad, CSE, DTU
25. Failure analysis of a low-pressure stage steam turbine blade, **7.Pooja Rani** and **3.Atul K. Agrawal**, Mechanical, DTU
26. Hydrogenic impurity effect on the optical properties of Ga_{1-x}Al_xAs quantum wire under terahertz field, **6.Priyanka**, **3.Rinku Sharma**, Manoj Kumar and Pradumn Kumar, Applied Physics, DTU
27. Impact of sustainability reporting and performance on organization legitimacy, **3.Varsha Sehgal**, **3.Naval Garg** and Jagvinder Singh, DSM, DTU
28. Investigation of combustion and emission characteristics of an SI engine operated with compressed biomethane gas, and alcohols, **6.Pradeep Kumar Meena**, **3.Amit Pal** and **6.Samsher Gautam**, Mechanical, DTU

29. Knowledge-Infused Learning, **7.1.Manas Gaur**, CSE, DTU
30. Microplastics in the Ecosystem: An Overview on Detection, Removal, Toxicity Assessment, and Control Release, **6.Bhamini Pandey**, **7.Jigyasa Pathak**, **3.Poonam Singh**, Ravinder Kumar, Amit Kumar, Sandeep Kaushik and Tarun Kumar Thakur, Applied Chemistry, DTU
31. Multimedia information hiding method for AMBTC compressed images using LSB substitution technique, **3.Rajeev Kumar** and Aruna Malik, CSE, DTU
32. Nano-inspired smart medicines targeting brain cancer: diagnosis and treatment, **7.Raksha Anand**, **7.Lakhan Kumar**, **7.Lalit Mohan** and **3.Navneeta Bharadvaja**, Biotechnology, DTU
33. Natural polyphenols: a promising bioactive compounds for skin care and cosmetics, **3.Navneeta Bharadvaja**, **7.Shruti Gautam** and **7.Harshita Singh**, Biotechnology, DTU
34. Neural Underpinnings of Decoupled Ethical Behavior in Adolescents as an Interaction of Peer and Personal Values, Manvi Jain, **8.Karsheet Negi**, Pooja S. Sahni and Jyoti Kumar, Design, DTU
35. Offload 802.11 scanning to low power device, **6.Vishal Bhargava** and **3.N.S. Raghava**, CSE and Electronics, DTU
36. Open Source Software Based Electronic Health Record Management System, **8.Javteshwar Singh GILL**, **8.Himanshu MITTAL**, **8.Kunal BANSAL** and **3.Varsha SISAUDIA**, IT, DTU
37. Performance of adaptive radial basis functional neural network for inverter control, **3.Alka Singh** and **6.Amarendra Pandey**, Electrical, DTU
38. Plant integrated proportional integrating based control design for electric vehicle charger, **6.Aakash Kumar Seth** and **3.Mukhtiar Singh**, Electrical, DTU
39. Polarization Reversal of Oblique Electromagnetic Wave in Collisional Beam-Hydrogen Plasma, **8.Rajesh Gupta**, Ruby Gupta and **3.Suresh C. Sharma**, Applied Physics, DTU

40. Predictive linguistic cues for fake news: a societal artificial intelligence problem, Sandhya Aneja, **7.1.Nagender Aneja** and Ponnurangam Kumaraguru, CSE, DTU
41. Prospects of Nanostructure-based Electrochemical Sensors for Drug Detection: A Review, Manika Chaudhary, Ashwani Kumar, Arti Devi, Beer Pal Singh, **3.Bansi D. Malhotra**, Kushagr Singhal, Sangeeta Shukla, Srikanth Ponnada, Rakesh K Sharma, Carmen A Vega-Olivencia, Shrestha Tyagi and Rahul Singhal, Biotechnology, DTU
42. Radius of-Spirallikeness of order for some Special functions, Sercan Kazimoğlu and **6.Kamaljeet Gangania**, Applied Mathematics, DTU
43. Recent Advances in Various Types of Forging - A Research Review, **6.S. Wangchuk** and **3.Dr. AK Madan**, Mechanical, DTU
44. Retraction Note: Indian smart city ranking model using taxicab distance-based approach, **3.Kapil Sharma** and **6.Sandeep Tayal**, IT, DTU
45. Review on Chloride Ingress in Concrete: Chloride Diffusion and Predicting Corrosion Initiation Time, **3.Pradeep K. Goyal** and **6.Andualem E. Yadeta**, Civil, DTU
46. Role of bioactive compounds in the treatment of hepatitis: A review, Arpita Roy, Madhura Roy, Amel Gacem, **8.Shreeja Datta**, Md. Zeyauallah, Khursheed Muzammil, Thoraya A. Farghaly, Magda H. Abdellattif, Krishna Kumar Yadav and Jesus Simal-Gandara, Biotechnology, DTU
47. SAR and Optical Pixel Level Fusion Methods and Evaluations, **6.Sanjay Singh** and **3.K. C. Tiwari**, Electronics and Civil, DTU
48. Seismic Analysis of Sagging Elasto-flexible Cable using Placement Model, Pankaj Kumar, Sanjay Tiwari, S.K. Jain and **3.Ritu Raj**, Civil, DTU
49. Sensitivity Investigation of Junctionless Gate-all-around Silicon Nanowire Field-Effect Transistor-Based Hydrogen Gas Sensor, **3.Rishu Chaujar** and **6.Mekonnen Getnet Yirak**, Applied Physics, DTU
50. SHARP THIRD HANKEL DETERMINANT BOUND FOR $S(a)$, **6.NEHA VERMA** AND **3.S. SIVAPRASAD KUMAR**, Mathematics, DTU

51. SMOTE-LASSO-DeepNet Framework for Cancer Subtyping from Gene Expression Data, **7.Yashpal Singh** and **3.Seba Susan**, IT, DTU
52. Spatiotemporal Activity Mapping for Enhanced Multi-Object Detection with Reduced Resource Utilization, **6.Shashank** and **3.Indu Sreedevi**, Electronics, DTU
53. Study of third harmonic generation in InxGa1-xAs semi-parabolic 2-D quantum dot under the influence of Rashba spin-orbit interactions (SOI): Role of magnetic field, confining potential, temperature & hydrostatic pressure, **6.Suman Dahiya**, Siddhartha Lahon and **3.Rinku Sharma**, Applied Physics, DTU
54. Toeplitz determinants on bounded starlike circular domain in \mathbb{C}_n , **7.Surya Giri** and **3.S. Sivaprasad Kumar**, Applied Mathematics, DTU
55. TRANSFORMER-BASED NAMED ENTITY RECOGNITION FOR FRENCH USING ADVERSARIAL ADAPTATION TO SIMILAR DOMAIN CORPORA, **8.Arjun Choudhry**, **8.Pankaj Gupta**, **8.Inder Khatri**, **8.Aaryan Gupta**, Maxime Nicol, Marie-Jean Meurs and **3.Dinesh Kumar Vishwakarma**, IT, DTU
56. Twin core photonic crystal fiber based temperature sensor with improved sensitivity over a wide range of temperature, **6.Vishal Chaudhary** and **3.Sonal Singh**, Electronics, DTU

1. Vice Chancellor

1.1. Ex Vice chancellor

2. Pro Vice Chancellor

2.1. Ex Pro Vice Chancellor

3. Faculty

3.1. Ex Faculty

4. Teaching-cum-Research Fellow

4.1. Alumni

5. Asst. Librarian

5.1 Others

6. Research Scholar

6.1. Ex Research Scholar

7. PG Scholar

7.1. Ex PG Scholar

8. Undergraduate Student

8.1. Ex Undergraduate Student



A noise robust kernel fuzzy clustering based on picture fuzzy sets and KL divergence measure for MRI image segmentation

Inder Khatri¹ · Dharendra Kumar¹ · Aaryan Gupta¹

Accepted: 1 November 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

The image segmentation task becomes complex due to spatially distributed noise and vague boundary structure between the regions. For handling the vagueness present in images, fuzzy set theory-based clustering is popular for segmentation. Recently, the picture fuzzy set theoretic clustering methods have been investigated for image segmentation in literature. Few works based on picture fuzzy clustering have been reported to handle the vagueness and noise present in the image during segmentation. In literature, most research work uses smoothing for handling noise in the segmentation, which results in the loss of fine structure of images. Moreover, the non-linearity present in the data is also not addressed well, resulting in the loss of crucial features. In this research work, we have presented a picture fuzzy set-based clustering method termed as kernel fuzzy clustering based on picture fuzzy sets and KL divergence measure (KFPKL) to address the problem of noise, vagueness, and non-linear structure present in images. The picture fuzzy set can handle the vagueness present in data. We have included a KL divergence measure-based term in the proposed optimization problem to dampen the effect of noise in the segmentation process. To capture the non-linear structure present in images, the kernel distance measure is used in the proposed optimization problem. The experiments have been carried out on several synthetic image datasets and two publicly available brain MRI datasets. The comparison with the state-of-the-art methods shows that the proposed approach provides better segmentation performance in terms of average segmentation accuracy and Dice score.

Keywords Picture fuzzy sets · Picture fuzzy clustering · Kernel distance measures · Image segmentation · Magnetic resonance imaging · KL divergence

1 Introduction

Medical imaging devices are considered a handy tool for medical practitioners. These devices delineate the internal organs, which aids in the faster diagnosis of medical disorders. Some of the commonly used medical imaging modalities are X-ray scan, Computed Tomography (CT) scan, Medical Resonance Imaging (MRI), Endoscopy, Positron Emission Tomography (PET), and Ultrasound.

Among these, MRIs are usually considered the safest and most efficient as they provide high-contrast images with a multi-dimensional view. Unlike other modalities, they do not involve harmful ionizing radiations during image acquisition. Due to these advantages, MRI modality is preferred in investigating delicate body parts like soft brain tissues consisting of White Matter (WM), Gray Matter (Gray Matter), and Cerebrospinal Fluid (CSF).

With a large amount of MR data produced daily, faster and more accurate analysis of this data is desired, which requires skilled radiologists. However, the manual analysis of these images by radiologists is time-consuming, and further, the varying knowledge level of radiologists makes the analysis of these images inconsistent. Also, the MR images are often corrupted by additive noise, intensity non-uniformity (INU), and partial volume effect, making manual segmentation challenging. Hence, all these factors create a requirement for automated algorithms to support radiologists in providing quicker and more efficient analysis of these images. Several approaches for automated medical

✉ Dharendra Kumar
dhirendrakumar@dtu.ac.in

Inder Khatri
inderkhatri999@gmail.com

Aaryan Gupta
aryan227227@gmail.com

¹ Department of Applied Mathematics, Delhi Technological University, Delhi, India

image segmentation have been introduced in the literature, which can be broadly categorized into Region Growing [49], Threshold-based method [8], level set methods [2, 40, 61], Deep Neural Networks [9, 25, 46], Markov Random Fields, [24, 69, 73] Atlas Guided Approaches [47], Bayesian Approaches, graph cut [14], and Clustering techniques [4, 5, 15, 26, 30, 32, 35, 43–45, 63]. Among these, clustering techniques are found to be the most effective due to a good trade-off they offer between the time complexity and segmentation quality which is a major drawback of most of the other approaches.

Clustering is a well-known unsupervised learning technique used to group unlabeled data based on certain similarity measures. Some common clustering approaches are Centroid-based, Density-based, Distribution-based, and Hierarchical clustering [31]. Centroid-based clustering algorithms are the easiest to implement, and are found to be efficient for a variety of applications including data mining [6], image segmentation [11], document classification [34], data compression [42], bio-informatics [33], etc. Centroid-based clustering is further classified into Hard C-Means (HCM) and Fuzzy C-Means (FCM). In HCM, data points are given a membership value of either 0 or 1 to a given cluster resulting in the crisp membership matrix (1 if it belongs to a cluster, otherwise 0) [41]. However, in real-world applications, where a data point can fit in more than one cluster, HCM allocates datapoint to only one cluster. FCM, on the other hand, resolves this issue by using fuzzy set theory, where the membership matrix is soft. Thus, membership values could be any real value in between 0 and 1. The membership value of each data point corresponding to a cluster is inversely proportional to its distance to the prototype of that cluster [7].

In MRI images, voxels at boundaries may belong to more than one tissue; thus, a given voxel generally carries the information of different tissue classes. In the presence of imaging artifacts such as noise and intensity non-uniformity, the image quality deteriorates, inducing the non-linearity and uncertainty in MRI images. This makes the intensity distribution of MR images complex, against which most of the Fuzzy set theory based methods discussed in the literature are not robust. To overcome this, several clustering methods based on advanced fuzzy sets, such as intuitionistic fuzzy sets (IFS) and picture fuzzy sets (PFS), have been proposed in literature. These advanced fuzzy sets-based clustering methods carry the additional information that can model the non-linearity and uncertainty associated with MR images. However, existing methods based on advanced fuzzy sets have certain limitations, such as optimal parameter selection, over smoothing while handling the noise, and computational overhead. This leads to minimal preservation of shapes and structures and thus

make them less robust to imaging artifacts. Details of these segmentation methods are discussed in Section 2.

Recently, picture fuzzy set theory based clustering techniques have been investigated due to the better representational capability of PFS compared to the FS and IFS. Motivated by this, we formulate an optimization problem for clustering using a picture fuzzy set theoretic framework that solves the problem of noise, uncertainty, and non-linearity in the MRI image. In this work, we have proposed kernel fuzzy clustering using the picture fuzzy set theoretic approach via KL divergence measure, referred to as KFPKL. We utilize the Gaussian kernel distance measure to compute the distance measure in higher dimensional feature space. Due to the transformation of data points in higher dimensional feature space, the Gaussian kernel can solve the problem of inherent non-linearity in data without increasing the computational complexity. We also incorporate the KL divergence measure in the proposed optimization problem, which gives a better trade-off between robustness to noise and shape preservation.

In summary, the main contributions of this paper are as follows:

- The optimization problem for clustering is formulated using picture fuzzy set theoretic framework for handling the noise and non-linearity.
- Picture fuzzy set theory based clustering method is developed which increases the representational capability of clusters for better decision making.
- The proposed clustering approach aims at diminishing the effect of the noise in the segmentation process using Kullback-Leibler (KL) based divergence measure.
- To handle the non-linear structures, the kernel distance measures are utilized.
- The proposed clustering approach is robust to noise and non-linear structure present in image for segmentation task.

The rest part of this paper is organized as follows. Related works are discussed in Section 2. Section 3 describes the preliminaries. The proposed method is described in the Section 4. The experimental results on various synthetic and real data along with statistical analysis are presented in Section 5. Finally, Section 6 includes the conclusion.

2 Related works

A wide range of medical image segmentation methods has been discussed in the literature. These approaches can be broadly categorized into Region Growing method [49], Threshold-based method [8], level set method [2, 40, 61], Deep Neural Networks [9, 25, 46], Markov Random Fields,

[24, 69, 73] Atlas Guided Approaches [47], Bayesian Approaches, graph cut [14], and Clustering techniques [4, 5, 15, 26, 30, 32, 35, 43–45, 63].

Region growing is a region-based segmentation technique that integrates the neighboring pixels to form a segmented region based on the initial seed location and the similarity measure. In other words, region-growing segmentation is a process of integrating pixels into larger groups based on predefined seed pixels with some expanding criteria and exit conditions. Due to the iterative procedure over the pixels in the spatial domain, the region growing based segmentation methods often suffer from high time and space complexity. Another way of segmentation is threshold-based method which divides the images into several clusters, based on some threshold values calculated in advance using the pixel intensity histogram. Several threshold-based segmentation methods have been proposed in the literature [8]. A salient feature of threshold-based segmentation is that with the help of the threshold values, one can determine the optimal number of clusters, making the segmentation more precise. However, these methods do not consider the spatial information of pixels, making them vulnerable to additive noise. Also, their performance largely depends on the threshold values, making them prone to failure in case of any error in threshold calculation.

Level set-based segmentation is another vital approach for image segmentation, where the images are represented as contours given by the intersection of the surface with the plane. The segmented region in the image is obtained with the help of a numerical solution for processing topological changes of contours which updates the surface with forces derived from the image [57]. A major drawback of level set-based segmentation approaches is their high time complexity. Deep learning-based segmentation frameworks have gained much attention in recent years as they promise high accuracy. Various segmentation architectures like DeepLabv3, EfficientNet, TransUnet, MedT, Swin-Unet have shown state-of-the-art performance on several benchmark tasks. However, these methods rely on labeled training images and high-end computing resources. Further, training deep learning models takes plenty of time, which makes them expensive for image segmentation tasks.

In contrast to the aforementioned segmentation approaches, the clustering-based methods provide a good trade-off between the time complexity and segmentation performance. As discussed in the introduction (Section 1), fuzzy set theory based clustering methods are found to be the most effective, particularly for the segmentation of medical images. Since the performance of the FCM clustering for image segmentation task deteriorates in the presence of noise, therefore several improvements over the conventional FCM method have been proposed to suppress the

noise in segmentation process [1, 10, 13, 52]. Ahmad et al. [1] proposed an improved FCM algorithm termed FCM.S that captures the neighbors' pixel deviation from centroids to dampen the noise via a spatial regularization term. One major drawback of FCM.S is calculation of regularization term in each iteration, making it highly inefficient in terms of time-complexity. Chen et al. addressed this shortcoming of FCM.S with pre-computed mean and median image in the spatial regularization (termed as FCM.S1 and FCM.S2, respectively) [13]. In this way, FCM.S1 and FCM.S2 significantly reduced the computation time, however these were not much robust due to the involvement of spatial parameters and over-smoothing of images that loses fine image details. Further, Szilgayi et al. [52] presented the Enhanced FCM (EnFCM) that uses the histogram levels instead of pixel intensity values in the objective function. This avoids the repeated calculation for the same intensity pixel levels and thus, makes the segmentation very faster. Subsequently, Cai et al. [10], extended the EnFCM to a Fast and generalized FCM method (FGFCM). FGFCM incorporated neighborhood-dependent adaptive weights, which resulted in improved performance and better results than EnFCM. Similarly, Guo et al. [62] proposed a noise-detecting Fuzzy C means (NDFCM) to handle the problem of the in-homogeneous noise distribution in the MR images. NDFCM efficiently suppresses the noise and is able to preserve the sharp image details. It contains fewer input parameters and have a relatively low time complexity compared to the FGFCM. Another research work [74], proposed a kernel generalized fuzzy c-means clustering with spatial information termed KGFCM for image segmentation.

Most of the methods discussed above depend on specific input parameters, which are required to be fine-tuned for optimal performance. Also, these methods utilize smoothing to handle the noise during the segmentation process, which ultimately loses the image's fine details. To address these issues, Krindis et al. [35] proposed a fuzzy local information clustering method (FLICM), which uses the deviation of the neighboring pixel from the centroid's intensity, weighted by a fuzzy factor and spatial distance of neighbors. FLICM is parameter-free, and its segmentation performance on noisy images is also satisfactory. However, this method is highly complex since the local information is calculated in each iteration. Further, the loss function converges during step-wise optimization, but the objective loss does not minimize, making the optimization unstable. Motivated by FLICM, Gong et al. [22] proposed its variant termed as KWFLICM that replaced the Euclidean distance with kernel metric and introduced a trade-off weighted fuzzy factor to use the neighbor information adaptively. Although the KWFLICM performs better than the FLICM, it inherits the unstable optimization problem from FLICM. Zhao et al. [70]

proposed another variant, namely a Neighborhood Weighted FCM algorithm (NWFCM), which uses a neighborhood-weighted distance instead of Euclidean distance measure. NWFCM is more robust to noise and faster than FLICM and KWFLICM. Modified versions of the KWFLICM have been reported to improve the noise resistance. Wu et al. [67] proposed an extended KWFLICM algorithm, namely fuzzy local information c-means clustering algorithm utilizing total Bregman divergence driven by polynomial kernel function (TKWFLICM). Another work by Wu et al. [64] introduced a novel noise distance measure that combines entropy-like kernel divergence and normalized variance (NEKWFLICM), which improved the weighting term of the local information in KWFLICM by making it kernel fuzzy weighted.

Recently Lei et al. [39] proposed a fast and robust fuzzy c-means algorithm (FRFCM) utilizing the reconstruction operation as a pre-processing step to suppress the noise. To avoid the heavy-computational distance calculation between the neighbor pixels and centroids for neighborhood information constraints, FRFCM instead leveraged the membership filtering as a post-processing step. With no regularization term in the objective function, FRFCM significantly reduced the time complexity for segmentation in presence of various noises. FRFCM, however is brittle to the high-intensity noise and fails to preserve the sharp edges and shapes when encountered with high-noise samples. Zhang et al. [72] proposed DSFCM_N, which modeled the deviation between the original pixel values and measured noisy pixels value as residual. They considered the residual term as sparse and thus introduced it with l_1 norm constraint in the objective function. Like FRFCM, DSFCM_N also did not show reliable performance when tested with higher noise samples. Similarly, Wang et al. proposed Weighted Residual Fuzzy C-means (WRFCM), which used weighted l_2 -norm fidelity to make the residual estimation more reliable and obtained better results than the previous works [58]. With the same motivation, a few other recent works like SRFCM [60], and LRFCM [59] also utilized the residual modeling. SRFCM employed a three-step iterative optimization algorithm constructed by the Lagrangian multiplier method, hard-threshold operator, and normalization operator. It further used morphological reconstruction operation to suppress the noise and a tight wavelet approximated feature space to filter the images. Like SRFCM, LRFCM also uses morphological reconstruction and wavelet transformation with Residual noise modeling using l_0 norm constraint.

In many real-world applications, the measurement of data is imprecise. The fuzzy-based clustering method can not handle the impreciseness of data as the obtained fuzzy partition matrix is proportional to the distance between the inaccurate data and the cluster centroid. Hence,

this imprecision in data acquisition creates uncertainty in defining the membership value while the fuzzy clustering process. To deal with uncertainty in defining membership value, Atanassov proposed intuitionistic fuzzy set theory [3] which increases the representational capability of the real data. Integrating the intuitionistic fuzzy set in the clustering process exploits the non-membership degree and membership degree, leading to more precise, efficient handling of noise and quicker convergence in contrast to the traditional FCM [27]. In literature, many research works advocated IFS theory based clustering as it can handle the ambiguity while determining the membership values [12, 36–38, 56, 68, 71].

Xu et al. [68] suggested a fuzzy clustering of data represented in terms of IFS, which utilizes intuitionistic Euclidean distance measure [53]. Chaira [12] introduced the concept of IFS theory to address the problem of uncertainty in defining the membership value, also termed as hesitation degree in the conventional FCM algorithm. Also, it increases the significant data points in a given cluster through the entropy term defined for IFS. Dubey et al. [20] tried to solve the issues of pixel intensity variations using IFS representation of pixel intensity values. The performance of the IFS theory-based clustering method for the segmentation of the MRI images process deteriorates in the presence of noise. It is observed that the neighboring pixels of a given pixel in an image are similar, which leads to the idea of considering local spatial information to handle noisy pixels in the image segmentation problem.

Huang et al. [26] focused on the automatic optimal parameter selection and introduced neighborhood information-based IFCM algorithm with genetic algorithm (NIFCMGA). The method showed robust performance against noise and outlier in medical image segmentation but took considerable computation time as it utilizes genetic algorithm. Verma et al. [56] proposed a parameter-free method to handle noise and outlier in the medical data termed as improved IFCM (IIFCM). IIFCM used both local spatial and grey level information together for MRI segmentation. The method was found to be efficient on noisy data, but it is a considerably time-consuming method. Further, Kumar et al. [37] introduced a membership-dependent flexible neighborhood constraint to the IFCM method and proposed IFCM with spatial neighborhood information (IFCM-SNI). The spatial neighborhood information term preserved the fine image details and can deal with high-intensity various noises. Their model gives better results on noisy MRI images than the counter ones.

It is observed that when the data is complex, the IFS representation of real-world information corrupts. Thus, the IFS set-theoretic approach for clustering becomes inefficient to the need for uneven illumination and complex image segmentation problem [65]. To improve

the representation of data, Cuong et al. [18] extended the Intuitionistic Fuzzy Sets to Picture Fuzzy Set(PFS) by further detailing the degree of hesitancy to degree of refusal and neutrality. Motivated by the fact that PFS is a more generalized version of IFS theory, Son et al. [51] suggested a picture fuzzy c- means (PFCM) clustering method based on PFS theory. Thong et al. [54] suggested another variant of fuzzy clustering of PFS (FC-PFS) that treated the degree of neutrality and refusal identically. However, this assumption of FC-PFS does not appears reasonable, and the same reflects in its performance. Wu et al. [66] improved the robustness of the PFCM algorithm by incorporating the spatial context into the segmentation algorithm (referred to as IPFCM). Further, Wu et al. [65] modified the IPFCM method termed as AIPFCM method by introducing the adaptive weights in the optimization problem of the IPFCM method to enhance the segmentation performance. [65] also proposed a robust adaptive entropy weighted picture

fuzzy clustering with spatial information (APFCMS), which handles the noise in the image segmentation process. In most of the methods, the spatial neighborhood information term could not efficiently handle the problem of noise due to the involvement of image smoothing directly or indirectly in the segmentation problem. The smoothing operation during the segmentation process overlooks the delicate image structures and edges in an image. Also, the major issue with most of the existing image segmentation methods are non-linear structures present in the image and optimal spatial regularization parameter value selection. Table 1 summarizes the advantages and disadvantages of all the related methods.

The research problems discussed above motivate us to present an effort to solve the problems in related works with the help of proposed kernel fuzzy clustering using the picture fuzzy set theoretic approach via KL divergence measure to handle the noise and non-linearity present in the

Table 1 Summary of related methods

Method	Advantage	Disadvantage
FCM.S [13]	Proposed a Fuzzy Clustering framework with neighbor's information to tackle noise	Not able to preserve fine details, dependency on spatial parameter for good performance
FLICM [35]	Proposed a parameter-free algorithm for robust segmentation	Optimization was not found to be stable and was not converging
EnFCM [52]	Used histogram levels instead of spatial images for clustering, which significantly reduced time complexity	Dependent on spatial parameters and fine details not preserved
FGFCM [10]	Improved EnFCM by giving the adaptive weights to the neighbor pixel in the window on the basis of spatial and grayscale distance	Dependent on two input parameter and the performance gain was not significant
KWFLICM [22]	Improved FLICM by using Kernel Distance and Fuzzy weighted tradeoff	Optimization was not found to be stable and was not converging, and was time inefficient
NDFCM [23]	Proposed a parameter auto-tuning algorithm to tackle noise intensity inhomogeneity in the image	Sensitive to the input parameters, also dependent upon three parameter which need to be finetuned
IPFCM [65]	Extended FCM.S to the Picture Fuzzy Sets Framework with a symmetric picture fuzzy regularizing term	Lacks adequate robustness to noise and outliers and was dependent on spatial parameter alpha
KGFCM [74]	Added a membership Constraint for faster convergence and used kernel metric	Not robust to different type of noises
APFCMS [65]	Improved IPFCM by adding adaptive weights to Neighborhood information optimized with algorithm and making the membership value neighborhood weighted in the cost function	Fine details not preserved and over smoothing of images
IFCMSNI [37]	Proposed a novel Spatial Neighborhood Information term in IFS clustering	Not robust to the high strength of noise
IIFCM [56]	Used IFS Clustering with neighborhood information dependent on the spatial and grayscale distance of neighbor pixels	High time complexity due to the intensive internal operation
FRFCM [39]	Used Morphological Reconstruction as pre-processing and membership filtering as post-processing step	Not stable and difficult to obtain a tradeoff between noise robustness and over smoothing
DSFCMN [72]	Considered the deviation due to noise to be sparse as found in real-life scenarios and used l1 norm constraint	Performance drops when used for high-level noise images
WRFCM [58]	Improved DSFCMN by using l2 norm fidelity constraint and introducing weight term for residuals	Performance drops when used for high-level noise images

image refereed as KFPKL. The proposed KFPKL method is a PFS theory-based clustering method that incorporates the Gaussian kernel distance measure. It helps to cluster the data points robustly as the distance calculation is done in a higher dimensional space using the kernel trick. Due to the transformation of data points in higher dimensional feature space, the kernel trick can solve the problem of inherent non-linearity in data without increasing the computational complexity. Further, to check the efficacy of the proposed KFPKL method, experiments are performed on various real and synthetic datasets. The performance of the proposed method is compared quantitatively in terms of average segmentation accuracy (ASA) and Dice score (DS) with well-known existing methods such as FCM_S [13], FLICM [35], EnFCM [52], FGFCM [10], KWFLICM [22], NDFCM [23], IPFCM [65], KGFCM [74], APFCMS [65], IFCMSNI [37], IIFCM [56], FRFCM [39], DSFCMN [72] and WRFCM [58].

3 Preliminaries and notations

The description of notations used in this work and related definitions are presented here in this section.

Definition 1 Fuzzy set: A Fuzzy set is a set in which each member element will have the fractional membership via a membership function $\mu_A : X \rightarrow [0, 1]$ which gives the degree of belongingness [69]. If A is a fuzzy set defined over a set X , it can be represented as:

$$A = \{(x, \mu_A(x)) : x \in X\} \quad (1)$$

Definition 2 Picture fuzzy set (PFS): A Picture fuzzy set B , is an extension of fuzzy set over X which is represented as [18]:

$$B = \{(x, \mu_B(x), \eta_B(x), \gamma_B(x)) : x \in X \text{ and } 0 \leq \mu_B(x) + \eta_B(x) + \gamma_B(x) \leq 1\} \quad (2)$$

where $\mu_B : X \rightarrow [0, 1]$, $\eta_B : X \rightarrow [0, 1]$, $\gamma_B : X \rightarrow [0, 1]$ are positive membership value, neutral and negative membership functions of an element x in the set B .

The refusal degree $\xi_B(x)$ of an element x in the set B is represented as $\xi_B(x) = 1 - (\mu_B(x) + \eta_B(x) + \gamma_B(x))$. In case $\xi_B(x) = 0$, the PFS B reduces to IFS B and when $\eta_B(x) = 0$; $\xi_B(x) = 0$ then the PFS B is reduced to FS B for all x in B .

3.1 Kernel methods

Kernel-based machine learning techniques have proved to be efficient for several classification and feature extraction techniques such as Support Vector Machines (SVM) [17, 55],

Kernel Principal Component Analysis (KPCA) [50] and Kernel Fisher Discriminant (KFD) [48]. Generally, feature vectors associated with real data are not linearly separable, and further learning non-linear boundaries for classification is challenging. According to Cover's theorem [16], the transformation of data points from original low dimensional space to high dimensional space can be beneficial as data points are linearly separable in higher dimensional space. But, at the same time, the computational complexity for learning the classifier for high dimensional data is high as it involves a non-linear mapping ϕ from the original space to higher dimensional space. The problem of computational complexity can be address using a kernel function $K(x_i, x_j)$ which satisfies Mercer's condition [29] given as:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (3)$$

where x_i represents data in original space, $\phi(x_i)$ denotes the data in higher dimensional space. From (3), it is clear that the kernel function facilitates the ease of calculating the required metric without the transformation of data in higher dimension space. The use of kernels has a wide application in many areas as they map data onto a high-dimensional feature space to enhance the representation capability of linear machine learning model. The linear kernel function, polynomial kernel function of degree p , sigmoid kernel function, and the Gaussian radial basis function (RBF) are a few important kernel functions that are presented as

$$K(x_i, x_j) = x_j^T x_i$$

$$K(x_i, x_j) = \left(1 + x_j^T x_i\right)^p$$

$$K(x_i, x_j) = \tanh\left(\alpha\left(x_j^T x_i\right) + \beta\right)$$

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\sigma^2}\right)$$

respectively. Let elements \mathbf{x}_i and \mathbf{x}_j mapped to higher dimensional space through a non-linear transform ϕ then Euclidean distance in kernel space can be derived as follows:

$$\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 = \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^T \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|$$

$$\begin{aligned} \|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 &= \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_i) + \phi(\mathbf{x}_j)^T \phi(\mathbf{x}_j) \\ &\quad - 2\phi(\mathbf{x}_i)\phi(\mathbf{x}_j) \end{aligned}$$

and according to (3) the above equation can be written as:

$$\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 = K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j) \quad (4)$$

In case of Gaussian RBF kernel, this can be written as:

$$\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 = 2(1 - K(\mathbf{x}_i, \mathbf{x}_j)) \quad (5)$$

Similarly, the Euclidean distance for other kernel space can be calculated by replacing the corresponding kernel function in (4).

3.2 Picture fuzzy clustering

Son [51] proposed picture fuzzy clustering by modifying the conventional fuzzy c-mean to incorporate the advantage of the PFS theory. This increases the interpretability of the clustering and increases the representational capabilities. Let X represent the set of N data points to be clustered in c groups, then the optimization problem of picture fuzzy clustering can be given as [51]:

$$\min J(\mathbf{U}, \mathbf{V} : \mathbf{X}) = \sum_{i=1}^c \sum_{j=1}^N \left(\frac{\mu_{ij}}{1 - \eta_{ij} - \xi_{ij}} \right)^m \|\mathbf{x}_j - \mathbf{v}_i\|^2 \quad (6)$$

with the following constraints

$$0 \leq \mu_{ij}, \eta_{ij}, \xi_{ij} \leq 1$$

$$0 \leq \mu_{ij} + \eta_{ij} + \xi_{ij} \leq 1 \quad 1 \leq i \leq c \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \frac{\mu_{ij}}{1 - \eta_{ij} - \xi_{ij}} = 1, \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \left(\eta_{ij} + \frac{\xi_{ij}}{c} \right) = 1, \quad 1 \leq j \leq N$$

Where \mathbf{x}_j , \mathbf{v}_i , and m represent the j^{th} datapoint, i^{th} cluster centroid and degree of fuzziness, respectively. The membership degree, neutrality degree and refusal degree corresponding to j^{th} data point to i^{th} cluster are denoted by μ_{ij} , η_{ij} and ξ_{ij} , respectively. These notations have similar meanings throughout the manuscript. The optimization problem (6) can be solved using the Lagrange method of undetermined multipliers, and the solution can be given as:

$$\mu_{ij} = (1 - \eta_{ij} - \xi_{ij}) \frac{\|\mathbf{x}_j - \mathbf{v}_i\|^{\frac{-2}{m-1}}}{\sum_{k=1}^c \|\mathbf{x}_j - \mathbf{v}_k\|^{\frac{-2}{m-1}}} \quad (7)$$

$$\eta_{ij} = 1 - \xi_{ij} + \frac{\frac{c-1}{c \sum_{k=1}^c \xi_{kj}}}{\sum_{k=1}^c \frac{\mu_{ij}}{\mu_{kj}} \left(\frac{\|\mathbf{x}_j - \mathbf{v}_i\|}{\|\mathbf{x}_j - \mathbf{v}_k\|} \right)^{\frac{2}{m+1}}} \quad (8)$$

$$\mathbf{v}_i = \frac{\sum_{j=1}^N \left(\frac{\mu_{ij}}{1 - \eta_{ij} - \xi_{ij}} \right)^m \mathbf{x}_j}{\sum_{j=1}^N \left(\frac{\mu_{ij}}{1 - \eta_{ij} - \xi_{ij}} \right)^m} \quad (9)$$

$$\xi_{ij} = 1 - (\mu_{ij} + \eta_{ij}) - (1 - (\mu_{ij} + \eta_{ij})^\alpha)^{\frac{1}{\alpha}} \quad (10)$$

Using the above solution an alternating optimization algorithm can be designed and segmentation results can be obtained.

3.3 Fuzzy clustering on picture fuzzy sets (FC_PFS)

Thong et al. [54] proposed the variant of picture fuzzy clustering termed FC_PFS. In the FC_PFS method, the optimization problem consists of either the refusal degree

or the neutral degree. This condition results in improved performance of PFS as the other two degrees are also updated throughout the optimization. The actual membership value in the PFS framework is represented as $\mu_{ij}(2 - \xi_{ij})$. Let X represents the set of N data points to be clustered in c groups, then the optimization problem of the FC_PFS method can be given as [54]:

$$\min J(\mathbf{U}, \mathbf{V} : \mathbf{X}) = \sum_{i=1}^c \sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij}))^m \|\mathbf{x}_j - \mathbf{v}_i\|^2 + \sum_{i=1}^c \sum_{j=1}^N \eta_{ij} (\log \eta_{ij} + \xi_{ij}) \quad (11)$$

with the following constraints

$$0 \leq \mu_{ij}, \eta_{ij}, \xi_{ij} \leq 1$$

$$0 \leq \mu_{ij} + \eta_{ij} + \xi_{ij} \leq 1 \quad 1 \leq i \leq c \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \mu_{ij}(2 - \xi_{ij}) = 1, \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \left(\eta_{ij} + \frac{\xi_{ij}}{c} \right) = 1, \quad 1 \leq j \leq N$$

The optimization problem (11) can be solved using Lagrange method of undetermined multipliers and the solution can be given as:

$$\mu_{ij} = \frac{1}{(2 - \xi_{ij})} \frac{\|\mathbf{x}_j - \mathbf{v}_i\|^{\frac{-2}{m-1}}}{\sum_{k=1}^c \|\mathbf{x}_j - \mathbf{v}_k\|^{\frac{-2}{m-1}}} \quad (12)$$

$$\eta_{ij} = \frac{\exp(-\xi_{ij})}{\sum_{k=1}^c \exp(-\xi_{kj})} \left(1 - \frac{\sum_{k=1}^c \xi_{kj}}{c} \right) \quad (13)$$

$$\mathbf{v}_i = \frac{\sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij}))^m \mathbf{x}_j}{\sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij}))^m} \quad (14)$$

$$\xi_{ij} = 1 - (\mu_{ij} + \eta_{ij}) - (1 - (\mu_{ij} + \eta_{ij})^\alpha)^{\frac{1}{\alpha}} \quad (15)$$

Using the above solution an alternating optimization algorithm can be designed and segmentation results can be obtained.

4 Proposed kernel fuzzy clustering based on picture fuzzy sets and KL divergence measure method

As discussed earlier, the presence of non-linear structures and the noise in the image make the task of image segmentation challenging. Many fuzzy set-theoretic based methods in the literature address noise and non-linearity for image segmentation tasks. It has been observed that the segmentation performance mainly depends on the proper selection of regularization parameter values for noisy image segmentation problems. In this work, we have presented a picture fuzzy set theoretic clustering method for image

segmentation in the presence of the above mentioned problems. Motivated by the picture fuzzy set theory-based clustering method [54], we have formulated an optimization problem (16) termed as kernel fuzzy clustering for picture fuzzy set using the Kullback Liebler divergence measure (KFKPL). The optimization problem consists of three terms aiming to overcome the drawback of the existing method. The first term utilizes the kernel distance measure to handle the problem of non-Euclidean structures or non-linear structures. Further, the second term aims to diminish the noise effect in the segmentation process using Kullback-Leibler (KL) based divergence measure with a regularization parameter γ . This term incorporates the neighborhood information to correctly assign the pixel cluster label and introduce the fuzziness in the model. The proposed optimization takes advantage of PFS over FS and IFS. To preserve the consistency of the PFS based optimization problem with the FS and IFS, the true membership is denoted by $\mu_{ij}(2 - \xi_{ij})$. The third term $\sum_{i=1}^c \sum_{j=1}^N (\eta_{ij}(\log \eta_{ij} + \xi_{ij}))$ tries to reduce the entropy information corresponding to PFS set for obtaining good cluster points. In this way, the optimization problem helps the algorithm reduce the neutrality degree η_{ij} and refusal degree ξ_{ij} of an element to become a cluster member. Further, the constraint $\sum_{i=1}^c \left(\eta_{ij} + \frac{\xi_{ij}}{c}\right) = 1$ ensures at least one out of the neutrality degree η_{ij} and refusal degree ξ_{ij} always exist in the model and the constraint $\sum_{i=1}^c \mu_{ij}(2 - \xi_{ij}) = 1$ indicates the sum of true membership values to different cluster is always 1. The optimization problem of the proposed KFKPL can be given as follows:

$$\begin{aligned} \min J_{\gamma}^{\phi}(\mathbf{U}, \mathbf{V} : \mathbf{X}) = & \sum_{i=1}^c \sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij})) \|\phi(\mathbf{x}_j) - \phi(\mathbf{v}_i)\|^2 \\ & + \gamma \sum_{i=1}^c \sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij})) \log \frac{(\mu_{ij}(2 - \xi_{ij}))}{\pi_{ij}} \\ & + \sum_{i=1}^c \sum_{j=1}^N (\eta_{ij}(\log \eta_{ij} + \xi_{ij})) \end{aligned} \quad (16)$$

Where $\pi_{ij} = \frac{1}{|N_j|} \sum_{k \in N_j} (\mu_{ik}(2 - \xi_{ik}))$

Subject to the constraints

$$0 \leq \mu_{ij}, \eta_{ij}, \xi_{ij} \leq 1$$

$$0 \leq \mu_{ij} + \eta_{ij} + \xi_{ij} \leq 1 \quad 1 \leq i \leq c \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \mu_{ij}(2 - \xi_{ij}) = 1, \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \left(\eta_{ij} + \frac{\xi_{ij}}{c}\right) = 1, \quad 1 \leq j \leq N$$

The mapping utilized for the transformation of samples to higher dimensional space is represented by ϕ . The optimization problem (16) for the Gaussian radial basis

function (RBF) kernel function using (5) can be rewritten as:

$$\begin{aligned} \min J_{\gamma}^{\phi}(\mathbf{U}, \mathbf{V} : \mathbf{X}) = & \sum_{i=1}^c \sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij})) (1 - K(\mathbf{x}_j, \mathbf{v}_i)) \\ & + \gamma \sum_{i=1}^c \sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij})) \log \frac{(\mu_{ij}(2 - \xi_{ij}))}{\pi_{ij}} \\ & + \sum_{i=1}^c \sum_{j=1}^N (\eta_{ij}(\log \eta_{ij} + \xi_{ij})) \end{aligned} \quad (17)$$

Where $\pi_{ij} = \frac{1}{|N_j|} \sum_{k \in N_j} (\mu_{ik}(2 - \xi_{ik}))$

Subject to the constraints

$$0 \leq \mu_{ij}, \eta_{ij}, \xi_{ij} \leq 1$$

$$0 \leq \mu_{ij} + \eta_{ij} + \xi_{ij} \leq 1 \quad 1 \leq i \leq c \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \mu_{ij}(2 - \xi_{ij}) = 1, \quad 1 \leq j \leq N$$

$$\sum_{i=1}^c \left(\eta_{ij} + \frac{\xi_{ij}}{c}\right) = 1, \quad 1 \leq j \leq N$$

Where the kernel distance measure is denoted by $K(\mathbf{x}_j, \mathbf{v}_i) = \exp\left(\frac{-\|\mathbf{x}_j - \mathbf{v}_i\|^2}{\sigma^2}\right)$ and the regularization parameter is denoted by γ . $\mathbf{v}_i \quad \forall i = 1, 2, \dots, c$ and $\mathbf{x}_j \quad \forall j = 1, 2, \dots, N$ represent the centroids of clusters and data points respectively. After solving the optimization problem (17), the membership value μ_{ij} , cluster center \mathbf{v}_i , neutrality degree η_{ij} and refusal degree ξ_{ij} using Lagrange's method of the undetermined multiplier can respectively be obtained as:

$$\mu_{ij} = \frac{\pi_{ij} \exp\left(-\frac{(1 - K(\mathbf{x}_j, \mathbf{v}_i))}{\gamma}\right)}{\sum_{l=1}^c \pi_{lj} \exp\left(-\frac{(1 - K(\mathbf{x}_j, \mathbf{v}_l))}{\gamma}\right)} \frac{1}{(2 - \xi_{ij})} \quad (18)$$

$$\mathbf{v}_i = \frac{\sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij})) K(\mathbf{x}_j, \mathbf{v}_i) \mathbf{x}_j}{\sum_{j=1}^N (\mu_{ij}(2 - \xi_{ij})) K(\mathbf{x}_j, \mathbf{v}_i)} \quad (19)$$

$$\eta_{ij} = \frac{\exp(-\xi_{ij})}{\sum_{l=1}^c \exp(-\xi_{lj})} \left(1 - \sum_{l=1}^c \frac{\xi_{lj}}{c}\right) \quad (20)$$

$$\xi_{ij} = 1 - (\mu_{ij} + \eta_{ij}) - (1 - (\mu_{ij} + \eta_{ij})^{\alpha})^{\frac{1}{\alpha}} \quad (21)$$

Outline of the iterative procedure for finding the solution of the proposed method is given below.

Figure 1 shows a demonstrative example of image segmentation using the proposed approach. The left part of the Fig. 1 shows the clean image, whereas the right part depicts the same image with noise. Image contains four regions i.e., 0, 1, 2 and 3. For demonstration, three windows of size 5×5 from both clean and noisy image are selected with intensities values shown at the bottom of image. It can be observed that the final assignment made by the proposed

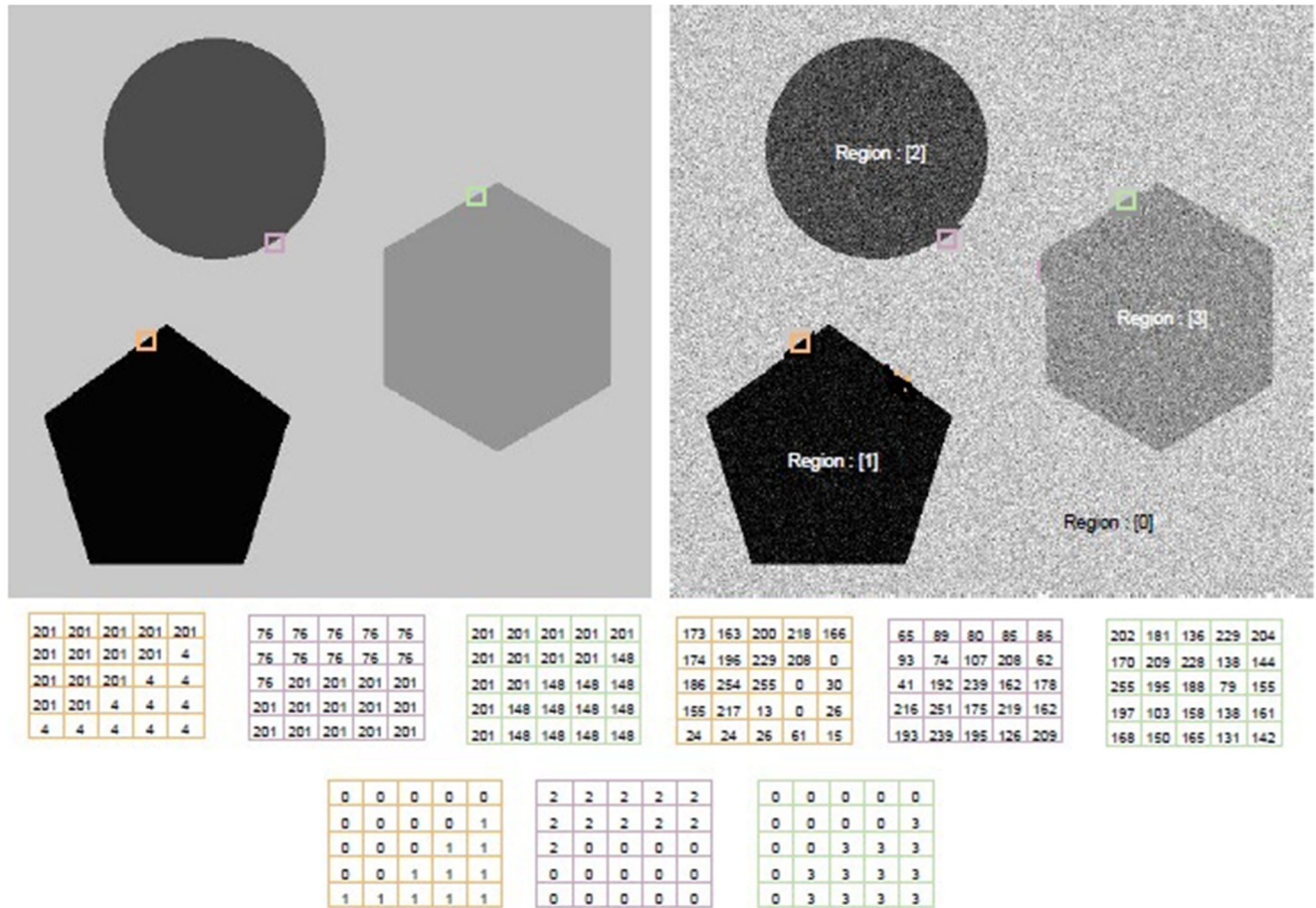


Fig. 1 Demonstrative example

Require: Set values for the number of clusters c , the value of regularization parameter γ , the value of fuzzy negation parameter α and the error ϵ .

- 1: Initialize the membership $v_i^0 = \{v_i^{(0)}\}_{c \times 1}$ using the conventional FCM
- 2: $k \leftarrow 1$
- 3: **repeat**
- 4: Update the membership matrix $U^{(k)} = \{\mu_{ij}^{(k)}\}_{c \times N}$, $1 \leq i \leq c$ and $1 \leq j \leq N$ using (18)
- 5: Update the η_{ij} , $1 \leq i \leq c$ and $1 \leq j \leq N$ using (20)
- 6: Update the ξ_{ij} , $1 \leq i \leq c$ and $1 \leq j \leq N$ using (21)
- 7: Compute the centers $v_i^{(k)}$, $1 \leq i \leq c$ using (19)
- 8: $k \leftarrow k + 1$
- 9: **until** $\|U^{(k+1)} - U^{(k)}\| + \|\eta^{(k+1)} - \eta^{(k)}\| + \|\xi^{(k+1)} - \xi^{(k)}\| < \epsilon$
- 10: **return** the centers of clusters v_i , the membership degrees μ_{ij} , neutrality degree η_{ij} and refusal degree ξ_{ij} .

Algorithm 1 Proposed KFPKL algorithm.

approach to different regions for each of the noise windows (depicted with three different colors) is consistent with the ground truth (see the bottom part of Fig. 1).

4.1 Sensitivity analysis on parameter

The proposed KFPKL method requires Yager's negation parameter α , regularization parameter γ , and kernel metric parameter σ as input. The method requires parameter α 's value to be in the range $[0, 1]$. Figure 2 shows variation in ASA for five BrainWeb MR images with noise levels 0, 3, 5, 7, and 9 on varying the value of regularization parameter γ . It can be noted from Fig. 2 that the performance of the proposed KFPKL method is better when the regularization parameter γ values are in the range $[0, 0.5]$. On this reduced range of search space, grid search is used to find the optimal values with step size 0.1.

The selection of the optimal value of parameter σ corresponding to the Gaussian RBF kernel function is also crucial to the performance of the proposed KFPKL method. Similar to the sensitivity analysis of γ , we perform

Fig. 2 Variation in performance in terms of ASA on several MRI images with noise on different values of γ

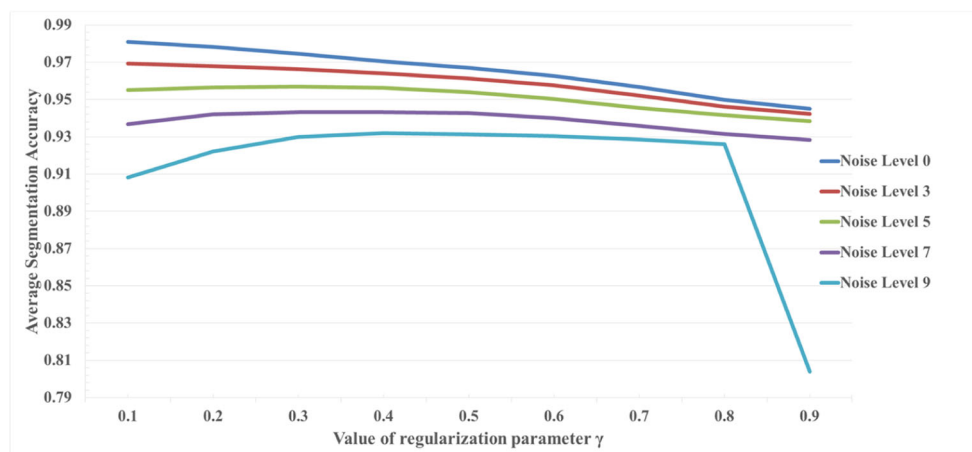


Fig. 3 Variation in performance in terms of ASA on several MRI images with noise on different values of α

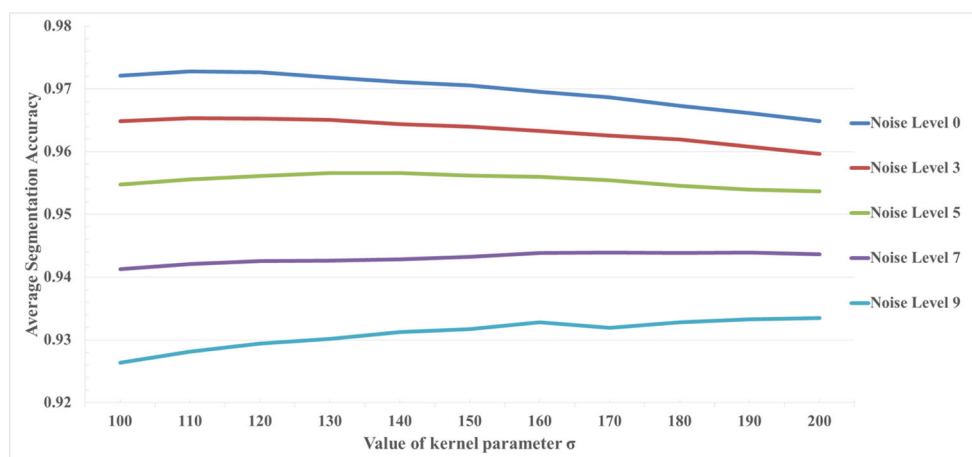
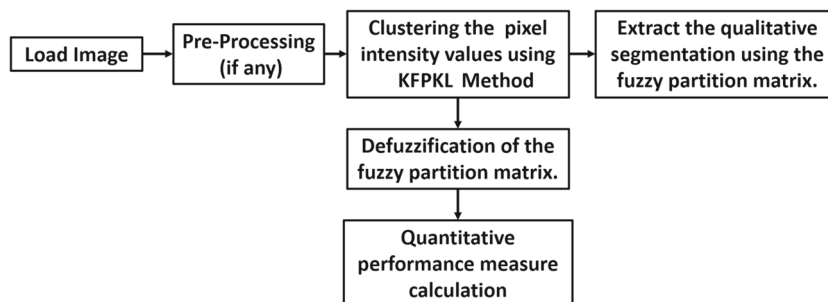


Fig. 4 Flow diagram for evaluating the clustering approach



sensitivity analysis for parameter σ by running the KFPKL algorithm for different values of σ in the range [100–200] with step size 10. Figure 3 shows the ASA accuracy for different values of sigma on the five BrainWeb images. From Fig. 3, it could be observed that our proposed KFPKL method is not sensitive to the kernel parameter σ , and thus, we use the value of sigma to be 150 in all our experiments.

The salient features of the proposed KFPKL method are as follows:

- The proposed method is utilizing the picture fuzzy set theory that handle the higher level of uncertainty that can not be handled with FS and IFS.
- The proposed method is efficient in performing the image segmentation in presence of noise and non-linearity.
- KL divergence based similarity measure is used to dampen the noise.
- The proposed method is less sensitive to the values of parameters involve for image segmentation process.
- The proposed method utilized the kernel distance measure which can handle the non-linear structures present in data.

5 Experimental setup and results

In order to validate the efficacy of the proposed KFPKL method, we have performed experiments on one synthetic dataset and two openly accessible datasets. Aim of the experiment is to show the effectiveness of the proposed KFPKL clustering approach for image segmentation task for different kind of images including natural and medical images. Experiments are designed to check the efficacy of the proposed method in comparison to other related methods such as FCMS, FGFCM, EnFCM, KWFLICM, FLICM, NDFCM, KGFCM, IPFCM, APFCMS, IIFCM, IFCMSNI, FRFCM, DSFCMN and WRFCM. For the comparison purpose, both quantitative and qualitative results are obtained on several images. The quantitative performance metrics such as average segmentation accuracy (ASA) and Dice score (DS) [13, 40, 52] can be calculated as:

$$ASA = \frac{\sum_{i=1}^c |X_i \cap Y_i|}{\sum_{j=1}^c |X_j|} \quad (22)$$

$$DS = \frac{2|X_i \cap Y_i|}{|X_i| + |Y_i|} \quad (23)$$

Where the number of clusters is represented by c , X_i denotes the pixels corresponding to the i^{th} class in the manually segmented ground truth image. Y_i represents the pixels assigned to the i^{th} class by the segmentation algorithm. The cardinality of X_i is represented by $|X_i|$. ASA depicts the overall segmentation performance of

a given method and lies in the range [0, 1]. On the other hand, DS indicates the region-wise segmentation performance, which lies in the range [0, 1]. The higher value of both ASA and DS corresponding to the given method shows the superior performance of that method. Also, computation time analysis is performed to show the superiority of the proposed KFPKL method over other related methods. Further, statistical test is conducted to

Table 2 List of parameter values associated with each methods

Method	Input parameters	Optimal parameter selection
FCM_S [13]	p, α	$p : 2$ $\alpha : 0.2 - 3(\text{step size } 0.2)$
FLICM [35]	m	$m : 2$
EnFCM [52]	p, α	$p : 2$ $\alpha : 0.2 - 3(\text{step size } 0.2)$
FGFCM [10]	m, λ_s, λ_g	$m : 2$ $\lambda_s : 0.2 - 5(\text{step size } 0.2)$ $\lambda_g : 0.2 - 5(\text{step size } 0.2)$
KWFLICM [22]	m	$m : 2$
NDFCM [23]	$m, \lambda_a, \lambda_s, \lambda_g$	$m : 2$ $\lambda_s : 0.2 - 5(\text{step size } 0.2)$ $\lambda_g : 0.2 - 5(\text{step size } 0.2)$ $\lambda_a : 0.2 - 10(\text{step size } 0.5)$
IPFCM [65]	m, α, α_1	$m : 2$ $\alpha_1 : 0.2 - 3(\text{step size } 0.2)$ $\alpha : 0.5 - 0.9(\text{step size } 0.1)$
KGFCM [74]	m, α, β	$m : 2$ $\alpha : 0.99$ $\beta : 2 - 5(\text{step size } 0.2)$
APFCMS [65]	$m, \lambda, \alpha_1, p, q$	$m : 2$ $\lambda : 10^6$ $\alpha_1 : 0.5 - 0.9(\text{step size } 0.1)$ $p : 3$ $q : 6$
IFCMSNI [37]	m, α, β	$m : 2$ $\alpha : 0.2 - 6(\text{step size } 0.2)$ $\beta : 0.5 - 0.9(\text{step size } 0.1)$
IIFCM [56]	m, λ	$m : 2$ $\lambda : 0.5 - 0.9(\text{step size } 0.1)$
FRFCM [39]	m	$m : 2$
DSFCMN [72]	m, λ	$m : 2$ $\lambda : (0.2 - 1) * \sigma(\text{step size } 0.2)$
WRFCM [58]	m, β, η	$m : 2$ $\beta : (0.01 - 0.1) * \sigma(\text{step size } 0.01)$ $\eta : 0.0001$
KFPKL	m, γ, α	$m : 2$ $\gamma : 0.1 - 1(\text{step size } 0.1)$ $\alpha : 0.5 - 0.9(\text{step size } 0.1)$

Fig. 5 Qualitative results obtained on synthetic gray image corrupted with Gaussian noise

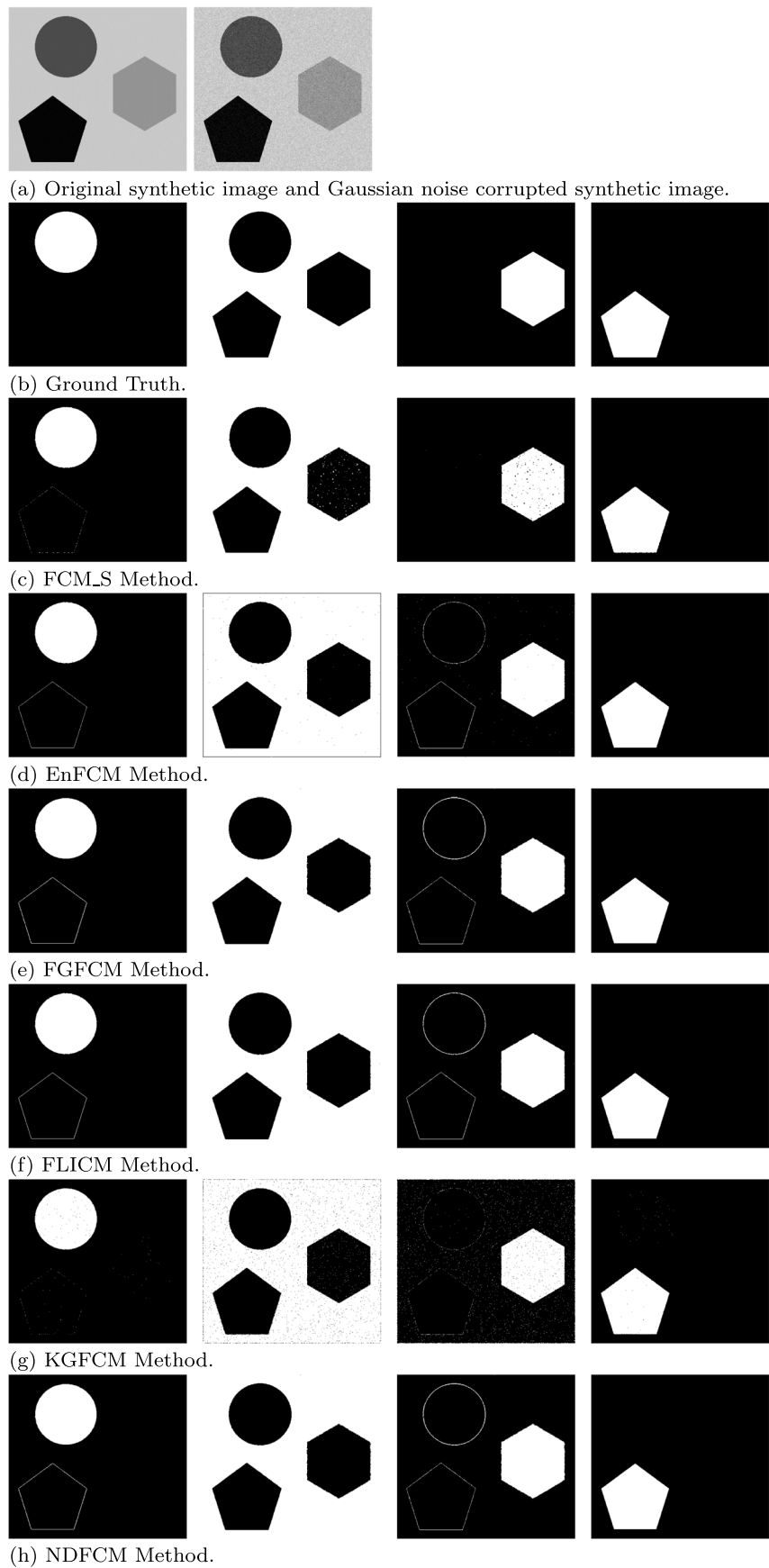


Fig. 5 (continued)

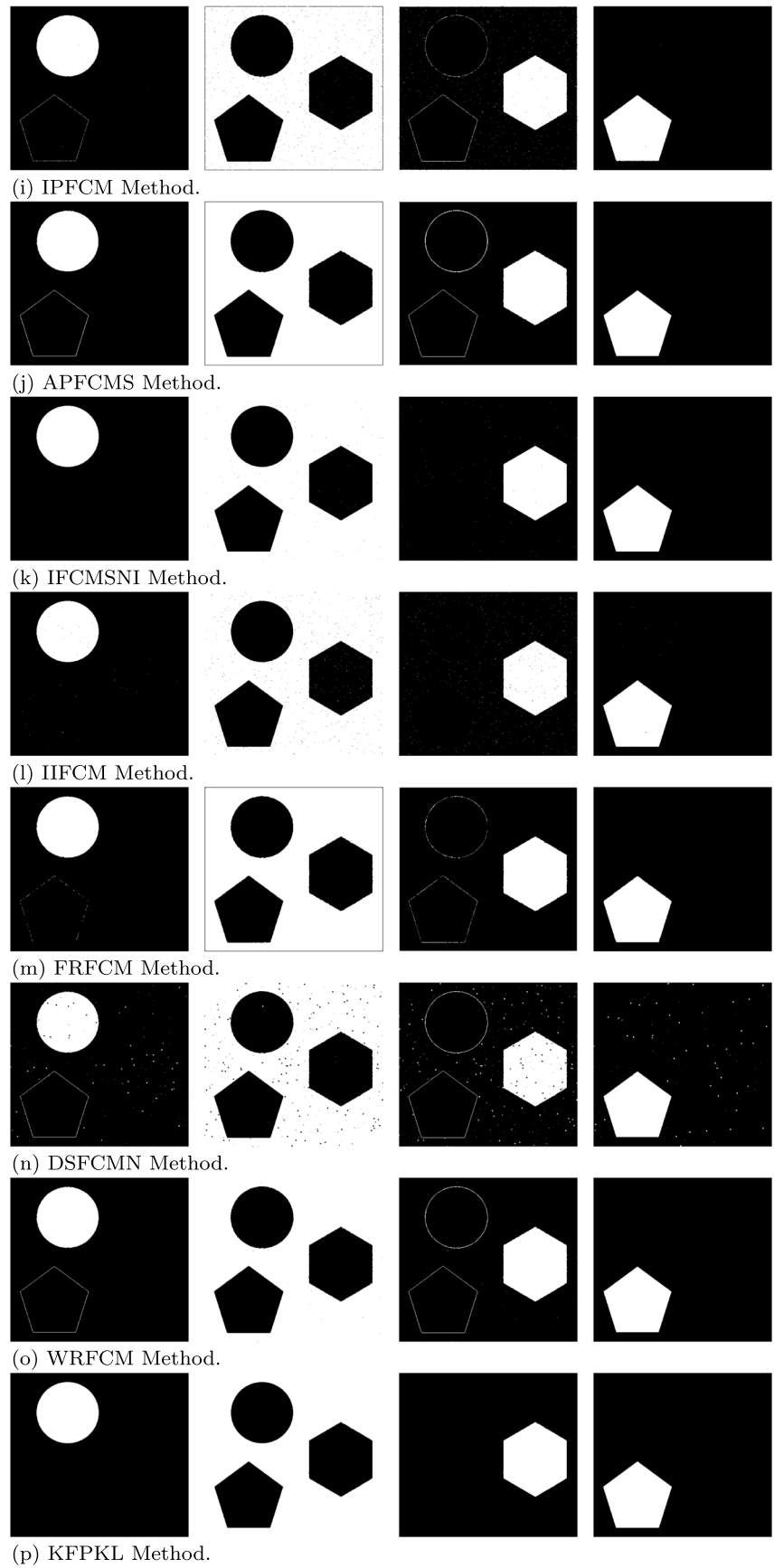


Table 3 Details of datasets utilized in the study

Dataset	Source	Size	Instances	Noise
Brain Web	http://www.brainweb.bic.mni.mcgill.ca/brainweb	$181 \times 217 \times 181$	9	RN (5%, 7%, 9%)
IBSR	https://www.nitrc.org/projects/ibsr	$256 \times 256 \times 63$	5	RN (5%)
Synthetic Image	Synthesized	420×456	6	GN (5%, 10%, 15%) RN (5%, 10% 15%)
Natural Images (Coin corrupted with noise)	Publicly available	300×246	2	SP (5%) GN (5%)

validate the effectiveness of the proposed approach for image segmentation task.

Figure 4 shows the flow diagram for evaluating the qualitative and quantitative segmentation using the clustering approach. First of all, we acquire the image, followed by pre-processing, which for MR images consists of skull stripping and tissue extraction, if required. After this, the pre-processed image is passed to the clustering algorithm which produces fuzzy partition matrix corresponding to the image. Further, we get the crisp labels by defuzzifying the fuzzy partition matrix which are used to obtain the quantitative results. To get the qualitative results, we can obtain the segmented image for different tissues regions using the fuzzy partition matrix.

Table 2 shows the list of the associated parameter corresponding to related methods along with the proposed method. We obtain the optimized parameter value for different techniques using the grid search and use the average segmentation accuracy as the metric for final parameter selection. We also report each parameter's step

size and range in Table 2. Please note that the symbols mentioned in Table 2 are consistent according to the details given in the corresponding research work.

5.1 Datasets

The performance of the proposed method is being validated and compared with other related methods using a synthetic gray image and two publicly available MRI datasets. The synthetic gray image has four regions with intensities 4, 76, 148 and 201 (See Fig. 5(a)). For experiment purpose, this gray image is adulterated with Gaussian noise and Rician noise.

BrainWeb dataset consisting of simulated MRI brain volumes is utilized. This dataset is obtained from the Montreal Neurological Institute of McGill University's, McConnell Brain Imaging Center¹ [15]. We have used T1-weighted brain volume data which are simulated with different intensity non-uniformity (0%, 20% and 40%) and noise (1%, 3%, 5%, 7% and 9%) of resolution $1 \times 1 \times 1\text{mm}$ with $181 \times 217 \times 181$ dimension.

Internet Brain Segmentation Repository (IBSR) is the second brain MRI dataset being used for validation purpose. It is also an openly accessible dataset acquired from the Internet Brain Segmentation Repository (IBSR)² with given ground truth or manually segmented images. Skull stripping for all the MRI images is done using the brain extraction tool.³ Table 3 shows the detailed description of the different datasets being used for validation of the proposed approach for image segmentation problem.

Table 4 Average segmentation accuracy for synthetic image dataset corrupted by Gaussian noise

Methods\Images	GN_005	GN_010	GN_015
FCM_S	99.31	98.09	95.45
FLICM	99.11	98.96	98.86
EnFCM	98.37	98.09	97.56
FGFCM	98.93	98.74	98.00
KWFLICM	99.97	99.90	99.72
NDFCM	98.92	98.69	97.90
IPFCM	98.33	97.60	96.17
KGFCM	95.46	98.07	97.50
APFCM_S	98.12	98.05	97.67
IFCMSNI	99.94	96.19	84.04
IIFCM	99.66	96.45	91.11
FRFCM	98.83	98.66	98.13
DSFCMN	98.36	95.76	94.45
WRFCM	99.13	98.36	97.89
KFPKL	99.99	99.96	99.93

¹BrainWeb [online], available: <http://www.brainweb.bic.mni.mcgill.ca/brainweb>.

²IBSR [online], available: <http://www.cma.mgh.harvard.edu/ibsr/>

³Brain Extraction Tool (BET) [online], available: <http://www.fmrib.ox.ac.uk/fs/>

Table 5 Dice Score for synthetic image dataset corrupted by Gaussian noise

Methods/Images	GN_005				GN_010				GN_015			
	Region 1	Region 2	Region 3	Region 4	Region 1	Region 2	Region 3	Region 4	Region 1	Region 2	Region 3	Region 4
FCM_S	0.9875	0.9869	0.9824	0.9968	0.9869	0.9831	0.9291	0.9884	0.9824	0.9446	0.8053	0.9751
FLICM	0.9831	0.9668	0.9921	0.9959	0.9994	0.9863	0.9569	0.9942	0.9832	0.9548	0.9867	0.9952
EnFCM	0.9901	0.9869	0.9448	0.9896	0.9882	0.9813	0.9361	0.9883	0.9872	0.9795	0.9181	0.9845
FGFCM	0.9878	0.9762	0.9695	0.9953	0.9875	0.9769	0.9621	0.9938	0.9856	0.9747	0.9355	0.9886
KWFLICM	0.9998	0.9999	0.9989	0.9997	0.9994	0.9997	0.9966	0.9988	0.9989	0.9991	0.9901	0.9981
NDFCM	0.9878	0.9762	0.9691	0.9952	0.9875	0.9768	0.9601	0.9934	0.9856	0.9748	0.9317	0.9877
IPFCM	0.9918	0.9896	0.9423	0.9891	0.9889	0.9846	0.9184	0.9841	0.9867	0.9798	0.8721	0.9737
KGFCM	0.9882	0.9801	0.9403	0.9892	0.9884	0.9815	0.9335	0.9882	0.9871	0.9791	0.9158	0.9841
APFCM_S	0.9880	0.9788	0.9381	0.9890	0.9878	0.9793	0.9351	0.9883	0.9872	0.9781	0.9222	0.9856
IFCMSNI	1.0000	0.9999	0.9973	0.9995	0.9995	0.9981	0.8650	0.9709	0.9834	0.9417	0.5919	0.8759
IIFCM	0.9994	0.9987	0.9868	0.9977	0.9874	0.9734	0.8755	0.9772	0.9618	0.9063	0.7274	0.9452
FRFCM	0.9989	0.9984	0.955	0.9915	0.9977	0.9958	0.9499	0.9905	0.9962	0.9933	0.9314	0.9868
DSFCMN	0.9781	0.9673	0.9573	0.9921	0.9401	0.9166	0.8849	0.9812	0.9376	0.9005	0.8508	0.9709
WRFCM	0.9883	0.9823	0.9751	0.9961	0.9902	0.9873	0.9445	0.9895	0.9889	0.9839	0.9286	0.9864
KFPKL	1.0000	1.0000	0.9993	0.9998	1.0000	0.9999	0.9982	0.9996	1.0000	0.9998	0.9974	0.9995

Table 6 Average segmentation accuracy for synthetic image dataset corrupted by Rician noise

Methods/Images	RN_05	RN_10	RN_15
FCM_S	100.00	99.99	99.53
FLICM	100.00	99.80	99.07
EnFCM	100.00	99.45	98.75
FGFCM	98.99	98.97	98.96
KWFLICM	100.00	100.00	99.98
NDFCM	98.99	98.97	98.96
IPFCM	99.80	99.61	98.74
KGFCM	100.00	99.93	99.24
APFCM_S	100.00	98.16	98.14
IFCMSNI	100.00	100.00	100.00
IIFCM	100.00	100.00	99.93
FRFCM	99.00	98.95	98.89
DSFCMN	99.00	98.13	97.81
WRFCM	99.17	99.16	99.15
KFPKL	100.00	100.00	100.00

5.2 Results on synthetic dataset

We demonstrate the robustness of the proposed method in handling different types of noise with varying intensity levels. We have conducted the experiments on synthetic image data with added Gaussian white noise (standard deviations of 5, 10 and 15) and Rician noise (standard deviations of 5, 10 and 15). The results measuring average segmentation accuracy and Dice score for these images are shown in Tables 4, 5, 6 and 7. The experimental results show that the proposed KFPKL method performs better than any of the mentioned methods in suppressing Gaussian white noise and Rician noise. The performance of all of the other mentioned methods degrade significantly with increase in noise level but the performance of proposed KFPKL method remained consistent and wasn't affected much by it.

The IFCMSNI method also performed well while segmenting the images corrupted Rician noise. The qualitative segmentation results obtained using the proposed KFPKL method and the other related methods corresponding to synthetic gray image corrupted with Gaussian white noise with standard deviation 5 are shown in Fig. 5. The proposed KFPKL method have performed better in comparison to the existing state of the art methods as shown in Fig. 5. The proposed approach suppresses the noise without losing fine details of image in the segmentation process. The similar observation can be noted from Fig. 6 for the synthetic image corrupted with Rician noise with standard deviation 5.

5.3 Qualitative results on coin image

This section presents the analysis on qualitative results obtained for the natural images corrupted with salt and

Table 7 Dice Score for synthetic image dataset corrupted by Rician noise

Methods\Images	RN_5				RN_10				RN_15			
	Region 1	Region 2	Region 3	Region 4	Region 1	Region 2	Region 3	Region 4	Region 1	Region 2	Region 3	Region 4
	Region 1	Region 2	Region 3	Region 4	Region 1	Region 2	Region 3	Region 4	Region 1	Region 2	Region 3	Region 4
FCM.S	1.0000	1.0000	1.0000	1.0000	0.9998	0.9998	0.9997	0.9999	0.9875	0.9803	0.9915	0.9994
FLICM	1.0000	1.0000	1.0000	1.0000	0.9967	0.9945	0.9947	0.9993	0.9877	0.9796	0.9733	0.9961
EnFCM	1.0000	1.0000	1.0000	1.0000	0.9988	0.9987	0.9791	0.9961	0.9961	0.9952	0.9539	0.9912
FGFCM	0.9881	0.9762	0.9721	0.9956	0.9879	0.9764	0.9712	0.9955	0.9878	0.9773	0.9701	0.9955
KWFLICM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9998	0.9999	0.9999	0.9999	0.9994	0.9998
NDFCM	0.9881	0.9762	0.9721	0.9956	0.9879	0.9764	0.9711	0.9955	0.9878	0.9773	0.9699	0.9954
IPFCM	0.9999	0.9999	0.9920	0.9985	0.9993	0.9992	0.9846	0.9971	0.9945	0.9937	0.9550	0.9914
KGFCM	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9971	0.9994	0.9970	0.9961	0.9722	0.9948
APFCM.S	1.0000	1.0000	1.0000	1.0000	0.9880	0.9789	0.9393	0.9893	0.9879	0.9793	0.9384	0.9891
IFCMSNI	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	1.0000	1.0000	0.9996	0.9999
IIFCM	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9972	0.9994
FRFCM	1.0000	1.0000	0.9604	0.9924	0.9997	0.9999	0.9588	0.9921	0.9988	0.9989	0.9569	0.9918
DSFCMN	0.9871	0.9824	0.9713	0.9951	0.9808	0.9758	0.9480	0.9887	0.9722	0.9620	0.9403	0.9889
WRFCM	0.9967	0.9880	0.9771	0.9811	0.9965	0.9880	0.9768	0.9818	0.9964	0.9881	0.9764	0.9821
KFPKL	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	1.0000	1.0000	0.9998	0.9999

pepper noise and Gaussian noise. Figures 7 and 8 show the segmentation results of the coin's image, corrupted with salt & paper noise of intensity of 5% and the Gaussian white noise with standard deviation 5, respectively. The results of KFPKL are more robust and precise in comparison to other related methods. KWFLICM, FLICM, FRFCM, and WRFCM also perform well in suppressing Gaussian and salt & pepper noise, as shown in Figs. 7 and 8. It can be observed that IFCMSNI, EnFCM, FCM.S, KGFCM, NDFCM, IPFCM, FGFCM, DSFCMN, and APFCM.S cannot suppress these noise efficiently in natural images. We observe that most of the mentioned methods do not entirely recognize the noise, the proposed KFPKL method reduces the noise's impact and obtains the correct segmentation result.

5.4 Results on simulated MRI dataset

The experimental results obtained on simulated MRI BrainWeb dataset for different performance measures are summarized in Tables 8, 9, 10, and 11. From these tables, it can be point out that the proposed KFPKL method gives best results in terms of ASA and DS in almost all the cases. Though in some cases the average segmentation accuracy and Dice scores of IFCMSNI, KGFCM, IIFCM and EnFCM are comparable to the proposed method but when observed on the whole dataset, the proposed method surely have an upper hand over other related methods. The comparison of DS for the proposed KFPKL method with existing methods corresponding to segmentation of WM, GM and CSF in MR images of the brain are listed in the Tables 9, 10 and 11 respectively. In case of GM, KFPKL performs better in 5 out of 9 scenarios. FLICM, IIFCM and IFCMSNI also performs better in one of the cases although the results are closed to KFPKL. In case of WM, the proposed KFPKL performs better in 6 out of 9 scenarios whereas the performance of all of the other methods were below the proposed method. For CSF, it could be observed that the KFPKL, KGFCM, IIFCM and IFCMSNI performed well and showed very similar performance. From the above observation and discussion on ASA and DS result metrics, it can be concluded that the performance of the proposed KFPKL method is superior on this dataset.

Figures 9, 10 and 11 show the variation in performance, in terms of ASA corresponding to all the methods considered in this work for 0%, 20% and 40% intensity non-uniformity, respectively. For a fixed INU level, we consider six different noise levels 0, 1, 3, 5, 7 and 9, and observe that KFPKL is effective in handling the noise. The ASA of all the methods decrease with increase in noise levels for a given intensity non-uniformity, but the performance of the KFPKL does not degrade much when compared to other related methods. We further show qualitative results

Fig. 6 Qualitative results obtained on synthetic gray image corrupted with Rician noise

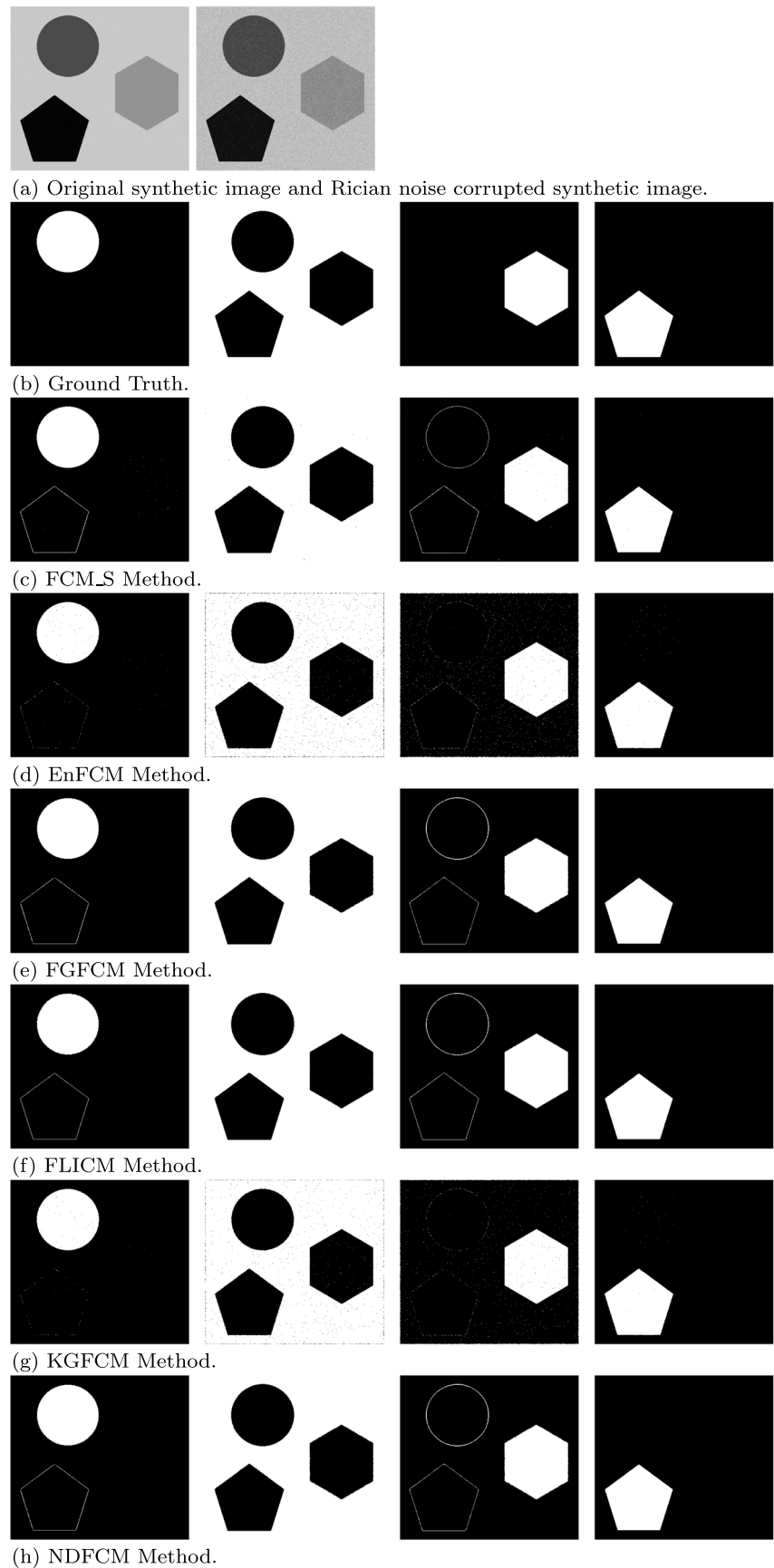
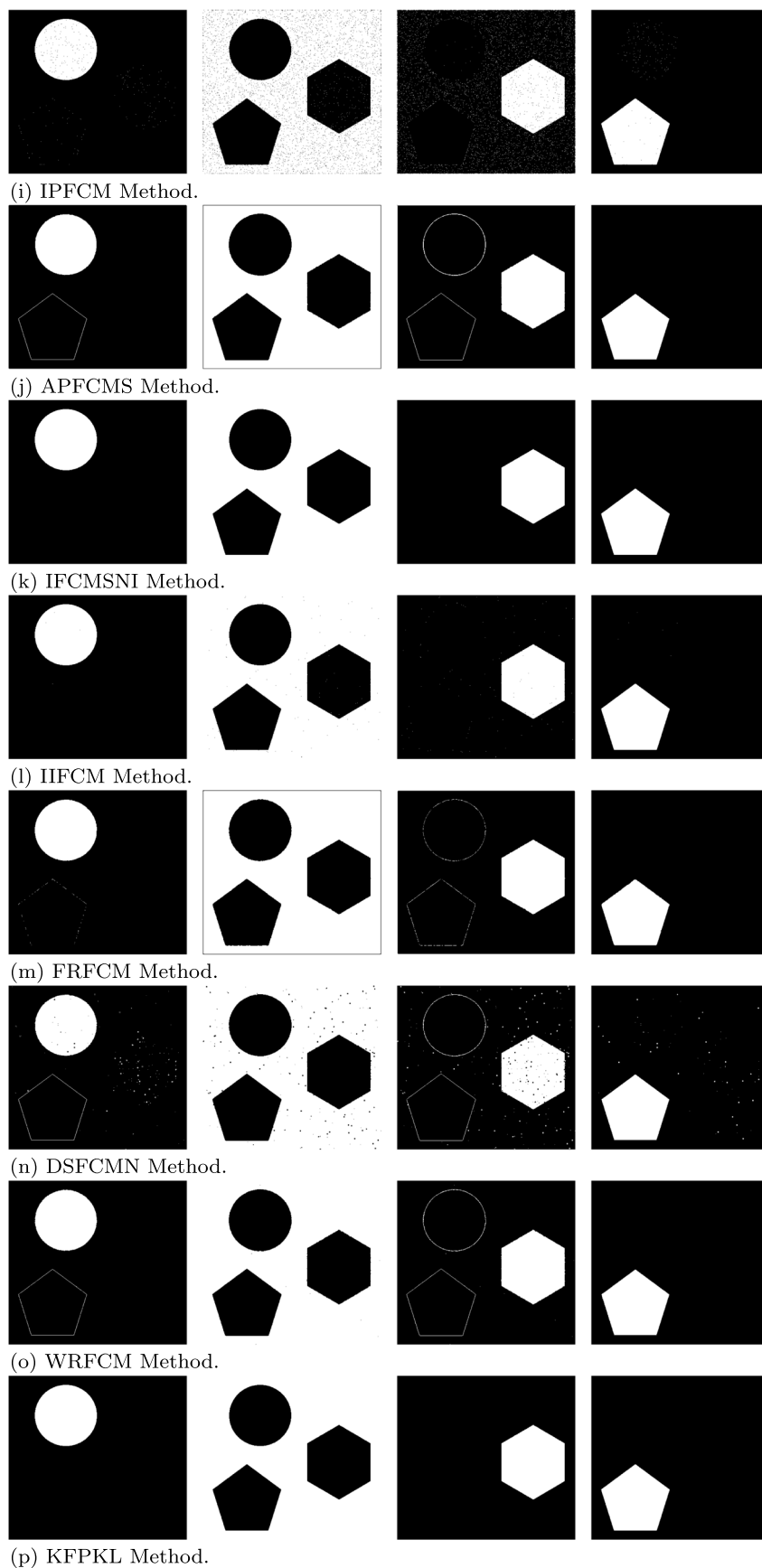


Fig. 6 (continued)



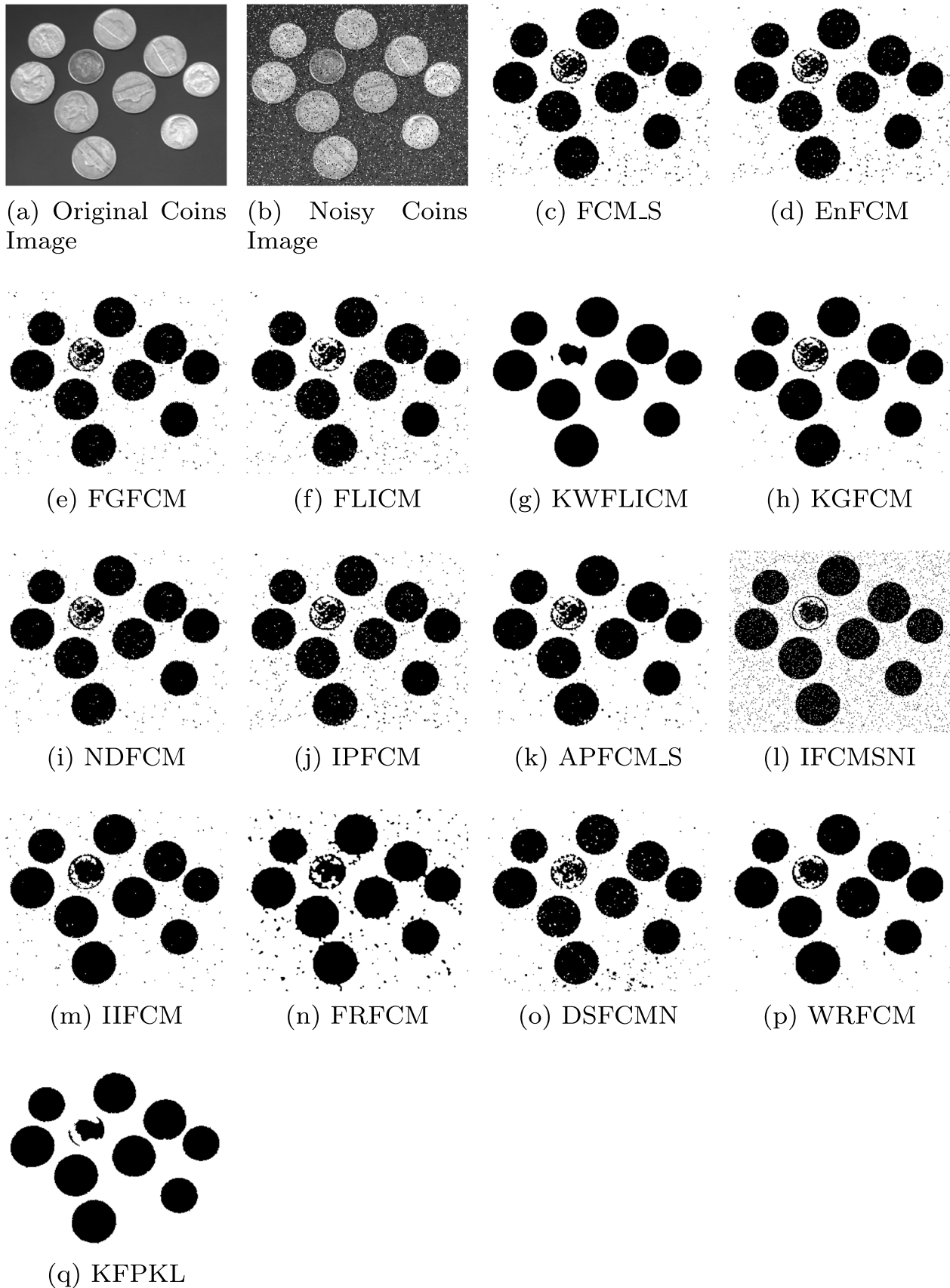
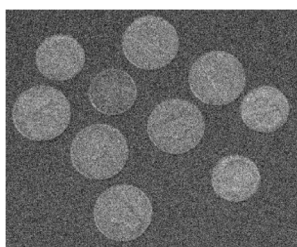


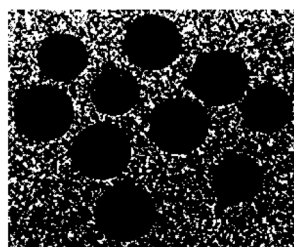
Fig. 7 Qualitative results obtained on coin image corrupted with SP noise



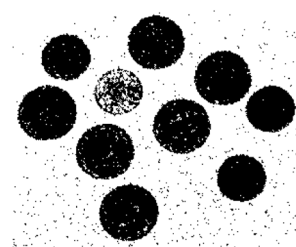
(a) Original Coins Image



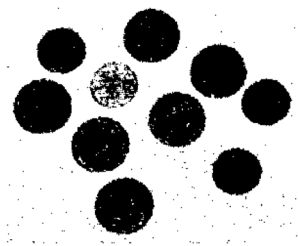
(b) Noisy Coins Image



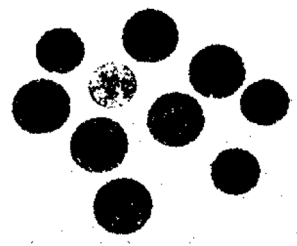
(c) FCM-S



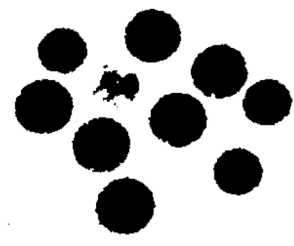
(d) EnFCM



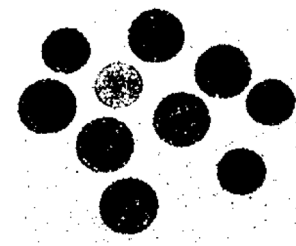
(e) FGFCM



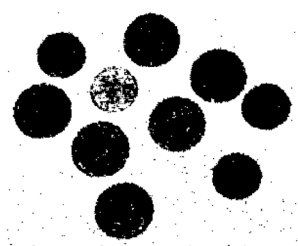
(f) FLICM



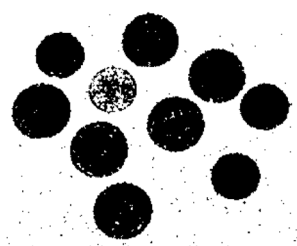
(g) KWFLICM



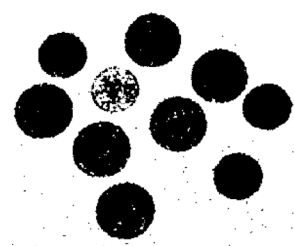
(h) KGFCM



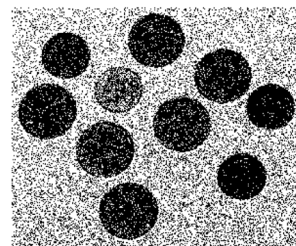
(i) NDFCM



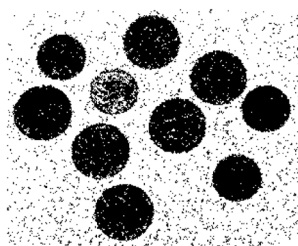
(j) IPFCM



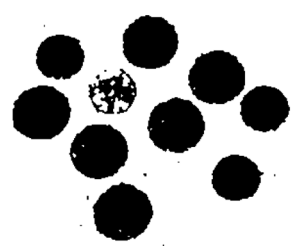
(k) APFCM-S



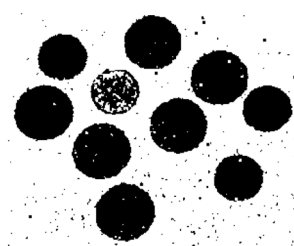
(l) IFCMSNI



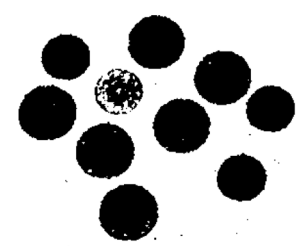
(m) IIFCM



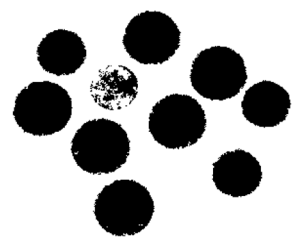
(n) FRFCM



(o) DSFCMN



(p) WRFCM



(q) KF PKL

Fig. 8 Qualitative results obtained on coin image corrupted with Gaussian noise

Table 8 Average segmentation accuracy for BrainWeb dataset

Methods\Images	Noise level 5			Noise level 7			Noise level 9		
	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40
FCMLS	0.9478	0.9312	0.9093	0.9387	0.9347	0.9030	0.9321	0.9196	0.8957
FLICM	0.9359	0.9284	0.9089	0.9387	0.9321	0.9036	0.9321	0.9197	0.8942
EnFCM	0.9504	0.9459	0.9267	0.9393	0.9332	0.9134	0.9267	0.9207	0.9017
FGFCM	0.9232	0.9160	0.8952	0.9177	0.9113	0.8933	0.9073	0.9068	0.8879
KWFLICM	0.9366	0.9289	0.9067	0.9283	0.9217	0.8979	0.9220	0.9123	0.8973
NDFCM	0.9228	0.9158	0.8944	0.9165	0.9099	0.8887	0.9070	0.9053	0.8824
IPFCM	0.9508	0.9481	0.9278	0.9395	0.9327	0.9127	0.9288	0.9226	0.9034
KGFCM	0.9559	0.9516	0.9306	0.9417	0.9355	0.9144	0.9302	0.9242	0.9041
APFCMS	0.9301	0.9243	0.9065	0.9255	0.9198	0.9007	0.9181	0.9165	0.8961
IFCMSNI	0.9511	0.9493	0.9345	0.9406	0.9389	0.9229	0.9347	0.9282	0.9096
IIFCM	0.9541	0.9513	0.9392	0.9397	0.9397	0.9243	0.9239	0.9241	0.9105
FRFCM	0.9201	0.9112	0.8953	0.9134	0.9024	0.8841	0.9027	0.8960	0.8730
DSFCMN	0.9329	0.9272	0.9072	0.9318	0.9240	0.9079	0.9202	0.9148	0.8926
WRFCM	0.9387	0.9313	0.9127	0.9335	0.9264	0.9032	0.9232	0.9175	0.8959
KFPKL	0.9559	0.9552	0.9377	0.9436	0.9394	0.9215	0.9334	0.9284	0.9118

obtained for different methods along with the proposed KFPKL corresponding to the MR image (with noise level = 9%) on an axial slice 90 shown in Fig. 12. The ground truth for the MR image corresponding to WM, GM, and CSF is also shown in Fig. 12 (b). It can be pointed out that the qualitative segmentation results obtained using the proposed method are robust to noise to a greater extent when compared to other related works.

5.5 Results on real MRI dataset

The experimental results obtained on real MRI dataset (corrupted by rician noise with standard deviation 5) are summarized in Tables 12, 13 and 14. From these tables, it is evident that the proposed KFPKL method is proficient for brain MR image segmentation task. Average segmentation accuracy of the proposed KFPKL method is the best in

Table 9 Dice score for WM BrainWeb dataset

Methods\Images	Noise level 5			Noise level 7			Noise level 9		
	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40
FCMLS	0.9636	0.9511	0.9251	0.9563	0.9512	0.9175	0.9511	0.9383	0.9123
FLICM	0.9576	0.9509	0.9325	0.9568	0.9502	0.9199	0.9511	0.9395	0.9104
EnFCM	0.9649	0.9607	0.9429	0.9573	0.9521	0.9338	0.9482	0.9427	0.9242
FGFCM	0.9528	0.9475	0.9291	0.9471	0.9418	0.9258	0.9398	0.9397	0.9215
KWFLICM	0.9664	0.9591	0.9361	0.9595	0.9531	0.9288	0.9534	0.9474	0.9297
NDFCM	0.9529	0.9476	0.9291	0.9475	0.9419	0.9234	0.9413	0.9395	0.9189
IPFCM	0.9651	0.9621	0.9435	0.9574	0.9529	0.9339	0.9516	0.9465	0.9266
KGFCM	0.9693	0.9651	0.9463	0.9591	0.9546	0.9432	0.9525	0.9474	0.9269
APFCMS	0.9552	0.9509	0.9361	0.9377	0.9328	0.9216	0.8996	0.9016	0.8919
IFCMSNI	0.9632	0.9565	0.9491	0.9683	0.9688	0.9624	0.9496	0.9419	0.9313
IIFCM	0.9701	0.9674	0.9569	0.9603	0.9574	0.9435	0.9478	0.9465	0.9338
FRFCM	0.9569	0.9489	0.9333	0.9518	0.9424	0.9242	0.9435	0.9366	0.916
DSFCMN	0.9572	0.9532	0.9339	0.9557	0.9496	0.9318	0.9489	0.9463	0.9218
WRFCM	0.9613	0.9552	0.9395	0.9575	0.9522	0.9311	0.9511	0.9472	0.9273
KFPKL	0.9703	0.9681	0.9532	0.9629	0.9604	0.9439	0.9587	0.9513	0.9337

Table 10 Dice score for GM BrainWeb dataset

Methods\Images	Noise level 5			Noise level 7			Noise level 9		
	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40
FCMLS	0.9039	0.8855	0.8902	0.8952	0.9011	0.8981	0.8903	0.8873	0.8858
FLICM	0.8772	0.8742	0.8668	0.8917	0.8903	0.8856	0.8902	0.8817	0.8861
EnFCM	0.9317	0.9262	0.9018	0.9076	0.9022	0.8927	0.8946	0.8916	0.8861
FGFCM	0.8537	0.8485	0.8389	0.8546	0.8506	0.8419	0.8426	0.8424	0.8393
KWFLICM	0.8853	0.8832	0.8769	0.8791	0.8774	0.8715	0.8725	0.8645	0.8704
NDFCM	0.8526	0.8473	0.8368	0.8502	0.8447	0.8304	0.8364	0.8362	0.8254
IPFCM	0.9321	0.9291	0.9031	0.9091	0.8964	0.8915	0.8853	0.8815	0.8771
KGFCM	0.9382	0.9335	0.9069	0.9212	0.9134	0.8861	0.9067	0.8989	0.8731
APFCMS	0.8704	0.8655	0.8551	0.8838	0.8765	0.8626	0.8183	0.8276	0.8125
IFCMSNI	0.9285	0.9162	0.9006	0.9417	0.9424	0.9307	0.9103	0.8973	0.8748
IIFCM	0.936	0.9331	0.917	0.9169	0.9171	0.8983	0.894	0.8956	0.8777
FRFCM	0.8931	0.8826	0.8636	0.8857	0.8718	0.8496	0.8702	0.8643	0.8367
DSFCMN	0.9116	0.9052	0.8796	0.9128	0.8999	0.8795	0.8961	0.8908	0.857
WRFCM	0.9191	0.9096	0.8872	0.9127	0.9043	0.8744	0.8989	0.8934	0.865
KFPKL	0.9389	0.9385	0.9161	0.9318	0.9212	0.8974	0.9267	0.9043	0.8818

comparison to other related methods. The proposed KFPKL method achieves the best Dice score in comparison to the related methods while segmenting GM in the real MRI images of brain. In case of WM, though the proposed KFPKL method do not perform the best in some of the cases like 001_24, 007_08 and 002_04 but the results were certainly comparable to the best performing method. When compared over the all dataset, the proposed KFPKL method has outperformed.

5.6 Time complexity

To show the computational efficacy of the proposed KFPKL method in comparison to other related methods, we have obtained the average segmentation time corresponding to image with size N . Let c be the number of segments produced and I be the iterations required. Further, w represents the size of filtering window, N_j represents the number of neighboring data points and L represents the

Table 11 Dice score for CSF BrainWeb dataset

Methods\Images	Noise level 5			Noise level 7			Noise level 9		
	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40	INU 0	INU 20	INU 40
FCMLS	0.9301	0.9108	0.8839	0.8952	0.9011	0.8981	0.8903	0.8873	0.8858
FLICM	0.9152	0.9061	0.8822	0.9185	0.9105	0.8771	0.9096	0.8959	0.8654
EnFCM	0.9424	0.9403	0.9312	0.9173	0.9102	0.8851	0.9007	0.8934	0.8694
FGFCM	0.8992	0.8903	0.8649	0.8901	0.8823	0.8597	0.8756	0.8766	0.8511
KWFLICM	0.9277	0.9181	0.8912	0.9188	0.9103	0.8819	0.9114	0.9003	0.8797
NDFCM	0.8991	0.8906	0.8646	0.8895	0.8817	0.8551	0.8776	0.8773	0.8478
IPFCM	0.9488	0.9518	0.9413	0.9177	0.9102	0.8844	0.9041	0.8965	0.8719
KGFCM	0.9531	0.9534	0.9417	0.9264	0.9321	0.9253	0.8903	0.8861	0.8864
APFCMS	0.9069	0.9004	0.8781	0.9186	0.9191	0.9098	0.8588	0.8874	0.8829
IFCMSNI	0.9529	0.9383	0.9220	0.9640	0.9638	0.9533	0.9459	0.9219	0.9236
IIFCM	0.9444	0.9439	0.9369	0.928	0.9378	0.9289	0.9179	0.9213	0.9168
FRFCM	0.8373	0.8325	0.8286	0.8312	0.8254	0.8204	0.8267	0.8242	0.811
DSFCMN	0.8773	0.8728	0.8676	0.8785	0.8726	0.862	0.8648	0.856	0.3712
WRFCM	0.8822	0.8782	0.866	0.8795	0.8719	0.8609	0.8669	0.8612	0.8515
KFPKL	0.9515	0.9557	0.9442	0.9237	0.9311	0.9247	0.9137	0.9163	0.9164

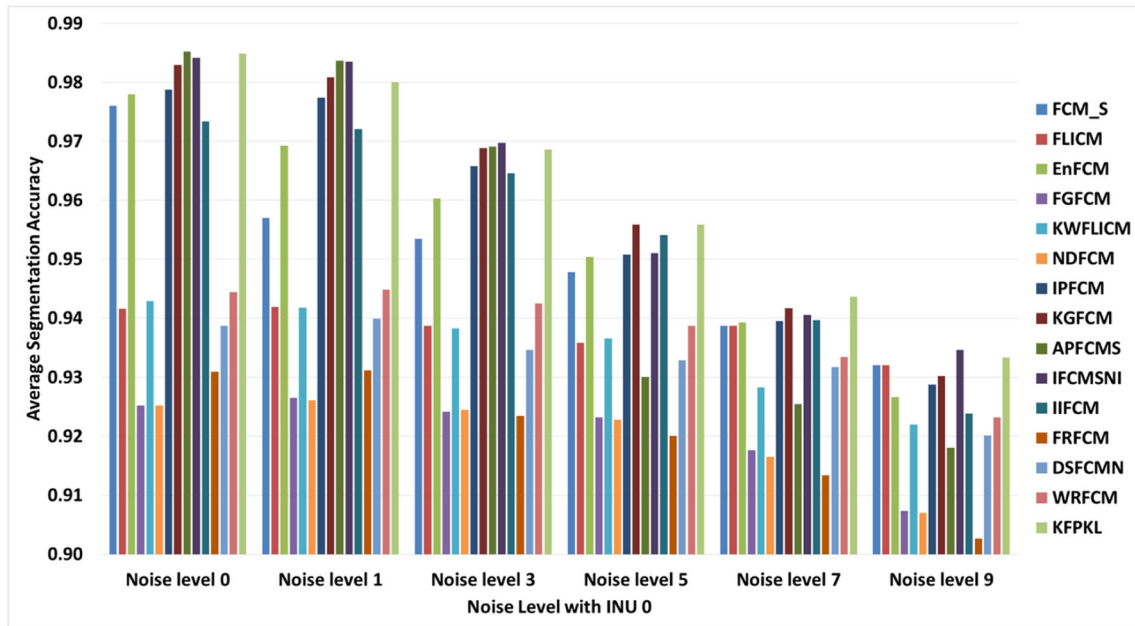


Fig. 9 Variation in performance in terms of ASA on images with 0% INU

pixel intensity levels in the image. The algorithmic time complexity and the computation time of all the methods for the BrainWeb image of size 181×217 are presented in Table 15. The average computation time corresponding to all the methods have been obtained in the same environment for fair comparison. It can be observed from the Table 15 that some of the methods having similar asymptotic time complexity but taking different average computation time (in sec.) for segmentation task. It can be pointed out the

methods for which the average computation time are less are converging faster in comparison to the rival methods i.e., taking less number of iteration to get the final segmentation result.

5.7 Statistical test

To further evaluate the performance of the proposed KFPKL method, we conduct statistical tests for the different

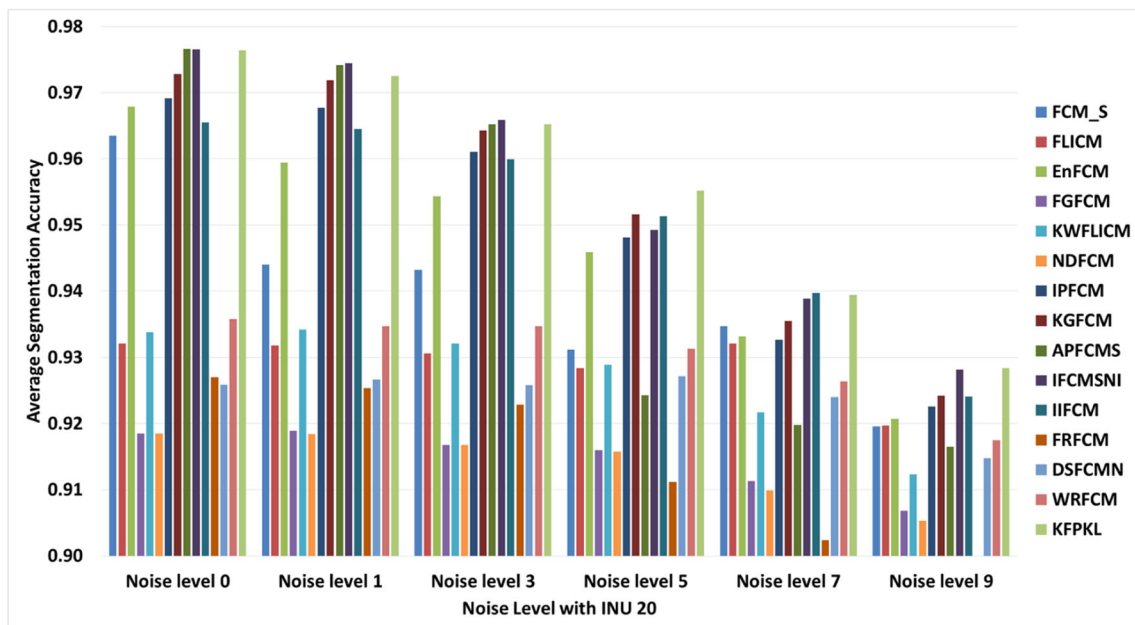


Fig. 10 Variation in performance in terms of ASA on images with 20% INU

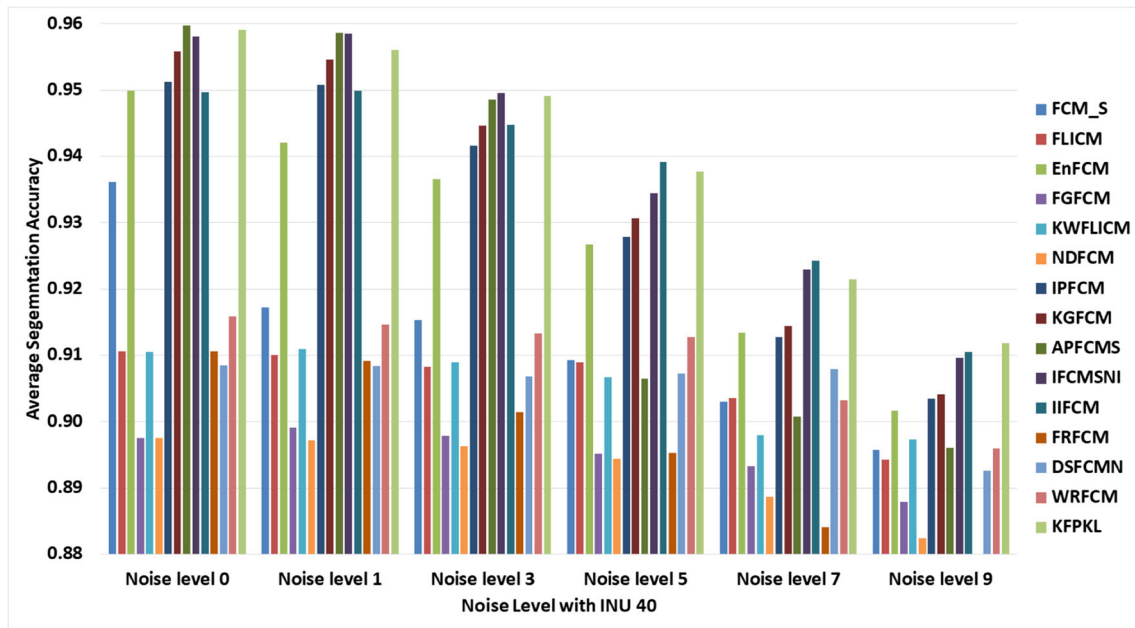


Fig. 11 Variation in performance in terms of ASA on images with 40% INU

methods on Brainweb and IBSR dataset. We use Friedman test, a two way parameter free statistical test used to find out if there is any considerable difference between the methods on the basis of observed results. It has two hypotheses i.e. the null-hypothesis (H_0) and alternative hypothesis (H_1) respectively. If there is no major difference in the performance of the proposed method and other related methods then null-hypothesis is satisfied and if the performance gap is larger than the alternative hypothesis is satisfied. Formally (H_0) and (H_1) for a performance measure M can be presented as:

$$\begin{aligned}
 H_0 : & \mu_{FCM_S} = \mu_{FLICM} = \mu_{EnFCM} = \mu_{FGFCM} = \mu_{KWFLICM} \\
 & = \mu_{NDFCM} = \mu_{IPFCM} = \mu_{KGFCM} = \mu_{APFCM_S} = \mu_{IFCMSNI} \\
 & = \mu_{IIFCM} = \mu_{FRFCM} = \mu_{DSFCMN} = \mu_{WRFCM} = \mu_{KFPKL} \\
 H_1 : & \mu_{FCM_S} \neq \mu_{FLICM} \neq \mu_{EnFCM} \neq \mu_{FGFCM} \neq \mu_{KWFLICM} \\
 & \neq \mu_{NDFCM} \neq \mu_{IPFCM} \neq \mu_{KGFCM} \neq \mu_{APFCM_S} \neq \mu_{IFCMSNI} \\
 & \neq \mu_{IIFCM} \neq \mu_{FRFCM} \neq \mu_{DSFCMN} \neq \mu_{WRFCM} \neq \mu_{KFPKL}
 \end{aligned}
 \tag{24}$$

We particularly use the average segmentation accuracy obtained from the different methods to rank them. In

Table 12 Average segmentation accuracy for IBSR dataset

Methods\Cases	100_23	191_03	001_24	007_08	002_04
FCM_S	0.6605	0.6793	0.6876	0.5694	0.4926
FLICM	0.6248	0.6714	0.6834	0.5386	0.5083
EnFCM	0.6497	0.6513	0.6799	0.5493	0.4502
FGFCM	0.6786	0.6642	0.6890	0.5454	0.4659
KWFLICM	0.6944	0.6785	0.6892	0.5608	0.4448
NDFCM	0.7044	0.6728	0.6944	0.5608	0.4970
IPFCM	0.6386	0.6464	0.6744	0.5421	0.4331
KGFCM	0.6343	0.5801	0.6876	0.5288	0.4272
APFCM_S	0.6517	0.6599	0.6838	0.5718	0.5177
IFCMSNI	0.6801	0.6633	0.6907	0.5719	0.5177
IIFCM	0.6574	0.6738	0.6894	0.5815	0.4241
FRFCM	0.6605	0.6355	0.6813	0.5427	0.4586
DSFCMN	0.6834	0.6586	0.6894	0.6151	0.4971
WRFCM	0.6355	0.6541	0.6714	0.5454	0.4873
KFPKL	0.7297	0.6871	0.7042	0.5846	0.5198

Table 13 Dice score for WM IBSR dataset

Methods\Cases	100_23	191_03	001_24	007_08	002_04
FCM_S	0.7804	0.7419	0.7874	0.6347	0.5216
FLICM	0.7847	0.7638	0.7932	0.6608	0.5199
EnFCM	0.7896	0.7652	0.7976	0.6776	0.4911
FGFCM	0.7876	0.7607	0.7906	0.6751	0.5034
KWFLICM	0.7876	0.7543	0.7914	0.6745	0.4758
NDFCM	0.7861	0.7546	0.7987	0.6829	0.5149
IPFCM	0.7895	0.7669	0.7968	0.6785	0.4842
KGFCM	0.7906	0.7681	0.7957	0.6795	0.4496
APFCM_S	0.7861	0.7538	0.7948	0.6896	0.4470
IFCMSNI	0.7849	0.7504	0.7954	0.6891	0.5987
IIFCM	0.7905	0.7540	0.7905	0.6816	0.4243
FRFCM	0.7612	0.7622	0.7811	0.6734	0.4975
DSFCMN	0.7839	0.6863	0.7501	0.6722	0.4362
WRFCM	0.7644	0.7471	0.7915	0.6749	0.4913
KFPKL	0.7954	0.7878	0.7491	0.6795	0.4421

Table 14 Dice Score for GM IBSR dataset

Methods\Cases	100_23	191_03	001_24	007_08	002_04
FCM_S	0.6937	0.7231	0.7258	0.5719	0.5522
FLICM	0.6403	0.6781	0.7135	0.5421	0.5922
EnFCM	0.6591	0.6521	0.7051	0.5435	0.5379
FGFCM	0.6927	0.6647	0.7147	0.5295	0.5451
KWFLICM	0.7081	0.6791	0.6939	0.5405	0.5342
NDFCM	0.7201	0.6736	0.7198	0.5431	0.5817
IPFCM	0.6467	0.6469	0.6991	0.5305	0.5154
KGFCM	0.5817	0.5711	0.7112	0.5082	0.5487
APFCM_S	0.6610	0.6602	0.7070	0.5549	0.4916
IFCMSNI	0.6877	0.6629	0.7124	0.5555	0.5221
IIFCM	0.6760	0.6810	0.7180	0.5768	0.5214
FRFCM	0.6689	0.6392	0.7134	0.5393	0.5251
DSFCMN	0.7017	0.6391	0.7389	0.6198	0.5166
WRFCM	0.6433	0.6567	0.6911	0.5201	0.5916
KFPKL	0.7441	0.7218	0.7373	0.5781	0.6124

Friedman test, the average rank R_j of j^{th} method for a given N number of images is obtained as:

$$R_j = \frac{1}{N} \sum_{i=1}^N r_i^j \quad (26)$$

where $r_i^j \in \{1, 2, \dots, k\} (1 \leq i \leq N, 1 \leq j \leq k)$ is rank value for i^{th} image and j^{th} method. The average Friedman ranking is obtained by using the ASA on 9 BrainWeb brain images (noise level 5, 7 and 9; INU 0, 20 and 40) and 5 real brain images for the different segmentation methods and is shown in Table 16 [19, 21]. The segmentation method with the lowest numerical value of rank is considered to be superior among all the methods being compared. From Table 16, it can be inferred that the Friedman ranking for the proposed KFPKL method is the least; hence, it performs better in comparison to other methods in terms of ASA. We have used the statistical hypothesis test proposed by Iman and Davenport for further investigation. Iman and Davenport [28] defined the statistic F_{ID} as:

$$F_{ID} = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} \quad (27)$$

which is distributed according to F-distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom, where χ_F^2 is the Friedman's statistic defined as $\frac{12N}{k(k+1)} \left[\sum_j R_j^2 - \frac{k(k+1)^2}{4} \right]$. In our experiments $k = 15$ and $N = 14$. The p -values obtained by Iman and Davenport statistic are $2.92E-19$ corresponding to the performance measures ASA, which advocates the rejection of null hypothesis H_0 as there is a significant difference among different segmentation methods at the significance level of 0.05.

However, these p -values obtained are not suitable for comparison with the control method. Although the obtained p -values conclude the rejection of null hypothesis H_0 , but are not appropriate for comparison with the control method. Hence adjusted p -values [19] are required to compare with the control method on the statistical ground. Adjusted p -values provide the correct correlation by considering the error accumulated with respect to a control method, the proposed KFPKL method (lowest rank for ASA). To calculate the adjusted p -values, some post-hoc procedures need to be defined, and one of the commonly used procedures is the Holm procedure. The adjusted p -values obtained from the post-hoc procedure are shown in Table 17. It can be observed that the performance of the proposed KFPKL method is significantly different from the other related methods in terms of ASA except for the IFCMSNI and IIFCM algorithms.

Based on the statistical test observation, further, the comparison of qualitative performance of the top 5 methods is presented in Fig. 13 corresponding to a sample image of BrainWeb MRI image corrupted with noise level 9. It can be observed that the qualitative segmentation result obtained using the proposed KFPKL method is far better than the other 4 top methods. The statistical test also verifies the same.

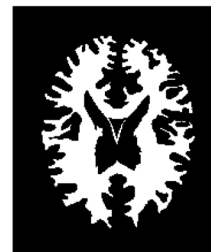
6 Conclusion

In this research, we presented a picture fuzzy set-based clustering to address the noise, vagueness, and non-linear structure in an image termed Kernel Fuzzy clustering for Picture fuzzy set using the Kullback Liebler divergence measure (KFPKL). To handle the vagueness associated with the data, KFPKL utilizes the picture fuzzy sets that improve clustering results by increasing the representational capability. The Gaussian kernel is utilized to deal with the non-linear structure present in data. Due to the transformation of data points in higher dimensional feature space, the kernel trick can solve the problem of inherent non-linearity present in data without increasing the computational complexity. The Kullback-Leibler (KL) divergence between the actual membership and the average neighborhood membership information in the PFS framework helps dampen the effect of noise in the segmentation process. The experiments are carried out on several synthetic image datasets and two publicly available brain MRI datasets. The comparison with the state-of-the-art method shows that the proposed KFPKL method provides better segmentation performance in terms of average segmentation accuracy and Dice score. Further, a statistical test is also conducted to show a significant difference in the performance of the proposed KFPKL and

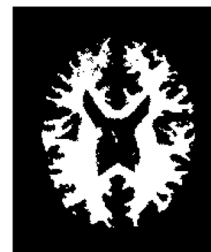
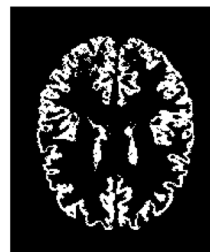
Fig. 12 Qualitative results on Brain Web MRI image



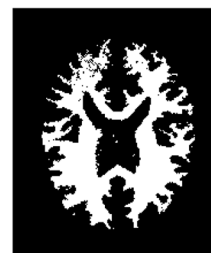
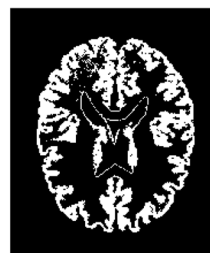
(a) Brain Web MRI image corrupted with 9% noise.



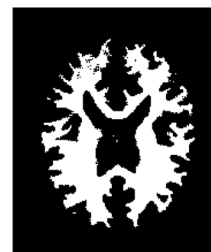
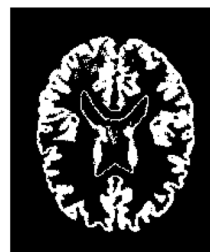
(b) Ground Truth.



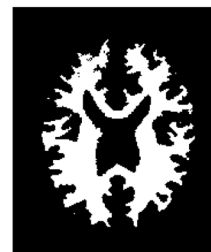
(c) FCM_S Method.



(d) EnFCM Method.



(e) FGFCM Method.

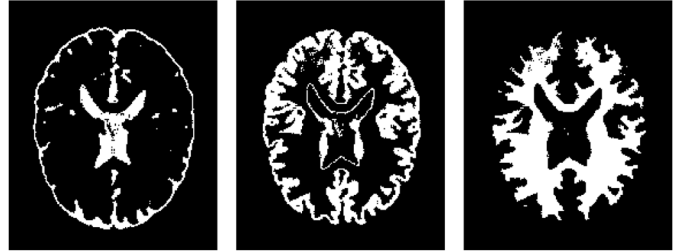


(f) FLICM Method.

Fig. 12 (continued)



(g) KWFLICM Method.



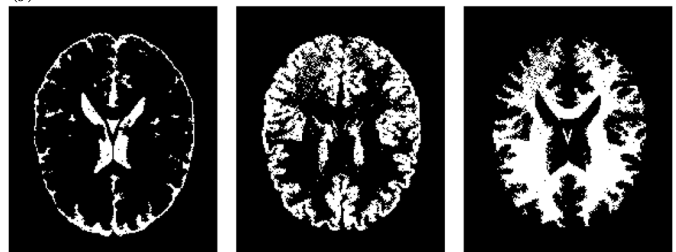
(h) NDFCM Method.



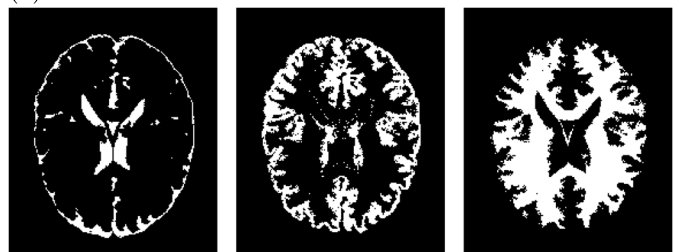
(i) IPFCM Method.



(j) APFCMS Method.



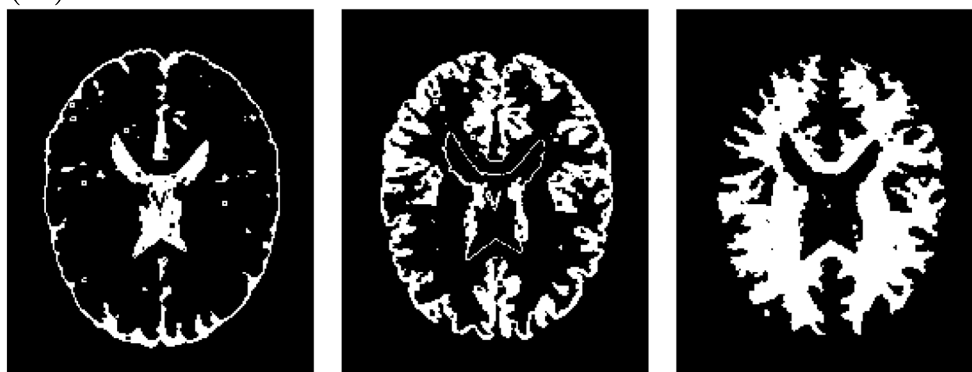
(k) IFCMSNI Method.



(l) IIFCM Method.



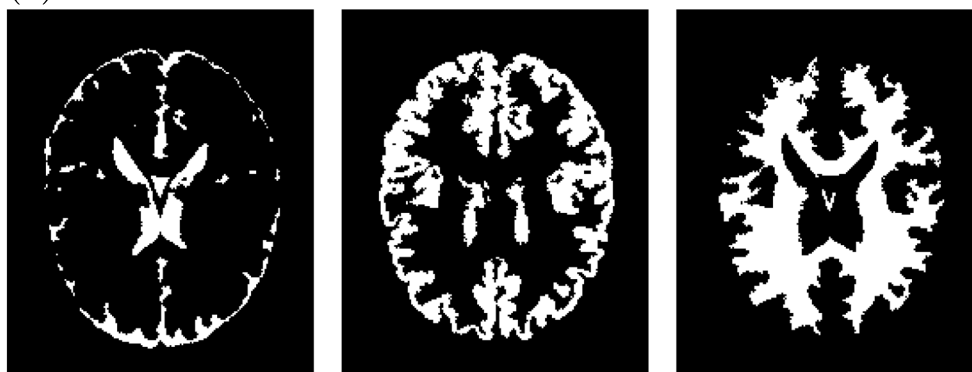
(m) FRFCM Method.



(n) DSFCMN Method.



(o) WRFCM Method.



(p) KFPKL Method.

Fig. 12 (continued)

Table 15 Comparison of asymptotic time complexity and computation time

Methods	Time Complexity	Time (in Sec.)	Methods	Time Complexity	Time (in Sec.)
FCM_S	$O(Nw + INc)$	1.7	APFCMS	$O(INc)$	4.5
FLICM	$O(INcN_j)$	6.5	IFCMSNI	$O(Nw + INc)$	5.8
EnFCM	$O(ILc)$	0.7	IIFCM	$O(Nw + INc)$	9.6
FGFCM	$O(Nw + ILc)$	3.1	FRFCM	$O(Nw + ILc)$	0.9
KWFLICM	$O(Nw + INcN_j)$	13.6	DSFCMN	$O(INcN_j)$	11.7
NDFCM	$O(Nw + ILc)$	4.3	WRFCM	$O(INcN_j)$	6.9
IPFCM	$O(INc)$	10.0	KFPKL	$O(INc)$	5.4
KGFCM	$O(INc)$	11.7			

Table 16 Average Friedman ranking of algorithms

Algorithm	Ranking	Algorithm	Ranking
KFPKL	1.46	KWFLICM	8.82
IFCMSNI	3.39	FLICM	8.86
IIFCM	4.46	WRFCM	9.54
KGFCM	6.86	APFCMS	9.82
FCM_S	7.00	NDFCM	10.68
EnFCM	7.86	FGFCM	11.25
IPFCM	8.14	FRFCM	13.46
DSFCMN	8.39		

Table 17 Adjusted p -value (Friedman)

Algorithm	unadjusted\$ p	\$ p _Holm\$	Algorithm	unadjusted\$ p	\$ p _Holm\$
FRFCM	1.25E-12	1.76E-11	DSFCMN	4.15E-05	2.90E-04
FGFCM	7.07E-09	9.19E-08	IPFCM	7.78E-05	4.67E-04
NDFCM	5.00E-08	6.00E-07	EnFCM	1.56E-04	7.78E-04
APFCMS	7.65E-07	8.41E-06	FCM_S	0.001	0.004
WRFCM	1.80E-06	1.80E-05	KGFCM	0.001	0.004
FLICM	1.22E-05	1.10E-04	IIFCM	0.076	0.152
KWFLICM	1.35E-05	1.10E-04	IFCMSNI	0.254	0.254

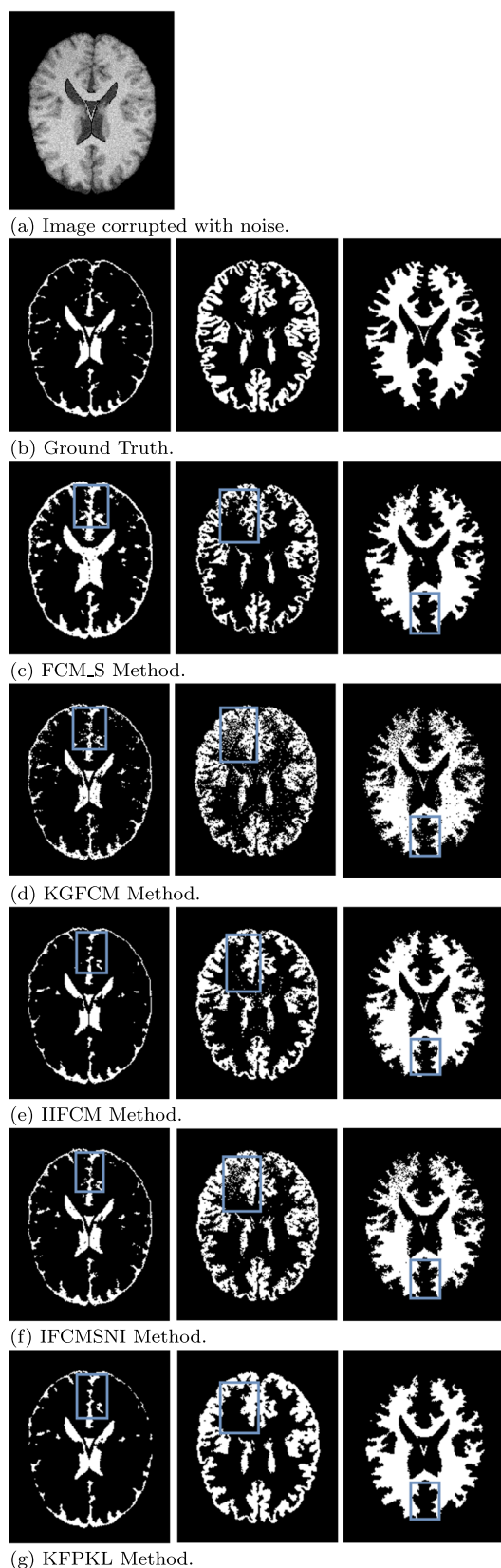


Fig. 13 Comparison of qualitative result obtained from top five methods based on the statistical test for BrainWeb MRI image corrupted with noise

state-of-the-art methods. Although the proposed method's performance is effective compared to other related methods, there are some issues that need further attention. The first issue is related to the selection of parameters crucial for getting accurate segmentation results. It will be an interesting future direction to develop methods that adaptively select the parameters. Also, another extension of fuzzy sets for developing a clustering method for image segmentation can be investigated in the future. Another open issue related to the clustering problem is to develop methods that automatically select the numbers of clusters.

Author Contributions There is equal contributions in this research from all the authors of this article.

Funding We have received no funding for this research.

Declarations

Ethics approval and consent to participate This article does not contain any studies with human participants or animals performed by any of the authors.

Consent for Publication Informed consent was obtained from all individual participants included in the study.

Conflict of Interests The authors declare that they have no conflict of interest.

References

1. Ahmed MN, Yamany SM, Mohamed N, Farag AA, Moriarty T (2002) A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. *IEEE Trans Med Imaging* 21(3):193–199
2. Alipour S, Shanbehzadeh J (2014) Fast automatic medical image segmentation based on spatial kernel fuzzy c-means on level set method. *Mach Vis Appl* 25(6):1469–1488
3. Atanassov KT (1986) Intuitionistic fuzzy sets. *Fuzzy Sets Syst* 20(1):87–96
4. Balafar MA, Ramli AR, Saripan MI, Mashohor S (2010) Review of brain mri image segmentation methods. *Artif Intell Rev* 33(3):261–274
5. Benaichouche A, Oulhadj H, Siarry P (2013) Improved spatial fuzzy c-means clustering for image segmentation using pso initialization, mahalanobis distance and post-segmentation correction. *Digit Signal Process* 23(5):1390–1400
6. Berkhin P (2006) A survey of clustering data mining techniques. In: *Grouping multidimensional data*. Springer, pp 25–71
7. Bezdek JC (1981) Objective function clustering. In: *Pattern recognition with fuzzy objective function algorithms*. Springer, pp 43–93
8. Bhargavi K, Jyothi S (2014) A survey on threshold based segmentation technique in image processing. *Int J Innov Res Dev* 3(12):234–239
9. Cai L, Gao J, Zhao D (2020) A review of the application of deep learning in medical image classification and segmentation. *Ann Trans Med* 8(11):713
10. Cai W, Chen S, Zhang D (2007) Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern Recogn* 40(3):825–838

11. Celenk M (1990) A color clustering technique for image segmentation. *Computer Vision Graph Image Process* 52(2):145–170
12. Chaira T (2011) A novel intuitionistic fuzzy c means clustering algorithm and its application to medical images. *Appl Soft Comput* 11(2):1711–1717
13. Chen S, Zhang D (2004) Robust image segmentation using fcm with spatial constraints based on new kernel-induced distance measure. *IEEE Trans Syst Man Cybern B Cybern* 34(4):1907–1916
14. Chen X, Nguyen BP, Chui CK, Ong SH (2016) Automated brain tumor segmentation using kernel dictionary learning and superpixel-level features. In: *Systems, man, and cybernetics (SMC), 2016 IEEE international conference on*. IEEE, pp 002,547–002,552
15. Cocosco CA, Kollokian V, Kwan RK-S, Evans AC (1997) “BrainWeb: Online Interface to a 3D MRI Simulated Brain Database” *NeuroImage*, vol.5, no.4, part 2/4, S425. In: *Proceedings of 3-rd International Conference on Functional Mapping of the Human Brain*, Copenhagen, May 1997
16. Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans Electron Comput* EC-14(3):326–334. <https://doi.org/10.1109/PGEC.1965.264137>
17. Cristianini N, Shawe-Taylor J (2000) *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press
18. Cuong BC, Kreinovich V (2013) Picture fuzzy sets-a new concept for computational intelligence problems. In: *2013 Third world congress on information and communication technologies (WICT 2013)*. IEEE, pp 1–6
19. Derrac J, García S, Molina D, Herrera F (2011) A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol Comput* 1(1):3–18
20. Dubey YK, Mushrif MM (2012) Segmentation of brain mr images using intuitionistic fuzzy clustering algorithm. In: *Proceedings of the eighth Indian conference on computer vision, graphics and image processing*. ACM, p 81
21. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc* 32(200):675–701
22. Gong M, Liang Y, Shi J, Ma W, Ma J (2012) Fuzzy c-means clustering with local information and kernel metric for image segmentation. *IEEE Trans Image Process* 22(2):573–584
23. Guo F, Wang XX, Shen J (2016) Adaptive fuzzy c-means algorithm based on local noise detecting for image segmentation. *IET Image Process* 10(4):272–279
24. Held K, Kops ER, Krause BJ, Wells WM, Kikinis R, Muller-Gartner HW (1997) Markov random field segmentation of brain mr images. *IEEE Trans Med Imaging* 16(6):878–886
25. Hesamian MH, Jia W, He X, Kennedy P (2019) Deep learning techniques for medical image segmentation: achievements and challenges. *J Digit Imaging* 32(4):582–596
26. Huang CW, Lin KP, Wu MC, Hung KC, Liu GS, Jen CH (2015) Intuitionistic fuzzy c-means clustering algorithm with neighborhood attraction in segmenting medical image. *Soft Comput* 19(2):459–470
27. Iakovidis D, Pelekis N, Kotsifakos E, Kopanakis I (2008) Intuitionistic fuzzy clustering with applications in computer vision. In: *Advanced concepts for intelligent vision systems*. Springer, pp 764–774
28. Iman RL, Davenport JM (1980) Approximations of the critical region of the fbietkan statistic. *Commun Stat-Theory Methods* 9(6):571–595
29. J Mercer B (1909) Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Phil Trans R Soc Lond A* 209(441–458):415–446
30. Ji ZX, Sun QS, Xia D (2014) A framework with modified fast fcm for brain mr images segmentation (retraction of vol 44, pg 999, 2011). *Pattern Recog* 47(12):3979–3979
31. Kameshwaran K, Malarvizhi K (2014) Survey on clustering techniques in data mining. *Int J Comput Sci Inf Technol* 5(2):2272–2276
32. Kannan S, Devi R, Ramathilagam S, Takezawa K (2013) Effective fcm noise clustering algorithms in medical images. *Comput Biol Med* 43(2):73–83
33. Kiran K, Srinivas K (2021) An efficient cluster system for bio-informatics data using amalgam of clustering methods. *Eur J Mol Clin Med* 7(10):1958–1971
34. Kotte VK, Rajavelu S, Rajasingh EB (2020) A similarity function for feature pattern clustering and high dimensional text document classification. *Found Sci* 25(4):1077–1094
35. Krinidis S, Chatzis V (2010) A robust fuzzy local information c-means clustering algorithm. *IEEE Trans Image Process* 19(5):1328–1337
36. Kumar D, Agrawal R, Verma H (2019) Kernel intuitionistic fuzzy entropy clustering for mri image segmentation. *Soft Comput* 24:4003–4026
37. Kumar D, Agrawal R, Kirar JS (2019) Intuitionistic fuzzy clustering method with spatial information for mri image segmentation. In: *2019 IEEE international conference on fuzzy systems (FUZZ-IEEE)*, pp 1–7
38. Kumar D, Verma H, Mehra A, Agrawal R (2019) A modified intuitionistic fuzzy c-means clustering approach to segment human brain mri image. *Multimed Tools Appl* 78(10):12,663–12,687
39. Lei T, Jia X, Zhang Y, He L, Meng H, Nandi AK (2018) Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. *IEEE Trans Fuzzy Syst* 26(5):3027–3041
40. Li C, Huang R, Ding Z, Gatenby JC, Metaxas DN, Gore JC (2011) A level set method for image segmentation in the presence of intensity inhomogeneities with application to mri. *IEEE Trans Image Process* 20(7):2007–2016
41. Lloyd S (1982) Least squares quantization in pcm. *IEEE Trans Inf Theory* 28(2):129–137
42. Moffat A, Stuiver L (1996) Exploiting clustering in inverted file compression. In: *Proceedings of data compression conference-DCC’96*. IEEE, pp 82–91
43. Olabarriaga SD, Smeulders AW (2001) Interaction in the segmentation of medical images: a survey. *Med Image Anal* 5(2):127–142
44. Pham DL, Xu C, Prince JL (2000) Current methods in medical image segmentation. *Ann Rev Biomed Eng* 2(1):315–337
45. Qiu C, Xiao J, Yu L, Han L, Iqbal MN (2013) A modified interval type-2 fuzzy c-means algorithm with application in mr image segmentation. *Pattern Recogn Lett* 34(12):1329–1338
46. Reddick WE, Glass JO, Cook EN, Elkin TD, Deaton RJ (1997) Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks. *IEEE Trans Med Imaging* 16(6):911–918
47. Rohlfing T, Brandt R, Menzel R, Maurer CR (2004) Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *Neuroimage* 21(4):1428–1442
48. Roth V, Steinlage V (2000) Nonlinear discriminant analysis using kernel functions. In: *Advances in neural information processing systems*, pp 568–574
49. Sato M, Lakare S, Wan M, Kaufman A, Nakajima M (2000) A gradient magnitude based region growing algorithm for accurate

- segmentation. In: Image processing, 2000. proceedings. 2000 international conference on. vol 3. IEEE, pp 448–451
50. Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput* 10(5):1299–1319
 51. Son LH (2015) Dpfc: a novel distributed picture fuzzy clustering method on picture fuzzy sets. *Expert Syst Appl Int J* 42(1):51–66
 52. Szilágyi L, Benyo Z, Szilágyi SM, Adam H (2003) Mr brain image segmentation using an enhanced fuzzy c-means algorithm. In: Engineering in medicine and biology society, 2003. Proceedings of the 25th annual international conference of the IEEE. vol 1. IEEE, pp 724–726
 53. Szmidi E, Kacprzyk J (2000) Distances between intuitionistic fuzzy sets. *Fuzzy Sets Syst* 114(3):505–518
 54. Thong PH et al (2016) Picture fuzzy clustering: a new computational intelligence method. *Soft Comput* 20(9):3549–3562
 55. Vapnik V (2013) The nature of statistical learning theory. Springer Science & Business Media
 56. Verma H, Agrawal R, Sharan A (2016) An improved intuitionistic fuzzy c-means clustering algorithm incorporating local information for brain image segmentation. *Appl Soft Comput* 46:543–557
 57. Vineetha G, Darshan G (2013) Level set method for image segmentation: a survey. *J Comput Eng* 8(6):74–78
 58. Wang C, Pedrycz W, Li Z, Zhou M (2020) Residual-driven fuzzy c-means clustering for image segmentation. *IEEE/CAA J Autom Sin* 8(4):876–889
 59. Wang C, Pedrycz W, Li Z, Zhou M, Zhao J (2021) Residual-sparse fuzzy c-means clustering incorporating morphological reconstruction and wavelet frame. *IEEE Trans Fuzzy Syst* 29:3910–3924. <https://doi.org/10.1109/TFUZZ.2020.3029296>
 60. Wang C, Pedrycz W, Zhou M, Li Z (2021) Sparse regularization-based fuzzy c-means clustering incorporating morphological grayscale reconstruction and wavelet frames. *IEEE Trans Fuzzy Syst* 29:1826–1840. <https://doi.org/10.1109/TFUZZ.2020.2985930>
 61. Wang L, Chen Y, Pan X, Hong X, Xia D (2010) Level set segmentation of brain magnetic resonance images based on local gaussian distribution fitting energy. *J Neurosci Methods* 188(2):316–325
 62. Wang XX, Guo F (2016) Adaptive fuzzy c-means algorithm based on local noise detecting for image segmentation. *IET Image Process* 10:272–279. <https://doi.org/10.1049/iet-ipr.2015.0236>
 63. Wang Z, Song Q, Soh YC, Sim K (2013) An adaptive spatial information-theoretic fuzzy clustering algorithm for image segmentation. *Comput Vis Image Underst* 117(10):1412–1420
 64. Wu C, Cao Z (2021) Noise distance driven fuzzy clustering based on adaptive weighted local information and entropy-like divergence kernel for robust image segmentation. *Dig Signal Proc* 111(102):963. <https://doi.org/10.1016/j.dsp.2021.102963>
 65. Wu C, Chen Y (2020) Adaptive entropy weighted picture fuzzy clustering algorithm with spatial information for image segmentation. *Appl Soft Comput* 86(105):888
 66. Wu C, Wu Q (2017) A robust image segmentation algorithm based on modified picture fuzzy clustering method on picture fuzzy sets. *J Xi'an Univ Posts Telecommun* 22(5):37–43
 67. Wu C, Zhang X (2021) A novel kernelized total bregman divergence-based fuzzy clustering with local information for image segmentation. *Int J Approx Reason* 136:281–305. <https://doi.org/10.1016/j.ijar.2021.06.004>. <https://www.sciencedirect.com/science/article/pii/S0888613X21000852>
 68. Xu Z, Wu J (2010) Intuitionistic fuzzy c-means clustering algorithms. *J Syst Eng Electron* 21(4):580–590
 69. Zadeh LA (1965) Fuzzy sets. *Inf Control* 8(3):338–353
 70. Zaixin Z, Lizhi C, Guangquan C (2014) Neighbourhood weighted fuzzy c-means clustering algorithm for image segmentation. *IET Image Proc* 8(3):150–161
 71. Zang W, Zhang W, Zhang W, Liu X (2017) A kernel-based intuitionistic fuzzy c-means clustering using a dna genetic algorithm for magnetic resonance image segmentation. *Entropy* 19(11):578
 72. Zhang Y, Bai X, Fan R, Wang Z (2018) Deviation-sparse fuzzy c-means with neighbor information constraint. *IEEE Trans Fuzzy Syst* 27(1):185–199
 73. Zhang Y, Brady M, Smith S (2001) Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 20(1):45–57
 74. Zhao F, Jiao L, Liu H (2013) Kernel generalized fuzzy c-means clustering with spatial information for image segmentation. *Digit Signal Proc* 23(1):184–199

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

A Review on Fabrication of Universal Drilling Machine

ROHAN KUMAR¹, SAATVIK SAGAR², AK MADAN³

^{1,2} Student, Mechanical Engineering, Delhi Technological University, New Delhi

³ Professor, Mechanical Engineering department, Delhi Technological University, New Delhi

Abstract

A lot of work is being done to get the final product out of immature products. These activities are carried out with the help of various machines. Mining is one of the most important industries. These operations are usually performed on two types of drilling machines, namely the Bench Drilling Machine and the other is a Hand Drill Machine. A bench piercing machine can be used for heavy tasks but to perform many tasks fails. It requires a lot of spinning machines to do the same. In the case of a hand drill, a great deal of effort from man is needed, and the machine survives. This world-class drilling rig is a combination of both a hand drill and a bench drill, which is designed to get the best results with reduced effort and time as well. This machine provides additional flexibility to travel in any direction due to its contact and construction. This machine can work on all sides provided by the operator. There are various connections and joints made in their construction that bring more flexibility to operation. A straight column is designed in such a way that it can rotate in a circular motion. The arm provides angular movement up and down and the latter provides precise stability and holds the engine and other joints to perform the function.

Introduction

The drilling machine is one of the most common and is one of the most common and useful machines employed in industry for producing, forming and finishing holes in a work piece. The unit essentially consists of:

1. The spindle which turns the tool (called drill) which can be advanced in the work piece either automatically or by hand
2. A work table which holds the workpiece rigidly in position.

Working principle: The Rotating edge of the drill exerts a large force on the work piece and the holes generated. The removal is by shearing and extrusion.

History : About 35,000 BCE, Homo sapiens discovered the benefits of the use of rotating tools. This would involve a sharp sickle twisted between the hands of a person to pierce a hole in something. This led to the piercing of the hand, a smooth stick that was sometimes attached to a sandbar, rubbed between the hands. This was used by many ancient civilizations around the world including Mayors. The earliest known artifacts, such as bone, ivory, shells, and anecdotes have been found, dating to the Upper Paleolithic period. Bow drill (strap-drill) is a first-of-its-kind machine drill, as it converts back and forth motion into a rotating motion, and can be traced back about 10,000 years ago. It was found that tying a rope to a stick, and then connecting the ends of the rope to the end of the rod (bow), allows the wearer to pierce quickly and effectively. Mainly used to create fires, bow-drills were also used in ancient wooden, stone, and dental works. Archaeologists have unearthed Neolithic tombs in Mehrah, Pakistan since the time of the Harappa, about 7,500-9,000 years ago, consisting of 9 adult teeth with 11 teeth already bound. There are hieroglyphs depicting Egyptian carpenters and bead makers in Thebes tomb using archery tools. The earliest evidence of such tools used in Egypt dates back to about 2500 BCE The use of bow-drills was widespread in Europe, Africa, Asia, and North America, in ancient times, and it is still used today. Over the years a small variety of bow and string piercings have been developed for the use of boring materials by means of building materials or fires.

PROBLEM DEFINITION

As we know, a drilling operation is a heart of many industries as well as it is also an important machine in domestic use. But if you want to have more number of holes on a same work piece, you have to use multi spindle drilling machine having a combination with radial drilling machine which become more costly and expensive. So it cannot be affordable for all. It also requires varying the positions of the work piece then it also affects the accuracy of the operation. If we are going to use a hand

drill machine: i.e. portable drilling machine, then we can obtain the flexibility in the operations but the stability during the operation and the working conditions there is a possibility of reducing the accuracy of the operation and the weight of the machine also affects the performance.

For such kinds of machines, the weight and vibrations also have to be considered. In portable drilling machines it is not easily possible to achieve same accuracy and efficiency of operation every time

Construction

The fig symbolizes the various parts of a universal digging machine.

This worldwide drilling machine consists of the following components:

1. Base Frame
2. Straight Arm
3. Robot Arm
4. Bearing
5. Links
6. Motor
7. Drill Chuck
8. Drill Tool
9. Battery
10. Connecting Cables
11. Switch.

Basic Framework A basic framework is nothing but a member used for the foundation and provides appropriate support for the meeting.

Vertical Arm: attached to a frame by means of a bearing. It rotates its axis to give direction to the other assembly to cover the working part of the circuit.

Robot Arm: connected with the help of a flexible join on a straight arm. This connects the car handle to the straight arm. Provides angular movement up and down the operation.

Bearing: provides support for a straight arm and allows it to rotate freely with its straight axis. Attached to the base frame

Linkage : this is used to connect or attach various parts of the system. These are flexible and can provide slide movement between links. These are used to make rotating parts.

Motor: is a feature that provides rotational movement on the chuck or tool handle to detect the output of the machine. Here we have used a 12V dc engine in this machine. Fig-8: Motor & Chuck

Drill Chuck: nothing but a tool handle. Hold the tool in the jaws. Attached to the motor to get the rotation movement and move it to the cutting tool.

Piercing tool: also known as drill bit. We will use bits that are limited in size but that are mostly used on this machine in this paper. We will use a bit of diameter 6mm, 7mm, 8mm & 10mm.

Battery: it is used to drive the motor to obtain desired output. Here we are going to use a rechargeable battery of 12V capacity.

Electric wires: wires are used to connect the battery to the motor.

Switch: it is used to control the ON-OFF action of the motor. When switch is at on position, the motor gets started & when switch is at OFF position motor get stopped

PRINCIPLE / VISION

This universal drilling rig simply works on the principle of converting electricity from a battery to a useful machine by rotating the chuck or spindle to gain cutting using the tool. It also refers to a sliding search engine that allows free rotation of links. This results in a higher degree of freedom for the machine or the arms of the robot. This will result in a higher operating capacity of the machine.

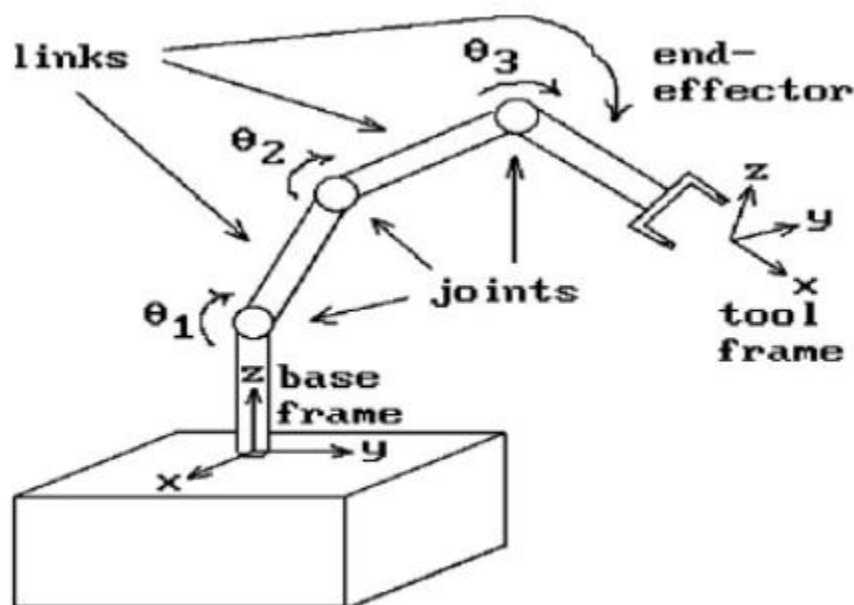


Fig - Universal Piercing Concept

Working

.As seen in fig. provide smooth pears for each member. This sliding palette offers more accessibility and a higher level of freedom. These members are brought together with the help of imprisoned members. As relaxation and strengthening of communication can easily happen. As the switch is operated by the operator and is NOT ON, it then connects the engine to the battery via power cords. As soon as the engine starts, it rotates the chuck or spindle mounted on the motor outlet. The chuck is used to hold the tool through the jaws. As the chuck or spindle starts, it leads to rotation of the tool. This tool rotation is also used to find the cut of a piece of work. Before starting the machine, it is necessary to make a decision and mark the size of the cut and the depth of the holes that should be drilled. Then position the machine so that you can cover the top points with their width. After determining the location where the foundation will be connected for the drill, the operator must start the engine. The operator must then provide directions to the marks. After completing tasks the operator should shut down the engine and slow down the engine. In this way the universal machine works with the least possible effort from the operator.

Advantage

Lightweight ,Portable, Easy to carry., Few efforts required.

CONCLUSION

This universal drilling machine provides better operational stability with respect to the portable drilling rig. It seems to be more profitable than conventional drilling rigs. This is a lightweight and portable too. So it provides better control during operation. The joints are made so that they can rotate everywhere and work better, so they work as we expected. It reduces the human effort required for mining operations and reduces the total energy consumption required to perform the same tasks. It also requires less space and is easier to manage.

References

- [1] A.M.TAKALE, V.R.NAIK, "Design & manufacturing of multi spindle drilling head (msdh) for its cycle time optimization", International Journal of Mechanical Engineering applications Research – IJMEAR, Vol 03, Issue 01; January-April 2012.
- [2] Manufacturing Processes by O.P. Khanna.
- [3] Mr. Jay M. Patel Mr. Akhil P. Nair Prof. Hiral U.Chauhan, "3-Directional Flexible Drilling Machine", IJSRD - International Journal for Scientific Research & Development| Vol. 3, Issue 01, 2015 | ISSN (online): 2321-0613, PP: 1262 to 1264.
- [4] A.S.Udgave, Prof.V.J.Khot, Design & development of multi spindle drilling head (msdh), IOSR Journal of Mechanical and Civil Engineering (IOSR-JMCE) ISSN: 2278-1684, PP: 60-69

A Switching NMOS Based Single Ended Sense Amplifier for High Density SRAM Applications

Bhawna Rawat

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India
bhawnarawat12@gmail.com

Poornima Mittal

Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India
poornimamittal@dtu.ac.in

The demand for single ended static random access memory is growing, driven by the decreasing technology node and increasing processing load. This mandates the need for a single ended sense amplifier to be used along with the memory. Consequently, a single ended latch based sense amplifier is proposed in this paper for high-speed, low power application. The sense amplifier is designed at 32 nm technology node and its functioning is analyzed at 1 V supply voltage, while the environment temperature is maintained at 27 °C. It is analyzed for its delay, temperature tolerance, variability tolerance, and area occupancy. The delay requirement of 0.2 ns for the proposed scheme is significantly lower in comparison to its other counter parts. While, its false read time is 0.3μs. In terms of power consumption, the proposed sensing topology is marginally higher than SPSS, but its leakage power is 1.4 times less than SPSS. The major advantage of the proposed SA is its reduced area footprint of 7.65 μm², which is 1.78 times better the best pre-existing topology in terms of area.

CCS CONCEPTS: Hardware ~Integrated circuits ~Semiconductor memory ~Static memory

Additional Keywords and Phrases: Single ended Sensing, Sense Amplifier, Low Variability, Temperature resilient, and Low Power Circuit

1 INTRODUCTION

The increasing demand for battery powered system-on-chip (SoC) applications has forced designers to lower the power requirement for a circuit. Especially, with the growing market for implantable bio-medical devices, the need to reduce power consumption has taken the center stage. Conventionally, for a SoC the major chunk of total power consumption is consumed by cache memory and its peripheral circuitry [Zhai et al. 2018]. The most orthodox method to lower power consumption and improve battery life for a SoC application is to lower its operational supply voltage (V_{DD}) [Rawat and Mittal 2022]. But, with the increasing processing load, cache memories tend to occupy more area and consume major portion of total power consumption thereby, limiting its performance. Whereas, designing a high density cache memory operational at lower V_{DD} , is limited by the following reasons. Firstly, designing a high density memory requires minimally sized transistors, which are highly vulnerable to process variations [Kim et al. 2011]. Secondly, as the V_{DD} decreases the current through transistor becomes exponentially dependent on threshold voltage (V_{TH}). Thereby, increasing the circuits' vulnerability to process variation [Rawat and Mittal 2021, Patel et al 2021]. Additionally, as cache memory is a large array based on static random access memory (SRAM) bit cell, it is unable to average out random variation effect due to multi-stage circuit

design [Jeong et al 2015]. Consequently, at lower technology node SRAM is directly exposed to the ill effects of random variation caused by device mismatch.

Conventionally, an SRAM bit cell is differential in nature, and its corresponding SA is also differential. This differential SA uses two bitlines to perform the sensing operation. This SA circuit is solely responsible for detection of a small differential signal on the bitline to yield a full swing signal at the output. The differential 6T SRAM cell was the de-facto for cache memory implementation. But, its performance was tremendously impacted by the decreasing technology node and lowering V_{DD} . Additionally, the inherent conflict between the read and write operation for the 6T cell resulted in an even drastic impact on its performance. Consequently, it can be inferred that achieving high stability for read and write operation simultaneously, along with high yield at low V_{DD} is difficult. To overcome the mentioned limitations and achieve the highlighted targets, various modified SRAM cell topologies were reported in literature [Sharma et al 2018, Cho et al 2020, Surana and Mekie 2019, Kumar and Ubhi 2019]. One mechanism adopted to improve the read and write stability simultaneously, is to isolate the read and write ports [Rawat and Mittal 2022]. But, this may result in increased area footprint for the cell. Another alternative is to decouple the read and write port for the cell. This is achieved with the help of additional transistors introduced within the cell topology [Aly, and Bayoumi 2007]. A common feature amongst the decoupled cell and dual port cell topologies is the utility of single bitline for read operation. The single ended read port for the cell enables improving its read stability, results in better yield, while keeping the cell area in check. Therefore, SRAM bit cells with single ended read operation are gaining popularity as they demonstrate improved performance at lower values of V_{DD} . A cache memory designed using single ended read port mandates designing a single ended sensing scheme. The functioning of the single ended SA is dependent on a single bit line. One of the most common sensing scheme for single ended cell is large-signal signal ended inverter [Fan et al 2012]. For a single ended SA the sensing margin is defined as the voltage difference between the LBL voltage levels and inverter trip voltage [Fan et al 2012].

In this paper, different pre-existing SA architectures are extensively studied to identify their weakness. Based on the learnings, a single ended voltage mode SA is designed for low bitline input, faster sensing, low power consumption, and greater reliability. The paper is organized into five sections, including this introduction section. In section II the different SA designs and their flaws are discussed. The proposed single-ended dynamic SA architecture is elaborated upon in section III. The performance of different SA topologies are compared and analyzed in section IV. In section V the findings of the brief are summarized.

2 PRE-EXISTING SINGLE ENDED SENSE AMPLIFIER TOPOLOGIES

Conventionally for single ended topologies, domino sensing scheme is a preferred choice [Chang et al. 2008, Ohbayashi et al. 2007, Warnock et al. 2012], but its performance deteriorates as the number of cells per bitline increases [Jeong et al. 2015]. Whereas, the pseudo differential sensing scheme is operational even when large number of bit cells are integrated on a single bitline [Chang et al. 2014]. But, its performance is limited by variability in reference voltage generation and strobe signal variations. The aforementioned problems can be resolved using techniques reported by researcher in [Qazi et al. 2011, Verma and Chandrakasan 2009]. But, these SA topologies increase the power consumption tremendously

and also cause unintentional couple between bitline and input of SA, thereby severely limiting the performance for the topology.

The differential SA circuit is solely responsible for detection of a small differential signal on the bitline, to yield a full swing signal at the output [Lai and Huang, 2008]. The performance of an SA is highly critical for an SRAM as it determines the minimum operating point, operational frequency, and power consumption for an SRAM based memory [Zhang et al. 2000]. Conventionally, the most important performance metric for an SA includes - sensing delay, minimum differential input voltage, and power consumption during the read operation [Houle 2007]. But, with the increasing popularity of single ended configuration of SRAM bit cell, it has become primal to investigate and design single ended SA for ease of integration. Another dimension of challenge for SA designers is to bridge the widening conflict between delay and power performance [Yang and Kim 2005, Shin et al. 2005], caused by the decreasing technology node and increasing demand for integration density. Another major concern in the nanometer domain is process variation; it severely effects the SA reliability [Dounavi et al. 2019]. It has been reported in literature by different researchers that aging in SA may result in speed degradation [Agbo et al. 2017], and off set voltage related concerns [Kraak et al. 2017, Kaark et al. 2017]. Amongst the popular conventional SA topologies, the voltage latch based SA offers a three times better offset tolerance. Thus, making it a better topology for the same area footprint [Patel and Sachdev 2018]. Different sensing schemes that have been reported to perform the read sensing operation for a single ended SRAM bit cell are as follows

2.1 Domino Sensing Scheme

Typically, a domino logic based circuit is composed of a static gate inserted between consecutive dynamic stages, which may be utilized for single ended bit-line sensing. A block diagram representation for domino based sensing scheme used for single ended read operation is depicted in Fig. 1(a). The local read bit-line (LBL) acts as the input to the domino logic. For a single-ended SRAM bit cell, the LBL selectively discharges (either for '1' or '0') turning ON the M2 transistor, and consequently charging node Z; which in turn is connected via an inverter to PMOS transistor M6. As Z is high, the M6 transistor is ON thereby charging the global bit-line (GBL). Thereby, completing the read sensing operation for the memory topology. In Fig. 1(a), one GBL is shared between two different sub banks, thus node Z may be changed by either of the two LBLs.

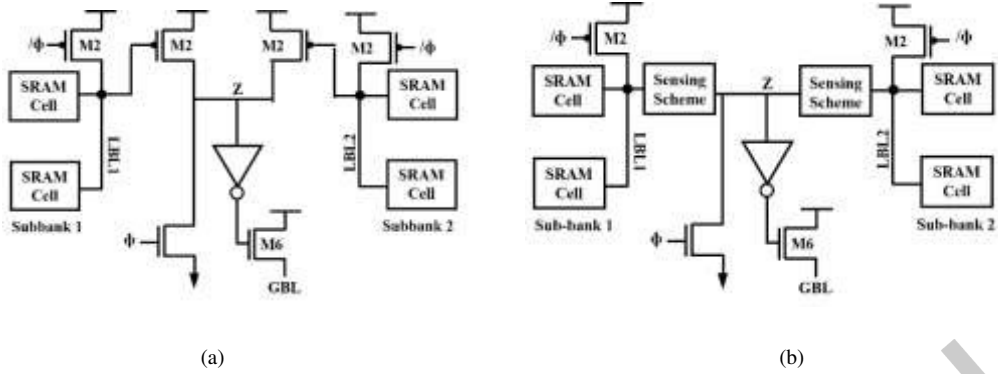


Fig. 1 Block diagram representation for (a) domino logic based sensing, and (b) modified sensing scheme

The extreme poor performance of dynamic PMOS is attributed to its high dependence on the bit line capacitance. If large number of cells are connected in the same column, the bit line capacitance increases drastically. This results in an extremely high delay, as the transition time from '1' to '0' is very high. Thus, making dynamic PMOS sensing scheme not a suitable alternative for large memories. This sensing scheme is simple and effective, only for memories with smaller bitline capacitance. As the size of the memory array increases, the performance for the domino sensing scheme deteriorates. The performance of the sensing scheme is dependent on the time consumed by the topology to detect the data stored in the cell. A large voltage swing is required on LBL, to charge node Z. If the bit-line capacitance is very high, the discharge time for the scheme increases, resulting in inferior performance and higher power consumption [Qazi et al. 2011]. This technique is best suited for smaller sub-banks. To improve the performance of the domino sensing scheme the block diagram for domino sensing scheme may be modified to replace the first PMOS transistor M2 with a dedicated sensing topology. The block diagram for the same is depicted in Fig. 1 (b). Two different pre-existing sensing topologies that may be utilized as the first stage for the modified domino sensing scheme are as follows.

2.2 AC Coupled Sensing Scheme

One of the topologies that has been reported for sensing scheme in Fig. 1(b) is AC coupled sensing scheme (ACSS), reported by [Qazi et al 2011]. The detailed circuit diagram for ACSS is depicted in Fig. 2(a). It is composed of a coupling capacitor, an inverter with a foot switch (M2), a PMOS transistor (M1) for equalization, and an output transistor M3. Additionally, an NMOS transistor (M4) is used to form the feedback connection between the GBL signal and the inverter input node X. The utility of the M2 transistor is that it enables the sensing scheme, only when a particular sub-bank is selected.

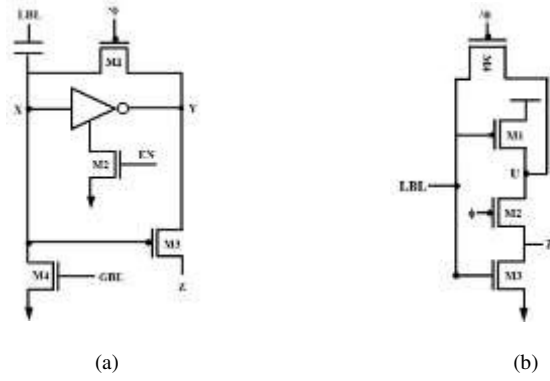


Fig. 2 Schematic diagram for (a) ACSS, and (b) SPSS topologies

The sensing operation for the ACSS topology is a two-step process – pre-charge and evaluation. During the pre-charge phase the enable signal ϕ is set high. Thus, the $\phi = '0'$ and the PMOS transistor M1 connects the input node (X) and output node (Y). Consequently, the voltage values at node X and Y are equalized. This equalization of the X and Y node biases the tri-gate inverter at the trip voltage; as its V_{in} and V_{out} are equal. Additionally, during the pre-charge phase the Z node is discharged to ground. Once the pre-charge step for sensing is completed the evaluation phase begins; the ϕ signal is set low. During this phase the data stored in the memory cell to be read is being sensed. During the read operation for an SRAM cell, the pre-charged LBL value is lowered due to read discharge current. This lowering of voltage level on LBL gets coupled with the X node; which was initially biased at the trip voltage (during the pre-charge phase). Consequently, a small decline in node X voltage, translates to a rapid rise in the voltage level at node Y; the voltage gain for an inverter is extremely high at the trip voltage [Rabaey et al 2002]. Lowering of value at node X also, turns ON the M3 PMOS transistor. Thus, the Y node is connected to the Z node via transistor M3 (which is in conducting state). The output waveform for the ACSS topology is depicted in Fig. 3(a).

As the ACSS topology utilizes the high gain of the inverter near the trip point. It is able to perform robustly even with a fairly smaller input swing, resulting in high sensing performance. It is also robust in terms of variability performance, as the bias condition tracks the trip point variation in the inverter. But, with the aforementioned merits there are limitations that have ill impacts on the performance of the ACSS. Firstly, biasing the inverter at the trip point results in a short circuit condition between the V_{DD} and the ground terminal. Thereby increasing its static power consumption. Second, for optimal functioning a large capacitor is required by the circuit which increases the area footprint for the circuit significantly [Chang et al. 2007].

2.3 Switching PMOS Sensing Scheme

Another circuit topology that was reported to be used as a replacement for the first stage dynamic PMOS SA in domino sensing scheme is the switching PMOS sensing scheme (SPSS). This scheme was reported by Jeong et al. 2016, the schematic diagram for the SPSS topology is depicted in Fig. 2(b). The circuit comprises of a pull up PMOS transistor (M1) and a pull down NMOS transistor (M3). Two additional PMOS

transistors connected in pass transistor configuration are added to the inverter; one between M1 and M3 and the other is connected between the drain of M1 and the input signal LBL.

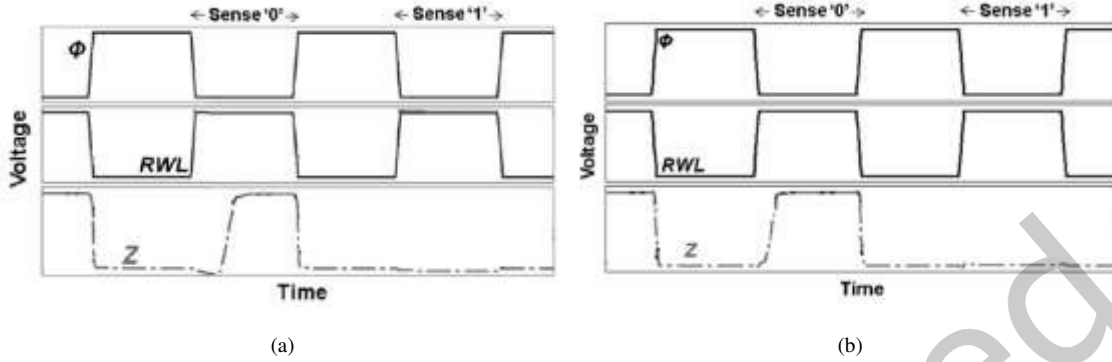


Fig. 3. Output waveform corresponding to (a) ACSS and (b) SPSS topologies [Jeong et al. 2015]

During the pre-charge phase ($\phi = '1'$), the M4 transistor is turned ON resulting in the M1 transistor in the diode connected topology. This results in LBL being pre-charged to $V_{DD} - V_{TH}$. In the evaluation phase ($\phi = '0'$), the M4 transistor is turned OFF, and the M2 transistor is turned ON. This completes the inverter circuit, with LBL as the input and the Z node as the output. For sensing '0' the discharge on LBL turns on transistor M4. The pre-charge level for LBL is $V_{DD} - V_{TH}$ only a small swing of LBL turns ON transistor M4, resulting in better performance. The output waveform corresponding to the SPSS topology is depicted in Fig. 3(b). The major limitation of the SPSS topology is its utility of large number of PMOS transistors in the circuits. The current capacity for a PMOS is lower in comparison to an equally sized NMOS transistor [Rabaey et al 2002]. Therefore, due to the dominance of PMOS transistors in the SPSS topology, its performance in terms of current carrying capacity, delay and area footprint suffer.

3 PROPOSED SENSE AMPLIFIER

3.1 Structure and Functioning of the proposed SA

In this paper a switching NMOS sensing scheme (SNSS) is proposed. This SNSS topology is designed as a modifications of the reported pre-existing sensing schemes explained in the previous section. This topology will be added as the sensing scheme in Fig. 1(b), which is a modification of domino sensing scheme, with the first dynamic stage replaced. The detailed structure for the proposed SNSS is depicted in Fig. 4(a). The circuit topology for SNSS consists of a pull up PMOS transistor (M1) and a pull down NMOS transistor (M3). Two NMOS transistors (M2 and M4) are added to the inverter topology between transistor M1 and M3. The utility of M2 is that, it is controlled by an additional control signal $/\phi$, which is activated only when a particular sub bank is selected.

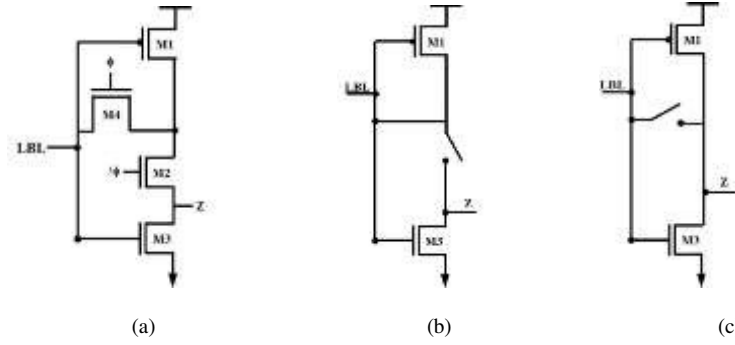


Fig. 4 Schematic diagram for (a) proposed sense amplifier design, equivalent-circuit configuration during (b) pre-charge phase, and (c) evaluation phase.

During the pre-charge phase ($\phi = '1'$), the LBL is charged to due to the pre-charge circuit in Fig. 1(b). This circuit eliminates the short circuit current condition between V_{DD} and ground, which gets created in the ACSS circuit. Whereas, in the evaluation phase ($\phi = '0'$), the M1-M3 inverter is in conducting state; its input being the bitline voltage and the output is sensed at node Z. The equivalent circuit topology during the pre-charge and evaluation phase is depicted in Fig. 4 (b) and (c) respectively. The output waveform corresponding to the functioning of the proposed SNSS topology is depicted in Fig. 5. The circuit is simulated for 1 V supply voltage and 27 °C environment temperature.

The performance of the proposed SNSS is tolerant to variations as the pre-charge voltage levels of LBL works in conjunction with the NMOS transistor (M2). The obtained output waveform for SNSS is depicted in Fig. 5(a). It can be inferred from the output waveform that SNSS performs the read sensing operation when '0' is stored in the SRAM cell. During the sense '0' operation, the read wordline is exerted and the bitline pre-charged voltage level is lowered due to read discharge current. During this phase the ϕ signal is high and the sensing inverter topology is biased at the trip voltage (explained in the previous section). Once the read wordline is exerted, and the bit line has discharge the ϕ signal is set low. Then the effective circuit for SNSS is depicted in Fig. 4(c). Now, when the previously pre-charged LBL experiences a decline in voltage level it is sufficient to turn ON transistor M1. As a consequence, the sensing performance for the proposed SA is increased. Whereas, if '1' is stored in the cell, no read discharge current is occurs. Thereby, no output is recorded by SNSS. The same may also be inferred from Fig. 5(a).

A small bit-line discharge can turn ON transistor M1, resulting in rise in the output level voltage. Thereby, improving its performance in comparison to the domino sensing scheme. Whereas, in comparison to SPSS, the proposed SNSS topology uses NMOS for switching, which improves its operational speed; NMOS device is faster in comparison to an equally sized PMOS transistor. Additionally, the SPSS topology is designed with stacked PMOS configuration for the pull up network, but this makes it highly skewed. Consequently, the PMOS width needs to be fairly large, so as to ensure fast operation. This also causes high levels of energy consumption for SPSS due to parasitic capacitance [Wong et al. 2015]. The local bit line (LBL) is the common bit line shared by a column; it acts as the input to the proposed sensing scheme. When the data stored in the memory cell to be read is '0', the discharge current flows through the read port of the cell and the LBL is discharged to '0'. This LBL is the input of the proposed

sensing scheme, thus for $LBL = '0'$ it charges Z to $'1'$, provided $\phi = '0'$. The Z node in turn, drives the inverter $M6$ labelled in Fig. 1(b) to yield the full swing output as the global bit line (GBL). The GBL is important, as it is final output for the sensing scheme. The LBLs are common amongst a memory column, but the GBL is unique for the entire cache. Therefore, the number of LBLs in cache memory is dependent on the array size and configuration, but there is only one GBL per cache memory.

SRAM bit cells are arranged in rows and columns to form the storage core for the cache memory. At the bottom of each column, the bit line is connected to the input of the SA. Thus, when a read operation is to be performed for a given cell, its respective column SA is enabled. For the proposed SNSS the same is applicable, one sensing scheme block is common for a given column. The proposed SA performs sensing operation only when the data stored in the bit cell is $'0'$. Whereas, if the data stored in the bit cell is $'1'$, no discharge current is registered. So, when a cell in a column has to perform read operation, the enable signal (ϕ) is set low. Additionally, the ϕ signal has to be operated with slight delay with respect to read wordline signal; this is done to ensure that the LBL has attained its desired value after the discharge current has passed through the circuit. Once the LBL value is set, then the ϕ signal for SA of that column is set to $'0'$. When the read operation is completed the ϕ signal is restored to $'1'$ and the SA is turned OFF. Thereby, lowering the power consumption for the proposed sense amplifier.

3.2 Delay Analysis for the proposed SA

The proposed SA is designed for single ended SRAM cell, which results in read discharge current for state $'0'$. Therefore, LBL is low only when the selected cell have state $'0'$ stored; for state $'1'$, no discharge event is registered. During the LBL discharge operation, the time required to raise the output node to 90% of V_{DD} after assertion of the read word line for the SRAM cell is defined as the sensing delay [Na et al. 2013]. The performance of the proposed SNSS topology at the different process corners is compared in Fig. 5(b). The proposed SNSS topology has the most inferior performance at the SS corner. This is because the performance of both the PMOS and NMOS transistor deteriorates at the SS corner. While, the best performance is observed at the FF corner due to the uplifted performance of the PMOS and NMOS transistors. The performance of the proposed SNSS topology at the TT corner is 0.2 ns, which is improved in comparison to the pre-existing topologies (explained in subsequent section).

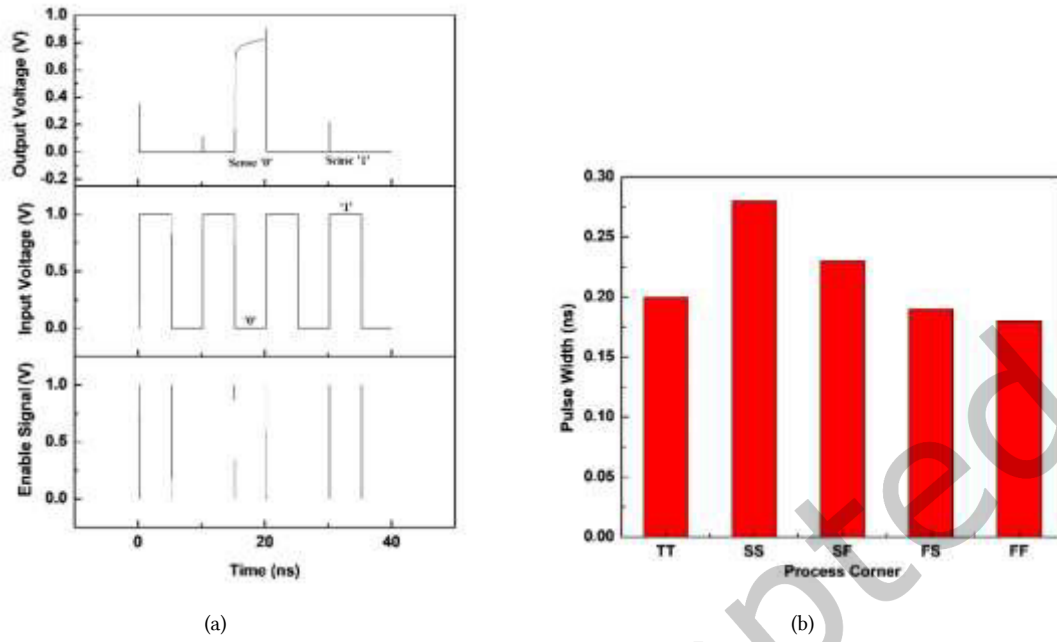


Fig. 5 (a) Output waveform corresponding to the proposed SNSS topology, and (b) Delay timing for the proposed sense amplifier topology at different process corner for $V_{DD} = 1$ V

3.3 Process Voltage and Temperature Tolerance of the proposed SA

In this sub-section the performance of the proposed SNSS is evaluated against process, voltage, and temperature variation. The analysis is performed to validate the reliability of the proposed technique when subjected to variations due to internal and external factors. For process variation analysis, statistical methods are employed to identify the impact of variation in transistor V_{TH} , caused due to process variations. Monte Carlo simulations are carried out for 10,000 data point, varying transistor V_{TH} in 6σ range around the mean V_{TH} value. The output waveform obtained for SNSS using Monte Carlo simulation is depicted in Fig. 6 (a). It can be inferred from the Monte Carlo simulation output for SNSS that process variation result in a minor variation in the performance of the bit cell and the reliability of the waveform is maintained. It is also observed that as the variation in the value of V_{TH} increases, the slope gradient of the transient analysis curve decreases. Thereby, resulting in a very insignificant increase in time required to attain the maximum output level for the SA. The same may also be inferred from Fig. 6 (a).

In any digital circuit, the voltage available to the circuit may vary within defined range; each discrete level is defined for a voltage range and not for a fixed value. Thus, when designing a digital circuit it is essential to ensure that it performs reliably when the operational V_{DD} for the circuit is varied slightly. The impact of voltage variation ($\pm 10\%$ of the operational V_{DD}) on the output waveform of the proposed SNSS topology is depicted in Fig 6(b). It can be inferred from Fig. 6(b) that the output waveform for SNSS, does observe a deviation in its performance. But, the range of variation in its performance is within a manageable limit and will not have any drastic impact on the overall performance of the cache memory.

The proposed sensing amplifier topology is designed at 32 nm node; in nanometer vicinity the reliability of a circuit is of primal concern. A circuit design is expected to demonstrate resilience to temperature variation. A system may register variation in temperature due to internal and external factors. Correspondingly, the performance of the proposed SNSS is evaluated for temperature variation from -10°C to 110°C . A section of the transient waveform for the SNSS is depicted in Fig. 6(c). It can be observed from Fig. 6(c), the output waveform for SNSS does register a shift in its performance. But the variation in performance of the SNSS is within manageable bounds and will not have a drastic impact on the overall performance of the circuit.

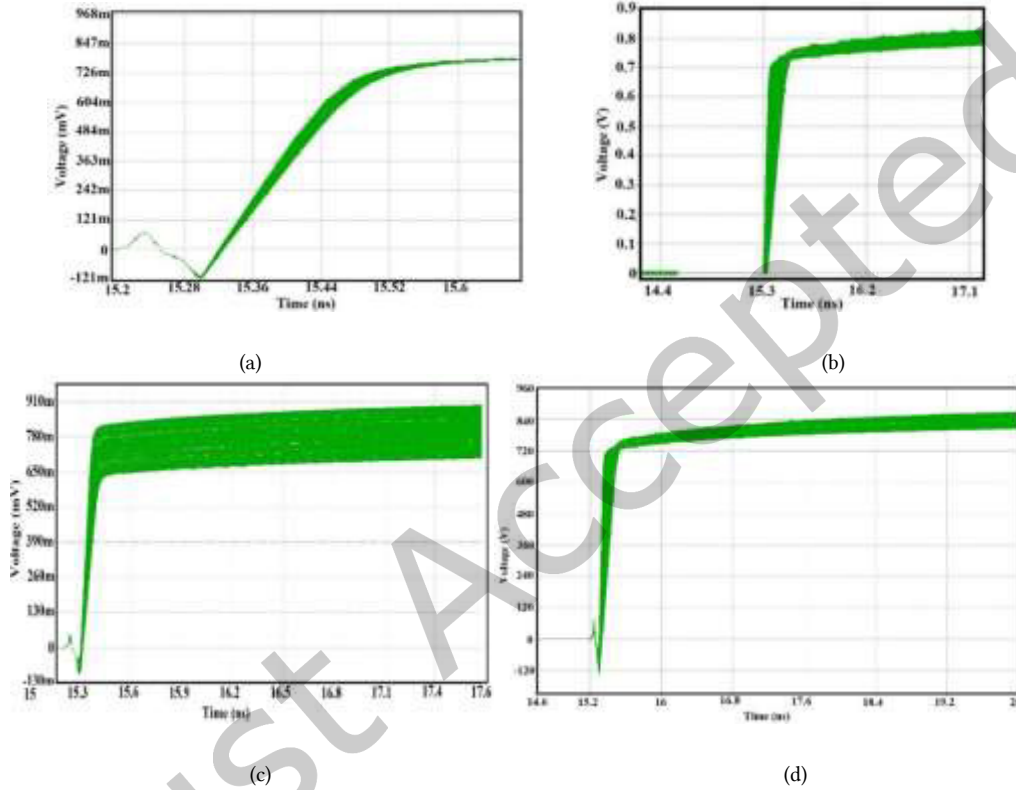


Fig. 6 Variation in the output waveform for the proposed SA for sensing 0 for (a) process variation, (b) Voltage variation between 0.9 to 1.1 V, (c) temperature variation from -10°C to 110°C , and (d) temperature variation from 0°C to 70°C

It was identified that the worst case output voltage for the proposed SNSS topology is $\sim 0.75\text{V}$ the same can be inferred from Fig. 5(a). While, the impact of process, voltage, and temperature variation on performance of the proposed SNSS topology is depicted in Fig. 6 (a), (b), and (c) respectively. The maximum variation in output waveform of SNSS topology due to process variation is $\sim 0.1\text{V}$ as can be inferred from Fig. 6(a). While, voltage variation results in altering the output waveform by $\sim 0.05\text{V}$ (as depicted in Fig. 6(b)). The maximum variation of $\sim 0.25\text{V}$ in performance of SNSS topology is caused by temperature variation, but here the maximum output voltage obtained is $\sim 0.9\text{V}$. Therefore, the minimum voltage that the output waveform may register is $\sim 0.66\text{V}$, which is within reliable voltage limits. Also, the

temperature range taken into consideration is -10°C to 110°C . This is an extremely wide range for evaluation of analysis, whereas most commercial electronic devices in 0°C to 70°C . Thus, when the cell was evaluated for this range, the variation in the performance of SNSS is considerably reduced. The same may be inferred from Fig. 6(d). It can be inferred from Fig. 6(d) that the maximum variation in caused due to temperature variation is $\sim 0.07\text{V}$. Thus, the proposed SA topology; SNSS may be deemed resilient to process, voltage, and temperature variations.

4 COMPARISON

For validation of the proposed single ended sense amplifier topology depicted in Fig. 4(a), it is designed using 32 nm technology node and is simulated for 1 V of V_{DD} . The models used for designing the circuit topology is based on the Predictive Technology Model.

4.1 Sensing Performance

The most essential aspect for an SA topology is its timing requirement. During the read operation for single ended SRAM bit cell, the time required for flipping the output of the SA after the ϕ signal has been set is referred to as the read time (T_s) for the single ended SA topology [Jeong et al 2015]. The proposed cell improves upon the delay performance of the existing SA is by using more NMOS transistor and aptly sizing them to have sufficiently large drive current while maintaining its area occupancy.

The sensing performance for the different SA topologies (explained in section 2) along with the proposed SA are graphically compared in Fig. 7(a). The sensing performance for each SA is determined at each process corner to analyze the impact of global variation on the performance of the topology. The simulations are performed at 1V supply voltage. For all the SAs, the best performance is observed at FF corner, owing to the better performance for both NMOS and PMOS. The dynamic PMOS based SA topology has the most inferior performance in comparison to others. The proposed SNSS has improved performance in comparison to the other topologies. In comparison to the dynamic PMOS, ACSS, and SPSS, the SNSS has improved performance by 5.25, 0.5, and 0.25 times respectively. The best performing pre-existing SA topology in terms of delay is SPSS with delay of 0.4 ns. Whereas, the proposed SNSS topology registers a delay of 0.2ns. Thus, the proposed SA topology may be deemed to have improved delay performance in terms of delay parameter.

The read operation for a single ended SRAM bit cell can be divided into stages. Firstly, the read wordline signal is asserted. Then, if the data content of the cell is '0', a read discharge current is registered, and the voltage at LBL is lowered from '1'. Otherwise, no discharge current is registered and LBL maintains its pre-charge value. Then, after a certain amount of time the ϕ signal is set low to enable the sensing scheme. If the LBL at this instance is '0', the designated cell is determined to have '0' stored in it. Otherwise, '1' is stored in the cell. Ideally, when '1' is stored in the cell, no current should flow in the circuit. But, OFF current due to state '1' stored in cell may cause an unintentional discharge of LBL. The time required for this unintentional discharge to falsely flip the output of the SA is referred to as false read time. Ideally, for reliable sensing operation, the value obtained for false read time should be significantly larger than the sensing delay for the topology. The false read time values obtained for the different sensing schemes are graphically compared in Fig. 7(b). It may be observed from Fig. 7(b) that at TT corner, the

dynamic PMOS SA topology has the highest false read time at 0.48 μ s. While, the performance of the proposed SNSS topologies is 0.3 μ s. The performance of ACSS and SPSS is inferior amongst the four SA topologies, as their false read time is lower in comparison to the other SA topologies.

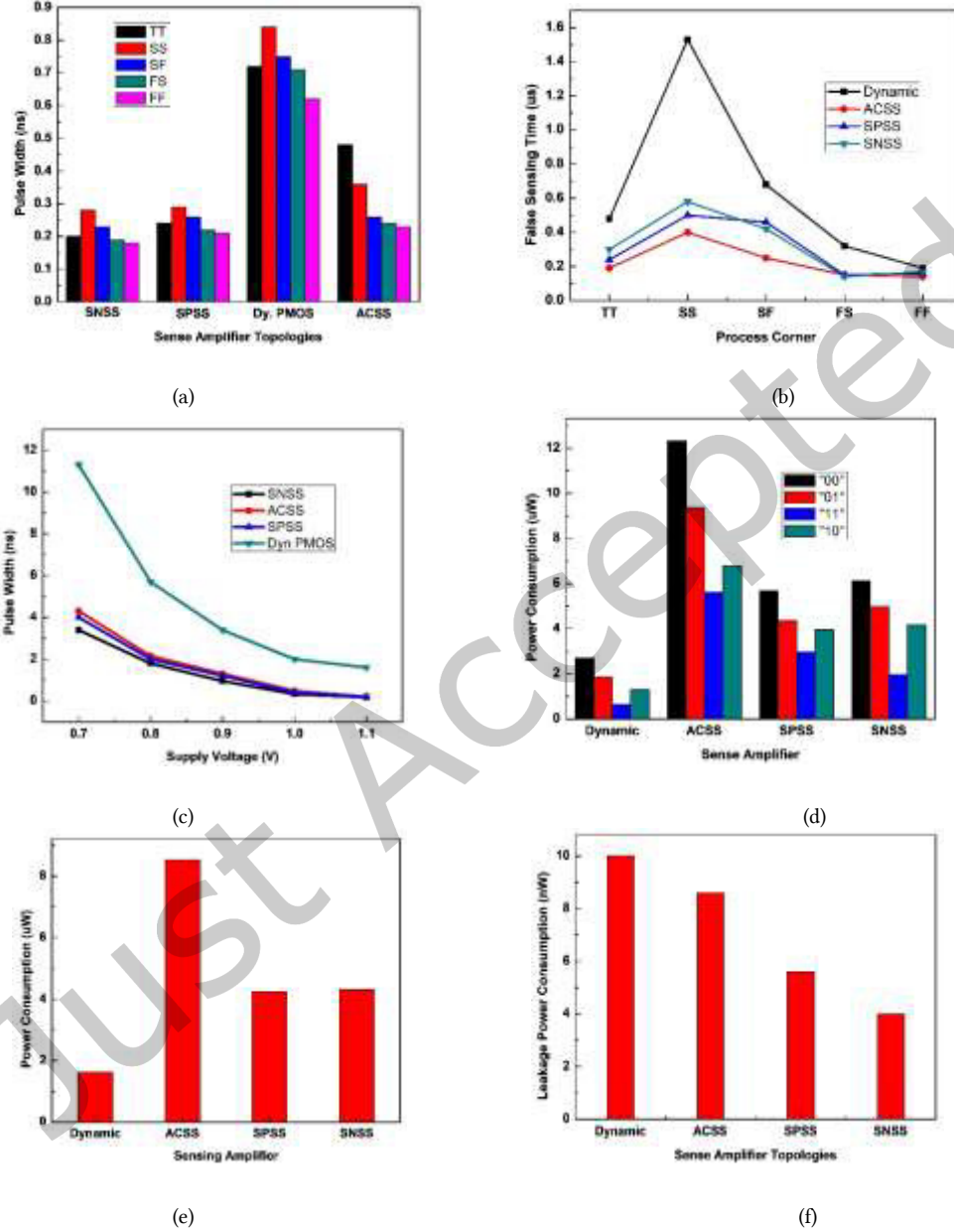


Fig. 7 Comparison of (a) delay timings at different process corners, (b) false read time at all process corners, (c) delay for varying V_{DD} , (d) power consumption corresponding to four data cases – 00, 01, 10, and 11, (e) average power consumption, and (f) leakage power consumption for all the SA topologies.

The SNSS topology has superior sensing performance even when evaluated for different V_{DD} values; the same can also be inferred from Fig. 7(c). The performance of all the SA topologies at different V_{DD} values is compared in Fig. 7(c). The highest delay is utilized by the dynamic PMOS topology. While, the remaining single ended SA techniques have comparable performance. But, amongst the ACSS, SPSS, and SNSS topology the least delay requirement for output generation is registered by SNSS.

The performance of the proposed SNSS topology is comparable against performance of SPSS and ACSS pre-existing topologies. Amongst pre-existing sensing scheme, the SPSS topology has the least delay performance. The SNSS topology improves upon the delay performance of the SPSS topology. This is achieved by replacing the stacking PMOS transistor (M2) in SPSS topology by an NMOS transistor (M2) in the SNSS topology. The major objective of changing the nature of M2 transistor is to achieve the faster operation using a smaller sized transistor; an NMOS has faster operation than an equally sized PMOS. Additionally, the SPSS topology has two PMOS transistors M1 and M2 stacked one over another. This configuration increases the delay of the circuit, while poorly impacting the area footprint for the SA (more number of PMOS implies larger n-wells in the layout). Thus, for the proposed SA, only one PMOS transistor; M1 is used and its sizing is also optimized to reduce area while achieving better delay performance.

4.2 Power Consumption

Power consumed by a single ended sensing scheme is dependent on the current requirement for the sensed state and the previous state. For instance, power is consumed during the pre-discharging of the current state only if '0' was sensed during the previous time cycle. Similarly, for the current time, period power is consumed only when the state being sensed is '0', for sensing '1' no power is consumed by the circuit. Therefore, in keeping with the condition of the present state ('0' or '1') and the previous state ('1' or '0'), four distant cases can be identified for power consumption calculation. Thus, for each SA topology power calculations are done corresponding to 00, 01, 10, and 11 cases. The two bit data pattern denotes the data state sensed by a given SA topology during the previous time cycle and the current time cycle. Consequently, "10" data state implies that the SA topology sensed data '1' during the previous time cycle and during the current time cycle it is sensing state '0'. The power consumption values obtained for the different SA topologies for the four different data cases are graphically compared in Fig. 7(d). Most power consumption for any of the single ended topology discussed in the paper is reported for "00" bit sequence. This is because for it two consecutive sensing operations. Whereas, the least power consumption for all topologies is observed for data sequence "11", as no sensing operation is performed (the single ended sensing topology only sense '0' state).

Additional to the power consumption for four different data cases, average power consumption for each SA topology is also calculated. The average power consumption for each SA is presented in Fig. 7(e). The average power consumption for the ACSS topology is the highest amongst others. Its average power consumption is 34.27, 43.06, and 40.72 % greater than dynamic PMOS, SPSS, and SNSS respectively. The higher power consumption of the ACSS is an implication of biasing the first-stage inverter to an intermediate voltage during pre-charge mode, resulting in a short circuit current to flow through the circuit. Whereas, the SPSS and SNSS topologies are able to evade this problem with the help of a switching

PMOS and NMOS transistor respectively. Therefore the average power consumption for the two topologies are drastically lower than the ACSS topology.

When the read operation is not being performed, the SA is disabled using signal ϕ . During this disabled stated the power consumption by the SA topology is referred to as its leakage power consumption. It is calculated as the product of leakage current in the circuit and V_{DD} . The leakage power consumption values obtained for the different SA topologies are graphically compared in Fig. 7(f). The leakage power is highest for the dynamic PMOS topology and it is 2.5 times the value for the SNSS topology. Amongst the pre-existing SA topologies the best performance is demonstrated by SPSS topology; its leakage power consumption is 5.6 nW. The performance of the SPSS topology is 1.4 times more than the SNSS topology. The predominant use of largely sized PMOS in its design is the culprit for the same; larger PMOS are necessary for higher drive current and smaller delay of the circuit.

4.3 Area

The different single ended SA topologies explained in section 2 and the proposed SNSS use a static inverter to drive a dynamic PMOS to eventually develop the correct value on GBL. The difference lies in which sensing technique is used. Consequently, the area footprint for the sensing scheme is also dependent on the SA topology employed. Layout for each sensing scheme is designed for area estimation of each SA. The area for the domino sensing with a dynamic PMOS SA is $5.314 \mu\text{m}^2$. The area for the ACSS, SPSS, and SNSS topologies are 16.3, 9.43, and $7.65 \mu\text{m}^2$ respectively. The area for the ACSS, SPSS, and SNSS is multi-fold larger than the dynamic PMOS SA, this is because the transistor count for each is fairly larger. The performance improvement of the proposed SNSS outweighs its larger area overhead in comparison to the dynamic PMOS sensing scheme.

Also, the area footprint of ACSS and SPSS is 8.65, and 1.78 times larger than the area requirement of the proposed SNSS technique. The SPSS technique has larger area as for the same performance larger PMOS are required in comparison to an NMOS transistor. Therefore, increasing the area footprint for SPSS. Also, more the number of PMOS transistor greater number of n-well are to be created in the layout design, which also increase the area footprint for the cell. Whereas, for the ACSS technique the large area is caused by the use of a large coupling capacitor and additional static inverter (which increases its transistor count) used within its circuit topology for sensing operation. Thus, the area footprint for the ACSS is largest amongst all the single ended SA techniques discussed in this paper.

5 CONCLUSION

In this paper a single ended switching NMOS based sense amplifier topology is proposed. The sense amplifier relies on the same transistor for its pre-charge and sensing operation. The performance of the proposed sensing topology is compared with pre-existing dynamic PMOS, ACSS, and SPSS topologies. The performance of the proposed sense amplifier is improved in comparison to the dynamic PMOS in terms of delay and power. The delay requirement of 0.2 ns for the proposed scheme is significantly lower in comparison to its other counter parts. Whereas, in terms of power also the proposed sensing topology performance reliably. The low power consumption of the proposed SA is because of its ability to evade short circuit condition. Additionally, the leakage power for SNSS is also found to be least amongst the

different SA topologies at 4 nW. The additional advantage the proposed SA has its lower area footprint of $7.65\mu\text{m}^2$. The SPSS circuit relied on multiple PMOS transistors in its circuit, which required a large n-well to be created. Thereby, increasing the area of the SPSS topology. On the contrary, the proposed SNSS relies on a single PMOS for its design, thereby reducing its area footprint, and increasing its integration density and economic feasibility.

REFERENCES

- J. Zhai, C. Yan, S.G. Wang, D. Zhou, H. Zhou, and X. Zeng. 2018. An Efficient Non-Gaussian Sampling Method for High Sigma SRAM Yield Analysis. *ACM Transactions on Design Automation of Electronic Systems*, vol. 23, no. 3, pp. 1-23, Article 36.
- B. Rawat, and P. Mittal. 2022. A Reliable and Temperature Variation Tolerant 7T SRAM Cell with Single Bitline Configuration for Low Voltage Application. *Circuits, Systems and Signal Processing*, vol. 41, pp. 2779-2801.
- D. Kim, V. Chandra, R. Aitken, D. Blaauw, and D. Sylvester. 2011. Variation-aware static and dynamic writability analysis for voltage-scaled bit-interleaved 8-T SRAMs. *IEEE/ACM International Symposium on Low Power Electronics and Design*, pp. 145-150.
- B. Rawat, and P. Mittal. 2021. A 32 nm single ended single port 7T SRAM for low power utilization. *Semiconductor Science and Technology*. vol. 36, no. 9, pp. 095006-095022.
- D. Patel, A. Neale, D. Wright, M. Sachdev. 2021. Body Biased Sense Amplifier With Auto-Offset Mitigation for Low Voltage SRAMs. *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 8, pp. 3265-3278.
- H. Jeong, T. Kim, K. Kang, T. Song, G. Kim, H.S. Won, and S.O. Jung. 2015. Switching pMOS Sense Amplifier for High Density Low Voltage Single - Ended SRAM. *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 6, 1555-1563.
- V. Sharma, S. Vishvakarma, S.S. Chouhan and K. Halonen. 2018. A write improved low power 12T SRAM cell for wearable wireless sensor nodes. *International Journal of Circuit Theory Application*. vol. 46, no. 12, pp. 2314-2333.
- K. Cho, J. Park, T.W. Oh and O.K. Jung. 2020. One sided Schmitt-Trigger Based 9T SRAM cell for near threshold operation. *IEEE Transactions on Circuits and Systems I: Regular Papers*. vol. 67, no. 5, pp. 1551-1561.
- N. Surana and J. Mekie. 2019. Energy Efficient Single-Ended 6-T SRAM for Multimedia Applications. *IEEE Transactions on Circuits and Systems II: Express Briefs*. vol. 66, no. 6, pp.1023-1027.
- M. Kumar and J.S. Ubhi. 2019. Design and analysis of CNTFET based 10T SRAM for high performance at nanoscale. *International Journal of Circuit Theory Applications*. vol. 47, no. 11, pp. 1775-1785.
- B. Rawat, P. Mittal. 2022. A comprehensive analysis of different 7T SRAM topologies to design a 1R1W bit interleaving enabled and half select free cell for 32 nm technology node. *Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences*, vol. 478, no. 2259.
- R.E. Aly, M.A. Bayoumi. 2007. Low-Power Cache Design Using 7T SRAM Cell. *IEEE Transactions on Circuits and Systems II: Express Briefs*. vol 54, no. 4, pp. 318-322.
- M.L. Fan, V.P.H. Hu, Y.N. Chen, P. Su, C.T. Chuang. 2012. Variability Analysis of Sense Amplifier for FinFET Subthreshold SRAM Applications. *IEEE Transactions on Circuits and Systems-II: Express Briefs*, vol. 59, no. 12, pp. 878-882
- L. Chang, R. K. Montoye, Y. Nakamura, K. A. Batson, R. J. Eickemeyer, R. H. Dennard, W. Haensch, and D. Jamsek. 2008. An 8 T-SRAM for variability tolerance and low-voltage operation in high-performance caches. *IEEE Journal on Solid-State Circuits*, vol. 43, no. 4, pp. 956-963.
- S. Ohbayashi, M. Yabuuchi, K. Nii, Y. Tsukamoto, S. Imaoka, Y. Oda, T. Yoshihara, M. Igarashi, M. Takeuchi, H. Kawashima, Y. Yamaguchi, K. Tsukamoto, M. Inuishi, H. Makino, K. Ishibashi, and H. Shinohara. 2007. A 65-nm SOC embedded 6 T-SRAM designed for manufacturability with read and write operation stabilizing circuits. *IEEE Journal on Solid-State Circuits*, vol. 42, no. 4, pp. 820-829.
- J. D. Warnock, Y.-H. Chan, S. M. Carey, H. Wen, P. J. Meaney, G. Gerwig, H. H. Smith, Y. H. Chan, J. Davis, P. Bunce, A. Pelella, D. Rodko, P. Patel, T. Strach, D. Malone, F. Malgioglio, J. Neves, D. L. Rude, and W. V. Huott. 2012. Circuit and physical design implementation of the microprocessor chip for the Enterprise system. *IEEE Journal on Solid-State Circuits*, vol. 47, pp. 151-163.
- H. Jeong, T. Kim, K. Kang, T. Song, G. Kim, H.S. Won, and S.O. Jung. 2015. Switching pMOS Sense Amplifier for High Density Low Voltage Single - Ended SRAM. *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 62, no. 6, 1555-1563.
- M.-F. Chang, J.-J. Wu, T. F. Chien, Y.-C. Liu, T.-C. Yang, W.-C. Shen, Y.-C. King, C.-J. Lin, K.-F. Lin, Y.-D. Chih, S. Natarajan, and J. Chang. 2014. Embedded 1Mb ReRAM in 28 nm CMOS with 0.27-to-1 V read using swing-sample-and-couple sense amplifier and self-boost-write-termination scheme. 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), pp. 332-333
- M. Qazi, K. Stawiasz, L. Chang, and A. P. Chandrakasan, 2011. A 512 kb 8T SRAM macro operating down to 0.57 V with an AC-coupled sense amplifier and embedded data-retention-voltage sensor in 45 nm SOI CMOS. *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 85-96.
- N. Verma and A. P. Chandrakasan. 2009. A high-density 45 nm SRAM using small-signal non-strobed regenerative sensing. *IEEE Journal on Solid-State Circuits*, vol. 44, pp. 163-173.
- Y. C. Lai and S.Y. Huang,. 2008. A resilient and Power-Efficient Automatic-Power Down Sense Amplifier for SRAM Design," *IEEE Transactions on circuits and Systems- II: Express briefs*, vol. 55, no. 10, pp. 1031-1035.
- K. Zhang, K. Hose, V. De, and B. Senyk. 2000. The scaling of data sensing schemes for high speed cache design in sub-0.18 μm technologies. *Symposium on VLSI Circuits Digest of Technical Papers*, Honolulu, HI, USA, pp. 226-227.

- R. Houle. 2007. Simple statistical analysis techniques to determine minimum sense amp set times. *Proceedings of IEEE Custom Integrated Circuits Conference*, San Jose, CA, USA, pp. 37–40.
- B. D. Yang and L. S. Kim. 2005. A low-power SRAM using hierarchical bit line and local sense amplifier. *IEEE J. Solid-State Circuits*, vol. 40, no. 6, pp. 1366–1376.
- J. L. Shin, B. Petrick, M. Singh, and A. Leon. 2005. Design and implementation of an embedded 512-KB level-2 cache subsystem. *IEEE Journal on Solid-State Circuit*, vol. 40, no. 9, pp. 1815–1820.
- H. M. Dounavi, Y. Sfikas, and Y. Tsiatouhas. 2019. Periodic Monitoring of BTI Induced Aging in SRAM Sense Amplifiers. *IEEE Transactions on Device and Materials Reliability*, vol. 19, no. 1, 64–72.
- I. Agbo, M. Taouil, D. Kraak, S. Hamdioui, H. Kukner, P. Weckx, P. Raghavan, and F. Catthoor. 2017. Integral impact of BTI, PVT variation, and workload on SRAM sense amplifier. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 4, pp. 1444–1454.
- D. Kraak, M. Taouil, I. Agbo, S. Hamdioui, P. Weckx, S. Cosemans, and F. Catthoor. 2017. Impact and mitigation of sense amplifier aging degradation using realistic workloads. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 25, no. 12, pp. 3464–3472.
- D. Kraak, I. Agbo, M. Taouil, S. Hamdioui, P. Weckx, S. Cosemans, F. Catthoor, W. Dehaene. 2017. Mitigation of sense amplifier degradation using input switching. *Proceedings of Design Automation & Test Europe Conference (DATE)*, pp. 858–863.
- D. Patel, and M. Sachdev. 2018. 0.23 V Sample Boost Latch Based Offset Tolerant Sense Amplifier,” *IEEE Solid State Circuits Letters*, vol. 1, no. 1–9.
- Rabaey J, Chandrakasan A and Nikolic B, *Digital Integrated Circuits: A Design Prespective*, 2nd ed. Englewood Cliffs, NJ: Prentice Hall. 2002:308–309.
- L. Chang, Y. Nakamura, R. Montoye, J. Sawada, A. Martin, K. Kinoshita, F. Gebara, K. Agarwal, D. Acharyya, W. Haensch, K. Hosokawa, and D. Jamsek. 2007. A 5.3 GHz 8 T-SRAM with operation down to 0.41 V in 65 nm CMOS. *IEEE Symp. VLSI Circuits Dig.* pp. 252–253.
- O.Y. Wong, H.Wong, W.S. Tam, and C.W. Kok. 2015. Parasitic capacitance effect on the performance of two-phase switched-capacitor DC-DC converters,” *IET Power Electronics*, vol. 8, no. 7, 1195–1208.
- T. Na, S.H. Woo, J. Kim, H. Jeong, and S.O. Jung. 2013. Comparative Study of various Latch-Type Sense Amplifiers. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 2, 425–429.



A₂B corroles: fluorescent signalling system for Hg²⁺ ion

ATUL VARSHNEY and ANIL KUMAR*

Department of Applied Chemistry, Delhi Technological University, Delhi 42, India

E-mail: anil_kumar@dce.ac.in; atulvarshney711992@gmail.com

MS received 2 May 2022; revised 30 October 2022; accepted 31 October 2022

Abstract. A series of four A₂B corroles (where A = *para*-nitrophenyl, and B = 2,3,4,5,6-pentafluorophenyl, 2,6-difluoro, 2,6-dichloro and 2,6-dibromophenyl group) have been synthesized and characterized. These four corroles were tested for the sensing ability towards Hg²⁺ ion. The LOD for these corroles are comparable to reported sensors for Hg²⁺ ions. All these four A₂B corroles exhibit different fluorescence quenching due to the electronic effect of the phenyl group at C₁₀ position, which has a different halogen atom at 2,6 position of the phenyl ring.

Keywords. A₂B corrole; Hg²⁺ ion; Sensor; LOD.

1. Introduction

Chemosensors (fluorescent molecular probes) have many applications in various fields, such as environmental science,¹ medicine,² aeronautics,³ national security.⁴ Signal transduction in the chemosensor could be constructed by the physical or chemical process for the recognition of an analyte.⁵ Also, light emission is used to detect the analyte at the single molecule.⁶ A classical chemosensor could be designed by a fluorophore, a receptor, and a spacer for analytical sensing.⁷ A good Chemosensor must have photostability, high affinity, and selectivity towards the analyte.⁸ In this respect, chemosensors developed for the analytes, such as metal ions, which were synthesized to contain polyamines and oxygen or sulphur donor atoms.⁹ In recent years, chemosensors have been employed to detect and quantify pollutant metal ions in clinical toxicology, waste management, and environmental chemistry.¹⁰

In this direction, corroles were explored as chemosensors because of their significant photophysical and chemical properties.^{11,12} Corroles are trianionic, tetrapyrrolic molecules that stabilize the metal ions in lower and higher oxidation states.¹³ One of the most important features of corrole, it emits and absorbs the light in the visible region with high fluorescent quantum yield and good photostability.^{14,15}

Due to their photostability, corroles are used as photocatalysts for endergonic reactions.¹⁶ Corroles exhibit phenomena such as phosphorescence,¹⁷⁻²⁰ fluorescence,^{21,22} photosensitizer,^{17,23-25} energy relaxation,^{26,27} and generation of singlet oxygen.^{14,28} Inner nitrogen atoms of corrole behave as donor atoms to recognize the metal atom in lower and higher oxidation state.^{14,15,29} Metal atoms have a significant role in environmental and biological fields.^{9,31-37} Very few reports are available in the literature about the corrole used as a metal ion sensor.³⁸⁻⁴³

Corrole have similar properties as porphyrin and is also used as the sensing material. Plaschke and co-workers 1995 demonstrated the use of 5,10,15,20-tetra(p-sulfonatophenyl) porphyrin-doped sol-gel films for fluorimetric determination of mercury ions.^{44a} Later, Chan and co-workers developed the mercury ion-selective optical sensor using 5,10,15,20-tetraphenylporphyrin.^{44b} The porphyrin dimer-based optical fiber chemical sensor for mercury ions^{45c} was also shown by Zhang and co-workers. In the same year, 2006, the free base corrole in a PVC matrix-based fluorescent chemical sensor for Hg²⁺ ion³⁸ was also developed. In addition, the selective detection of Hg²⁺ ion using cationic triazatetrazabenzcorrole by nucleic acid-induced aggregation³⁹ was shown by Zhou and co-workers. Bandyopadhyay's group

*For correspondence

Supplementary Information: The online version contains supplementary material available at <https://doi.org/10.1007/s12039-022-02114-5>.

synthesized the A₂B corroles and was demonstrated as fluorescence signaling system for sensing Hg²⁺ ion.⁴⁰ Santos and co-workers reported the 5,10,15-tris(pentafluorophenyl)corrole as Hg²⁺ ion sensor, which was observed by the naked eye through a change of color from purple to blue in acetonitrile and from green to yellow in toluene.⁴¹ Hg²⁺ metal ion binds with the free nitrogen atoms in the inner of the corrole due to ICT (intramolecular charge transfer) transitions as compared to other metal ions and is responsible for chelation enhancement of quenching (CHEQ) in the emission intensity.⁴⁵ Our group already explored the A₂B corroles as fluoride ion sensor.^{46b} In this article, we have synthesized the four A₂B corroles with different phenyl group containing halogen atoms at C₁₀ meso position of the corrole, 10-(2,3,4,5,6-pentafluorophenyl)-5, 15-bis(4-nitrophenyl)corrole (**1**), 10-(2, 6-Difluorophenyl)-5, 15-bis(4-nitrophenyl)corrole (**2**), 10-(2,6-dichlorophenyl)-5, 15-bis(4-nitrophenyl)corrole (**3**), and 10-(2,6-dibromophenyl)-5, 15-bis(4-nitrophenyl)corrole (**4**) (Scheme 1). Corrole **1**, **2** have already been reported in our recent publication,^{46b} and corrole **3**, **4** newly synthesized and characterized by UV- visible spectroscopy, NMR spectroscopy, and mass spectroscopy. These corroles behave as sensors towards Hg²⁺ in the order **4** > **3** > **1** > **2**. Also, it is obvious to evaluate the halogen atom effect on the sensing ability of A₂B Bis p-nitro corrole towards Hg²⁺ ion.

2. Experimental

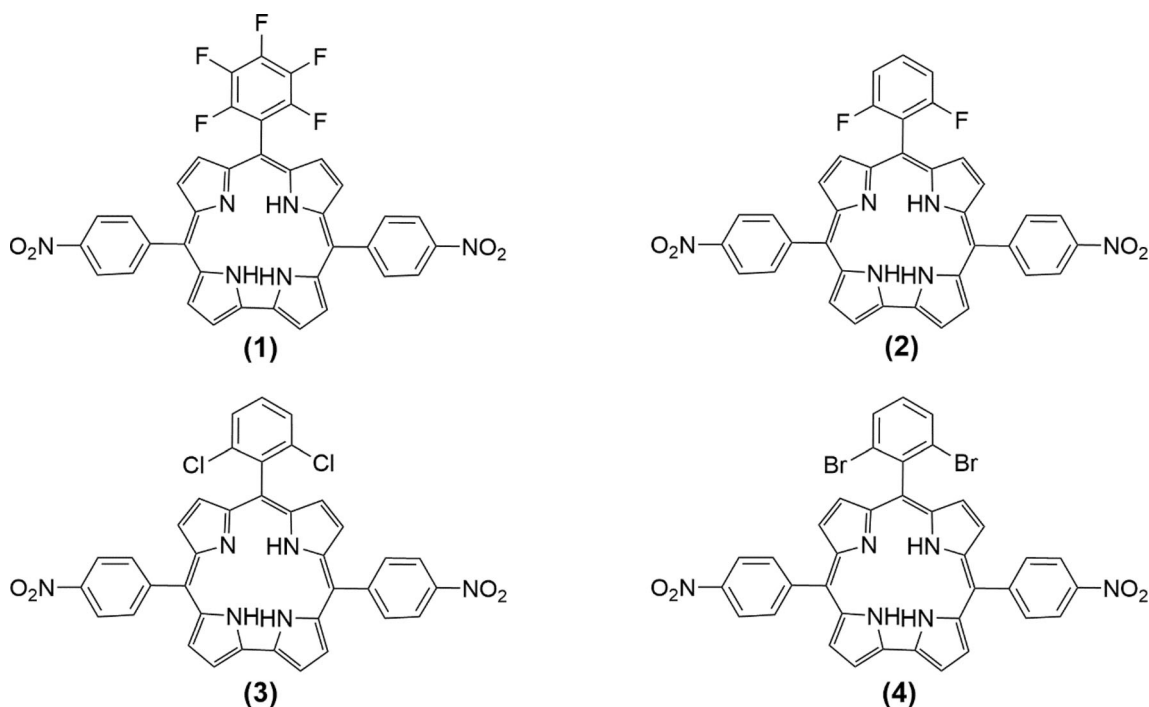
1-4, A₂B corroles, were purified by column chromatography using silica gel as an adsorbent. UV-visible spectra were collected at room temperature from UV-1800 Shimadzu spectrophotometer. The fluorescence spectra of corroles were recorded from HORIBA spectrophotometer under excitation wavelength 440 nm. Slits were set to 1.5 mm for the entrance,

2 mm for exit. Quantum yield of **1-4** corrole was measured by the experimental in triplet and calculations have done using a reference of meso-tetraphenylporphyrin ($\Phi = 0.13$).^{44d} For fluorescence quenching experiments, dry toluene solution was used for A₂B corrole and methanol solution for Hg²⁺ ion.

2.1 Syntheses

2.1a Synthesis of 5-(4-nitophenyl)dipyrromethane: Preparation of 5-(4-nitophenyl)dipyrromethane from pyrrole and 4-nitrobenzaldehyde according to reported in literature.⁵⁸

2.1b General method for the synthesis of 1-4, A₂B corroles: All A₂B corroles were synthesized as reported methods in the literature.^{40,46–49} 1 equivalent of 5-(4-nitophenyl)dipyrromethane and 0.5 equivalent of respective aldehyde were dissolved in 100 mL methanol, and 5 mL of 36% HCl_{aq} was



Scheme 1. A₂B corroles used as sensor towards Hg²⁺ ion.

added to this solution. The solution was kept for stirring for up to 2 h at room temperature. The resultant reaction mixture was extracted by chloroform (CHCl_3), washed twice with distilled water, and dried over anhydrous sodium sulphate (Na_2SO_4). Filtered off the reaction mixture and diluted with chloroform. 1.5 equivalent of p-chloranil was added into the diluted reaction mixture and stirred overnight at room temperature. TLC was used to monitor the completion of the reaction. Then the reaction mixture was collected, and evaporated to dryness. The pure corrole was obtained using silica gel in column chromatography.

2.1c *5,15-Bis(4-nitrophenyl)-10-(pentafluorophenyl) A₂B corrole (1)*: The green color solution was obtained by using column chromatography with 6:4; hexane: dichloromethane as eluent. UV-vis in toluene λ_{max} ($\epsilon/\text{M}^{-1}\text{cm}^{-1}$) 442(6797), 598(2303). Rest analytical data of **1** corrole reported in our previous article.^{46b}

2.1d *5,15-Bis(4-nitrophenyl)-10-(2,6-difluorophenyl) A₂B corrole (2)*: The green color solution was obtained by column chromatography with 1:1; hexane: dichloromethane as eluent. UV-vis in toluene λ_{max} ($\epsilon/\text{M}^{-1}\text{cm}^{-1}$) 442(5272), 598(1505). Rest analytical data of **2** corrole reported in our published article.^{46b}

2.1e *5,15-Bis(4-nitrophenyl)-10-(2,6-dichlorophenyl) A₂B corrole (3)*: The green color solution was obtained by column chromatography with 3:7; hexane: dichloromethane as eluent. UV-vis in toluene λ_{max} ($\epsilon/\text{M}^{-1}\text{cm}^{-1}$) 429(9090), 598(2404). ^1H NMR (400 MHz, CDCl_3) δ 9.08 (d, $J = 4.2$ Hz, 2H), 8.91 (d, $J = 4.7$ Hz, 2H), 8.74 (d, $J = 8.7$ Hz, 4H), 8.64 (d, $J = 2.3$ Hz, 2H), 8.59 (d, $J = 8.7$ Hz, 4H), 8.49 (d, $J = 4.7$ Hz, 2H), 7.85 (d, $J = 8.7$ Hz, 2H), 7.77–7.72 (m, 1H) (Figure S1, SI). HRMS: Calcd for $\text{C}_{37}\text{H}_{22}\text{Cl}_2\text{N}_6\text{O}_4$ m/z found 684.1074, m/z theo 684.1080 (Figure S3, SI).

2.1f *5,15-Bis(4-nitrophenyl)-10-(2,6-dibromophenyl) A₂B corrole (4)*: The intense green color solution was obtained by column chromatography with 1:9; Hexane: dichloromethane as eluent. UV-vis in toluene λ_{max} ($\epsilon/\text{M}^{-1}\text{cm}^{-1}$) 430(47731), 596(14070). ^1H NMR

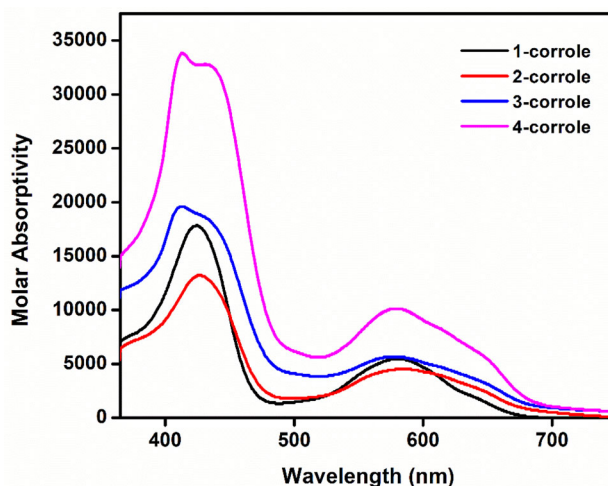


Figure 1. UV-visible spectra of **1-4** A₂B corroles in toluene (9.9 μM).

(400 MHz, CDCl_3) δ 9.07 (d, 2H), 8.91 (d, 2H), 8.74 (d, 4H), 8.63 (d, 2H), 8.60 (d, 4H), 8.48 (d, 2H), 8.07 (d, 2H), 7.58 (t, 1H) (Figure S2, SI). HRMS: Calcd for $\text{C}_{37}\text{H}_{22}\text{Br}_2\text{N}_6\text{O}_4$ m/z found $[\text{M}+1]$ 772.9992, m/z theo 771.9912 (Figure S4, SI).

3. Results and Discussion

Corrole **1** and **2** have already been reported in our previous article.^{46b} Corrole **3** and **4** were synthesized with previous synthetic protocols.^{47,48} The corrole **1** differ from corrole **2** by the number of fluorine atoms at the phenyl ring at C₁₀ position of corrole. Corrole **2**, **3**, **4** differ by the halogen atom at both *ortho* positions of phenyl ring, as shown in Scheme 1. Free base corroles were characterized by different spectroscopic techniques and explained in the experimental section. A₂B corroles, **1-4**, have different electron density due to the electronic effect of halogen atoms. The electronic effect exerted by the halogen atom were examined for the sensing ability of A₂B corrole towards Hg^{2+} ion.

Photophysical characterizations of corroles **1-4** were observed in the toluene at room temperature are shown

Table 1. Photophysical data of **1-4** A₂B corroles in toluene at room temperature.

Probe	$\lambda_{\text{max}}/\text{nm}$ ($\epsilon/\text{M}^{-1}\text{cm}^{-1}$)	λ_{em} (nm)	Stoke's shift (cm^{-1})	FWHM	QY%
1	442(6797), 598(2303)	672	7743	46.06	0.10
2	442(5272), 598(1505)	680	7919	45.34	0.11
3	429(9090), 598(2404)	686	8733	52.97	0.12
4	430(47731), 596(14070)	687	8700	50.85	0.13

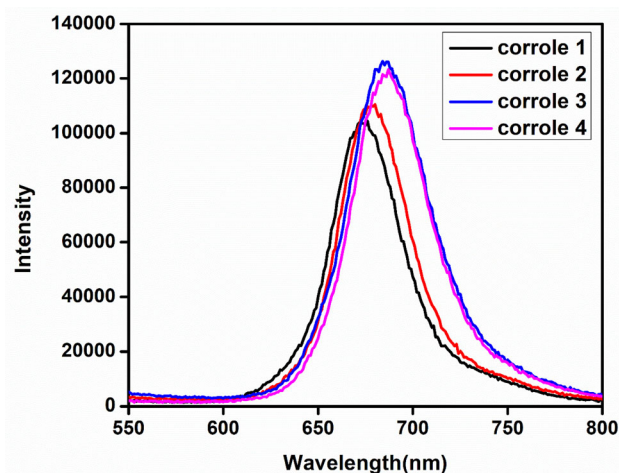


Figure 2. Emission spectra of **1-4** A₂B corroles in toluene at 440 nm excitation.

in Table 1. B bands were observed at 428–444 nm, and Q bands appeared at 594–598 nm in the toluene solution (Figure 1). The transition of non-bonding electron $n \rightarrow \sigma^*$, $n \rightarrow \pi^*$ is one of the reasons for absorbance in

UV-visible spectra of corrole. It is predicted that halogen atoms have non-bonding electrons with the ease of availability of non-bonding electron decreasing as $\text{Br} > \text{Cl} > \text{F}$.⁵⁰ Corrole **4** has the highest molar absorption coefficient due to the availability of non-bonding electrons of Br atom that is higher than that from the Cl and F atoms. Corrole **3** has a higher molar absorption coefficient compared to corrole **1** and corrole **2** due to the availability of non-bonding electrons of Cl atom higher than F atoms. Corrole **1** has a higher molar absorption coefficient than corrole **2** because corrole **1** have higher non-bonding electrons of five F atoms compared to two F atoms of corrole **2**, respectively. So we observed molar absorption coefficient of corroles $4 > 3 > 1 > 2$ at 428–444 nm and 594–598 nm. Also, in general, a bathochromic shift occurred by increasing the dielectric constant of the solution. But according to previous literature, in the case of bis(nitro)substituted corroles, the opposite trend⁴⁹ were recorded. Due to this behaviour of these corroles, bis(nitro)substituted corroles, have unique property

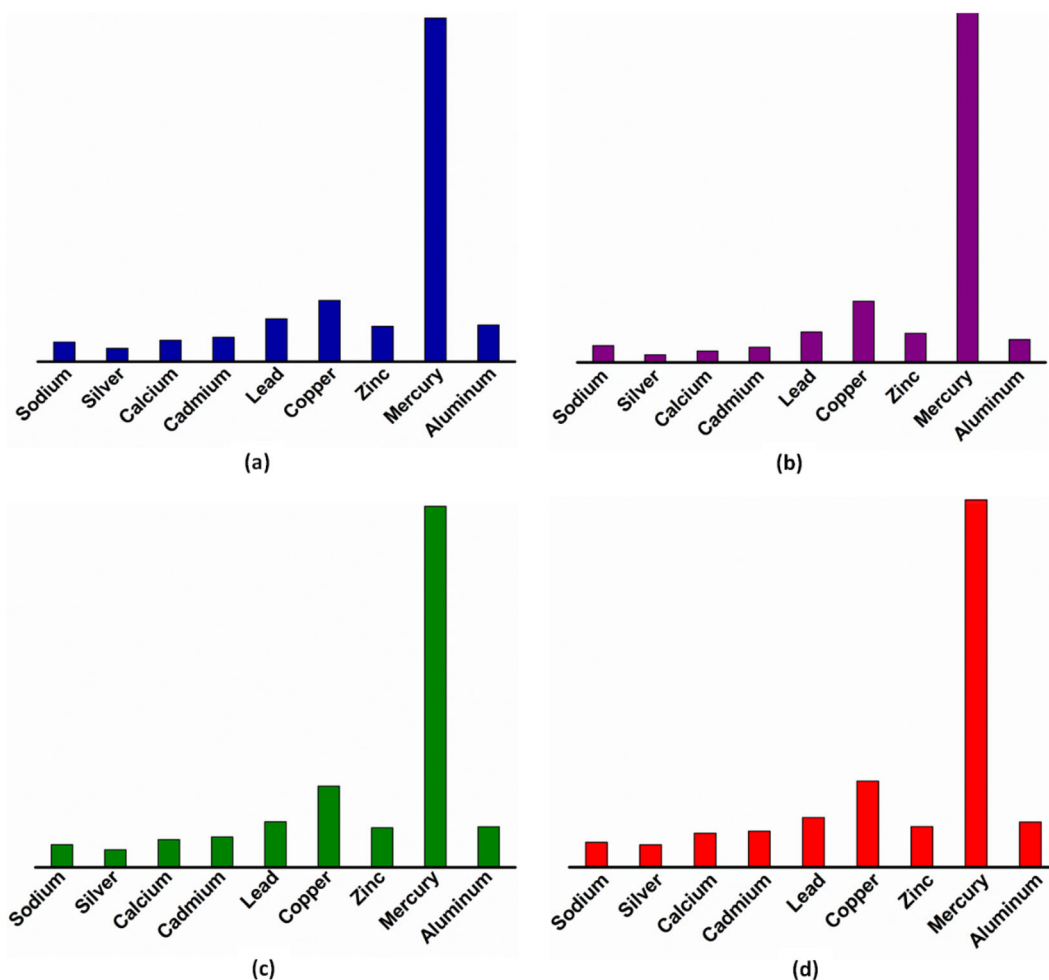


Figure 3. Fluorescence response of **1-4** of 9.9×10^{-6} M to various cations (3.3×10^{-7} M) in toluene represented as a–d, respectively.

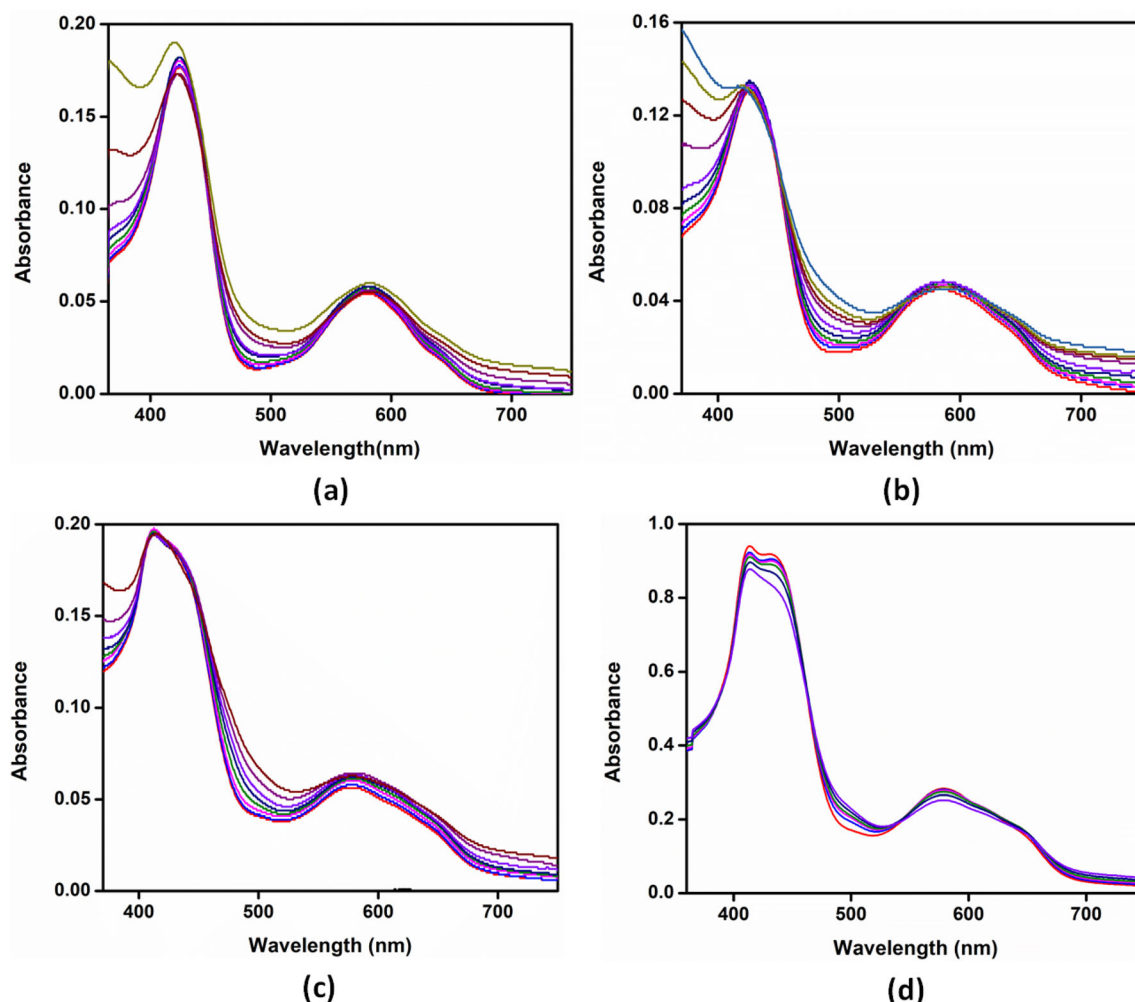


Figure 4. Change in UV–visible spectrum during the addition of methanolic solution of Hg(II) acetate (0 to 100 μL) in a toluene solution of **1-4** A₂B corroles in the aerobic condition represented as **a**, **b**, **c**, **d** respectively.

compare to other substituted corroles i.e. without nitro group. The steady state fluorescence emission spectra of **1-4** A₂B corroles were recorded in toluene. We observed strong emission bands within the range of 673–687 nm (Figure 2). We have calculated Stoke's shift with the help of emission spectra of **1-4** A₂B corroles. The Stoke's shift was calculated and found within the range 230–257 nm. This may be because of a change in the electronic nature of the excited state as compared to the ground state.

Further, quantum yield of **1-4** A₂B corroles were also calculated by employing the Eq. (1). In which, tetraphenylporphyrin was used as a reference material.

$$QY_S = QY_R \times \left(\frac{I_S}{I_R} \right) \times \left(\frac{A_R}{A_S} \right) \times \left(\frac{\eta_S}{\eta_R} \right)^2 \quad (1)$$

Where, QY_S and QY_R are the quantum yields of the sample and reference (Quantum yield of *meso*-tetraphenylporphyrin = 0.13). I_S and I_R are the integrated area under the PL spectrum of sample and

reference (*meso*-tetraphenylporphyrin). A_R and A_S are the absorbance, η_R and η_S are the refractive indexes of the solvents of reference (*meso*-tetraphenylporphyrin) and sample, respectively. The quantum yield of corrole **3** and **4** was higher than the reported 5,15-bis(nitrophenyl) A₂B corroles due to stronger charge transfer.^{46b,49b,51}

5,15-bis(nitrophenyl) A₂B corroles also detect other metal ions but shows higher sensing ability with Hg²⁺ ion even at lower concentration (i.e., $<10^{-6}$ – 10^{-9} M).⁴⁰ The fluorescence quenching of A₂B corroles towards Hg²⁺ ion with the change in halogen atom at phenyl ring at C₁₀ position was also examined. For fluorescence study, we have chosen toluene as a solvent for corroles and methanol for analyte.⁴⁰ The photostability of corroles, **1-4**, and remarkable solubility in toluene made it a better choice for fluorescence study.

The fluorescence response of **1-4** were determined for different cations in order to check the specific sensing of **1-4** by carrying out fluorescence titration.

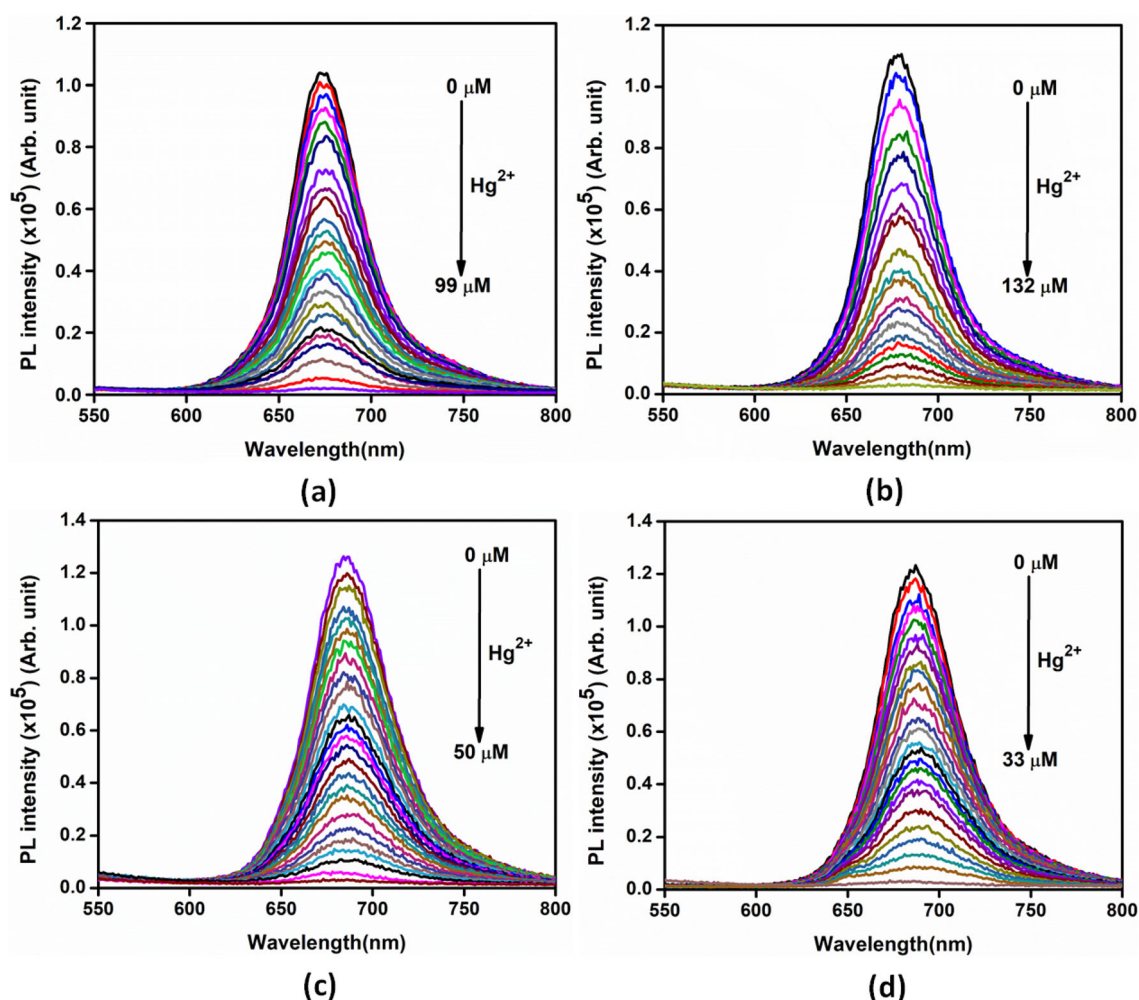


Figure 5. The decay in fluorescence emission intensity during the titration of methanolic solution of Hg(II) acetate (3.3×10^{-7} M) in a toluene solution of **1-4** A₂B corroles (9.9×10^{-6} M) respectively under the excitation at $\lambda = 440$ nm in aerobic condition.

The results obtained from fluorescence titration are shown in Figure 3, which clearly indicates that quenching of fluorescence intensity is lesser influenced by these cations relative to that of Hg^{2+} ion. The corroles, **1-4** show peculiar selectivity and high affinity for Hg^{2+} ion.

When a solution of Hg^{2+} ion prepared in methanol is added to the toluene solution of A₂B corroles, a hypsochromic shift of ICT transition occurs. The hypsochromic shift occurs due to the binding of Hg^{2+} ion through lone pair of the inner nitrogen atom of corrole. Hg^{2+} ion serves as a quencher through the spin-orbital coupling effect.⁵²

The binding of Hg^{2+} ion with **1-4** A₂B corroles were also easily observed by the color change of the toluene solution of corroles through the naked eye as well as under UV lamp. The green color of **1-4** A₂B corroles changed into yellowish brown color during the addition of Hg^{2+} ion. In UV-lamp, without Hg^{2+} ion A₂B corroles are highly fluorescent pink color, and

after addition of Hg^{2+} ion it transforms into colourless solution with no fluorescent. Also, the molar absorption coefficient of **1-4** A₂B corroles continuously decreases with increasing the concentration of Hg^{2+} ion and shown in Figure 4.

In addition, when we recorded fluorescence spectra of a titration between different concentrations of methanol solution of Hg^{2+} ion into toluene solution of **1-4** corroles, quenching of fluorescence emission intensity continuously decreased, as shown Figure 5. This is due to the non-radiative relaxation path from an excited state to ground state, which signifies the absence of an emission band and represents the new fluorophore formation of Hg(II)-corrole. Excessive addition of Hg^{2+} ion, the saturation point was observed. It is evident that all the **1-4** A₂B corrole have different saturation points because of the number of halogen atoms and different halogen atom present on phenyl ring at C₁₀ position of corrole. The saturation point of **1, 2, 3, 4** corrole are 99 μM , 132 μM ,

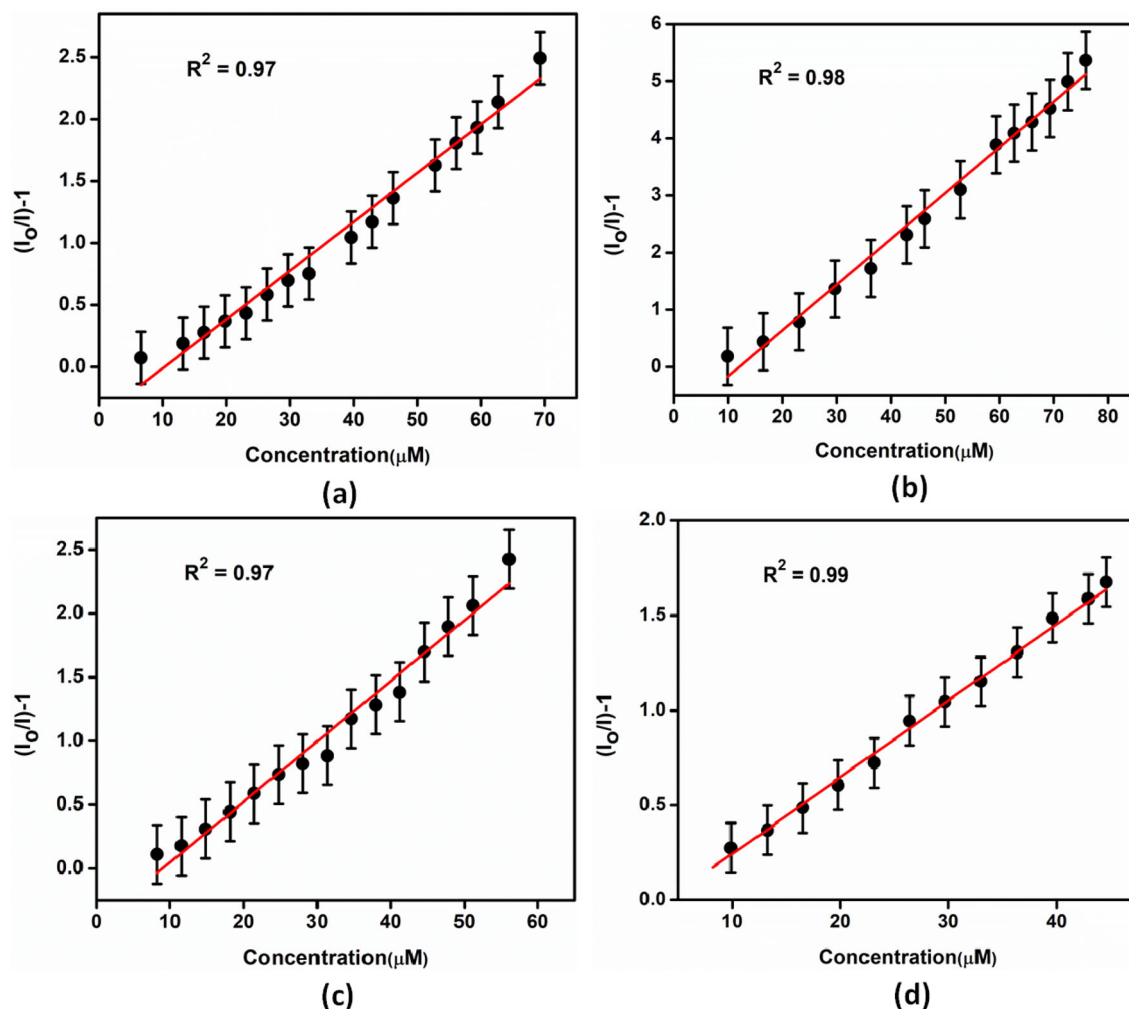


Figure 6. The Stern–Volmer plot of compounds **1–4** A₂B corroles respectively with positive deviation during increasing the concentration of a methanolic solution of Hg(II) acetate inset the value of linearity constant (R^2).

Table 2. Comparison of the LOD with other reported sensors.

S. No.	LOD towards Hg^{2+}	Used conc. of Hg^{2+} for sensing	Reference
1.	-	1.2×10^{-7} to 1.0×10^{-4} M	[38]
2.	-	0 to 60×10^{-9} M	[39]
3.	-	2.49×10^{-6} M	[40]
4.	0.02 (ppm)	10^{-4} - 10^{-6} M	[41]
5.	2.64-38.61 (μM)	3.3×10^{-7} M	this work

Table 3. Comparison of **1–4** probes for Hg^{2+} sensing

Probe	K_{SV} (M^{-1})	LoD (μM)	Linearity range (μM)	R^2
1	0.39×10^5	21.93	6.6 - 69.3	0.97
2	0.81×10^5	38.61	9.9 - 75.9	0.98
3	0.49×10^5	8.72	8.25 - 56.1	0.97
4	0.40×10^5	2.64	9.9 - 44.55	0.99

50 μM , 33 μM , respectively (Figure 4a, 4b, 4c, 4d). Corrole **4** showed the highest quenching ability towards Hg^{2+} corroles. The quenching order of A_2B corroles was **4**>**3**>**1**>**2** under identical conditions. It is well shown in Figure 4, the order of quenching through the increasing concentration of Hg^{2+} ion used for titration.

In general static interaction or dynamic interaction, or both are responsible for quenching the PL intensity in the presence of quencher.⁵⁴ But in this case of A_2B corrole static as well as dynamic interactions are responsible for quenching in the presence of Hg^{2+} ion.⁴⁰ Further investigation by the well-known Stern-Volmer relation (Eq.2) was examined. The S-V plot for the quenching of PL intensity of A_2B corrole in the presence of various concentrations of Hg^{2+} ion (Figure 5).

$$\frac{I_0}{I} = 1 + K_{SV} [\text{Hg}^{2+}] \quad (2)$$

Where I_0 represents the initial PL intensity without Hg^{2+} ion and I represent the final PL intensity in the presence of quencher (Hg^{2+} ion), $[\text{Hg}^{2+}]$ signifies the molar concentration of methanol solution of Hg^{2+} ion added into toluene solution of A_2B corrole and K_{SV} represent the PL quenching constant [M^{-1}]. We observed linear relation with positive deviation between $(I_0/I)-1$ and molar concentration of Hg^{2+} ion in the S-V plot for a particular range of concentrations (Figure 6), which is similar to earlier reported in the literature for quenching by metal ions.⁵⁵

From the Stern–Volmer plot, the calculated K_{SV} (10^5 M^{-1}) for **1**, **2**, **3**, **4** A_2B corroles are 0.39, 0.81, 0.49, 0.40, respectively. The limit of detection (LoD) of **1–4** corroles was found 21.93 μM , 38.61 μM , 8.72 μM , 2.64 μM , respectively, using formula $3\sigma/K$, which is comparable to other reported sensors in the literature for Hg^{2+} ions and tabulated in Table 2. Where σ represent the standard deviation and K represents the slope of the Stern–Volmer plot. The linear detection ranges for **1**, **2**, **3**, and **4** are 6.6×10^{-6} to $69.3 \times 10^{-6} \text{ M}$, 9.9×10^{-6} to $75.9 \times 10^{-6} \text{ M}$, 8.25×10^{-6} to $56.1 \times 10^{-6} \text{ M}$ and 9.9×10^{-6} to $44.55 \times 10^{-6} \text{ M}$, respectively. The linearity constant (R^2) values for **1–4** A_2B corrole are 0.97, 0.98, 0.97, 0.99, respectively. The values of K_{SV} , LoD, linearity range, and R^2 are tabulated in Table 3. We observed that the order of LoD for **1–4** A_2B corroles towards Hg^{2+} ion is **4** > **3** > **1** > **2**.

Fluorescence titration was carried out for corroles, **1–4**, with the addition of Hg^{2+} ion. The significant quenching of fluorescence intensity of corroles was

observed, and quenching is more appreciable in the presence of a higher concentration of Hg^{2+} ion. The S-V plot (Figure 5) signifies the dynamic and static interactions through $(I_0/I)-1$ and the concentration of Hg^{2+} ion. The formation of long-lived charge-separated states in nonpolar solvents also quenches the fluorescence in corrole-fullerene.⁵⁶ This is the reason that in the presence of Hg^{2+} ion, electronic spectral changes occurred in A_2B corrole. The order of sensing depends on the electron-donating property of the halogen atom, which is present at C_{10} position of corrole. The increasing order of electron-donating efficiency of halogen atom is $\text{F} < \text{Cl} < \text{Br}$.⁵⁷ It is clear from the above discussion that the sensing efficiency of A_2B corrole depends upon the behaviour of halogen atom.

4. Conclusions

The synthesized the A_2B corroles (**1–4**) were explored as Hg^{2+} ion sensor with different sensing efficiency. Quenching of the fluorescence emission intensity with the addition concentration of Hg^{2+} ion was studied. The sensing order of **1–4** corrole towards Hg^{2+} ion is **4**>**3**>**1**>**2**, which demonstrates the electronic effect of halogen atoms present on the phenyl ring of C_{10} position of corrole. The sensing order is proved by the value of LoD as well as the amount of Hg^{2+} used for fluorescence titration. We also conclude that increase in the halogen atom on phenyl group will increase the sensing efficiency of substituted 5,15-bis(nitrophenyl) free base corrole as in case of **1** corrole and **2** corrole. But in the case of **2**, **3**, **4**, halogen atom fluorine (F), chlorine (Cl), bromine (Br) are present at 2,6 position of phenyl ring, respectively at C_{10} position of corrole. We observed sensing efficiency of **4** higher than **3** corrole and sensing efficiency of **3** corrole higher than not only for **2** corrole but also higher than **1** towards Hg^{2+} ion. This may be due to the electron-donating property of halogen atom. To increase the sensing efficiency of 5,15-bis(nitrophenyl) A_2B corrole towards Hg^{2+} ion, we need to increase the electron donating atom on the corrole. Though a majority of Hg^{2+} ion-containing pollutants are in the aqueous phase, sensors have been developed using different matrices. But corrole is completely dissolved in toluene and mercury acetate in methanol. The solvent toluene and methanol have previously been used for this quenching study. The water-soluble corrole-based sensor is the current interest of our lab.

Supplementary Information (SI)

All spectral data (^1H NMR, HRMS) of the new compounds are available at <http://www.ias.ac.in/chemsci>.

Acknowledgements

Authors thanks Delhi Technological University, Delhi, for its facilities.

References

- (a) Niessner R 1991 Chemical sensors for environmental analysis *Trends Anal. Chem.* **10** 310; (b) Szmazinski H and Lakowicz J R 1995 Fluorescence lifetime-based sensing and imaging *Sens. Actuat. B* **29** 16; (c) Wang L, Wang T, Xia L, Dong L, Bian G and Chen H 2004 Direct Fluorescence Quantification of Chromium(VI) in Wastewater with Organic Nanoparticles Sensor *Anal. Sci.* **20** 1013; (d) Ma C, Zeng F, Huang L and Wu S 2011 FRET-sed Ratiometric Detection System for Mercury Ions in Water with Polymeric Particles as Scaffolds *J. Phys. Chem. B* **115** 874; (e) Su S, Wu W, Gao J, Lu J and Fan C J 2012 Nanomaterials-based sensors for applications in environmental monitoring *Mater. Chem.* **22** 18101; (f) Kim H N, Ren W X, Kim J S and Yoon J 2012 Fluorescent and colorimetric sensors for detection of lead, cadmium, and mercury ions *Chem. Soc. Rev.* **41** 3210
- (a) Arnold A M and Small G W 2005 Noninvasive Glucose Sensing *Anal. Chem.* **77** 5429; (b) Aslan K, Zhang J, Lackowicz J R and Geddes C D 2004 Saccharide Sensing Using Gold and Silver Nanoparticles-A Review *J. Fluoresc.* **14** 391; (c) Ballerstadt R, Evans C, McNichols R and Gowda A 2006 Concanavalin A for in vivo glucose sensing: A biotoxicity review *Biosens. Bioelectron.* **22** 275; (d) Kondepoti V R and Heise H M 2007 Recent progress in analytical instrumentation for glycemic control in diabetic and critically ill patients *Anal. Bioanal. Chem.* **388** 545; (e) Pickup J C, Hussain F, Evans N D, Rolinski O J and Birch D J S 2005 Fluorescence-based glucose sensors *Biosens. Bioelectron.* **20** 2555; (f) Wang J 2008 Electrochemical Glucose Biosensors *Chem. Rev.* **108** 814
- (a) Pouya S, Koochesfahani M M, Snee P, Bawendi M G and Nocera D G 2005 Single quantum dot (QD) imaging of fluid flow near surfaces *Exp. Fluids* **39** 784; (b) Pouya S, Koochesfahani M M, Greytak A B, Bawendi M G and Nocera D G 2008 Experimental evidence of diffusion-induced bias in near-wall velocimetry using quantum dot measurements *Exp. Fluids* **44** 1035; (c) Hu H, Jin Z, Nocera D, Lum C and Koochesfahani M 2010 Experimental investigations of micro-scale flow and heat transfer phenomena by using molecular tagging techniques *Meas. Sci. Technol.* **21** 085401; (d) Ji H F, Shen Y, Hubner J P, Carroll B F, Schmehl R H, Simon J A and Schanze K S 2000 Temperature-Independent Pressure-Sensitive Paint Based on a Bichromophoric Luminophore *Appl. Spectrosc.* **54** 856; (e) Carroll B F, Hubner J P, Schanze K S and Bedlek-Anslow J M 2001 Principal Component Analysis of Dual-luminophore Pressure/Temperature Sensitive Paints *J. Visual.* **4** 121; (f) Carroll B F, Abbott J D, Lukas E W and Morris M J 1996 Step response of pressure-sensitive paints *AIAA J.* **34** 521
- (a) Cable M L, Kirby J P, Sorasaene K, Gray H B and Ponce A 2007 Bacterial Spore Detection by $[\text{Tb}^{3+}(\text{-macrocycle})(\text{dipicolinate})]$ Luminescence *J. Am. Chem. Soc.* **129** 1474; (b) Yung P T, Lester E D, Bearman G and Ponce A 2007 An automated front-end monitor for anthrax surveillance systems based on the rapid detection of airborne endospores *Biotechnol. Bioeng.* **98** 864; (c) Royo S, Martinez-Manez R, Sancenon F, Costero A M, Parra M and Gil S 2007 Chromogenic and fluorogenic reagents for chemical warfare nerve agents' detection *Chem. Commun.* **46** 4839; (d) Jiang Y, Zhao H, Zhu N, Lin Y, Yu P and Mao L 2008 A simple assay for direct colorimetric visualization of trinitrotoluene at picomolar levels using gold nanoparticles *Angew. Chem. Int. Ed.* **47** 8601; (e) Gao D, Wang Z, Liu B, Ni L, Wu M and Zhang Z 2008 Resonance Energy Transfer-Amplifying Fluorescence Quenching at the Surface of Silica Nanoparticles toward Ultrasensitive Detection of TNT *Anal. Chem.* **80** 8545; (f) Ai K, Zhang B and Lu L 2008 Europium-Based Fluorescence Nanoparticle Sensor for Rapid and Ultrasensitive Detection of an Anthrax Biomarker *Angew. Chem. Int. Ed.* **121** 310; (g) Oh W K, Jeong Y S, Song J and Jang J 2011 Fluorescent europium-modified polymer nanoparticles for rapid and sensitive anthrax sensors *Biosens. Bioelectron.* **29** 172
- (a) Bell T and Hext N M 2004 Supramolecular optical chemosensors for organic analytes *Chem. Soc. Rev.* **33** 589; (b) De Silva A P, Gunaratne H Q N, Gunlaugsson T, Huxley A J M, McCoy C P, Rademacher J T and Rice T E 1997 Signaling Recognition Events with Fluorescent Sensors and Switches *Chem. Rev.* **97** 1515; (c) Czarnik A W 1994 Chemical Communication in Water Using Fluorescent Chemosensors *Acc. Chem. Res.* **27** 302; (d) Rudzinski C M, Hartmann W K and Nocera D G 1998 Lanthanide-ion modified cyclodextrin supramolecules *Coord. Chem. Rev.* **171** 115; (e) Rudzinski C M, Young A M and Nocera D G 2002 A Supramolecular Microfluidic Optical Chemosensor *J. Am. Chem. Soc.* **124** 1723; (f) Swager T M 1998 The Molecular Wire Approach to Sensory Signal Amplification *Acc. Chem. Res.* **31** 201; (g) Mohr G J 2004 Chromo- and Fluororeactants: Indicators for Detection of Neutral Analytes by Using Reversible Covalent-Bond Chemistry *Chem. Eur. J.* **10** 1082; (h) De Silva A P, Fox D B, Moody T S and Weir S M 2001 The development of molecular fluorescent switches *Trends Biotechnol.* **19** 29
- (a) Dahan M, Levi S, Luccardini C, Rostaing P, Riveau B and Triller A 2003 Diffusion dynamics of glycine receptors revealed by single-quantum dot tracking *Science* **302** 442; (b) Shimizu K T, Neuhauser R G, Leatherdale C A, Empedocles S A, Woo W K and Bawendi M G 2001 Blinking statistics in single semiconductor nanocrystal quantum dots *Phys. Rev. B* **63** 205316; (c) Deniz A A, Dahan M, Grunwell J R, Ha

- T, Faulhaber A E, Chemla D S, Weiss S and Schultz P G 1999 Single-pair fluorescence resonance energy transfer on freely diffusing molecules: Observation of Förster distance dependence and subpopulations *Proc. Natl. Acad. Sci. U.S.A.* **96** 3670; (d) Ha T, Ting A Y, Liang J, Caldwell W B, Deniz A A, Chemla D S, Schultz P G and Weiss S 1999 Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism *Proc. Natl. Acad. Sci. U.S.A.* **96** 893
7. Lodeiro C, Pina E, Parola A J, Bencini A, Bianchi A, Bazzicalupi C, Ciattini S, Giorgi C, Masotti A, Valtancoli B and Melo J S 2001 Exploring the Photocatalytic Properties and the Long-Lifetime Chemosensor Ability of $\text{Cl}_2[\text{Ru}(\text{Bpy})_2\text{L}]$ ($\text{L} = 2,5,8,11,14\text{-Pentaaza}[15]-2,2'\text{-bipyridilophane}$) *Inorg. Chem.* **40** 6813
8. Xu Z, Yoon Y and Spring D R 2010 Fluorescent chemosensors for Zn^{2+} *Chem. Soc. Rev.* **39** 1996
9. Lodeiro C, Capelo J L, Mejuto J C, Oliveira E, Santos H M, Pedras B and Nunez C 2010 Light and colour as analytical detection tools: A journey into the periodic table using polyamines to bio-inspired systems as chemosensors *Chem. Soc. Rev.* **39** 2948
10. (a) Chen P and He C 2004 A General Strategy to Convert the MerR Family Proteins into Highly Sensitive and Selective Fluorescent Biosensors for Metal Ions *J. Am. Chem. Soc.* **126** 728; (b) Henary M M and Fahrni C J 2002 Excited State Intramolecular Proton Transfer and Metal Ion Complexation of 2-(2'-Hydroxyphenyl)benzoxoles in Aqueous Solution *J. Phys. Chem. A* **106** 5210; (c) Walkup G K, Imperiali B 1996 Design and Evaluation of a Peptidyl Fluorescent Chemosensor for Divalent Zinc *J. Am. Chem. Soc.* **118** 3053
11. Aviv I and Gross Z 2007 Corrole-based applications *Chem. Commun.* **20** 1987
12. Santos C I M, Barata J F B, Calvete M J F, Vale L S H P, Dini D, Meneghetti M, Neves M G P M S, Faustino M A F, Tome A C and Cavaleiro J A S 2014 Synthesis and Functionalization of Corroles. An Insight on Their Nonlinear Optical Absorption Properties *Curr. Org. Synth.* **11** 29
13. Barata J F B, Neves M G P M S, Faustino M A F, Tome A C and Cavaleiro J A S 2017 Strategies for Corrole Functionalization *Chem. Rev.* **117** 3192
14. Ventura B, Degli Esposti A, Koszarna B, Gryko D T and Flamigni L 2005 Photophysical Characterization of Free-base Corroles, Promising Chromophores for Light Energy Conversion and Singlet Oxygen Generation *New J. Chem.* **29** 1559
15. Ding T, Aleman E A, Modarelli D A and Ziegler C J 2005 Photophysical Properties of a Series of Free-Base Corroles *J. Phys. Chem. A* **109** 7411
16. Mahammed A and Gross Z 2015 Metallocorroles as Photocatalysts for Driving Endergonic Reactions, Exemplified by Bromide to Bromine Conversion *Angew Chem. Int. Ed.* **54** 12370
17. Vestfrid J, Goldberg I and Gross Z 2014 Tuning the Photophysical and Redox Properties of Metallocorroles by Iodination *Inorg. Chem.* **53** 10536
18. Rabinovich E, Goldberg I and Gross Z 2017 Gold(I) and Gold(III) Corroles *Chem. Eur. J.* **17** 12294
19. Palmer JH, Durrell AC, Gross Z, Winkler JR and Gray HB 2010 Near-IR Phosphorescence of Iridium(III) Corroles at Ambient Temperature *J. Am. Chem. Soc.* **132** 9230
20. Palmer JH, Day MW, Wilson AD, Henling LM, Gross Z and Gray HB 2008 Iridium Corroles *J. Am. Chem. Soc.* **130** 7786
21. Lemon CM, Halbach RL, Huynh M and Nocera DG 2015 Photophysical Properties of β -Substituted Free-Base Corroles *Inorg. Chem.* **54** 2713
22. Shi L, Liu H-Y, Shen H, Hu J, Zhang G-L, Wang H, et al. 2009 Fluorescence properties of halogenated mono-hydroxyl corroles: the heavy-atom effects *J. Porphyr. Phthalocyan.* **13** 1221
23. Zhang L, Liu Z-Y, Zhan X, Wang L-L, Wang H and Liu H-Y 2015 Photophysical properties of electron-deficient free-base corroles bearing *meso*-fluorophenyl substituents *Photochem. Photobiol. Sci.* **14** 953
24. Kowalska D, Liu X, Tripathy U, Mahammed A, Gross Z, Hirayama S and Steer RP 2009 Ground- and Excited-State Dynamics of Aluminum and Gallium Corroles *Inorg. Chem.* **48** 2670
25. Liu X, Mahammed A, Tripathy U, Gross Z and Steer R P 2008 Photophysics of Soret-excited tetrapyrroles in solution. III. Porphyrin analogues: Aluminum and gallium corroles *Chem. Phys. Lett.* **459** 113
26. Stensitzki T, Yang Y, Berg A, Mahammed A, Gross Z and Heyne K 2016 Ultrafast electronic and vibrational dynamics in brominated aluminum corroles: Energy relaxation and triplet formation *Struct. Dyn.* **3** 043210
27. Mahammed A, Tumanskii B and Gross Z 2011 Effect of bromination on the electrochemistry, frontier orbitals, and spectroscopy of metallocorroles *J. Porphyr. Phthalocyan.* **15** 1275
28. Shao W, Wang H, He S, Shi L, Peng K, Lin Y, et al. 2012 Photophysical Properties and Singlet Oxygen Generation of Three Sets of Halogenated Corroles *Phys. Chem. B* **116** 14228
29. Paolesse R 2000 In *The Porphyrin Handbook* K M Kadish, K M Smith and R Guillard (Eds.) 2nd ed. (Academic Press: New York) Ch.11 p. 202
30. Aviv-Harel I and Gross Z 2011 Coordination chemistry of corroles with focus on main group elements *Coord. Chem. Rev.* **255** 717
31. (a) Hussain S M, Hess K L, Gearhart J M, Geiss K T and Schlager J J 2005 In vitro toxicity of nanoparticles in BRL 3A rat liver cells *Toxicol. In Vitro* **19** 975; (b) Landsdown A B G 2007 Critical Observations on the Neurotoxicity of Silver *Crit. Rev. Toxicol.* **37** 237
32. Donnell EE, Han S, Hilty C, Pierce KL and Pines A 2005 NMR Analysis on Microfluidic Devices by Remote Detection *Anal. Chem.* **77** 8109
33. Lione A 1985 Aluminum intake from non-prescription drugs and saccharate *J. Gen. Pharmacol.* **16** 223
34. Crisponi G, Nurchi VM, Bertolas V, Remelli M and Faa G 2012 Chelating agents for human diseases related to aluminium overload *Coord. Chem. Rev.* **256** 89
35. (a) Chen X, Nam X-W, Jou M, Kim Y, Kim S-J, Park S and Yoon J 2008 Hg^{2+} Selective Fluorescent and Colorimetric Sensor: Its Crystal Structure and Application to Bioimaging *Org. Lett.* **10** 5235; (b) Nolan E M and Lippard S J 2008 Tools and Tactics for the Optical

- Detection of Mercuric Ion *Chem. Rev.* **108** 3443; (c) Tamayo A, Pedras B, Lodeiro C, Escriche L, Casabo J, Capelo J L, Covelo B, Kivekas R and Sillampaa R 2007 Exploring the Interaction of Mercury(II) by N_2S_2 and NS_3 Anthracene-Containing Macrocyclic Ligands: Photophysical, Analytical, and Structural Studies *Inorg. Chem.* **46** 7818; (d) Mameli M, Lippolis V, Caltagirone C, Capelo J L, Nieto-Faza O and Lodeiro C 2010 Hg^{2+} Detection by New Anthracene Pendant-Arm Derivatives of Mixed N/S- and N/S/O-Donor Macrocycles: Fluorescence, Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry and Density Functional Theory Studies *Inorg. Chem.* **49** 8276
36. (a) Rocha A, Marques M M B and Lodeiro C 2009 Synthesis and characterization of novel indole-containing half-crowns as new emissive metal probes *Tetrahedron Lett.* **50** 4930; (b) Jou M J, Chen X, Swamy K M K, Kim H N, Kim H-J, Lee S and Yoon J 2009 Highly selective fluorescent probe for Au^{3+} based on cyclization of propargylamide *Chem. Commun.* **46** 7218; (c) Park C S, Lee J Y, Kang E-J, Lee J-E and Lee S S 2009 A highly selective fluorescent chemosensor for silver(I) in water/ethanol mixture *Tetrahedron Lett.* **50** 671; (d) Tamayo A, E. Oliveira E, Covelo B, Casabo J, Escriche L and Lodeiro C 2007 Exploring the Interaction of Anthracene-Containing Macrocyclic Chemosensors with Silver(I) and Cadmium(II) Ions – Photophysical and Structural Studies *Z. Anorg. Allg. Chem.* **633** 1809
 37. (a) Capelo J L, Maduro C and Mota A M 2006 Evaluation of focused ultrasound and ozonolysis as sample treatment for direct determination of mercury by FI-CV-AAS. Optimization of parameters by full factorial design *Ultrason. Sonochem.* **13** 98; (b) Nolan E M and Lippard S J 2008 Tools and Tactics for the Optical Detection of Mercuric Ion *Chem. Rev.* **108** 3443
 38. He C-L, Ren F-L, Zhang X-B and Han Z-X 2006 A fluorescent chemical sensor for $Hg(II)$ based on a corrole derivative in a PVC matrix *Talanta* **70** 364
 39. Zhou Y, Deng M, Du Y, Yan S, Huang R, Weng X, Yang C, Zhang X and Zhou X 2011 A fluorescent chemical sensor for $Hg(II)$ based on a corrole derivative in a PVC matrix *Analyst* **136** 955
 40. Pariyar A, Bose S, Chhetri S S, Biswas A N and Bandyopadhyay P 2012 Fluorescence signaling systems for sensing $Hg(ii)$ ion derived from A_2B -corroles *Dalton Trans.* **41** 3826
 41. Santos CIM, Oliveira E, Fernandez-Lodeiro J, Barata JFB, Santos SM, Faustino MAF, et al. 2013 Corrole and Corrole Functionalized Silica Nanoparticles as New Metal Ion Chemosensors: A Case of Silver Satellite Nanoparticles Formation *Inorg. Chem.* **52** 8564
 42. Adinarayana B, Thomas AP, Yadav P, Kumar A and Srinivasan A 2016 Bipyricorrole: A Corrole Homologue with a Monoanionic Core as a Fluorescence Zn^{II} Sensor *Angew. Chem. Int. Ed.* **55** 969
 43. Li Y, Chen M, Han Y, Feng Y, Zhang Z and Zhang B 2020 Fabrication of a New Corrole-Based Covalent Organic Framework as a Highly Efficient and Selective Chemosensor for Heavy Metal Ions *Chem. Mater.* **6** 2532
 44. (a) Plaschke M, Czolk R and Ache H J 1995 Fluorimetric determination of mercury with a water-soluble porphyrin and porphyrin-doped sol-gel films *Analyt. Chim. Acta* **304** 107; (b) Chan W H, Yang R H and Wang K M 2001 Development of a mercury ion-selective optical sensor based on fluorescence quenching of 5, 10, 15, 20-tetraphenylporphyrin *Analyt. Chim. Acta* **444** 261; (c) Zhang X B, Guo C C, Li Z Z, Shen G L and Yu R Q 2002 An optical fiber chemical sensor for mercury ions based on a porphyrin dimer *Anal. Chem.* **74** 821; (d) Gouterman M 1978 Optical Spectra and Electronic Structure of Porphyrins and Related Rings In *The Porphyrins* Vol. III D Dolphin (Ed.) (New York: Academic Press) Ch.1.
 45. Jensen WB 2003 The place of zinc, cadmium, and mercury in the periodic table *J. Chem. Edu.* **80** 952
 46. (a) Yadav O, Varshney A and Kumar A 2017 Manganese(III) mediated synthesis of A_2B Mn(III) corroles: A new general and green synthetic approach and characterization *Inorg. Chem. Commun.* **86** 168; (b) Yadav O, Varshney A, Kumar A, Ratnesh R K and Mehata M S 2018 A_2B corroles: Fluorescence signaling systems for sensing fluoride ions *Spectrochim. Acta Part A* **202** 207; (c) Varshney A, Kumar A, Yadav S 2021 Catalytic activity of bis p-nitro A_2B (oxo)Mn(V) corroles towards oxygen transfer reaction to sulphides *Inorg. Chim. Acta* **514** 120013
 47. (a) Gross Z, Galili N, Saltsman I 1999 The First Direct Synthesis of Corroles from Pyrrole *Angew. Chem. Int. Ed.* **38** 1427; (b) Paolesse R, Jaquinod L, Nurco D J, Mini S, Sagone F, Boschi T and Smith K M 1999 5,10,15-Triphenylcorrole: a product from a modified Rothmund reaction *Chem. Commun.* **14** 1307; (c) Paolesse R, Nardis S, Sagone F and Khoury R G 2001 Synthesis and Functionalization of meso-Aryl-Substituted Corroles *J. Organomet. Chem.* **66** 550
 48. Santos CIM, Oliveira E, Lodeiro JF, Barata JFB, Santos SM, Faustino MAF, et al. 2013 Corrole and Corrole Functionalized Silica Nanoparticles as New Metal Ion Chemosensors: A Case of Silver Satellite Nanoparticles Formation *Inorg. Chem.* **52** 8564
 49. (a) Ventura B, Degli Esposti A, Koszarna B, Gryko D T and Flamigni L 2005 Photophysical characterization of free-base corroles, promising chromophores for light energy conversion and singlet oxygen generation *New J. Chem.* **29** 1559; (b) Ding T, Alem E A, Mordarelli D A and Ziegler C J 2005 Photophysical Properties of a Series of Free-Base Corroles *J. Phys. Chem. A* **109** 7411; (c) Paolesse R 2000 Synthesis and application of corroles In *The Porphyrin Handbook* K M Kadish, K M Smith and R Guilard (Eds.) Vol 2 (Academic Press: New York) ch. 11 p. 202
 50. (a) Metrangolo P, Meyer F, Pilati T, Resnati G and Terraneo G 2008 Halogen Bonding in Supramolecular Chemistry *Angew. Chem. Int. Ed.* **47** 6114; (b) Politzer P, Murray J S and Clark T 2010 Halogen bonding: an electrostatically-driven highly directional noncovalent interaction *Phys. Chem. Chem. Phys.* **12** 7748; (c) Parisini E, Metrangolo P, Pilati T, Resnati G and Terraneo G 2011 Halogen bonding in halocarbon-protein complexes: a structural survey *Chem. Soc. Rev.* **40** 2267; (d) Erdelyi M 2012 Halogen bonding in

- solution *Chem. Soc. Rev.* **41** 3547; (e) Cavallo G, Metrangolo P, Milani R, Pilati T, Priimagi A, Resnati G and Terraneo G 2016 The Halogen Bond *Chem. Rev.* **116** 2478
51. Mahammed A, Weaver J J, Gray H B, Abdelas M and Gross Z 2003 How acidic are corroles and why? *Tetrahedron Lett.* **44** 2077
 52. McClure D S 1952 Spin-Orbit Interaction in Aromatic Molecules *J. Chem. Phys.* **20** 682
 53. (a) Chae M Y and Czarnik A W 1992 Fluorometric chemodosimetry. Mercury(II) and silver(I) indication in water via enhanced fluorescence signalling *J. Am. Chem. Soc.* **114** 9704; (b) Rurack K, Kollmannsberger M, ReschGenger U and Daub J 2000 A Selective and Sensitive Fluoroionophore for Hg^{II} , Ag^{I} , and Cu^{II} with Virtually Decoupled Fluorophore and Receptor Units *J. Am. Chem. Soc.* **122** 968; (c) Prodi L, Bargossi C, Montalti M, Zaccheroni N, Su N, Bradshaw J S, Izatt R M and Savage P B 2000 An Effective Fluorescent Chemosensor for Mercury Ions *J. Am. Chem. Soc.* **122** 6769; (d) Moon S Y, Cha N R, Kim Y H and Chang S-K 2004 New Hg^{2+} -Selective Chromo- and Fluoroionophore Based upon 8-Hydroxyquinoline *J. Org. Chem.* **69** 181; (e) Moon S-Y, Yoon N J, Park S M and Chang S-K 2005 Diametrically Disubstituted Cyclam Derivative Having Hg^{2+} -Selective Fluoroionophoric Behaviors *J. Org. Chem.* **70** 2394
 54. Fraiji L K, Hayes D M and Werner T C 1992 Static and dynamic fluorescence quenching experiments for the physical chemistry laboratory *J. Chem. Educ.* **69** 424
 55. (a) Sharma P and Mehata M S 2020 Colloidal MoS_2 quantum dots based optical sensor for detection of 2,4,6-TNP explosive in an aqueous medium *Optic. Mater.* **100** 109646; (b) Sharma P and Mehata M S 2020 Rapid sensing of lead metal ions in an aqueous medium by MoS_2 quantum dots fluorescence turn-off *Mater. Res. Bull.* **131** 110978; (c) Sharma V and Mehata M S 2021 Rapid optical sensor for recognition of explosive 2,4,6-TNP traces in water through fluorescent ZnSe quantum dots *Spectrochim. Acta Part A* 119937; (d) Sharma V and Mehata M S 2021 Synthesis of photoactivated highly fluorescent Mn^{2+} -doped ZnSe quantum dots as effective lead sensor in drinking water *Mater. Res. Bull.* **134** 111121
 56. D'Souza F, Chitta R, Ohkubo K, Tasior M, Subbaiyan N K, Zandler M E *et al.* 2008 Corrole–Fullerene Dyads: Formation of Long-Lived Charge-Separated States in Nonpolar Solvents *J. Am. Chem. Soc.* **130** 14263
 57. Clark D T, Murrell J N and Tedder J M 1963 The magnitudes and signs of the inductive and mesomeric effects of the halogens *J. Chem. Soc.* 1250
 58. (a) Choi K and Hamilton A D 2001 A Dual Channel Fluorescence Chemosensor for Anions Involving Intermolecular Excited State Proton Transfer *Angew. Chem. Int. Ed.* **40** 3912; (b) Wang Q, Xie Y, Ding Y, Lie X and Zhu W 2010 Colorimetric fluoride sensors based on deprotonation of pyrrole–hemiquinone compounds *Chem. Commun.* **46** 3669; (c) Amendola V, Fabbrizzi L and Mosca L 2010 Anion recognition by hydrogen bonding: urea-based receptors *Chem. Soc. Rev.* **39** 3889; (d) Ahmed N, Suresh V, Shirinfar B, Geronimo I, Bist A, Hwang I C and Kim K S 2012 Fluorogenic sensing of CH_3CO_2^- and H_2PO_4^- by ditopic receptor through conformational change *Org. Biomol. Chem.* **10** 2094

An Embodied Conversational Agent to Minimize the Effects of Social Isolation During Hospitalization

Full research paper

Jemma Smith

School of Computer Science
University of Technology Sydney
Sydney, Australia
Email: jemmasmith0305@gmail.com

Aashish Bhandari

School of Computer Science
Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
Email: aashish.bhandari2010@gmail.com

Berkan Yuksel

School of Computer Science
University of Technology Sydney
Sydney, Australia
Email: berkan.yuksel@alumni.uts.edu.au

A. Baki Kocaballi

School of Computer Science
University of Technology Sydney
Sydney, Australia
Email: baki.kocaballi@uts.edu.au

Abstract

Social isolation and loneliness contribute to the development of depression and anxiety. Comorbidity of mental health issues in hospitalized patients increases the length of stay in hospital by up to 109% and costs the healthcare sector billions of dollars each year. This study aims to understand the potential suitability of embodied conversational agents (ECAs) to reduce feelings of social isolation and loneliness among hospital patients. To facilitate this, a video prototype of an ECA was developed for use in single-occupant hospital rooms. The ECA was designed to act as an intelligent assistant, a rehabilitation guide, and a conversational partner. A co-design workshop involving five healthcare professionals was conducted. The thematic analysis of the workshop transcripts identified some major themes including improving health literacy, reducing the time burden on healthcare professionals, preventing secondary mental health issues, and supporting higher acceptance of digital technologies by elderly patients.

Keywords Embodied Conversational Agent, Personalization, Mental Health, Social Isolation, Loneliness, Co-design

1 Introduction

Pressures on resources in acute hospitals mean that the focus of attention is the primary illness or condition for which the patient has been admitted, and secondary concerns such as prevention of mental health deterioration may have less priority to be dealt with. Volunteer support workers offer some support but can be limited by the nature of the patient's condition and volunteer availability. Hospital staff generally do not have time to sit with patients and provide companionship, and therefore, there is often no emotional outlet for patients in long-term hospitalisation (Siddiqui et al. 2018). This affects not only the patients but also the staff and the healthcare sector with longer lengths of stay, increased admissions, and the higher level of care needed by socially isolated patients costing the USA Medicare system an extra \$6.7USD Billion per year (Shaw et al. 2017). Recently, the COVID-19 pandemic has further compounded social isolation for inpatients due to greatly reduced visitation (Vlake et al. 2021).

Recent studies have analysed potential applications of conversational agent (CA) technology for patients with mental illnesses such as depression, anxiety (Fitzpatrick et al. 2017), and PTSD (Vlake et al. 2021). None of this research has explored the application of CA technology in a preventative capacity for people currently in hospital due to other medical problems. Therefore, the current study focuses on understanding the suitability of using a CA in hospital rooms to minimise the effects of social isolation experienced by the long stay patients.

1.1 The Effects of Hospitalisation on Patients' Mental Health

Walker et al. (2018) conducted a systematic review to analyse the prevalence of depression in general hospital patients and found that the prevalence of depression was between 5% and 34%, with an average of 12. They concluded that screening for depression should therefore occur in all hospitalised patients. (Purssell et al. 2020) in their systematic review found a higher correlation between anxiety (1.28 times higher) and depression (1.45 times higher) in isolated patients compared to non-isolated patients. The findings of these two reviews suggest that social isolation may be a primary contributory factor in mental health decline in hospitalised patients.

Purssell et al. (2020) observed that infectious patients have higher rates of depression and anxiety due to their separation from others. The COVID-19 pandemic has provided an unprecedented amount of data available for analysis due to the high rates of patients requiring isolated hospitalisation and a high degree of post-discharge care. Vlake et al. (2021) found that even three months after discharge, patients were still experiencing decreased mental state, with 13% having probable PTSD, 20% having probable anxiety and 24% having probable depression. While this data is useful in identifying the potential impact of isolated hospitalisation, there exists possible confounding influences on the results, such as social stigma and worry for the safety of close contacts. It is therefore not feasible to only rely on data collected during the COVID-19 pandemic as the healthcare environment differs significantly from pre- and post-COVID-19.

The main cause of mental illness in the older population is existential loneliness and isolation. Sundström et al. (2019) examined the cause of existential loneliness in older people in residential care, palliative care, hospital care and in-home care. They concluded that creating relationships with people in care needs to be an important part of a healthcare professional's role and can mitigate some of the impacts of existential loneliness and decrease the length of stay for the patients. The problem with this conclusion is that it assigns all the responsibility to a healthcare professional who often does not have time to create meaningful relationships with patients, as their primary role is to care for the physical needs of all people in their care.

It has been seen by an early pilot study by Schubert et al. (1992) that the length of stay (LOS) in the hospital increased with higher Geriatric Depression Scale (GDS) scores. This study concluded that depression could delay medical recovery from physical illness and complicate discharge planning. Siddiqui et al. (2018) critically analysed the effect mental illness has on the LOS of patients in hospitals and found that comorbidity of mental illness unfavourably affected the LOS of hospital patients by up to 109%. This not only leads to higher healthcare spending (Shaw et al. 2017) but also increases the risk of mental health degradation. Similar to Sundström et al. (2019), the solutions proposed by the authors to combat the problem rely solely on already-existing systems, such as upskilling healthcare teams and improving LOS reporting, which have proven unhelpful in the past due to high pressures already placed on healthcare staff.

1.2 Conversational Agents used as Mental Health Support

Laranjo et al. (2018) reviewed 17 studies that focused on using unconstrained natural language input CAs for healthcare. They found that the most common conditions were related to mental health, and

CAs have the capability to support health across many application areas. (Laranjo et al. 2018) also found some evidence that using CAs reduced depression symptoms in the users and increased awareness of symptoms and triggers. Fitzpatrick et al. (2017) conducted a randomized controlled trial (RCT) delivering cognitive behavioural therapy (CBT) using a CA to young adults with depression and anxiety symptoms. In this study, the control group only received self-help information from an eBook about depression (Fitzpatrick et al. 2017). The group using the CA experienced significant improvement in their depression symptoms, whereas the control group did not. The authors concluded that CAs could be a “feasible, engaging and effective way to deliver CBT” (Fitzpatrick et al. 2017). These findings are supported by (Gaffney et al. 2019), who reported promising data that CAs can aid in the treatment of mental health problems. Contrary to Fitzpatrick et al. (2017), Gaffney et al. (2019) concluded that current study methods in this area are not robust enough to demonstrate the efficacy and efficiency of CAs. Gaffney recommends further research to demonstrate equivalence to other treatment methods to increase the validity.

To develop a CA for mental health treatment and recovery, it is important to analyse why patients find CA technology helpful. For a CA to appropriately respond to mental health disclosure statements, it may be necessary for these systems to have the ability to display sympathy and empathy towards the user. Liu and Sundar (2018) analysed and compared people’s reactions to different kinds of empathic expressions by CAs. They found that users (including those initially sceptical) preferred empathetic and sympathetic responses from an interactive CA over reading text lacking in emotion. These findings support the proposal that a CA with empathetic capabilities could be accepted by hospital patients who are typically older and maybe initially sceptical of CA technology.

In the literature, majority of studies examining the effects of CAs focus on the young adult population. Fitzpatrick’s RCT included people aged 18–28 years. Similarly, the studies discussed in the systematic reviews by Laranjo et al. (2018) and Gaffney et al. (2019) and the experiments run by Liu and Sundar (2018) did not report outcomes specific to older people. Bennion et al. (2020) evaluated the usability, acceptability, and effectiveness of CAs to facilitate problem-solving in older adults receiving mental health treatment. Their findings suggest that older adults have a high adherence rate to CA technology, with only 12% of participants discontinuing use during the trial. While this study involved small numbers and no control group, and while no consideration was given to the influence of natural recovery, the high adherence and the high problem-resolution rates observed provide good support for the potential use of CAs by elderly patients and people experiencing long-term hospitalisation.

In a study by Ring et al. (2015), a virtual companion was installed for a week in the homes of 14 elderly people to offer them ongoing social assistance via sympathetic feedback. The results from their study showed significant reductions in loneliness based on self-reported mood. Jegundo et al. (2020) also expressed that embodied conversational agents offer a viable method of delivering support care to senior citizens.

1.3 Conversational Personality and Embodiment of CAs

To improve the acceptability of CA technology, recent research has focused on the impact of human-like characteristics such as personality, tone of voice and visual avatars. Face-to-face interactions allow ECAs to develop an intimate, harmonious relationship with the user that fosters bonding (ter Stal et al. 2020a).

Wolters et al. (2016) found that CAs should be able to change the tone of voice and interaction methods based on the patient’s cognitive ability. These new features typify the Embodied Conversational Agent (ECA), a CA with added human-like features such as avatars and gestures. Several prior research supports the adoption of animations of the agent’s embodiment, showing that animations favourably impact users’ perceptions of the agent and interaction time (ter Stal et al. 2020a).

Valtolina and Hu (2021) created and tested the validity and acceptability of using a customised CA called “Charlie” as a companion for elderly people experiencing social isolation and loneliness in their homes. The study findings show very high acceptability among the elderly users who viewed Charlie as a polite, intelligent, reliable, and helpful companion, and there was a notable decrease in the loneliness experienced by the user. These results provide a basis for future research into companion-type ECAs to mitigate feelings of isolation and loneliness.

Philip et al. (2017) studied using ECAs in patients with major depressive disorders. The participants were all patients at a sleep clinic. They found that ECAs effectively communicate empathy without judgement, which gains the patients’ trust and leads to self-disclosure of sensitive information. Ho et al. (2018) showed that CAs could inspire self-disclosure at a similar if not higher rate than humans. Ermolina and Tiberius (2021) found during their Delphi study that current commercial Voice Controlled Personal Assistants (VIPAs) with human-like characteristics are not technologically developed enough

to act as healthcare assistants safely. They did, however, conclude that with further development, VIPA technology can be an effective tool to support staff and patients, especially elderly patients, in a healthcare environment.

1.4 Summary

Increased length of hospital stay can be linked to comorbidity of mental illness (Schubert et al. 1992), which further prolongs recovery. Current strategies for mitigating this effect, such as staff training and reporting, have proven unsuccessful. A more effective and cost-efficient approach is required to address the precursors of mental illness (such as social isolation). Current literature shows supporting evidence for including CAs in mental health treatment, tracking and recovery. Older adults may gain the same, if not more, benefits as younger people from using CAs in a health context (Bennion et al. 2020; Crabb et al. 2012). Older adults have a high adherence rate to CA technology and gain similar emotional, psychological, and relational benefits to younger people. Broadly, some improvement in emotional distress has been demonstrated through CAs.

Although Bickmore et al. (2018) and Kocaballi et al. (2020) warn of the dangers of commercially available technologies in a mental health context, there is emerging evidence that newer ECA technologies can effectively support people in a companionship role (Valtolina and Hu 2021; Wolters et al. 2016) and prevent the development of more serious mental illness.

The rest of this article is organized as follows: The method of the co-design workshop is discussed in section 2. Section 3 provides insights into the results of the workshop. The discussion, including findings, limitations, and future work, has been presented in section 4. Finally, section 5 concludes the study.

2 Method

The study employed a co-design workshop method with healthcare professionals, using a video prototype of an ECA demonstrating some key features. The workshop included a demonstration of the video prototype, construction of a mind map, imagination, and discussion of potential solutions. Thematic analysis was performed to analyse the workshop transcripts. The video prototype was built with an online video production tool called Animaker. Using a video prototype allowed for the demonstration of the functionality of the ECA while eliminating the complexity of developing a fully functional prototype. The following subsections will present each phase in detail.

2.1 Prototype Design

The prototype included four design features: Embodiment, Proactivity, Personalisation and Rehabilitation. In addition, there were some safety considerations.

2.1.1 Embodiment

The CA was designed to be an ECA (Embodied Conversational Agent), which incorporates human-like characteristics in the form of an avatar - named "Jackie". Since no guidelines exist for the appropriate appearances of ECAs (ter Stal et al. 2020b), we selected Jackie and her appearance as generic as possible. By providing the avatar with a human-like resemblance, the interface of our agent feels more connected than a physical device.

Jackie displays emotions through facial expressions and tone of voice. She has a full range of facial expressions, four of which are displayed in Figure 1. She can alter her tone of voice when discussing different subjects to create a human-like interaction. Jackie can also perform basic movements and actions such as waving, sitting, running, and jumping and exercises such as squats and bicep curls. The name "Jackie" was chosen for the sole reason that people tend to trust names that are easy to pronounce and familiar (Newman et al. 2014).

2.1.2 Proactivity

Jackie can "listen" to the sounds in the room and initiate a conversation with the patient if there is no sign of visitors or if the patient is asleep. She begins light conversations with patients with conversation starters.

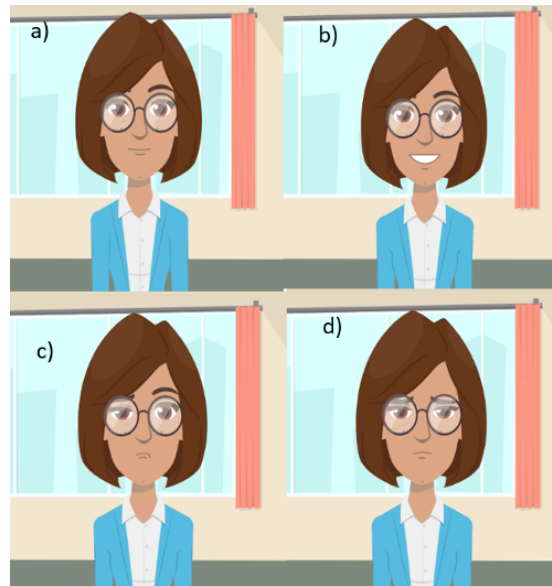


Figure 1: Jackie's facial expression: a) Resting, b) Happy, c) Confused and d) Sad

2.1.3 Personalisation

Jackie can be programmed with personal information about the patient's life, such as the names of family members and friends, their careers, and significant life events. These can be manually input into the system during admission or at an appropriate time when Jackie is being introduced to the patient. Jackie can be programmed with basic information about the patient's condition, if applicable, to aid in health literacy. For example, if a patient has a knee replacement, Jackie can provide the patient with pre-approved generic information about the procedure, the recovery, and the expected outcomes.

2.1.4 Rehabilitation

As some rehabilitation exercises can be done without a healthcare professional's aid, Jackie can motivate and guide the patient through pre-programmed rehabilitation exercises approved by their healthcare staff. She can remind the patient that it is time to do their activities, encourages them through every step, and gathers information about their pain and mobility to be sent to their healthcare professional.

2.2 Sample Conversation

Three short sample conversations were created for the video prototype, and one introduction video would be played for patients when introduced to Jackie. These conversation helps generate further discussion. The three conversations made were:

1. A short casual conversation about a recent visit from a family member, which demonstrates Jackie's ability to remember information told to her regarding specific family members and when they were planning to visit,
2. An instructional conversation that took the patient through their rehabilitation exercises, and
3. A more profound conversation where the patient self-discloses their worries about going home and being isolated. This highlighted how Jackie could encourage patients to talk about emotional topics they may be too embarrassed to initiate with a healthcare staff member.

2.3 Co-design workshop and Survey

A co-design workshop with five healthcare professionals with many years (10-39) of clinical experience from a range of disciplines was conducted to evaluate and envision the design elements of Jackie. Table 1 shows the workshop participants' details. The participants were recruited through personal and professional connections. Eligibility criteria included 5+ years working in a professional healthcare setting and experience caring for patients in long-term hospitalisation. All participants were sent an information sheet about the project and the proposed solution. A short survey was sent out with questions relating to their current job, relevant work experience in healthcare, experience with witnessing social isolation in healthcare settings and initial reactions and reservations to the proposed

ECA solution. The answers to these questions informed other discussion topics in the co-design workshop.

Participant (P)	Healthcare role	Areas of Experience (years of experience)
P1	Occupational Therapist	Rehabilitation, cardiology, and orthopaedics (35)
P2	Speech Pathologist	Neurology, aged care, and rehabilitation (22)
P3	Registered Nurse	Neurology, palliative care, aged care (10)
P4	General Practitioner	Nursing home and aged care home visits (39)
P5	Registered Nurse	Rehabilitation, Drug health recovery (20)

Table 1. Co-design workshop participants

The co-design workshop was held online through video chat and a collaborative brainstorming tool called Mind Meister. The first mind map aimed to gather information about current social isolation experiences and mitigation strategies that healthcare facilities may have. The workshop then moved into solution ideation, where participants individually provided feedback on the proposed solution and offered solutions to problems identified in the first section. This stage also focused on any negative impacts the ECA may have on patients and any considerations to ensure patient safety. Then, the participants were shown the video prototype and were asked to discuss its features. The workshop was audio recorded. Following the co-design workshop, the audio recording was transcribed and analysed. Thematic analysis (Braun and Clarke 2006) was used to generate sub-themes and themes.

3 Results

Responses to survey questions from the five workshop participants were analysed, along with responses extracted through workshop participation. Inductive coding of themes identified 15 codes from the raw data (survey responses and workshop quotes). The analysis generated three main and 15 sub-themes (Table 2).

3.1 Current healthcare environment

The participants' comments indicated that current isolation mitigation strategies used in hospitals have limited effectiveness. Language barriers, limited visiting hours, busy staff and insufficient volunteer resources were all identified as barriers to success. P3 noted that patients often wander the ward looking for social interaction: *"Visiting hours can be very limiting. So overnight, people can become very isolated. We often have cases of people wandering the hallways trying to talk to people ... and there's just no facilities available through a night shift to help someone who ... just feels very alone."* In contrast, P1 noted that some patients will avoid talking to healthcare staff and volunteers: *"Often patients don't want to be 'a bother' – so they don't talk, and they don't even talk to the people who are there to listen and help."* Some participants explained that the only current mitigation strategies are visitors, volunteers, and socialisation between patients. P1 highlighted that visitors are not an option for all patients: *"[It is] hard for rural patients or those without visitors."* P3 described that socialisation between patients is often not possible: *"High doses of medications can make it difficult for patients to mobilise or concentrate."* Furthermore, P4 stated that none of the mitigation strategies is possible if the patient must be physically isolated: *"[Physical] Isolation [may be necessary] for medical reasons, e.g., infectious disease, immunosuppressed"*. Finally, P4 brought up that current admission procedures do not measure a patient's level of loneliness or potential for social isolation: *"[social isolation] is not systematically assessed."*

3.2 The ECAs' potential impact on patients, staff, and the healthcare system

The participants' comments suggest that a companion tool such as Jackie has the potential to be integrated into some healthcare settings. Although some comments refer to the demonstrated features of ECA including proactivity, personalisation, and rehabilitation, the participants' focus remained mostly on the overall role of such an interactive technology as a new actor.

Current healthcare environment	The ECAs' potential impact on patients, staff, and the healthcare system	Safety and privacy factors to address before implementation
Medical staff time (1)	Medical staff time (1)	Medical staff time (1)
Shared rooms (2)	Shared rooms (2)	Shared rooms (2)
Issues current patients experience (3)	Features of ECA/Things it could do (9)	Safety concerns of ECA (13)
Current methods to solve issues (4)	Health areas of most help (10)	Data Privacy (14)
Systematic assessment of isolation (5)	Technology (11)	Challenges (15)
Visitors (6)	COVID-19 (12)	
Language (7)	Language (7)	
Shared patient areas (8)		

Table 2. Thematic analysis theme groups

During a discussion about when a patient may instigate a conversation with Jackie, P1 noted that patients often have questions about their condition that they may not ask the healthcare staff: “... patients don’t always ask questions when they are told something about their condition – but they may have questions that they want to ask later – and they don’t know who/when to ask.”

Improving physical health was deemed an important factor in shortening the length of stay for people in hospital. P3 noted that this technology could be a positive encouragement and aid rehabilitation for patients: “Thinking about stroke rehab, there’s definitely a place for something like this that’s set up by the physio, encourages people, reminds them, tracks their progress.” P1 added that this technology could also improve patients’ health literacy: “and it will all build health literacy which I just think is a huge potential for this.” P3 stated that people of all ages can learn to use new technology if they are provided with help and support while they are learning: “My experience working with people in neuro and aged care settings is that people of all ages can learn to use technology with the right supports.”

During the workshop, it was found that the acceptability of new technology is high among patients and staff. During a conversation about current technology, P2 stated, “Allied Health already use iPads.” To which P3 replied, “And older people love their iPads.” Followed by P2 adding, “They just think they’re the coolest thing ever.” These quotes show an optimistic view of such technologies’ potential to positively impact patients, staff, and the health system, by providing as-needed companionship, supporting rehabilitation, and reducing the burden on busy staff.

4 Discussion

Our participants’ comments were overall positive about the suitability of ECAs for the patients, staff, and healthcare system. This aligns with data found in the current literature, which shows that, with the introduction of additional safety measures, CAs could help improve the physical and mental health of patients in the hospital, decrease their length of stay, free up nursing staff time, and reduce healthcare spending (Bennion et al. 2020; Fitzpatrick et al. 2017; Gaffney et al. 2019; Shaw et al. 2017; Siddiqui et al. 2018; Vaidyam et al. 2021). Current research suggests that commercial CA technology is not yet safe enough for use in situations where the patient expresses mental health issues or crisis statements (Bickmore et al. 2018; Ermolina and Tiberius 2021; Kocaballi et al. 2020); however, some recent CAs appear to be providing a much safer and more improved experience for the users (Bennion et al. 2020; Gaffney et al. 2019; Philip et al. 2017).

The main aim of hospitalisation (in general medical wards) is to focus on the patient’s physical health and get them physically well enough to go home. Mental health tends to be a consideration only once symptoms have emerged. Nurses and other healthcare professionals have limited capacity to meet the companionship needs that could mitigate social isolation and subsequent mental health deterioration. Our literature review and the outcomes of the co-design workshop suggest that psychosocial assessment is not considered at the time of general hospital admission. Nor is there a systematic process for measuring patients’ level of social isolation throughout their hospital stay. During the co-design

workshop, the participants were asked to identify measures currently being used to prevent and/or mitigate social isolation and loneliness. Even though healthcare staff is aware that social isolation is a problem, almost nothing is being done to address it.

The purpose of this study was to explore an ECA solution across the "normal" healthcare environment (i.e., pre- and post-COVID-19 pandemic); however, data collected during the pandemic has proven useful in demonstrating the importance of providing alternative companionship methods to isolated patients, with the mental health of patients coming out of hospital after being treated for COVID-19 dangerously low (Vlake et al. 2021). As it was observed by a healthcare professional that patients are reluctant to talk to those who are there to listen and assist, the proactivity feature of the agent can aid in encouraging interaction. In the case of infectious diseases where physical isolation is a must, using a virtual agent can be advantageous.

The participant's feedback on our video prototype is aligned with the outcomes reported by (Valtolina and Hu 2021), (Bennion et al. 2020) and (Wolters et al. 2016) that ECA technology can support people in an emotional capacity and relieve feelings of isolation and loneliness. Therefore, using ECAs in actual hospital settings may prove useful to provide companionship, aid recovery, and shorten the length of stay. However, larger-scale interventions with fully functional ECAs should be conducted to evaluate the efficacy and subjective user experience.

Although none of the co-design workshop participants has experience in hospital or healthcare sector financial management, they agreed that the ECA could shorten the length of stay for hospital patients by targeting social isolation, which would positively impact health sector expenditure. There was also group consensus that the additional features of the ECA, such as education, reminders, and encouragement to perform rehabilitation exercises, could improve health literacy, increase mood and motivation to recover and consequently decrease the length of stay in the hospital.

An important topic of discussion during the co-design workshop was the point at which a medical professional must take over the conversation between the ECA and the patient. Healthcare staff has limited time to interact with patients, so the initial aim was to have the ECA responsible for all mental health conversations, as literature shows that the use of CAs for mental health tracking and recovery can have positive outcomes (Fitzpatrick et al. 2017; Gaffney et al. 2019; Laranjo et al. 2018). Despite this, the participants of the co-design workshop were wary of the technology handling safety-critical conversation topics. To reduce the uncertainty, there was an agreement amongst the participants regarding the disclosure of safety-critical statements; the ECA can notify staff, and a mental health professional can take over. This may help staff identify patients who potentially require additional support.

All the participants in the co-design workshop had questions regarding whether the ECA is safe to use for the patient's physical and mental health. These are the same questions asked in the observational study conducted by Bickmore et al. (2018), which found that commonly available multi-purpose CAs cannot yet be relied on for actionable medical advice. Once the safety and security risks of the technology were discussed in the workshop, the ECAs' potential role switched from a "therapist" providing advice to a hospital companion providing friendship and motivation, with a focus on decreasing social isolation and preventing mental health decline, aligned with the research conducted by Valtolina and Hu (2021), and Wolters et al. (2016).

4.1 Limitations

This study has some limitations. First, the prototype used in the workshop was a video prototype. Although it demonstrated a high-fidelity prototype in action, it did not provide the participants with the first-person experience of using a fully working product. Second, our co-design workshop only included healthcare professionals. Another workshop with patients with prior long-term hospital stay experience and other types of healthcare professionals would complement our study. Third, all the participants were from a similar age group (40 to 60 years old). Therefore, it is possible that this limited variation in technological proficiency and understanding.

4.2 Future work

This study was a proof-of-concept study, obtaining a preliminary understanding of the potential suitability of using ECAs during hospitalisation. Future research can focus on developing a fully working ECA with unconstrained natural language input capabilities for increased accessibility and usability. The ECA should undergo rigorous risk analysis and safety testing before the interventions and should be monitored throughout to ensure the ECA does not give responses that could worsen the physical or mental condition of the patient. To ensure the implementation of the ECA does not hinder the healthcare

staff's ability to provide the necessary care to their patients, future development should ensure that the ECA can be set up easily and the patient's data can be input without posing an imposition to the staff.

While not explored in this study, an opportunity for future development is integrating this technology with Wi-Fi-enabled medical equipment to aid patients' physical health. Voice-activated ECAs could provide an accessibility solution for patients with limited mobility and hand function. This has the potential to create hospital rooms that cater to a much more comprehensive range of conditions and disabilities.

5 Conclusion

The evidence shows that long-term hospital stays can lead to feelings of social isolation and loneliness, significantly contributing to depression and anxiety. Mental health issues such as these have negatively impacted people's physical well-being. Prior studies have investigated if CAs can be helpful for people already experiencing mental illness; however, there has been limited research into the potential application of CAs in a preventative capacity, especially among long-stay patients. Current mitigation methods have limited ineffectiveness, and an additional focus on preventive measures is needed. Our workshop study suggests that ECAs can be helpful technological solutions during hospitalisation. However, privacy and safety remain major concerns similar to other CA applications in healthcare.

6 References

- Bennion, M.R., Hardy, G.E., Moore, R.K., Kellett, S., and Millings, A. 2020. Usability, Acceptability, and Effectiveness of Web-Based Conversational Agents to Facilitate Problem Solving in Older Adults: Controlled Study. *Journal of Medical Internet Research* 22, e16794.
- Bickmore, T.W., Trinh, H., Olafsson, S., O'leary, T.K., Asadi, R., Rickles, N.M., and Cruz, R. 2018. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *Journal of Medical Internet Research* 20, e11510.
- Braun, V., and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 77-101.
- Crabb, R.M., Cavanagh, K., Proudfoot, J., Learmonth, D., Rafie, S., and Weingardt, K.R. 2012. Is computerised cognitive-behavioural therapy a treatment option for depression in late-life? A systematic review. *British Journal of Clinical Psychology* 51, 459-464.
- Ermolina, A., and Tiberius, V. 2021. Voice-Controlled Intelligent Personal Assistants in Health Care: International Delphi Study. *Journal of Medical Internet Research*, (23), e25312.
- Fitzpatrick, K.K., Darcy, A., and Vierhile, M. 2017. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR mental health* 4, e19.
- Gaffney, H., Mansell, W., and Tai, S. 2019. Conversational Agents in the Treatment of Mental Health Problems: Mixed-Method Systematic Review. *JMIR Mental Health* 6, e14166.
- Ho, A., Hancock, J., and Miner, A.S. 2018. Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations with a Chatbot. *The Journal of Communication* 68, 712-733.
- Jegundo, A.L., Dantas, C., Quintas, J., Dutra, J., Almeida, A.L., Caravau, H., Rosa, A.F., Martins, A.I. and Pacheco Rocha, N., 2020. Perceived usefulness, satisfaction, ease of use and potential of a virtual companion to support the care provision for older adults. *Technologies*, 8(3), p.42.
- Kocaballi, A.B., Quiroz, J.C., Rezazadegan, D., Berkovsky, S., Magrabi, F., Coiera, E., and Laranjo, L. 2020. Responses of Conversational Agents to Health and Lifestyle Prompts: Investigation of Appropriateness and Presentation Structures. *Journal of Medical Internet Research* 22, e15823.
- Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y. and Coiera, E., 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9), pp.1248-1258.
- Liu, B., and Sundar, S.S. 2018. Should Machines Express Sympathy and Empathy? Experiments with a Health Advice Chatbot. *Cyberpsychology, Behavior, and Social Networking* 21, 625-636.

- Newman, E.J., Sanson, M., Miller, E.K., Quigley-Mcbride, A., Foster, J.L., Bernstein, D.M., and Garry, M. 2014. People with Easier to Pronounce Names Promote Truthiness of Claims. *PLOS ONE* 9, e88671.
- Philip, P., Micoulaud-Franchi, J., Sagaspe, P., Sevin, E.D., Olive, J., Bioulac, S., and Sauteraud, A. 2017. Virtual human as a new diagnostic tool, a proof-of-concept study in the field of major depressive disorders. *Scientific Reports* 7, 42656.
- Purssell, E., Gould, D. And Chudleigh, J. 2020. Impact of isolation on hospitalised patients who are infectious: systematic review with meta-analysis. *BMJ open* 10, e030371.
- Ring, L., Shi, L., Totzke, K. and Bickmore, T., 2015. Social support agents for older adults: longitudinal affective computing in the home. *Journal on Multimodal User Interfaces*, 9(1), pp.79-88.
- Schubert, D.S.P., Burns, R., Paras, W. And Sioson, E. 1992. Increase of Medical Hospital Length of Stay by Depression in Stroke and Amputation Patients: A Pilot Study. *Psychotherapy and Psychosomatics* 57, 61-66.
- Shaw, J.G., Farid, M., Noel-Miller, C., Joseph, N., Houser, A., Asch, S.M., Bhattacharya, J. And Flowers, L. 2017. Social Isolation and Medicare Spending: Among Older Adults, Objective Social Isolation Increases Expenditures while Loneliness Does Not. *Journal of Aging and Health* 29, 1119-1143.
- Siddiqui, N., Dwyer, M., Stankovich, J., Peterson, G., Greenfield, D., Si, L. And Kinsman, L. 2018. Hospital length of stay variation and comorbidity of mental illness: a retrospective study of five common chronic medical conditions. *BMC Health Services Research* 18.
- Sundström, M., Blomqvist, K., Edberg, A. And Rämgård, M. 2019. The context of care matters: older people's existential loneliness from the perspective of healthcare professionals—A multiple case study. *International Journal of Older People Nursing* 14, e12234.
- ter Stal, S., Broekhuis, M., van Velsen, L., Hermens, H. and Tabak, M., 2020a. Embodied conversational agent appearance for health assessment of older adults: explorative study. *JMIR human factors*, 7(3), p.e19987.
- ter Stal, S., Kramer, L.L., Tabak, M., op den Akker, H. and Hermens, H., 2020b. Design features of embodied conversational agents in eHealth: a literature review. *International Journal of Human-Computer Studies*, 138, p.102409.
- Vaidyam, A.N., Linggonegoro, D. And Torous, J. 2021. Changes to the Psychiatric Chatbot Landscape: A Systematic Review of Conversational Agents in Serious Mental Illness: Changements du paysage psychiatrique des chatbots: une revue systématique des agents conversationnels dans la maladie mentale sérieuse. *The Canadian Journal of Psychiatry* 66, 339-348.
- Valtolina, S. And Hu, L. 2021. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness.
- Vlake, J.H., Wesselius, S., Van Genderen, M.E., Van Bommel, J., Boxma-De Klerk, B. And Wils, E. 2021. Psychological distress and health-related quality of life in patients after hospitalisation during the COVID-19 pandemic: A single-center, observational study. *PLoS ONE* 16.
- Walker, J., Burke, K., Wanat, M., Fisher, R., Fielding, J., Mulick, A., Puntis, S., Sharpe, J., Esposti, M.D., Harriss, E., Frost, C. And Sharpe, M. 2018. The prevalence of depression in general hospital inpatients: a systematic review and meta-analysis of interview-based studies. *Psychological Medicine* 48, 2285-2298.
- Wolters, M.K., Kelly, F. And Kilgour, J. 2016. Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. *Health Informatics Journal* 22, 854-866.

Analyze the SATCON Algorithm's Capability to Predict Tropical Storm Intensity across the West Pacific Basin

Monu Yadav¹ and Laxminarayan Das^{1*}

^{1*}Department of Applied Mathematics, Delhi Technological
University, , New Delhi, India.

*Corresponding author(s). E-mail(s): lnidas@dce.ac.in;
Contributing authors: yadavm012@gmail.com;

Abstract

A group of algorithms for estimating the current intensity (CI) of tropical cyclones (TCs), which use infrared and microwave sensor-based images as the input of the algorithm because it is more skilled than each algorithm separately, are used to create a technique to estimate the TC intensity which is known as SATCON . In the current study, an effort was undertaken to assess how well the SATCON approach performed for estimating TC intensity throughout the west pacific basin from year 2017 to 2021. To do this, 26 TCs over the west pacific basin were analysed using the SATCON-based technique, and the estimates were compared to the best track predictions provided by the Regional Specialized Meteorological Centre (RSMC), Tokyo. The maximum sustained surface winds (Vmax) and estimated central pressures (ECP) for various “T” numbers and types of storm throughout the entire year as well as during the pre-monsoon (March-July) and post-monsoon (July-February) seasons have been compared. When compared to weaker and very strong TCs, the ability of the SATCON algorithm to estimate intensity is determined to be rather excellent for mid-range TCs. We demonstrate that SATCON is more effective in the post-monsoon across the west pacific basin than in the pre-monsoon by comparing the algorithm results.

Keywords: SATCON algorithm, West pacific basin, Estimated intensity,

1 Introduction

Tropical cyclone (TC) observation by meteorological satellites has largely reduced the challenge of detection. A constellation of geostationary (GEO) and polar-orbiting platforms regularly scans the tropics, and sensors with better spatial and spectrum sampling are used. Numerous types of multispectral photography can be used to qualitatively track and record the location, genesis, occurrence, and dissipation of TCs. Estimating the present TC intensity from space-based platforms is a little more complicated. It is possible to perform subjective analysis of TC cloud patterns using infrared (IR) images employing trained analysts and empirically supported guidelines. In order to analyse the CI and anchor TC intensity catalogues (or “best tracks”) in the absence of in situ intensity observations, operational TC centres have depended extensively on the time-tested Dvorak technique [1, 2] for many years. As demonstrated by crowdsourcing techniques, even inexperienced analysts can fairly accurately estimate the CI [3]. The inherent subjectivity in the interpretation of the images and restrictions on the capacity to detect structured convective structure beneath the normally massive and dense TC cirrus canopy, however, pose difficulties to IR-based cloud pattern recognition techniques [4, 5]. Techniques that make use of cloud-penetrating microwave (MW) sensors [6–9] can be helpful in this area, but they also have drawbacks.

Obtaining accurate CI estimations is crucial for various reasons: The operational TC forecast process begins with the CI; it is one of the key input variables required to initialise both dynamical and statistical TC forecast models; and it is crucial for understanding TC climatologies and trends to have precise best-track intensities [10]. Forecasters (or best-track analysts) frequently struggle with the issue of competing satellite-based CI estimations with significant spread/uncertainty. Taken as a solution, a common conservative strategy is to average the estimations (simple consensus). A “smarter” consensus procedure, depending on the situational performance of each consensus attribute, is preferred since it further minimises the CI estimate uncertainty.

Multiple satellite fitted sensor respond data based observation techniques are combined into an ensemble model known as SATCON created by Cooperative Institute for Meteorological Satellite Studies (CIMSS). Below are basic explanations of SATCON’s methodology.

Each attribute has situational strengths and weaknesses, which are represented by their separate intensity estimation error distributions, from which the individual attribute weights utilised in the SATCON process of finding a CI are created. Therefore, each attribute’s performance behaviour can be categorised into situational bins. For example, using the IR images of the scene type, the ADT technique [11] estimates the intensity of TC. When the eye scene is clear, it provides the best estimation; nevertheless, if it is not clear, the estimation is poor or the outcome may be compromised. To best weight all of the available intensity estimates into a single, superior consensus estimate,

SATCON uses this situational information. The two TC intensity measurements, MSLP and MSW, each have distinct performance traits, leading to various SATCON weighting algorithms for each metric.

Sharing information between sensors is another aspect of the SATCON process. Each SATCON attribute contains distinctive parametric data that the other coinciding attribute might use to evaluate the situational bins and conceivably modify the intensity estimates. For example, when an eye is observable in the IR, then ADT generates estimations of TC eye size [12].

The Automated Tropical Cyclone Forecasting system (ATCF; [13]) can be utilised by operational TC centres to provide additional sources of input to the SATCON process. These sources can include storm motion and the environmental pressure used in the pressure > wind member. For storms that significantly differ from an average TC motion of roughly 11 kts ($1kts \approx 0.51ms^{-1}$), the methodology from [14] can be used to make minor modifications to the final predicted MSW values.

In the current work, the authors made an effort to evaluate the SATCON algorithm's performance in estimating TCs intensity throughout the west pacific basin by contrasting it with the parameters supplied by RSMC Tokya. Section 2 describes the data and technique used based on the statistics of 2017-2021. Section 3 discusses the findings, while section 4 enumerates the study's conclusions.

2 Data and Methodology

We verified the "SATCON" output data by comparing it to the data provided by RSMC, Tokya of TC intensity for all TCs between 2017 and 2021 over the West Pacific (particularly taking into account these storms that affect Japan). The SATCON algorithm's data collects from UW-CIMSS (<http://tropic.ssec.wisc.edu/misc/satcon>) and RSMC provided data collects from the RSMC, Tokyo (jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/RSMC_HP.htm), to determine the optimal track parameters for TCs. The number of TC cases included in the study is listed in Table 1 [15–19]. 26 TCs are therefore investigated in the current study.

Multiple satellite fitted sensor respond data based observation techniques are combined into an ensemble model known as SATCON. Which is the studies of Geostationary-based Advanced Dvorak Technique and the Passive Microwave signal based advance sounding and imaging unit designed by the CIMSS. It provides a consensus intensity estimation of TCs across all the basins. It makes use of a statistically determined weighting system that maximises (minimises) to be evaluated consensus intensity for a variety of TC structures (weaknesses). The intensity computation is built from a series of formulae dependent on the number of attribute available, and the SATCON weights are proportional to the RMSE attribute values for the selected scenarios.

The three-part equation for SATCON [10] is

Table 1: In this study, following TCs were considered over the West Pacific Basin

SI. No.	Cyclone Name	Season	Date	Maximum Wind Speed (knots)
2021				
1	Surigae	Pre-Monsoon	12-30 Apr.	120 kts
2	IN-FA	Pre-Monsoon	15-30 Jul.	85 kts
3	Chanthu	Post-Monsoon	05-20 Sept.	115 kts
4	Rai	Post-Monsoon	11-21 Dec.	105 kts
2020				
1	Vongfone	Pre-Monsoon	08-18 May	85 kts
2	Maysak	Post-Monsoon	27 Aug. - 07 Sept.	95 kts
3	Haishen	Post-Monsoon	30 Aug. - 10 Sept.	105 kts
4	Goni	Post-Monsoon	26 Oct. - 06 Nov.	115 kts
5	Molave	Post-Monsoon	22-29 Oct.	90 kts
2019				
1	Nari	Pre-Monsoon	24-28 Jul.	35 kts
2	Danas	Pre-Monsoon	14-23 Jul.	45 kts
3	Lekima	Post-Monsoon	02-15 Aug.	105 kts
4	Wutip	Post-Monsoon	08 Feb. - 02 Mar.	105 kts
5	Hagibis	Post-Monsoon	04-14 Oct.	105 kts
6	Halong	Post-Monsoon	01-10 Nov.	115 kts
2018				
1	Jelawat	Pre-Monsoon	24 Mar. - 01 Apr.	105 kts
2	Prapiroon	Pre-Monsoon	28 Jun. - 05 Jul.	65 kts
3	Maria	Pre-Monsoon	03-13 Jul.	105 kts
4	Shanshan	Post-Monsoon	02-11 Aug.	70 kts
5	Trami	Post-Monsoon	20 Sept. - 03 Oct.	105 kts
6	Kong-Rey	Post-Monsoon	28 Sept. - 07 Oct.	115 kts
2017				
1	Noru	Pre-Monsoon	19 July-12 Aug.	95 kts
2	Talim	Post-Monsoon	08-22 Sept.	95 kts
3	Sanvu	Post-Monsoon	26 Aug. - 06 Sept.	80 kts
4	Lan	Post-Monsoon	15-23 Oct.	100 kts
5	Hato	Post-Monsoon	19-24 Aug.	75 kts

$$SATCON = \frac{W_1 W_2 (W_1 + W_2) E_3 + W_1 W_3 (W_1 + W_3) E_2 + W_2 W_3 (W_3 + W_2) E_1}{W_1 W_2 (W_1 + W_2) + W_1 W_3 (W_1 + W_3) + W_3 W_2 (W_3 + W_2)}$$

where E_n is the attribute n's intensity estimations and W_n is the attribute n's weight (RMSE). The weights of attributes 1, 2, and 3 are W_1 , W_2 , and W_3 , and the intensity estimations of attributes 1, 2, and 3 are E_1 , E_2 , and E_3 .

The situational RMSE values for each of the attributes used to calculate the intensity estimate are known as attribute weights. The SATCON weighting structure's composition is intended to give more weight to a situational dependent attribute with the highest efficiency (among the available attributes). For instance, the equation above shows how higher RMSEs (weights) of E_1 and E_2 are added to E_3 . Thus adding greater weight to the specific estimation E_3 , if

E_3 is the best-performing attribute in a given context. For those more uncertain estimates, less weight (relatively smaller RMSEs) is to be allotted (E_1 and E_2) [10].

One of the finest methods for forecasting the TC intensity over the Atlantic and North Indian Oceans is the SATCON.

To evaluate the accuracy of the intensity forecasting and the effectiveness of the CIMSS-SATCON algorithm, 26 TCs are used to validate the method (table 1). Comparison of RSMC, Tokyo (jma.go.jp/jma/jma-eng/jma-center/rsmc-hp-pub-eg/RSMC_HP.htm) provided intensity estimation data with SATCON intensity estimates.

Between the estimation of ECP and Vmax based on RSMC, Tokya provided data, and SATCON calculation, various variables, which are root mean square difference (RMSD), actual mean difference (bias), and mean absolute difference (MAD), are determined. These variables are estimated for the various stages of a TC's "T" number, as specified in the RSMC, Tokyo-provided intensity data, inside each three-hourly observation that is at 00, 03, 06, 09, 12, 15, and 21 UTC throughout the whole time period of a TC. The mean MAD, RMSD, and bias of intensity estimations across the West Pacific basin are estimated for various "T" numbers during the different seasons as well as for the entire year based on all TCs taken into account. The student's t-test is used to determine whether there are any significant differences between the mean values over the West Pacific basin during the pre- and post-monsoon seasons.

When compared to the information provided by RSMC, Tokyo, the capability of SATCON has also been evaluated for various stages of TCs. Table 2 displays the various TC stages used in RSMC, Tokyo.

Stage	Maximum Sustained Wind (knots; kts)
Tropical Storm	34-48 kts
Severe Tropical Storm	48-64 kts
Typhoon	64-85 kts
Very Strong Typhoon	85-105 kts
Violent Typhoon	105-130 kts

Table 2: Different stage of TCs with maximum sustained wind used in RSMC, Tokya

3 Results and discussion

3.1 Over the period of entire year, the capability of the "SATCON" algorithm over Japan (West Pacific)

3.1.1 Capability of the SATCON algorithm for various "T" number stages

Table 3 compares the capaability of SATCON TC MSW and MSLP calculation to intensity estimation data provided by RSMC for TCs across the west Pacific

Table 3: Calculated different parameters (in terms of MSW and MSLP) based on SATCON and RSMC, Tokya provided data for TCs during the year 2017-2021

Best track T No.	Total no. of cases	Best track intensity range	Best track intensity (MSW) (A)	SATCON intensity range	SATCON intensity MSW (B)	BIAS (A-B)	Mean absolute difference	RMSD
2.0	94	30-40	35	40-67	50.75	-15.75	15.4	17.98
2.5	99	40-50	42.68	40-70	52.97	-10.29	10.81	13.57
3.0	97	50-60	52.89	45-74	57.5	-4.61	5.13	12.18
3.5	166	60-70	62.65	46-92	8.21	-5.56	7.71	13.03
4.0	114	70-75	71.84	62-98	76.11	-4.27	10.08	12.83
5.0	123	75-80	80	69-100	83.47	-3.47	8.37	10.09
5.5	133	80-95	86.85	73-114	88.36	-1.51	6.93	11.72
6.0	98	95-105	97.5	90-122	105.04	-7.54	11.05	10.38
6.5	43	105-115	105.7	112-138	125.15	-19.45	20.18	21.76
7.0	16	115-125	115.93	122-144	136.47	-20.57	21.06	22.85
MSLP (hPa)								
Best track T No.	Total no. of cases	Best track intensity range	Best track intensity (MSLP) (A)	SATCON intensity range	SATCON intensity MSLP (B)	BIAS (A-B)	Mean absolute difference	RMSD
2.0	21	996-1003	1005.85	984-1004	1002.03	3.82	4.92	6.28
2.5	35	983-1002	972.54	981-1004	970.59	1.95	4.01	6.83
3.0	158	983-996	971.86	980-1000	969.5	2.36	5.58	7.09
3.5	116	978-991	970.58	967-999	966.34	4.24	7.16	9.43
4.0	104	975-988	964.67	957-994	959.5	5.17	9.78	11.13
5.0	82	964-982	988.18	950-980	981.43	6.75	8.16	9.72
5.5	45	955-964	962.45	939-976	956.38	6.07	9.09	10.73
6.0	36	946-956	958.57	934-970	955.36	3.21	7.89	8.29
6.5	22	932-946	898.11	922-940	891.69	6.42	8.95	8.47
7.0	17	922-930	865.35	920-935	867.83	-2.48	6.72	7.43

basin during 2017-2021. The bias progressively declines as the “T” number rises, but it gradually rises after T5.5, being roughly 16-10 knots (kts) for T2.0-T2.5, approximately 6-4 kts for T3.0-T5.0, and about 2 kts for T5.5. Due to the small sample size, the results for T6.0-T7.0 bias increasing with increasing in T number from roughly 8 to 21 kts may not be indicative. According to the student’s t-test for the “T” numbers T2.0-T5.5 and T6.0-T7.0, the difference is significant with a 99% level of confidence.

After the T3.5, the MAD is approximately 10-14 kts, and the MAD is approximately 12-16 kts for T2.0-T3.0. Due to the small sample size, the higher MAD value in the T6.5 range could not be indicative. The MAD values for T5.0 and above across the west Pacific (7-10 kts) are consistent with [10, 20] observations.

As a consequence, the intensity is estimated to be overestimated (negative bias) by approximately 2 hPa for T7.0, approximately 2-5 hPa for T2.0-T3.5 and approximately 5-7 hPa for more than T3.5. For the range of T2.0-T6.5, the underestimate is statistically significant at a 99% level of confidence. For T2.0-T2.5, the MAD is approximately 5 hPa, and for T3.0-T7.0, it is approximately 5-10 hPa. For T2.0-T3.0, the RMSD is about 6-7 hPa, and for T3.5-T7.0, it is approximately 8-11 hPa.

3.1.2 Results of the SATCON technique for the various TC categories

The SATCON method and the intensity estimation data provided from the RSMC, Tokyo were used to analyse the average characteristics of tropical storms to violent typhoons over the west pacific basin between the year 2017-2021 in terms of the MSW (knots) and MSLP (hPa). It demonstrates that as TC goes to a higher category, the bias steadily reduces, being around 10-8 knots for a tropical storm to a severe tropical storm and 8-4 knots for a severe tropical storm to very strong typhoon, and violent typhoon. Due to the small sample size, the bias value for the violent typhoon category, 12.21 kts, may not be indicative (table 4). Accordingly, the bias is reduced for stronger TCs, with the exception of violent typhoons, which is consistent with [10, 20] findings. Although the MAD is for typhoons, severe typhoons, and tropical storms, approximately 9-12 kts, and approximately 7-11 kts for very strong and violent typhoons (figure 1). For all TC categories, the overestimation is statistically significant at a 99% level of confidence.

The MAD value for typhoons and intense typhoons (8-10 kts) is consistent with [10, 20]’s prior findings. A slight increase in MAD values of intensity over the west pacific basin is recorded for tropical storm, severe tropical storm and very strong typhoon compared with [10, 20] conclusion. For tropical storms, severe tropical storms, typhoons, and very strong typhoons, the RMSD values over the west Pacific are approximately 11-14, and for violent typhoons, they are less than 10 kts. The Violent typhoon RMSD estimates across the west pacific (<10 kts) are consistent with [10, 20] earlier observations. However, compared to [10, 20] findings, the RMSD values over the west pacific for

Table 4: Compared the SATCON and RSMC, Tokya provided data (in terms of MSW and MSLP) of TCs (stage wise) over the west pacific basin during the year 2017-2021

Category	Total no. of cases	Best track range	Best track MSW (A)	SATCON range	SATCON MSW (B)	BIAS (A-B)	Mean absolute difference	RMSD
Violent Typhoon	34	105-130	110.8	116-144	123.01	-12.21	7.43	9.24
Very Strong Typhoon	79	85-105	90.89	73-138	101.36	-10.47	10.97	12.5
Typhoon	167	64-85	72.23	62-100	84.07	-11.84	9.61	11.43
Severe Tropical Storm	108	48-64	55.73	46-92	68.21	-12.48	11.06	13.74
Tropical Storm	251	34-48	39.59	40-74	53.93	-14.34	10.89	12.96
MSLP (hPa)								
Category	Total no. of cases	Best track range	Best track MSLP (A)	SATCON range	SATCON MSLP (B)	BIAS (A-B)	Mean absolute difference	RMSD
Violent Typhoon	28	920-928	925.56	922-940	928.45	-2.89	6.94	7.24
Very Strong Typhoon	75	932-966	962.63	930-972	957.17	5.46	8.41	9.58
Typhoon	168	964-984	981.81	953-983	975.73	6.08	9.08	10.43
Severe Tropical Storm	106	983-990	980.58	965-990	976.3	4.28	7.83	9.12
Tropical Storm	267	983-1002	1003.29	984-1004	1002.03	1.26	6.14	7.01

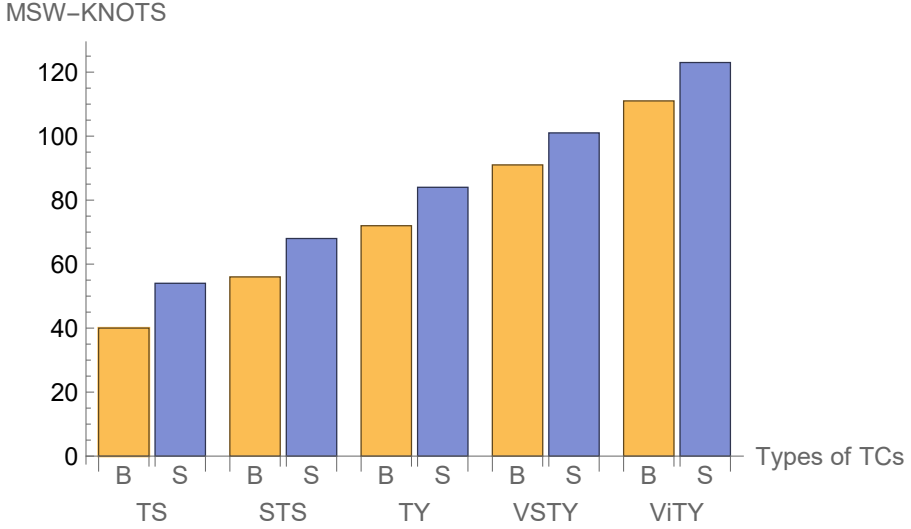


Fig. 1: Compared the intensity estimation by SATCON and RSMC provided data for all TCs(stahe wise) over the west pacific basin during the year 2017-2021. B stands for RSMC provided data, S stands for SATCON algorithm data, ViTY stands for Violent Typhoon, VSTY for Very Strong Typhoon, TY for Typhoon, STS for Severe Tropical Storm, and TS for Tropical Storm

the tropical storm, severe tropical storm, typhoon and very strong typhoon category are marginally greater (11-14 kts).

As a result, the intensity is understated (negative bias) in terms of MSLP by about -3 hPa for violent typhoons and 1-6 hPa for all other types of storms. For tropical storms, the MAD is roughly 6 hPa, and for all other storm types, it is between 7-9 hPa. All storm types have RMSD values between 7 and 11 hPa.

The SATCON algorithm shows the overestimation of the intensity of TCs during the begining stage of formation and up to T2.5, it may be seen from this. But after that, it is discovered that its performance is fairly good in measuring the intensity of stronger TCs (more than Severe Tropical storm). In the SATCON method, creates a single estimate from several TC intensity estimations derived from objective intensity algorithms. The major component of the SATCON model, ADT 9.0, feeds continuous inputs into the model every 30 minutes, whilst the microwave sounder satellite feeds irregular intensity inputs into the model, which are then extrapolated to hourly estimations. The final SATCON estimate is produced by combining these interpolated estimates with ADT estimates. An objective method evolved from the original Dvorak Technique is used by the ADT to calculate intensity. Up to T2.5 in the first development phase, the cloud organisation pattern is not clearly specified. At this time, the Dvorak Technique is unable to comprehend the intricate details of cloud patterns. Because of this, both the ADT 9.0 technique and SATCON

overstate the intensity estimations based on the methodology's pre-defined fixed cloud pattern, primarily the central dense overcast (CDO) and eye pattern, regardless of whether it is a curved band or shear pattern. It goes without saying that the shear pattern TCs have a maximum strength of T3.0 and that majority of the TCs over the west pacific originate from shear patterns under the influence of monsoon circulation. As a result, when the intensity is T2.0 or higher, the ADT 9.0 version and SATCON are utilised globally. Additionally, the SATCON algorithm is reasonable good for T3.0 and more because for TCs whose intensity is more than T3.0, they show clear cloud pattern i.e. either eye pattern or CDO. In addition, SATCON used the new ADT 9.0 methodology, which integrates infrared sensor, short-wave infrared imaging sensor, visible imaging sensor, and microwave images to find phenomena that the original Dvorak Technique was unable to find, such as secondary eye-wall formation, double eyewall structure, the centre in the presence of cirrus canopy, coiling of convective clouds (in the presence of cirrus) around the centre, and eye-wall replacement cycle [21].

Given the foregoing, forecasters can utilize the SATCON technique to estimate intensity in the case of stronger TCs (T3.0 or more). As cloud organization patterns are not clearly defined in the beginning stage, and the automated approach of ADT (a attribute of SATCON) selects pre-established patterns, overestimating of the intensity results, it is not suitable to cyclogenesis and the begining phase of TC formation. Forecasters can, however, accurately estimate TC intensities based on SATCON data by using the bias, RMSD, and MAD calculated in this study.

3.2 Capability of SATCON algorithm in various seasons

3.2.1 The pre-monsoon season's capabilities of the SATCON algorithm

SATCON TC MSW (kts) and MSLP (hPa) estimates' capability in comparison to RSMC Tokyo intensity estimate data for TCs developed over the west Pacific during the pre-monsoon are shown in tables 5 and 6. With the exception of very strong typhoons, the bias value stays high during all phases of TCs at roughly 11-18 kts. A very strong typhoon has a bias value of less than 1 kts (figure 2). The smaller sample size may be the cause of the very strong typhoon's unrepresentative value. According to the student's t-test for all types of TCs, the difference is significant at a 99% level of confidence.

For tropical storms, severe tropical storms, and typhoons, the MAD is approximately 12-19 kts, and for very strong typhoons, it is approximately 4 kts. The RMSD ranges between 13 and 19 kts for tropical storms, severe tropical storms, and typhoons, and between 4 and 5 kts for very strong typhoons. This runs counter to [10, 20] past findings. The average SATCON intensity, in turn, overestimates the MSW in the pre-monsoon season by around 11 kts and underestimates the average MSLP estimations by nearly 7 hPa, as demonstrated in Table 5.

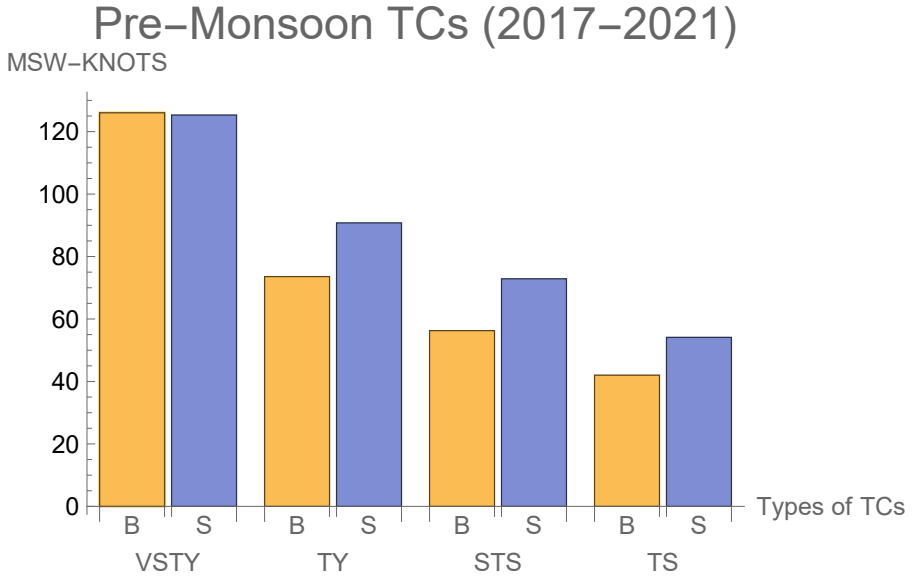


Fig. 2: Compared the intensity estimation by SATCON and RSMC provided data for pre-monsoon season. B stands for RSMC provided data, S stands for SATCON algorithm data, VSTY for Very Strong Typhoon, TY for Typhoon, STS for Severe Tropical Storm, and TS for Tropical Storm

3.2.2 The post-monsoon season’s capabilities of the SATCON algorithm

Tables 5 and 6 show the capability of SATCON’s algorithm of TCs MSW (kts) and MSLP (hPa) estimations compared to RSMC, Tokya provided data of intensity estimates for TCs developed across the west Pacific during the year 2017-2021’s post-monsoon. When the strength rises, the bias steadily decreases between 2 and 7 knots for tropical storms, severe typhoons, typhoons, and between 11 to 13 knots for extremely strong and violent typhoons (figure 3). The student’s t-test for all forms of TC indicates that the difference is significant at a 99% level of confidence. The bias value for the typhoon stage is consistent with [10, 20] past research.

The MAD for a tropical storm is approximately 11 knots, for TC categories such as a severe tropical storm, typhoon, very strong typhoon, and for violent typhoon, it is between 8 and 10 knots. The results of [10, 20] are supported by the MAD values throughout the west Pacific for severe tropical storms, typhoons, very strong typhoons, and violent typhoon stage (8-10 kts). However, compared to [10, 20] findings, the MAD values for the tropical storm stage are a little bit higher (11 kts). For all storm types, the RMSD across the west Pacific is approximately 10-14 kts. This runs counter to [10, 20] earlier research

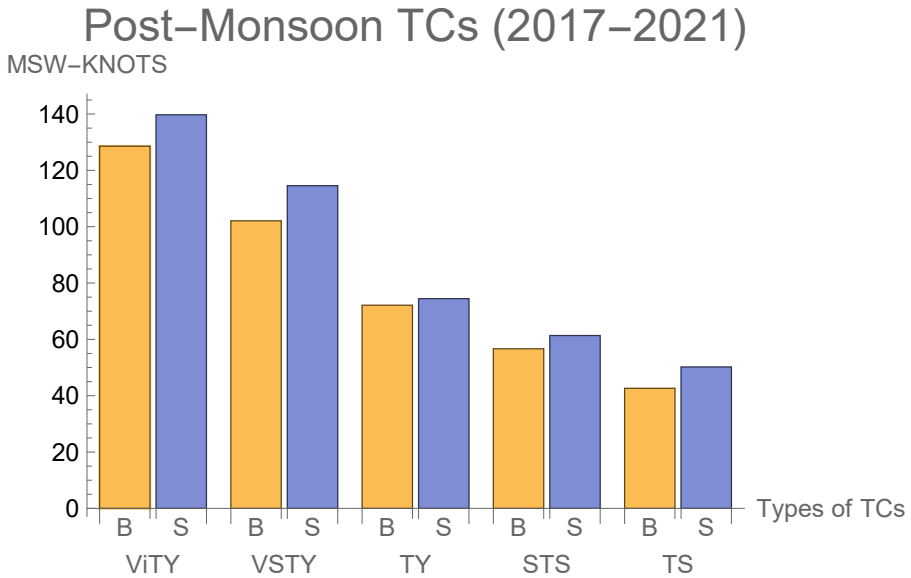


Fig. 3: Compared the intensity estimation by SATCON and RSMC provided data for post-monsoon season. B stands for RSMC provided data, S stands for SATCON algorithm data, ViTY stands for Violent Typhoon, VSTY for Very Strong Typhoon, TY for Typhoon, STS for Severe Tropical Storm, and TS for Tropical Storm

conclusions. Table 5 demonstrates that the average SATCON intensity underestimates the average MSLP estimates by around 5 hPa while overestimating the average MSW during the post-monsoon season by nearly 9 kts.

4 Conclusion

The key takeaways from the results and discussions above are listed below.

Tropical storm, severe tropical storm, typhoon, very strong typhoon, and violent typhoon types of TCs were examined in terms of intensity ('T' number) estimates across the west pacific basin from 2017 to 2021 using data from RSMC, Tokyo and the SATCON algorithm. As TCs progress through the initial phase of development, the range of overestimation of SATCON intensity estimation decreases. The result for T6.0-T7.0 may not be representative due to the sample size.

When we compared the SATCON algorithm's output with the data provided by the RSMC, we found that during the pre-monsoon, the SATCON algorithm overestimated tropical storms by about 13 kts, severe tropical storms by about 17 kts, and typhoons by about 19 kts. During the post-monsoon, the SATCON algorithm overestimated tropical storm by about 11 kts, and severe

Table 5: Compared the SATCON and RSMC, Tokya provided data (MSW/MSLP) for TCs over the west pacific basin during the year 2017-2021 as pre-monsoon, post-monsoon, and annual

Season	Total no. of cases	Best track MSW (A)	SATCON MSW (B)	BIAS (A-B)	Mean absolute difference	RMSD
Pre-Monsoon	158	64.85	75.59	-10.74	11.53	13.96
Post-Monsoon	308	59.51	66.26	-6.75	9.38	12.21
Annual Season	466	62.18	70.92	-8.74	10.83	13.07

MSLP (hPa)						
Season	Total no. of cases	Best track MSW (A)	SATCON MSW (B)	BIAS (A-B)	Mean absolute difference	RMSD
Pre-Monsoon	156	984.13	976.3	7.83	8.55	10.59
Post-Monsoon	359	959.39	958.16	1.23	5.34	7.24
Annual Season	515	971.76	967.23	4.53	6.86	8.79

Category of TC	Season	Best Track interval	Best Track MSW (A)	SATCON interval	SATCON MSW(B)	BIAS (A-B)	Mean Absolute difference	RMSD
Violent Typhoon	Pre-Mon	-	-	-	-	-	-	-
	Post-Mon	105-130	128.57	120-144	139.73	-11.16	8.68	13.34
Very strong Typhoon	Pre-Mon	80-110	126.03	90-120	125.30	0.73	4.26	4.92
	Post-Mon	85-105	102.08	112-132	114.57	-12.49	9.52	14.19
Typhoon	Pre-Mon	64-85	73.58	79-100	90.77	-17.19	18.44	18.24
	Post-Mon	65-90	72.12	72-98	74.46	-2.34	8.24	10.18
Severe Tropical Storm	Pre-Mon	48-64	56.25	69-81	72.87	-16.62	16.54	17.08
	Post-Mon	44-63	56.66	57-87	61.36	-4.69	9.24	12.47
Tropical Storm	Pre-Mon	34-48	42.01	16-70	54.11	-12.1	12.19	13.48
	Post-Mon	30-50	42.62	39-74	50.19	-6.92	10.58	12.84

Table 6: Compared the SATCON and RSMC, Tokya provided data for TCs (stage wise) over the west pacific basin as pre-monsoon and post-monsoon during the year 2017-2021

tropical storm, typhoon, very strong typhoon and violent typhoon by about 9kts.

We demonstrate that SATCON is more effective in the post-monsoon across the west pacific basin than in the pre-monsoon by comparing the algorithm results.

Acknowledgements

The RSMC Tokya and CIMSS-SATCON are thanked by the authors for providing the information used in this article. The authors appreciate the anonymous peer reviewers’ insightful criticism, which helped the paper’s quality.

Author Statement

Monu Yadav: Conceptualization, investigation, data curation, methodology, validation, preparation of tables/figures. Laxminarayan Das: Supervision, reviewing and editing.

References

[1] Dvorak, V.F.: Tropical cyclone intensity analysis and forecasting from satellite imagery. Monthly Weather Review **103**(5), 420–430 (1975). [https://doi.org/10.1175/1520-0493\(1975\)103\(0420:TCIAAF\)2.0.CO;2](https://doi.org/10.1175/1520-0493(1975)103(0420:TCIAAF)2.0.CO;2)

[2] Dvorak, V.F.: Tropical Cyclone Intensity Analysis Using Satellite Data vol. 11. US Department of Commerce, National Oceanic and Atmospheric Administration ..., ??? (1984)

- [3] Hennon, C.C., Knapp, K.R., Schreck, C.J., Stevens, S.E., Kossin, J.P., Thorne, P.W., Hennon, P.A., Kruk, M.C., Rennie, J., Gad  a, J.-M., Striegl, M., Carley, I.: Cyclone center: Can citizen scientists improve tropical cyclone intensity records? *Bulletin of the American Meteorological Society* **96**(4), 591–607 (2015). <https://doi.org/10.1175/BAMS-D-13-00152.1>
- [4] Velden, C., Harper, B., Wells, F., Beven, J.L., Zehr, R., Olander, T., Mayfield, M., Guard, C.C., Lander, M., Edson, R., Avila, L., Burton, A., Turk, M., Kikuchi, A., Christian, A., Caroff, P., McCrone, P.: The dvorak tropical cyclone intensity estimation technique: A satellite-based method that has endured for over 30 years. *Bulletin of the American Meteorological Society* **87**(9), 1195–1210 (2006). <https://doi.org/10.1175/BAMS-87-9-1195>
- [5] Knaff, J.A., Brown, D.P., Courtney, J., Gallina, G.M., Beven, J.L.: An evaluation of dvorak technique–based tropical cyclone intensity estimates. *Weather and Forecasting* **25**(5), 1362–1379 (2010). <https://doi.org/10.1175/2010WAF2222375.1>
- [6] Brueske, K.F., Velden, C.S.: Satellite-based tropical cyclone intensity estimation using the noaa-klm series advanced microwave sounding unit (amsu). *Monthly Weather Review* **131**(4), 687–697 (2003). [https://doi.org/10.1175/1520-0493\(2003\)131<0687:SBTCIE>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<0687:SBTCIE>2.0.CO;2)
- [7] Demuth, J.L., DeMaria, M., Knaff, J.A., Haar, T.H.V.: Evaluation of advanced microwave sounding unit tropical-cyclone intensity and size estimation algorithms. *Journal of Applied Meteorology* **43**(2), 282–296 (2004). [https://doi.org/10.1175/1520-0450\(2004\)043<0282:EOAMSU>2.0.CO;2](https://doi.org/10.1175/1520-0450(2004)043<0282:EOAMSU>2.0.CO;2)
- [8] Bankert, R., Cossuth, J.: Tropical cyclone intensity estimation via passive microwave data features. 32nd Conf. on Hurricanes and Tropical Meteorology, San Juan, PR, Amer. Meteor. Soc., 10C. 1 (2016)
- [9] Jiang, H., Tao, C., Pei, Y.: Estimation of tropical cyclone intensity in the north atlantic and northeastern pacific basins using trmm satellite passive microwave observations. *Journal of Applied Meteorology and Climatology* **58**(2), 185–197 (2019). <https://doi.org/10.1175/JAMC-D-18-0094.1>
- [10] Velden, C.S., Herndon, D.: A consensus approach for estimating tropical cyclone intensity from meteorological satellites: Satcon. *Weather and Forecasting* **35**(4), 1645–1662 (2020)
- [11] Olander, T.L., Velden, C.S.: The advanced dvorak technique (adt) for estimating tropical cyclone intensity: Update and new capabilities. *Weather and Forecasting* **34**(4), 905–922 (2019). <https://doi.org/10.1175/>

WAF-D-19-0007.1

- [12] Kossin, J.P., Knaff, J.A., Berger, H.I., Herndon, D.C., Cram, T.A., Velden, C.S., Murnane, R.J., Hawkins, J.D.: Estimating hurricane wind structure in the absence of aircraft reconnaissance. *Weather and Forecasting* **22**(1), 89–101 (2007). <https://doi.org/10.1175/WAF985.1>
- [13] Sampson, C.R., Schrader, A.J.: The automated tropical cyclone forecasting system (version 3.2). *Bulletin of the American Meteorological Society* **81**(6), 1231–1240 (2000). [https://doi.org/10.1175/1520-0477\(2000\)081\(1231:TATCFS\)2.3.CO;2](https://doi.org/10.1175/1520-0477(2000)081(1231:TATCFS)2.3.CO;2)
- [14] Schwerdt, R.W., Ho, F.P., Watkins, R.R.: Meteorological criteria for standard project hurricane and probable maximum hurricane windfields, gulf and east coasts of the united states (1979)
- [15] Tokyo, R.: Annual report on activities of the rsmc tokyo- typhoon center 2021. RSMC Tokyo (2021)
- [16] Tokyo, R.: Annual report on activities of the rsmc tokyo- typhoon center 2020. RSMC Tokyo (2020)
- [17] Tokya, R.: Annual report on activities of the rsmc tokyo- typhoon center 2019. RSMC Tokyo (2019)
- [18] Tokya, R.: Annual report on activities of the rsmc tokyo- typhoon center 2018. RSMC Tokyo (2018)
- [19] Tokya, R.: Annual report on activities of the rsmc tokyo- typhoon center 2017. RSMC Tokyo (2017)
- [20] Herndon, D., Velden, C.: An update on the CIMSS Satellite Consensus (SATCON) tropical cyclone intensity algorithm. 33rd Conf. on Hurricanes and Tropical Meteorology, Ponte Verdi, FL, Amer. Meteor. Soc., 284 (2018)
- [21] Olander, T.L., Velden, C.S.: The advanced dvorak technique (adt) for estimating tropical cyclone intensity: Update and new capabilities. *Weather and Forecasting* **34**(4), 905–922 (2019)

An Emotion-guided Approach to Domain Adaptive Fake News Detection using Adversarial Learning (Student Abstract)

Arkajyoti Chakraborty^{*1}, Inder Khatri^{*1}, Arjun Choudhry^{*1}, Pankaj Gupta¹, Dinesh Kumar Vishwakarma¹, Mukesh Prasad²

¹ Biometric Research Laboratory, Delhi Technological University, New Delhi, India

² School of Computer Science, FEIT, University of Technology Sydney, Sydney, Australia

{arkajyotichakraborty_2k19ep022, dinesh}@dtu.ac.in, {inderkhatri999, choudhry.arjun}@gmail.com, mukesh.prasad@uts.edu.au

Abstract

Recent works on fake news detection have shown the efficacy of using emotions as a feature for improved performance. However, the cross-domain impact of emotion-guided features for fake news detection still remains an open problem. In this work, we propose an emotion-guided, domain-adaptive, multi-task approach for cross-domain fake news detection, proving the efficacy of emotion-guided models in cross-domain settings for various datasets.

Introduction

Over the years, our reliance on social media as an information source has increased, leading to an exponential increase in the spread of *fake news*. To counter this, researchers have proposed various approaches for fake news detection (FND). Models trained on one domain often perform poorly on datasets from other domains due to the domain shift (Figure 1(1)). Some works show the efficacy of domain adaptation for cross-domain FND by extracting domain-invariant features (Figure 1(2)) for classification (Zhang et al. 2020). However, adapting domains does not ensure that features in different classes align correctly across domains, which sometimes has a negative impact on performance. Some works have shown a correlation between fake news and their intrinsic emotions (Guo et al. 2019; Choudhry, Khatri, and Jain 2022) (Figure 1(3)), having successfully used it for fake news detection. However, these works are restricted to in-domain settings and don't consider cross-domain evaluation. We propose the use of emotion-guided multi-task models for improved cross-domain fake news detection, experimentally proving its efficacy, and present an emotion-guided domain adaptive approach for improved cross-domain fake news detection by leveraging better feature alignment across domains due to the use of emotion labels (Figure 1(4)).

Proposed Methodology

Datasets, Emotion Annotation & Preprocessing

We use the FakeNewsAMT & Celeb (Pérez-Rosas et al. 2018), Politifact¹, and Gossipcop² datasets. We annotate them with the core emotions from Ekman's (Ekman 1992)

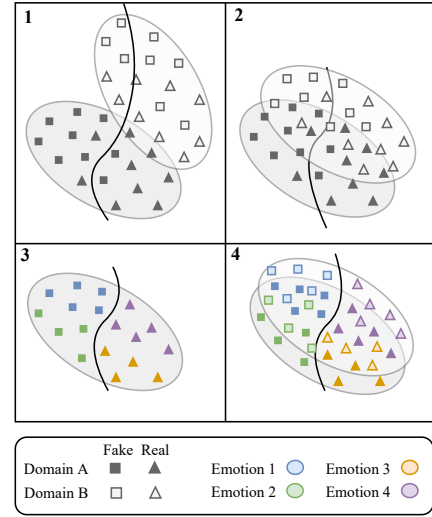


Figure 1: (1) Cross-domain texts not aligned. (2) Domain adaptation for improved alignment. (3) Emotion-guided classification. (4) Emotion-guided domain adaptation.

(6 emotions: *Joy, Surprise, Anger, Sadness, Disgust, Fear*) and Plutchik's (Plutchik 1982) (8 emotions: *Joy, Surprise, Trust, Anger, Anticipation, Sadness, Disgust, Fear*) emotion theories. We use the Unison model (Colneric and Demsar 2018) for annotating the datasets with emotion tags. During preprocessing, we convert text to lower case, remove punctuation, and decontract verb forms (eg. "I'd" to "I would").

Emotion-guided Domain-adaptive Framework

We propose the cumulative use of domain adaptation and emotion-guided feature extraction for cross-domain fake news detection. Our approach aims to improve the feature alignment between different domains using adversarial domain adaptation by leveraging the correlation between the emotion and the veracity of a text (as shown in Figure 1(4)). Figure 2 shows our proposed framework. We use an LSTM-based multi-task learning (MTL) feature extractor which

¹<https://www.politifact.com/>

²<https://www.gossipcop.com/>

^{*}These authors contributed equally.

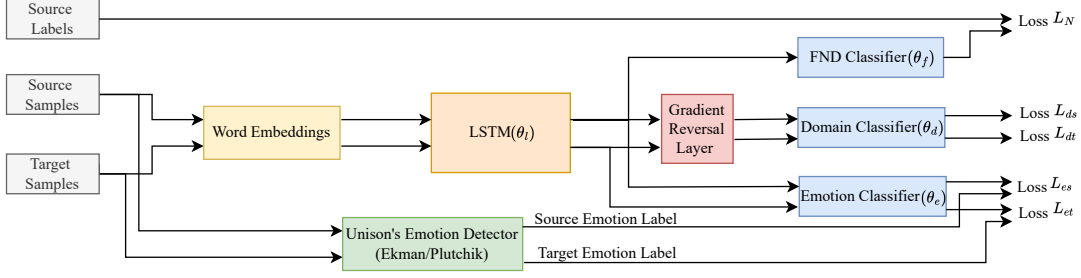


Figure 2: Graphical representation of our emotion-guided domain-adaptive framework for cross-domain fake news detection.

is trained by the cumulative losses from fake news classifier, emotion classifier, and discriminator (aids in learning domain-invariant features). LSTM can be replaced with better feature extractors. We use it specifically for easier comparison to non-adapted emotion-guided and non-adapted single-task models. The domain classifier acts as the discriminator. Fake news classification loss, emotion classification loss, adversarial loss, and total loss are defined as:

$$L_{FND} = \min_{\theta_l, \theta_f} \sum_{i=1}^{n_s} L_f^i \quad (1)$$

$$L_{emo} = \min_{\theta_l, \theta_e} \left(\sum_{i=1}^{n_s} L_{es}^i + \sum_{j=1}^{n_t} L_{et}^j \right) \quad (2)$$

$$L_{adv} = \min_{\theta_d} \left(\max_{\theta_l} \left(\sum_{i=1}^{n_s} L_{ds}^i + \sum_{j=1}^{n_t} L_{dt}^j \right) \right) \quad (3)$$

$$L_{Total} = (1-\alpha-\beta) * L_{FND} + \alpha * (L_{adv}) + \beta * (L_{emo}) \quad (4)$$

where n_s and n_t are number of samples in source and target sets; θ_d , θ_f , θ_e , and θ_l are parameters for discriminator, fake news classifier, emotion classifier, and LSTM feature extractor; L_{ds} and L_{dt} are binary crossentropy loss for source and target classification; L_{es} and L_{et} are crossentropy loss for emotion classification; L_f is binary crossentropy loss for Fake News Classifier; α and β are weight parameters in L_{Total} . We optimized α and β for each setting.

Experimental Results & Discussion

Each model used for evaluation was optimized on an in-domain validation set. Table 1 illustrates our results proving the efficacy of using emotion-guided models in non-adapted cross-domain settings. Table 2 compares non-adaptive models, domain adaptive models, and our emotion-guided domain adaptive models in various settings. MTL (E) and MTL (P) refer to emotion-guided multi-task frameworks using Ekman’s and Plutchik’s emotions respectively. STL refers to single-task framework. DA refers to domain-adaptive framework with a discriminator. Non-DA refers to a non-adapted model. Some findings observed are:

Emotions-guided non-adaptive multi-task models outperform their single-task counterparts in cross-domain settings, as seen in Table 1, indicating improved extraction of features that are applicable across different datasets.

Emotion-guided domain-adaptive models improve performance in cross-domain settings. Table 2 shows the advantage of emotion-guided adversarial domain-adaptive

Source	Target	Accuracy Non-DA STL	Accuracy Non-DA MTL(E)	Accuracy Non-DA MTL(P)
FAMT	Celeb	0.420	0.520	0.530
Celeb	FAMT	0.432	0.471	0.476

Table 1: Cross-domain evaluation of non-adaptive models on FakeNewsAMT (FAMT) & Celeb datasets. Emotion-guided models (MTL (E) and MTL (P)) outperform their corresponding STL models in cross-domain settings.

Source	Target	Accuracy Non-DA STL	Accuracy DA STL	Accuracy DA MTL(E)	Accuracy DA MTL(P)
FAMT	Celeb	0.420	0.560	0.540	0.600
Celeb	FAMT	0.432	0.395	0.501	0.551
Politi	Gossip	0.527	0.585	0.698	0.671
Celeb	Gossip	0.488	0.525	0.555	0.587
FAMT	Gossip	0.451	0.790	0.805	0.795
FAMT	Politi	0.363	0.621	0.704	0.621

Table 2: Cross-domain evaluation of non-adaptive, adaptive and emotion-guided adaptive models on various datasets.

models over their non-adaptive counterparts. This shows the scope for improved feature extraction even after adversarial adaptation, and emotion-guided models act as a solution.

References

- Choudhry, A.; Khatri, I.; and Jain, M. 2022. An Emotion-Based Multi-Task Approach to Fake News Detection (Student Abstract). *AAAI*, 36(11).
- Colneric, N.; and Demsar, J. 2018. Emotion Recognition on Twitter: Comparative Study and Training a Unison Model. *IEEE Transactions on Affective Computing*, 11(3).
- Ekman, P. 1992. An argument for basic emotions. *Cognition & Emotion*, 6.
- Guo, C.; Cao, J.; Zhang, X.; Shu, K.; and Yu, M. 2019. Exploiting Emotions for Fake News Detection on Social Media. *ArXiv*, abs/1903.01728.
- Pérez-Rosas, V.; Kleinberg, B.; Lefevre, A.; and Mihalcea, R. 2018. Automatic Detection of Fake News. In *COLING. ACL*.
- Plutchik, R. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5).
- Zhang, T.; Wang, D.; Chen, H.; Zeng, Z.; Guo, W.; Miao, C.; and Cui, L. 2020. BDANN: BERT-Based Domain Adaptation Neural Network for Multi-Modal Fake News Detection. In *IJCNN*.



ANN prediction approach analysis for performance and emission of antioxidant-treated waste cooking oil biodiesel

N. Kumar¹ · K. Yadav^{1,2} · R. Chaudhary³

Received: 9 August 2022 / Revised: 11 October 2022 / Accepted: 7 November 2022

© The Author(s) under exclusive licence to Iranian Society of Environmentalists (IRSEN) and Science and Research Branch, Islamic Azad University 2022

Abstract

For the purpose of lowering hazardous emissions and enhancing performance of diesel engine, waste cooking oil biodiesel has emerged as a feasible and promising biofuel. In this research paper, 300 and 400 ppm doses of tert-butylhydroquinone (TBHQ) and diphenylamine (DPA) antioxidants were added to waste cooking oil biodiesel of 20% volume to evaluate performance and emission parameters in unmodified diesel engine. An artificial neural network model was developed to predict brake thermal efficiency (BTE), brake specific energy consumption (BSEC), nitrogen oxide emission (NO_x), carbon monoxide emission (CO), hydrocarbon emission (HC), and smoke opacity by considering load, blends, and type of antioxidant in different doses as input. Prediction and validation were carried out using the findings of the experiments. The quasi-Newton method algorithm was used to predict data that best fits with linear regression analysis. The result showed at full load, BTE and BSEC have R^2 values of 0.985 and 0.995, respectively. The recommended ANN model's accuracy and performance were acceptable. At full load, the brake thermal efficiency increased, and brake specific energy consumption was reduced for fuel blend with antioxidant in respect of without antioxidant blend. NO_x emission was reduced by 2.32, 5.24, 7.35, and 12.44% for 300-doses DPA blend, 300-doses TBHQ blend, 400 doses TBHQ blend, and 400 doses DPA antioxidant blend, respectively, compared to without antioxidant blend. The adoption of ANN to predict performance and emission can speed up and lower the running cost of understanding output behavior.

Editorial responsibility: Shahid Hussain.

✉ K. Yadav
khushbu.mnnit@gmail.com

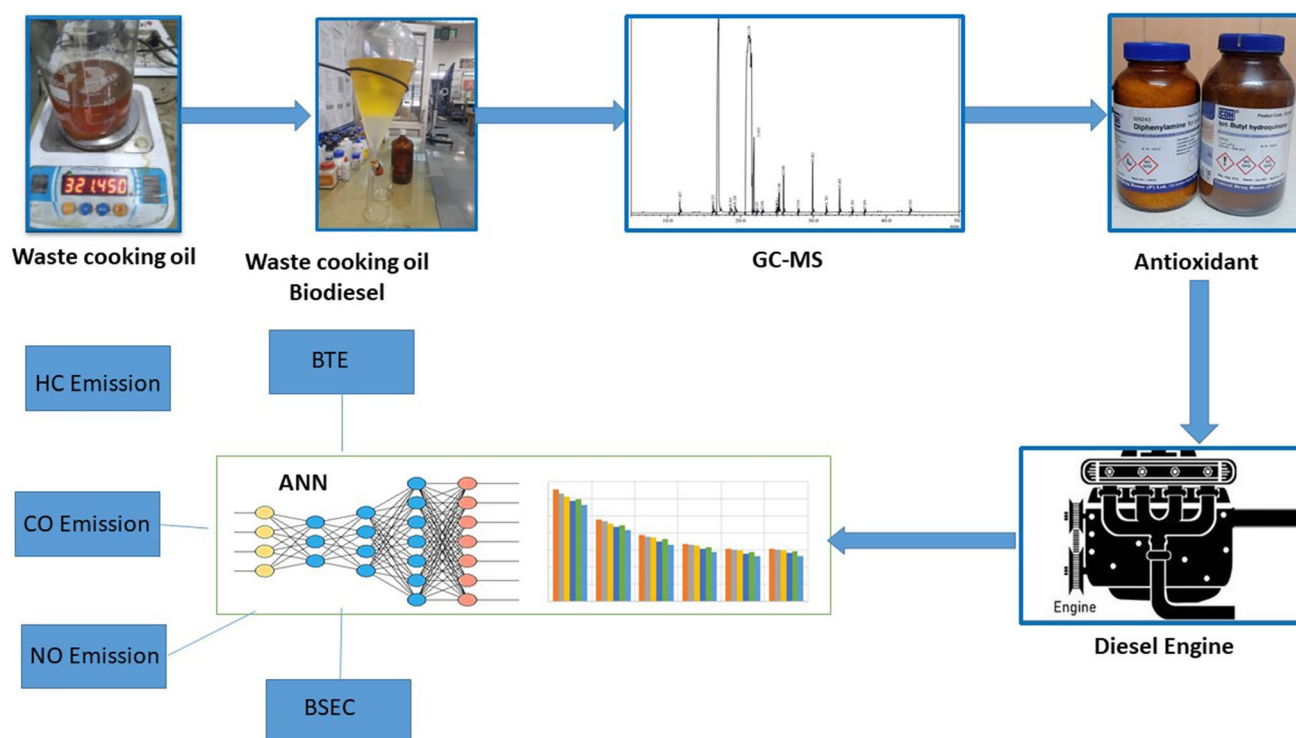
¹ Centre for Advanced Studies and Research in Automobile Engineering, Delhi Technological University, Delhi 110042, India

² Mechanical Engineering Department, Amity University, Uttar Pradesh, Noida 201301, India

³ Department of Mechanical Engineering, Delhi Technological University, Delhi 110042, India



Graphical Abstract



Keywords Artificial neural network · Biodiesel · Characterization · DPA and TBHQ antioxidants · Emission analysis · Performance analysis

Introduction

Increasing energy demand, as well as the environmental harm caused by pollutant emissions from fossil-fuel combustion, has prompted researchers to look for alternative fuels to replace traditional petroleum fuels. However, the combustion of diesel engines produces a lot of NO_x and smoke, which affects the health of humans. Therefore, biodiesel is considered an alternative fuel, as it is renewable and can be used in unmodified diesel engines along with reduced harmful emissions such as carbon monoxide, hydrocarbon compared to fossil fuel. (Pali et al. 2015; Dueso et al. 2018; Uğuz et al. 2019). Waste cooking oil (WCO) is considered as one of the potential feedstock for the production of biodiesel (Chhetri et al. 2008; Sidharth and Kumar 2020). The utilization of waste cooking oil helps to reduce the problem of used oil disposal and reduces raw material costs. Therefore, waste cooking oil gained worldwide interest for converting it into biodiesel (Zareh et al. 2017; Nagarajan and Narayanasamy 2021). Conversion of waste cooking oil into biodiesel reduces the cost of processing waste treatment including waste cooking oil and reduces environmental contamination.

Converting waste cooking oil into biodiesel also helps to alleviate the energy issue (Sonthalia and Kumar 2021).

Biodiesel, fatty acid methyl esters (FAME) contains saturated and unsaturated fatty acid. Unsaturated fatty acids present in FAME are susceptible to autoxidation. Autoxidation deteriorates the quality of biodiesel; if the quality of biodiesel deteriorates due to fuel autoxidation, it may be unsuitable for use as a fuel in an engine. Autoxidation of biodiesel may cause oil blockage, filtration difficulties, engine corrosion, and engine performance instability (Fu et al. 2016; Bharti and Singh 2020). Antioxidants have been suggested by several studies as a way to increase the biodiesel oxidation stability. Antioxidants terminate a process of oxidation by scavenging free radicals. Most of the commonly used synthetic antioxidants are TBHQ, BHT, BHA, PY, and PG having phenolic compounds (Liu et al. 2019; Yadav et al. 2022). Biodiesel from waste cooking oil was used with TBHQ, PY, and BHT, antioxidants to evaluate oxidation stability by FTIR and DSC techniques. TBHQ was more effective compared to BHT and PY antioxidants (Uğuz et al. 2019). TBHQ, BHA, GA, MT, PG, OG, PA VC, AP, D-TBHQ antioxidants were tested with rubber seed biodiesel at different temperatures. BHA and



TBHQ show better results than others at room temperature (Ni et al. 2020).

Despite several advantages, biodiesel has few drawbacks when utilized in diesel engines. Higher alcohol consumption and lower oil yield during the production process are a few of them and NO_x emission during combustion of fuel (Pikula et al. 2020). In comparison with diesel, NO_x emissions rise when biodiesel is blended at a higher proportion. The utilization of waste cooking oil biodiesel increases NO_x emission, shown in previous studies (Abed et al. 2018). Performance and emission of N-phenyl-1, titanium oxide (TiO₂), and 4-phenylenediamine (NPPD) antioxidant nanoparticles were evaluated for palm biodiesel. Both antioxidants can mitigate NO_x compared to a diesel with optimum performance. Antioxidants with biodiesel reduce NO_x by terminating free radicals, decomposing peroxides, and preventing free radical chain reactions (Reddy and Wani 2021). PPDA, AT, and LA antioxidants (150 mg concentration), when mixed in 20% blend of Annona biodiesel (A20), effectively reduce NO_x emission by 24.7, 22, and 23.8% compared to A20 biodiesel with additive (Rajendran 2020).

The performance and emission analysis of biodiesel treated with antioxidants has been the focus of several studies. However, these tests are costlier and require a huge amount of time. As a solution, the application of computer software gives the same level of efficiency with less number of tests. Artificial neural network (ANN) is getting attention to predict engine performance and emission by minimizing the number of experiments (Uslu and Celik 2018). Hosseini et al. (2020) studied the ANN model to evaluate engine performance, emission, and vibration levels by utilizing alumina as an additive, 30, 60, 90 ppm, 5%, and 10% biodiesel blend in diesel. Selected input parameters were fuel density, fuel blend, engine speed, lower heating value, intake manifold pressure, fuel viscosity, consumption of fuel, exhaust gas temperature, relative humidity, oil temperature, and ambient air pressure for targeted output parameters power output, torque, UHC, CO₂, CO, NO, RMS, and engine vibrations. Performance and emission of the proposed model were found satisfactory with an R-Value nearby 0.999 for training, validation, and testing. Kumar et al. (2020) develops ANN model to predict performance and emission of a ternary blend, diesel-palm biodiesel-decanol additive. Selected output was BTE, NO_x, CO, BSFC, HC, CO₂, EGT, ignition delay period, and smoke opacity. Results show R-value greater than 0.99 for all ANN predicted output.

Considering available literature, waste cooking oil biodiesel fits waste to fuel production, which eliminates disposal issues of waste cooking oil, prevents land and water pollution. Utilization of waste cooking oil biodiesel helps to enhance performance of the engine and reduces emission (Chen et al. 2020; Chaudhary 2022). However, compared to the one using petro-diesel, NO_x emission will be somewhat

enhanced. Increased NO_x emission significantly controlled by oxidizing peroxides, removing free radicals, and stopping the chain reaction of free radicals. The antioxidants added to biodiesel demonstrated a more notable reduction in NO_x emissions (Reddy and Wani 2021). In this, present study, an aromatic antioxidant (DPA)-treated biodiesel fuel blend was used to analyze performance and emission of the engine. Most commonly used synthetic antioxidants (TBHQ)-treated biodiesel fuel blend was used for comparison of analysis. Traditional methods for analyzing performance and emission behavior of engine are costly and very time consuming. Therefore, researchers are looking toward to find out alternative ways to analyze it by cheaper and faster methods. An artificial technique like ANN was found best-fitted technique to analyze performance and emission of the engine by developing models (Ayd et al. 2020). The aim of this research work was to find aromatic antioxidant impact on biodiesel in terms of performance and emission, specially NO_x emission reduction and to create an ANN model for estimating performance and emission analysis in relation with input load, blend percentage, and antioxidants. The present study shows an ANN model application by utilizing data from experiments and prediction of performance and emission. This model is applicable to analyze output parameters under different conditions.

This research work was initiated in 2021 at CASRAE (Centre for Advanced Studies and Research in Automobile Engineering), Department of mechanical engineering, Delhi technological university, Delhi, India, and this research paper was finalized in 2022.

Materials and methods

Waste cooking oil was obtained from a nearby canteen. All other chemicals were obtained from a local supplier with analytical grade of 99 percent purity. In this section, the production of biodiesel, antioxidant treatment of biodiesel, and blend preparation are explained. BTE, BSEC, EGT, NO_x, HC, CO, and opacity were evaluated for different test fuel blends for unmodified diesel engine. Prediction and validated of performance and emission parameters are performed by ANN approach.

Production and preparation of antioxidant-treated biodiesel

The first stage of biodiesel production is to measure free fatty acid (FFA) concentration of the oil. After that, either the esterification (FFA is greater than 2 wt%) or direct transesterification process (FFA is less than 2 wt%) takes place (Tomar and Kumar 2020). The ASTM-D644 standard is used to calculate FFA. Oil is titrated using N/10 KOH



solution to the mixture of phenolphthalein and isopropyl alcohols. The FFA concentration of waste cooking oil was measured less than 2%. Biodiesel can be produced only by one-step transesterification with an alkaline catalyst. In order to transesterify the oil, methanol (20% weight-to-weight) and KOH pellets (1% weight-to-weight) were combined. Oil is transesterified in the presence of a base as a catalyst in the last phase of production. As crude glycerol is heavier than water, it sinks to the bottom and is subsequently filtered out. As a result, a clean biodiesel layer was produced by washing the leftover residue (methyl ester) in hot water (around 38 °C) until. Figure 1 represents the flow chart for production process of biodiesel. A similar approach was used by Sidharth and Kumar (2020). Figure 2 shows a physical representation of biodiesel production at different stages.

Preparation of test blend

Antioxidants additives, tert-butylhydroquinone (TBHQ), and diphenylamine (DPA) were purchased from local vendor. TBHQ has 97% assay, 166.22 g/mol molecular weight, 127–129 °C, white to light tan color, fully soluble, phenolic type and DPA has 98% assay, 166.23 g/mol molecular weight, 51–55 °C, white to light yellow color, fully soluble, aromatic type. For experimental testing, five test fuel blends were prepared, named 80D20B, B20(300DPA), B20(400DPA), B20(300TBHQ), and B20(400TBHQ). Representation of the fuel blend is given in Table 1. Mixing of antioxidants in neat waste cooking oil biodiesel was carried out in an ultrasonicator at a frequency of 40 kHz for 45 min.

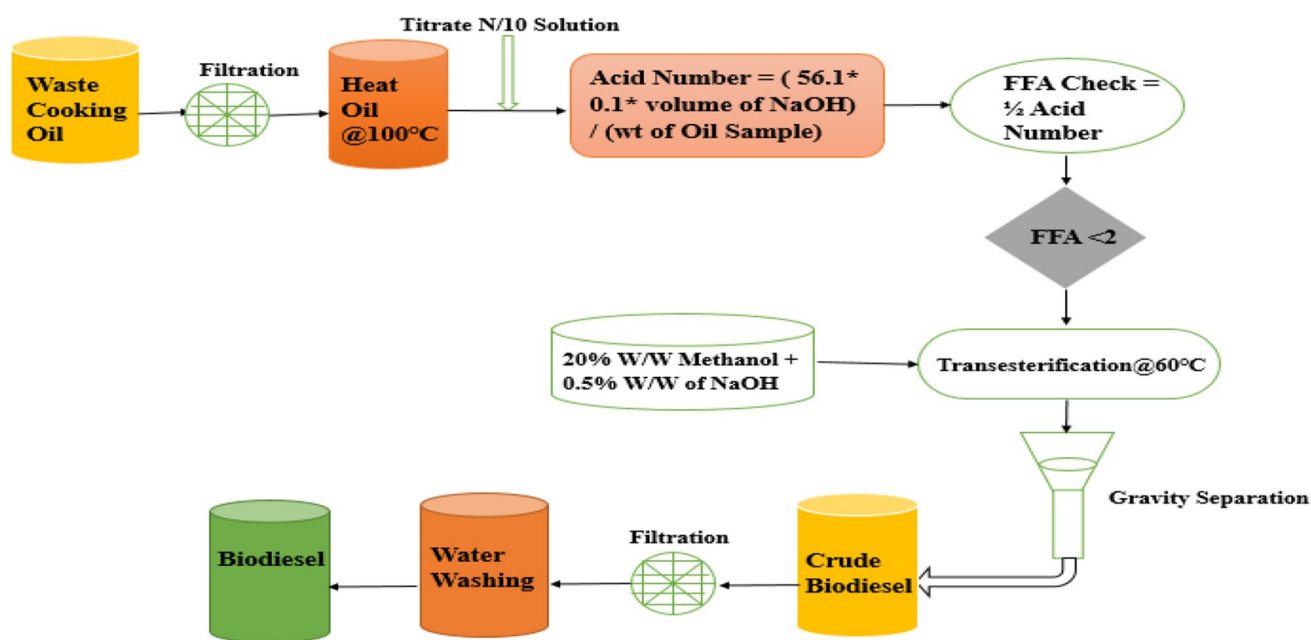


Fig. 1 Flow chart for production process of biodiesel

Fig. 2 Photographic representation of methyl ester and glycerol separation, water washing, and biodiesel



Table 1 Representation of test fuel blends

S. no	Test fuel blends	Representation
1	80D20B	80% diesel 20% biodiesel in volume percentage
2	B20(300DPA)	80% diesel 20% (300 ppm of DPA in biodiesel)
3	B20(400DPA)	80% diesel 20% (400 ppm of DPA in biodiesel)
4	B20(300TBHQ)	80% diesel 20% (300 ppm of TBHQ in biodiesel)
5	B20(400TBHQ)	80% Diesel 20% (400 ppm of TBHQ in biodiesel)

Table 2 Engine test rig specifications

Parameter	Specifications
Type	Single cylinder, 4-stroke, direct injection type
Model	DAF8, Kirloskar
Rated power	3.5 KW
Rated speed	1500 rpm
Length of stroke and diameter of bore	110×95 mm
Injector type	Six holed solenoid
Cooling	Air-cooled
Compression ratio	17.5:1
Lubrication type	Forced feed
Injection timing and pressure fuel	23° and 200 bar before top dead center

Experiment setup

The engine testing was performed on Kirloskar's unmodified diesel engine. Engine having 4-stroke, direct injection, single-cylinder working with constant rpm (1500) and 3.5KW rated power output, detailed specifications are given in Table 2. A schematic layout of diesel engine configuration for experimental trial is shown in Fig. 3. On the electric loading unit, a voltmeter, an ammeter, and rpm indicator were used to measure current, voltage, and speed of the output shaft of the engine. A computer unit and data acquisition system attached to the engine test setup. A data acquisition system was used to measure, control, and monitor the mass flow rates of fuel, temperature, and engine load at varying load conditions. An electric bulb setup was used to apply engine loading. Exhaust gas temperature of the engine was controlled by attaching a thermocouple at the exhaust. An AVL gas analyzer (AVL-1000) and AVL-480 smoke meter were used to measure exhaust gas emissions and smoke opacity coming out from engine combustion. The load applied

to the engine was varied by 20% of 100% load through electric loading at 1500 rpm. Volumetric mass flow rate of fuel was measured using a burette and stopwatch. The performance of a baseline diesel was assessed first, and then other diesel biodiesel blends were tested and compared. All test was performed in ambient conditions.

ANN application and data prediction approach

Nowadays researchers show their keen interest to use ANN in the automobile sector (Ayd et al. 2020). An ANN relates the inputs and outputs of a system, and it has been successfully applied to map nonlinear input and output correlations in a variety of fields. An ANN network has a layer of input nodes, a layer of output nodes, and one or more layers of hidden nodes connecting them. All three layers are having a specific number of small individuals, called neurons. The neurons transfer their signal of communication to other neurons via a communication link that is associated with specific weight (Uslu and Celik 2018). In this machine learning, ANN approach uses two major steps: 1. training step (set of data provided with weight and bias) and 2. prediction step (training of neural network according to input data provided) (Barnawal and Kumar 2021). A detailed process of ANN approach is given through the flow chart in Fig. 3.

ANN model structure

In this research paper, prediction of performance and emission of waste cooking biodiesel blend have been done using ANN. Data set used as input in ANN model were taken from engine trial at different loads. Engine load, blend percentage, dose of antioxidants were used as input to the model. Evaluated parameters were BSEC, BTE, NO_x, EGT, HC, CO, and opacity. The architecture diagram of ANN is shown in Fig. 4. In this ANN model, 70% of data were used for training purposes, 15% of data for testing purposes, and 15% of data were utilized as specific samples. In ANN modeling, for training and optimization of data, the “quasi-Newton method” was used.

Data prediction and validation by ANN

Linear regression analysis is found to be standard method to test the loss of the model between scaled output of the neural network to their corresponding targets. This analysis gives three parameters to individual output variables. These three parameters were y-intercept, slope, and correlation coefficient between scaled output and target. Slope and y-intercept refer to as a and b. For best fit, slope approaches one and y-intercept approaches zero and correlation approaches (R^2) one. For a perfect fit, the slope between output and targeted value should be one and the y-intercept is zero shows there



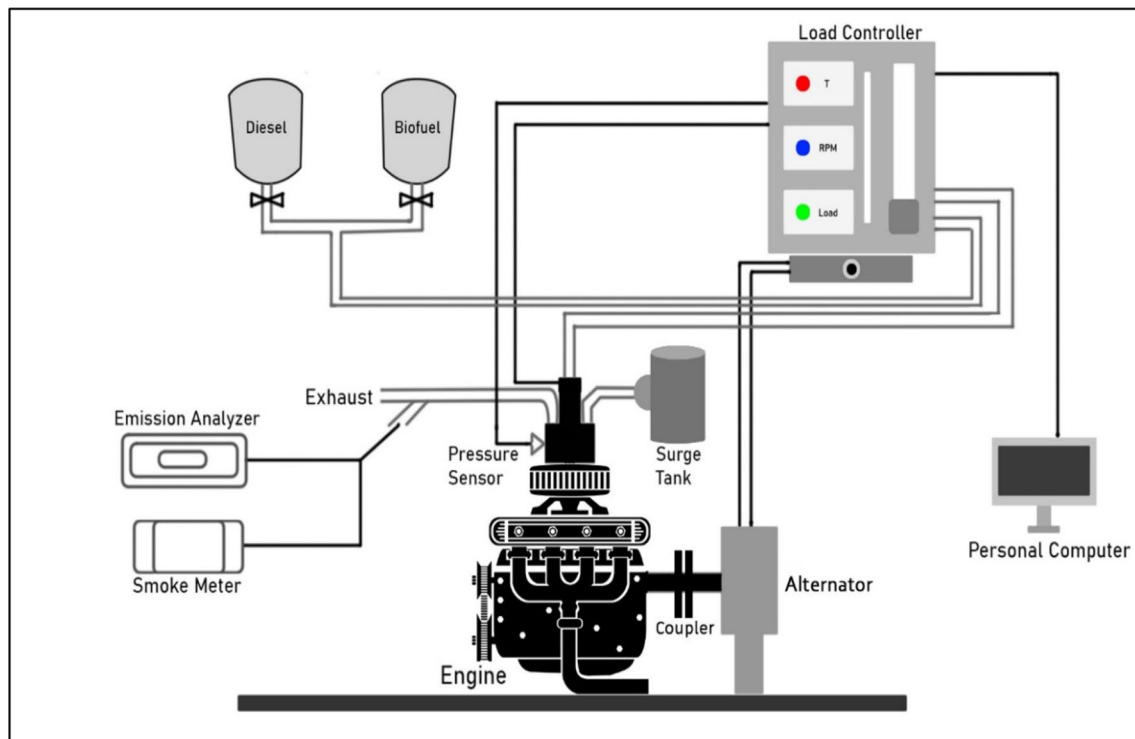


Fig. 3 Schematic diagram of diesel engine setup for experimental trial

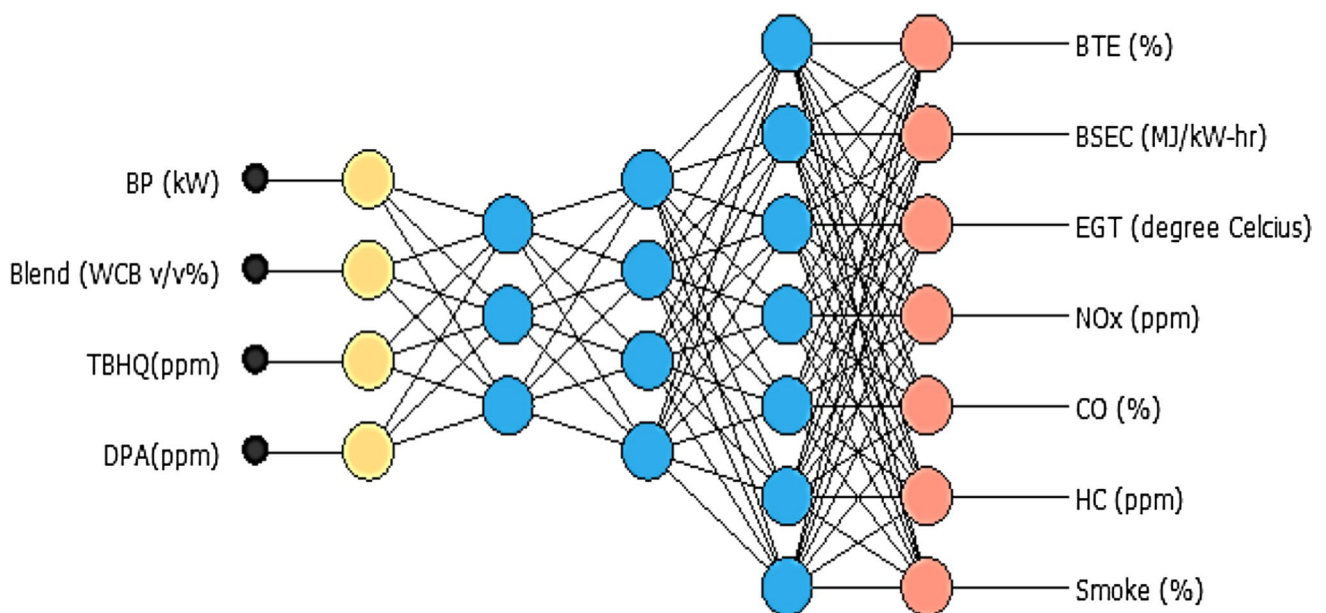


Fig. 4 ANN architecture



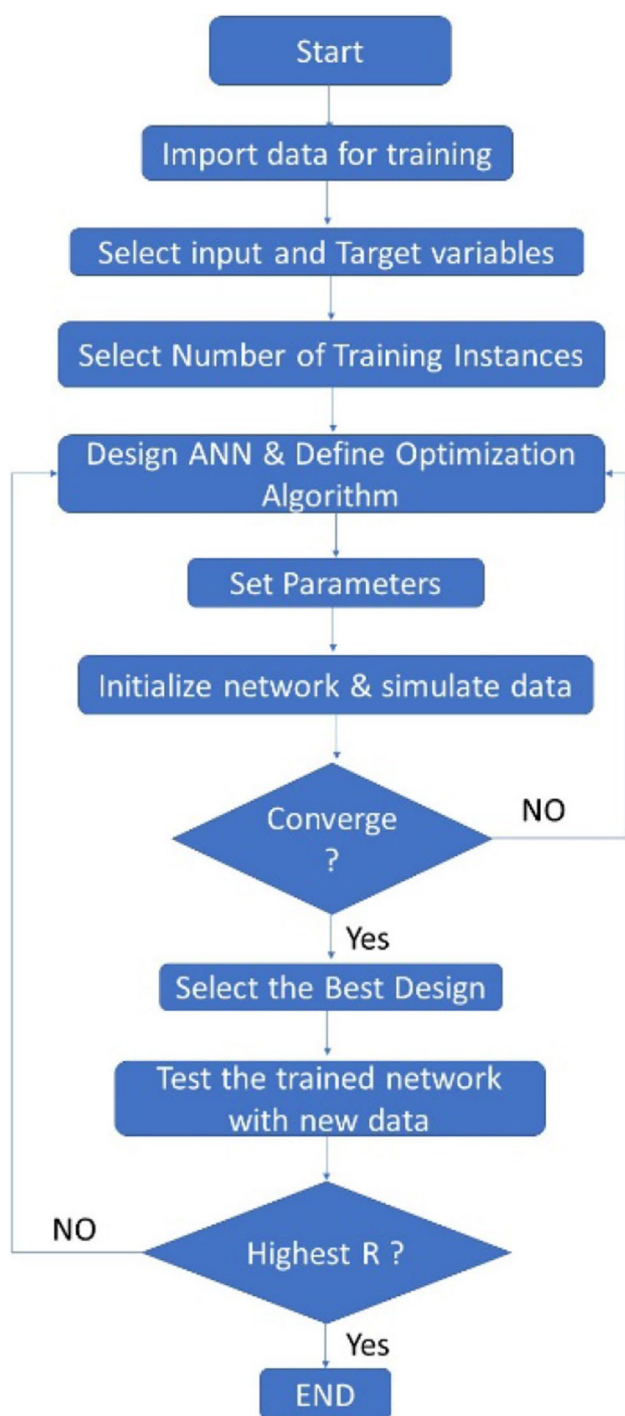


Fig. 5 Flow chart of artificial neural network

is no gap between output and target. When scaled output and targeted out are perfectly correlated, correlation coefficient will be one. Flow chart of ANN is represented in Fig. 5.

Results and discussion

Physico-chemical properties and characterization of biodiesel

The physicochemical qualities of the fuel are heavily dependent on the engine's performance, combustion, and emission characteristics. Biodiesel and biodiesel with antioxidant blends having different fuel characteristics such as viscosity, density, and calorific values are shown in given Table 3. Properties of diesel fuel were found in range as per ASTM methods, but biodiesel had shown higher viscosity in range. 20% blending of diesel in biodiesel gave possible range of viscosity within ASTM range. Slight increases in viscosity of biodiesel was found by adding antioxidant in 20% blended fuel. All tested physicochemical properties of antioxidant-treated biodiesel blends were found within range as per ASTM standard. The chemical composition of waste cooking oil biodiesel (WCB) was investigated using GC–MS. Fatty acid methyl esters can be separated, identified, quantified, and analyzed using GC. At varying retention periods, distinct fatty acid methyl esters were identified. It was shown that a higher unsaturated fatty acid content leads to inferior oxidative stability and a low cetane number.

The gas chromatography/mass spectroscopy was performed using GC-MS machine, shown in Fig. 6. By using this machine, fatty acid methyl ester of waste cooking oil biodiesel was attained. The composition of all acids is represented in Table 4. Major compounds were to be 67.72% palmitic acid (9,12-octadecadienoic acid, methyl) ester and 22.10% linoleic acid (hexadecanoic acid, methyl ester). Total unsaturated and saturated fatty acids of WCB were 68.45 and 31.55, respectively. The quantity and group of methyl ester present in biodiesel determine the sustainability, and suitability of fuel.

Experimental performance and data prediction by ANN

Variation of brake specific energy consumption

Braking specific fuel consumption is the amount of fuel used by the engine per unit of production of brake power, which is represented in kg/kWh. BSFC depends on various fuel properties like viscosity, calorific value, density. Figure 7 depicts the variation of brake specific energy consumption



Table 3 Physicochemical properties of diesel, biodiesel, and test fuel blends

Fuel properties	Diesel	WCB	80D20B	80D20(WCB + 300 DPA)	80D20(WCB + 400 DPA)	80D20(WCB + 300 TBHQ)	80D20(WCB + 400 TBHQ)	Measuring Apparatus
Kinematic viscosity (cSt) @40°C	2.55	5.34	3.12344	3.13271	3.15721	3.14582	3.15124	Visco bath, petrotest
Density @15 °C (g/m ³)	834.4	882.3	848.2	849.5	849.6	849.6	849.7	Anton par, DMA 4500
Lower heating value (MJ/Kg)	42.871	38.92	41.8392	41.7164	41.6437	41.7812	41.6534	Oxygen bomb calorimeter parr
Oxidation stability		1.74						Rancimat

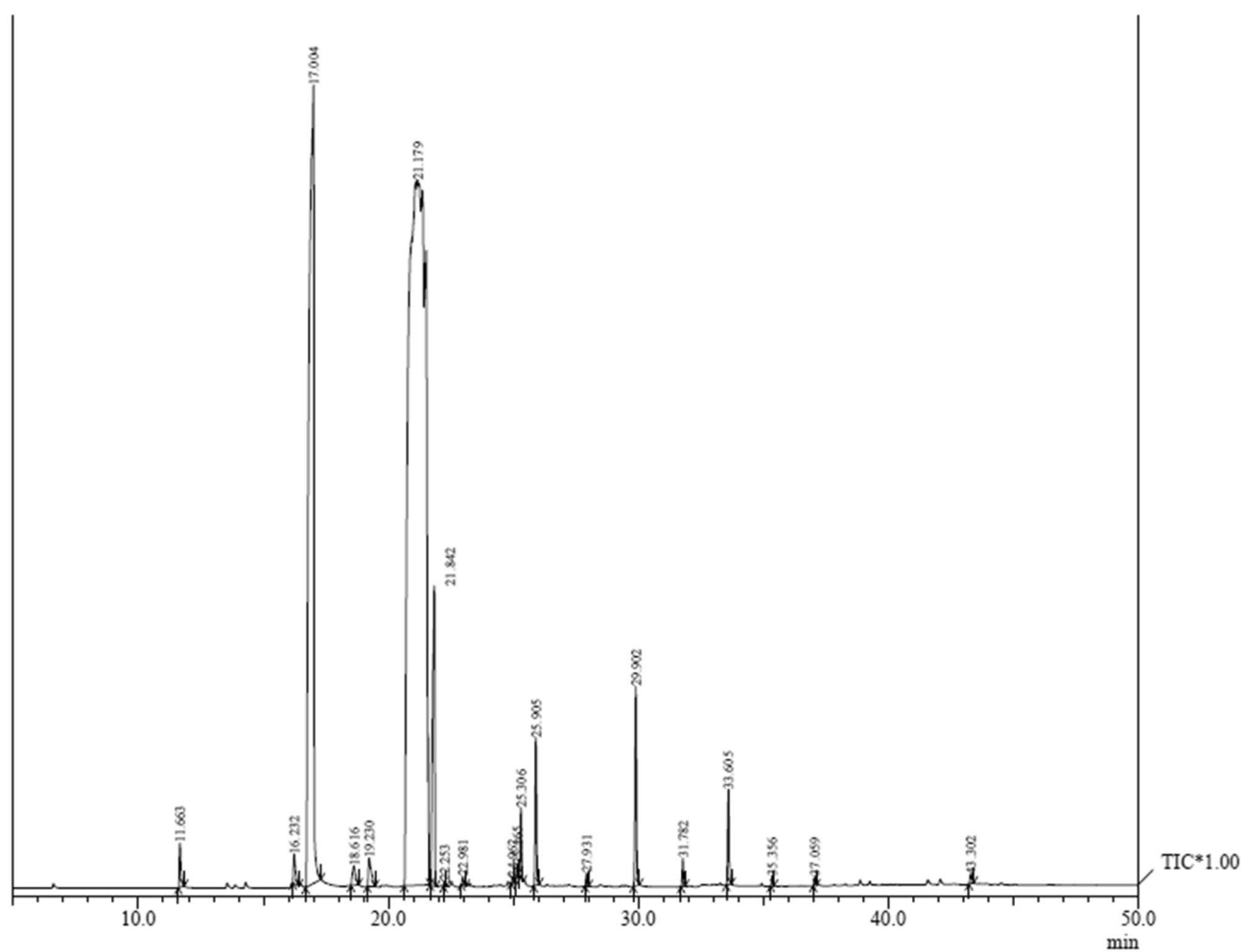
**Fig. 6** GC–MS chromatograms of WCB

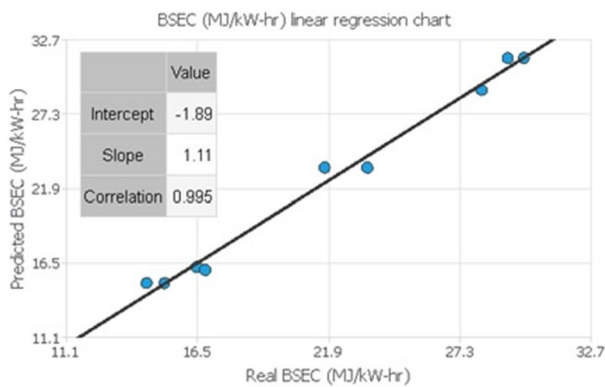
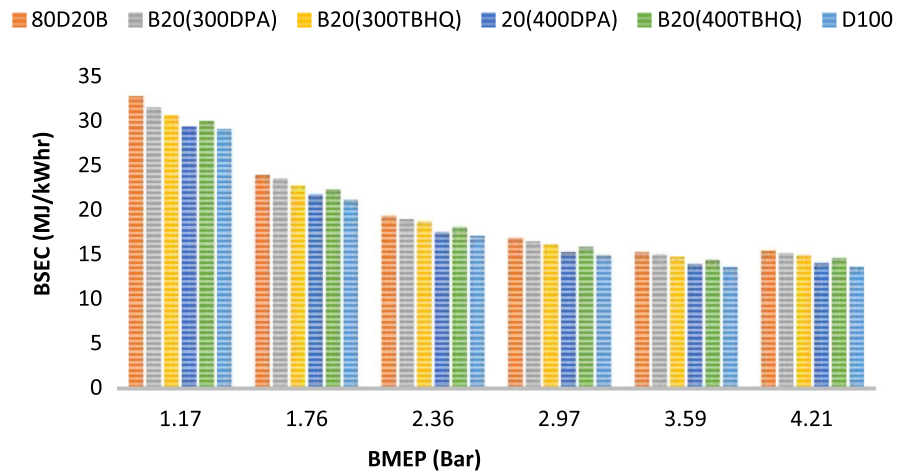
Table 4 Unsaturated and saturated fatty acid composition of WCB

Peak	R. Time	Area%	Compound name	Chemical formula	Molecular weight	Molecular structure	Type
1	11.663	0.41	Methyl tetradecanoate	C15H30O2	242		Saturated
2	16.232	0.35	Methyl palmitoleate	C17H32O2	268		Saturated
3	17.004	22.10	PALMITIC ACID METHYL ESTER	C17H34O2	270		Saturated
4	18.616	0.35	9-Heptadecenoic acid	C18H34O2	282		Saturated
5	19.230	0.48	Margaric acid methyl ester	C18H36O2	284		Saturated
6	21.179	67.72	Methyl octadeca-9,12-dienoate	C19H34O2	294		Unsaturated
7	21.842	3.72	Stearic acid, methyl ester	C19H38O2	298		Saturated
8	22.253	0.03	METHYL HEXADECATRIENOATE	C17H28O2	264		Unsaturated
9	22.981	0.07	Methyl trans-9,trans-11-octadecadienoate	C19H34O2	294		Unsaturated
10	24.962	0.04	Linoleoyl chloride	C18H31ClO	298		Unsaturated
11	25.165	0.09	Methyl octadeca-9,12-dienoate	C19H34O2	294		Unsaturated
12	25.306	0.50	11-Eicosenoic acid, methyl ester, (Z)-	C21H40O2	324		Unsaturated
13	25.905	1.17	ARACHIDIC ACID METHYL ESTER	C21H42O2	326		Saturated
14	27.931	0.10	Methyl heneicosanoate	C22H44O2	340		Saturated
15	29.902	1.63	Behenic acid methyl ester	C23H46O2	354		Saturated
16	31.782	0.22	Methyl tricosanoate	C24H48O2	368		Saturated
17	33.605	0.74	Methyl lignocerate	C25H50O2	382		Saturated
18	35.356	0.07	Methyl pentacosanoate	C26H52O2	396		Saturated
19	37.059	0.07	Methyl hexacosanoate	C27H54O2	410		Saturated
20	43.302	0.11	Clionasterol	C29H50O	414		Saturated
		68.45					Unsaturated
		31.55					Saturated

(BSEC) with brake mean effective pressure (BMEP) for different ppm antioxidant-treated fuel blends (B20 + antioxidant ppm) compared with diesel. It has been observed from the graph that BSEC decreases as BMEP increases. BSEC for diesel was lower compared to all other tested blends due to lower density, higher calorific value, higher volatility, and lower viscosity of fuel (Saravanan et al. 2019). 80D20B blend depicts higher BSEC due to higher viscosity, higher density, and lower calorific value. In comparison with 80D20B, B20(300DPA), B20(300TBHQ), B20(400TBHQ), B20(400DPA), and D100 exhibit 1.6, 3.2, 8.9, and 11.67%

lower BSEC. Because of the friction reduction characteristic of amine in antioxidants, all antioxidant-treated biodiesel blends exhibit a considerable reduction in BSEC. Similar results were shown in previous studies (Senthur Prabu et al. 2017; Reddy and Wani 2021). Figure 8 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.995 for BSEC.



Fig. 7 Variation of BSEC (MJ/kWh) and BMEP (bar)**Fig. 8** Predicted BSEC (MJ/kWh) vs. real BSEC (MJ/kWh)

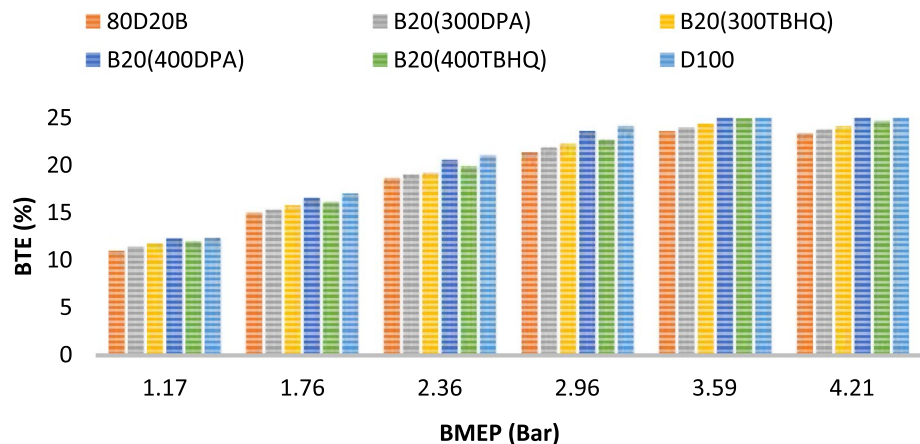
Variation of brake thermal efficiency

Brake power output to the energy supplied as mass of fuel and calorific value is known as brake thermal efficiency (BTE). Figure 9 depicts the variation of BTE with BMEP due to lean air–fuel mixture improving combustion (Raman

and Kumar 2020). D100 has higher BTE; on other hand, 80D20B has lower BTE on all loads. Lower BTE of 80D20B is due to combined effect of low calorific fuel, higher viscosity, and shorter ignition delay period (Senthur Prabu et al. 2017). In comparison with 80D20B, B20(300DPA), B20(300TBHQ), B20(400 TBHQ), B20(400DPA), and D100, BTE increase by 1.6, 3.1, 5.5, 9.7, and 13.38%. All antioxidant-treated biodiesel blends show a remarkable increase in BTE due to a decrease in BSEC and higher power output. Similar trends are in line with previous studies carried out by Nagappan et al. (2021). Figure 10 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.985 for BTE.

Variation of NOx emission

A greater combustion temperature, a longer combustion period, and a high oxygen content in the fuel are the main contributors to nitrogen oxide (NOx) formation. Figure 11 depicts variation between NOx (%) emission with BMEP (bar). D100 shows the minimum emission among all other

Fig. 9 Variation of BTE (%) and BMEP (bar)

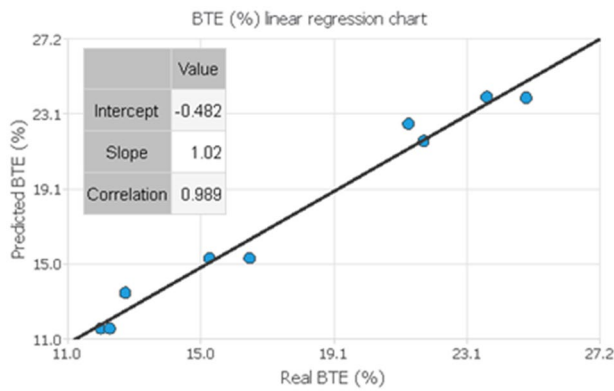


Fig. 10 Predicted BTE (%) vs. real BTE (%)

fuel blends. 80D20B shows higher NO_x due to higher oxygen content, 12% higher than diesel, reduction in ignition delay period, and higher temperature of combustion (Nagappan et al. 2021). In comparison with 80D20B, NO_x levels decreased by 2.32, 5.24, 7.35, 12.44, and 16.15% in B20(300DPA), B20(300TBHQ), B20(400TBHQ), B20(400DPA), and D100. All antioxidant-treated fuel blends show lower NO_x than 80D20B due to concentration reduction of free radicals, quenching of free radicals, and scavenging of free radical behavior of antioxidants (Jeyakumar and Narayanasamy 2020). Figure 12 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.988 for NO_x.

Variation of hydrocarbon carbon emission

The hydrocarbon formation is greatly affected by the fuel characteristics and fuel spray characteristics in a diesel

engine. Figure 13 shows the variation of HC emission (ppm) with BMEP (bar). D100 shows higher emissions than all other fuel blends. 80D20B shows lower HC emissions due to rich oxygen content of biodiesel and cetane number leading to complete combustion (Adam et al. 2018). In comparison with D100, B20(400DPA), B20(400TBHQ), B20(300TBHQ), B20(300DPA), and 80D80WCB show 1.9, 5.8, 9.8, 13.72, and 17.6% reduction in HC. Figure 14 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.946 for HC.

Variation of CO emission

The main reason for CO emissions formation is incomplete combustion, which is caused by inadequate air and low flame temperature. Figure 15 depicts the variation of CO emission for all test fuel blends, biodiesel blends and biodiesel blends with antioxidant. The increased value of CO specific emissions was obtained for pure

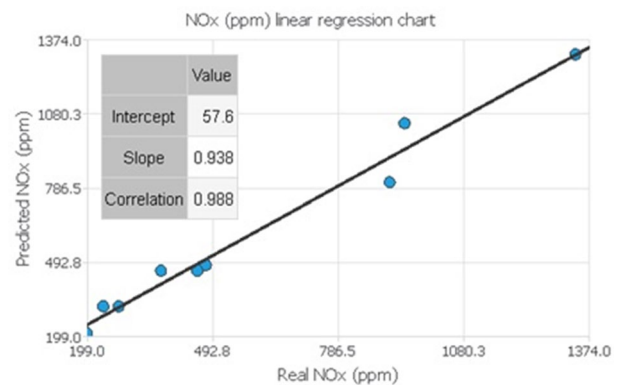


Fig. 12 Predicted NO_x (ppm) vs. real NO_x (ppm)

Fig. 11 Variation of NO_x (ppm) and BMEP (bar)

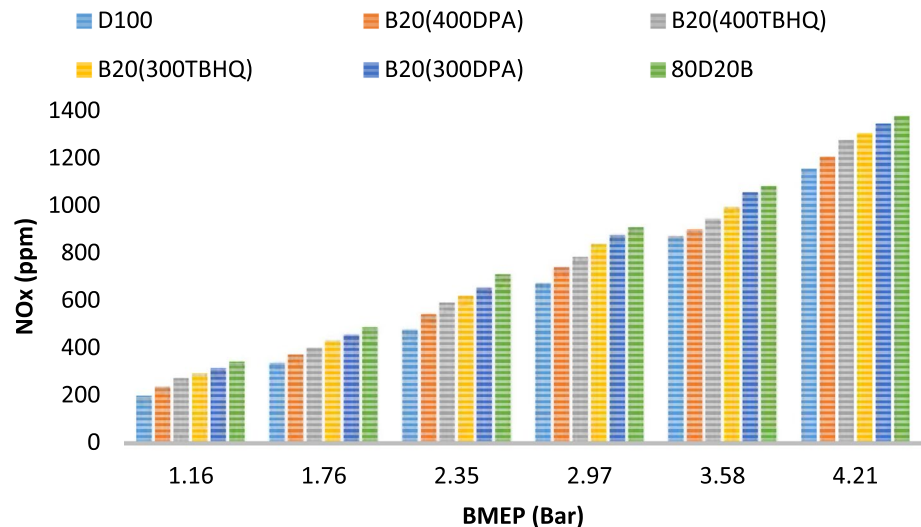
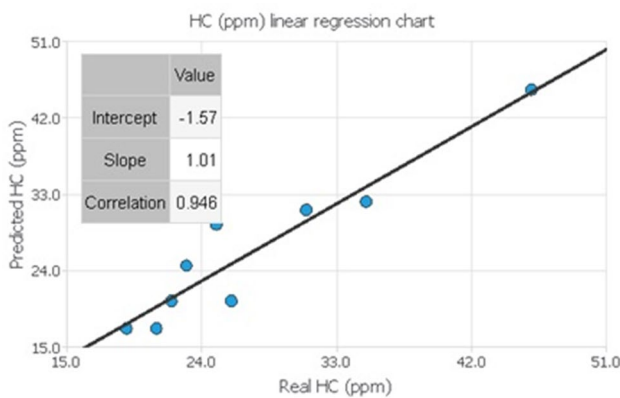
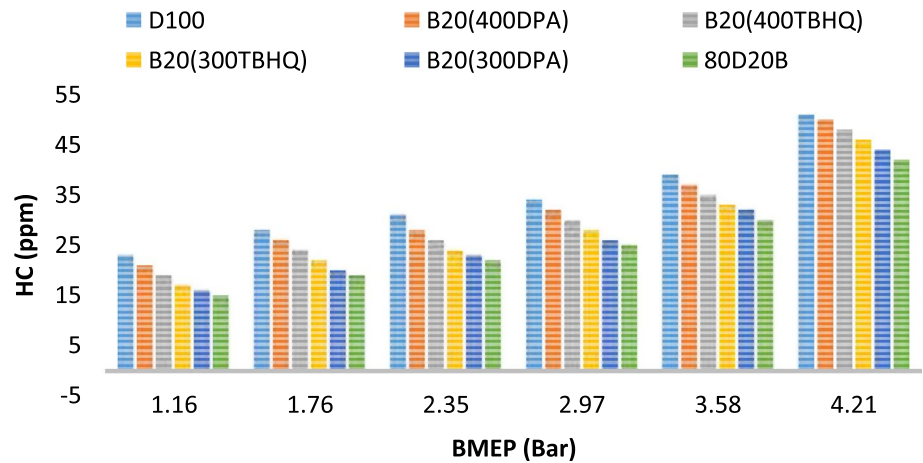


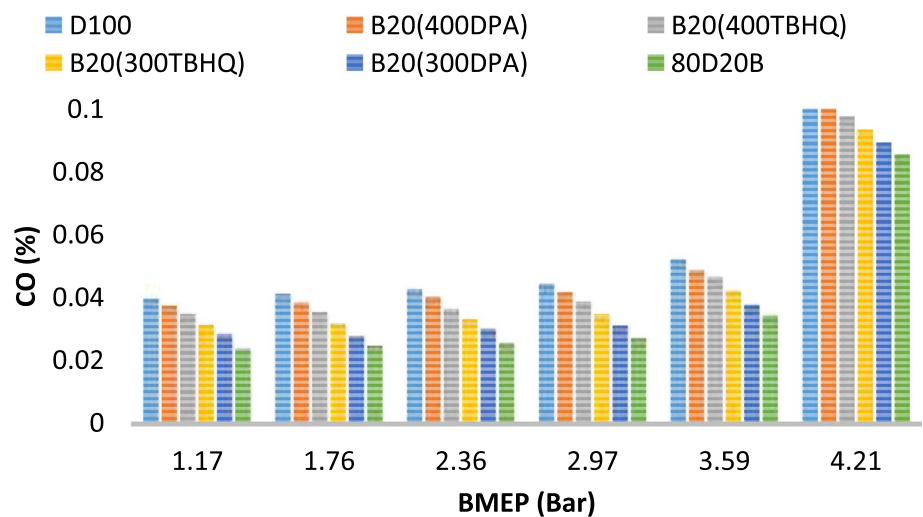
Fig. 13 Variation of HC (ppm) and BMEP (bar)**Fig. 14** Predicted HC (ppm) vs. real HC (ppm)

diesel. Utilization of biodiesel blend shows lower emissions than D100 due to higher oxygen availability in the fuel blend (Dueso et al. 2018). In comparison with

D100, the HC levels of B20(400DPA), B20(400TBHQ), B20(300TBHQ), B20(300DPA), and 80D80WCB are reduced by 5.45, 11.27, 15, 18.8, and 22.18%, respectively. The rise in CO emissions caused by antioxidant blends is mostly due to the antioxidant's ability to prevent CO from being converted to CO₂ (Adam et al. 2018). Figure 16 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.945 for CO.

Variation of Smoke opacity

Figure 17 depicts variation in smoke opacity vs BMEP for all tested fuel blends. The smoke emission for D100 is higher due to presence of increasing aromatic content and ignition delay (Tomar and Kumar 2020). Biodiesel fuel blend without antioxidants shows lower emission due to rich oxygen content leads to better combustion and reduction in smoke

Fig. 15 Variation of CO (%) and BMEP (bar)

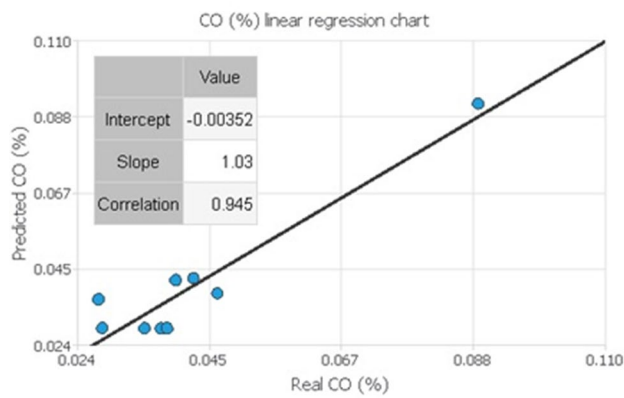


Fig. 16 Predicted CO (%) vs. real CO (%)

emission. All antioxidant biodiesel blends have lower emissions than diesel. In comparison with D100, B20(400DPA), B20(400TBHQ), B20(300TBHQ), B20(400DPA), and 80D80WCB exhibit 3.32, 8.64, 14.06, 16.87, and 20.46% lower HC. Figure 18 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.983 for smoke opacity.

Variation of EGT

Figure 19 depicts variation between EGT (ppm) emission with BMEP (bar). It has been observed from the graph conventional diesel has lower EGT than all other fuel blends at all loading conditions. On the other hand, 80D20B shows

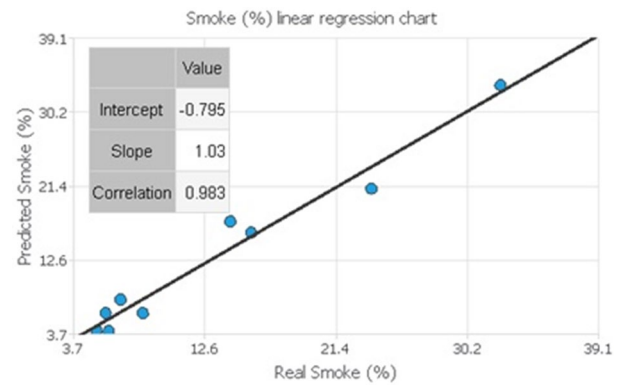


Fig. 18 Predicted smoke opacity (%) vs. real smoke opacity (%)

highest EGT emission, this reduction is due to minimum ignition delay period and rich oxygen content present in biodiesel (Nagappan et al. 2021). In comparison with 80D20B, EGT were reduced by 3.39, 5.3, 8.25, 11, and 12.5% in B20(300DPA), B20(300TBHQ), B20(400TBHQ), 80D20(WCB + 00DPA), and D100. All antioxidant-treated fuel blends show lower EGT than 80D20B due to concentration reduction of free radicals, quenching of free radicals, and scavenging of free radical behavior of antioxidants (Jeyakumar and Narayanasamy 2020). Figure 20 illustrates ANN simulation of research findings and software interpretation. According to the interpretation of ANN, R^2 values were found 0.979 for EGT.

Fig. 17 Variation of smoke opacity (%) and BMEP (bar)

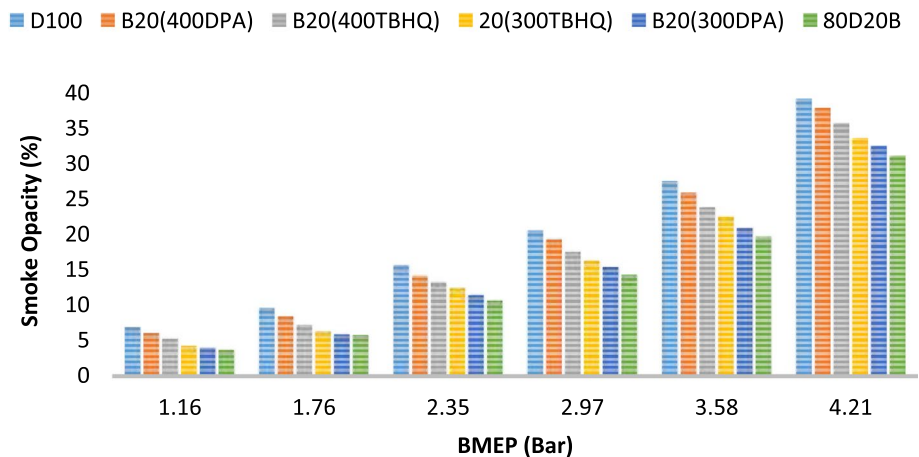


Fig. 19 Variation of EGT (ppm) and BMEP (Bar)

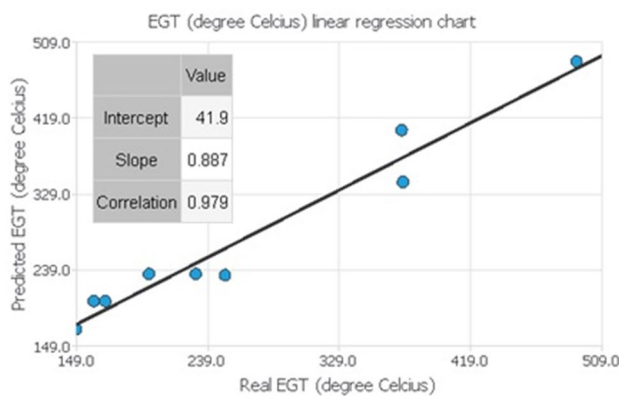
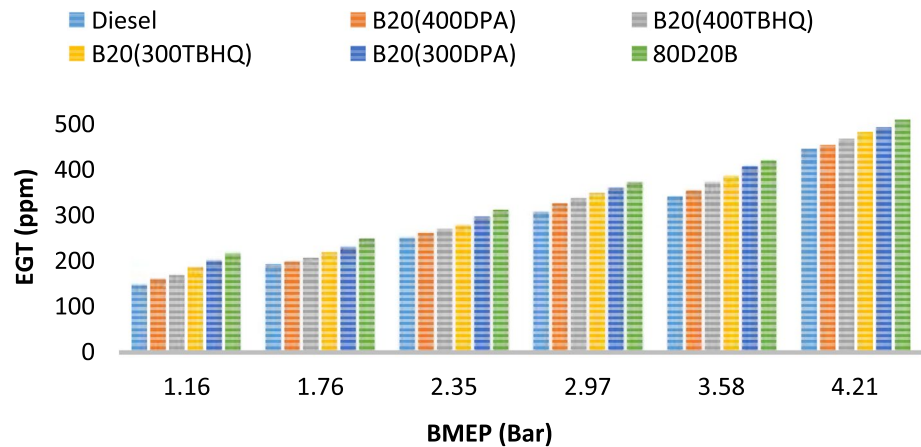


Fig. 20 Predicted EGT (ppm) vs. real EGT (ppm)

Conclusion

In the present research work, two distinct antioxidants were tested to enhance the viability of waste cooking oil biodiesel. WCB physicochemical properties were tested as per ASTM D6751. GC–MS test were also conducted, showing higher unsaturation compounds present in WCB. An ANN modeling was performed to estimate correlation between observed and experimented data. The following observations were drawn:

- Antioxidant-treated biodiesel blends show increasing BTE and decreasing BSEC at all loads. At peak load, BTE increased by 13.38% and BSEC decreased by 11.67% for 400 ppm DPA antioxidant fuel blend compared to without antioxidant. ANN results also approach each other.

- NO_x emissions were lowest for 400 ppm antioxidant fuel blend, i.e., 16.15% reduction compared to without antioxidant. Other emissions slightly increase compared to biodiesel fuel blend without antioxidants but were found to be lower than diesel.
- DPA is a promising antioxidant addition, and it may be combined with WCB of 20% (v/v) blend to significantly reduce NO_x emissions.
- Regression R-value for performance and emission using ANN was 0.985, 0.995, 0.979, 0.988, 0.946, 0.945, and 0.983 for BTE, BSEC, NO_x, EGT, HC, CO, and smoke, respectively. As an outcome, the current study found that ANN is a useful tool for predicting performance and emission of internal combustion engines.

Acknowledgements This research shows that out of several available antioxidants aromatic antioxidant, DPA found effective to reduce NO_x emission in comparison with synthetic antioxidants. Artificial neural network approach was found significant to use for prediction of performance and emission of engine analysis. The future scope of this research study includes application of artificial neural network approach for performance and emission parameters.

Data availability All data generated or analyzed during this study are included in this published article.

References

- Abed KA, El Morsi AK, Sayed MM et al (2018) Effect of waste cooking-oil biodiesel on performance and exhaust emissions of a diesel engine. *Egypt J Pet* 27:985–989. <https://doi.org/10.1016/j.ejpe.2018.02.008>
- Adam IK, Heikal M, Aziz ARA, Yusup S (2018) Mitigation of NO_x emission using aromatic and phenolic antioxidant-treated biodiesel blends in a multi-cylinder diesel engine.



- Environ Sci Pollut Res 25:28500–28516. <https://doi.org/10.1007/s11356-018-2863-8>
- Ayd M, Uslu S, Çelik MB (2020) Performance and emission prediction of a compression ignition engine fueled with biodiesel-diesel blends: a combined application of ANN and RSM based optimization. *Fuel*. <https://doi.org/10.1016/j.fuel.2020.117472>
- Barnawal SK, Kumar N (2021) Experimental investigation and artificial neural network modeling of performance and emission of a CI engine using orange peel oil-diesel blends. *Energy Sour, Part A Recover Util Environ Eff* 00:1–15. <https://doi.org/10.1080/15567036.2021.1967520>
- Bharti R, Singh B (2020) Green tea (*Camellia assamica*) extract as an antioxidant additive to enhance the oxidation stability of biodiesel synthesized from waste cooking oil. *Fuel* 262:116658. <https://doi.org/10.1016/j.fuel.2019.116658>
- Chaudhary V (2022) Influence of diethyl ether on exergy and emission characteristics of diesel engine with waste cooking oil methyl ester. *Int J Environ Sci Technol* 19:4931–4946. <https://doi.org/10.1007/s13762-021-03347-6>
- Chen RH, Ong HC, Wang WC (2020) The optimal blendings of diesel, biodiesel and gasoline with various exhaust gas recirculations for reducing NOx and smoke emissions from a diesel engine. *Fuel* 263:116751. <https://doi.org/10.1016/j.fuel.2019.116751>
- Chhetri A, Watts K, Islam M (2008) Waste cooking oil as an alternate feedstock for biodiesel production. *Energies* 1:3–18. <https://doi.org/10.3390/en1010003>
- Dueso C, Muñoz M, Moreno F et al (2018) Performance and emissions of a diesel engine using sunflower biodiesel with a renewable antioxidant additive from bio-oil. *Fuel* 234:276–285. <https://doi.org/10.1016/j.fuel.2018.07.013>
- Fu J, Turn SQ, Takushi BM, Kawamata CL (2016) Storage and oxidation stabilities of biodiesel derived from waste cooking oil. *Fuel* 167:89–97. <https://doi.org/10.1016/j.fuel.2015.11.041>
- Hosseini SH, Taghizadeh-Alisaraei A, Ghoobadian B, Abbaszadeh-Mayvan A (2020) Artificial neural network modeling of performance, emission, and vibration of a CI engine using alumina nano-catalyst added to diesel-biodiesel blends. *Renew Energy* 149:951–961. <https://doi.org/10.1016/j.renene.2019.10.080>
- Jeyakumar N, Narayanasamy B (2020) Effect of Basil antioxidant additive on the performance, combustion and emission characteristics of used cooking oil biodiesel in CI engine. *J Therm Anal Calorim* 140:457–473. <https://doi.org/10.1007/s10973-019-08699-3>
- Kumar AN, Kishore PS, Raju KB et al (2020) Decanol proportional effect prediction model as additive in palm biodiesel using ANN and RSM technique for diesel engine. *Energy* 213:119072. <https://doi.org/10.1016/j.energy.2020.119072>
- Liu W, Lu G, Yang G, Bi Y (2019) Improving oxidative stability of biodiesel by cis-trans isomerization of carbon-carbon double bonds in unsaturated fatty acid methyl esters. *Fuel* 242:133–139. <https://doi.org/10.1016/j.fuel.2018.12.132>
- Nagappan B, Devarajan Y, Kariappan E et al (2021) Influence of antioxidant additives on performance and emission characteristics of beef tallow biodiesel-fuelled C.I engine. *Environ Sci Pollut Res* 28:12041–12055. <https://doi.org/10.1007/s11356-020-09065-9>
- Nagarajan J, Narayanasamy B (2021) Effects of natural antioxidants on the oxidative stability of waste cooking oil biodiesel. *Biofuels* 12:485–494. <https://doi.org/10.1080/17597269.2019.1711320>
- Ni, Li F, Wang H Z et al (2020) Antioxidative performance and oil-soluble properties of conventional antioxidants in rubber seed oil biodiesel. *Renew Energy* 145:93–98. <https://doi.org/10.1016/j.renene.2019.04.045>
- Pali HS, Kumar N, Alhassan Y (2015) Performance and emission characteristics of an agricultural diesel engine fueled with blends of Sal methyl esters and diesel. *Energy Convers Manag* 90:146–153. <https://doi.org/10.1016/j.enconman.2014.10.064>
- Pikula K, Zakharenko A, Stratidakis A et al (2020) The advances and limitations in biodiesel production: feedstocks, oil extraction methods, production, and environmental life cycle assessment. *Green Chem Lett Rev* 13:11–30. <https://doi.org/10.1080/17518253.2020.1829099>
- Rajendran S (2020) Effect of antioxidant additives on oxides of nitrogen (NOx) emission reduction from Annona biodiesel operated diesel engine. *Renew Energy* 148:1321–1326. <https://doi.org/10.1016/j.renene.2019.10.104>
- Raman R, Kumar N (2020) Experimental studies to evaluate the combustion, performance and emission characteristics of acetylene fuelled CI engine. *Int J Ambient Energy*. <https://doi.org/10.1080/01430750.2019.1709896>
- Reddy SNK, Wani MM (2021) An investigation on the performance and emission studies on diesel engine by addition of nanoparticles and antioxidants as additives in biodiesel blends. *Int Rev Appl Sci Eng* 12:111–118. <https://doi.org/10.1556/1848.2020.00157>
- Saravanan A, Murugan M, Reddy MS, Parida S (2019) Performance and emission characteristics of variable compression ratio CI engine fueled with dual biodiesel blends of Rapeseed and Mahua
- Senthur Prabu S, Asokan MA, Roy R et al (2017) Performance, combustion and emission characteristics of diesel engine fuelled with waste cooking oil bio-diesel/diesel blends with additives. *Energy* 122:638–648. <https://doi.org/10.1016/j.energy.2017.01.119>
- Sidharth KN (2020) Performance and emission studies of ternary fuel blends of diesel, biodiesel and octanol. *Energy Sour, Part A Recover Util Environ Eff* 42:2277–2296. <https://doi.org/10.1080/15567036.2019.1607940>
- Sonthalia A, Kumar N (2021) Comparison of fuel characteristics of hydrotreated waste cooking oil with its biodiesel and fossil diesel. *Environ Sci Pollut Res* 28:11824–11834. <https://doi.org/10.1007/s11356-019-07110-w>
- Tomar M, Kumar N (2020) Effect of multi-walled carbon nanotubes and alumina nano-additives in a light duty diesel engine fuelled with schleicher oleosa biodiesel blends. *Sustain Energy Technol Assessments* 42:100833. <https://doi.org/10.1016/j.seta.2020.100833>
- Uğuz G, Atabani AE, Mohammed MN et al (2019) Fuel stability of biodiesel from waste cooking oil: A comparative evaluation with various antioxidants using FT-IR and DSC techniques. *Biocatal Agric Biotechnol*. <https://doi.org/10.1016/j.bcab.2019.101283>
- Uslu S, Celik MB (2018) Prediction of engine emissions and performance with artificial neural networks in a single cylinder diesel



engine using diethyl ether. *Eng Sci Technol Int J* 21:1194–1201. <https://doi.org/10.1016/j.jestch.2018.08.017>

Yadav K, Kumar N, Chaudhary R (2022) Effect of synthetic and aromatic amine antioxidants on oxidation stability, performance, and emission analysis of waste cooking oil biodiesel. *Environ Sci Pollut Res*. <https://doi.org/10.1007/s11356-021-18086-x>

Zareh P, Zare AA, Ghobadian B (2017) Comparative assessment of performance and emission characteristics of castor, coconut and waste cooking based biodiesel as fuel in a diesel engine. *Energy* 139:883–894. <https://doi.org/10.1016/j.energy.2017.08.040>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Assessment of water surface profile in nonprismatic compound channels using machine learning techniques

Vijay Kaushik* and Munendra Kumar

Department of Civil Engineering, Delhi Technological University, Delhi, 110042, India.

*Corresponding author. E-mail: vijaykaushik_2k20phdce01@dtu.ac.in

ABSTRACT

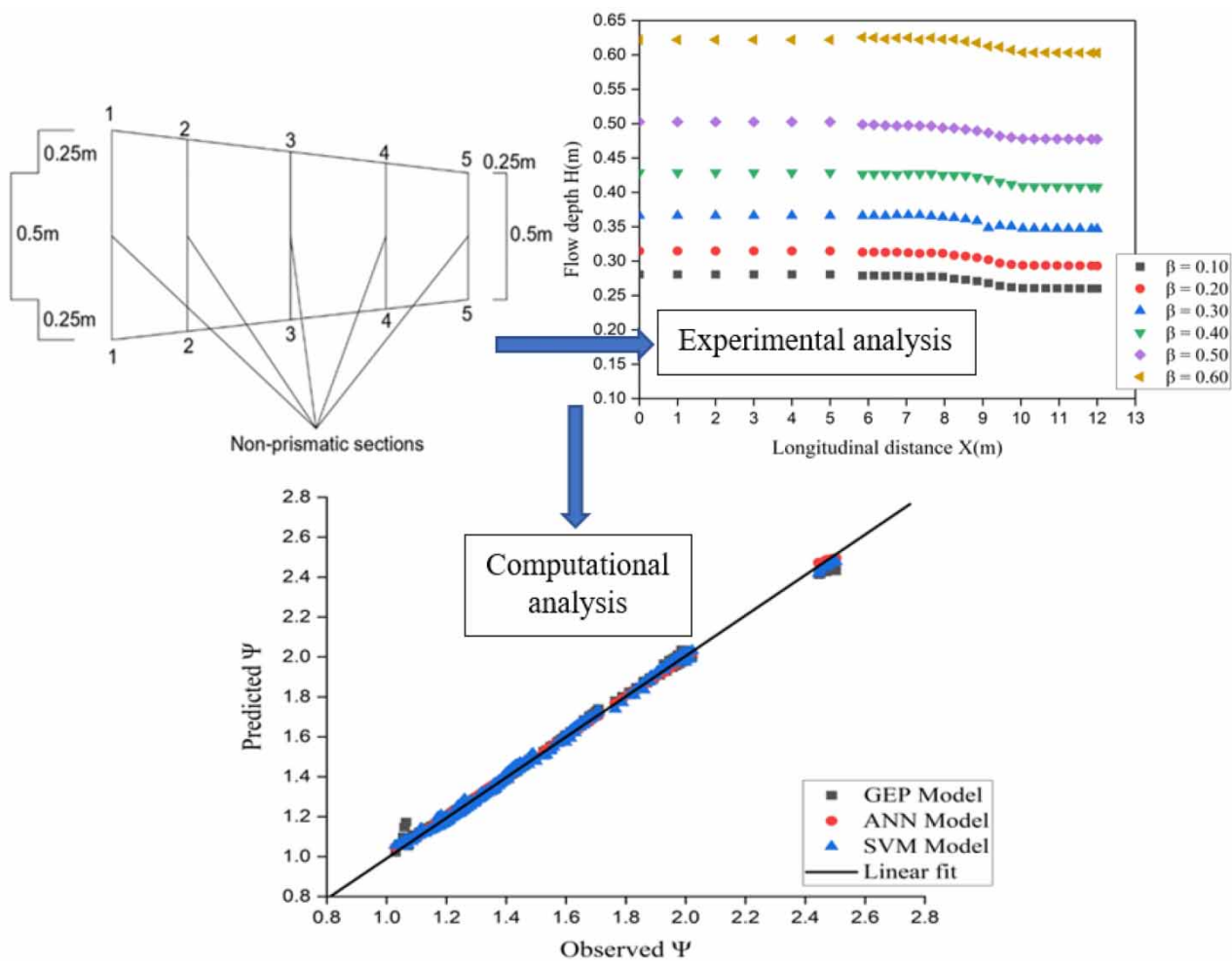
Accurate prediction of water surface profile in an open channel is the key to solving numerous critical engineering problems. The goal of the current research is to predict the water surface profile of a compound channel with converging floodplains using machine learning approaches, including Gene Expression Programming (GEP), Artificial Neural Networks (ANN), and Support Vector Machines (SVM), in terms of both geometric and flow variables, as past studies were more focused on geometric variables. A novel equation was also proposed using gene expression programming to predict the water surface profile. In order to evaluate the performance and efficacy of these models, statistical indices are used to validate the produced models for the experimental analysis. The findings demonstrate that the suggested ANN model accurately predicted the water surface profile, with coefficient of determination (R^2) of 0.999, root mean square error (RMSE) of 0.003, and mean absolute percentage error (MAPE) of 0.107%, respectively, when compared to GEP, SVM, and previously developed methods. The study confirms the application of machine learning approaches in the field of river hydraulics, and forecasting water surface profile of non-prismatic compound channels using a proposed novel equation by gene expression programming made this study unique.

Key words: geometric and flow parameters, machine learning techniques, nonprismatic compound channel, statistical analysis, water surface profile

HIGHLIGHTS

- The present study predicted the water surface profile in nonprismatic compound channels with the help of various machine learning approaches.
- The water surface profile is found to be affected by many nondimensional geometric and flow variables.
- The results suggest that the proposed ANN model have a high generalization capacity and do not show overtraining when applied to non-prismatic rivers.

GRAPHICAL ABSTRACT



NOTATION

The following symbols are used in this paper:

- B total width of compound channel;
- b width of the main channel;
- h height of the main channel;
- H flow depth;
- α width ratio (B/b);
- β relative flow depth $[(H - h)/H]$;
- δ aspect ratio (b/h);
- θ converging angle;
- X_r relative distance (x/L);
- L converging length;
- x distance between two consecutive sections;
- S_o longitudinal bed slope;
- S_e energy slope;
- Q_r discharge ratio (Q/Q_b);
- Q discharge at any depth;
- Q_b bankfull discharge;
- F_r Froude number;
- R_e Reynolds number;
- Ψ nondimensional water surface profile (H/h);

INTRODUCTION

Increased human settlements, buildings, and activities along river floodplains have resulted in severe repercussions during natural river floods due to the global population rise. River floods cause massive human casualties as well as economic damage. Flood catastrophes account for a third of all-natural disaster damages worldwide; flooding accounts for half of all fatalities, with trend analysis revealing that these percentages have dramatically grown (Berz 2000). Flood protection needs to predict the conveyance capacity of natural streams precisely. When water running through a channel exceeds the waterway's capacity, it results in flooding. Consequently, the requirement for precise flow parameter prediction during flood conditions to limit damage and save lives and property has piqued the interest of academics and engineers in recent years. Various methodologies and procedures have been used to aid precise measurement and forecast of river discharge, velocity distribution, shear stress distribution, and water surface level during overbank flows. Compound channels are the most common river feature during overbank flow. During the course of a river's flow, the geometry of the floodplain changes, resulting in a compound channel that is either converging or diverging. It is more challenging to replicate flow in a nonprismatic compound channel because more momentum is carried from the main channel to the floodplains. Sellin (1964), Myers & Elsayy (1975), Knight *et al.* (2010), and Khatua *et al.* (2012) have explored the flow models of straight and meandering prismatic two-stage channels, but little is known about nonprismatic compound channels. A converging channel shape causes the flow on floodplains to rise, while the flow on diverging floodplains is reduced (James & Brown 1977). Compound channels with symmetrically declining floodplains were studied by Bousmar & Zech (2002), Bousmar *et al.* (2004), Rezaei (2006), and Rezaei & Knight (2009) and found the extra loss of head and transfer of momentum from the main channel to floodplains. Asymmetric geometry with a greater convergence rate was examined by Proust *et al.* (2006). A greater convergence angle (22°) results in increased mass transfer and head loss. Chlebek *et al.* (2010) studied the flow behavior of skewed, two-stage converging, and diverging channels. They observed increased head losses due to the mass and momentum transfer, homogenization of the velocity on contracting floodplains, and increased velocity gradient on the expanding floodplains. Due to changes in the flow force from one subsection to another between the main channel and floodplains, there are noticeable variances in the flow distribution. In compound channels with nonprismatic floodplains, Rezaei & Knight (2011) investigated depth-averaged velocity, local velocity distributions, and boundary shear stress distributions at various convergence angles. The depth-averaged velocities show how contractions affect velocity distributions, specifically, an increase in velocity near the main channel walls, most notably in the second half of the convergence reach, and for high relative depth, the effects of the lateral flow that comes into the main channel. Yonesi *et al.* (2013) investigate the impact of floodplains' roughness on overbank flow in compound channels with nonprismatic floodplains. The velocity gradient between the main channel and the floodplain in the middle and end of the divergence stretch is lowered by raising the depth ratio or lowering the roughness ratio. The gradient of velocity increased as the angle of divergence increased. The gradient of shear stress rises as the surface roughness of the floodplain increases. Naik & Khatua (2016) developed a multivariate regression model to predict the water surface profile for different compound channels with the nonprismatic floodplain using nondimensional geometric factors. The constructed model has a high level of concordance with both the empirical evidence and the findings of other scholars. In nonprismatic compound channels, the effect of flow parameters on the water surface profile and variation of flow characteristics at higher flow rates has received much less attention. As a consequence, additional experiments are being carried out at higher flow rates on nonprismatic compound channel with converging floodplains to simulate the water surface profile.

It is complicated to evaluate the connections between the components that rely on other factors and those that are independent by developing a water surface profile model using mathematical, analytical, or numerical methods. These models end up being quite cumbersome and time-consuming as they progress. Not only is the amount of time spent conducting experiments cut in half as a result, but also the amount of time spent doing labor-intensive computations is reduced. Due to the increasing reliance on machine learning algorithms to estimate flow in compound channels, these channels are increasingly calculated using support vector machines (SVM), gene expression programming (GEP), artificial neural networks (ANN), fuzzy neural networks (ANFIS), and the M5 tree decision models (Seckin 2004; Unal *et al.* 2010; Sahu *et al.* 2011; Zahiri & Azamathulla 2014; Najafzadeh & Zahiri 2015; Parsaie *et al.* 2017). The GEP's capacity to generate mathematical correlations distinguishes it from other soft computing approaches such as ANN and SVM (Cousin & Savic 1997; Drecourt 1999; Savic *et al.* 1999; Whigham & Crapper 1999, 2001; Babovic & Keijzer 2002; Karimi *et al.* 2015). However, river engineering using the GEP method has received far less attention (Harris *et al.* 2003; Giustolisi 2004; Guven & Gunal 2008; Guven &

Aytek 2009; Azamathulla *et al.* 2013; Pradhan & Khatua 2017b). Parsaie *et al.* (2015) use the support vector machine (SVM) technique to predict the discharge in the compound open channel. The analysis of the error indices demonstrate that the SVM has the highest level of accuracy. Khuntia *et al.* (2018) developed an artificial neural network (ANN) model for predicting boundary shear stress distribution in straight compound channels using the most influential parameters such as width ratio, relative flow depth, aspect ratio, Reynolds number, and Froude number. Back-propagation neural network (BPNN) models performed well over global ranges of independent parameter values. For the estimation of discharge in diverging and converging compound channels, an equation has been created by Das *et al.* (2019). This equation encourages the usage of GEP. Mohanta & Patra (2021) have developed a model equation for calculating discharge in meandering compound channels, validating the use of GEP over the classic channel division technique. For meandering compound channels with relative roughness, Mohanta *et al.* (2021) used several AI methods, including multivariate adaptive regression splines (MARS), a group method of data handling Neural Network (GMDH-NN), and gene-expression programming (GEP), to develop model equations. Compared to GEP and MARS, the results show that the suggested GMDH-NN model accurately predicted the values. In order to simulate the flow of the Hablehroud River in north-central Iran, Esmaceli-Gisavandani *et al.* (2021) studied five hydrological models, including the soil and water assessment tool (SWAT), identification of unit hydrograph and component flows from rainfall, evapotranspiration, and streamflow (IHACRES), Hydrologiska Byrns Vattenbalansavdelning (HBV), Australian water balance model (AWBM). It was discovered that the calibration phase findings for SWAT, IHACRES, and HBV were good. Only the SWAT model, however, performed well and outperformed the other models throughout the validation phase. The GEP combination approach may combine model results from other, less accurate models to get a better indication of river flow. To forecast the snow depth (SD) one day in advance at the North Fork Jocko snow telemetry (SNOTEL) station in the city of Missoula, Montana State of the United States, Adib *et al.* (2021a, 2021b) proposed a novel algorithm that combines various wavelet transform (WT) approaches, including discrete wavelet transform (DWT), maximal overlap discrete wavelet transform (MODWT), and multiresolution-based MODWT (MODWT-MRA). The findings validated wavelet-based models' superiority over solo ones. This shows that the innovative wavelet-based model is a method that merits further investigation for its potential to deliver useful information across snow-covered areas. Adib *et al.* (2021a, 2021b) demonstrate the use of wavelet transforms, such as discrete wavelet transform, maximum overlap discrete wavelet transforms (MODWT), multiresolution-based MODWT (MODWT-MRA), and wavelet packet transform (WP), in combination with artificial intelligence (AI)-based models, such as multi-layer perceptron's, radial basis functions, adaptive neuro-fuzzy inference systems (ANFIS), and gene expression programming, to retrieve snow depth (SD) from the national snow and ice data center. According to the findings, the WP combined with ANFIS (WP-ANFIS) performed better than the other analyzed models in terms of statistical analysis. Mohseni & Naseri (2022) used ANN and SVM to estimate the water surface profile in compound channels with vegetated floodplains. According to the results, the SVM algorithm performed better than the ANN and regression models. According to sensitivity analysis, the water surface profile depended mainly on relative discharge and relative depth. Naik *et al.* (2022) proposed a novel equation using GEP to predict water surface profile in converging compound channels using geometric variables.

In past studies, the water surface profile of compound channels with converging floodplains was predicted in terms of geometric variables using the GEP technique only. Therefore, this research has been conducted to predict the water surface profile of compound channels with converging floodplains in terms of both geometric and flow parameters using machine learning techniques such as GEP, ANN, and SVM. Comparisons among these techniques and approaches from other researchers are made with the help of statistical analysis to evaluate the effectiveness of the developed models for predicting the water surface profile of nonprismatic compound channels.

MATERIALS AND METHODS

Data source

Numerous studies have investigated the shear force distribution, momentum transfer, and discharge methods of the flow in nonprismatic compound channels. Rezaei (2006), Naik & Khatua (2016), and the author's experimental data are used in the current work. Table 1 provides information on the experimental channel dimensions of these compound channels with various geometric properties. Figure 1 describes the flow chart of the methodology applied in the current study.

Table 1 | Details of experimental channel dimensions used in the present study

Verified Test Channel	Type of Channel	Angle of Convergent (θ)	Longitudinal Slope (S)	Cross-sectional Geometry	Total Channel Width (B) (m)	Main Channel Width (b) (m)	Main Channel Depth (h) (m)	Converging Length (X_c) (m)	Aspect ratio (δ)
Channel 1 Rezaei (2006)	Converging	11.31°	0.002	Rectangular	1.2	0.398	0.05	2	7.96
Channel 2 Rezaei (2006)	Converging	3.81°	0.002	Rectangular	1.2	0.398	0.05	6	7.96
Channel 3 Rezaei (2006)	Converging	1.91°	0.002	Rectangular	1.2	0.398	0.05	6	7.96
Channel 1 Naik & Khatua (2016)	Converging	5°	0.0011	Rectangular	0.9	0.5	0.1	2.28	5
Channel 2 Naik & Khatua (2016)	Converging	9°	0.0011	Rectangular	0.9	0.5	0.1	1.26	5
Channel 3 Naik & Khatua (2016)	Converging	12.38°	0.0011	Rectangular	0.9	0.5	0.1	0.84	5
Present Channel	Converging	4°	0.001	Rectangular	1.0	0.5	0.25	3.60	2

Experimental setup

The experiments were performed at the Hydraulics laboratory of the Department of Civil Engineering, Delhi Technological University. All experiments were conducted in a masonry flume of 12 m long, 1.0 m wide, and 0.8 m deep. In this flume, a compound cross section was constructed using brick masonry with a 0.5 m main channel wide and 0.25 m deep ([Figure 2](#)). The geometric characteristics of a two-stage channel are described in [Figure 2](#). The converging segment of the channel was built with the help of brick masonry, having a converging angle of $\theta = 4^\circ$. The compound channel has a prismatic section of 6 m long, a nonprismatic section of length of 3.6 m, and the rest is the downstream portion. The flume was run for six different flow rates. For each discharge, various flow characteristics, such as stage-discharge relationship, water surface profile, velocity distribution, shear stress distribution, etc., were measured in the prismatic section (PS) and various nonprismatic sections (NPS), as shown in [Figure 3](#). NPS1 and NPS5 are the sections at the start and end of the converging portion, whereas NPS3 is the middle of the converging section. The NPS2 represents the middle section between NPS1 and NPS3. Similarly, NPS4 represents the middle section between NPS3 and NPS5.

The subcritical flow regime was attained in several conditions of the two-stage channel with a longitudinal bed slope of 0.001. Based on data collected from in-bank and over-bank flows in the floodplains and main channel, Manning's n value was estimated. The system derives the water supply from an underground sump to an overhead tank in the experimental channel. The water from the channel is collected in a volumetric tank outfitted with a v-notch. This v-notch was calibrated to measure the discharge from the experimental channel. After that, it makes the flow back to the sump located underneath. [Figure 3](#) represents the experimental setup and the types of equipment that were employed in the research. [Figure 4](#) depicts a plan view of the nonprismatic cross-sections of [Rezaei \(2006\)](#), [Naik & Khatua \(2016\)](#), and the current channel. At the downstream end of the flume, a tailgate was installed to control the water surface profile and impose a specific flow depth in the flume portion. A point gauge with 0.1 mm precision was used to measure the water surface profile at a distance of 1.0 m and 0.3 m in the prismatic and nonprismatic portions, respectively. An Acoustic Doppler Velocimeter (ADV) was used to detect the average velocity of the cross-section and three-dimensional velocity distributions along the wetted perimeter at 2.5 and 10 cm vertical and horizontal intervals, respectively, as shown in grid form in [Figure 2](#). The data collected using the ADV were filtered using the Horizon ADV software. The lateral distributions of boundary shear stress were also measured using a Preston tube of 5 mm outer diameter at the same sections where the velocity distributions were tested. A digital manometer was used to measure the pressure difference. After that, [Patel's \(1965\)](#) calibration equations were used for calculating the shear stress values.

Theoretical background

River engineers need to make an accurate forecast of the water surface profile in overbank flows in order to successfully build drainage canals, flood defense systems, river training works, floodplain management, and other similar projects. It may be

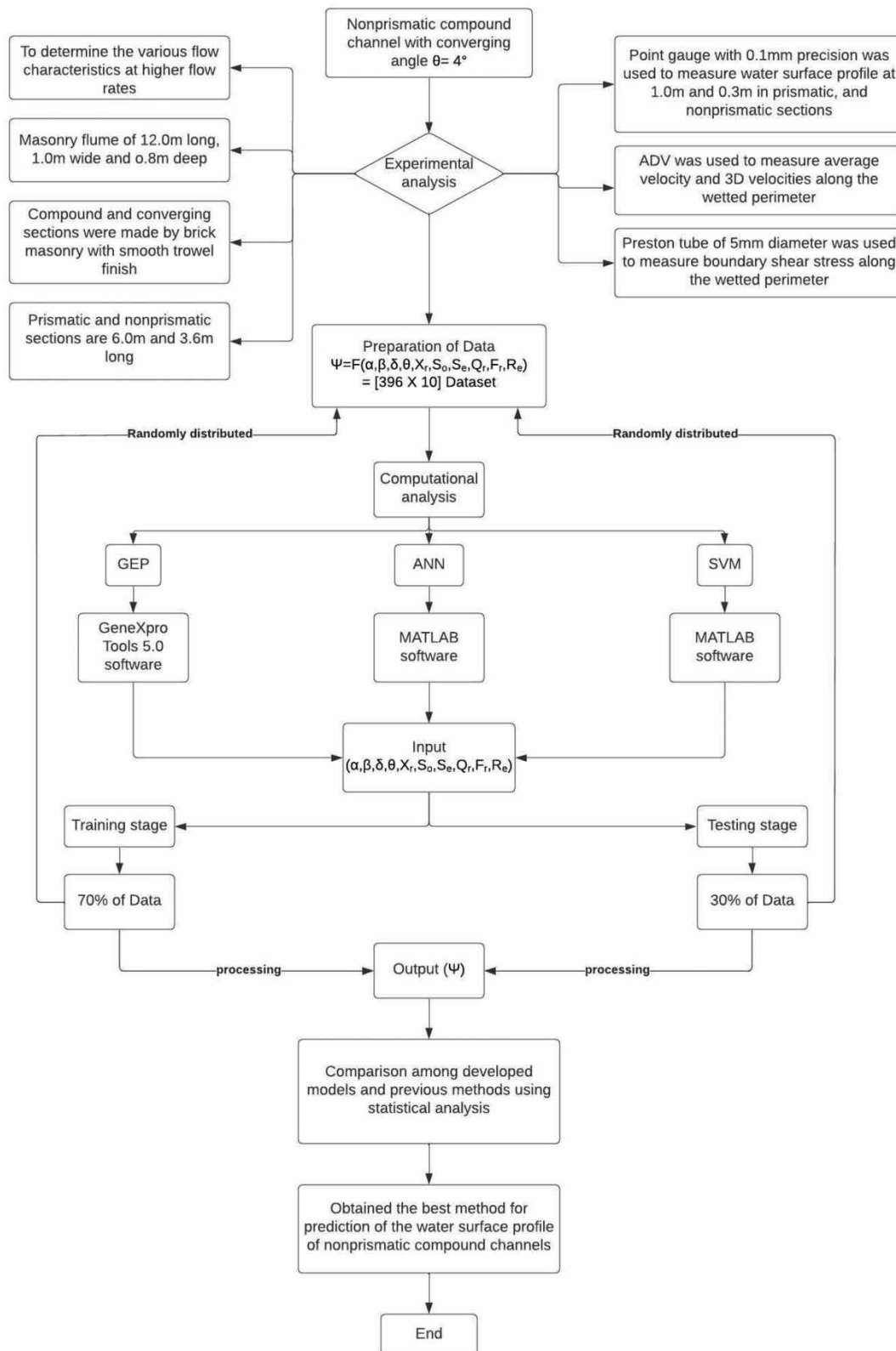


Figure 1 | Flow of methodology.

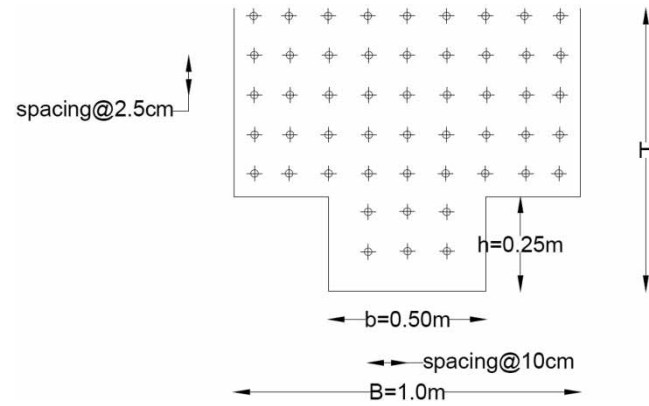


Figure 2 | Cross-section of a two-stage channel.

challenging to collect field data that is both sufficiently exact and comprehensive in natural rivers, especially when flood flow conditions are variable. Experiments conducted in a laboratory are essential in order to improve one's comprehension related to the water surface profile of compound channels that include both prismatic and nonprismatic floodplains. For the purpose of predicting the water surface profile of a converging compound channel, Naik & Khatua (2016) used multivariate analysis to suggest the Equations (1)–(21), while Naik *et al.* (2022) used the GEP technique to offer the Equation (22).

$$\Psi = 1.06\alpha^{0.22} \quad \text{for lower aspect ratio channel 1} \quad (1)$$

$$\Psi = 1.16\alpha^{0.22} \quad \text{for lower aspect ratio channel 2} \quad (2)$$

$$\Psi = 1.21\alpha^{0.29} \quad \text{for lower aspect ratio channel 3} \quad (3)$$

$$\Psi = 0.07\alpha + 1.78 \quad \text{for higher aspect ratio channel 1} \quad (4)$$

$$\Psi = 0.05\alpha + 1.28 \quad \text{for higher aspect ratio channel 2} \quad (5)$$

$$\Psi = 0.13\alpha + 1.25 \quad \text{for higher aspect ratio channel 3} \quad (6)$$

$$\Psi = -0.14X_r + 1.22 \quad \text{for lower aspect ratio channel 1} \quad (7)$$

$$\Psi = -0.15X_r + 1.32 \quad \text{for lower aspect ratio channel 2} \quad (8)$$

$$\Psi = -0.22X_r + 1.37 \quad \text{for lower aspect ratio channel 3} \quad (9)$$

$$\Psi = -0.15X_r + 2.01 \quad \text{for higher aspect ratio channel 1} \quad (10)$$

$$\Psi = -0.16X_r + 1.45 \quad \text{for higher aspect ratio channel 2} \quad (11)$$

$$\Psi = -0.21X_r + 1.67 \quad \text{for higher aspect ratio channel 3} \quad (12)$$

$$\Psi = -1.22 + 2.27\alpha^{0.22} + 0.18X_r \quad \text{for lower aspect ratio channel 1} \quad (13)$$

$$\Psi = -1.21 + 2.28\alpha^{0.22} + 0.19X_r \quad \text{for lower aspect ratio channel 2} \quad (14)$$

$$\Psi = -0.58 + 1.63\alpha^{0.29} + 0.18X_r \quad \text{for lower aspect ratio channel 3} \quad (15)$$

$$\Psi = -0.66 + 0.29\alpha + 0.12X_r \quad \text{for higher aspect ratio channel 1} \quad (16)$$

$$\Psi = 0.86 + 0.29\alpha + 0.11X_r \quad \text{for higher aspect ratio channel 2} \quad (17)$$

$$\Psi = 0.86 + 0.29\alpha + 0.12X_r \quad \text{for higher aspect ratio channel 3} \quad (18)$$

$$\Psi^*(\theta) = \frac{\text{Actual } \Psi}{\text{Eq. (12)}} \quad (19)$$

$$\frac{\text{Actual } \Psi}{\text{Eq. (12)}} = e^{0.0017\theta} \quad (20)$$

$$\Psi = e^{0.0017\theta}(-1.21 + 2.25\alpha^{0.22} + 0.18X_r) \quad (21)$$

$$\Psi = \left[\frac{3.72 + \delta\beta + S\beta}{3.72 - S - 0.12\alpha} + \frac{9.63 - \alpha\delta}{(5.73 + \theta) \times (\delta^2\beta + 2\delta X_r)} - 14.75\delta S - \alpha\delta S \right] \quad (22)$$

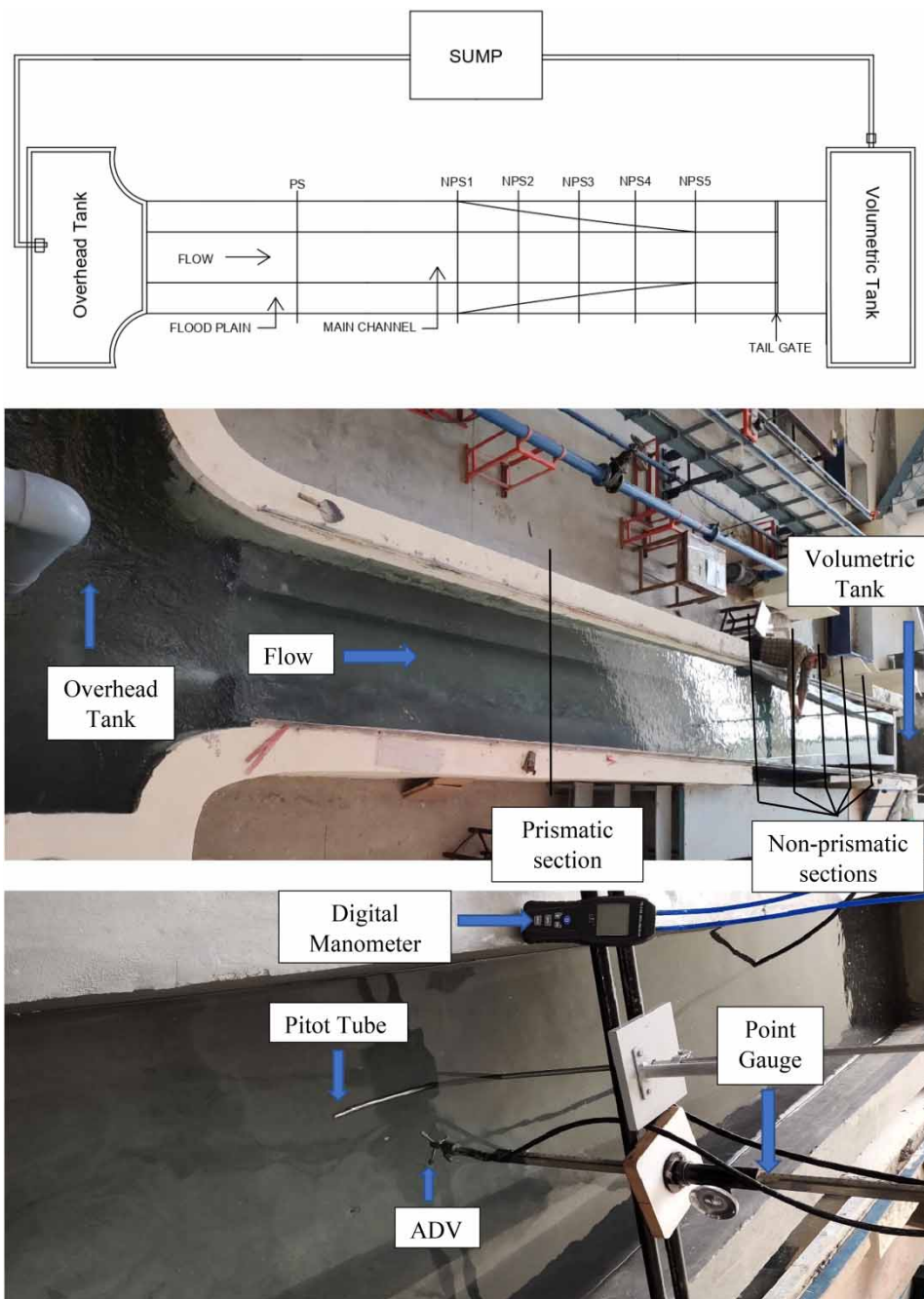


Figure 3 | Experimental setup.

The water surface profile of the compound channel with different converging floodplains was attempted to be predicted. The flow is found to be non-uniform in the nonprismatic section, contrary to be anticipated until it reaches the prismatic zone. These channels have a uniformly smooth surface on their main channel and floodplains. Manning's n values for all of these smooth surfaces are found to be 0.013. The majority of influencing factors, including width ratio (α), relative depth ratio (β), aspect ratio (δ), converging angle (θ), relative distance (X_r), longitudinal slope (S_o), energy slope (S_e), discharge ratio (Q_r), Froude's number (F_r) and Reynold's number (R_e), are taken into consideration when estimating the water surface profile of nonprismatic compound channels. A number of parameters are used to enable the model equation to be applied to diverse compound channels.

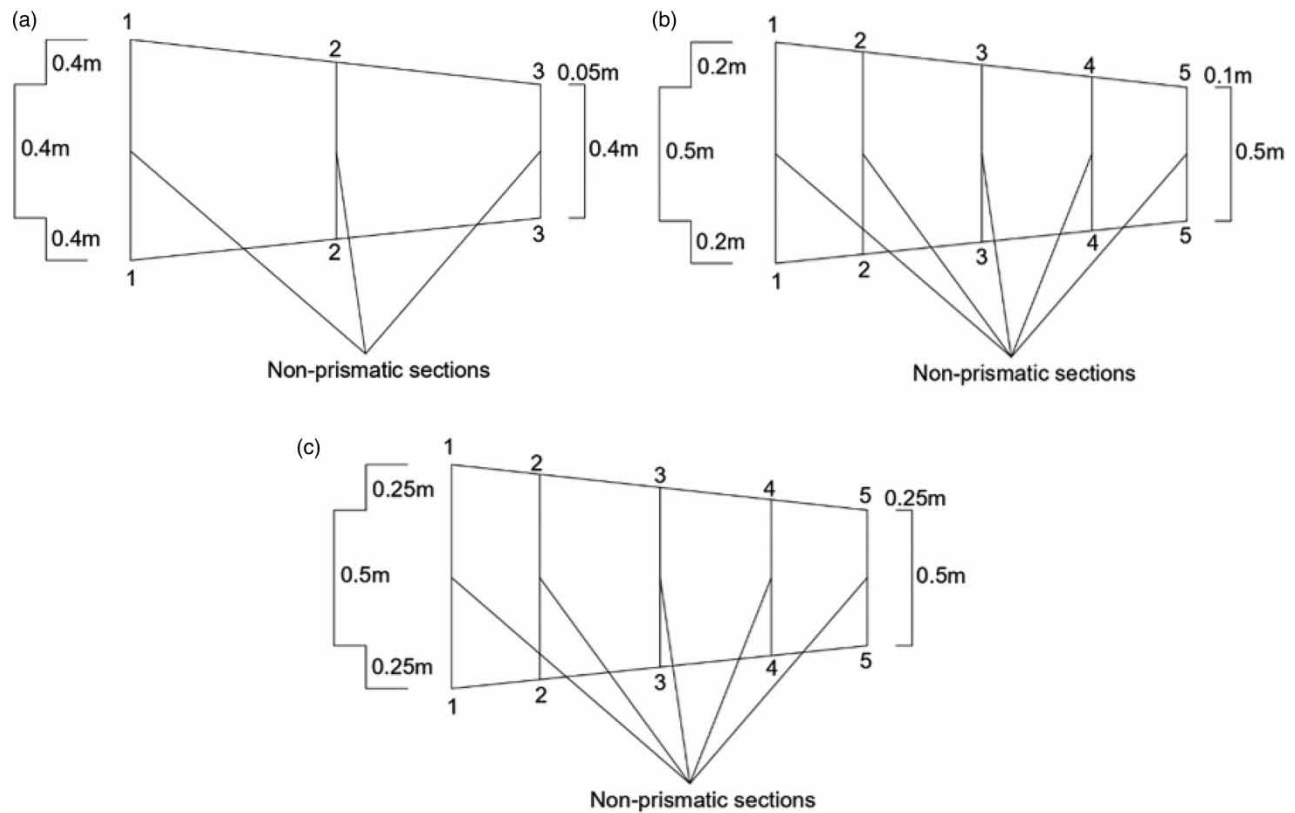


Figure 4 | Nonprismatic sections of (a) Rezaei (2006) (b) Naik & Khatua (2016) (c) Present channel.

The necessary dimensionless equation may be expressed as follows:

$$\Psi = F(\alpha, \beta, \delta, \theta, X_r, S_o, S_e, Q_r, F_r, R_e) \quad (23)$$

GEP model

Ferreira first put up the idea for the GEP approach in 1999. Genetic Programming (GP) and Genetic Algorithms (GA) are brought together in this approach. The plain and linear chromosomes are joined with structures of fixed length and branching that vary in size and form (similar to the parse tree in Genetic programming). Phenotype and genotype may be distinguished using this methodology due to the fact that all branch structures, despite their varying dimensions and configurations, are recorded in linear chromosomes of a predetermined length. Gene-expression programming, often known as GEP, is a multi-gene, one-of-a-kind coding language that permits the alteration of increasingly sophisticated equations by dividing them into many sub-equations. Gene generations, fitness-based gene selection, and the introduction of genetic diversity via using one or more genetic operators are also utilized in this process. The author develops an equation for the nondimensional water surface profile parameter (Ψ) of a non-prismatic compound channel by using a Gene-expression programming (GEP) technique with the assistance of [GeneXproTools 5.0 \(2014\)](#). The development of models is decided upon according to the appropriateness of the datasets used for training and testing. The selected models are recreated in GEP using one or more genetic operators, such as mutation or recombination. Previous studies provide a concise conceptual outline of GEP in their descriptions ([Mallick et al. 2020](#)). The relationship (Equation (23)) illustrates the water surface profile of nonprismatic compound channels varies as a function of the geometric and hydraulic factors. In this study, the modeling procedure uses Ψ as the target value and the ten independent factors ($\alpha, \beta, \delta, \theta, X_r, S_o, S_e, Q_r, F_r, R_e$) as input variables discussed in Equation (23). The structure of the model is built with the help of the four basic operators of arithmetic (+, −, ×, /). In all, 396 data sets are used, and they are dispersed at random over the two distinct stages of the modeling process. For the purpose of the present

investigation, training makes use of 70% of the data, while testing makes use of the remaining 30%. In this investigation, the root-mean-squared error (RMSE) served as the fitness function (E_i), and the fitness (f_i) was determined using Equation (24), which defined the total sum of all errors relative to the goal value. The initial model was developed using only one gene and two different head sizes as its starting point. The number of genes and heads was then gradually raised during each run, and the results of the training and testing datasets were recorded after each iteration. There was not a significant improvement in performance between the training data phase and the testing data phase for head lengths of more than twelve and more than six genes.

As a consequence of this, twelve were selected as head length to be included in the GEP model, and six genes were assigned to each chromosome. The addition served as the connecting function that was used to connect the genes. After 15,000 generations, the value of the fitness function and the coefficient of determination of the training and testing data had not changed, which was suggestive of the fact that generational progress may have reached its conclusion. Table 2 summarizes the primary characteristics that impact the success of GEP modeling and are used in the construction of a model for predicting the water surface profile of compound channels with converging floodplains. Figure 5 represents the GEP model for the water surface profile as an expression tree (ET). Within this representation, the input variables are denoted by d0 to d9, and the constant value for gene two is denoted by G2c0. In order to decode this expression tree, an algebraic equation (Equation (24)) was developed that links the output variables to the input variables.

In terms of the analytical form, the GEP is modeled according to the following:

$$\Psi = [Q_r - \frac{0.125\delta}{(\delta Q_r + \delta S_e) \times (\theta\delta - \alpha)} + F_r S_o - \alpha S_o + Q_r S_e S_o - \delta Q_r S_o - \alpha F_r S_e S_o + \alpha \delta F_r S_o + S_e - \frac{S_e^2}{(Q_r - 1)S_o} + 3.729\beta^2 + \frac{X_r\beta^2}{\alpha S_e R_e - 9.667 + \delta} - 5.671] \quad (24)$$

ANN model

An artificial neural network (ANN) is a significant computing approach that is making rapid strides in development. An artificial neural network is the association between easily handled components known as neurons or nodes, which are arranged

Table 2 | The GEP model's selection criteria and parameters

Description of Parameter	Parameter Setting
Function Set	+, −, ×, /
Number of Chromosomes	30
Head Size	12
Number of Genes	6
Gene Size	38
Linking Function	Addition
Fitness Function	RMSE
Program Size	80
Literals	37
Number of Generations	15,000
Constants per Gene	10
Data Type	Floating-point
Mutation	0.00138
Inversion	0.00546
Gene recombination rate	0.00277
Insertion sequence (IS) transposition rate	0.00546
Root insertion sequence (RIS) transposition rate	0.00546

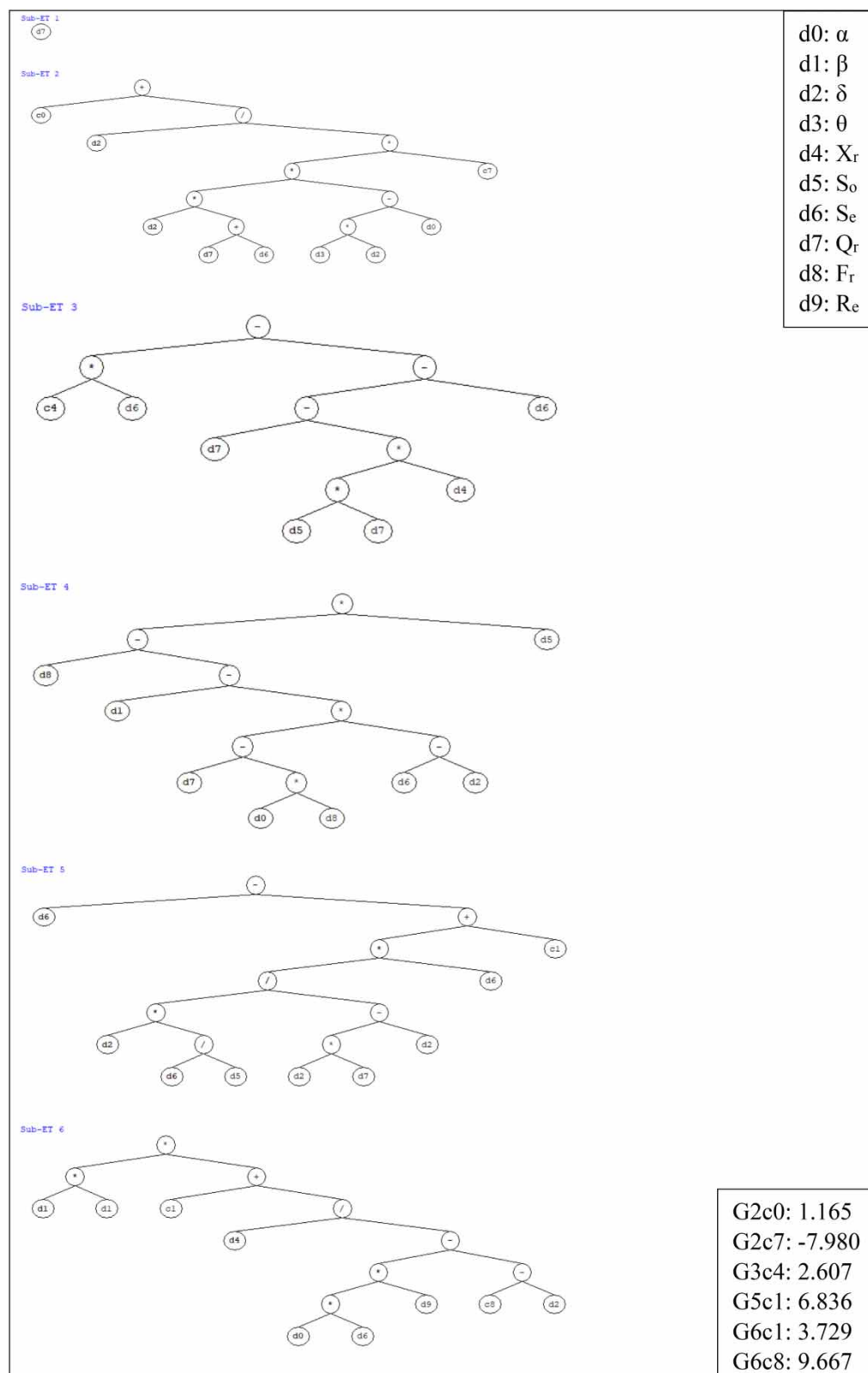


Figure 5 | GEP Expression Tree.

in a layered method. Figure 6 represents the architecture of an artificial neural network (ANN) which is a back-propagation (BP) neural network model with a feedforward design with three layers, consisting of I input neurons, m hidden neurons, and n output neurons. The data enters the network via the input layer. The hidden layer takes the data from the input layer, which is responsible for all the data processing, and the output layer receives data that has been processed from the network. The data from the output are sent to receptors on the outside. In an ANN, layers are connected to the resultant layers by means of interconnections between layers. These interconnections are referred to as weights and weighted values, respectively. To restrict the behavior of specified cost functions in an ANN, the weights of the interconnections are increased (Khuntia *et al.* 2018).

In this study, an ANN method was used to predict the water surface profile of a compound channel with converging flood-plains. For prediction purposes, a feedforward network is built on MATLAB using a backpropagation training technique. Figure 7 depicts the simulation process that the ANN goes through when it is operating inside the network. The network must first be trained before it can be used for any problem. During this phase, the weights and biases of each output neuron are adjusted in accordance with the training procedure so that the goal output of each output neuron is constrained. Training in ANNs is comprised of three different components: weights between neurons, which characterize the relative significance of the sources of input, a sigmoid transfer function, which controls the stage of the output from a neuron and an arrangement of learning laws, which depicts how the weights are changed throughout the training process. During training, a nonlinear function, which is most often a sigmoid function, is used (Govindaraju 2000a).

$$F(a) = \frac{1}{1 + e^{-a}} \quad (25)$$

where a = sum of weighted input value plus bias. The result is sent to the subsequent layer of nodes to be processed. There are four stages in feedforward back-propagation neural network (BPNN) approaches. These processes are as follows:

1. Sum the weighted input

$$Nod_z = \sum_{i=1}^n (W_{xz}k_x) + \epsilon_z \quad (26)$$

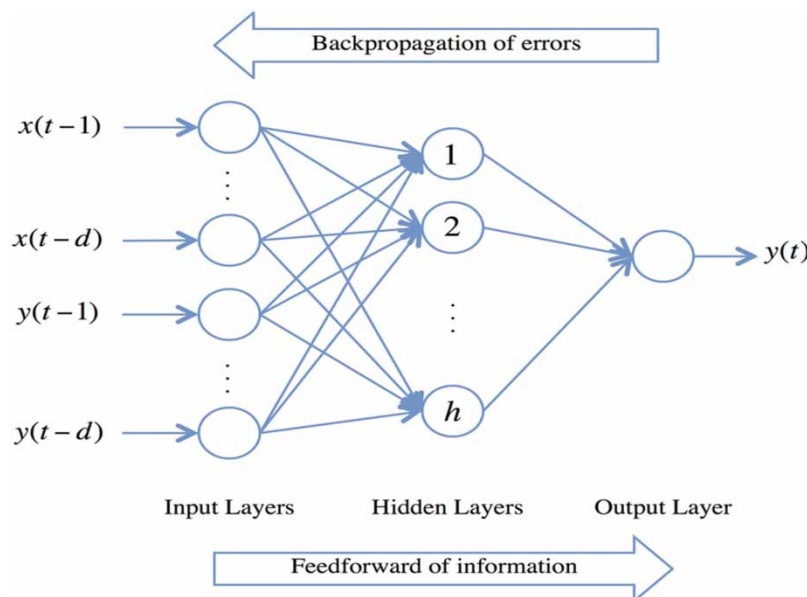


Figure 6 | ANN architecture.

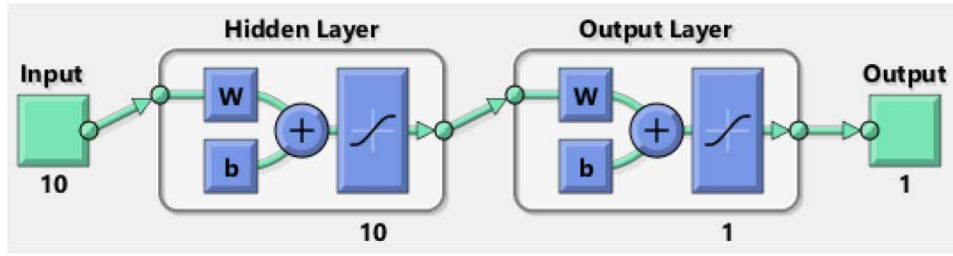


Figure 7 | Simulation process of ANN model.

where Nod_z = sum for the z th hidden node, n = total number of input nodes, W_{xz} = connection weight between the x th input and the z th hidden node, k_z = normalized input at the x th input node, and ϵ_z = bias value at the z th hidden node.

2. Transform the weighted input

$$Out_z = \frac{1}{1 + e^{-Nod_z}} \quad (27)$$

where Out_z = output from the z th hidden node.

3. Sum the hidden node output

$$Nod_y = \sum_{j=1}^m (W_{zy} Out_y) + \epsilon_y \quad (28)$$

where Nod_y = sum of y th output node, m = total number of hidden nodes, W_{zy} = connection weights between the z th hidden node and the y th output node, and ϵ_y = bias value at the y th output node.

4. Transform the weighted sum

$$Out_y = \frac{1}{1 + e^{-Nod_y}} \quad (29)$$

where Out_y = output at the y th output node.

In order to create the structure of the neural network, ten neurons were used for the input layers, ten neurons were used for the hidden levels, and one neuron was used for the output layer. Figure 8 shows the neural network parameters like gradient, mean, and validation checks of this system, along with the alterations that occur when the system is in the training stage. As the number of generations increases, the gradient and mean decreases, and the number of validation checks increases, leading to convergence of the model. Figure 9 depicts the error monitored in the training, testing, and validation phase; the error of training fell quickly at the beginning stage and then progressively slowed down after that. The network reached convergence after 43 generations and shows the minimum value of the error.

SVM model

SVMs, or support vector machines, are a group of similar supervised learning algorithms that may be used for classification and regression. A non-linear classifier offers superior accuracy in many different types of applications. First, in SVR, the input x is mapped into an m -dimensional feature space using a fixed (nonlinear) mapping. Next, a linear model is formed in this feature space utilizing the information obtained from the previous step (Parsaie *et al.* 2015). The naive way of making a non-linear classifier out of a linear classifier is to map our data from the input space X to a feature space F using a non-linear function $\varphi : x \rightarrow f$. In the space F the discriminate function is:

$$f(x) = w^T \varphi(x) + b \quad (30)$$

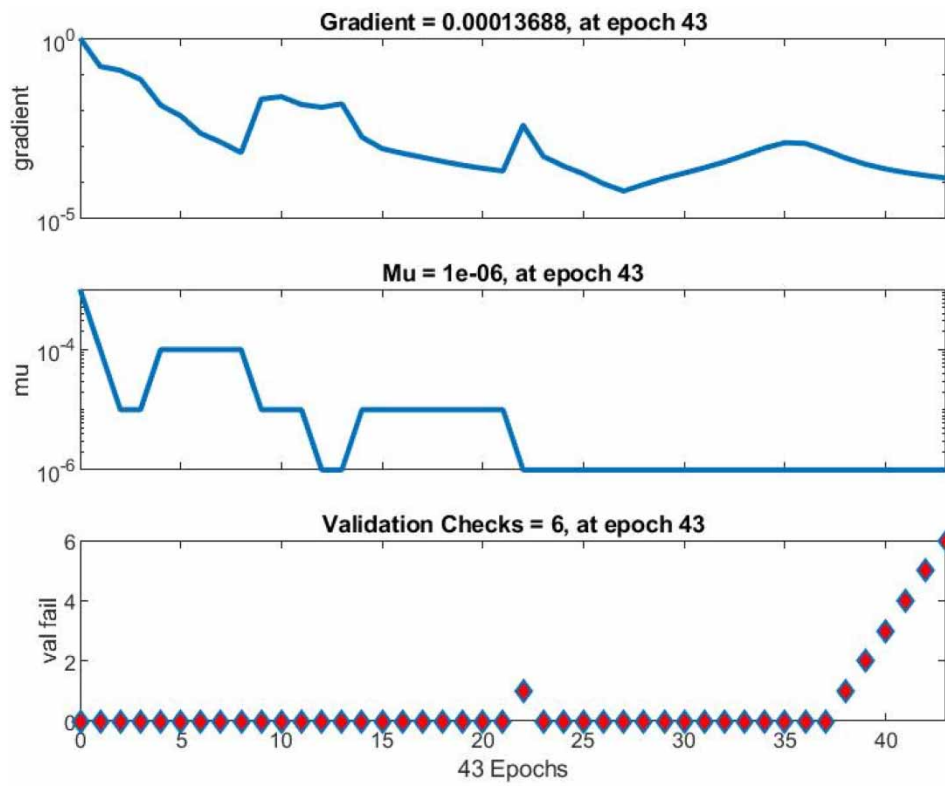


Figure 8 | Training parameters during ANN modeling.

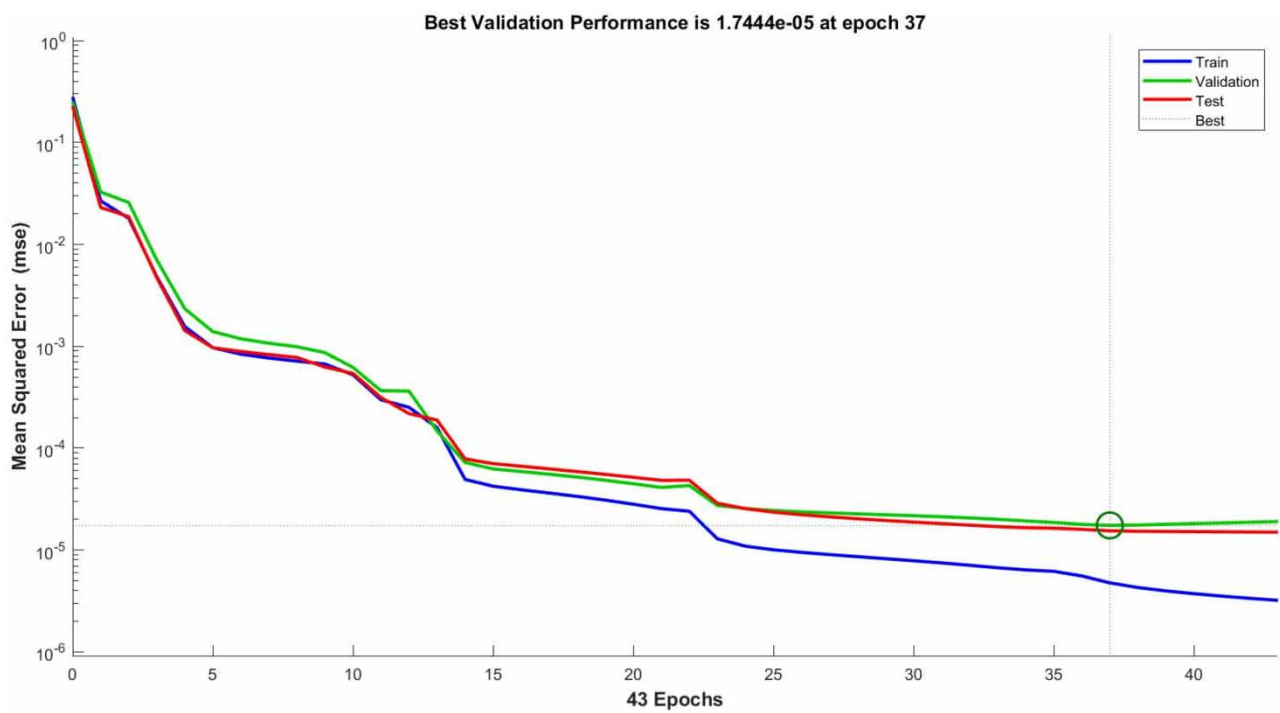


Figure 9 | Convergence curve for the training of BPNN.

The linear model, denoted by the mathematical notation $f(x, w)$, may be represented in the feature space as follows:

$$w = \sum_{i=1}^n \alpha_i x_i \quad (31)$$

$$f(x, w) = \sum_{i=1}^n \alpha_i x_i \varphi_i(x) + b \quad (32)$$

$$f(x) = \sum_{i=1}^n \alpha_i x_i^T x + b \quad (33)$$

In the feature space, F , this expression takes the form:

$$f(x) = \sum_{i=1}^n \alpha_i \varphi(x_i)^T \varphi(x) + b \quad 0 \leq \alpha_i \leq C \quad (34)$$

$$k(x, x') = \varphi(x)^T \varphi(x') \quad (35)$$

$$f(x) = \sum_{i=1}^n \alpha_i k(x, x_i) + b \quad (36)$$

Since SVM contains a large number of kernel functions, determining how to choose an effective kernel function is another research challenge. On the other hand, there are other helpful kernel functions that may be used in general.

Linear kernel: $k(x_i, x_j) = x_i^T x_j$

Polynomial kernel: $k(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \quad \gamma > 0$

RBF kernel: $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0$

Sigmoid kernel: $k(x_i, x_j) = \tanh(\gamma x_i^T x_j + r), \quad \gamma > 0$

where C, γ, r , and d are kernel parameters. It is common that the SVM generalization performance (estimation accuracy) is directly proportional to the quality of the settings for the meta-parameters C, γ , and r , as well as the kernel parameters. The complexity of the prediction (regression) model is controlled by the values chosen for C, γ , and r . The issue of optimum parameter selection is even more difficult because the complexity of the SVM model (and, therefore, its generalization performance) is dependent on all three parameters. In order to carry out the classification, kernel functions are called upon to modify the dimensionality of the input space. Data sets are used to develop the SVM model, which is comparable to the development of other neural network models. To accomplish this goal, a total of 396 data sets about the water surface profile of converging compound channels were used. In order to construct the SVM model, the acquired data is split into two categories in MATLAB software: the training data and the testing data. Equation (23) demonstrates the primary factors that affect the profile of the water's surface.

Statistical measures

For further testing of the accuracy of the developed models by GEP, ANN, and SVM approach, various types of error analysis, such as the coefficient of determination (R^2), mean absolute error (MAE), mean absolute percentage error (MAPE), and root

mean squared error (RMSE), are analyzed using the following equations (Naik *et al.* 2022).

$$R^2 = \frac{\sum_{i=1}^N (a_i - \bar{a})^2 (p_i - \bar{p})^2}{\sum_{i=1}^N (a_i - \bar{a})^2 \sum_{i=1}^N (p_i - \bar{p})^2} \quad (37)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |p_i - a_i| \quad (38)$$

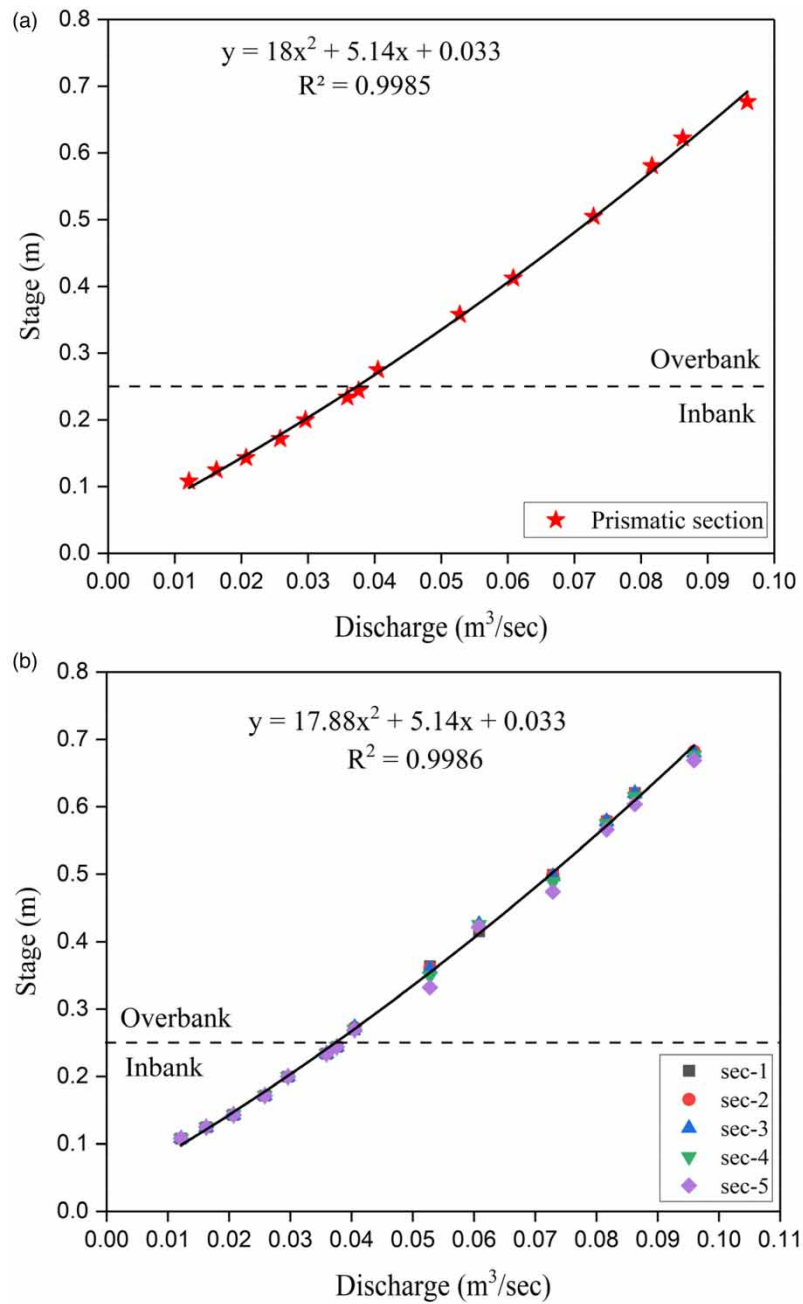


Figure 10 | Stage-discharge relationship for the nonprismatic compound channel (a) prismatic section (b) nonprismatic sections.

$$\text{MAPE}(\%) = \frac{1}{N} \sum_{i=1}^N \left(\frac{|p_i - a_i|}{a_i} \times 100 \right) \quad (39)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (p_i - a_i)^2} \quad (40)$$

where a and p are the actual and predicted values, respectively, \bar{a} and \bar{p} are the mean of actual and predicted values, respectively, and N is the number of datasets.

RESULTS AND DISCUSSION

Figure 10 represents the stage-discharge relationship for prismatic and nonprismatic sections of the nonprismatic compound channel with a converging angle of $\theta = 4^\circ$. The flow depth rises as the discharge increases. However, there is a minor fall in increment beyond bankfull depth due to interaction and additional momentum transfer between the main channel and floodplains. Due to the convergence of channel geometry, flow depth reduces for the same discharge in the converging portion from section 1 to section 5. Therefore, for in-bank and overbank flow in prismatic and nonprismatic parts, the best-fitted trend lines are found to be a polynomial function with high R^2 values. Figures 11 depict the water surface profile of the nonprismatic compound channel with the longitudinal distance for different relative flow depths. In the prismatic part of the flume, the water surface profile stays the same, but in the converging portion of the flume, there is a decreasing water level due to the flow acceleration (especially in the second half of the transition), and in the downstream part of the flume, the flow is nearly uniform with some fluctuations. Figure 12 shows the variation of the nondimensional water surface profile (Ψ) with nondimensional geometric and flow parameters such as width ratio, relative depth, aspect ratio, relative distance, energy slope, discharge ratio, Froude number, and Reynolds number for converging angle $\theta = 4^\circ$ and relative depths ranging from $\beta = 0.10$ to 0.60 , respectively. The water surface profile increases as the width ratio increases due to a rise in stage for a particular width ratio. For different converging angles, the shape of the water's surface follows the same pattern of change with respect to the width ratio. The water surface profile increases non-linearly as the relative flow depths rises. Along the several parts of the nonprismatic compound channel, the water surface profile follows a pattern of variation with an aspect ratio that decreases as it moves from section to section. As demonstrated in the figure, the relative distance between two points causes the impact of relative distance on nondimensional water surface profiles to decrease as the relative distance between the two points increases. This indicates that a converging transition rapidly increases the velocity head

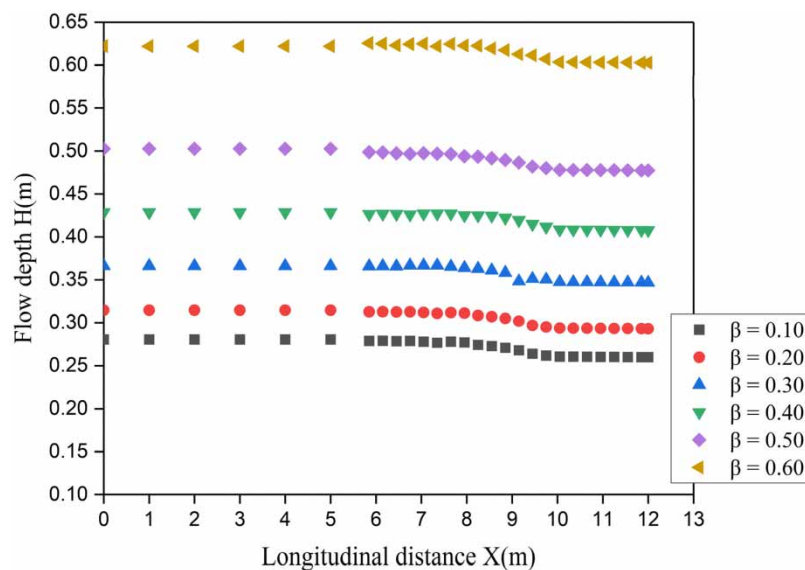


Figure 11 | Water surface profile for nonprismatic compound channel for different relative flow depths.

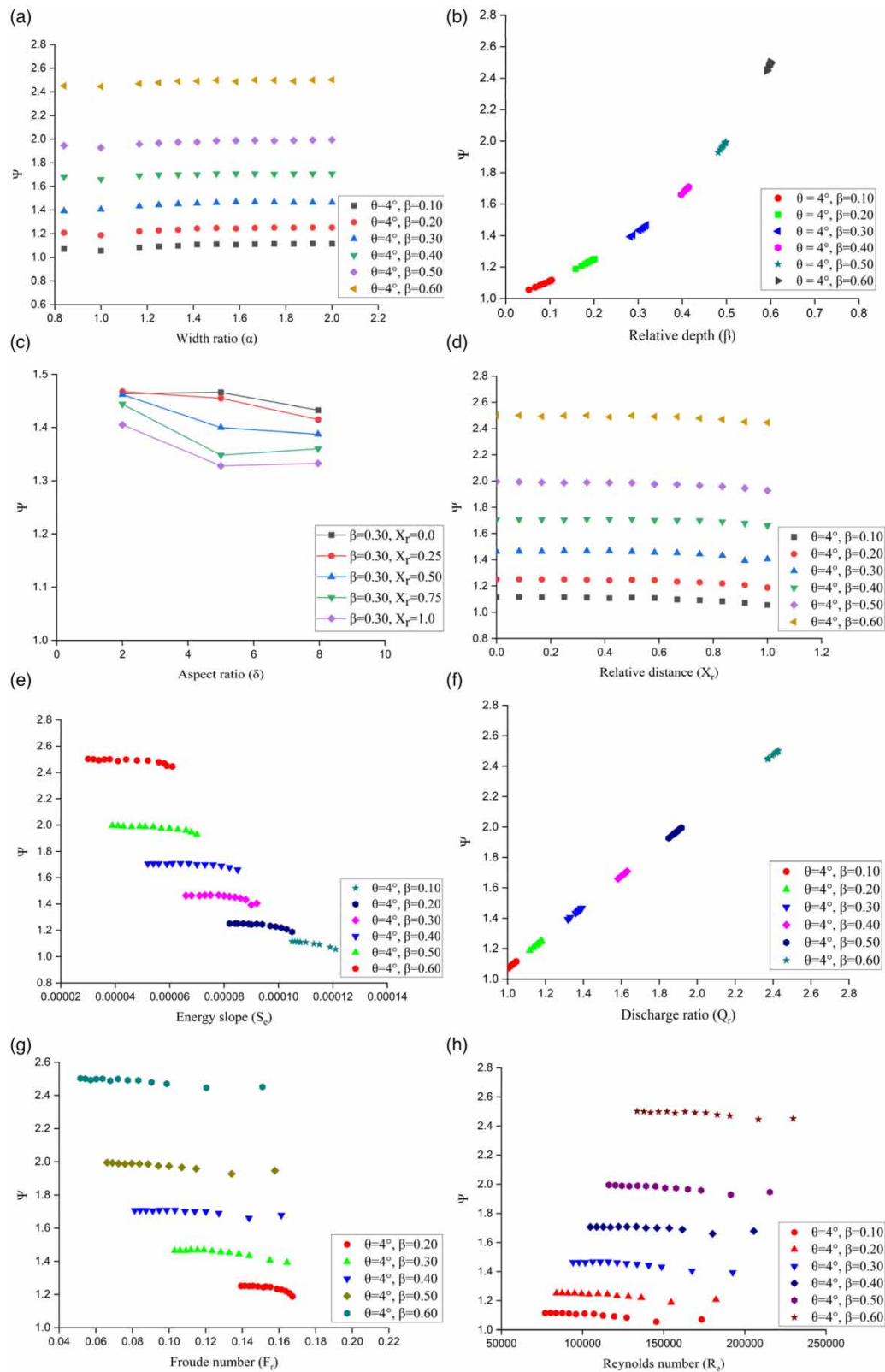


Figure 12 | Variation of nondimensional water surface profile with various parameters (a) width ratio (b) relative depth (c) aspect ratio (d) relative distance (e) energy slope (f) discharge ratio (g) Froude number (h) Reynolds number.

while simultaneously lowering the potential head. The drop is quite precipitous at the more acute angles of convergence of the floodplain. The water surface profile decreases as the energy slope increases due to the loss of energy along the flow length. It increases linearly with the discharge ratio due to a rise in flow depth with the increase of flow rate. The same pattern of variation is observed for all the relative depths; however, the decline is found to be steeper for higher converging angles due to more flow resistance from the channel convergence. The water surface elevation decreases as the Froude number, and Reynolds number increases with the same trend of variation for the different relative depths. The rise in velocity due to the convergence of channel geometry leads to flow acceleration in the nonprismatic sections, causing the water surface profile to decline.

The scatter plots comparing the predicted and observed Ψ values for each of the different ML techniques are shown in Figure 13. The fact that the values are so near to the line indicating good agreement, is a strong indicator of the generated GEP, ANN, and SVM models' capacity to make accurate predictions. Among all the three models, ANN predicted values are very close to the best-fitted line compared to GEP and SVM, having scatter values along the fitted line. Therefore, the ANN model best agrees with the experimental data with high R^2 values. Figure 14 compares developed models for predicting water surface profile with the previous methods developed by Naik *et al.* (2022) and Naik & Khatua (2016). It has been noticed that the constructed models are very close to the best-fitted line compared to previous methods and have a strong potential for generalization. They do not exhibit any signs of the phenomenon of overtraining. The effectiveness of the

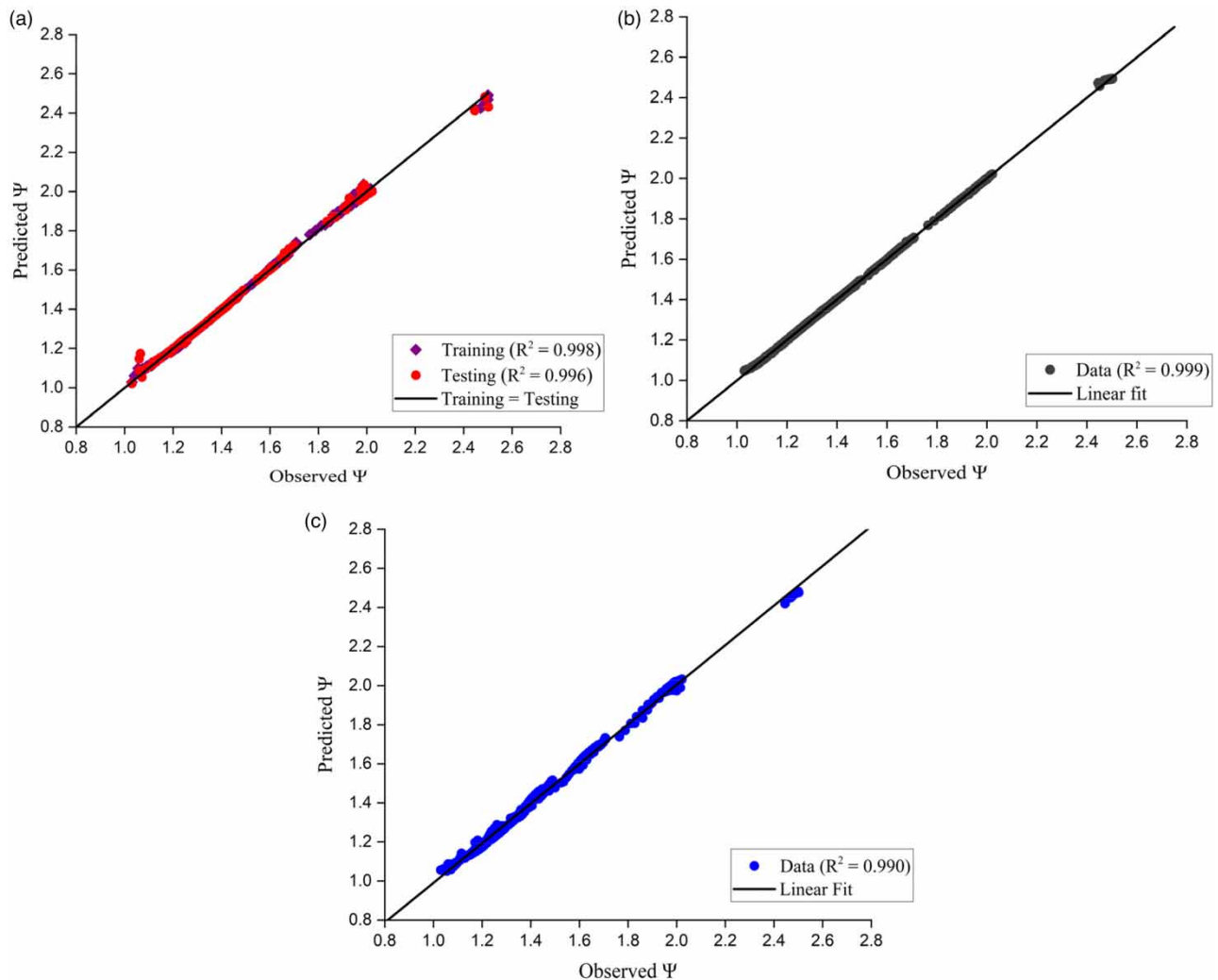


Figure 13 | Scatter plots of observed and predicted Ψ for various ML models (a) GEP (b) ANN (c) SVM.

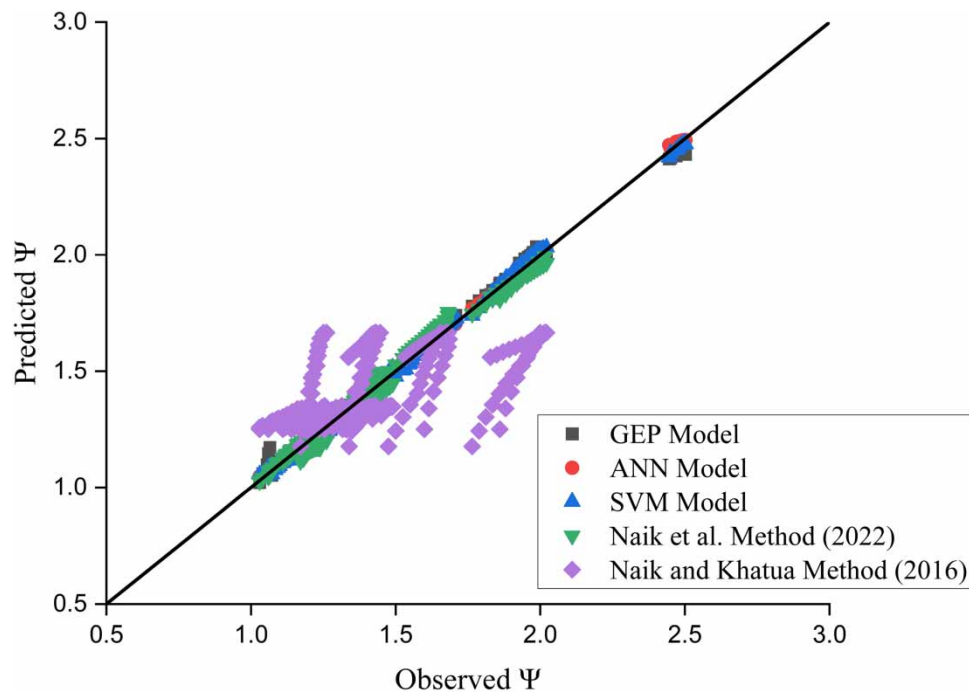


Figure 14 | Comparison of predicted value of Ψ for different models.

GEP, ANN, and SVM models were judged based on a number of statistical parameters, including R^2 , MSE, RMSE, MAE, and MAPE, which are shown in Table 3. From the table, it can be observed that ANN ($R^2 = 0.999$ and $RMSE = 0.003$) gives the best predicting results for the water surface profile of nonprismatic compound channels as compared to other models such as GEP ($R^2 = 0.998$ and $RMSE = 0.013$ in training, $R^2 = 0.996$ and $RMSE = 0.018$ in testing), SVM ($R^2 = 0.990$ and $RMSE = 0.017$) and other previously developed methodologies. ANN model with MAPE of 0.107 proves to be the most suitable method as compared to other methods in the prediction of water surface profile in nonprismatic compound channels. To reduce the losses due to flood, understanding the flow mechanism in prismatic and nonprismatic (additional momentum transfer should also account in flow modeling) reaches of the river is important in designing flood control and diversion structures. The models developed in the study can have a practical application to nonprismatic rivers such as the River Main in Northern Ireland, the Brahmaputra River in India, and other similar rivers. The findings of the study will be useful in the design of such structures and thereby reducing economic as well as human losses.

CONCLUSIONS

This study demonstrates the application of machine learning strategies, specifically Artificial Neural Networks (ANN), Gene-Expression Programming (GEP), and Support Vector Machine (SVM), to determine the water surface profile of a

Table 3 | Error analysis of predicted Ψ by various approaches

Statistical parameters	ANN Model	SVM Model	GEP Model		Naik et al. Method (2022)		Naik and Khatua Method (2016)
			Training	Testing	Training	Testing	
R^2	0.999	0.990	0.998	0.996	0.99	0.99	0.896
MSE	0.0001	0.0003	0.0002	0.0004	0.0008	0.0007	0.0019
RMSE	0.003	0.017	0.013	0.018	0.028	0.027	0.043
MAE	0.002	0.015	0.008	0.010	0.022	0.022	0.002
MAPE(%)	0.107	1.074	0.595	0.598	1.543	1.546	2.429

nonprismatic compound with converging floodplains. The proposed models are developed based on 396 high-quality laboratory datasets with dimensionless geometric and flow parameters for nonprismatic compound channels with different converging angles ($\theta = 1.91^\circ$ to 12.38°) and relative depths ($\beta = 0.10$ to 0.60). The following are some of the findings and inferences that may be drawn from this study:

The proposed model appears to be influenced by many parameters such as width ratio, relative flow depth, aspect ratio, converging angle, relative distance, longitudinal slope, energy slope, discharge ratio, Froude number, and Reynolds number. Flow depth rises as discharge increases up to bankfull depth, but beyond bankfull depth, a modest decrease in depth is seen at all converging angles owing to interaction and momentum transfer between the main channel and floodplains. Due to the convergence of the channel geometry, the flow depth decreases with the length of the channel, and the same tendency has been seen for greater relative depths and varied floodplain convergence angles. The nondimensional water surface profile is found to be increasing with width ratio, relative depth, discharge ratio and shows a decreasing trend with aspect ratio, energy slope, relative distance, Froude number, and Reynolds number. For all the converging angles, the same trend of variation is observed for the water surface profile in nonprismatic compound channels. The link between the nondimensional water surface profile and the nondimensional geometric and hydraulic variables of a converging compound channel is examined. It is found that there is a nonlinear relationship between all of the parameters.

In contrast to previous methods, such as Naik & Khatua (2016) and Naik *et al.* (2022), the developed models show better results in terms of R^2 , MAE, RMSE, and MAPE for various datasets. The findings demonstrated that, in accordance with the assessment criteria, all techniques (ANN, GEP, and SVM) could reasonably predict the water surface profile in nonprismatic compound channels. The ANN model showed better performance due to the highest R^2 (0.999), lowest RMSE (0.003), MAE (0.002), and MAPE (0.107). However, a novel equation was developed using the GEP method for estimating the water surface profile of nonprismatic compound channels, as the GEP method also showed good statistical performance. The model's restriction is that they can only be used to forecast the water surface profile of a compound channel with a converging floodplain with uniform roughness. Future studies must be focused on estimating the water surface profile of nonprismatic compound channels with rough floodplains and new techniques.

ACKNOWLEDGEMENTS

The authors acknowledge the support from the Department of Civil Engineering, Delhi Technological University, Delhi, India.

FUNDING

Not applicable.

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

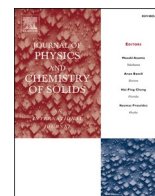
REFERENCES

- Aadib, A., Zaerpour, A. & Lotfird, M. 2021a On the reliability of a novel MODWT-based hybrid ARIMA-artificial intelligence approach to forecast daily snow depth (Case study: the western part of the Rocky Mountains in the U.S.A). *Cold Reg. Sci. Technol.* **189**, 103342. <https://doi.org/10.1016/j.coldregions.2021.103342>.
- Aadib, A., Zaerpour, A., Kisi, O. & Lotfird, M. 2021b A rigorous wavelet-packet transform to retrieve snow depth from SSMIS data and evaluation of its reliability by uncertainty parameters. *Water Resour. Manage.* **35** (9), 2723–2740. <https://doi.org/10.1007/s11269-021-02863-x>.
- Azamathulla, H. M., Ahmad, Z. & Ghani, A. A. 2013 An expert system for predicting Manning's roughness coefficient in open channels by using gene expression programming. *Neural Comput. Appl.* **23** (5), 1343–1349.
- Babovic, V. & Keijzer, M. 2002 Rainfall runoff modelling based on genetic programming. *Hydrol. Res.* **33** (5), 331–346.
- Berz, G. 2000 Flood disasters: lessons from the pastworries for the future. *Proc. Inst. Civ. Eng. Water Marit. Energy* **142** (1).
- Bousmar, D. & Zech, Y. 2002 Periodical turbulent structures in compound channels. In: *River Flow International Conference on Fluvial Hydraulics*, Louvain-la-Neuve, Belgium, pp. 177–185.

- Bousmar, D., Wilkin, N., Jacquemart, J. H. & Zech, Y. 2004 Overbank flow in symmetrically narrowing floodplains. *J. Hydraul. Eng. ASCE* **130** (4), 305–312.
- Chlebek, J., Bousmar, D., Knight, D. W. & Sterling, M. A. 2010 Comparison of overbank flow conditions in skewed and converging/diverging channels. In *River Flows International Conference*, pp. 503–511.
- Cousin, N. & Savic, D. A. 1997 A Rainfall-Runoff Model Using Genetic Programming. *Centre for Systems and Control Engineering*, Rep. No. 97, 3.
- Das, B. S., Devi, K. & Khatua, K. K. 2019 Prediction of discharge in converging and diverging compound channel by gene expression programming. *J. Hydraul. Eng.* doi:10.1080/09715010.2018.1558116.
- Drecourt, J. P. 1999 Application of neural networks and genetic programming to rainfall runoff modeling. *Water Resour. Manage.* **13** (3), 219–231.
- Esmaili-Gisavandani, H., Lotfirad, M., Sofla, M. S. D. & Ashrafzadeh, A. 2021 Improving the performance of rainfall-runoff models using the gene expression programming approach. *J. Water Clim. Change* **12** (7), 3308–3329. <https://doi.org/10.2166/wcc.2021.064>.
- Gepsoft, G. 2014 Version 5.0.
- Giustolisi, O. 2004 Using genetic programming to determine Chezy resistance coefficient in corrugated channels. *J. Hydroinf.* **6** (3), 157.
- Govindaraju, R. S. 2000a Artificial neural networks in hydrology. I: preliminary concepts. *J. Hydraul. Eng.* **5** (2), 115–123.
- Guvén, A. & Aytekin, A. 2009 New approach for stage-discharge relationship: gene-expression programming. *J. Hydraul. Eng.* **14** (8), 812–820.
- Guvén, A. & Gunal, M. 2008 Genetic programming approach for prediction of local scour downstream of hydraulic structures. *J. Irrig. Drain. Eng.* **134** (2), 241–249.
- Harris, E. L., Babovic, V. & Falconer, R. A. 2003 Velocity predictions in compound channels with vegetated floodplains using genetic programming. *Int. J. River Basin Manage.* **1** (2), 117–123.
- James, M. & Brown, R. J. 1977 Geometric parameters that influence floodplain flow. In *U.S. Army Engineer Waterways Experimental Station*, June, Vicksburg Miss. Research report H-77.
- Karimi, S., Shiri, J., Kisi, O. & Shiri, A. A. 2015 Short-term and long-term streamflow prediction by using ‘wavelet-gene expression’ programming approach. *ISH J. Hydraul. Eng.* **22** (2), 148–162.
- Khatua, K. K., Patra, K. C. & Mohanty, P. K. 2012 Stage-discharge prediction for straight and smooth compound channels with wide floodplains. *J. Hydraul. Eng. ASCE* **138** (1), 93–99.
- Khuntia, J. R., Devi, K. & Khatua, K. K. 2018 Boundary shear stress distribution in straight compound channel flow using artificial neural network. *J. Hydraul. Eng.* **23** (5), 04018014. doi:10.1061/(asce)he.1943-5584.0001651.
- Knight, D. W., Tang, X., Sterling, M., Shiono, K. & McGahey, C. 2010 Solving open channel flow problems with a simple lateral distribution model. *River Flow* **1**, 41–48.
- Mallick, M., Mohanta, A., Kumar, A. & Patra, K. C. 2020 Gene-expression programming for the assessment of surface mean pressure coefficient on building surfaces. *Build. Simul.* **13**, 401–418.
- MATLAB R 2019a [Computer Software]. MathWorks, Natick, MA.
- Mohanta, A. & Patra, K. C. 2021 Gene-expression programming for calculating discharge in meandering compound channels. *Sustainable Water Resour. Manage.* **7**, 33. <https://doi.org/10.1007/s40899-021-00504-0>.
- Mohanta, A., Pradhan, A., Mallick, M. & Patra, K. C. 2021 Assessment of shear stress distribution in meandering compound channels with differential roughness through various artificial intelligence approach. *Water Resour. Manage.* **35** (13), 4535–4559. <https://doi.org/10.1007/s11269-021-02966-5>.
- Mohseni, M. & Naseri, A. 2022 Water surface profile prediction in compound channels with vegetated floodplains. *Proc. Inst. Civ. Eng. Water Manage.*, 1–12.
- Myers, W. R. C. & Elsayy, E. M. 1975 Boundary shears in channel with flood plain. *J. Hydraul. Div. ASCE* **101** (7), 933–946.
- Naik, B. & Khatua, K. K. 2016 Water surface profile computation for compound channels with narrow flood plains. *Arabian J. Sci. Eng.* **42** (3), 941–955. doi:10.1007/s13369-016-2236-x.
- Naik, B., Kaushik, V. & Kumar, M. 2022 Water surface profile in converging compound channel using gene expression programming. *Water Supply* **22** (5), 5221–5236. <https://doi.org/10.2166/ws.2022.172>.
- Najafzadeh, M. & Zahiri, A. 2015 Neuro-fuzzy GMDH-based evolutionary algorithms to predict flow discharge in straight compound channels. *J. Hydraul. Eng.* **20** (12), 04015035.
- Parsaie, A., Yonesi, H. A. & Najafian, S. 2015 Predictive modeling of discharge in compound open channel by support vector machine technique. *Model. Earth Syst. Environ.* **1** (1–2). doi:10.1007/s40808-015-0002-9.
- Parsaie, A., Yonesi, H. & Najafian, S. 2017 Prediction of flow discharge in compound open channels using adaptive neuro fuzzy inference system method. *Flow Meas. Instrum.* **54**, 288–297.
- Patel, V. C. 1965 Calibration of the Preston tube and limitations on its use in pressure gradients. *J. Fluid Mech.* **231**, 85–208.
- Pradhan, A. & Khatua, K. K. 2017b Gene expression programming to predict Manning’s n in meandering flows. *Can. J. Civ. Eng.* **45** (4), 304–313.
- Proust, S., Rivière, N., Bousmar, D., Paquier, A. & Zech, Y. 2006 Flow in the compound channel with abrupt floodplain contraction. *J. Hydraul. Eng.* **132** (9), 958–970.
- Rezaei, B. 2006 Overbank Flow in Compound Channels with Prismatic and non-Prismatic Floodplains. Ph.D. Thesis, University of Birmingham, Birmingham, UK.

- Rezaei, B. & Knight, D. W. 2009 [Application of the Shiono and Knight Method in the compound channel with non-prismatic floodplains](#). *J. Hydraul. Res.* **47** (6), 716–726.
- Rezaei, B. & Knight, D. W. 2011 [Overbank flow in compound channels with non-prismatic floodplains](#). *J. Hydraul.* **137**, 815–824.
- Sahu, M., Khatua, K. K. & Mahapatra, S. S. 2011 [A neural network approach for prediction of discharge in straight compound open channel flow](#). *Flow Meas. Instrum.* **22** (5), 438–446.
- Savic, D. A., Walters, G. A. & Davidson, J. W. 1999 [A genetic programming approach to rainfall-runoff modelling](#). *Water Resour. Manage.* **13** (3), 219–231.
- Seckin, G. 2004 [A comparison of one-dimensional methods for estimating discharge capacity of straight compound channels](#). *Can. J. Civ. Eng.* **31** (4), 619–631.
- Sellin, R. H. J. 1964 [A laboratory investigation into the interaction between flow in the channel of a river and that of its flood plain](#). *LaHouille Blanche* **7**, 793–801.
- Unal, B., Mamak, M., Seckin, G. & Cobaner, M. 2010 [Comparison of an ANN approach with 1-D and 2-D methods for estimating discharge capacity of straight compound channels](#). *Adv. Eng. Software* **41** (2), 120–129.
- Whigham, P. A. & Crapper, P. F. 1999 Time series modelling using genetic programming: an application to rainfall-runoff models. *Adv. Genet. Program* **3**, 89–104.
- Whigham, P. A. & Crapper, P. F. 2001 [Modelling rainfall-runoff using genetic programming](#). *Math. Comput. Modell.* **33** (6–7), 707–721.
- Yonesi, H. A., Omid, M. H. & Ayyoubzadeh, S. A. 2013 [The hydraulics of flow in nonprismatic compound channels](#). *J. Civ. Eng. Urbanism* **3** (6), 342–356. [https://doi.org/10.1061/\(ASCE\)0733-9429\(2000\)126:4\(299\)](https://doi.org/10.1061/(ASCE)0733-9429(2000)126:4(299)).
- Zahiri, A. & Azamathulla, H. M. 2014 [Comparison between linear genetic programming and M5 tree models to predict flow discharge in compound channels](#). *Neural Comput. Appl.* **24** (2), 413–420.

First received 13 October 2022; accepted in revised form 10 December 2022. Available online 17 December 2022



Augmented thermoelectric performance of LiCaX (X = As, Sb) Half Heusler compounds via carrier concentration optimization

Sangeeta, Mukhtiyar Singh *

Department of Applied Physics, Delhi Technological University, Delhi, 110042, India

ARTICLE INFO

Keywords:

Half Heusler compounds
DFT
Thermoelectrics
AIMD
Lattice thermal conductivity

ABSTRACT

The present study is focussed on the detailed physical insight into the structural, thermal, and dynamical properties of 8-valence electron Half Heusler (HH) compounds LiCaX (X = As, Sb) using Density functional theory. The thermal and dynamic stabilities of the compounds are assessed via *ab-initio* molecular dynamic simulations and phonon dispersion calculations, respectively. The Tran-Blaha modified Becke Johnson potential is used to accurately predict the band gap of investigated compounds. It is found that they are indirect band gap semiconductors with band gaps of 2.52 eV (LiCaAs) and 2.09 eV (LiCaSb). The transport parameters are obtained for *p*-type and *n*-type doping at temperatures ranging from 300 K to 800 K by solving the Boltzmann Transport equation. The deformation potential theory is employed to calculate the temperature dependent relaxation time for both compounds. The results of the various thermoelectric parameters obtained using actual values of time-dependent relaxation time are compared with that calculated under constant relaxation time approximation. The maximum power factor is 10.95×10^{11} (4.99×10^{11}) $\text{Wm}^{-1}\text{K}^{-2}\text{s}^{-1}$ for *p*-type (*n*-type) LiCaAs and 12.53×10^{11} (5.30×10^{11}) $\text{Wm}^{-1}\text{K}^{-2}\text{s}^{-1}$ for *p*-type (*n*-type) LiCaSb at optimized carrier concentration. The obtained low lattice thermal conductivities for LiCaSb ($0.66 \text{ Wm}^{-1}\text{K}^{-1}$) and LiCaAs ($0.88 \text{ Wm}^{-1}\text{K}^{-1}$) are explicated in terms of different phonon modes. The Figure of Merit at 800 K for *p*-type (*n*-type) LiCaAs is as high as 0.90 (0.73) and 0.93 (0.84) for LiCaSb at optimum carrier concentration $\sim 10^{20} \text{ cm}^{-3}$ ($\sim 10^{19} \text{ cm}^{-3}$), which has been experimentally realized in other 8-valence electron Li-based HH compounds. The good thermoelectric performance of *p*-type LiCaX in comparison to *n*-type suggests that *p*-type LiCaX alloys are viable candidates for high temperature energy harvesting applications.

1. Introduction

Thermoelectric (TE) materials based solid-state electronic devices convert waste heat into useable electrical energy. These materials have been shown to hold great potential for clean and green energy harvesting [1–3] over other renewable energy resources such as wind, geothermal, and solar which have their own constraints pertaining to location and weather [4]. The efficiency of a TE material is given by a dimensionless quantity known as the Figure of Merit (FOM), which is defined by $ZT = S^2\sigma T / (\kappa_e + \kappa_{\text{lattice}})$, where σ , S , κ_{lattice} , and κ_e , respectively, represents the electrical conductivity, Seebeck coefficient, lattice thermal conductivity, and electronic thermal conductivity. An efficient TE material should possess a high ZT value [5,6]. The tuning of σ , S , and κ_e to maximize the ZT is a tedious task as these parameters are mutually dependent [7]. It is established that enhancement in ZT can be achieved by tuning these parameters independently in two ways: one is to

increase the power factor (PF i.e., $S^2\sigma$) through carrier concentration optimization [8], and another is to suppress the κ_{lattice} by increasing phonon scattering [9].

Most of the traditional TE materials are based upon lead chalcogenides and bismuth tellurides. Although these materials possess high ZT [10] but they are not eco-friendly. The plausible alternatives are Zintl materials [11], skutterudites [12], and Half Heusler (HH) compounds [13–15]. The HH compounds have attracted major scientific attention in recent years owing to their excellent band structure tunability, as well as high mechanical and thermal stability [16–18]. These compounds exist in chemical composition XYZ (or, 1:1:1 ratio), where X and Y are transition metal elements, and Z is a p-block element. Numerous studies are existing in the literature that shows enhancement in TE performance of HH alloys realized through optimization of the carrier concentration via doping [19], heavy element substitution [20,21], and strain [22]. Rausch et al. [20] have synthesized *p*-type HH alloy

* Corresponding author.

E-mail addresses: msphysik09@gmail.com, mukhtiyarsingh@dtu.ac.in (M. Singh).

$\text{Ti}_{0.3}\text{Zr}_{0.35}\text{Hf}_{0.35}\text{CoSb}_{1-x}\text{Sn}_x$ and attained a maximum ZT value of 0.8 (for $x = 0.15$) corresponding to an optimal carrier concentration of $1.4 \times 10^{21} \text{ cm}^{-3}$ thereby proving that the carrier concentration can be an efficient way for improving TE properties. El-Khouly et al. [19] have reported that Hf–Ti co-doping in FeVSb increased the ZT value by 20% as compared to pristine compound at 873 K. Recently, Shen et al. [23] and Serrano-Sánchez et al. [24] have demonstrated improved TE performance of NbFeSb and NbCoSn, respectively, by doping with heavy elements as it optimizes the PF and reduces the κ_{lattice} by causing high scattering of phonons.

Among the anticipated HH alloys, the Li-based 8-VEC HH compounds exhibit good mechanical, dynamical stability, and transport properties [15,25–28]. Few studies have been carried out in recent times focussing on their TE properties [29–31]. Very recently, Xiong et al. have achieved a significant enhancement in the figure of merit using carrier concentration optimization in Li-based 8-VEC HH compounds at 673 K. They have concluded that this compound may be very promising for TE energy harvesting if the thermal conductivity can be reduced significantly [32]. This motivated us to look for a unique combination of Li-based 8-VEC HH compounds with low thermal conductivity. This can be achieved by using heavy p-block elements such as Antimony as it may increase the phonon–phonon scattering rates and reduces the phonon group velocities. We observed that all the reported studies on Li-based 8-VEC HH compounds lack a crucial thorough investigation of lattice thermal conductivity. Moreover, as these Li-based compounds are not experimentally synthesized yet, a clear guidance for the experimentalist is missing in reported works, i.e., the thermal stability at higher temperatures and phonon dynamics.

We used the density functional theory (DFT), and density functional perturbation theory (DFPT) based methods to probe the electronic structure and phonon dispersions of LiCaX ($X = \text{As}, \text{Sb}$) compounds. We assessed the thermal stability at different temperatures using *ab-initio* molecular dynamics simulations (AIMD), which suggests that these compounds can be used for high temperature applications. We then employed the semiclassical Boltzmann transport theory and Phono3py code to calculate TE parameters at different temperatures for a given carrier concentration. In order to predict the accurate ZT, a true value of κ_{lattice} and temperature dependent relaxation time are obtained. We understood the role of acoustic and optical phonon modes in determining κ_{lattice} . This work shows that the ZT values of the LiCaAs, and LiCaSb could be significantly enhanced with optimized carrier concentration, which results in making these Li-based 8-VEC HH compounds very promising candidates for high temperature TE applications.

2. Computational details

The structural properties of LiCaX were obtained using Density functional theory (DFT) as implemented in the Vienna *Ab-initio* Simulation Package (VASP) [33]. The Perdew-Burke-Ernzerhof generalized gradient approximation (GGA) was used for the exchange-correlation potential [34]. The projector augmented plane wave approach was employed for electron-ion interactions [35]. The Brillouin zone (BZ) was sampled using a Gamma-centred $11 \times 11 \times 11$ k-point mesh, and a plane-wave kinetic energy cut-off of 650 eV was used. Geometric structure was fully optimized with a convergence threshold of 10^{-8} eV for energy and -0.001 eV/Å for the force. The *ab-initio* molecular dynamics (AIMD) simulation was employed, over a $2 \times 2 \times 2$ supercell, to check the high temperature stability of the proposed materials. These calculations were carried out in a canonical ensemble (NVT) at temperatures 400 and 800 K using a time step of 1 fs, which lasted for 5000 fs.

The electronic structure and transport properties were calculated using the Full potential linearized augmented plane wave method implemented in WIEN2K code [36]. The cut-off energy was set to -6 Ry for the core and valence states, and the plane wave cut-off was decided by $R_{\text{MT}} \times K_{\text{max}} = 7$. The Kohn-Sham equation was solved self-consistently

with an energy convergence of 0.0001 Ry/cell. The Tran Blaha modified Becke Johnson (TB-mBJ) potential was used for accurate prediction of band gap and transport properties [37]. The TE parameters were estimated via solving the Boltzmann Transport equation (BTE) under the constant relaxation time approximation, which is implemented in the BoltzTraP code [38]. In order to ensure precise TE properties, a dense k-mesh of $34 \times 34 \times 34$ was used for the BZ. The temperature dependent relaxation time was calculated using Deformation Potential theory to obtain true ZT.

The Phonopy code [39] was used to obtain the second order harmonic interatomic force constants (IFCs) and phonon structure via the DFPT method [40] using a $2 \times 2 \times 2$ supercell on an $8 \times 8 \times 8$ k-mesh. With the same supercell, the Phono3py code [41] was used to calculate anharmonic third-order IFCs. Interactions up to the fifth-closest neighbour were taken into consideration for the third-order IFCs. The Boltzmann transport equation was solved self consistently to obtain the lattice thermal conductivity using a dense $20 \times 20 \times 20$ q-mesh.

3. Results and discussion

3.1. Structural properties and thermal stability

The HH LiCaX crystallizes in the cubic structure having a space group of $F\bar{4}3m$ (no. 216), where Li atoms occupy positions 4b ($1/2, 1/2, 1/2$), Ca atoms occupy positions 4a (0, 0, 0), and X (As and Sb) occupies 4c ($1/4, 1/4, 1/4$). Thus, the crystal structure can be viewed as zinc blende type structure (CaX), partially filled by Li^+ ions [42]. Thomas et al. have conducted extensive *ab-initio* investigations for the crystal structures of a variety of HHs and have observed the above-mentioned structure to be the most favourable for I–II–V HHs [43]. To further verify this, we have carried out the total energy calculations of different possible configurations (Type I, Type II, and Type III) whose Wyckoff positions are listed in Table 1. The total energies obtained from the self-consistent calculations for these configurations for both the compounds are plotted as a function of cell volume (Fig. 1) and curves are fitted in accordance with the Birch-Murnaghan equation of states [44]. The results show that the Type I configuration corresponds to the minimum energy and thus the most stable for both HH compounds. The calculated lattice constants of LiCaAs (6.67 Å) and LiCaSb (6.96 Å) are found to be in good agreement with previous studies [31,45]. The computed volume, bulk modulus, its pressure derivative, and ground state energy for both compounds are presented in Table 2.

The thermal stability is a crucial parameter for any material to be used for TE applications. We have probed the thermal stability of LiCaX by AIMD simulations using canonical ensemble at 400 K and 800 K, as shown in Fig. 2. It is visible that the temperature fluctuations are very trivial throughout the 5000 fs AIMD simulations at 400 K and 800 K. These results indicate that the structures are thermally stable at ambient as well as elevated temperatures.

We have also obtained the melting temperature (T_m) of LiCaX using an empirical relation [46]: $T_m = 553 \text{ K} + (5.91/\text{GPa}) C_{11} \pm 300 \text{ K}$. This relation has also been used for other similar Li-based 8 VEC HH compound [47]. We have used the elastic package, implemented in WIEN2K software, to calculate the elastic constants for cubic systems in order to obtain the value of C_{11} . The obtained results are presented in Table 3. The value of the bulk modulus (B) has obtained by fitting the Birch-Murnaghan equation of states [44].

Table 1

Possible Wyckoff positions of atoms for the LiCaX ($X = \text{As}, \text{Sb}$) HH compounds.

	Type I	Type II	Type III
Li	4b	4c	4a
Ca	4a	4a	4b
X	4c	4b	4c

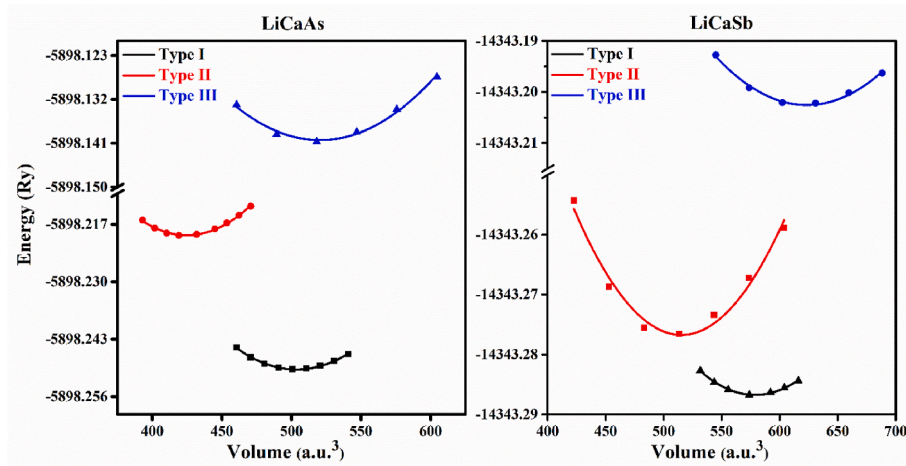


Fig. 1. Calculated total energy (in Ry) versus volume (in a.u.³) for Type I, Type II, and Type III configurations of LiCaAs and LiCaSb.

Table 2

Calculated optimized volume (V_0), bulk modulus (B), pressure derivative (B_p), and total energy (E_0) of both LiCaAs and LiCaSb compounds.

	V_0 (a.u. ³)	B (GPa)	B_p	E_0 (Ry)
LiCaAs	501.94	37.5093	3.5402	-5898.249798
LiCaSb	573.23	30.6543	3.5741	-14343.286715

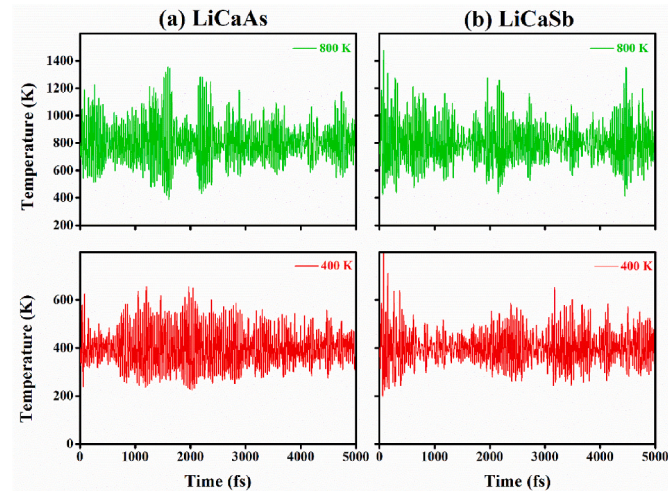


Fig. 2. The AIMD simulations of temperature fluctuations versus time for (a) LiCaAs and (b) LiCaSb at 400 K and 800 K.

Table 3

Calculated elastic constants C_{ij} (GPa), Bulk modulus B (GPa) and melting temperature T_m (K).

	C_{11}	C_{12}	C_{44}	B	T_m
LiCaAs	62.305	25.252	28.220	37.026	921.22
LiCaSb	53.730	18.346	24.996	30.137	870.54

3.2. Dynamical stability

To confirm the dynamical stability, we have calculated frequency-dependent phonon band structure using DFPT as presented in Fig. 3. The presence of all positive frequency phonon modes shows that the structures under investigation are stable. In their primitive cell, there are three atoms that give rise to nine different phonon modes, three of which

are acoustic (low frequency phonon modes) and six of which are optical (high frequency phonon modes). The highest frequencies for LiCaAs and LiCaSb are 10 THz and 8.5 THz, respectively. There is a band gap from 6.5 THz to 7.5 THz in LiCaAs, between 5.6 THz and 6.8 THz and slightly from 3.3 THz to 4.0 THz in LiCaSb. It can be seen that in the entire frequency range the dispersion of phonon modes shrinks with an increase in atomic weight of the X atom. This suggests a reduction in corresponding group velocity, which further leads to low lattice thermal conductivity [48].

It can be seen from Fig. 4 that the optical phonon modes above the band gap are predominantly contributed from the vibrations of Li atoms. This characteristic is the same for both structures LiCaAs and LiCaSb. The primary factor separating the phonon DOS of these two structures is the vibrations of As and Sb atoms. Owing to the large atomic mass of Sb, its vibrational frequencies mainly dominate up to 3 THz in LiCaSb. On the other hand, in the case of LiCaAs, the vibrational frequencies dominate up to 5 THz. Also, the minor contribution of Ca ions can be seen in the low frequency modes; however, the mid frequency phonon modes are largely contributed by the vibrations of Ca ions.

The Grüneisen parameter (γ) as a function of frequency is shown in Fig. 5, its value is positive for all modes. It is a measure of anharmonicity, therefore large γ reflects the strong anharmonicity, thus the low lattice thermal conductivity. The average value of γ for LiCaAs and LiCaSb is 1.28 and 1.20, respectively, smaller than that of other HH alloys [48,49] but comparable to the promising TE material PbTe [50].

3.3. Electronic structure properties

Fig. 6 shows the energy band structure of LiCaAs and LiCaSb, which is obtained using TB-mBJ potential. The conduction band (CB) minima and the valence band (VB) maxima of both alloys are found at the X-point and the Γ -point in the BZ, respectively. Therefore, both are found to be indirect band gap semiconductors. The band gaps of LiCaAs and LiCaSb are calculated to be 1.82 (2.52) eV and 1.52 (2.09) eV using GGA (TB-mBJ), respectively. The calculated value of the band gap for LiCaAs is comparable to previous DFT calculations [31,45]. The band gap of LiCaX decreases from As to Sb. Thus, a higher value of the Seebeck coefficient can be expected for LiCaAs. Also, the VB edges are flat indicating strongly localized holes, on the contrary, the CB edges are dispersed showing free electrons. Therefore, we foresee superior TE performance for *p*-type compounds in comparison to the *n*-type ones.

The total and partial density of states (DOS) of LiCaX HH alloys are depicted in Fig. 7. It is noticeable that the VB is primarily dominated by X (i.e., As and Sb) atoms and the CB is majorly composed of Ca atoms for LiCaX HH alloy. Li atoms give negligible contribution both in CB and VB. It is also found that the DOS in the VB is higher than that in the CB. As a

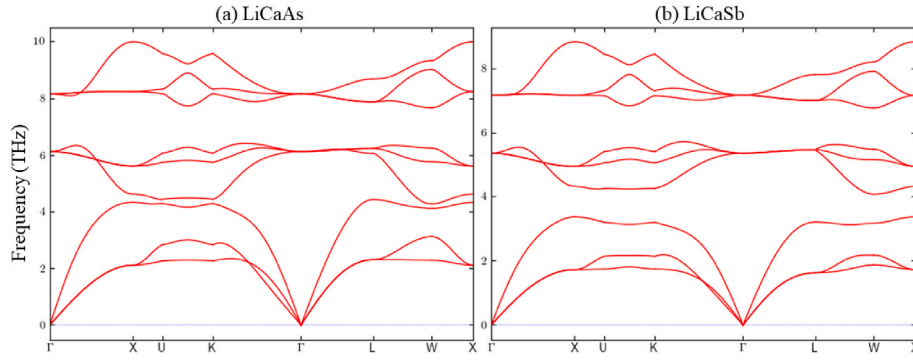


Fig. 3. The obtained phonon dispersion for (a) LiCaAs and (b) LiCaSb.

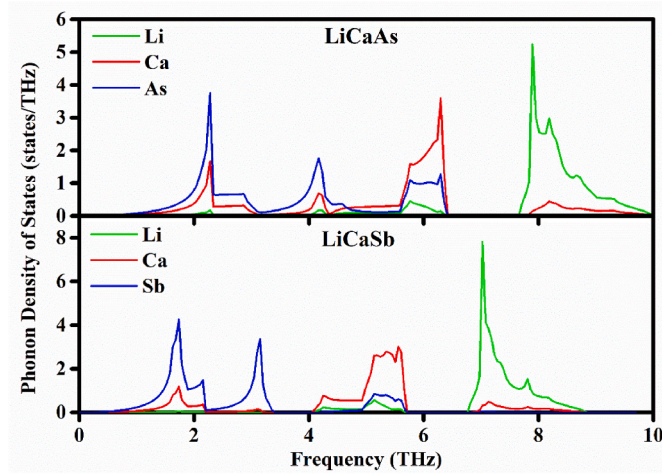


Fig. 4. Calculated phonon density of states of LiCaAs and LiCaSb.

result, we anticipate superior TE performance in the *p*-type doping than in the *n*-type counterpart.

4. Thermoelectric properties

The TE parameters S , σ , and κ_e are calculated by the following relations [38]:

$$S_{\alpha\beta}(\mu, T) = \frac{1}{eT\sigma_{\alpha\beta}(\mu, T)} \int \sigma_{\alpha\beta}(\epsilon)(\epsilon - \mu) \left(-\frac{\partial f(T, \epsilon, \mu)}{\partial \epsilon} \right) d\epsilon \quad (1)$$

$$\kappa_{\alpha\beta}^e(\mu, T) = \frac{1}{e^2 T \Omega} \int \sigma_{\alpha\beta}(\epsilon)(\epsilon - \mu)^2 \left(-\frac{\partial f(T, \epsilon, \mu)}{\partial \epsilon} \right) d\epsilon \quad (2)$$

$$\sigma_{\alpha\beta}(\mu, T) = \frac{1}{\Omega} \int \sigma_{\alpha\beta}(\epsilon) \left(-\frac{\partial f(T, \epsilon, \mu)}{\partial \epsilon} \right) d\epsilon \quad (3)$$

where Ω is the reciprocal space volume, f is the Fermi distribution function, e is the electron charge, ϵ is the carrier energy. $\sigma_{\alpha\beta}(\epsilon)$ is conductivity tensor can be expressed as:

$$\sigma_{\alpha\beta}(\epsilon) = \frac{1}{N} \sum_{i,k} \sigma_{\alpha\beta}(i, k) \frac{\delta(\epsilon - \epsilon_{i,k})}{d\epsilon} \quad (4)$$

The transport properties are calculated as a function of temperature and with varying carrier concentration (n) ranging from $1 \times 10^{19} \text{ cm}^{-3}$ to $1 \times 10^{22} \text{ cm}^{-3}$, which is an optimum range for good TE performance [6], with constant relaxation time for the HH alloys LiCaX ($X = \text{As}$ and Sb). Considering that these compounds are stable at higher temperatures, thereby we calculated the transport properties up to 800 K.

4.1. Seebeck coefficient

Fig. 8 depicts the S as a function of carrier concentration ($1 \times 10^{19} \text{ cm}^{-3}$ to $1 \times 10^{22} \text{ cm}^{-3}$) at temperatures ranging from 300 K to 800 K. For degenerate semiconductors, it can be expressed as [51]:

$$S = \frac{8\pi^2 k_B^2}{3eh^2} m^* T \left(\frac{\pi}{3n} \right)^{\frac{2}{3}} \quad (5)$$

The absolute value i.e., $|S|$, decreases with the increase in carrier concentration, which is consistent with the fact that S is inversely proportional to n . The S of *p*-type LiCaAs and LiCaSb is significantly greater than that for the *n*-type because of flat bands in the VB and dispersed bands in CB (that leads to a higher effective mass of holes than that of electrons). This indicates that the TE performance of the *p*-type LiCaX is better than that of the *n*-type. We calculated the effective mass (m^*) of hole and electron carriers using the relation $m^* = \hbar^2 / [\partial^2 E / \partial k^2]$. The calculated value of m^* of holes is $11.86 m_e$, and $5.55 m_e$, respectively, for

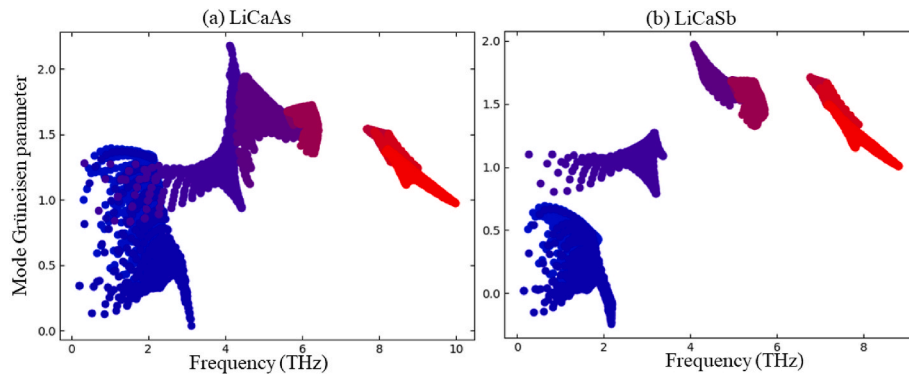


Fig. 5. Calculated Grüneisen parameter for (a) LiCaAs and (b) LiCaSb.

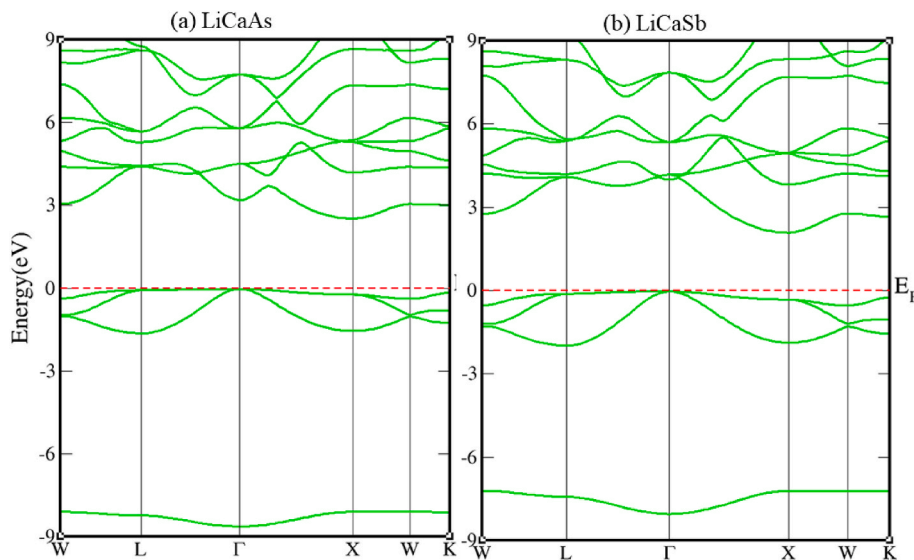


Fig. 6. Calculated electronic band structures of (a) LiCaAs and (b) LiCaSb along the high symmetry directions of the BZ as determined using the Tran-Blaha modified Becke-Johnson potential. The Fermi energy is set to 0 eV.

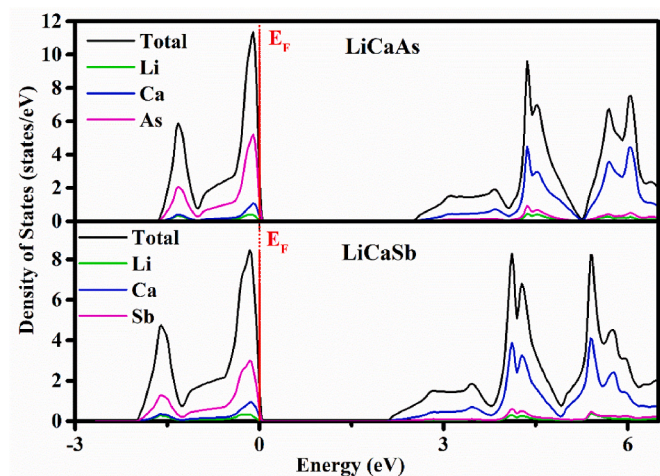


Fig. 7. Calculated total and partial density of states of LiCaAs and LiCaSb. The Fermi energy is set to 0 eV.

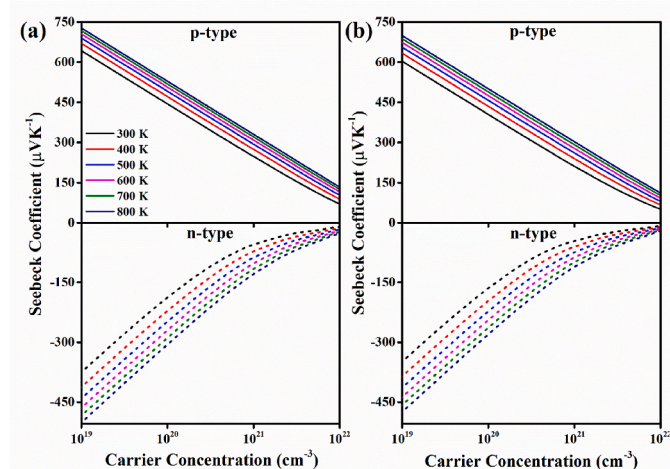


Fig. 8. Calculated Seebeck coefficient of (a) LiCaAs and (b) LiCaSb as a function of carrier concentration and for temperatures ranging from 300 K to 800 K.

LiCaAs, and LiCaSb, whereas the corresponding effective mass of electron is $0.63 m_e$ and $0.54 m_e$, here m_e is the mass of free electron. The maximum obtained value of S for p -type LiCaAs, and LiCaSb, respectively, are $727.31 \mu\text{VK}^{-1}$ and $704.08 \mu\text{VK}^{-1}$ at 800 K, whereas the corresponding S values for n -type are $-501.54 \mu\text{VK}^{-1}$ and $-475.14 \mu\text{VK}^{-1}$, respectively. The larger S is obtained for p -type variants, throughout the studied carrier concentration range and temperature, of both compounds, which might be because of the higher effective mass of holes than electrons. The optimum carrier concentration for which maximum $|S|$ obtained is $1 \times 10^{19} \text{ cm}^{-3}$.

4.2. Electrical and electronic thermal conductivity

The magnitude of σ/τ increases with the carrier concentration for both n -type and p -type nature of investigated compounds. Fig. 9 shows that σ/τ increases sharply beyond carrier concentration $\sim 1 \times 10^{21} \text{ cm}^{-3}$. Also, it can be noted that electrical conductivity shows weak dependence on temperature. The value of σ/τ is higher in n -type than in p -type due to parabolically dispersed CB edges.

The thermal conductivity includes both lattice and electronic contribution; its electronic component is described by Wiedemann Franz law [52] ($\kappa_e = L\sigma T$ where L is Lorentz number) that states that κ_e varies linearly with the temperature and electrical conductivity. The obtained electronic contribution to thermal conductivity, with respect to relaxation time i.e., κ_e/τ presented in Fig. 10, reveals that for the same carrier concentration, the values for LiCaSb are greater than those for LiCaAs, and the change in κ_e/τ with n is similar to that of σ/τ .

4.3. Power factor

The optimization of the carrier concentration substantially improves the PF of LiCaX HH compounds as given in Fig. 11. With the increase in carrier concentration the value of PF first increases, reaches a maximum value and then starts decreasing. We estimated that the optimum carrier concentration for n -type compounds is relatively lower than p -type. For LiCaAs, the optimal p -type and n -type PF/ τ values are around $10.95 \times 10^{11} \text{ Wm}^{-1}\text{K}^{-2}\text{s}^{-1}$ at carrier concentration of $5 \times 10^{21} \text{ cm}^{-3}$ and $4.99 \times 10^{11} \text{ Wm}^{-1}\text{K}^{-2}\text{s}^{-1}$ at $9 \times 10^{20} \text{ cm}^{-3}$, respectively, whereas corresponding values for LiCaSb are $12.53 \times 10^{11} \text{ Wm}^{-1}\text{K}^{-2}\text{s}^{-1}$ at $4 \times 10^{21} \text{ cm}^{-3}$ and $5.30 \times 10^{11} \text{ Wm}^{-1}\text{K}^{-2}\text{s}^{-1}$ at $8 \times 10^{20} \text{ cm}^{-3}$. The comparison of PF/ τ shows that p -type LiCaSb has a relatively higher value. Therefore, one can expect better TE performance in case of p -type LiCaSb

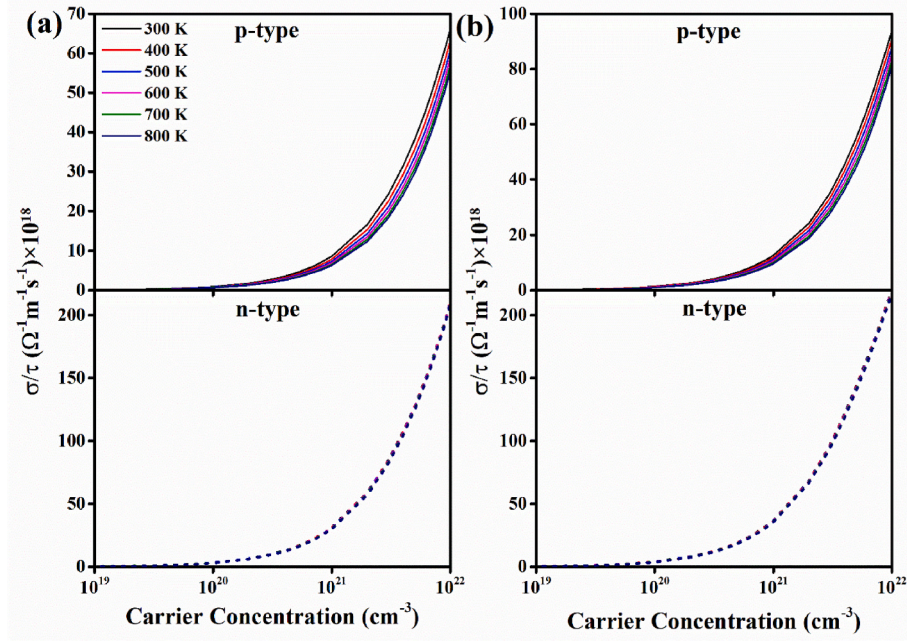


Fig. 9. Calculated electrical conductivity of (a) LiCaAs and (b) LiCaSb as a function of carrier concentration and for temperatures ranging from 300 K to 800 K.

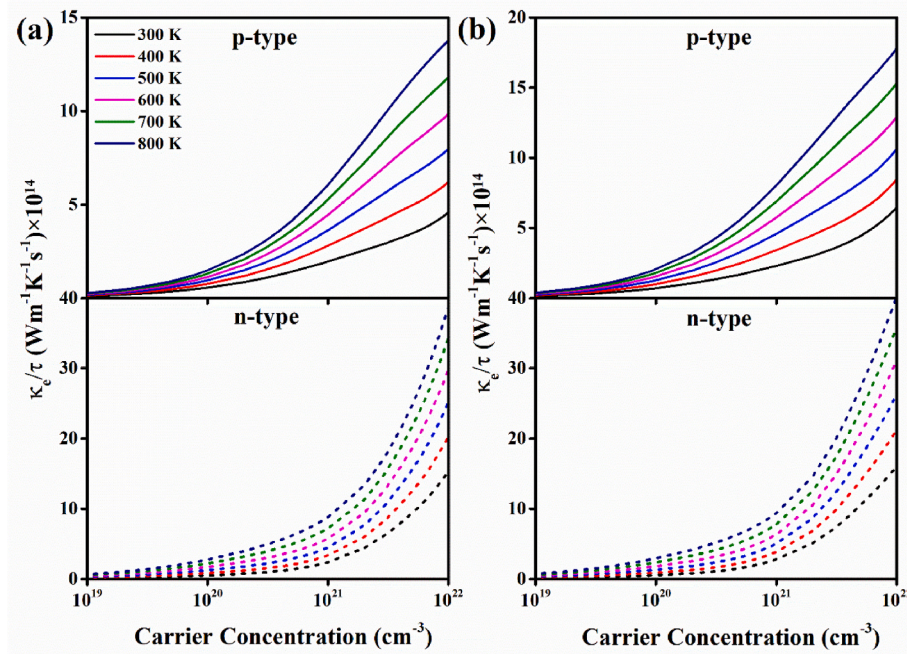


Fig. 10. Calculated electronic thermal conductivity of (a) LiCaAs and (b) LiCaSb as a function of carrier concentration and for temperatures ranging from 300 K to 800 K.

compound.

4.4. Lattice thermal conductivity

To understand the behaviour of κ_{lattice} in LiCaAs and LiCaSb, it is crucial to understand the phonon dynamics of both HH alloys. From Fig. 3, it can be noticed that the LiCaSb has overall lower frequencies than that of LiCaAs. Therefore, low κ_{lattice} is expected for LiCaSb. This also reflects in its low value of lattice thermal conductivity than that of LiCaAs. The calculated temperature dependent lattice contribution to the thermal conductivity of these HH alloys is shown in Fig. 12 (a). The

room temperature κ_{lattice} values for LiCaSb, and LiCaAs are $1.42 \text{ W m}^{-1} \text{ K}^{-1}$ and $1.90 \text{ W m}^{-1} \text{ K}^{-1}$, respectively, but these values decrease $0.66 \text{ W m}^{-1} \text{ K}^{-1}$ and $0.88 \text{ W m}^{-1} \text{ K}^{-1}$ at 800 K. These compounds have an intrinsically low value of κ_{lattice} than that of TaFeSb based ($\kappa_{\text{lattice}} = 3.1 \text{ W m}^{-1} \text{ K}^{-1}$ for $\text{Ta}_{0.84}\text{Ti}_{0.16}\text{FeSb}$ and $2.3 \text{ W m}^{-1} \text{ K}^{-1}$ for $\text{Ta}_{0.74}\text{V}_{0.1}\text{Ti}_{0.16}\text{FeSb}$), which is amongst the high-performance TE material [53], and $\text{Ti}_{0.5}\text{Hf}_{0.5}\text{NiSn}$ with $3.2 \text{ W m}^{-1} \text{ K}^{-1}$ [50] HH alloys. These obtained values of κ_{lattice} at room temperature are comparable to conventional promising TE candidates like PbTe ($2.2 \text{ W m}^{-1} \text{ K}^{-1}$) [54] and Bi_2Te_3 ($1.6 \text{ W m}^{-1} \text{ K}^{-1}$) [55]. The value of lattice thermal conductivity of LiCaSb is lower than that of LiCaAs.

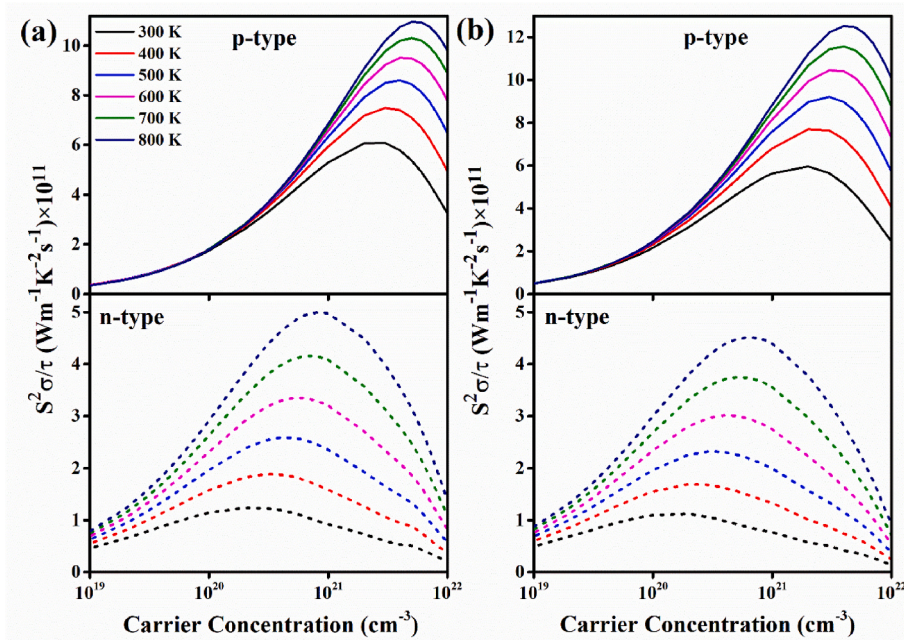


Fig. 11. Calculated power factor of (a) LiCaAs and (b) LiCaSb as a function of carrier concentration and for temperatures ranging from 300 K to 800 K.

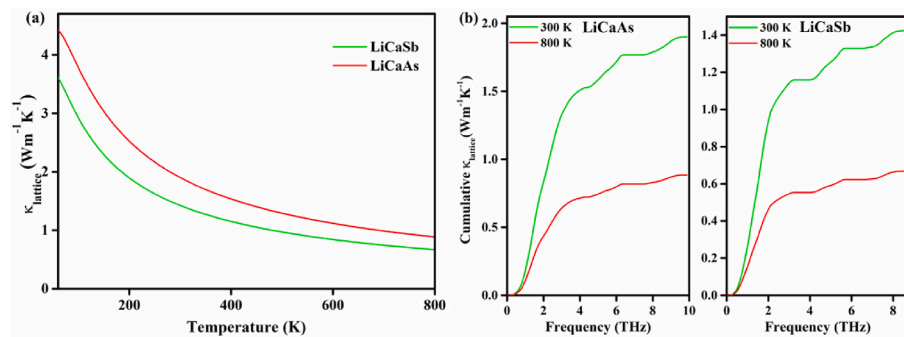


Fig. 12. (a) The lattice thermal conductivity of LiCaAs and LiCaSb as a function of temperature (b) The cumulative lattice thermal conductivity as a function of frequency in LiCaAs and LiCaSb at different temperatures 300 K and 800 K.

Further, we have calculated the cumulative thermal conductivity of LiCaAs and LiCaSb at 300 K and 800 K as a function of frequency (Fig. 12 (b)). It increases rapidly in the low-frequency acoustic phonon modes (1–4 THz), and this part of the phonons contributes more than 80% of the lattice thermal conductivity. Therefore, the high frequency optical phonon modes have a minor contribution (less than 20%) to the thermal conductivity. The relative contribution of these phonon modes of both the compounds to κ_{lattice} is presented in Table 4.

Table 4

Relative contributions of acoustic and optical phonon modes to the lattice thermal conductivity (κ_{lattice}) at room temperature and 800 K for both compounds LiCaAs and LiCaSb.

	Temperature (K)	κ_{lattice} (Wm ⁻¹ K ⁻¹)	Contribution of acoustic modes (%)	Contribution of optical modes (%)
LiCaAs	300	1.90	80.26	19.73
	800	0.88	81.03	18.96
LiCaSb	300	1.42	81.49	18.50
	800	0.66	83.17	17.18

4.5. Figure of merit (ZT)

The dependence of transport parameters on the carrier concentration and temperature suggests that a high Figure of Merit can be obtained by tuning the carrier concentration and increasing the temperature. As depicted in Fig. 13, the ZT is calculated as a function of carrier concentration and for temperature range 300 K–800 K in HH alloys LiCaAs and LiCaSb by incorporating the PF and thermal conductivity. It is observed that the room temperature ZT values are high for *p*-type LiCaAs and LiCaSb than that of *n*-type. The calculated values of ZT for both *p*-type and *n*-type compounds at 300 K and 800 K are listed in Table 5.

It can be seen that the ZT improved at higher temperatures at optimized carrier concentrations ($\sim 10^{21}$ cm⁻³ for *p*-type carriers and $\sim 10^{20}$ cm⁻³ for *n*-type carriers for both the investigated compounds). The similar values of carrier concentrations have been experimentally observed in other Li-based 8 VEC HH compounds [32]. It is obvious that as we move towards higher atomic weight element, the κ_{lattice} suppressed by 24% that results in augmentation of TE performance of LiCaSb. However, the final ZT value of the two compounds is nearly identical because of the large *S* value of LiCaAs compared to LiCaSb. The low κ_{lattice} and high PF of LiCaX compounds result in high ZT value at higher temperatures. This shows that these HH alloys are worthy candidates for further experimental investigations. The TE properties of the *p*-type

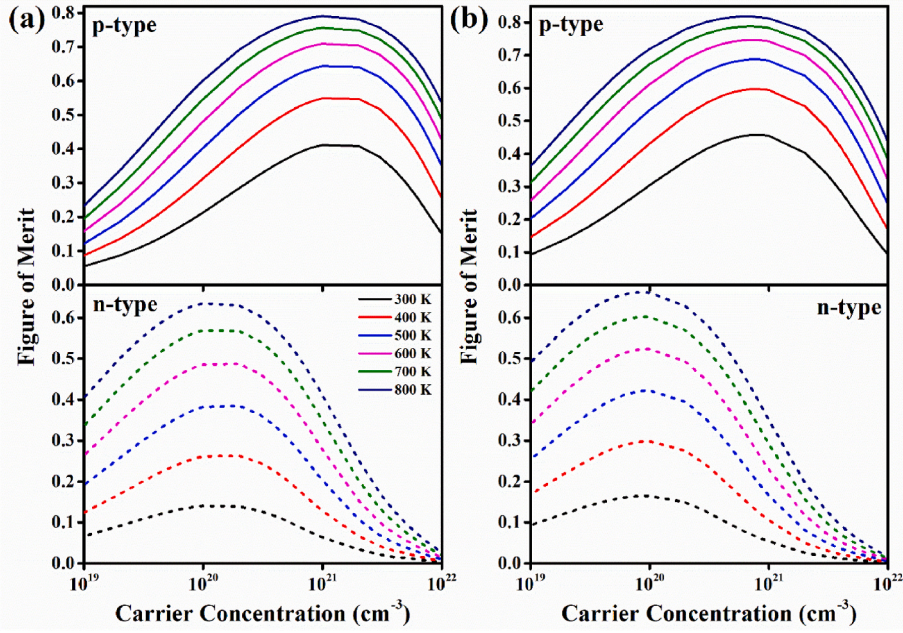


Fig. 13. Calculated ZT values of (a) LiCaAs and (b) LiCaSb as a function of carrier concentration within constant relaxation time approximation (i.e., $\tau = 1 \times 10^{-14}$ s) at different temperatures ranging from 300 K to 800 K.

Table 5

Figure of merit of LiCaAs and LiCaSb for both type of carriers at optimized carrier concentration.

		Figure of merit at optimized carrier concentration			
		LiCaAs		LiCaSb	
		300 K	800 K	300 K	800 K
With constant τ	p-type	0.41	0.79	0.46	0.82
		1×10^{21} cm ⁻³	1×10^{21} cm ⁻³	8×10^{20} cm ⁻³	7×10^{20} cm ⁻³
	n-type	0.14	0.64	0.16	0.66
		1×10^{20} cm ⁻³	1×10^{20} cm ⁻³	1×10^{20} cm ⁻³	9×10^{19} cm ⁻³
With calculated temperature dependent τ	p-type	0.84	0.90	0.89	0.93
		4×10^{20} cm ⁻³	4×10^{20} cm ⁻³	1×10^{20} cm ⁻³	1×10^{20} cm ⁻³
	n-type	0.58	0.73	0.62	0.84
		3×10^{19} cm ⁻³	4×10^{19} cm ⁻³	2×10^{19} cm ⁻³	2×10^{19} cm ⁻³

compounds are superior to those of *n*-type; therefore, hole doping seems to be good for improving the TE performance of these compounds.

Further to determine the figure of merit accurately, the temperature dependent τ of the charge carriers, i.e., electrons and holes has evaluated using the deformation potential theory developed by Bardeen and Shockley [56], according to which τ is given by the following:

$$\tau = \frac{2\sqrt{2}\pi C_{\beta} \hbar^4}{3(k_B T m^*)^{3/2} E_{\beta}^2} \quad (6)$$

where, C_{β} is the elastic modulus along β axis, m^* is the effective mass of charge carriers, k_B is the Boltzmann's constant and E_{β} is the deformation potential constant along β axis. The m^* values are written in text (section 4.1) and other parameters- C_{β} , E_{β} , and τ of charge carriers i.e., electrons and holes for both compounds at different temperatures are listed in Table 6.

When the temperature dependent τ is used the increase in FOM is observed as shown in Fig. 14 in comparison to constant $\tau = 1 \times 10^{-14}$ s. The calculated FOM values at optimized carrier concentration for both *p*-type and *n*-type compounds at 300 K and 800 K are listed in Table 5.

Table 6

The elastic modulus (C_{β}), deformation potential constant (E_{β}), and relaxation time (τ) at 300 K, 600 K and 800 K of electrons and holes for LiCaAs and LiCaSb.

Carriers type	LiCaAs		LiCaSb	
	electron	hole	electron	hole
C_{β} (eV/Å ³)	0.070146	0.070146	0.063437	0.063437
E_{β} (eV)	1.73353	0.20233	1.58507	0.20083
τ (s)	300 K	2.56×10^{-13}	2.30×10^{-13}	3.49×10^{-13}
	600 K	9.06×10^{-14}	8.14×10^{-14}	1.23×10^{-13}
	800 K	5.88×10^{-14}	5.29×10^{-14}	8.02×10^{-14}

The variation of ZT with the temperature at the optimum carrier concentration using constant τ approximation and temperature dependent τ of charge carriers is also shown in Fig. 15. It is found that the optimum carrier concentration for *p*-type LiCaAs and LiCaSb offers better TE performance than the *n*-type counterparts.

5. Conclusion

We have used first-principles calculations in conjunction with semi-classical Boltzmann approach, density functional perturbation technique and *ab-initio* molecular dynamics to explore electronic, transport, phononic, and thermal properties of 8-VEC Li-based HH compounds LiCaX (X = As, Sb). We have also calculated temperature dependent relaxation time using deformation potential theory. We have presented a comparative analysis of the Figure of Merit obtained using actual values of temperature dependent relaxation time and that calculated under constant relaxation time approximation. Our calculations have validated the high temperature stability and dynamical stability of these compounds. We found that both LiCaAs and LiCaSb are indirect band gap semiconductors having band gap, respectively, of 2.52 eV and 2.09 eV estimated using TB-mBJ potential. The value of the Seebeck coefficient is have been found to be higher in *p*-type LiCaX than that of *n*-type due to flat VB edges. TE performance is slightly enhanced with an increase in atomic weight of X atom owing to low $\kappa_{lattice}$ and significant power factor. The room temperature $\kappa_{lattice}$ values for LiCaSb and LiCaAs have found to be $1.42 \text{ Wm}^{-1}\text{K}^{-1}$ and $1.90 \text{ Wm}^{-1}\text{K}^{-1}$, respectively, but these values have decreased to $0.66 \text{ Wm}^{-1}\text{K}^{-1}$ and $0.88 \text{ Wm}^{-1}\text{K}^{-1}$ at

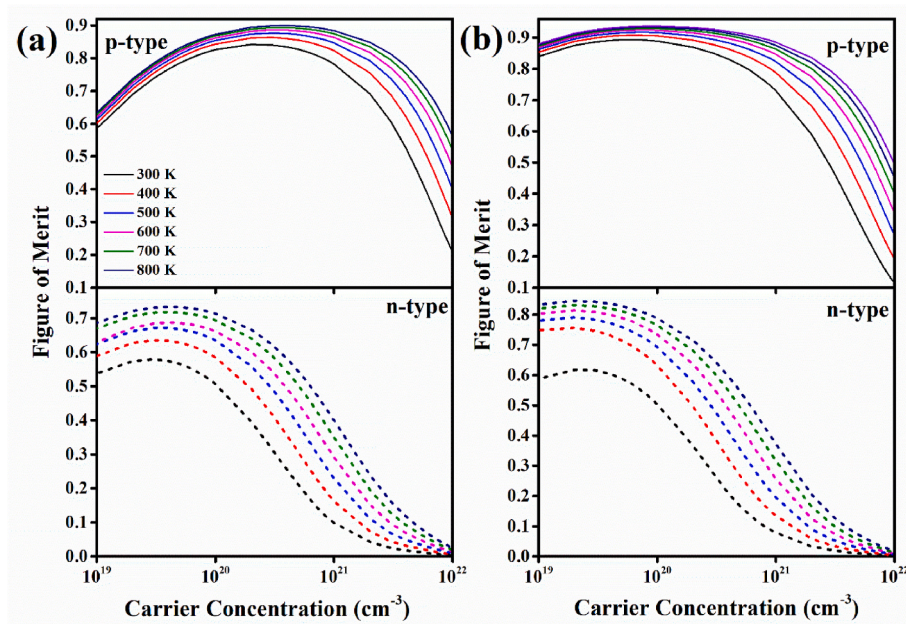


Fig. 14. Calculated ZT values of (a) LiCaAs and (b) LiCaSb as a function of carrier concentration using temperature dependent relaxation time of carriers at different temperatures ranging from 300 K to 800 K.

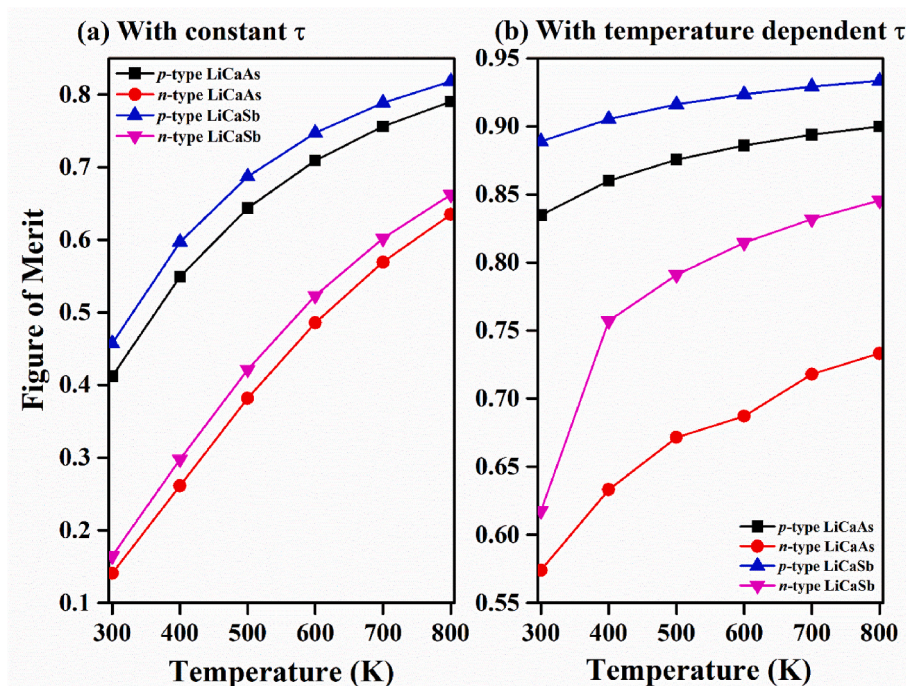


Fig. 15. The variation in Figure of Merit with temperature (a) within constant relaxation time approximation (b) with temperature dependent relaxation time of carriers for p-type and n-type LiCaAs and LiCaSb.

800 K. The remarkably low κ_{lattice} of 8-VEC HH compounds LiCaX (X = As, Sb), which has been understood in terms of different phonon modes, and optimization of carrier concentration resulted into an improved ZT at higher temperature. The optimized carrier concentration ($\sim 10^{20} \text{ cm}^{-3}$ for p-type carriers and $\sim 10^{19} \text{ cm}^{-3}$ for n-type carriers) of the investigated compounds have found to be comparable to experimentally estimated value for other 8-VEC Li-based HH. Our calculations have predicted that the p-type HH alloys LiCaX are promising TE materials. We are optimistic that this work could leads to future experiment to investigate the TE properties of proposed Li-based half-Heusler alloys.

Author statement

Sangeeta: Investigation, Methodology, Software, Data curation, Writing – original draft preparation. **Mukhtiyar Singh:** Supervision, Conceptualization, Visualization Writing- Reviewing and Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

Financial support to Sangeeta from Delhi Technological University, Delhi, is thankfully acknowledged. The authors also acknowledge the National Supercomputing Mission (NSM) for providing computing resources of 'PARAM SEVA' at IIT, Hyderabad, which is implemented by C-DAC and supported by the Ministry of Electronics and Information Technology (MeitY) and Department of Science and Technology (DST), Government of India.

References

- [1] M. Massetti, F. Jiao, A.J. Ferguson, D. Zhao, K. Wijeratne, A. Würger, J. L. Blackburn, X. Crispin, S. Fabiano, Unconventional thermoelectric materials for energy harvesting and sensing applications, *Chem. Rev.* 121 (2021) 12465–12547, <https://doi.org/10.1021/acs.chemrev.1c00218>.
- [2] X. Zhu, Y. Yu, F. Li, A review on thermoelectric energy harvesting from asphalt pavement: configuration, performance and future, *Construct. Build. Mater.* 228 (2019), 116818, <https://doi.org/10.1016/j.conbuildmat.2019.116818>.
- [3] G. Schierning, Bring on the heat, *Nat. Energy* 3 (2018) 92–93, <https://doi.org/10.1038/s41560-018-0093-4>.
- [4] P.A. Østergaard, N. Duic, Y. Noorollahi, H. Mikulic, S. Kalogirou, Sustainable development using renewable energy technology, *Renew. Energy* 146 (2020) 2430–2437, <https://doi.org/10.1016/j.renene.2019.08.09>.
- [5] G. Tan, M. Ohta, M.G. Kanatzidis, Thermoelectric power generation: from new materials to devices, *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* 377 (2019), 20180450, <https://doi.org/10.1098/rsta.2018.0450>.
- [6] G.J. Snyder, E.S. Toberer, Complex thermoelectric materials, *Nat. Mater.* 7 (2008) 105–114, <https://doi.org/10.1038/nmat2090>.
- [7] T. Zhu, Y. Liu, C. Fu, J.P. Heremans, G.J. Snyder, X. Zhao, Compromise and synergy in high-efficiency thermoelectric materials, *Adv. Mater.* 29 (2017), 1605884, <https://doi.org/10.1002/adma.201605884>.
- [8] Y. Pei, A.D. Lalonde, N.A. Heinz, X. Shi, S. Iwanaga, H. Wang, L. Chen, G.J. Snyder, Stabilizing the optimal carrier concentration for high thermoelectric efficiency, *Adv. Mater.* 23 (2011) 5674–5678, <https://doi.org/10.1002/adma.201103153>.
- [9] H. Kim, G. Park, S. Park, W. Kim, Strategies for manipulating phonon transport in solids, *ACS Nano* 15 (2021) 2182–2196, <https://doi.org/10.1021/acsnano.0c10411>.
- [10] K. Kaur, Enamullah, S.A. Khandy, J. Singh, S. Dhiman, Traditional thermoelectric materials and challenges, in: *Woodhead Publishing Series in Electronic and Optical Materials, Thermoelectricity and Advanced Thermoelectric Materials*, Woodhead Publishing, 2021, pp. 139–161, <https://doi.org/10.1016/B978-0-12-819984-8.00009-6>.
- [11] S.M. Kazuilarich, S.R. Brown, J.G. Snyder, Zintl phases for thermoelectric devices, *J. Chem. Soc., Dalton Trans.* 21 (2007) 2099–2107, <https://doi.org/10.1039/b702266b>.
- [12] G. Rogl, P. Rogl, Skutterudites, a most promising group of thermoelectric materials, *Curr. Opin. Green Sustain. Chem.* 4 (2017) 50–57, <https://doi.org/10.1016/j.cogsc.2017.02.006>.
- [13] C. Fu, S. Bai, Y. Liu, et al., Realizing high figure of merit in heavy-band p-type half-Heusler thermoelectric materials, *Nat. Commun.* 6 (2015) 8144, <https://doi.org/10.1038/ncomms9144>.
- [14] S.S. Essaoud, A.S. Jbara, First-principles calculation of magnetic, structural, dynamic, electronic, elastic, thermodynamic and thermoelectric properties of Co₂ZrZ (Z = Al, Si) Heusler alloys, *J. Magn. Magn. Mater.* 531 (2021), 167984, <https://doi.org/10.1016/j.jmmm.2021.167984>.
- [15] Vikram, B. Sahni, C.K. Barman, A. Alam, Accelerated discovery of new 8-electron half-Heusler compounds as promising energy and topological quantum materials, *J. Phys. Chem. C* 123 (2019) 7074–7080, <https://doi.org/10.1021/acs.jpcc.9b01737>.
- [16] J.-W.G. Bos, Recent Developments in Half-Heusler Thermoelectric Materials, *Thermoelectric Energy Conversion*, Woodhead Publishing, 2021, pp. 125–142, <https://doi.org/10.1016/B978-0-12-818535-3.00014-1>.
- [17] L. Huang, Q. Zhang, B. Yuan, X. Lai, X. Yan, Z. Ren, Recent progress in half-Heusler thermoelectric materials, *Mater. Res. Bull.* 76 (2016) 107–112, <https://doi.org/10.1016/j.materresbull.2015.11.032>.
- [18] T. Fang, X. Zhao, T. Zhu, Band structures and transport properties of high-performance half-Heusler thermoelectric materials by first principles, *Mater* 11 (2018) 847, <https://doi.org/10.3390/ma11050847>.
- [19] A. El-Khouly, A. Novitskii, I. Serhienko, A. Kalugina, A. Sedegov, D. Karpenkov, A. Voronin, V. Khovaylo, A.M. Adam, Optimizing the thermoelectric performance of FeVbSb half-Heusler compound via Hf–Ti double doping, *J. Power Sources* 477 (2020), 228768, <https://doi.org/10.1016/j.jpowsour.2020.228768>.
- [20] E. Rausch, B. Balke, T. Deschauer, S. Ouardi, C. Felser, Charge carrier concentration optimization of thermoelectric p-type half-Heusler compounds, *Apl. Mater.* 3 (2015), 041516, <https://doi.org/10.1063/1.4916526>.
- [21] S.A. Khandy, J.-D. Chai, Strain engineering of electronic structure, phonon, and thermoelectric properties of p-type half-Heusler semiconductor, *J. Alloys Compd.* 850 (2021), 156615, <https://doi.org/10.1016/j.jallcom.2020.156615>.
- [22] S.A. Khandy, K. Kaur, S. Dhiman, J. Singh, V. Kumar, Exploring thermoelectric properties and stability of half-Heusler PtXSn (X = Zr, Hf) semiconductors: a first principle investigation, *Comput. Mater. Sci.* 188 (2021), 110232, <https://doi.org/10.1016/j.commatsci.2020.110232>.
- [23] J. Shen, L. Fan, C. Hu, T. Zhu, J. Xin, T. Fu, D. Zhao, X. Zhao, Enhanced thermoelectric performance in the n-type NbFeSb half-Heusler compound with heavy element Ir doping, *Mater. Today Phys* 8 (2019) 62–70, <https://doi.org/10.1016/j.mtphys.2019.01.004>.
- [24] F. Serrano-Sánchez, T. Luo, J. Yu, W. Xie, C. Le, G. Auffermann, A. Weidenkaff, T. Zhu, X. Zhao, J.A. Alonso, B. Gault, C. Felser, C. Fu, Thermoelectric properties of n-type half-Heusler NbCoSn with heavy-element Pt substitution, *J. Mater. Chem.* 8 (2020) 14822–14828, <https://doi.org/10.1039/D0TA04644B>.
- [25] S.H. Shah, S.H. Khan, A. Laref, G. Murtaza, Optoelectronic and transport properties of LiBZ (B = Al, In, Ga and Z = Si, Ge, Sn) semiconductors, *J. Solid State Chem.* 258 (2018) 800–808, <https://doi.org/10.1016/j.jssc.2017.12.014>.
- [26] J. Barth, G.H. Fecher, M. Schwind, A. Beleanu, C. Felser, A. Shkablo, A. Weidenkaff, J. Hanss, A. Reller, M. Köhne, Investigation of the thermoelectric properties of LiAlSi and LiAlGe, *J. Electron. Mater.* 39 (2010) 1856–1860, <https://doi.org/10.1007/s11664-010-1076-9>.
- [27] Y. Dhakshayani, G. Suganya, G. Kalpana, DFT studies on electronic, magnetic and thermoelectric properties of half Heusler alloys XCaB (X = Li, Na, K and Rb), *J. Cryst. Growth* 583 (2022), 126550, <https://doi.org/10.1016/j.jcrysgro.2022.126550>.
- [28] U. Chopra, M. Zeeshan, S. Pandey, R. Dhawan, H.K. Singh, J.V. D Brink, C. Kandal, First-principles study of thermoelectric properties of Li-based Nowotony–Juza phases, *J. Phys. Condens. Matter* 31 (2019), 505504, <https://doi.org/10.1088/1361-648X/ab4015>.
- [29] M.K. Yadav, B. Sanyal, First principles study of thermoelectric properties of Li-based half-Heusler alloys, *J. Alloys Compd.* 622 (2015) 388–393, <https://doi.org/10.1016/j.jallcom.2014.10.025>.
- [30] Anuradha, K. Kaur, R. Singh, R. Kumar, Search for thermoelectricity in Li-based half-Heusler alloys: a DFT study, *Mater. Res. Express* 5 (2018), 014009, <https://doi.org/10.1088/2053-1591/aaa507>.
- [31] A. Azouaoui, A. Hourmatallah, N. Benzakour, K. Bouslykhane, First-principles study of optoelectronic and thermoelectric properties of LiCaX (X = N, P and As) half-Heusler semiconductors, *J. Solid State Chem.* 310 (2022), 123020, <https://doi.org/10.1016/j.jssc.2022.123020>.
- [32] J.-L. Xiong, F. Yu, J. Liu, X.-C. Liu, Q. Liu, K. Liu, S.-Q. Xia, LiAlTt (Tt = Si, Ge): experimental and theoretical reinvestigation on the thermoelectric properties of 8-valence-electron half-Heusler compounds, *ACS Appl. Energy Mater.* 5 (2022) 3793–3799, <https://doi.org/10.1021/acsaem.2c00279>.
- [33] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, *Comput. Mater. Sci.* 6 (1996) 15–50, [https://doi.org/10.1016/0927-0256\(96\)00008-0](https://doi.org/10.1016/0927-0256(96)00008-0).
- [34] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. B* 77 (1996) 18, <https://doi.org/10.1103/PhysRevLett.77.3865>.
- [35] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Phys. Rev. B* 54 (1996) 16, <https://doi.org/10.1103/PhysRevB.54.11169>.
- [36] K. Schwarz, P. Blaha, G.K.H. Madsen, Electronic structure calculations of solids using the WIEN2k package for material sciences, *Comput. Phys. Commun.* 147 (2002) 71–76, [https://doi.org/10.1016/S0010-4655\(02\)00206-0](https://doi.org/10.1016/S0010-4655(02)00206-0).
- [37] F. Tran, P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, *Phys. Rev. Lett.* 102 (2009) 5–8, <https://doi.org/10.1103/PhysRevLett.102.226401>.
- [38] G.K. Madsen, D.J. Singh, BoltzTraP. A code for calculating band-structure dependent quantities, *Comput. Phys. Commun.* 175 (2006) 67–71, <https://doi.org/10.1016/j.cpc.2006.03.007>.
- [39] S. Baroni, S. de Gironcoli, A. Dal Corso, P. Giannozzi, Phonons and related crystal properties from density-functional perturbation theory, *Rev. Mod. Phys.* 73 (2001) 515–562, <https://doi.org/10.1103/RevModPhys.73.515>.
- [40] A. Togo, I. Tanaka, First principles phonon calculations in materials science, *Scripta Mater.* 108 (2015) 1–5, <https://doi.org/10.1016/j.scriptamat.2015.07.021>.
- [41] A. Togo, L. Chaput, I. Tanaka, Distributions of phonon lifetimes in Brillouin zones, *Phys. Rev. B* 91 (2015), 094306, <https://doi.org/10.1103/PhysRevB.91.094306>.
- [42] K. Kuriyama, K. Nagasawa, K. Kushida, Growth and band gap of the filled tetrahedral semiconductor LiMgN, *J. Cryst. Growth* 237–239 (2002) 2019–2022, [https://doi.org/10.1016/S0022-0248\(01\)02249-7](https://doi.org/10.1016/S0022-0248(01)02249-7).
- [43] G. Thomas, Comparative ab initio study of half-Heusler compounds for optoelectronic applications, *Phys. Rev. B* 82 (2010), 125210, <https://link.aps.org/doi/10.1103/PhysRevB.82.125210>.
- [44] F. Birch, Finite elastic strain of cubic crystals, *Phys. Rev.* 71 (1947) 809–824, <https://link.aps.org/doi/10.1103/PhysRev.71.809>.
- [45] H. Mehnane, B. Bekkouch, S. Kacimi, A. Hallouch, M. Djermouni, A. Zaoui, First-principles study of new half Heusler for optoelectronic applications, *Superlattice. Microst.* 51 (2012) 772–784, <https://doi.org/10.1016/j.spmi.2012.03.020>.
- [46] M.E. Fine, L.D. Brown, H.L. Marcus, Elastic constants versus melting temperature in metals, *Scripta Metall.* 18 (1984) 951–956, [https://doi.org/10.1016/0036-9748\(84\)90267-9](https://doi.org/10.1016/0036-9748(84)90267-9).

- [47] H.Y. Wu, Y.H. Chen, C.R. Deng, X.Y. Han, Z.J. Liu, Electronic, elastic and dynamic properties of the filled tetrahedral semiconductor LiMgN under pressures, *J. Solid State Chem.* 231 (2015) 1–6, <https://doi.org/10.1016/j.jssc.2015.07.047>.
- [48] X. Ye, Z. Feng, Y. Zhang, G. Zhao, D.J. Singh, Low thermal conductivity and high thermoelectric performance via Cd underbonding in half-Heusler PCdNa, *Phys. Rev. B* 105 (2022), 104309, <https://doi.org/10.1103/PhysRevB.105.104309>.
- [49] S.N.H. Eliassen, A. Katre, G.K.H. Madsen, C. Persson, O.M. Løvvik, K. Berland, Lattice thermal conductivity of $\text{Ti}_x\text{Zr}_y\text{Hf}_{1-x-y}\text{NiSn}$ half-Heusler alloys calculated from first principles: key role of nature of phonon modes, *Phys. Rev. B* 95 (2017), 045202, <https://doi.org/10.1103/PhysRevB.95.045202>.
- [50] G.A. Slack, *The Thermal Conductivity of Nonmetallic Crystals*, Solid State Physics, vol. 34, Academic Press, 1979, pp. 1–71, [https://doi.org/10.1016/S0081-1947\(08\)60359-8](https://doi.org/10.1016/S0081-1947(08)60359-8).
- [51] M. Dutta, T. Ghosh, K. Biswas, Electronic structure modulation strategies in high-performance thermoelectrics, *Apl. Mater.* 8 (2020), 040910, <https://doi.org/10.1063/5.0002129>.
- [52] M. Jonson, G. Mahan, Mott's formula for the thermopower and the Wiedemann-Franz law, *Phys. Rev. B Condens. Matter* 21 (1980), <https://doi.org/10.1103/PhysRevB.21.4223>, 21 4223.
- [53] H. Zhu, J. Mao, Y. Li, et al., Discovery of TaFeSb-based half-Heuslers with high thermoelectric performance, *Nat. Commun.* 10 (2019) 270, <https://doi.org/10.1038/s41467-018-08223-5>.
- [54] S. Ju, T. Shiga, L. Feng, J. Shiomi, Revisiting PbTe to identify how thermal conductivity is really limited, *Phys. Rev. B* 97 (2018), 184305, <https://doi.org/10.1103/PhysRevB.97.184305>.
- [55] M.-K. Han, Y. Jin, D.-H. Lee, S.-J. Kim, Thermoelectric properties of Bi_2Te_3 : CuI and the effect of its doping with Pb atoms, *Mater* 10 (2017) 1235, 10.3390%2Fma10111235.
- [56] J. Bardeen, W. Shockley, Deformation potentials and mobilities in non-polar crystals, *Phys. Rev.* 80 (1950) 72–80. <https://link.aps.org/doi/10.1103/PhysRev.80.72>.

Capital Structure Study: A Systematic Review and Bibliometric Analysis

Vision

1–17

© 2022 MDI

Reprints and permissions:

in.sagepub.com/journals-permissions-india

DOI: 10.1177/09722629221130453

journals.sagepub.com/home/vis

Anjali Sisodia¹  and G. C. Maheshwari¹

Abstract

The capital structure study is extant and wide. It touches every sector of the economy with its extreme relevance. Its importance cannot be restricted to empirical analysis and financial data study rather on knowing its significance an attempt has been made to highlight its features, the areas covered the countries in which studies are undertaken, its relation with determinants and its evolution from the year 2010 to 2021 in the form of review analysis. In short, this article showcases the review comprising bibliometric analysis followed by selected papers systematic review on capital structure.

Key Words

Bibliometric Analysis, Capital Structure, Financing, Systematic Literature Review

Introduction

‘Capital structure decision is the explicit fusion of debt and equity which an organization uses to back up its operating and investment decisions.’ (Kumar et al., 2017). It is considered to be a significant feature of the firm success. Firms are found to become bankrupt due to their inadequate capital structure. ‘The theory of capital structure has been dominated by the search for optimal capital structure’ (Myers, 1998). The optimum capital is demanded to maintain the borrowing level to a certain extent which is possible through a trade-off between the cost of financial distress and tax on borrowed money (Myers, 1993). The capital structure theories include trade-off theory, pecking order theory and market timing theory. There is no optimum capital structure that exists in reality. Madan (2007) states that capital structure is benefitted from the use of leverage when there is sufficient profit in the company. Capital structure is the result of cumulative outcomes which can be achieved by monitoring the equity market (Baker & Wurglar, 2002). In the absence of tax, the value of the firm will not change with the moderate amount of debt but it will decline when the debt usage is high and there exists an optimum leverage ratio in context to capital structure in the presence of tax (Robichek & Myers, 1966). Myers (1993) says capital structure has the inverse relation between debt usage and profitability. Deb and Banerjee (2018) found

concerning to Indian firm following almost zero leverage policy perform far better than the levered firms. Shivdasani and Zenner (2005) hold an opinion that the choice of debt equity is based on the credit rating of the company. He found that high-rated company focuses on their financial flexibility or the buyback of their shares while the low rated prefers to strengthen their credit rating.

The study on the capital structure has its inception from the year 1958 when the seminal study of Modigliani and Miller on the capital structure was propounded. Later it was observed as a prominent study conducted vividly in developed economies and also catches pace in developing economies. The study on the mentioned topic has various dimensions in terms of theory, determinants impact and its role on the firm value. The earlier researchers’ main focus has been identified in the area of capital structure theories, its determinants and country-specific factors. There has been immense fragmentation identified while summarizing the previous findings. This article aims to consolidate and generate a comprehensive overview of the importance of the topic of capital structure and map the areas already touched. It further attempts to specifically shed light on a systematic literature review in capital structure in various horizons. There has been a surge in demonstrating the determinants of capital structure and its impact on the debt ratio, various capital structure theories are reviewed and

¹DSM, Delhi Technological University (DTU), New Delhi, Delhi, India

Corresponding author:

Anjali Sisodia, DSM, Delhi Technological University (DTU), New Delhi, Delhi 110042, India.

E-mail: anjalis376@gmail.com

related to capital structure financing. The necessity and importance of capital structure have been identified and presented in literatures for several years but despite its coverage to every economic segment and influence on every sector, its adversity has proved inevitable. This article aims to link the existing dots on capital structure and identify the research gap in this direction followed by a suggestion for future research direction.

The study of capital structure is largely influenced by the seminal work of Modigliani and Miller (1958) which further resulted in various theories and discussions in the last few decades. Table 1 here attempts to highlight the progression of such studies on capital structure done so far by eminent authors which have discovered several new concepts and findings on this subject and created the direction for future researchers to discover the gap still not identified.

Previous Review of Literature on Capital Structure

The capital structure is found to be influenced by different phases of the business cycle due to upsurges and slumps in the economy. Thus decisions related to financial decision-making are always considered very crucial. Despite having such importance, the review of literature on this topic is scant. A few eminent work has been highlighted and named hereunder which has fostered the capital structure study globally.

- (1) Harris and Raviv (1991);
- (2) Lugi and Sorin (2009);
- (3) Miglo (2010); and
- (4) Iqbal et al. (2012).

The inference derived from the above studies shows a major focus being levied on theories and determinants of capital structure but the idea to cover the study on capital structure, its findings in terms of reputed journals, the authors and organisations involved and to be precise conducting a bibliometric study on this subject followed with the systematic review of the literature with cluster and citation analysis is the first of its kind conducted till date to the best of authors knowledge. Thus it proclaims the immense need for such study to be conducted on this well accepted important subject which determines the gap in this direction not by just citing theoretically but by understanding its coverage through bibliometric study and recognising the dearth in various areas which are needed to be overcome and analysing the huge existing data on this subject through cluster analysis which enables to identify the future gap to be covered by the researchers.

Aim of the Study

The aim of the study is to highlight the below mentioned questions which are derived from the review of past research on capital structure.

Table 1. Eminent Studies on Capital Structure.

Sr.no	References	Findings
1	Modigliani and Miller (1958)	It has a significant contribution in the area of capital structure with the origin of 'Irrelevance theory' which states that capital structure has no impact on firm value.
2	Modigliani and Miller (1963)	It analysed the impact of tax shield on interest expense
3	Kraus and Litzenberger (1973)	The study introduced classical 'Trade-off theory'. It covers the concept of tradeoff between cost of financial distress and benefits derived from debt tax shield.
4	Stiglitz (1973)	It developed the concept of pecking order. This study states that leverage ratio is the unexpected resultant of profits and investments made by firm.
5	Jensen and Meckling (1976)	Introduced 'Agency cost theory' and analysed the impact of debtholder-shareholder and manager shareholder conflict on capital structure financing.
6	Miller (1977)	Propounds the significance of personal and corporate tax in the financial decision making.
7	Ross (1977)	It developed the 'Signaling theory' of capital structure and promoted the debt issue as positive indicator in the performance in capital structure financing
8	Bradley et al. (1984)	Introduced the well-known 'Static trade-off theory'
9	Kane et al. (1984)	It introduced the Dynamic trade off theory which includes trade off theory along with the impact of uncertainty, cost, taxes and tax benefits.
10	Myers and Majluf (1984)	It propounded the 'Pecking order theory' and the major role of information asymmetry towards choice between internal fund, debt and equity for capital structure financing.
11	Fischer et al. (1989)	It initiated the transaction cost concept and shown its impact on leverage in the capital structure of the firm.
12	Harris and Raviv (1991)	It initiated the concept of 'Control driven theory'
13	Baker and Wurgler (2002)	It predicts the long run impact of market value fluctuations on the capital structure. It states that firm issue equity when market is overvalued and issue debt when undervalued.

Source: The authors.

- RQ1: What are the ongoing publications and top studies undertaken in the field of capital structure in terms of time, journals, disciplines, authors, affiliated countries and institutions, and method of study adopted in the mentioned area?
- RQ2: What is the influential anatomy of capital structure research? How has capital structure research evolved over the years?
- RQ3: What are the prominent keynotes and which are the popular studies conducted in the reputed journals?
- RQ4: What are the break and avenues for future research in the capital structure?

Research Design

The study conducted in this research is clearly described and classified with the help of tables and figures with the appropriate headings. It intends to cover the complete analysis framework of the research undertaken vividly shown in Figure 2, along with a brief mention of the techniques adopted for data search incorporating certain inclusion criteria. It further elaborates capital structure study in 594 research papers in the form of progression of publication, avenues of publication and recognition in terms of prolific authors along with their institutions and countries, top disciplines and sample statistics. The keyword analysis is followed by the content study conducted on the 58 research papers under Table 7, then discussion and future research avenues under Table 8 following conclusion of the study.

Methodologies for Research Analysis

Knowledgebase Portal, Keywords and Inclusion Criterion

The knowledgebase portal used for the review of literature was from the reputed Web of Science (WOS) core collection by clarivate analytics. Although both Scopus and WOS are the reputed databases used by the researchers in their respective disciplines but the author in this study has preferred the WOS database as Scopus carries a vast database which creates the difficulty in evaluation of quality journals and identification of eminent literature on the topic while WOS covers research in top tier journals of high recognition in the publication field with an impact factor associated with it which limits the vast literature on the subject to qualify in this database and it is considered appropriate for bibliometric analysis (Korom, 2019). The search was conducted in October 2021 to extract the relevant papers related to the study undertaken covering the period from 1 January 2000 to 1 January 2021. The search conducted included Social Sciences Citation Index (SSCI) publications from the WOS database for the period

2000–2021. The search criterion used ‘capital structure’ as the keyword under the title search bar. The capital structure search has been done on the title tab to retrieve data covering all the eminent papers which has word capital structure included in its title. The study on the mentioned subject is vast, after applying the filters in terms of WOS categories considering SSCI publication and using the word capital structure in double quotes to find the records that contain the exact phrase and avoid records that contain all the words randomly which may or may not be close together. It provided 643 documents from the mentioned search criterion. The records were further refined based on document type which considered articles only and in the English language. The refined search gave 594 results. The final database used for the review after the removal of duplicates was 594 articles. Further 58 articles were closely observed to understand the pattern of capital structure studies. The articles for close analysis were selected by reading their abstracts and in case required the full paper was studied. Figure 1 represents the flow of the study undertaken.

Analysis Method

The method employed in the present study is a bibliometric technique followed by a structured review explaining the pattern of methods, theories and constructs in the form of tables and figures which intends to give researchers insightful information from the content furnished on capital structure (Paul & Criado, 2020). Thus to depict the concrete findings of the research undertaken, a bibliometric study is combined with further investigation of the selected research papers content analysis. The VOS viewer software is being adopted for the analysis as it maps the connectedness of items and is a simple, well-explanatory software tool for the construction and visualisation of bibliometric networks. The VOS viewer is well-explanatory software in the analysis and visualization of bibliometric networks developed by Nees Jan Van Eck and Ludo Waltman of Leiden University.

Findings

The progression of publications on capital structure is shown in Figure 3. The graph shows there has been a maximum in the year 2019 counted 75 followed by 67 records in 2020.

Avenues of Publication

The capital structure study is conducted in many other affiliated institutions and other databases but when restricted to the WOS core collection and in that further attempt to identify the ABDC ranking and CAB rating journals the list is reduced to very few. Table 2 lists the data of the top 10 journals with A* and A ranking along with

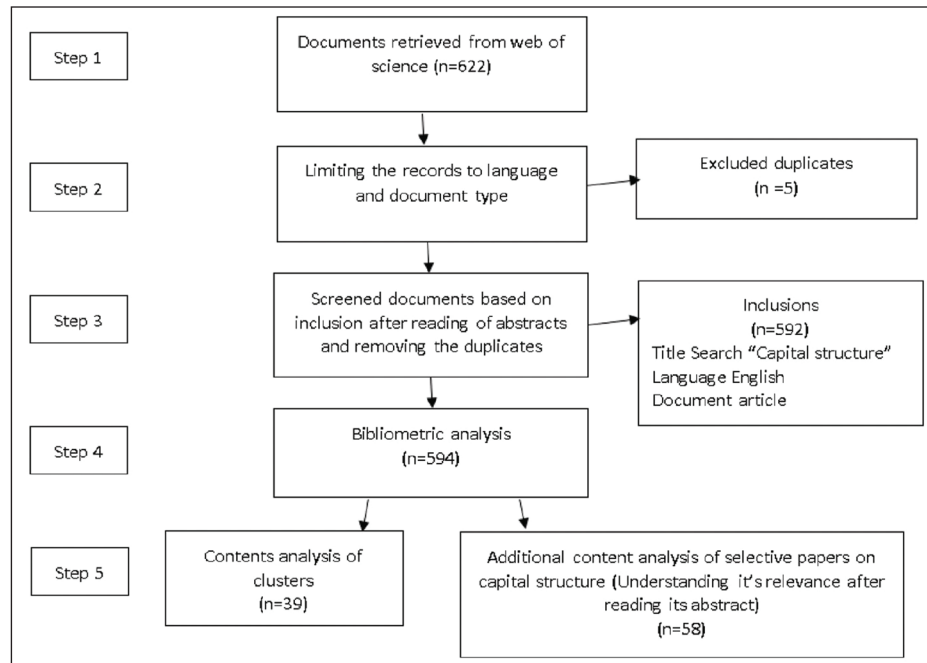


Figure 1. Prisma Diagram Showing Data Retrieval Process.

Source: The authors.

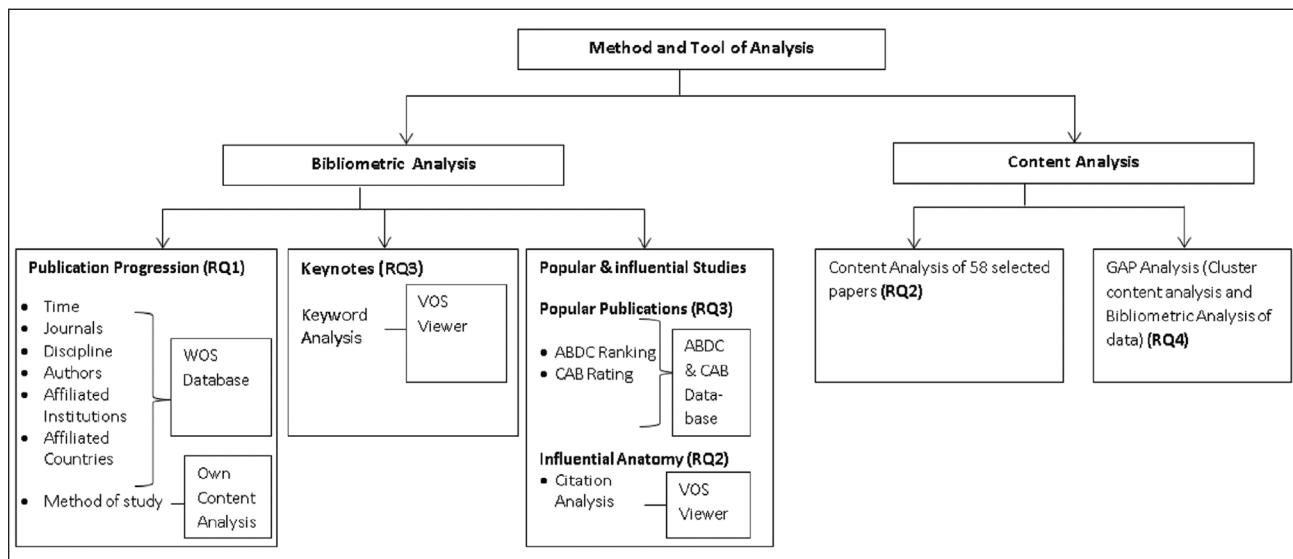


Figure 2. Analysis Framework.

Source: The authors.

CAB rating which reveals such information which is about capital structure study. The 594 publications are dispersed across 202 journals. Table 2 acknowledges the most prominent journals publishing in the area of capital structure. The top 10 journals have published 149 of the total articles studied representing 73.76% of the total. The Journal of financial economics includes the maximum

number of studies in the concerned area which totals 26 followed by the Journal of Banking and finance. The capital structure study has a considerable space in every sector of the economy and so its presence spread across prominent journals. The journals rankings as per well-appreciated body in the academic fraternity with the name Australian Business Dean Council ranking A* and

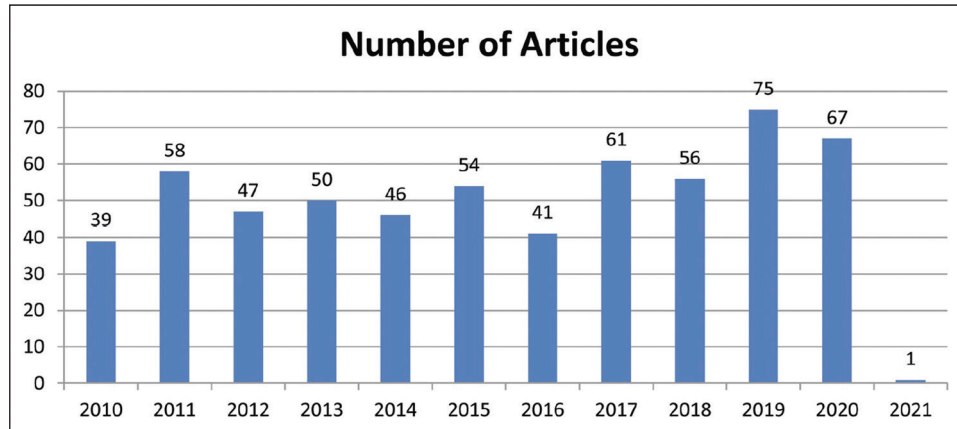


Figure 3. The Publication Progression of 594 Papers during the Period 2010–2021.

Source: Retrieved from WOS.

Table 2. Publication Avenues in Foremost Journals.

Journal	ABDC Ranking	ABS Rating	Publishing House	TP
Journal of Corporate Finance	A*	3	Elsevier	35
Journal of Financial Economics	A*	4	Elsevier	26
Journal of Banking Finance	A*	3	Elsevier	19
Journal of Financial and Quantitative Analysis	A*	4	Cambridge University Press	14
Small Business Economics	A	3	Springer Nature	13
Review of Financial Studies	A*	4	Oxford University Press	11
Review of Finance	A*	3	Oxford University Press	9
Financial Management	A	3	Wiley Online Library	7
Journal of Finance	A*	4	Wiley Online Library	7
Applied Economics Letters	B	1	Taylor and Francis	8

Source: The authors.

A ranking and further another reputed body showcasing the rating of journals, named Chartered Association of Business School (CABS) rating is represented in Table 2.

Profuse Authors with their Affiliated Institutions and Countries

The data set analysed shows the spread of 1,221 authors affiliated with 48 organisations across 75 countries. Table 3 showcases the contributions made in terms of a number of publications. Serrasqueiro, records six publication followed by Yang and Chipeta. The highest citations scored which counts 136 by Dang, Viet following Di Pietro, Fillipo with 88 citations and Serrasqueiro 85. The table also accommodates the statistics related to top institutions to authors on the mentioned subject. The National Bureau of Economic Research is contributing the highest number of publications which are 1,190 along with the highest citations, that is, 14. The State University System of Florida with 11 publications and 555 citations stands second in the list and third is the University of Texas System with 10 publications and 549

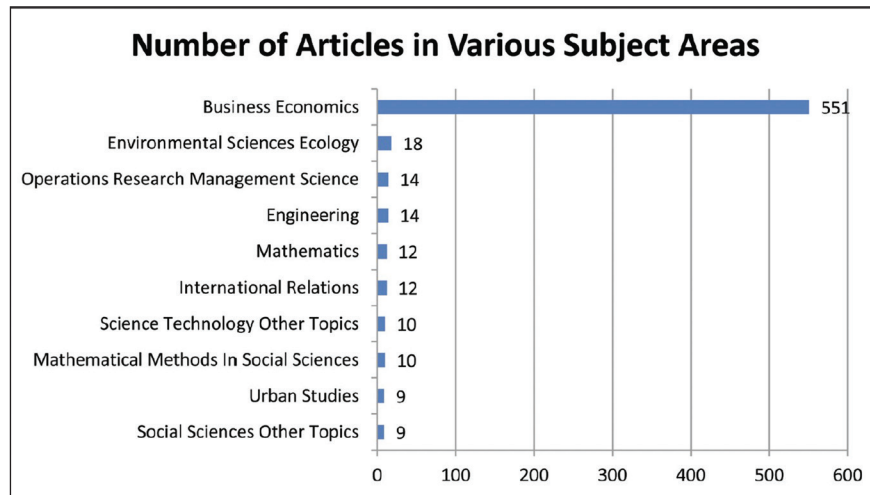
citations. Thus, reflecting the lack of studies in developing economies. Table 3 also covers the contributions made by the countries affiliated with the authors of capital structure. The USA has the highest number of publications which record 185 and also reports the highest number of citations 5,920 then succeeding with People's Republic of China with 105 publications and 1,537 citations and the third is England with 60 publications and 907 citations. Thus the major reason of the persistence problem of improper financing of firms and inadequate capital structure design in the developing countries is the lack of attention in this regard. The countries like India do not chart in the list of under 10 thus demanding a lot of attention and quality study on the concerned subject.

The capital structure study though conducted majorly in the field of business economics is not just confined to it but is fragmented in other fields like environmental sciences, engineering, operations research, international relations, mathematics and many more. Figure 4 represents the top four research areas in which the subject is majorly investigated.

Table 3. Profuse Authors with Affiliated Institutions and Countries Publishing on Capital Structure.

Top Authors			Top Institutions			Top Countries		
Author	TP	TC	Institution	TP	TC	Country	TP	TC
Serrasqueiro, Zelia	6	85	National Bureau of Economic Research	14	1190	USA	185	5920
Yang, Jinqiang	5	17	State University System of Florida	11	555	Peoples R China	105	1537
Chipeta, Chimwemwe	4	22	University of Texas System	10	549	England	60	907
Dang, Viet	4	136	Shanghai University of Finance Economics	9	80	Australia	33	727
Di Pietro, Filippo	4	88	Cornell University	8	219	Taiwan	32	344
Nunes, Paulo Macas	4	35	Tsinghua University	8	39	Italy	28	435
Wang, Wei	4	39	University of California System	8	445	Spain	28	254
Zhang, Hong	4	29	University of Manchester	8	166	Germany	26	695
Acedo-ramirez, Miguel A.	3	29	Hunan University	7	65	Canada	24	647
A. N. Bany-ariffin	3	29	Pennsylvania Commonwealth System of Higher Education Pcshe	7	139	France	18	168

Source: The authors.

**Figure 4.** Capital Structure Research in 594 Papers Distributed in Major Disciplines.

Source: The authors.

Sample Statistics

The study on capital structure is being carried on in several domains. Based on the data studied the capital structure study can be classified under three heads which are majorly empirical, conceptual and theoretical. The empirical study on capital structure is undertaken through observation or experimentation. It is majorly substantiated numerically in terms of numbers. It can be quantitative or qualitative. The theoretical study on the subject is based on certain predefined theories rather than experience or practice like

predicting or determining the behaviour of capital structure financing theories while conceptual studies are found to be dealing with original thoughts. In this research analysis available information on a topic associating the financing pattern with the theories or (theory) of capital structure is represented under theoretical study. Figure 5 shows the study percentages covered in the area of capital structure. The 594 articles classification under three heads was done. The maximum number of studies is empirical which marks 70%. Theoretical and conceptual represent the least space in the study. It shows the dearth of conceptual study majorly

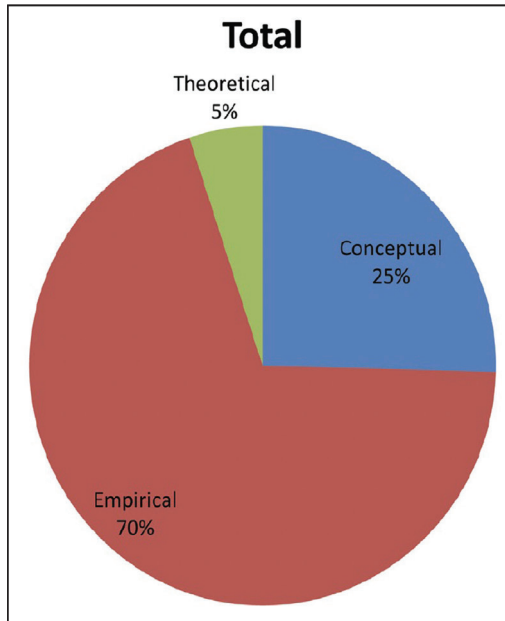


Figure 5. Study Percentages.

Source: The authors.

Table 4. Major Keywords of Capital Structure based on the Occurrence.

Keyword	Occurrences
Capital structure	362
Leverage	61
Trade-off theory	37
Pecking order theory	31
Speed of adjustment	22
Corporate governance	20
g32	18
Panel data	18
Smes	16
China	15

Source: The authors.

in this direction and also the reason for the existence of the capital structure problem which cannot be resolved just by going through and analysing the predefined data sets of the company and associating the increase or downfall of any content in capital structure with the capital structure theories rather a conceptual framework is highly demanded to meet unpredictable scenarios which are always not based on earlier patterns and statistics. The data in terms of percentage reveals the dearth of study in various economies. Thus the reason for a shortage of study could not be given to the size of the economy rather there is a lot in terms of money which is needed to be monitored and managed in terms of capital structure study and this cannot be neglected.

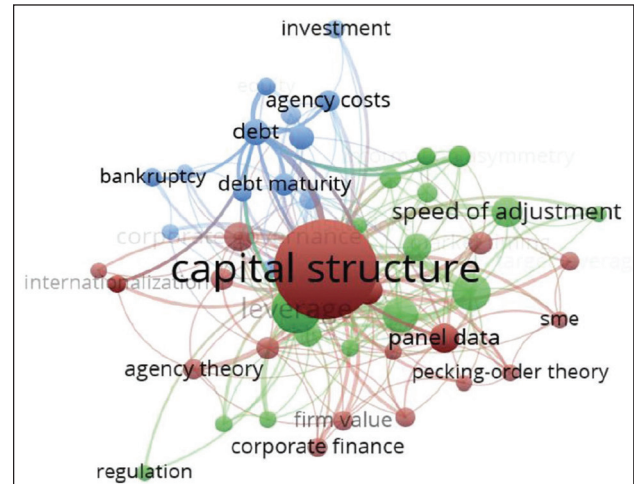


Figure 6. Keyword Analysis using VOS Viewer.

Source: The authors.

Table 5. Articles Classified in Each Cluster based on Citation Network Analysis.

Cluster 1 (Role of External Financing in Capital Structure Dynamics)	Cluster 2 (Factors Affecting Capital Structure Decisions of a Firm)	Cluster 3 (Capital Structure Determinants and Firm Performance)
Bae et al. (2011)	Ampenberger et al. (2013)	Campello and Giambona (2013)
Berk et al. (2010)	Baghai et al. (2014)	Deangelo et al. (2011)
Bhamra et al. (2010)	Buettner et al. (2012)	Hackbarth and Mauer (2012)
Chang et al. (2014)	Céspedes et al. (2010)	Klasa et al. (2018)
Chemmanur et al. (2013)	Detthamrong et al. (2017)	Matsa (2010)
Cho et al. (2014)	Feld et al. (2013)	Morellec et al.(2012)
Fan et al. (2012)	Jiraporn et al. (2012)	Rampini and Viswanathan (2013)
Graham and Leary (2011)	Kayo and Kimura (2011)	Rauh and Sufi (2010)
Graham et al. (2015)	Kesternich and Schnitzer (2010)	Serfling (2016)
Hovakimian and Li (2011)	Margaritis and Psillaki (2010)	Uysal (2011)
Kieschnick and Moussawi (2018)	Shivdasani and Stefanescu (2010)	Xu (2012)
Lemmon and Zender (2010)	Le and Phan (2017)	
Öztekin and Flannery (2012)		
Öztekin (2015)		
Welch (2011)		
Zhang et al. (2015)		

Source: The authors.

Keyword Analysis

The co-occurrence of author, keywords is analysed and is mentioned in Table 4. The analysis is undertaken with a minimum of five occurrences of the keyword. Out of the 1,234 keywords, 47 meet the threshold. For each of the 47 keywords, the total strength of the co-occurrence links with the other keywords will be calculated. The 47 keywords with the greatest total link strength are selected. The co-occurrence of the keyword network is shown in Figure 6 with the help of VOSviewer software. It is a software tool used for visualization of the network in the bibliometric studies.

Cluster Content Analysis on the Basis of Citation Network

The Table 5 shows the division of three clusters with a common theme representing them followed by a thorough content analysis of each clusters.

Cluster 1 (Role of External Financing in Capital Structure Dynamics)

Cluster 1 is the biggest among the three clusters comprising 16 documents focusing on the various dimensions of external financing about the firm's capital structure requirement. The institutional environment has a significant impact on the capital structure of firms in both developing and developed countries. The country's legal and tax system, the corruption level and the choice of capital suppliers speak a lot about the variation in the leverage ratio.

It is further notified that countries having efficient legal system promotes more debt usage majorly short term while usage of higher long-term debt is related to bankruptcy codes and deposit insurance. The countries have higher debt usage where there is higher tax gain on its leverage (Fan et al., 2012). The trend analysis from the year 1945 to 1975 in US corporations' shows a drastic increase in debt consumption. Such increase affected firms of all sizes. It was stated that firm characteristics were not the reason for this increase but a change in government borrowing, macroeconomic uncertainty and financial sector development were the significant partners of this hike in debt usage by the US corporations (Graham et al., 2015). On an investigation of the firm's capital structure adjustment across countries, it has been found that legal and financial traditions have a prominent correlation with firm adjustment speeds, while the institutional arrangement is narrowly related with firm capital structure adjustment speeds (Öztekin & Flannery, 2012). The enquiry of international determinants of capital structure from the sample of firms from 37 countries reveals firm size, tangibility, industry, leverage, profits and inflation as the reliable determinants. On further examination, it has been found

that the quality of country's institutions affects leverage and the adjustment speed towards target leverage significantly. It signifies that premium quality institutions promotes smoother leverage adjustment while the places where there are laws in favour of debt holders than stockholders their leverage is found to hold a higher position (Öztekin, 2015). The study claiming its uniqueness in capital structure and capital issuing literature generates two distinct perspectives. It highlights certain problems in capital structure research on the ambiguity of considering non-financial liabilities as debt and not to be considered as equity. It further quantifies that equity issuing activity should not be related equivalently to the change in capital structure procedure (Welch, 2011). The observation drawn from the investigation of the stakeholder theory of capital structure concerning to a firm's relationship with its employees reveals that firms that believes in employee benefits are found to maintain low debt ratios. It has been notified that the ability to follow a fair employee treatment is an important financing factor (Bae et al., 2012). It is stated that firms optimal capital structure is largely influenced by the trade-off between human costs and the tax benefits of debt by the analysis made on an optimal labour contracts for a levered firm in perfect competitive capital and labour markets. It is observed that employees are fixed under this contract and resulting in increased bankruptcy costs (Berk et al., 2010). The study conducted with review of literatures from the year 2005 onwards documents three dimensions of capital structure variations under cross-firm, cross-industry and within firm specifically. The explanation of such variation is supported by the trade-off and pecking order theories of capital structure which highlights the shortcomings empirically (Graham & Leary, 2011).

Cluster 2 (Factors Affecting Capital Structure Decisions of a Firm)

This cluster is composed of 12 articles representing the various determinants of capital structure and its relation with the firm value. It gives the diversified literature on the investigation of German family firms over the period 1995–2006, it was ascertained that German family firms were found to have lower debt ratio than non-family firms (Ampenberger et al., 2013). The driving force to conduct such study was to know the impact of founding families on the capital structure of their firms and the result reveals the vital role of management in the creation of family impact on capital structure. This cluster defines the progression of studies on capital structure determinants. It started with ownership structure in earlier 2010 and then covered the impact of corporate governance, company taxation, hierarchical determinants, impact of pensions and the impact on firm value in the later years. It is reported that the firms affected by the rating agencies have lower debt spreads than the firms not affected with the same

rating representing the conservative impact concerning to debt usage in the firm (Baghai et al., 2014). The analysis on the effectiveness of the limitation of tax deductibility of interest expenses for corporations reveals that such thin capitalization rules reduce the internal loans usage for tax planning and promote higher debt consumption in the capital structure of the multinational corporations of foreign subsidiaries located in OECD countries (Buettnner et al., 2012). The analysis conducted in selected Latin American countries on the capital structure determinants reveals the similar debt levels to that of the US firms. Although it is believed that Latin American firms experience lower tax benefits and higher bankruptcy costs. It explores the relation of ownership concentration with firm's capital structure. It reports a positive relation between leverage and ownership concentration and also with growth as it is believed that issue of equity exploits the concentration of ownership and disturbs the control rights. It is further found that larger firms with more tangible assets are less profitable and more leveraged. Hence apart from firm's specific determinants of capital structure, there are various other factors such as theoretical, legal and technical which determine the capital structure financing of the firm (Céspedes et al., 2010).

The attempt made to identify the implication of capital structure of corporate pension plans benefits reveals that leverage ratios for firms with pension plans report 35% increase in the presence of pension assets and liabilities into the capital structure and thus impacting tax shields drastically. It presents a modest effect in lowering a firm's corporate tax rate. It stimulates the choice of capital structure finance and fosters a less conservative approach towards a choice of leverage (Shivdasani & Stefanescu, 2010). The study to investigate the effect of capital structure on firm performance in Vietnam reveals negative relation between debt ratios and firm performance. It highlights that developing market like Vietnam has less benefit from debt usage than the financial distress cost. Such a market is found to be influenced by severe information asymmetry generating an adverse role of debt (Le & Phan, 2017).

Cluster 3 (Capital Structure Determinants and Firm Performance)

This Cluster comprises 11 documents. The area of discussion under this cluster is around the factors affecting the capital structure decisions of a firm. On the review of literature, it has been obtained that the examination of firms' adjustment towards the priority structure identified as leverage, credit conditions and firm fundamentals reveals financially unconstrained firms with lesser growth opportunities prefer senior debt than constrained firms with or without growth opportunities who prefer junior

debt (Hackbarth & Mauer, 2012). The firm with the external finance constraint has the incentive to avail cash flow demands of debt service to substantiate its convincing position with workers (Matsa, 2010). A dynamic trade-off model was developed to elaborate the manager-shareholder conflicts in capital structure choice. It reveals the manager's behaviour towards owning a fraction of the firm's equity to capture free cash flow as the private benefits and maintain control over financing decisions. It further attempts to detangle the use of the low leverage puzzle and explains leverage ratio dynamics with the use of data on leverage choices (Morellec et al., 2012). An investigation on the relationship between collateral and capital structure was examined through a study of a dynamic model of investment, capital structure, and leasing and risk management depending on firms' need to collateralise promises through tangible assets. It shows that financing and risk management both demand promises to be paid with limited collateral while leasing demands strong collateral, costly financing and allows greater leverage. Thus it represents that more constrained firms hedge less and lease more and the firms aiming for productivity reduce the benefits of hedging low cash flows and restrict firms not to hedge (Rampini & Viswanathan, 2013). An enquiry into the relatedness of the cost associated with discharging workers affecting capital structure decisions shows that firm's debt ratio proportion reduces in the presence of labour protection laws. It eventually increases the degree of operating leverage, earnings variability and employment becomes more rigid. This proves to substantiate that higher firing costs eliminate the financial leverage through increasing financial distress costs (Serfling, 2016).

Citation, Co-citation and Co-authorship

The Table 6 shows the citation analysis of authors, documents, countries and organisation is an important parameter to measure and identify the most popular author and document, the country engaged on that particular subject and the organisation associated with it by the count of citations it has. Here the author named Oeztekin oezde has received a maximum citation for his work on the mentioned subject. The document which has the maximum citation is Fan et al. (2012) and the country which is highly recognised and cited is the USA and the organisation which has got highest number of citations is NBER (National Bureau of Economic Research). This data pronounces the quality consideration given in research of the capital structure study by the respective organisation, countries and authors expressed through the citations analysis which drives the attention of other countries and organisation in this research area.

Table 6. Citation Analysis.

Citation Analysis										
Author			Documents			Country			Organisation	
Author	Documents	Citations	Document	Citations	Country	Documents	Citations	Organization	Documents	Citations
El Ghoul, Sadok	3	100	Robb and Robinson (2014)	210	Australia	33	735	Cent University Finance and Economics	5	235
Flannery, Mark J.	3	268	Fan et al. (2012)	315	Canada	24	615	Cornell University	8	221
Giambona, Erasmo	3	101	Öztekin and Flannery (2012)	214	England	60	910	Duke University	5	593
Guedhami, Omrane	3	100	Bae et al. (2011)	180	France	18	168	National Bureau of Economic Research	6	448
Jiraporn, Pornsit	3	96	Chen et al. (2010)	192	Germany	26	699	NBER	8	714
Öztekin, Oezde	3	350	Rauh and Sufi (2010)	174	Italy	28	435	Northwestern University	5	461
Overesch, Michael	3	143	Lemmon and Zender (2010)	168	Peoples R China	105	1548	Purdue University	5	167
Schmid, Thomas	3	111	Gropp and Heider (2010)	212	Spain	28	256	University of British Columbia	6	251
Seller, Michael J.	3	71	Matsa (2010)	175	Taiwan	32	350	University of Florida	5	303
Xu, Jin	3	117	Margaritis and Psillaki (2010)	208	USA	185	5879	University of Tennessee	5	169

Source: The authors.

Reviewed Papers

Table 7. Review of 58 Research Papers.

Study	Type of Paper	Context	Source
Tran, D. V.; Hassan, M. K.; Paltrinieri, A.; Nguyen, T. D.	Conceptual	Determinants of the bank capital structure are the same as that of non-financial firms.	Singapore Economic Review
Do, T. K.; Lai, T. N.; Tran, T. T. C.	Empirical	The foreign ownership proves to be positively influencing the optimum capital structure formation of a firm.	Finance Research Letters
Im, H. J.; Kang, Y.; Shon, J.	Empirical	The study reports the impact of uncertainty considerably higher than any other determinants of leverage targets.	Journal of Corporate Finance
Lim, S. C.; Macias, A. J.; Moeller, T.	Empirical	The poor collateralizability of goodwill results in less debt financing.	Journal of Banking and Finance
Dai, N.; Piccotti, L. R.	Empirical	The required return on equity is matched with the target debt ratio.	Financial Management
Ayotte, K.	Empirical	Relation between capital structure and disagreement of investor in asset valuation is discussed.	Journal of Legal Studies
Fenyves, Veronika; Peto, Karoly; Szenderak, Janos; Harangi-Rakos, Monika	Conceptual	The study conducted in V4 countries: the Czech Republic, Hungary, Poland and Slovakia in order to analyse the capital structure of agricultural companies. The strong influence has been reported by such sector and its company size in context to capital structure.	Agricultural Economics (zemedelska ekonomika)
Ghasemzadeh, Morteza; Heydari, Mehdi; Mansourfar, Gholamreza	Empirical	The relation between earnings volatility and capital structure is brought into light along with the impact of financial distress as a moderating variable.	Emerging Markets Finance and Trade
Frank, Murray Z.; Shen, Tao	Conceptual	This article shows the financial actions of firms to adjust leverage.	Journal of Banking and Finance
Lambrinoudakis, Costas; Skiadopoulos, George; Gkionis, Konstantinos	Empirical	An aspect related to the financial flexibility is brought into picture which suggests that the expectation of firm specific investment shocks also affects firms leverage.	Journal of Banking and Finance
Harris, Christopher; Roark, Scott	Empirical	The level of debt in the capital structure is found to be related with the operating cash flow. Firms having higher cash flow volatility have interest and hence increases firm value by utilizing the debt consumption in the capital structure adequately..	Finance Research Letters
Doku, James Ntiamoah; Kpekpena, Fred Agbenya; Boateng, Prince Yeboah	Empirical	The capital structure measured as capital to asset ratio is found evident as a positive driver of bank performance	African Development Review (revue africaine de developpement)
Lemmon, Michael L.; Zender, Jaime F.	Conceptual	The capital structure choice studied in the presence of asymmetric information with respect to the standard pecking order and trade off theories.	Journal of Financial and Quantitative Analysis
Matemilola, B. T.; Bany-Ariffin, A. N.; Azman-Saini, W. N. W.; Nassir, Annur Md.	Theoretical	The top level managers (CEO) experience proves beneficial in the capital structure decision as it optimizes the benefit of tax shield available on debt interest and hence increases firm value by utilizing the debt consumption in the capital structure adequately.	Research in International Business and Finance
Alan, Yasin; Gaur, Vishal	Empirical	The application of asset based lending in the context of capital structure decision in the banks is being demonstrated.	M&om-Manufacturing & Service Operations Management
Dragota, Ingrid-Mihaela; Dragota, Victor; Curmei-Semenescu, Andreea; Pele, Daniel Traian	Conceptual	The relation between the capital structure and religion has been investigated to understand the influence of different cultural values on the financial variables.	Acta Oeconomica
Wang, Xiaoqiao; Johnson, Lewis; Wang, Jin	Conceptual	The firms located centrally have proper information access and has lower leverage ratios than remotely located ones.	Canadian Journal of Administrative Sciences (revue canadienne des sciences de l administration)
Kieschnick, Robert; Moussawi, Rabih	Conceptual	It is found that the amount of debt usage depends on the governance function. The more the concentration of power in the hands of insider, lesser is the debt usage as it ages.	Journal of Corporate Finance

(Table 7 continued)

(Table 7 continued)

Study	Type of Paper	Context	Source
Wang, Huamao; Xu, Qing; Yang, Jinqiang	Empirical	The impact of neglecting the liquidity risk has been reported which results in over leveraging, early bankruptcy, and many more.	European Journal of Finance
Yang, Shenggang; He, Feiying; Zhu, Qi; Li, Shihao	Conceptual	The company incorporating the corporate social responsibility strategies has higher leverage levels than the ones that do not. CSR bridged the gap of information asymmetry between firms and creditors.	Asia-pacific Journal of Accounting and Economics
Miglo, Anton	Conceptual	The influence of asymmetric information in terms of timings of earnings on capital structure is discussed.	North American Journal of Economics and Finance
Oliveira, Mauro; Kadapakkam, Palani-Rajan; Beyhaghi, Mehdi	Empirical	It accommodates the impact on the capital structure of the supplier by increasing the level of leverage in order to curb the negotiation from distressed customer's end.	Journal of Corporate Finance
Abdulla, Yomna	Empirical	It supports that a single theory is not sufficient and cannot be dependent upon to explain the capital structure behavior.	International Journal of Islamic and Middle Eastern Finance and Management
Ardalan, Kavous	Conceptual	It focuses on the assumptions which make the capital structure foundation relevant as it believes that results of the numerical models are dependent on the assumption laid and thus changes accordingly.	Research in International Business and Finance
Tung Lam Dang; Thi Hong Hanh Huynh; Manh Toan Nguyen; Thi Minh Hue Nguyen	Empirical	The studies points out the major role of analyst characteristics and information transparency on capital structure choices.	Applied Economics
Rastad, Mandi	Empirical	It has been observed that when the equity shares crosses the conversion criterion for a convertible bond then firm expects the decline in leverage in the future.	Journal of Corporate Finance
Reinartz, Sebastian J.; Schmid, Thomas	Empirical	An analysis has been conducted of energy utilities on a global sample. Production flexibility increases financial leverage in the light of reduced expected cost of financial distress and higher present value of tax shields.	Review of Financial Studies
Eun, Cheol S.; Wang, Lingling	Conceptual	International sourcing has negative influence on financial leverage.	Review of Finance
Tan, Yingxian; Yang, Zhaojun	Conceptual	Impact of contingent convertible bonds on capital structure is investigated.	North American Journal of Economics and Finance
Li, Xuefeng	Empirical	The significant impact of determinant of capital structure of Chinese firms reveals that leverage increases with growth opportunity.	Journal of Computational and Theoretical Nanoscience
Faccio, Mara; Xu, Jin	Empirical	The corporate and personal income taxes are found to be the significant determinants of the capital structure.	Journal of Financial and Quantitative Analysis
Tchuigoua, Hubert Tchakoute	Empirical	It states that apart from profitability the other variables have the same impact in both profit microfinance institutions and non-profit finance institutions.	Journal of Financial Services Research
Koksal, Bulent; Orman, Cuneyt	Empirical	The study conducted in a developing economy, Turkey to understand the applicability of capital structure theories in almost every sort of companies from manufacturing to non-manufacturing from small to large and from public traded to private firms.	Small Business Economics
Wong, Kit Pong	Empirical	The capital structure behaviour is examined from the managers risk aversion or regret perspective.	Finance Research Letters
Danis, Andras; Rettl, Daniel A.; Whited, Toni M.	Empirical	It reports positive correlation between profitability and leverage at the time of optimum level of leverage in firm.	Journal of Financial Economics
Cvijanovic, Dragana	Empirical	The impact of real estate prices on firm capital structure is examined. It reveals that change in debt structure with respect to collateral value appreciation.	Review of Financial Studies

(Table 7 continued)

(Table 7 continued)

Study	Type of Paper	Context	Source
Dang, Viet Anh; Kim, Minjoo; Shin, Yongcheol	Empirical	A negative impact has been reported of the global finance crisis on the speed of leverage adjustment.	International Review of Financial Analysis
Chod, Jiri; Zhou, Jianer	Empirical	It determines the influence of resource flexibility on capital structure shows that resource flexibility is negatively related to the cost of borrowing and positively related to debt.	Management Science
Robb, Alicia M.; Robinson, David T.	Empirical	The analysis of capital structure choices made in the firm from the data set of Kaufman firm survey shows the strong reliance on the external debt sources such as bank finance and less on family funding sources.	Review of Financial Studies
Chemmanur, Thomas J.; Cheng, Yingmei; Zhang, Tianming	Empirical	It was found that leverage and a positive significant impact on average employee pay and CEO compensation.	Journal of Financial Economics
Chung, Y. Peter; Na, Hyun Seung; Smith, Richard	Empirical	Firms tend to increase leverage in the presence of growth opportunities or when there is decline in equity value. Acquisition of firms cannot be dependent on any capital structure policy rather it is a result of financial slack due to rapid growth created.	Journal of Corporate Finance
Rampini, Adriano A.; Viswanathan, S.	Empirical	The collateral plays a vital role in capital structure financing. Leasing collateralizes financing with greater leverage and firms facing adverse cash flow shocks prefers to sell and lease back assets.	Journal of Financial Economics
Harding, John P.; Liang, Xiaozhong; Ross, Stephen L.	Empirical	Banks capital structure is analysed in the light of capital requirements, deposit insurance and franchise value.	Journal of Financial Services Research
Bartoloni, Eleonora	Empirical	The method applied is Granger Causality framework. The framework explains that firms leverage does not result innovation output rather leverage is caused by successful innovation and a firm operating profitability.	Empirica
Matemilola, B.; Bany-Ariffin, A.; McGowan, Carl	Empirical	The study conducted on the capital structure research in South Africa accommodates significant managerial skills and ability as the unobserved firm specific capital structure determinants.	Managerial Finance
Dudley, Evan.	Empirical	It shows that firms move towards their target capital structure in the course of investment as per the trade-off theory.	Journal of Corporate Finance
Jiraporn, Pornsit; Chintrakarn, Pandej; Liu, Yixin	Theoretical	It is found that companies having higher CEO dominance has a negative impact of changes in capital structure on firm performance.	Journal of Financial Services Research
Paligorova, Teodora; Xu, Zhaoxia	Conceptual	The study explores pyramidal firms and their motivation behind debt usage and the firm financing is found to be influenced by the expropriation activities of owners having control rights.	Journal of Corporate Finance
Rauh, Joshua D.; Sufi, Amir	Empirical	The firm's production and assets used in it are the important determinants of capital structure in the cross section.	Review of Finance
Bortolotti, Bernardo; Cambini, Carlo; Rondi, Laura; Spiegel, Yossi	Empirical	There is a significant role of control and regulation in the firm's financing decision. The privately controlled firm's debt usage is designed in a way which enables efficient regulatory outcomes.	Journal of Economics and Management Strategy
Pindado, Julio; de la Torre, Chabela	Empirical	Capital structure is largely influenced by the managerial ownership and ownership concentration.	International Review of Finance
Bertomeu, Jeremy; Beyer, Anne; Dye, Ronald A.	Conceptual	A perspective that jointly determines capital structures its voluntary disclosure policy and its cost of capital.	Accounting Review
Gao, Wenlian; Ng, Lilian; Wang, Qinghai	Conceptual	The study conducted in United States explains the variation of capital structure due to location of corporate headquarters. The results support the significant role of local culture and executives social interactions on corporate financial policies.	Financial Management

(Table 7 continued)

(Table 7 continued)

Study	Type of Paper	Context	Source
Stretcher, Robert; Johnson, Steve	Theoretical	Application of theory in practice in terms of capital structure choice to make informed decisions by managers.	Managerial Finance
Rauh, Joshua D.; Sufi, Amir	Empirical	The ignorance of debt issues in capital structure variation is brought into light.	Review of Financial Studies
Berk, Jonathan B.; Stanton, Richard; Zechner, Josef	Empirical	The firm optimal capital structure determined on the grounds of trade off between human costs and tax benefits derived from debt usage.	Journal of Finance
Margaritis, Dimitris; Psillaki, Maria	Empirical	The study undertaken on French manufacturing firms enquires the relationship between capital structure, ownership structure and firm performance.	Journal of Banking and Finance
Shibata, Takashi; Nishihara, Michi	Empirical	It examines the investment and capital structure financing decisions of a firm under management shareholder conflicts caused due to asymmetric information. It reveals such conflict increases investment and decreases social welfare by debt financing.	Journal of Economic Dynamics and Control
Saito, Richard; Hiramoto, Eduardo	Empirical	It tests the Brazilian companies experiencing foreign activities have different capital structure in contrast to companies experiencing local activities. It examines the relation between international activity and debt financing.	Academia Revista Latinoamericana de Administracion

Source: The authors.

Discussion and Gap Direction

The capital structure study undertaken provides tremendous valuable information. It reveals the publication in terms of numbers from the year 2010 to 2021. There have been maximum studies in the year 2019 in comparison to the earlier years. The position of increase and decrease in the studies which can find a place in the reputed database is being recorded. Then further the leading journals were identified based on their ABDC ranking and ABS categories to represent the importance of the subject. Later it was decided to understand the expansion of the subject in various disciplines. It was clear that business economics holds the maximum study in its domain but at the same time this could not be neglected that the capital structure subject holds its place in other disciplines too which are majorly environmental sciences, operations research and engineering. The authors having the maximum publications and citations, the top institutions and the countries having maximum publications and citations, are recorded. After that, the types of the study were ascertained which revealed the information related to the maximum area of empirical studies. It was then decided to understand the various associated areas concerning capital structure, for that keyword analysis was conducted which gave us an appropriate picture of the diversification of capital structure study in terms of leverage, a trade-off, pecking order theory, corporate governance and many other important fields. Finally, a regressive analysis of 58 research papers was done after reading their abstracts to understand their uniqueness.

The citation analysis marks the contribution of eminent authors and the documents. It identifies the most cited

study is from the USA and the organisation which is most popular on this subject is NBER. Thus encouraging the study in this direction in developing countries where it is still uncovered. The cluster analysis conducted brings various studies on the topic which were further divided into three clusters. Cluster 1 highlights the various relationships between maintaining low debt ratios and employee benefits, macro-economic uncertainty in the context of capital structure financing. The cluster 2 represents that ownership structure, rating agencies, corporate governance and the impact of pension assets and liabilities on the capital structure is considered significant and cluster 3 discusses the various factors affecting the capital structure decisions of a firm. It reveals the role of debt, manager–shareholder conflicts in capital structure choice. It highlights that the financing decision is determined by the growth opportunities, collaterals, risk management and laws prevailing.

All such information collectively speaks a lot about the subject of capital structure. There is a dearth of studies including the amalgamation of theory with some models. There have been a great financial crisis impacting the economy globally, the loss by them was neither negligible nor can be buried. If we try to understand the pattern of study over the years then we will notice a tremendous increase in its importance from the year 2011 which shows a decline in between and then again it shows a rapid increase in 2019. This reveals the importance of the study in recent years. The researchers are rigorously attempting to showcase its necessity and thus striving for prolific journals. The gap identified from this review highlights the vast fragmentation on this topic. The focus is more on

Table 8. Research Gap and Future Directions.

Research Gap	Future Direction
Scant studies in developing countries	To overcome the differences between financing decision of developed and developing countries
Lack of multi country studies	The studies from countries like South America, Oceania and African region to be added
Lack of model based study	The experimentation with model creation with conceptual background to be undertaken
Scant study on cross country comparisons	Studies on comparison of cross-country aspects to be conducted
Determining new capital structure determinants	Exploring impact of non-tested capital structure determinants

Source: The authors.

substantiating findings with decades-old theories. The theories are very explanatory, complete and specific but the problem of capital structure financing in every organisation and economy is existing and is still untouched in developing economies. The developed countries have taken maximum initiative towards the analysis of this subject which is marked by the bibliometric study. This study attempts to highlight the gap and future research required in this direction.

Limitation

There are certain limitations that are to be addressed despite following all the necessary bibliometric and systematic literature review protocols. Firstly the study includes the article which is only articles and excludes studies in a language other than English thus acknowledging the exclusion of some important literature from conference proceedings, book chapters and in other languages. Secondly, it has considered WOS database and no other databases which limited its literature and data. Thus the author urges future researchers to consider and develop more elaborative studies and includes databases like Scopus, IEEE and ACM in their data to enhance this research by overcoming such limitations.

Conclusion

The crisis in the ignorance or mismanagement of capital structure has given rise to vulnerable situations in the industry. This study on capital structure is undertaken with the aim to connect the existing literature to derive a valuable output. The capital structure should be relooked and methods should be framed in new frames. The capital structure study should be designed in every area considering its relevance and it is suggested that this should not be

completely empirical with the existing data set and neither the conclusion should be derived by associating the outcomes with the pre-defined capital structure theories. There is a scarcity of methods which involve an amalgamation of theory with model formation, and more conceptual study is highly demanded to overcome the lack of knowledge in terms of capital structure formation. An understanding of appropriate capital structure financing is needed globally. Hence, there should be some capital structure studies that are to be undergone with some predictions and concrete concepts to absorb the disaster caused in the absence of capital structure study.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors received no financial support for the research, authorship and/or publication of this article.

ORCID iD

Anjali Sisodia  <https://orcid.org/0000-0001-7224-2605>

References

- Ampenberger, M., Schmid, T., Achleitner, A. K., & Kaserer, C. (2013). Capital structure decisions in family firms: Empirical evidence from a bank-based economy. *Review of Managerial Science*, 7(3), 247–275.
- Bae, K. H., Kang, J. K., & Wang, J. (2011). Employee treatment and firm leverage: A test of the stakeholder theory of capital structure. *Journal of Financial Economics*, 100(1), 130–153.
- Baghai, R. P., Servaes, H., & Tamayo, A. (2014). Have rating agencies become more conservative? Implications for capital structure and debt pricing. *The Journal of Finance*, 69(5), 1961–2005.
- Baker, M., & Wurgler, J. (2002). Market timing and capital structure. *The Journal of Finance*, 57(1), 1–32.
- Berk, J. B., Stanton, R., & Zechner, J. (2010). Human capital, bankruptcy, and capital structure. *The Journal of Finance*, 65(3), 891–926.
- Bhamra, H. S., Kuehn, L. A., & Strebulaev, I. A. (2010). The aggregate dynamics of capital structure and macroeconomic risk. *The Review of Financial Studies*, 23(12), 4187–4241.
- Bradley, M., Jarrell, G. A., & Kim, E. (1984). On the existence of optimal capital structure: Theory and evidence. *Journal of Finance*, 39(3), 857–878.
- Buettner, T., Overesch, M., Schreiber, U., & Wamser, G. (2012). The impact of thin-capitalization rules on the capital structure of multinational firms. *Journal of Public Economics*, 96(11–12), 930–938.
- Campello, M., & Giambona, E. (2013). Real assets and capital structure. *Journal of Financial and Quantitative Analysis*, 48(5), 1333–1370.
- Céspedes, J., González, M., & Molina, C. A. (2010). Ownership and capital structure in Latin America. *Journal of Business Research*, 63(3), 248–254.

- Chang, C., Chen, X., & Liao, G. (2014). What are the reliably important determinants of capital structure in China?. *Pacific-Basin Finance Journal*, 30, 87–113.
- Chemmanur, T. J., Cheng, Y., & Zhang, T. (2013). Human capital, capital structure, and employee pay: An empirical analysis. *Journal of Financial Economics*, 110(2), 478–502.
- Chen, H. L., Hsu, W. T., & Huang, Y. S. (2010). Top management team characteristics, R&D investment and capital structure in the IT industry. *Small Business Economics*, 35(3), 319–333.
- Cho, S. S., El Ghouli, S., Guedhami, O., & Suh, J. (2014). Creditor rights and capital structure: Evidence from international data. *Journal of Corporate Finance*, 25, 40–60.
- DeAngelo, H., DeAngelo, L., & Whited, T. M. (2011). Capital structure dynamics and transitory debt. *Journal of Financial Economics*, 99(2), 235–261.
- Deb, S. G., & Banerjee, P. (2018). Low leverage policy: A boon or bane for Indian shareholders. *Journal of Asia Business Studies*, 12(4), 489–507.
- Detthamrong, U., Chancharat, N., & Vithessonthi, C. (2017). Corporate governance, capital structure and firm performance: Evidence from Thailand. *Research in International Business and Finance*, 42, 689–709.
- Fan, J. P., Titman, S., & Twite, G. (2012). An international comparison of capital structure and debt maturity choices. *Journal of Financial and Quantitative Analysis*, 47(1), 23–56.
- Feld, L. P., Heckemeyer, J. H., & Overesch, M. (2013). Capital structure choice and company taxation: A meta-study. *Journal of Banking & Finance*, 37(8), 2850–2866.
- Fischer, E. O., Heinkel, R., & Zechner, J. (1989). Dynamic capital structure choice: Theory and tests. *Journal of Finance*, 44(1), 19–40.
- Graham, J. R., & Leary, M. T. (2011). A review of empirical capital structure research and directions for the future. *Annual Review of Financial Economics*, 3, 309–345.
- Graham, J. R., Leary, M. T., & Roberts, M. R. (2015). A century of capital structure: The leveraging of corporate America. *Journal of Financial Economics*, 118(3), 658–683.
- Gropp, R., & Heider, F. (2010). The determinants of bank capital structure. *Review of Finance*, 14(4), 587–622.
- Hackbarth, D., & Mauer, D. C. (2012). Optimal priority structure, capital structure, and investment. *The Review of Financial Studies*, 25(3), 747–796.
- Harris, C., & Roark, S. (2019). Cash flow risk and capital structure decisions. *Finance Research Letters*, 29, 393–397.
- Hovakimian, A., & Li, G. (2011). In search of conclusive evidence: How to test for adjustment to target capital structure. *Journal of Corporate Finance*, 17(1), 33–44.
- Iqbal, J., Muhammad, S., Muneer, S., & Jahanzeb, A. (2012). A critical review of capital structure theories. *Information Management and Business Review*, 4(11), 553–557.
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360.
- Jiraporn, P., Kim, J. C., Kim, Y. S., & Kitsabunnarat, P. (2012). Capital structure and corporate governance quality: Evidence from the Institutional Shareholder Services (ISS). *International Review of Economics & Finance*, 22(1), 208–221.
- Kane, A., Marcus, A. J., & McDonald, R. L. (1984). How big is the tax advantage to debt? *Journal of Finance*, 39(3), 841–853.
- Kayo, E. K., & Kimura, H. (2011). Hierarchical determinants of capital structure. *Journal of Banking & Finance*, 35(2), 358–371.
- Kesternich, I., & Schnitzer, M. (2010). Who is afraid of political risk? Multinational firms and their choice of capital structure. *Journal of International Economics*, 82(2), 208–218.
- Kieschnick, R., & Moussawi, R. (2018). Firm age, corporate governance, and capital structure choices. *Journal of Corporate Finance*, 48, 597–614.
- Klasa, S., Ortiz-Molina, H., Serfling, M., & Srinivasan, S. (2018). Protection of trade secrets and capital structure decisions. *Journal of Financial Economics*, 128(2), 266–286.
- Korom, P. (2019). A bibliometric visualization of the economics and sociology of wealth inequality: A world apart? *Scientometrics*, 118(3), 849–868.
- Kraus, A., & Litzenberger, R. H. (1973). A state-preference model of optimal financial leverage. *Journal of Finance*, 28(4), 911–922.
- Kumar, S., Colombage, S., & Rao, P. (2017). Research on capital structure determinants: A review and future directions. *International Journal of Managerial Finance*, 13(2), 106–132.
- Le, T. P. V., & Phan, T. B. N. (2017). Capital structure and firm performance: Empirical evidence from a small transition country. *Research in International Business and Finance*, 42, 710–726.
- Lemmon, M. L., & Zender, J. F. (2010). Debt capacity and tests of capital structure theories. *Journal of Financial and Quantitative Analysis*, 45(5), 1161–1187.
- Lugi, P., & Sorin, V. (2009). A review of the capital structure theories. *Annals of University of Oradea, Economics Science Series*, 18(3), 315–320.
- Madan, K. (2007). An analysis of the debt-equity structure of leading hotel chains in India. *International Journal of Contemporary Hospitality Management*, 19(5), 397–414.
- Margaritis, D., & Psillaki, M. (2010). Capital structure, equity ownership and firm performance. *Journal of Banking & Finance*, 34(3), 621–632.
- Matsa, D. A. (2010). Capital structure as a strategic variable: Evidence from collective bargaining. *The Journal of Finance*, 65(3), 1197–1232.
- Miglo, A. (2010). *The pecking order, trade off, signalling and market timing theories of capital structure: A review* (Working Paper). University Library of Munich.
- Miller, M. H. (1977). Debt and taxes. *Journal of Finance*, 32(2), 261–275.
- Modigliani, F., & Miller, M. H. (1958). The cost of capital, corporation finance and the theory of investment. *American Economic Review*, 48(3), 261–297.
- Modigliani, F., & Miller, M. H. (1963). Corporate income taxes and the cost of capital: A correction. *American Economic Review*, 53(3), 433–443.
- Morellec, E., Nikolov, B., & Schürhoff, N. (2012). Corporate governance and capital structure dynamics. *The Journal of Finance*, 67(3), 803–848.
- Myers, S. C. (1993). Still searching for optimal capital structure. *Journal of Applied Corporate Finance*, 6(1), 4–14.
- Myers, S. C., & Majluf, N. S. (1984). Corporate financing and investment decision when firms have information investors do not have. *Journal of Financial Economics*, 13(2), 187–221.
- Öztekin, Ö. (2015). Capital structure decisions around the world: Which factors are reliably important? *Journal of Financial and Quantitative Analysis*, 50(3), 301–323.

- Öztekin, Ö., & Flannery, M. J. (2012). Institutional determinants of capital structure adjustment speeds. *Journal of Financial Economics*, 103(1), 88–112.
- Paul, J., & Criado, A. R. (2020). The art of writing literature review: What do we know and what do we need to know? *International Business Review*, 29(4), 101717.
- Rampini, A. A., & Viswanathan, S. (2013). Collateral and capital structure. *Journal of Financial Economics*, 109(2), 466–492.
- Rauh, J. D., & Sufi, A. (2010). Capital structure and debt structure. *The Review of Financial Studies*, 23(12), 4242–4280.
- Robb, A. M., & Robinson, D. T. (2014). The capital structure decisions of new firms. *The Review of Financial Studies*, 27(1), 153–179.
- Robichek, A. A., & Myers, S. C. (1966). Problems in the theory of optimal capital structure. *Journal of Financial and Quantitative Analysis*, 1(2), 1–35.
- Ross, S. A. (1977). The determination of financial structure: The incentive-signalling approach. *The Bell Journal of Economics*, 8(1), 23–40.
- Serfling, M. (2016). Firing costs and capital structure decisions. *The Journal of Finance*, 71(5), 2239–2286.
- Shivdasani, A., & Stefanescu, I. (2010). How do pensions affect corporate capital structure decisions? *The Review of Financial Studies*, 23(3), 1287–1323.
- Shivdasani, A., & Zenner, M. (2005). How to choose a capital structure: Navigating the debt-equity decision. *Journal of Applied Corporate Finance*, 17(1), 26–35.
- Stiglitz, J. E. (1973). Taxation, corporate financial policy and the cost of capital. *Journal of Public Economics*, 2(1), 1–34.
- Uysal, V. B. (2011). Deviation from the target capital structure and acquisition choices. *Journal of Financial Economics*, 102(3), 602–620.
- Welch, I. (2011). Two common problems in capital structure research: The financial-debt-to-asset ratio and issuing activity versus leverage changes. *International Review of Finance*, 11(1), 1–17.
- Xu, J. (2012). Profitability and capital structure: Evidence from import penetration. *Journal of Financial Economics*, 106(2), 427–446.
- Zhang, H., Gao, S., Seiler, M. J., & Zhang, Y. (2015). The effect of credit crunches and equity financing restrictions on the capital structure adjustments of Chinese listed real estate companies. *Emerging Markets Finance and Trade*, 51(sup5), S21–S32.

About the Authors

Anjali Sisodia (anjalis376@gmail.com) is currently pursuing Doctorate program as a Part-time Research Scholar from the prestigious Delhi Technological University, Delhi, India. She has worked as a visiting faculty in IGDTUW, Delhi, CVS college Delhi University and has more than 10 years of teaching experience in colleges of high repute in and across Delhi, NCR. Her areas of interest include corporate finance, cost accountancy and financial accounting.

Girish Chandra Maheshwari (gcmaheshwari2004@yahoo.com) is a professor of Management at Delhi Technological University. He has been Dean of Faculty of Management Studies, The MS University of Baroda. His teaching assignments spanned over four decades included stints at Delhi University (1972–1983), Indian Institute of Technology Delhi (1983–1984) and culminating at The MS University of Baroda (1984–2012). He has been author of several books (13) on Management and Accounting of which five are published by NCERT. He has published research papers (60) in national (50) and international journals. He is on the editorial boards of several research journals. His areas of interest cover Financial and Management Accounting, Accounting Theory.

Cervical Cancer Screening on Multi-class Imbalanced Cervigram Dataset using Transfer Learning

Manisha Saini*

Department of Computer Science and Engineering
Delhi Technological University
New Delhi, India
Email: manisha.saini44@gmail.com

Seba Susan

Department of Information Technology
Delhi Technological University
New Delhi, India
Email: seba_406@yahoo.in

Abstract—Image classification from a multi-class imbalanced dataset is challenging because it is difficult to detect all the minority classes present in the datasets. In this paper, the authors have extended their recently introduced work on a novel deep learning neural network for binary-class imbalanced datasets, called VGGIN-Net, by applying it to classify different grades of cervical cancer from a multi-class imbalanced dataset having a small number of samples. The experimental results prove that the proposed approach along with the data augmentation and rejection resampling is effective for multi-class imbalanced datasets. As analyzed through extensive experiments on a benchmark cervical screening dataset, the proposed approach in comparison to the other state-of-the-art approaches is vividly proven to be a more efficient method.

I. INTRODUCTION

Cervical cancer is one of the most common and deadly types of cancer that occurs in females. The death count due to this cancerous disease has increased tremendously worldwide, especially in developing countries, in the past few decades [1]. Screening for cancer is a very crucial aspect to cure it at its early stages. For cervical cancer screening, the first primary task is the detection of the cancer type which can be any of three known types. Type 1 cervix does not require screening but women having Type 2 and Type 3 cervix require screening for further cancer detection. However, manual screening and detection of the type of cervical cancer is problematic, time-consuming and tedious due to the high probability of occurrence of manual errors. So, an automated screening approach can increase the efficiency of cancer detection tasks.

Most biomedical datasets are imbalanced. There are several approaches available to tackle the brutal consequences of not considering the minority class while screening for cancer at its early stages. Undersampling, oversampling, and a hybrid of these two are the most common approaches available in literature to address the issues that occur due to biased datasets [2] [3] [4]. In the case of multi-class imbalanced classification, it is equally important to detect all the minority classes present in imbalanced datasets to increase the overall

efficiency of the model.

Saini and Susan (2020) applied Deep convolutional generative adversarial network (DCGAN) to tackle the class imbalance effect in the BreakHis dataset, and also proposed a modified VGG16 architecture [5]. In a subsequent work, the authors proposed another approach for tackling imbalance in multi-class settings using deep features extracted from ResNet in combination with the non-linear χ^2 SVM classifier and Bag-of-Visual-Words (BOVW) approach [6]. They explored the effect of data augmentation on minority classes and emphasized that the Inception-v3 model along with weighted SVM showed improved performance in comparison to other networks [7]. Matsuo et al. (2019) established the superior performance of deep learning models over the Cox proportional hazard regression model for the survival prediction of women with cervical cancer. They also discussed the improvement in the performance of the deep learning models after adding more features as compared to the Cox proportional hazard regression model [8].

In this paper, the authors have extended their recently introduced work on a novel deep learning neural network for binary-class imbalanced datasets, called VGGIN-Net [9], by applying it to classify different grades of cervical cancer in a multi-class imbalanced data scenario. The major contributions of the paper are (1) Transfer learning of VGGIN-Net deep neural architecture facilitates transfer of knowledge from a large dataset to the smaller, multi-class imbalanced cervical cancer dataset (2) data augmentation was applied successfully which had an overall impact on the performance of the deep learning network, along with a random undersampling strategy, that also helps to avoid overfitting or underfitting problems to a great extent. (3) Extensive set of experiments were performed to demonstrate that the proposed approach achieves better performance in terms of various evaluation measures [10], such as accuracy (weighted average and micro-average), precision, recall, F1 score, geometric mean, index balanced accuracy, in comparison to the state-of-the-art deep

TABLE I
DISTRIBUTION OF SAMPLES OF INTEL AND MOBILEODT CERVIGRAM DATASET.

	Type 1	Type 2	Type 3	Total
Train	249	781	450	1480
Additional Train	1189	3564	1976	6729
Test	87	265	160	512
Total	1525	4610	2586	8721

TABLE II
HYPERPARAMETER DESCRIPTION AND VALUES FOR TRAINING VGGIN-NET NETWORK.

Hyperparameter Description	Hyperparameter Values
Learning rate	0.0001
Optimizer	SGD with 0.9 momentum
Number of Epochs	300
Steps per Epoch	50
Batch Size	512
Image Size	299 x 224 using bilinear resizing 224 x 224 after random cropping
Data Augmentation	RandAugment with m=8, n=2
Network Input Size	224 x 224 x 3
Block4Pool Layer Output Size	14 x 14 x 512
Naïve Inception Block Layers	1x1 Conv2D - 64 filters
	3x3 Conv2D - 128 filters
	5x5 Conv2D - 32 filters
	3x3 MaxPooling2D
Dropout	0.4
Naïve Inception Block Output Size	14 x 14 x 736
Activation	ReLU (all Conv layers)
	Softmax (final dense layer)

learning architectures. Further, the entire paper is organized as follows: in section II, we have given the literature review of existing works done by various researchers using different deep learning techniques, primarily, for classifying different cervix types. In section IV, we have explained in detail the proposed methodology. In section III, we have described the dataset and necessary hyperparameters used in our experiments. A comparative analysis of the proposed approach with the various state-of-the-art approaches is presented in section V, and section VI includes the future works and conclusion.

II. LITERATURE REVIEW

Various researchers are using machine learning and deep learning in the biomedical domain [11] [12] [13] [14] [15]. Akter et al. (2021) used decision tree, random forest and XGBoost machine learning models to predict cervical cancer [16]. Kudva et al. (2020) used a shallow layer convolutional neural network (CNN) formulated with convolutional, fully connected layers, RELU, and pooling layers for the classification of cervical images into cancerous and non-cancerous categories [17]. Park et al. in 2021 proved that the ResNet deep learning model has shown an improvement in performance in comparison to other machine learning models such as XGB, SVM, and RF for the identification of cervical cancer [18]. Ghoneim et al. proposed a model based on CNN to extract the deep features, and then deep features were passed to the Extreme learning machine (ELM) for the classification of cervical cancer [19]. Li et al. (2019) used the Inception-V3 network to perform classification, and further applied fine-tuning, which

had an overall impact on the performance [20]. Yu et al. (2020) analyzed and compared hand-crafted features and deep learning methods on the multistate colposcopy image (MSCI) dataset. They proposed a gated recurrent convolutional neural network (C-GCNN) approach that includes time series along with the combined multistate cervical images [21]. Tripathi et al. in [22] proved that the ResNet-152 architecture gives higher accuracy as compared to other deep learning methods on the SIPAKMED Papsmear dataset. Guo et al. performed a comparative analysis of various models:- RetinaNet, VGG, and Inception; according to the analysis it was found that RetinaNet outperformed the other models [23]. Adem used softmax classification in [24] with a stacked autoencoder as one of the deep learning methods to classify the datasets. Various methods have been applied before to screen cervical cancer and detect different cervix types from the cervigram images as discussed, which briefly range from traditional machine learning and convolutional deep neural networks to various pre-trained deep neural networks along with the hybrid integration of features of deep learning and machine learning. The advantage of using pre-trained networks trained on the large-scale ImageNet dataset is the overall reduction of training time and adaptability to work well on small to medium datasets because of the knowledge transfer from robust large-scale ImageNet dataset.

III. EXPERIMENTAL SETTINGS

Intel and MobileODT had proposed the cervical cancer screening dataset which was made publicly available on Kaggle [31]. The dataset consists of image samples of different cervix types as shown in Fig. 2 which are grouped according to different cervix types. As observed in the figure, there are visual similarities between different types, rendering the problem to be a challenging task. This cervigram dataset consists of 3 classes (Type 1, Type 2 and Type 3). Type 1 refers to cervixes that are completely ectocervical, are fully visible, and may have small or large components. Type 2 cervixes have an endocervical component but are still fully visible and may or may not have an ectocervical component which may be small or large. In Type 3, there is an endocervical component present that is not fully visible and it may have an ectocervical component which can be small or large. Table I illustrates the imbalanced class distribution of image samples from the following classes: Type 1, Type 2 and Type 3; which is also depicted in Fig. 3 as a histogram of class populations. Images present were sourced from the train, additional train, and test splits. We combined the additional train directory with train and used it for training. This dataset has a very limited set of sample images. Images are artificially added on the fly at the mini-batch level in our training pipeline through data augmentation. Also, since the number of samples is less, training a CNN model from scratch would not make the model learn enough to classify into different types due to the limited set of sample images present in the dataset. Hence, data augmentation on transfer learned pre-trained models helps to make the model learn

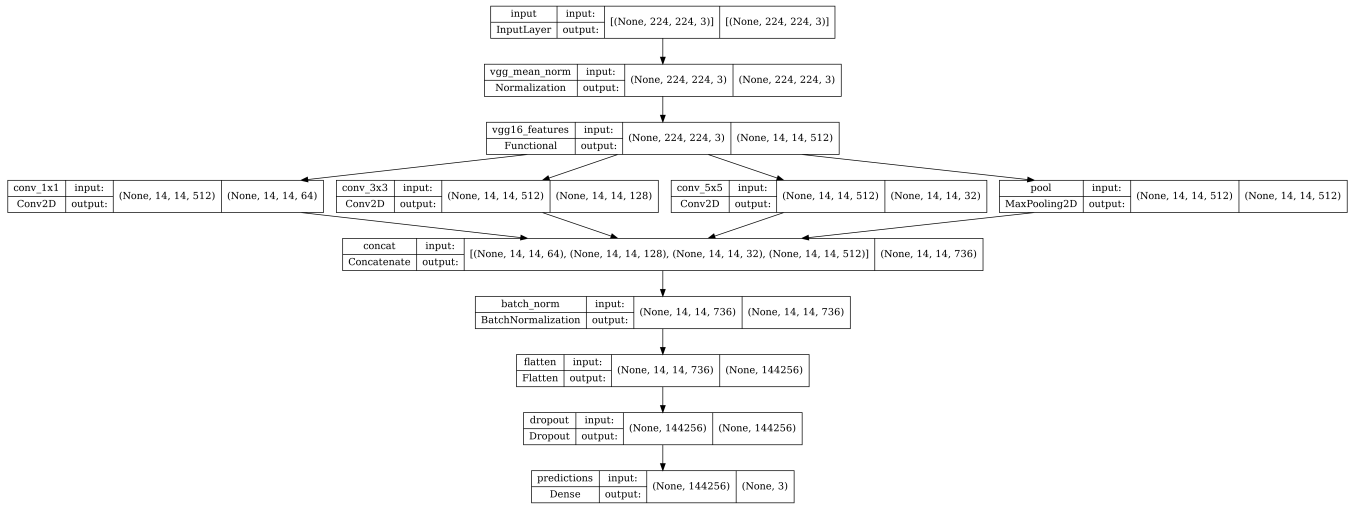


Fig. 1. The VGIN-Net architecture consisting of different layers in the network.

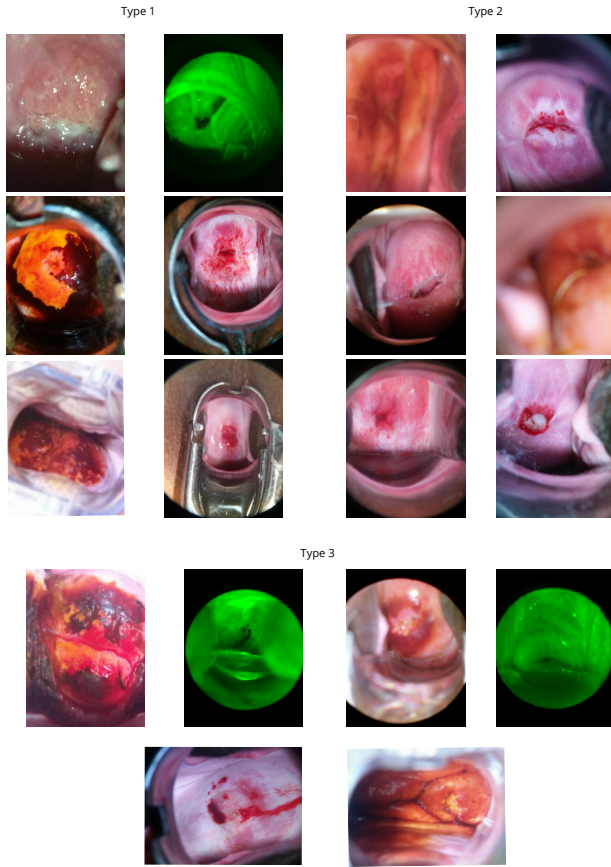


Fig. 2. Samples from Cervigram Dataset are grouped based on different Cervix types.

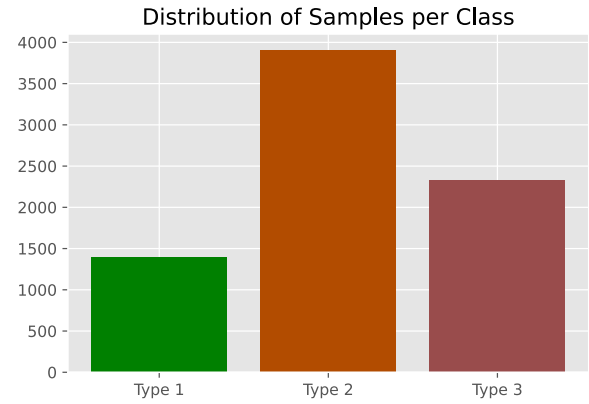


Fig. 3. Distribution of different classes (cervix types) from the cervigram dataset.

IV. METHODOLOGY

We have designed a novel approach to correctly classify the cervix type from cervigram images for a multi-class imbalanced dataset. We have extended our work [20], VGIN-Net for multi-class imbalanced datasets. Earlier we applied the proposed VGIN-Net model to the binary class problem on a Breast cancer dataset [25]. The VGIN-Net network architecture as shown in Fig.1 is based on the transfer learning approach, which is formed by freezing and concatenating all the layers till the block4 pool layer of the VGG16 pre-trained model [26] along with the naïve Inception block module [27]. Further, we have added the batch normalization [28], flatten, dropout, and dense layers [29] in the proposed architecture and constructed the 24-layer architecture by stacking the appropriate layers of VGG16 layers with the naïve Inception block and a few dense layers. In the proposed model, we use regularization in the form of dropout and data augmentation. We have used

features in a better way and generalize more suitably for the classification task [32].

TABLE III
COMPARISON OF THE PROPOSED VGGIN-NET BASED APPROACH WITH OTHER STATE-OF-THE-ART PRE-TRAINED MODELS USING IMBALANCE EVALUATION MEASURES.

Model		Precision				Recall				F1 Score				Index Balanced Accuracy				Geometric Mean			
Accuracy		Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg
VGGI6	0.71	0.59	0.73	0.76	0.72	0.63	0.77	0.66	0.71	0.61	0.75	0.71	0.71	0.56	0.54	0.59	0.56	0.76	0.73	0.77	0.75
InceptionV3	0.71	0.64	0.75	0.68	0.71	0.67	0.72	0.71	0.71	0.65	0.74	0.7	0.71	0.6	0.54	0.59	0.56	0.78	0.73	0.78	0.75
ResNet50V2	0.73	0.65	0.75	0.74	0.73	0.59	0.8	0.69	0.73	0.62	0.77	0.72	0.73	0.53	0.57	0.6	0.58	0.74	0.75	0.79	0.76
Xception	0.59	0.39	0.74	0.61	0.64	0.62	0.48	0.77	0.59	0.48	0.58	0.68	0.6	0.49	0.38	0.6	0.47	0.7	0.63	0.77	0.69
InceptionResNetV2	0.63	0.48	0.72	0.63	0.65	0.66	0.58	0.71	0.63	0.55	0.65	0.67	0.64	0.55	0.44	0.57	0.5	0.75	0.67	0.76	0.71
DenseNet121	0.72	0.61	0.74	0.75	0.72	0.62	0.78	0.67	0.72	0.62	0.76	0.71	0.72	0.55	0.55	0.59	0.56	0.74	0.78	0.75	0.75
EfficientNet-B0	0.56	0.37	0.66	0.62	0.6	0.63	0.49	0.64	0.56	0.47	0.56	0.63	0.57	0.49	0.35	0.51	0.42	0.7	0.6	0.72	0.65
VGGIN-Net	0.75	0.74	0.76	0.73	0.75	0.66	0.8	0.7	0.75	0.7	0.78	0.72	0.74	0.61	0.58	0.61	0.59	0.79	0.76	0.79	0.77

TABLE IV
COMPARISON OF THE PROPOSED VGGIN-NET BASED APPROACH WITH OTHER STATE-OF-THE-ART PRE-TRAINED MODELS WITHOUT APPLYING REJECTION RESAMPLING USING IMBALANCED EVALUATION MEASURES.

Model Accuracy	Precision				Recall				F1 Score				Index Balanced Accuracy				Geometric Mean				
	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	
VGGI6	0.71	0.65	0.74	0.72	0.71	0.68	0.77	0.64	0.71	0.66	0.75	0.68	0.71	0.61	0.55	0.56	0.56	0.79	0.74	0.75	0.75
InceptionV3	0.72	0.7	0.71	0.76	0.73	0.6	0.83	0.61	0.72	0.65	0.77	0.68	0.72	0.55	0.54	0.54	0.54	0.75	0.73	0.75	0.74
ResNet50V2	0.72	0.69	0.73	0.74	0.72	0.55	0.83	0.65	0.72	0.61	0.77	0.69	0.72	0.5	0.56	0.57	0.55	0.72	0.74	0.76	0.75
Xception	0.6	0.53	0.6	0.61	0.59	0.28	0.77	0.47	0.6	0.36	0.68	0.53	0.58	0.24	0.36	0.39	0.35	0.51	0.59	0.64	0.59
InceptionResNetV2	0.67	0.69	0.66	0.7	0.68	0.4	0.84	0.54	0.67	0.51	0.74	0.61	0.66	0.37	0.46	0.47	0.45	0.62	0.67	0.7	0.67
DenseNet121	0.74	0.76	0.72	0.8	0.75	0.59	0.87	0.62	0.74	0.66	0.79	0.7	0.74	0.54	0.57	0.56	0.56	0.75	0.74	0.76	0.75
EfficientNet-B0	0.66	0.57	0.65	0.69	0.65	0.33	0.82	0.57	0.66	0.42	0.73	0.63	0.64	0.3	0.45	0.49	0.44	0.56	0.66	0.71	0.66
VGGIN-Net	0.71	0.83	0.68	0.73	0.72	0.6	0.84	0.56	0.71	0.69	0.75	0.63	0.7	0.56	0.49	0.49	0.5	0.76	0.69	0.71	0.71

TABLE V
ABLATION EXPERIMENTS TO DETERMINE VERACITY OF THE PROPOSED VGGIN-NET BASED APPROACH.

Model		Precision				Recall				F1 Score				Index Balanced Accuracy				Geometric Mean			
Accuracy		Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg
Proposed Approach	0.75	0.74	0.76	0.73	0.75	0.66	0.8	0.7	0.75	0.7	0.78	0.72	0.74	0.61	0.58	0.61	0.59	0.79	0.76	0.79	0.77
Proposed Approach with multiple Inception blocks	0.75	0.76	0.76	0.72	0.75	0.67	0.8	0.72	0.75	0.71	0.78	0.72	0.75	0.62	0.59	0.62	0.6	0.8	0.76	0.79	0.78
Proposed Approach with Adam	0.72	0.67	0.73	0.73	0.72	0.71	0.78	0.62	0.72	0.69	0.76	0.67	0.72	0.65	0.65	0.54	0.56	0.81	0.74	0.75	0.75
Proposed Approach with SGDR	0.74	0.7	0.74	0.78	0.74	0.63	0.83	0.65	0.74	0.66	0.78	0.71	0.74	0.58	0.58	0.58	0.58	0.77	0.76	0.77	0.76

TABLE VI
EXPERIMENTS TO COMPARE THE PERFORMANCE OF DIFFERENT PRE-TRAINED NETWORKS WITH AND WITHOUT TRANSFER LEARNING.

Model Accuracy	Precision				Recall				F1 Score				Index Balanced Accuracy				Geometric Mean				
	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	Type 1	Type 2	Type 3	Avg	
VGGI6	0.40	0.30	0.0	0.47	0.2	0.71	0.0	0.89	0.4	0.42	0.0	0.61	0.26	0.47	0.0	0.5	0.24	0.68	0.0	0.7	0.33
(w/o Transfer Learning)																					
InceptionV3	0.39	0.28	0.0	0.5	0.2	0.8	0.0	0.82	0.39	0.42	0.0	0.62	0.26	0.48	0.0	0.52	0.24	0.68	0.0	0.72	0.33
(w/o Transfer Learning)																					
ResNet50V2	0.38	0.29	0.0	0.46	0.19	0.74	0.0	0.83	0.38	0.41	0.0	0.59	0.26	0.47	0.0	0.48	0.23	0.68	0.0	0.68	0.33
(w/o Transfer Learning)																					
EfficientNet-B0	0.36	0.26	0.0	0.46	0.19	0.77	0.0	0.72	0.36	0.38	0.0	0.57	0.24	0.43	0.0	0.45	0.21	0.65	0.0	0.67	0.32
(w/o Transfer Learning)																					
VGGIN-Net	0.39	0.35	0.0	0.41	0.19	0.69	0.0	0.88	0.39	0.46	0.0	0.56	0.25	0.51	0.0	0.4	0.21	0.71	0.0	0.56	0.25
(w/o Transfer Learning)																					
VGGI6	0.71	0.59	0.73	0.76	0.72	0.63	0.77	0.66	0.71	0.61	0.75	0.71	0.71	0.56	0.54	0.59	0.56	0.76	0.73	0.77	0.75
(w/ Transfer Learning)																					
InceptionV3	0.71	0.64	0.75	0.68	0.71	0.67	0.72	0.71	0.71	0.65	0.74	0.7	0.71	0.6	0.54	0.59	0.56	0.78	0.73	0.78	0.75
(w/ Transfer Learning)																					
ResNet50V2	0.73	0.65	0.75	0.74	0.73	0.59	0.8	0.69	0.73	0.62	0.77	0.72	0.73	0.53	0.57	0.6	0.58	0.74	0.75	0.79	0.76
(w/ Transfer Learning)																					
EfficientNet-B0	0.56	0.37	0.66	0.62	0.6	0.63	0.49	0.64	0.56	0.47	0.56	0.63	0.57	0.49	0.35	0.51	0.42	0.7	0.6	0.72	0.65
(w/ Transfer Learning)																					
VGGIN-Net	0.75	0.74	0.76	0.73	0.75	0.66	0.8	0.7	0.75	0.7	0.78	0.72	0.74	0.61	0.58	0.61	0.59	0.79	0.76	0.79	0.77
(w/ Transfer Learning)																					

the RandAugment [30] approach for data augmentation for our multi-class imbalanced task, which was found better as opposed to random combinations of flip, rotate, shift, and zoom which was used with this model on the binary classification task of BreakHis [25].

The VGGIN-Net architecture for the current work is obtained by freezing layers till the block 4 pool layer of the VGG16 model (pre-trained on ImageNet dataset) at the lower level and concatenating the layers of the naïve Inception module at the higher level which was initialized using Xavier uniform distribution. Other layers such as the dense layer along with batch normalization, flatten, and dropout layers were also added to make the network suitable for the classification

task. Xavier distribution was used to draw randomly initialized weights from a truncated normal distribution centred on zero mean and with a standard deviation equal to the square root of two divided by the sum of the number of input and output units in the layer. The naïve Inception block is composed of multiple convolutional layers with 1x1, 5x5, and 3x3 filters (64, 128, and 32 used as filters respectively) along with max pooling layer outputs which are concatenated. Batch Normalization normalizes data input values by standardizing them into a zero mean, unit variance distribution at a batch level which is also known to be very effective for tackling the vanishing gradient problem. Apart from batch normalization, the flatten layer added in the VGGIN-Net architecture concatenates all the features into a dimension of size 144256 which is followed

by Dropout regularization with a rate of 0.4. Dropout as a regularization technique helps to avoid overfitting by randomly dropping output values from a fraction of neurons during training.

All our experiments were conducted on Google Cloud TPU hardware [30], access to which was granted through the TensorFlow Research Cloud program. Each of our deep learning models was trained on the TPU v3-8 accelerator with the help of 128 GB high bandwidth memory and the TensorFlow Keras framework [33]. While training the models, we have considered 50 steps per epoch, the number of epochs as 300, and 512 as the batch size. Similar to the training setting described in [34] we set the learning rate to 0.0001 and allowed the models to train using the SGD algorithm with 0.9 as momentum. As a part of the data pipeline for model training, the training images were resized to 299 * 224, RandAugment [30] was applied with $m=8$ and $n=2$, and the images were further randomly cropped into images of size 224 * 224. During testing and evaluation, images are centrally cropped to 224 * 224 size. The exact set of hyperparameters is tabulated in Table II and the source code for the experiments is available in our GitHub repository.¹

V. RESULTS AND DISCUSSION

We have performed the comparative analysis between various pre-trained networks along with our proposed approach on the multi-class imbalanced dataset containing cervigram images for the cervical cancer screening task. The models used for comparison are pre-trained on the large-scale ImageNet dataset, and further fine-tuned on the cervix dataset using similar training settings as that of the VGGIN-Net model. Table III illustrates an analysis between our proposed VGGIN-Net model² and other state-of-the-art CNN architectures such as VGG16 [26], InceptionV3 [35], ResNet50 [36], ResNet50V2 [37], Xception [38], InceptionResNetV2 [39], DenseNet121 [40] and EfficientNet-B0 [41]. Performance metrics that are popularly used with imbalanced datasets such as Precision, Recall, F1 Score, Index Balanced Accuracy and Geometric Mean are used to determine which approach can tackle imbalance to a greater extent. Our analysis shows that among all the models VGGIN-Net gives significant performance improvement over other pre-trained models. Table IV, have shown the comparison of our proposed VGGIN-Net based approach with other state-of-the-art pre-trained models without applying the rejection resampling technique. It was observed that there is a significant drop in the results which proves that the application of rejection resampling is important for the given classification task. Further, we have conducted a study to show the efficacy of different hyperparameters chosen in our proposed approach, as shown in Table V. Additionally, we have shown the efficacy of transfer learning in our proposed approach by tabulating the results of different CNN models trained from scratch as shown in Table VI.

¹<https://github.com/SainiManisha/cervical-cancer-screening>

²<https://github.com/SainiManisha/vggin-net>

VI. CONCLUSION

Cervical cancer screening from a multi-class imbalanced dataset is a challenging task especially when the samples of minority classes are few in number. The situation is complicated due to visual similarities that exist between the Type 1, Type 2 and Type 3 stages of cancer. We have proposed a novel approach for cervical cancer screening using our recently introduced model VGGIN-Net, along with data augmentation and random crop and rejection resampling techniques to combat the challenges faced by multi-class imbalanced classification tasks. We have done comparative analysis of various pre-trained networks on the cervical cancer dataset based on various evaluation metrics such as accuracy, precision, recall, F1 score, geometric mean and index balanced accuracy. The experimental results show that the proposed approach is demonstrating better results as compared to the state-of-the-art pre-trained models. In future work, we shall create different deep learning architectures for multi-class imbalanced datasets for object detection and segmentation by exploring the effect and usage of focal loss to handle the class imbalance problem.

REFERENCES

- [1] T. Dong, C. Yang, B. Cui, T. Zhang, X. Sun, K. Song, L. Wang, B. Kong, and X. Yang, "Development and validation of a deep learning radiomics model predicting lymph node status in operable cervical cancer," *Frontiers in Oncology*, vol. 10, p. 464, 2020.
- [2] T. Zhang, J. Chen, F. Li, K. Zhang, H. Lv, S. He, and E. Xu, "Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions," *ISA transactions*, vol. 119, pp. 152–171, 2022.
- [3] S. Susan and A. Kumar, "Hybrid of intelligent minority oversampling and pso-based intelligent majority undersampling for learning from imbalanced datasets," in *International Conference on Intelligent Systems Design and Applications*. Springer, 2018, pp. 760–769.
- [4] —, "The balancing trick: Optimized sampling of imbalanced datasets—a brief survey of the recent state of the art," *Engineering Reports*, vol. 3, no. 4, p. e12298, 2021.
- [5] M. Saini and S. Susan, "Deep transfer with minority data augmentation for imbalanced breast cancer dataset," *Applied Soft Computing*, vol. 97, p. 106759, 2020.
- [6] —, "Bag-of-visual-words codebook generation using deep features for effective classification of imbalanced multi-class image datasets," *Multimedia Tools and Applications*, vol. 80, no. 14, pp. 20821–20847, 2021.
- [7] —, "Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-v3," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2019, pp. 409–420.
- [8] K. Matsuo, S. Purushotham, B. Jiang, R. S. Mandelbaum, T. Takiuchi, Y. Liu, and L. D. Roman, "Survival outcome prediction in cervical cancer: Cox models vs deep-learning model," *American journal of obstetrics and gynecology*, vol. 220, no. 4, pp. 381–e1, 2019.
- [9] M. Saini and S. Susan, "Vggin-net: Deep transfer network for imbalanced breast cancer dataset," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.
- [10] L. Pereira and N. Nunes, "Performance evaluation in non-intrusive load monitoring: Datasets, metrics, and tools—a review," *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, vol. 8, no. 6, p. e1265, 2018.
- [11] M. M. Badža and M. Č. Barjaktarović, "Classification of brain tumors from mri images using a convolutional neural network," *Applied Sciences*, vol. 10, no. 6, p. 1999, 2020.
- [12] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.
- [13] S. S. Yadav and S. M. Jadhav, "Deep convolutional neural network based medical image classification for disease diagnosis," *Journal of Big Data*, vol. 6, no. 1, pp. 1–18, 2019.

- [14] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyás, "3d deep learning on medical images: a review," *Sensors*, vol. 20, no. 18, p. 5097, 2020.
- [15] S. P. Singh, L. Wang, S. Gupta, B. Gulyas, and P. Padmanabhan, "Shallow 3d cnn for detecting acute brain hemorrhage from medical imaging sensors," *IEEE Sensors Journal*, vol. 21, no. 13, pp. 14 290–14 299, 2020.
- [16] L. Akter, M. Islam, M. S. Al-Rakhami, M. Haque *et al.*, "Prediction of cervical cancer from behavior risk using machine learning techniques," *SN Computer Science*, vol. 2, no. 3, pp. 1–10, 2021.
- [17] V. Kudva, K. Prasad, and S. Gurusware, "Automation of detection of cervical cancer using convolutional neural networks," *Critical Reviews™ in Biomedical Engineering*, vol. 46, no. 2, 2018.
- [18] Y. R. Park, Y. J. Kim, W. Ju, K. Nam, S. Kim, and K. G. Kim, "Comparison of machine and deep learning for the classification of cervical cancer based on cervicography images," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.
- [19] A. Ghoneim, G. Muhammad, and M. S. Hossain, "Cervical cancer classification using convolutional neural networks and extreme learning machines," *Future Generation Computer Systems*, vol. 102, pp. 643–649, 2020.
- [20] C. Li, D. Xue, X. Zhou, J. Zhang, H. Zhang, Y. Yao, F. Kong, L. Zhang, and H. Sun, "Transfer learning based classification of cervical cancer immunohistochemistry images," in *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*, 2019, pp. 102–106.
- [21] Y. Yu, J. Ma, W. Zhao, Z. Li, and S. Ding, "Msci: A multistate dataset for colposcopy image classification of cervical cancer screening," *International Journal of Medical Informatics*, vol. 146, p. 104352, 2021.
- [22] A. Tripathi, A. Arora, and A. Bhan, "Classification of cervical cancer using deep learning algorithm," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2021, pp. 1210–1218.
- [23] P. Guo, S. Singh, Z. Xue, R. Long, and S. Antani, "Deep learning for assessing image focus for automated cervical cancer screening," in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 2019, pp. 1–4.
- [24] K. Adem, S. Kilicarslan, and O. Comert, "Classification and diagnosis of cervical cancer with softmax classification with stacked autoencoder," 2019.
- [25] "Breast cancer histopathological database (breckhis) - laboratório visão robótica e imagem," Oct 2019. [Online]. Available: <https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/>
- [26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
- [29] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.
- [30] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "RandAugment: Practical automated data augmentation with a reduced search space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 702–703.
- [31] I. MobileODT, "Intel and mobileodt cervical cancer screening." [Online]. Available: <https://www.kaggle.com/c/intel-mobileodt-cervical-cancer-screening/data>
- [32] I. Z. Mukti and D. Biswas, "Transfer learning based plant diseases detection using resnet50," in *2019 4th International conference on electrical information and communication technology (EICT)*. IEEE, 2019, pp. 1–6.
- [33] G. Zaccane, M. R. Karim, and A. Menshawy, *Deep learning with TensorFlow*. Packt Publishing Ltd, 2017.
- [34] C. A. Ferreira, T. Melo, P. Sousa, M. I. Meyer, E. Shakibapour, P. Costa, and A. Campilho, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," in *International conference image analysis and recognition*. Springer, 2018, pp. 763–770.
- [35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [37] —, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.
- [38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
- [39] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [40] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.
- [41] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.



Characterization, utility, and interrelationship of household organic waste generation in academic campus for the production of biogas and compost: a case study

Pradeep Kumar Meena¹ · Amit Pal¹ · Samsher²

Received: 22 August 2022 / Accepted: 23 October 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

This case study presents the experimental outcomes of a zero organic waste campus. Three hundred fifty families live on this academic campus, whose solid wastes from their homes were not being properly segregated. For easy segregation, 750 dustbins of two types (labeled as organic and inorganic waste) were distributed in the canteen, mess, and all residential apartments, holding capacity up to 13 kg of garbage. A total of 24 sample sets of household organic waste were studied in 12 months period with a sample size of 1620 waste bags. Seventy-three types of organic waste were found to be 518.53, 263.57, and 249.94 kg, respectively, in the form of raw vegetable waste (RVW), fruit waste (FW), and mixed cooked waste (MCW). Regression method is applied, and the result suggests that the coefficient of determination (R^2) of these variables (RVW, FW), (RVW, MCW), and (FW, MCW) was observed to be 0.90, 0.91, and 0.94, respectively, with $p < 0.05$. Compostable and digestible wastes demonstrate the great potential to generate compost and biogas, and each compostable waste contains nutrients such as nitrogen (N), phosphorus (P), and potassium (K), which are essential for the overall growth of plants. Our present study demonstrated that one ton of digested waste collected from the academic campus and transferred in the biogas plant could generate 50 m³ of biogas/day, which can produce 160 kW of green electric energy and 200 kg of organic compost.

Keywords Solid waste · Organic waste · Zero organic waste · Compost · Biogas

Abbreviations

RVW	Raw vegetable waste
FW	Fruit waste
MCW	Mixed cooked waste
MSW	Municipal solid waste
R^2	Coefficient of determination
N	Nitrogen

✉ Pradeep Kumar Meena
pradeep_2k18phdme08@dtu.ac.in

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

² Harcourt Butler Technical University, Kanpur, India

P	Phosphorus
K	Potassium
Sample Sets (Sample Size)	S1(10) S2(15), S3(20), S4(25), S5(30), S6(35), S7(40), S8(45), S9(50), S10(55), S11(60), S12(65), S13(70), S14(75), S15(80), S16(85), S17(90), S18(95), S19(100), S20(105), S21(110), S22(115), S23(120), and S24(125)

1 Introduction

World generates approximately 2.01 billion metric tons of municipal solid waste (MSW) every year (Chen et al., 2016). Therefore, management of the proper disposal is of utmost importance. Organic wastes may be valorized by the use of cellulose fiber-rich wastes to produce board, binder-less board, paper, or to convert these organic wastes to clean fuels and/or petrochemical substitutes via pyrolysis (Fahmy et al., 2017), (Fahmy et al., 2020), (Fardous Mobarak et al., 1982a, 1982b), (Fahmy & Mobarak, 2013), (Fahmy et al., 1982a, b). Organic wastes may be also converted chemically—by hydrolysis—to sugars, which may be fermented to give bioethanol (Fardous Mobarak, 1983), (Mobarak et al., 1982a, 1982b), (Fahmy, 1982a, b), (Fahmy & El-Shinnawy, 1975), (El-Shinnawy et al., 1983).

Otherwise, the hazardous chemical released by waste can leak and reach our food through the soil. In addition, the burning of garbage at landfills also produces toxic gases into the air, including dioxin. Due to the lack of proper disposal practice of MSW and the emission of harmful gases, MSW is a significant contributor to global warming (Rubio-Romero et al., 2013). World population is expected to increase by 2 billion people in the next 30 years, crossing 9 billion in 2050. Growing population leads to an increase in the volume of MSW, which poses serious environmental challenges, including underground water and pollution (Vijayan & Parthiban, 2020). MSW problem is alarming for modern societies and developing countries because of environmental challenges associated with waste production, poor waste assortment, transport, unscientific treatment, and proper disposal (Ikhlayel, 2018).

Due to this, there is uncontrolled growth in MSW, and the lack of proper logistics in developing countries like India has become a severe and significant environmental problem. India is a large country with 1.3 billion people and produces millions of solid wastes from households (Pradeep Kumar Meena, Sumit Sharma 2022). Due to this, there is uncontrolled growth in MSW, and the lack of proper logistics in developing countries like India has become a severe and significant environmental problem (Song et al., 2015). Therefore, it is essential to improve MSW management for the people's health safety and sustainable environment (Chaturvedi et al., 2018). An estimated 70% of solid waste contains organic waste such as food waste, kitchen waste, green waste, and sewage sludge which can be recycled (Zhang et al., 2020). Due to lack of knowledge, technology, and resources, most developing countries, including India, cannot use organic wastes to their full potential. Organic solid debris is present in large quantities, mainly in-residence apartments, vegetable markets, and at food outlets (Forster-Carneiro et al., 2008).

Biogas is a renewable fuel energy source that contains methane, carbon dioxide, and other trace compounds produced from the anaerobic digestion of organic feed stocks, including solid organic debris available from household waste. Generated biogas can be used for various applications such as cooking food, domestic heating purposes, lighting fuel, running IC engines, and vehicle fuel. Composting is an aerobic process that requires

oxygen, optimal moisture content, and porosity to stabilize the organic wastes. Therefore, the generated composite from organic waste can be utilized for various applications, including improving soil structure, water-holding capacity, infiltration rate—moreover, composite increases microbial activity in the soil and the diversity of microorganisms. Therefore, the production of biogas and compost from organic waste could be an alternate energy source due to its lower production cost and renewable nature. Furthermore, food residual is suitable for composting and biogas combined with the green waste of the campus (Mason et al., 2004). Thousands of students and staff live on academic campuses with their families, whose homes, messes, and canteens produce large amounts of organic and inorganic waste every day, which can be recycled. University is a place where people work in research with new ideas to develop new technology necessary for their progress (Kelly et al., 2006).

It is the necessity of the time that university campuses must initiate a program and periodically educate people on proper segregation and disposal of organic and inorganic waste generated from the campuses. Problem can be solved by studying how much waste is produced daily, directly or indirectly, on campus with teachers and administrative staff's help involving students (Smyth et al., 2010). Design of university waste management systems was initiated in industrialized countries about 20 years ago (Armijo de Vega et al., 2003). For example, Massey University (New Zealand) conducted a study on adequately implementing the university campus's zero waste program (Mason et al., 2003). Solid waste management systems can be successful by using reliable data, accurate planning tools, and design (Ogwueleka, 2013). Recycling program is one of the most popular methods in the USA, with approximately 80% of schools and universities participating in an institutional program (Gallardo et al., 2016). Universities in India are also like a society where thousands of students attend daily, dealing with various activities that directly and indirectly generate a significant amount of waste (Lukman et al., 2009). However, little information is available on how zero organic waste systems can be implemented on university campuses or large city societies in India and other countries. In 2019, Delhi Technological University (DTU) installed a biogas plant on campus for biogas production. In the present study, a waste management plan is designed for proper management and utilization of waste generated from the canteen and mess of the University. However, due to improper segregation, the amount of organic waste utilized was 30%–40% less because about 300–400 kg was not adequately segregated. Therefore, to make DTU a zero organic waste campus, two bags labeled with organic and inorganic waste are distributed to all households on the University campus with the help of volunteers. As a result, the segregated organic waste was utilized for optimal biogas production and the slurry (compost) residual part (used as fertilizer). India is the most populous country and has more than 1000 universities, where the proposed waste management system in the manuscript can be implemented for proper disposal and management (Fig. 1).

2 Materials and methods

Delhi Technological University, at latitude 28.7496°N, and longitude 77.1174°E, was established in 1941 as Delhi Polytechnic by the Government of India. This University has eight boys hostels, six girls hostels, and about 350 residential apartments for its employees (DTU, 2009). University has 9045 students till 2018–19; undergraduate students 7170,

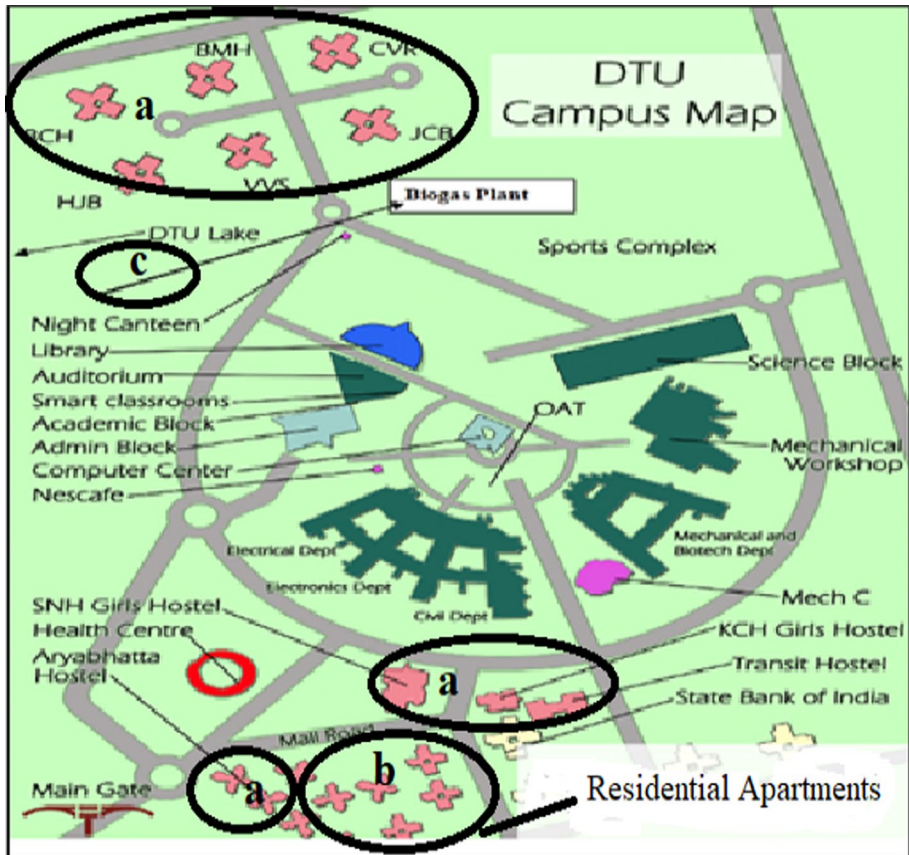


Fig. 1 Location of **a** boys and girls hostel, **b** residential apartments, **c** biogas plant in Delhi Technological University Campus (DTU, 2014)

postgraduates 898, doctoral students 395 and around 1000–1200 teaching and non-teaching staff (University, 2019).

First, to initiate the zero organic waste processes, approximately 750 dustbins with a capacity of 13 kg were distributed to all residential apartments, canteens of the university campus, which can be seen in Fig. 2a. There were two types of dustbins, one in blue color in which dry waste is collected and the other in green in which wet organic waste is collected. Solid organic waste generated from residential apartments, canteens, and campuses through these dustbins is collected daily at one location. After that, the accumulated solid organic waste is separated as digestive organic waste and compost organic waste. Digestive organic waste is separated from approximately 1 ton of organic waste per day and introduced into a biogas digester, producing biogas.

Some studies have found that the methane (CH_4) concentration in potato, wet sugarcane bagasse, semi-dry sugarcane bagasse, and sugarcane waste was 35.64%, 96.06%, 91.39%, and 58.74%, respectively (Romero et al., 2020). Biogas can be produced from different waste such as cow dung with vegetable, fruit waste, and their mixture (Vats et al., 2019). More than 90% of CH_4 gas is obtained in the biogas purification process by removing CO_2 and H_2S from the raw biogas. A Genset of 15kVA is powered by biogas



Fig. 2 a Dustbin distribution on the university campus. b Biogas plant in DTU campus, c Compost made from biogas slurry and compost made from pits

generated from digestive waste. It operates sixteen mercury lights of 300 W, eight fans of 50 W, and a biogas plant, as shown in Fig. 2b. Biogas is not only used in power, engine running, and cooking, but it also contributes to reducing environmental pollution arising from MSW as 400 kg (CH_4) is equivalent to about 1000 kg (CO_2), which is used in biogas (Shane & Gheewala, 2017). After the biogas production, the slurry from the biogas plant is used to make compost from the composting machine. Remaining composted organic waste, which is not used for biogas production, is dumped in the pit and composted in 30–60 days, as shown in Fig. 2c. Later, this fertilizer is dried and used in the garden of the university campus. About 200 kg of manure is produced daily from this biogas plant used in the campus garden spread over 164 acres, and farmers and nurseries nearby also use this fertilizer. It means that 100% of the organic waste

generated on campus is used, and the campus has been converted to a zero organic waste system.

2.1 Sample collections and waste profile

Every day, many people, including students, employees, and visitors, enter the campus and eat in a mess, canteens, and cafeteria. In addition, thousands of students live in hostels for whom food is prepared four times a day in the mess and canteen. Moreover, more than three hundred flats are available where staff resides based on their grade pay. Most professors and higher officials reside in the university campus with an average family of 4–5 members. Organic wastes such as tea powder, leftover food, fruit waste, and vegetable waste are routinely used for biogas production; hence, it is called digestive organic wastes. Spinach sticks, cabbage, green leaves, dry leaves, orange peel, fibrous vegetables, fruits, etc., are used to make compost; hence it is called organic compost waste. A total of 24 sample sets were collected in a span of 12 months; thus, each 8-sample set was collected with an interval of 4 months. Sample size of each sample set is studied by incrementing five garbage bags. As shown in Fig. 3, more than 1000 kg of organic waste is generated from canteens, messes, and residential apartments every day. While weighing, $\pm 10\%$ accuracy has been taken for each household organic waste, shown in Table 1.

Sample sets S1 to S8 were collected between January 2020 and April 2020. In this period, the total sample size was 220 waste bags, in which a total of 90.94 kg of RVW, 32.67 kg of FW, and 25.74 kg MCW were found, as shown in Table 2. During this period, 36 types of organic waste were found as RVW, FW, and MCW in S1 to S8 sample sets. S1, S2, and S3 were collected in January with sample sizes of 10, 15, and 20 garbage bags, respectively. In total, 45 sample sizes, 16.95 kg of RVW, 6.51 kg of FW, and 5.53 kg of MCW were obtained, and the highest waste content among these samples: cabbage, green fenugreek, and tea leaves, as shown in Fig. 4a and Table 1(a). Total sample sizes of S4, S5 in February were 55 garbage bags in which RVW, FW, and MCW were obtained to be 22.71 kg, 3.42 kg, and 2.86 kg, respectively, and the maximum amount of spinach and pea peel waste was observed in these samples.

Similarly, the S6, S7, and S8 were interpreted in the order of March and April with 35, 40, and 45 sample sizes. Total sample size was 120 garbage bags, and RVW, FW, and MCW were obtained to be 51.28 kg, 22.74 kg, and 17.35 kg, respectively. A large quantity of cabbage, mango, and cooked mixed waste content was received in S6, S7, and S8 sample sets, respectively. Collection of sample sets S9 to S16 was performed between May and August 2020. The smallest sample size was 50 garbage bags, which was S9, and the largest sample size was 85 garbage bag collections, which belong to S16. Thus, the total sample size in this period was 540 garbage bags containing 186.96 kg of RVW, 94.65 kg of FW, and 90.85 kg of MCW was found. During this time, 39 types of organic waste were found as RVW, FW, and MCW in S9 to S16. Collections of S9 and S10 taken in May, with sample sizes of 50 and 55, respectively, with a total sample size of 105, yielded 40.5 kg of RVW, 19.75 kg of FW, and 18.5 kg of MCW, with the highest number of mixed debris, potato, and tomato waste were found. S11, S12 are taken in June with sample sizes of 60 and 65. Out of 125 sample sizes, 46.8 kg of RVW, 22.8 kg of FW, 21.4 kg of MCW were found. Table 1(b) and Fig. 4b show that the highest number of mixed debris, tea leaves, watermelon, mango, and cucumber waste were found in these samples. S13 and S14 were reviewed in July, with 145 sample sizes containing 48 kg of RVW, 27.5 kg of FW, 26.05 kg of MCW, respectively, and the highest amount of waste, mixed waste in these samples, tea

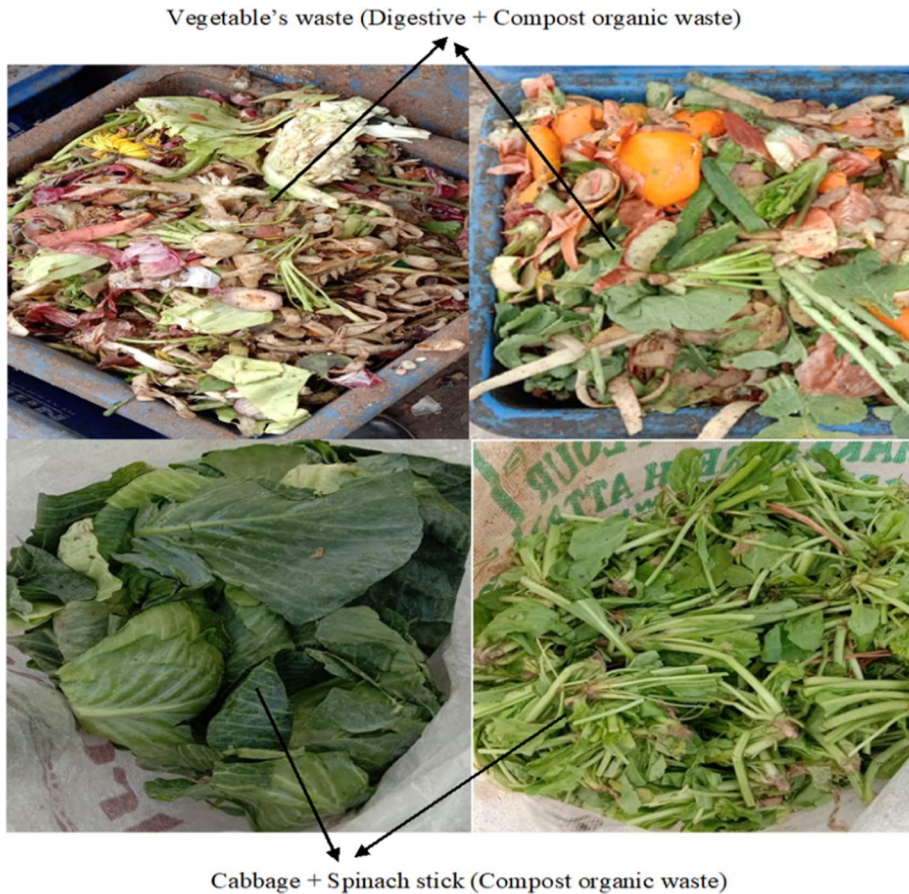


Fig. 3 Various types of vegetable wastes, fruit wastes, and mixed wastes are generated daily on the university campus

leaves, watermelon, and garlic was found. Similarly, S15 and S16 were reviewed in August, with a total sample size of 165 waste bags containing 51.66 kg of RVW, 24.6 kg of FW, 24.9 kg of MCW, and the most mixed waste, tea leaves, waste of onion, mango and potato were found. Sample sets S17 to S24 were studied from September 2020 to December 2020, with the smallest sample size being 90 and the largest sample size being 125 garbage bags. Total sample size in this period was 860 waste bags, of which 240.63 kg of RVW, 136.25 kg of FW, and 133.35 kg of MCW were found. During this, 55 types of organic wastes were found in RVW, FW, and MCW. S17 and S18 have 90 and 95 waste bags collected in September, with a total sample size of 185 consisting of 53.2 kg of RVW, 32.4 kg of FW, and 29 kg of MCW. Other mixed scraps, potato, banana, tea leaf, and lemon waste were highest in these samples, shown in Table 1(c) and Fig. 4c.

Similarly, S19 and S20 were studied in October, with a total sample size of 205 waste bags containing 55.32 kg of RVW, 30.9 kg of FW, and 30.1 kg of MCW. In which mixed waste, orange, rice, beet, and lime were found to have the highest waste material. Collections S21 and S22 were taken in November with sample sizes of 110

Table 1 Sample-set (S1 to S24) contains 1620 waste bags containing various household organic waste collected between January 2020 and December 2020

Types of household Organic waste		(a). S1 to S8 with a total sample size of 220 garbage bags collected between January 2020 and April 2020									
		S1(10)	S2(15)	S3(20)	S4(25)	S5(30)	S6(35)	S7(40)	S8(45)		
		(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)		(kg)
Beetroot		—	—	0.53 ± 0.053	—	—	—	—	—	—	—
Broccoli		—	—	—	—	—	—	1.95 ± 0.195	—	—	—
Chard		0.26 ± 0.026	—	—	—	—	—	—	—	—	—
Cauliflower		—	—	1.48 ± 0.148	—	0.2 ± 0.02	—	—	—	—	—
Zucchini		—	—	—	—	—	—	1.15 ± 0.115	—	—	—
Green fenugreek		0.46 ± 0.046	1.73 ± 0.173	—	1.35 ± 0.125	1.62 ± 0.162	—	—	—	0.8 ± 0.08	—
Spinach		1.1 ± 0.11	1.3 ± 0.13	1.35 ± 0.135	2.38 ± 0.212	1.42 ± 0.142	3.46 ± 0.346	1.05 ± 0.105	—	1.25 ± 0.125	—
Pumpkin		0.5 ± 0.05	0.7 ± 0.07	—	—	—	—	—	—	0.4 ± 0.04	—
Onion		—	0.42 ± 0.042	0.60 ± 0.060	—	0.61 ± 0.061	2.73 ± 0.273	1.19 ± 0.119	—	1.35 ± 0.135	—
Potato		0.52 ± 0.052	—	0.31 ± 0.031	1.44 ± 0.134	—	0.33 ± 0.033	0.65 ± 0.065	—	1.3 ± 0.13	—
Carrot		—	0.21 ± 0.021	0.25 ± 0.025	1.12 ± 0.102	1.53 ± 0.153	0.76 ± 0.076	1.85 ± 0.185	—	0.65 ± 0.065	—
Ladyfinger		—	0.3 ± 0.03	—	—	—	—	—	—	—	—
Green coriander		—	1.19 ± 0.119	0.16 ± 0.016	—	—	0.49 ± 0.049	1.75 ± 0.175	—	0.95 ± 0.095	—
Cabbage		1.15 ± 0.115	—	—	0.65 ± 0.065	—	4.31 ± 0.431	—	—	1.9 ± 0.19	—
Radish		0.4 ± 0.04	—	—	0.9 ± 0.09	1.29 ± 0.129	3.1 ± 0.31	2.07 ± 0.207	—	2.91 ± 0.291	—
Tomatoes		0.16 ± 0.016	—	0.78 ± 0.078	0.3 ± 0.02	—	0.2 ± 0.02	—	—	0.45 ± 0.045	—
Turnip		—	—	0.25 ± 0.025	0.2 ± 0.02	—	—	0.56 ± 0.056	—	—	—
Capsicum		—	—	0.39 ± 0.039	—	—	0.3 ± 0.03	—	—	0.85 ± 0.085	—
Bottle Gourd		—	—	0.15 ± 0.015	0.21 ± 0.021	—	—	—	—	1.75 ± 0.175	—
Pea peel		—	—	—	1.96 ± 0.196	5.53 ± 0.553	1.72 ± 0.172	4.12 ± 0.412	—	2.98 ± 0.298	—
Cucumber		—	—	0.3 ± 0.03	—	—	—	—	—	—	—
Banana		0.7 ± 0.07	0.45 ± 0.045	0.83 ± 0.083	0.7 ± 0.07	0.2 ± 0.02	2.34 ± 0.234	2.36 ± 0.236	—	2.5 ± 0.25	—
Orange		—	0.53 ± 0.053	0.54 ± 0.054	0.4 ± 0.04	—	—	—	—	—	—
Papaya		0.61 ± 0.061	—	0.5 ± 0.05	0.45 ± 0.045	—	—	—	—	1.48 ± 0.148	—
Apple		0.29 ± 0.029	—	0.14 ± 0.014	0.32 ± 0.032	0.25 ± 0.025	—	0.5 ± 0.05	—	0.9 ± 0.09	—
Pineapple		—	—	—	—	0.83 ± 0.083	—	—	—	0.55 ± 0.055	—

Table 1 (continued)

(a). S1 to S8 with a total sample size of 220 garbage bags collected between January 2020 and April 2020									
Types of household Organic waste	S1(10) (kg)	S2(15) (kg)	S3(20) (kg)	S4(25) (kg)	S5(30) (kg)	S6(35) (kg)	S7(40) (kg)	S8(45) (kg)	
Pomegranate	–	–	–	–	–	1.58 ± 0.158	0.7 ± 0.07	2.17 ± 0.217	
Plum	–	–	0.7 ± 0.07	–	–	0.37 ± 0.037	–	0.3 ± 0.03	
Kumquats	–	0.57 ± 0.057	–	–	–	–	–	0.55 ± 0.055	
Kiwi	0.25 ± 0.025	–	–	–	–	–	0.76 ± 0.076	–	
Mango	–	–	–	–	–	–	4.43 ± 0.443	1.25 ± 0.125	
Limes	–	–	0.4 ± 0.04	0.20 ± 0.020	0.27 ± 0.027	–	–	–	
Tea leaves	0.66 ± 0.066	–	1.85 ± 0.185	1.11 ± 0.111	0.52 ± 0.052	2.39 ± 0.239	0.86 ± 0.086	2.8 ± 0.28	
Bread	0.45 ± 0.045	–	–	–	–	–	0.8 ± 0.08	0.5 ± 0.05	
Rice	0.4 ± 0.04	–	–	0.4 ± 0.04	–	–	0.55 ± 0.055	0.6 ± 0.06	
Mixed waste	–	1.22 ± 0.122	.95 ± 0.095	–	0.63 ± 0.063	0.92 ± 0.092	4.20 ± 0.420	3.73 ± 0.373	
(b). S9 to S16 with a total sample size of 540 garbage bags collected between May 2020 to August 2020									
Types of household Organic waste	S9(50) (kg)	S10(55) (kg)	S11(60) (kg)	S12(65) (kg)	S13(70) (kg)	S14(75) (kg)	S15(80) (kg)	S16(85) (kg)	
Garlic	1.30 ± 0.130	0.75 ± 0.075	–	1.35 ± 0.135	–	3.5 ± 0.35	1.65 ± 0.165	2.3 ± 0.23	
Asparagus	0.7 ± 0.07	1.35 ± 0.135	0.65 ± 0.065	–	0.85 ± 0.085	–	0.45 ± 0.045	1.45 ± 0.145	
Chard	–	–	1.85 ± 0.185	1.45 ± 0.145	0.4 ± 0.04	0.55 ± 0.055	1.5 ± 0.15	0.35 ± 0.035	
Cucumber	1.75 ± 0.175	–	–	3.35 ± 0.335	–	0.65 ± 0.065	1.85 ± 0.185	0.71 ± 0.071	
Chicory	0.68 ± 0.068	–	1.55 ± 0.155	1.75 ± 0.175	0.5 ± 0.05	–	2.7 ± 0.27	0.95 ± 0.095	
Green bean	0.20 ± 0.020	1.05 ± 0.105	1.25 ± 0.125	0.95 ± 0.095	1.40 ± 0.140	–	1.45 ± 0.145	2.65 ± 0.265	
Broad bean	0.35 ± 0.035	–	0.25 ± 0.025	–	0.3 ± 0.03	1.75 ± 0.175	0.35 ± 0.035	0.4 ± 0.04	
Lettuce	0.25 ± 0.025	1.23 ± 0.123	1.45 ± 0.145	1.65 ± 0.165	–	1.5 ± 0.15	0.35 ± 0.035	0.55 ± 0.055	
Tomato	0.5 ± 0.05	3.5 ± 0.35	2.25 ± 0.225	–	1.50 ± 0.150	–	0.55 ± 0.055	0.6 ± 0.06	

Table 1 (continued)

Types of household Organic waste		(b). S9 to S16 with a total sample size of 540 garbage bags collected between May 2020 to August 2020									
		S9(50)	S10(55)	S11(60)	S12(65)	S13(70)	S14(75)	S15(80)	S16(85)		
		(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)		(kg)
Pea		1.95 ± 0.195	—	—	2.75 ± 0.275	1.70 ± 0.170	1.65 ± 0.165	2.15 ± 0.215	1.5 ± 0.15		
Rhubarb		0.25 ± 0.025	0.85 ± 0.085	0.45 ± 0.045	—	0.5 ± 0.05	0.45 ± 0.045	—	—		
Radish		0.55 ± 0.055	1.55 ± 0.155	0.65 ± 0.065	0.45 ± 0.045	—	1.5 ± 0.15	—	1.8 ± 0.18		
Celery		0.85 ± 0.085	1.15 ± 0.115	0.85 ± 0.085	0.3 ± 0.03	3.30 ± 0.33	1.85 ± 0.185	0.65 ± 0.065	—		
Courgette		1.15 ± 0.115	—	1.45 ± 0.145	1.85 ± 0.185	—	2.7 ± 0.27	1.1 ± 0.11	—		
Potatoes		3.05 ± 0.305	2.47 ± 0.247	1.25 ± 0.125	2.7 ± 0.27	2.60 ± 0.260	1.45 ± 0.145	1.2 ± 0.12	4.5 ± 0.45		
Bottle gourd		2.67 ± 0.267	—	0.25 ± 0.025	0.3 ± 0.03	—	0.35 ± 0.035	—	1.65 ± 0.165		
Onion		2.25 ± 0.225	—	0.23 ± 0.023	0.4 ± 0.04	1.9 ± 0.19	0.35 ± 0.035	3.5 ± 0.35	0.45 ± 0.045		
Ladyfinger		0.85 ± 0.085	0.5 ± 0.05	0.27 ± 0.027	0.55 ± 0.055	—	0.50 ± 0.050	0.5 ± 0.05	1.5 ± 0.15		
Capsicum		—	0.35 ± 0.035	—	2.65 ± 0.265	0.35 ± 0.035	2.65 ± 0.265	0.4 ± 0.04	1.60 ± 0.160		
Zucchini		—	1.65 ± 0.165	—	0.35 ± 0.035	2.6 ± 0.26	0.4 ± 0.04	0.75 ± 0.075	0.85 ± 0.085		
Arugula		—	0.85 ± 0.085	2.70 ± 0.270	—	2.85 ± 0.285	—	0.7 ± 0.07	—		
Brinjal		—	2.15 ± 0.215	1.65 ± 0.165	0.65 ± 0.065	0.8 ± 0.08	0.65 ± 0.065	1.4 ± 0.14	—		
Sen		—	1.35 ± 0.135	3.5 ± 0.35	—	0.45 ± 0.045	1.7 ± 0.17	1.55 ± 0.155	1.5 ± 0.15		
Jackfruit		—	0.45 ± 0.045	—	0.85 ± 0.085	0.65 ± 0.065	1.2 ± 0.12	—	1.6 ± 0.16		
Watermelon		2.15 ± 0.215	2.35 ± 0.235	2.75 ± 0.275	2.60 ± 0.285	3.8 ± 0.38	2.5 ± 0.25	2.7 ± 0.27	1.65 ± 0.165		
Banana		1.45 ± 0.145	0.75 ± 0.075	1.65 ± 0.165	1.90 ± 0.190	2.5 ± 0.25	1.8 ± 0.18	1.65 ± 0.165	2.25 ± 0.225		
Mango		—	2.25 ± 0.225	2.85 ± 0.285	0.75 ± 0.075	2.1 ± 0.21	2.55 ± 0.255	2.95 ± 0.295	0.55 ± 0.055		
Limes		1.5 ± 0.15	0.45 ± 0.045	1.35 ± 0.135	1.8 ± 0.18	1.35 ± 0.135	2.1 ± 0.21	0.65 ± 0.065	1.8 ± 0.18		
Pineapple		0.8 ± 0.08	0.95 ± 0.095	0.85 ± 0.085	0.45 ± 0.045	0.4 ± 0.04	—	1.8 ± 0.18	0.65 ± 0.065		
Apple		0.55 ± 0.055	0.35 ± 0.035	0.65 ± 0.065	1.35 ± 0.135	—	0.45 ± 0.045	—	1.15 ± 0.115		
Pomegranate		0.7 ± 0.07	—	—	1.15 ± 0.115	0.6 ± 0.06	—	1.4 ± 0.14	1.6 ± 0.16		
Melon		1.7 ± 0.17	3.05 ± 0.305	0.65 ± 0.065	1.25 ± 0.125	2.55 ± 0.255	1.85 ± 0.185	0.85 ± 0.085	1.45 ± 0.145		

Table 1 (continued)

Types of household waste		(b). S9 to S16 with a total sample size of 540 garbage bags collected between May 2020 to August 2020										
		S9(50)	S10(55)	S11(60)	S12(65)	S13(70)	S14(75)	S15(80)	S16(85)			
		(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)			(kg)
Strawberries		–	–	–	–	–	0.65 ± 0.065	–	0.35 ± 0.035			
Papaya		0.4 ± 0.04	0.35 ± 0.035	0.55 ± 0.055	0.25 ± 0.035	1.35 ± 0.135	0.55 ± 0.055	0.55 ± 0.055	–			
Kiwi		–	–	–	–	–	0.4 ± 0.04	0.6 ± 0.06	–			
Tea leaves		3.3 ± 0.33	3.8 ± 0.38	3.55 ± 0.355	4.8 ± 0.48	4.6 ± 0.46	4.65 ± 0.465	3.65 ± 0.365	2.9 ± 0.29			
Bread		0.4 ± 0.04	0.75 ± 0.075	1.25 ± 0.125	1.45 ± 0.145	1.9 ± 0.19	1.7 ± 0.17	1.6 ± 0.16	1.4 ± 0.14			
Rice		0.9 ± 0.09	1.75 ± 0.175	0.5 ± 0.05	1.5 ± 0.15	1.4 ± 0.14	2.1 ± 0.21	1.9 ± 0.19	2.65 ± 0.265			
Mixed waste		4.55 ± 0.455	3.05 ± 0.305	4.85 ± 0.485	3.5 ± 0.35	4.6 ± 0.46	5.1 ± 0.51	5.3 ± 0.53	5.5 ± 0.55			
Types of household waste		(c). S17 to S24 with a total sample size of 860 garbage bags collected between September 2020 to December 2020										
		S17(90)	S18(95)	S19(100)	S20(105)	S21(110)	S22(115)	S23(120)	S24(125)			
		(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)			(kg)
Broccoli		0.4 ± 0.04	0.3 ± 0.03	0.4 ± 0.04	–	0.45 ± 0.045	1.45 ± 0.145	1.22 ± 0.122	–			
Brussels sprouts		0.55 ± 0.055	0.5 ± 0.05	0.9 ± 0.09	–	0.9 ± 0.09	1.5 ± 0.15	1.4 ± 0.14	–			
Cabbage		0.95 ± 0.095	–	–	0.9 ± 0.09	2.52 ± 0.252	1.12 ± 0.112	1.6 ± 0.16	1.8 ± 0.18			
Cauliflower		0.6 ± 0.06	1.9 ± 0.19	1.1 ± 0.11	0.45 ± 0.045	1.9 ± 0.19	0.5 ± 0.05	0.90 ± 0.09	1.4 ± 0.14			
Grapefruit		0.71 ± 0.071	–	0.5 ± 0.055	1.75 ± 0.175	1.6 ± 0.16	0.45 ± 0.045	0.75 ± 0.75	0.90 ± 0.09			
Kale		–	0.3 ± 0.03	–	1.4 ± 0.14	0.50 ± 0.05	–	0.9 ± 0.09	–			
Leeks		–	–	0.9 ± 0.09	–	0.75 ± 0.75	0.4 ± 0.04	0.35 ± 0.035	0.75 ± 0.75			
Lemons		0.55 ± 0.055	3.5 ± 0.35	0.9 ± 0.09	0.45 ± 0.045	0.35 ± 0.035	0.95 ± 0.095	1.4 ± 0.14	0.9 ± 0.09			
Parsnips		0.25 ± 0.025	0.7 ± 0.07	0.3 ± 0.03	0.9 ± 0.09	–	1.1 ± 0.11	0.5 ± 0.05	0.8 ± 0.08			
Rutabagas		–	0.6 ± 0.06	–	1.14 ± 0.114	1.8 ± 0.18	1.9 ± 0.19	1.85 ± 0.185	0.6 ± 0.06			
Tangelos		0.35 ± 0.035	0.95 ± 0.095	–	0.9 ± 0.09	1.4 ± 0.14	1.8 ± 0.18	0.45 ± 0.045	0.8 ± 0.08			

Table 1 (continued)

Types of household Organic waste		(c). S17 to S24 with a total sample size of 860 garbage bags collected between September 2020 to December 2020									
		S17(90)	S18(95)	S19(100)	S20(105)	S21(110)	S22(115)	S23(120)	S24(125)		
		(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)	(kg)		(kg)
Tangerines		0.85 ± 0.085	0.6 ± 0.06	1.5 ± 0.15	1.6 ± 0.16	0.83 ± 0.083	0.55 ± 0.055	1.35 ± 0.135	1.43 ± 0.143		
Turnips		1.65 ± 0.165	0.35 ± 0.035	–	0.50 ± 0.05	1.85 ± 0.185	0.70 ± 0.070	0.83 ± 0.083	1.5 ± 0.15		
Beetroot		3.5 ± 0.35	0.5 ± 0.050	0.3 ± 0.03	2.75 ± 0.275	1.4 ± 0.14	0.35 ± 0.035	1.5 ± 0.15	0.95 ± 0.095		
Carrots		–	–	–	–	1.35 ± 0.135	1.5 ± 0.15	1.4 ± 0.14	0.85 ± 0.085		
Chicory		–	0.6 ± 0.06	0.4 ± 0.04	0.95 ± 0.095	0.75 ± 0.075	–	2.1 ± 0.21	1.5 ± 0.15		
Potatoes		4.04 ± 0.404	0.25 ± 0.025	0.45 ± 0.045	1.9 ± 0.19	2.1 ± 0.21	–	1.5 ± 0.15	0.75 ± 0.075		
Morel Mushrooms		–	1.85 ± 0.185	1.75 ± 0.175	–	–	–	0.6 ± 0.06	0.5 ± 0.05		
Parsnips		0.9 ± 0.09	–	–	0.35 ± 0.035	1.5 ± 0.15	0.75 ± 0.075	1.45 ± 0.145	0.75 ± 0.075		
Rhubarb		–	–	–	–	0.5 ± 0.05	1.8 ± 0.18	0.8 ± 0.08	2.15 ± 0.215		
Sorrel		0.35 ± 0.035	2.85 ± 0.0285	1.4 ± 0.14	1.5 ± 0.15	–	1.43 ± 0.143	0.9 ± 0.09	1.4 ± 0.14		
Spinach		–	–	2.25 ± 0.225	1.35 ± 0.135	–	0.8 ± 0.08	1.5 ± 0.15	1.6 ± 0.16		
Spring Greens		0.6 ± 0.06	0.95 ± 0.095	0.35 ± 0.035	1.8 ± 0.18	1.4 ± 0.14	1.75 ± 0.175	0.75 ± 0.075	1.4 ± 0.14		
Spring Onions		1.5 ± 0.15	1.6 ± 0.16	0.95 ± 0.095	1.25 ± 0.125	–	1.14 ± 0.114	1.4 ± 0.14	1.35 ± 0.135		
Watercress		2.35 ± 0.235	0.45 ± 0.045	–	1.4 ± 0.14	–	1.35 ± 0.135	–	0.9 ± 0.09		
Green fenugreek		–	–	0.45 ± 0.045	.8 ± 0.08	1.5 ± 0.15	0.5 ± 0.05	0.55 ± 0.055	0.7 ± 0.07		
Pumpkin		1.35 ± 0.135	–	0.9 ± 0.09	1.45 ± 0.145	1.45 ± 0.145	1.4 ± 0.14	–	1.52 ± 0.152		
Onion		0.45 ± 0.045	1.25 ± 0.125	1.35 ± 0.135	1.5 ± 0.15	1.5 ± 0.15	1.8 ± 0.18	0.4 ± 0.04	1.45 ± 0.0145		
Ladyfinger		0.9 ± 0.09	–	–	–	0.6 ± 0.06	–	–	0.55 ± 0.055		
Green coriander		–	–	1.9 ± 0.019	0.9 ± 0.09	–	0.6 ± 0.06	0.8 ± 0.08	0.55 ± 0.055		
Radish		–	0.9 ± 0.09	1.5 ± 0.15	0.8 ± 0.08	–	–	–	1.4 ± 0.14		
Tomatoes		1.45 ± 0.145	1.6 ± 0.16	1.35 ± 0.135	0.3 ± 0.03	0.9 ± 0.09	1.6 ± 0.16	1.45 ± 0.0145	0.8 ± 0.08		
Capsicum		1.6 ± 0.16	1.35 ± 0.135	1.8 ± 0.18	–	0.8 ± 0.08	–	–	0.4 ± 0.04		
Bottle Gourd		1.8 ± 0.18	1.7 ± 0.17	1.25 ± 0.125	1.48 ± 0.0148	1.45 ± 0.0145	–	0.55 ± 0.055	1.5 ± 0.15		

Table 1 (continued)

Types of household Organic waste	(c). S17 to S24 with a total sample size of 860 garbage bags collected between September 2020 to December 2020									
	S17(90) (kg)	S18(95) (kg)	S19(100) (kg)	S20(105) (kg)	S21(110) (kg)	S22(115) (kg)	S23(120) (kg)	S24(125) (kg)		
Pea peel	—	—	—	—	0.8±0.08	2.26±0.226	1.4±0.14	1.41±0.141		
Banana	3.25±0.325	1.65±0.165	2.5±0.25	1.25±0.125	1.7±0.17	1.65±0.165	0.9±0.09	1.95±0.195		
Apple	2.8±0.28	1.25±0.125	1.45±0.145	1.65±0.165	1.5±0.15	0.65±0.065	0.5±0.05	0.35±0.035		
Limes	2.6±0.26	1.45±0.145	2.±0.20	2.75±0.275	1.45±0.145	1.3±0.13	2.55±0.255	1.5±0.15		
Pineapple	2.5±0.25	2.75±0.275	0.65±0.065	1.45±0.145	2.5±0.25	1.5±0.15	1.75±0.175	3.05±0.305		
Plum	0.65±0.065	0.35±0.035	0.5±0.05	—	0.65±0.065	1.45±0.145	1.15±0.115	1.65±0.165		
Pomegranate	0.5±0.05	1.1±0.11	1.25±0.125	1.5±0.15	1.65±0.165	1.75±0.175	1.3±0.13	1.7±0.17		
Sugarcane	—	—	—	0.5±0.05	1.3±0.13	1.15±0.115	1.65±0.165	1.3±0.13		
Orange	1.65±0.165	2.5±0.25	2.75±0.275	2.5±0.25	1.55±0.155	1.7±0.17	1.5±0.15	2.25±0.225		
Papaya	1.25±0.125	2.1±0.21	1.65±0.165	0.65±0.065	1.45±0.145	2.55±0.255	1.45±0.145	1.25±0.125		
Carobs	0.35±0.035	—	0.65±0.065	—	0.35±0.035	0.5±0.05	1.5±0.15	1.65±0.165		
Kiwi	—	0.5±0.05	—	1.15±0.115	1.5±0.15	0.35±0.035	0.65±0.065	0.5±0.05		
Persimmon,	1.45±0.145	—	—	0.35±0.035	—	1.5±0.15	—	0.7±0.07		
Pear	0.55±0.055	0.65±0.065	0.5±0.05	—	1.15±0.115	0.5±0.05	1.7±0.17	0.65±0.065		
Raspberries	—	—	—	0.5±0.05	—	—	0.35±0.035	0.5±0.05		
Blackberries	—	0.55±0.055	0.35±0.035	1.3±0.13	0.5±0.05	—	0.7±0.07	0.6±0.06		
Grapes	—	—	1.1±0.11	—	0.5±0.05	—	0.5±0.05	0.9±0.09		
Tea leaves	3.5±0.35	3.8±0.38	3.4±0.34	4.5±0.45	4.1±0.41	4.33±0.433	4.83±0.483	5.41±0.541		
Bread	2.15±0.215	2.25±0.225	2.05±0.205	2.55±0.255	2.95±0.295	3.95±0.395	3.4±0.34	3.42±0.342		
Rice	2.5±0.25	2.9±0.29	2.6±0.26	3.3±0.33	3.7±0.37	4.5±0.45	4.22±0.422	3.2±0.32		
Mixed waste	5.4±0.54	6.5±0.65	6.2±0.62	5.5±0.55	5.52±0.552	6.72±0.672	6.5±0.65	7.5±0.75		

Table 2 Total amount of raw vegetables waste (RVW), fruits waste (FW), and mixed cooked waste (MCW) wastes were released from the sample size

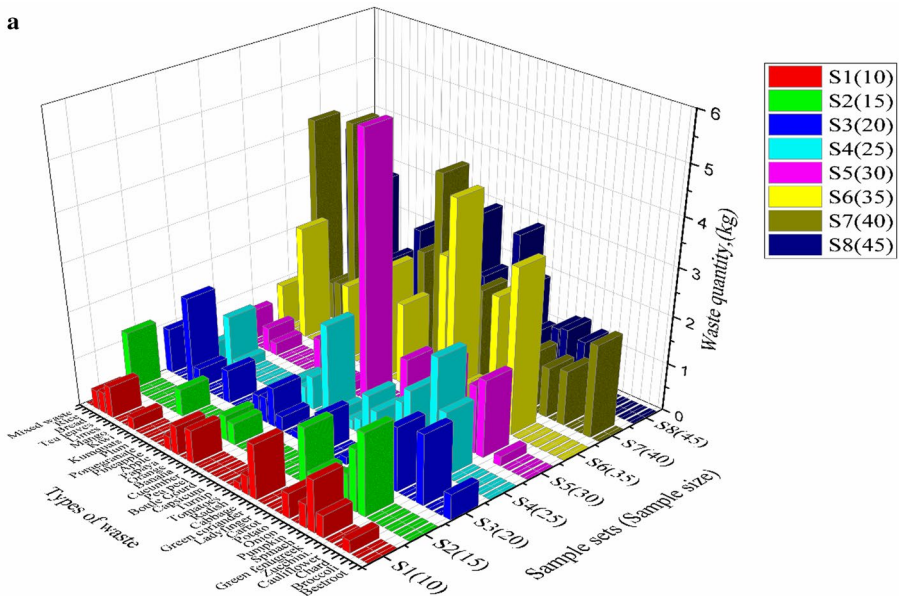
Months	Sample set (Sample size)	Vegetable waste (X), kg	Fruit waste (Y), kg	Mixed waste (Z), kg
January	S1 (10)	4.55	1.85	1.51
January	S2 (15)	5.85	1.55	1.22
January	S3 (20)	6.55	3.11	2.8
February	S4 (25)	10.51	1.87	1.71
February	S5 (30)	12.2	1.55	1.15
March	S6 (35)	17.4	4.29	3.31
March	S7 (40)	16.34	8.75	6.41
April	S8 (45)	17.54	9.7	7.63
May	S9 (50)	19.3	9.25	9.15
May	S10 (55)	21.2	10.5	9.35
June	S11 (60)	22.5	11.3	10.15
June	S12 (65)	24.3	11.5	11.25
July	S13 (70)	22.65	14.65	12.5
July	S14 (75)	25.35	12.85	13.55
August	S15 (80)	24.75	13.15	12.45
August	S16 (85)	26.91	11.45	12.45
September	S17 (90)	27.65	17.55	13.55
September	S18 (95)	25.55	14.85	15.45
October	S19 (100)	24.85	15.35	14.25
October	S20 (105)	30.47	15.55	15.85
November	S21 (110)	32.85	17.75	16.27
November	S22 (115)	31.45	16.55	19.5
December	S23 (120)	32.55	18.15	18.95
December	S24 (125)	35.26	20.5	19.53

and 115 with a total sample size of 225, yielding 64.3 kg of RVW, 34.3 kg of FW, and 35.77 kg of MCW. Other wastes, including tea leaves, rice, bread, pineapple, and orange, were the highest waste in these samples. S23 and S24 were collected in December with a total sample size of 245 bags containing 67.81 kg of RVW, 38.65 kg of FW, and 38.48 kg of MCW yielded the highest waste amounts of mixed waste, tea leaves, rice, bread, limes, and oranges. There were mainly three types of waste in each sample, RVW, FW, and MCW. In every model, total vegetable waste from RVW, complete fruit waste from FW, entire mixed waste from MCW to Shown in Table 2. Mixed waste means organic waste, which is very difficult to segregate, and all types of cooked waste are placed in the MCW category. Due to the reduced sample size, S1, S2, and S3 were studied in a month, and in April, due to the COVID19 pandemic, only one sample set could be collected. In the rest of every month, about two models have been gathered and analyzed.

2.2 Co-Composting and biogas production materials

N, P, and K are essential constituents of fertilizers. N forms the building blocks photosynthesis. P is necessary for flowering, fruiting, crop maturity, seed production, and root

a



b

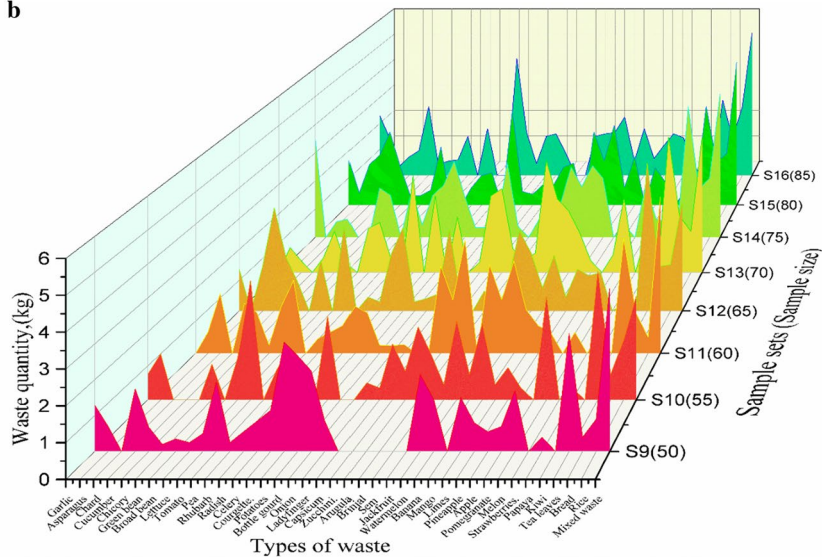


Fig. 4 Quantity (kg) of different types of household organic waste collected in garbage bags **a** Sample set S1 to S8 taken from January 2020 to April 2020, **b** S9 to S16 taken from May 2020 to August 2020, **c** S17 to S24 taken from September 2020 to December 2020

development. K helps plants withstand extreme cold and hot temperatures, resist pests, keep roots healthy, and plants to tolerate stresses such as drought (NOBLE RESEARCH INSTITUTE n.d.).

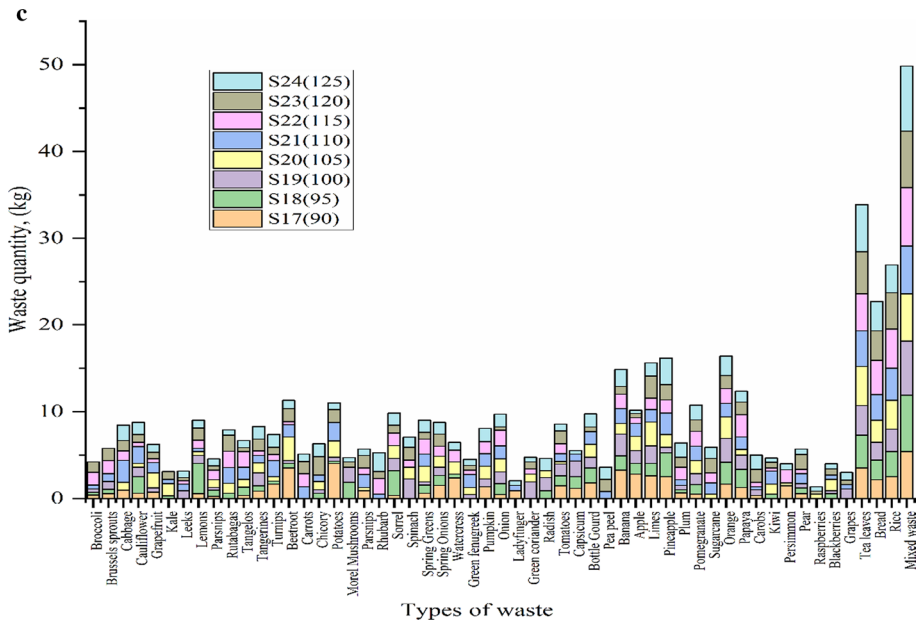


Fig. 4 (continued)

These three nutrients are found in compost made from household vegetable and fruit waste, as shown in Table 3. These wastes and the wastes found in the sample can be used to compost and make biogas. Value of N-P-K in the fertilizer used in gardens is 1–1–1 for tree plant growth, 1–1–2, 1–2–2 for flower-fruit development, 2–1–1 for leaf growth, and 1–2–1 for root development (Espace pour la vie montreal n.d.). Two hundred grams of pineapple waste, orange waste, pumpkin, and spinach waste can produce 965 cm³, 612 cm³, 373 cm³, 269 cm³ biogas weekly, respectively (Sagagi et al., 2010). For the growth of plants, the digested manure of pumpkin, spinach, and pineapple is more effective than the undigested manure. Digested waste of spinach and pineapple takes an average of three days for the seeds to germinate, while their undigested waste takes four days. 62.52 kg of vegetable and fruit waste (VFW) was taken in the ratio of 2.2:2.8, resulting in an average composition of 57.58% methane (Masebinu et al., 2018).

A mixture of 78% vegetable waste, 4% tuber waste, and 18% fruit waste with a total weight of 160 kg is treated in a 200-L biogas digester for 14 weeks, with the best amount of 65% methane content found in biogas, which means the mixture of VFW waste gives a good amount of methane gas (Sitorus et al., 2013). 3 kg of kitchen waste, vegetable waste, and fruit waste mixed with 9 L of water produced 0.0000080m³/day, 0.0000066m³/day, and 0.0000022m³/day biogas, respectively, at 30-degree Celsius temperature (Ojolo et al., 2007). Each fruit and vegetable waste from household waste can produce a biogas yield. Lemon, mixed food waste, and cooked meat have the highest biogas yield, seen in Fig. 5a. Fibrous vegetables and fruits can be used to make biogas. But due to their fibrous properties, it starts to choke the biogas digester after some time, which can be seen in Fig. 5b, which reduces the efficiency of the biogas plant. Therefore, fibrous organic waste is used more for compost than biogas. After the raw biogas

Table 3 Value of N-P-K (%) in compost made from dry household waste

Materials	N (%)	P (%)	K (%)
Spinach	4.73	0.22	4.06
Cabbage	2.88	0.17	2.73
Mustard	6.12	0.27	3.92
Carrot	1.91	0.13	2.59
Cucumber	2.0	0.07	2.31
Orange	0.67	0.03	0.98
Pineapple	0.78	0.06	2.35
Apple	0.39	0.07	1.8
Banana	0.56	0.05	2.1
Watermelon	2.67	0.2	3.83
Sweet melon	1.97	0.15	3.15
Barley (Grain)	1.75	0.75	0.5
Beet (Root)	0.25	0.1	0.5
Brewery grain (Wet)	0.9	0.5	0.05
Brigham tea (Ash)	0	0	5.94
Castor bean (Pomace)	6.0	2.5	1.25
Coffee grounds	2.08	0.32	0.28
Corn (grain)	1.65	0.65	0.4
Cottonseed	3.15	1.25	1.15
Cowpeas (green forage)	0.45	0.12	0.45
Cowpeas (seed)	3.1	1.0	1.2
Crabgrass (green)	0.66	0.19	0.71
Eggs	2.25	0.4	0.15
Beans—garden beans	0.25	0.08	0.3
Grapes	0.15	0.07	0.3
Lemon culls	0.15	0.06	0.26
Mussels	0.9	0.12	0.13
Oak leaves	0.8	0.35	0.15
Oats (Grain)	2.0	0.8	0.6
Pea pods (Ash)	0	1.79	9
Peanuts	3.6	0.7	0.45
Pine needles	0.46	0.12	0.03
Potato	0.35	0.15	0.5
Pumpkin (flesh)	0.16	0.07	0.26
Raw sugar residue	1.14	8.33	0
Sweet potato	0.25	0.1	0.5
Tea leaves (grounds)	4.15	0.62	0.4
Tomato	0.2	0.07	0.35

Source: (The nutrient company, n.d.)

is generated, it is converted into enriched biogas with the help of H_2S , CO_2 scrubber, as can be seen in Table 4, where the biogas analyzer shows the composition (%) of raw biogas and composition (%) of biogas after the purification process.

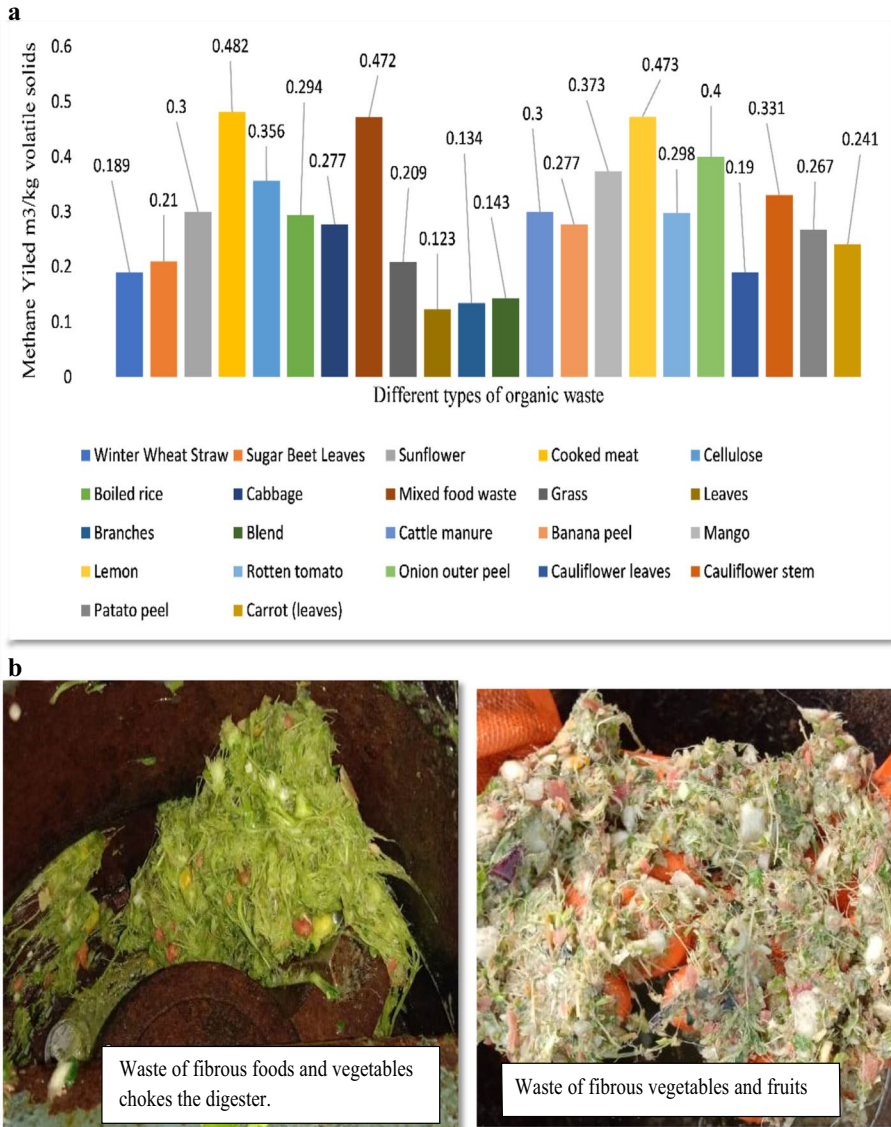


Fig. 5 **a** Methane (CH_4) yield of different organic waste/feedstock (Dhar et al., 2017). **b** Fibrous vegetable and fruit waste

Table 4 Composition of raw and purified biogas checked by biogas analyzer

Component	Raw biogas (Composition)	Biogas after purification (Composition)
Methane (CH_4)	48.5%	94.5%
Carbon dioxide (CO_2)	47.5%	2.5%
Hydrogen sulfide (H_2S)	78 ppm	0 ppm
Oxygen (O_2)	4.0%	3.0%

2.3 Regression analysis

Sample sizes in most of the houses were mostly raw vegetable waste (RVW), fruit waste (FW), and mixed cooked waste (MCW). It means that these three wastes depend on each other. So, the interrelationship of three wastes was evaluated by the regression method. First, let's determine the relationship between RVW and FW by the single variable regression method at a 95% confidence level.

Regression line FW on RVW,

$$FW - \overline{FW} = \frac{\text{Cov}(\text{RVW}, \text{FW}) (\text{RVW} - \overline{\text{RVW}})}{\sigma_{\text{RVW}}^2} \quad (1)$$

where

Cov = Coefficient of variation;

$$\text{Cov}(\text{RVW}, \text{FW}) = \frac{1}{n} \sum \text{RVW} \times \text{FW} - \overline{\text{RVW}} \times \overline{\text{FW}} \quad (2)$$

Standard deviation of RVW (σ_{RVW})

$$\sigma_{\text{RVW}} = \sqrt{\frac{1}{n} \sum \text{RVW}^2 - (\overline{\text{RVW}})^2} \quad (3)$$

Mean of vegetable waste, $\overline{\text{RVW}} = \frac{\sum \text{RVW}}{n}$; Mean of fruit waste, $\overline{\text{FW}} = \frac{\sum \text{FW}}{n}$ where n = no. of observations.

ⁿ Using Eqs. (1), (2) and (3), regression line FW on RVW is

$$FW = 0.6375 \text{RVW} - 2.792 \quad (4)$$

Standard error for FW (FW^*).

$$FW^* = \sqrt{\frac{\sum (FW - \widehat{FW})^2}{n - 2}}, \quad (5)$$

Coefficient of determination (R^2) for FW is

$$R^2 = \frac{\sum (\widehat{FW} - \overline{FW})^2}{\sum (FW - \overline{FW})^2}, \quad (6)$$

where FW is real value, \widehat{FW} is an estimated value, and \overline{FW} is the mean value.

Regression line RVW on FW.

$$\text{RVW} - \overline{\text{RVW}} = \frac{\text{Cov}(\text{RVW}, \text{FW}) (\text{FW} - \overline{\text{FW}})}{\sigma_{\text{FW}}^2} \quad (7)$$

Standard deviation of FW (σ_{FW})

$$\sigma_{FW} = \sqrt{\frac{1}{n} \sum FW^2 - (\overline{FW})^2} \quad (8)$$

Using Eqs. (2), (7), and (8), regression line RVW on FW is

$$RVW = 1.4159FW + 6.0559 \quad (9)$$

Standard error for RVW (RVW*).

$$RVW^* = \sqrt{\frac{\sum (RVW - \widehat{RVW})^2}{n - 2}}, \quad (10)$$

Coefficient of determination (R^2) for RVW-

$$R^2 = \frac{\sum (\widehat{RVW} - \overline{RVW})^2}{\sum (RVW - \overline{RVW})^2} \quad (11)$$

where \widehat{RVW} is an estimated value, RVW is real value, and \overline{RVW} is the mean value.

Similarly, the wastes (RVW, MCW) and (FW, MCW) have been calculated by applying the regression method, seen in Table 5.

3 Result and discussion

Almost every household uses vegetables, cooked food, and fruits in their daily meals, so this waste is found in nearly every sample collected from households. It means that when the use of vegetables and cooked food is less in a family, the consumption of fruits and other food items will increase; hence the quantity of fruit and mixed waste increases. Similarly, the number of vegetables and mixed waste increases when the fruit is less used in the family, and the vegetables and other things are consumed more in the diet. Mixed waste means waste that cannot be separated appropriately. Like other organic waste, mixed waste also originates from houses every day, such as cooked rice, bread, used tea, cooked vegetables, and lentils which belong to other diverse waste categories. It means that RVW, FW, and MCW waste are generated more or less from households every day depending on the consumption rate. With the help of the regression analysis (RVW, FW), (RVW, MCW), and (FW, MCW), the relationship between RVW, FW, and MCW waste was estimated, and the result summary output can be seen in Table 5.

3.1 Raw vegetable waste (RVW) and Fruit waste (FW) depend on each other

Total observations for all calculations considered for 24 sample sets, and the confidence level is 95%. In Fig. 6a, the regression line is $FW = 0.6375RVW - 2.792$ when FW on RVW, the linear predicted fruit waste line is shown at a different value of raw vegetable waste, and the value shown around the straight line is the actual value of fruit waste.

The coefficient of determination (R^2), the standard error, and the p -value are approximately 0.9026, 1.8811, and $1.31E-12$, as shown in Table 5(a). Value of p is much lower than $\alpha = 0.05$. These data indicate that the model obtained is approximately 90% accurate with significance. When RVW depends on FW, the regression line is

Table 5 Result summary output of regression statistics when (RVW, FW), (RVW, MCW), and (FW, MCW) wastes depend on each other

(a) Relation between RVW and FW waste		Summary output					
		Regression statistics					
		Multiple R	R Square	Adjusted R Square	Standard error	Observations	Intercept
When FW on RVW, FW = 0.6375RVW-2.792		0.95	0.9026	0.8982	1.8811	24	(RVW)
When RVW on FW, RVW = 1.4159FW + 6.0559		0.95	0.9026	0.8982	2.8035	24	Intercept (FW)
							Standard error
							<i>t</i> Stat
							<i>p</i> -value
(b) Relation between RVW and MCW waste		Summary output					
		Regression statistics					
		Multiple R	R Square	Adjusted R Square	Standard error	Observations	Intercept
When RVW on MCW, RVW = 1.3955MCW + 7.0722		0.9571	0.9161	0.9123	2.6018	24	(MCW)
When MCW on RVW, MCW = 0.6565RVW-3.77		0.9571	0.9161	0.9123	1.7845	24	Intercept (RVW)
							Standard error
							<i>t</i> Stat
							<i>p</i> -value
(c) Relation between FW and MCW waste		Summary output					
		Regression statistics					
		Multiple R	R Square	Adjusted R Square	Standard error	Observations	Intercept
When FW on MCW, FW = 0.9529MCW + 1.0587		0.9739	0.9486	0.9463	1.3665	24	(MCW)
When MCW on FW, MCW = 0.9956FW-0.5139		0.9739	0.9486	0.9463	1.3968	24	Intercept (FW)
							Standard error
							<i>t</i> Stat
							<i>p</i> -value

$RVW = 1.4159FW + 6.0559$. Similarly, the vegetable waste line's predicted value in Fig. 6b is a different value of fruit waste and the real deal of the vegetable waste around the straight line. Table 5(a) shows that the coefficient of determination (R^2), p -value, and standard error are 0.90, $1.31E-12$, and 2.8035, respectively. It means that the model found is highly significant, up to 90%.

3.2 Raw vegetable waste (RVW) and mixed cooked waste (MCW) depend on each other

When RVW depends on MCW, the regression line is $RVW = 1.3955MCW + 7.0722$, and R^2 , Standard Error, and p -Value are 0.9161, 2.6018, and $2.51E-13$, respectively. Similarly, when MCW depends on RVW, the regression line is $MCW = 0.6565RVW - 3.77$, where R^2 , Standard Error, and p -value are 0.9161, 1.7845, and $2.51E-13$, respectively, as shown in Table 5(b). Value of p is much less than $\alpha = 0.05$. It means that the model is approximately 91% correct with highly significant in both conditions. Figure 6c, d shows the linear raw vegetable waste (RVW) line and the linear mixed cooked waste line (MCW). Linear RVW line is found at different MCW values, and the actual value of RVW around the linear RVW line is shown in Fig. 6c; similarly, the deal of linear MCW is found at different RVW values, and the actual value of MCW is shown around the linear MCW line in Fig. 6d.

3.3 Fruit waste (FW) and mixed cooked waste (MCW) depend on each other

When FW destruction depends on RVW, the equation of $FW = 0.9529MCW + 1.0587$ regression line is found. As shown in Table 5(c), value of R^2 , standard error, and p -value are 0.9486, 1.3665, and $1.13E-15$, respectively. Similarly, when MCW depends on FW, the regression line is $MCW = 0.9956FW - 0.5139$, where R^2 , standard error, and p -value are 0.9486, 1.3968, and $1.13E-15$, respectively. Value of p is much less than $\alpha = 0.05$. It means that model is accurate up to 94% and highly significant in both conditions. In Fig. 6e–f, the linear fruit waste (FW) line and linear mixed cooked waste lines (MCW) are shown as linear FW waste line, which is found at different MCW values. Actual value of FW around the linear FW line. Similarly, the linear MCW line is located at different FW values. Real deal of MCW is around the linear MCW line. So, it can be seen that the model derived from (FW, MCW) waste is the most highly significant because it has the highest value of R^2 .

4 Discussion

Outcome of this case study is that the household waste collected in one year was found in all the sample sizes, mainly fruits, raw vegetables, and mixed cooked waste. Organic waste in the form of RVW, FW, and MCW was 518.53 kg, 263.57 kg, and 249.94 kg, respectively, collected during the study, which was used daily to make 50 m^3 of biogas, 160 kW of green electric energy, and 200 kg of organic compost. "A case study was done at the University of Benin Ugbowo campus; in this study, it was found that approximately 10,144.4 kg of MSW is generated per month. Due to this, about 968.30 m^3 of biogas is produced. In this study, it was found that 41% of the waste is used for making direct compost and biogas. Reaming Energy is generated by recyclable 59% MSW

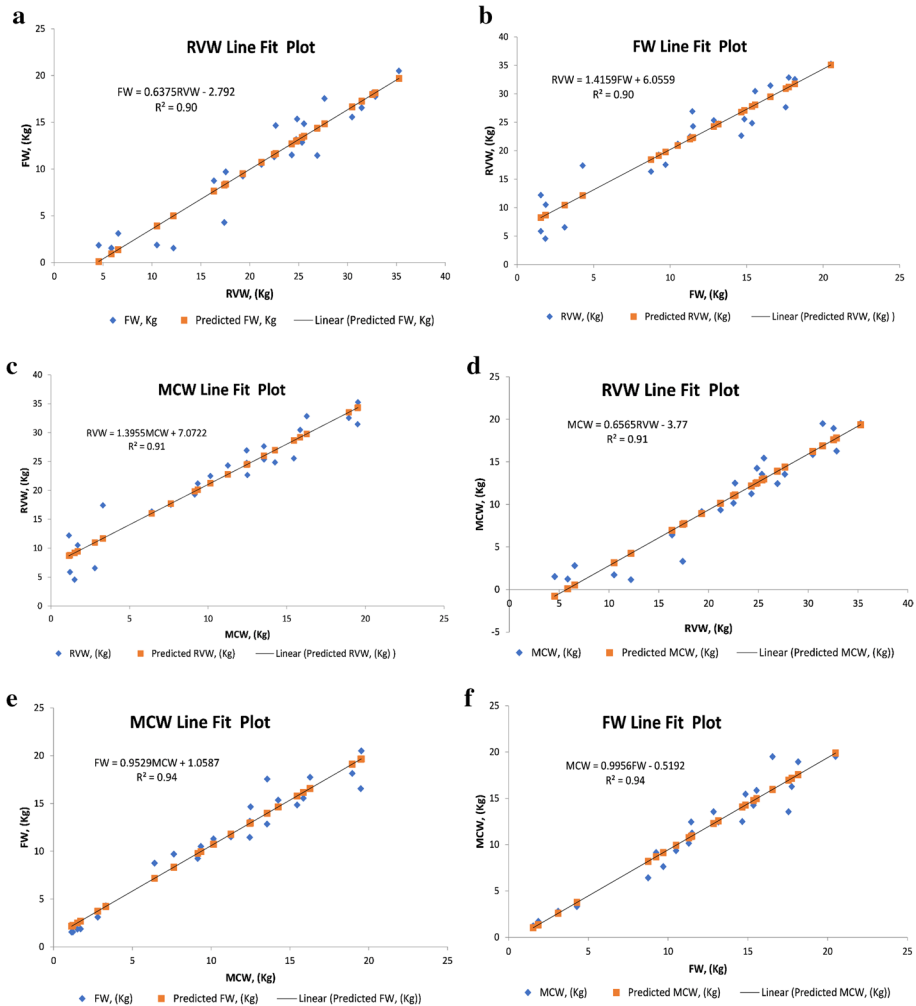


Fig. 6 Graphical representation of the linear regression line **a** when FW on RVW waste **b** RVW on FW **c** RVW on MCW **d** MCW on RVW **e** FW on MCW **f** MCW on FW waste

and segregation” (Kubeyinje, Centre, Energy, and Akingba 2022). In energy conversion, organic and biodegradable waste can be done either directly or by recycling. “In a study conducted on the East Coast of Malaysia, a total of 338 households were interviewed, and data collection was done using the Statistical Package for The Social Sciences (SPSS) technique. The Chi-square goodness of fit test was also used to determine the categorical variables. The study’s findings showed that 18.3% of homes disposed of plastic products as waste, and 74.3% of households disposed of food debris as waste. Study also revealed that while 49.7% of families did not separate their waste, 50.3% of them did. 95.9% of those surveyed were aware that diseases like malaria and diarrhea are brought on by poor waste management. According to the Chi-square test, $p < 0.05$, there were relationships between respondents’ waste segregation practices and their location, age, and dwelling type. Additionally, associations between location and the

perception of poor waste management leading to disease were discovered (Chi-square test, $p < 0.05$). This study is from Algiers City, also Algeria's capital. The survey of MSW coming out in the city has been done with the help of multiple regression and correlation analyses. The results of the multiple regression analysis show that the size of the settlement, specifically its area ($p = 0.0006$) and population ($p = 0.0028$), as well as the characteristics of the waste management companies, precisely the number of collection routes ($p = 0.0001$) and employees ($p = 0.0026$), has an impact on waste management (Kebaili et al., 2022). Correlation between dependent and independent variables can be established using regression and other statics tools. Similarly, this case study also used regression analyses between collecting FW, RVW, and CW. While applying the regression method, all the models have $p < 0.05$ and got the R^2 value between 0.9 and 0.94, which means whatever we get is accurate up to 90–94%.

5 Conclusion

In the present study, we evaluated each collected sample set and the amount of RVW, FW, and MCW generated from 73 types of household waste in the DTU campus. Sample S1 to S24 sample sets with a total of 1620 sample sizes were collected throughout the study. A total of 518.53 kg RVW, 263.57 kg FW, and 249.94 kg of MCW were accumulated and segregated into digestive and compost wastes. The coefficient of determination (R^2) of these (RVW, FW), (RVW, MCW), and (FW, MCW) were observed to be 0.90, 0.91, and 0.94 with $p < 0.05$, and the standard error value lies between 1.3665 and 2.8035. Our data suggest that all waste is interdependent and significant in biogas production. Fifty cubic meters of biogas was produced from one ton of digestive waste every day, which was adequate to produce 160 kW of green electric energy and 200 kg of compost. Composting waste is composted by placing it in the pit. In this way, the organic waste generated on the university campus was adequately utilized and managed. Every organic waste contains N-P-K nutrients, which could be transformed into biogas production and compost formation. Citizens of every country of the world should properly mark and throw organic and inorganic waste separately; therefore, waste can be segregated appropriately and recycled as green energy. Our study supports the recommendation that a fair number of biogas plants should be installed in societies and university campuses of metros worldwide; therefore, organic waste released from households could be used for green energy production. In this way, a vast problem that existed because of solid waste generated in metros can be resolved to a great extent.

Acknowledgements The authors would like to thank Dr. Anil Kumar for his valuable support in modeling and Delhi Technical University administration for providing the necessary finance and logistics to conduct this research on campus.

Author's contribution All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by [PKM], [AP] and [S]. The first draft of the manuscript was written by [PKM], and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The study was supported under the Making DTU a Zero Organic Waste Campus Project. Grant no. DTU/Council/BOM-AC/Notification/31/2018/5738.

Declarations

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

Ethical approval As authors, we would like to tell you that this is our original work, and this paper has not been submitted anywhere except in this journal.

References

- Armijo de Vega, C., Ojeda-Benítez, S., & Ramírez-Barreto, M. E. (2003). Mexican educational institutions and waste management programmes: A University case study. *Resources, Conservation and Recycling*, 39(3), 283–296. [https://doi.org/10.1016/S0921-3449\(03\)00033-8](https://doi.org/10.1016/S0921-3449(03)00033-8)
- Chaturvedi, H., Das, P., & Kaushal, P. (2018). Comparative study of different Biological Processes for non-segregated Municipal Solid Waste (MSW) leachate treatment. *Environmental Technology and Innovation*, 9, 134–139. <https://doi.org/10.1016/j.eti.2017.11.008>
- Chen, P., Xie, Q., Addy, M., Zhou, W., Liu, Y., Wang, Y., et al. (2016). Utilization of municipal solid and liquid wastes for bioenergy and bioproducts production. *Bioresource Technology*, 215, 163–172. <https://doi.org/10.1016/j.biortech.2016.02.094>
- Dhar, H., Kumar, S., & Kumar, R. (2017). A review on organic waste to energy systems in India. *Biore-source Technology*, 245(August), 1229–1237. <https://doi.org/10.1016/j.biortech.2017.08.159>
- DTU. (2009). How to reach Delhi College of Engineering. <https://web.archive.org/web/20090401001146/http://dce.edu/campus/location.htm>
- DTU. (2014). DTU CAMPUS MAP. https://en.wikipedia.org/wiki/The_Times_of_India
- El-Shinnawy, N. A., Heikal, S. O., & Fahmy, Y. (1983). Saccharification of cotton bolls by concentrated sulphuric acid, (January 1983). Espace pour la vie montreal. (n.d.). THE GREEN PAGES. <https://m.espacepourlavie.ca/en/ratio-fertilizer-n-p->
- Fahmy, T. Y. A., Fahmy, Y., Mobarak, F., El-Sakhawy, M., & Abou-Zeid, R. E. (2020). Biomass pyrolysis: Past, present, and future. *Environment, Development and Sustainability*, 22(1), 17–32. <https://doi.org/10.1007/s10668-018-0200-5>
- Fahmy, T. Y. A., & Mobarak, F. (2013). Advanced binderless board-like green nanocomposites from unbarked cotton stalks and mechanism of self-bonding. *Cellulose*, 20(3), 1453–1457. <https://doi.org/10.1007/s10570-013-9911-9>
- Fahmy, Y. (1982). Pyrolysis of agricultural residues. I. Prospects of lignocellulose pyrolysis for producing chemicals and energy sources. *Cellulose chemistry and technology*, 16(January 1982), 347–355.
- Fahmy, Y., & El-Shinnawy, N. (1975). Saccharification of cotton stalks. *Research and industry*, 20(January 1975), 7–10.
- Fahmy, Y., Fahmy, T. Y. A., Mobarak, F., El-Sakhawy, M., & Fadl, M. H. (2017). Agricultural Residues (Wastes) for Manufacture of Paper, Board, and Miscellaneous Products: Background Overview and Future Prospects. *International Journal of ChemTech Research*, 10(2), 424–448. <https://doi.org/10.5281/zenodo.546735>
- Fahmy, Y., Mobarak, F., & Schweers, W. (1982). Pyrolysis of agricultural residues. II. Yield and chemical composition of tars and oils produced from cotton stalks, and assessment of lignin structure. *Cellulose chemistry and technology*, 16(January 1982), 453–459.
- Forster-Carneiro, T., Pérez, M., & Romero, L. I. (2008). Thermophilic anaerobic digestion of source-sorted organic fraction of municipal solid waste. *Bioresource Technology*, 99(15), 6763–6770. <https://doi.org/10.1016/j.biortech.2008.01.052>
- Gallardo, A., Edo-Alcón, N., Carlos, M., & Renau, M. (2016). The determination of waste generation and composition as an essential tool to improve the waste management plan of a university. *Waste Management*, 53, 3–11. <https://doi.org/10.1016/j.wasman.2016.04.013>
- Ikhlayel, M. (2018). Development of management systems for sustainable municipal solid waste in developing countries: A systematic life cycle thinking approach. *Journal of Cleaner Production*, 180, 571–586. <https://doi.org/10.1016/j.jclepro.2018.01.057>
- Kebaili, F. K., Baziz-berkani, A., Aouissi, H. A., & Mihai, F. (2022). Characterization and Planning of Household Waste Management : A Case Study from the MENA Region, 1–13.

- Kelly, T. C., Mason, I. G., Leiss, M. W., & Ganesh, S. (2006). University community responses to on-campus resource recycling. *Resources, Conservation and Recycling*, 47(1), 42–55. <https://doi.org/10.1016/j.resconrec.2005.10.002>
- Kubeyinje, B., Centre, N., Energy, F., & Akingba, O. (2022a). Analysis for investigation of characterization and composition, (June).
- Kubeyinje, B., Centre, N., Energy, F., Akingba, O., Haider, S., Campus, V., et al. (2022b). Household solid waste management practices and perceptions among residents in the East Coast of Malaysia. *BMC Public Health*, 12(June), 1–20. <https://doi.org/10.1186/s12889-021-12274-7>
- Lukman, R., Tiwary, A., & Azapagic, A. (2009). Resources. *Conservation and Recycling towards Greening a University Campus : The Case of the University of Maribor, Slovenia*, 53, 639–644. <https://doi.org/10.1016/j.resconrec.2009.04.014>
- Masebinu, S. O., Akinlabi, E. T., Muzenda, E., Aboyade, A. O., & Mbohwa, C. (2018). Experimental and feasibility assessment of biogas production by anaerobic digestion of fruit and vegetable waste from Joburg Market. *Waste Management*, 75, 236–250. <https://doi.org/10.1016/j.wasman.2018.02.011>
- Mason, I. G., Brooking, A. K., & Oberender, A. (2003). *Implementation of a Zero Waste Program at a University Campus*, 38, 257–269. [https://doi.org/10.1016/S0921-3449\(02\)00147-7](https://doi.org/10.1016/S0921-3449(02)00147-7)
- Mason, I. G., Oberender, A., & Brooking, A. K. (2004). Source separation and potential re-use of resource residuals at a university campus. *Resources, Conservation and Recycling*, 40(2), 155–172. [https://doi.org/10.1016/S0921-3449\(03\)00068-5](https://doi.org/10.1016/S0921-3449(03)00068-5)
- Mobarak, F., Fahmy, Y., & Schweers, W. (1982a). Production of phenols and charcoal from bagasse by a rapid continuous pyrolysis process. *Wood Science and Technology*, 16(1), 59–66. <https://doi.org/10.1007/BF00351374>
- Mobarak, F. (1983). Rapid Continuous Pyrolysis of Cotton Stalks for Charcoal Production. *Holz-forschung*, 37(5), 251–254. <https://doi.org/10.1515/hfsg.1983.37.5.251>
- Mobarak, F., Fahmy, Y., & Augustin, H. (1982b). Binderless lignocellulose composite from bagasse and mechanism of self-bonding. *Holzforschung*, 36(3), 131–136. <https://doi.org/10.1515/hfsg.1982.36.3.131>
- Noble Research Institute. (n.d.). Back to Basics: The Roles of N, P, K and Their Sources. <https://www.noble.org/news/publications/ag-news-and-views/2007/january/back-to-basics-the-roles-of-n-p-k-and-their-sources/>
- Ogwueleka, T. C. (2013). Survey of household waste composition and quantities in Abuja, Nigeria. *Resources, Conservation and Recycling*, 77, 52–60. <https://doi.org/10.1016/j.resconrec.2013.05.011>
- Ojolo, S. J., Dinrifo, R. R., & Adesuyi, K. B. (2007). Comparative study of biogas production from five substrates. *Advanced Materials Research*, 18–19, 519–525. <https://doi.org/10.4028/www.scientific.net/amr.18-19.519>
- Meena, P. K., Sumit Sharma, A. P. (2022). Evaluation of In-House Compact Biogas Plant Thereby Testing Four-Stroke Single-Cylinder Diesel Engine. In *Introduction to Artificial Intelligence for Renewable Energy and Climate* (pp. 277–343). Scrivener Publishing.
- Romero, H. I., Vega, C., Feijóo, V., Villacreses, D., & Sarmiento, C. (2020). ScienceDirect Methane production through anaerobic co-digestion of tropical fruit biomass and urban solid waste. *Energy Reports*, 6, 351–357. <https://doi.org/10.1016/j.egy.2020.11.170>
- Rubio-Romero, J. C., Arjona-Jiménez, R., & López-Arquillos, A. (2013). Profitability analysis of biogas recovery in Municipal Solid Waste landfills. *Journal of Cleaner Production*, 55, 84–91. <https://doi.org/10.1016/j.jclepro.2012.12.024>
- Sagagi, B., Garba, B., & Usman, N. (2010). Studies on biogas production from fruits and vegetable waste. *Bayero Journal of Pure and Applied Sciences*, 2(1), 115–118. <https://doi.org/10.4314/bajopas.v2i1.58513>
- Shane, A., & Gheewala, S. H. (2017). Missed environmental benefits of biogas production in Zambia. *Journal of Cleaner Production*, 142, 1200–1209. <https://doi.org/10.1016/j.jclepro.2016.07.060>
- Sitorus, B., & Sukandar, & Panjaitan, S. D. (2013). Biogas recovery from anaerobic digestion process of mixed fruit—vegetable wastes. *Energy Procedia*, 32, 176–182. <https://doi.org/10.1016/j.egypro.2013.05.023>
- Smyth, D. P., Fredeen, A. L., & Booth, A. L. (2010). Resources, Conservation and Recycling Reducing solid waste in higher education : The first step towards ‘ greening ’ a university campus. *Resources, Conservation and Recycling*, 54(11), 1007–1016. <https://doi.org/10.1016/j.resconrec.2010.02.008>
- Song, Q., Li, J., & Zeng, X. (2015). Minimizing the increasing solid waste through zero waste strategy. *Journal of Cleaner Production*, 104, 199–210. <https://doi.org/10.1016/j.jclepro.2014.08.027>
- University, D. T. (2019). *National Institutional Ranking Framework, Ministry of Human Resource Development Government of India*. <https://nirfcdn.azureedge.net/2019/pdf/OVERALL/IR-O-U-0098.pdf>




- Vats, N., Khan, A. A., & Ahmad, K. (2019). Effect of substrate ratio on biogas yield for anaerobic co-digestion of fruit vegetable waste & sugarcane bagasse. *Environmental Technology and Innovation*, 13, 331–339. <https://doi.org/10.1016/j.eti.2019.01.003>
- Vijayan, D. S., & Parthiban, D. (2020). Effect of Solid waste based stabilizing material for strengthening of Expansive soil- A review. *Environmental Technology and Innovation*, 20, 101108. <https://doi.org/10.1016/j.eti.2020.101108>
- Zhang, D., Xu, Z., Wang, G., Huda, N., Li, G., & Luo, W. (2020). Insights into characteristics of organic matter during co-biodrying of sewage sludge and kitchen waste under different aeration intensities. *Environmental Technology and Innovation*, 20, 101117. <https://doi.org/10.1016/j.eti.2020.101117>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Article

Circuit Complexity in Z_2 EEFT

Kiran Adhikari ¹, Sayantan Choudhury ^{2,3,4,*} , Sourabh Kumar ^{5,6}, Saptarshi Mandal ^{7,8}, Nilesh Pandey ⁹, Abhishek Roy ¹⁰, Soumya Sarkar ¹¹ , Partha Sarker ¹² and Saadat Salman Shariff ^{13,14} 

- ¹ Department of Physics, RWTH Aachen University, Otto-Blumenthal-Straße, 52074 Aachen, Germany
- ² Centre For Cosmology and Science Popularization (CCSP), SGT University, Delhi-NCR, Gurugram 122505, India
- ³ National Institute of Science Education and Research, Bhubaneswar 752050, India
- ⁴ Homi Bhabha National Institute, Training School Complex, Anushakti Nagar, Mumbai 400085, India
- ⁵ Department of Physics and Astronomy, University of Calgary, Calgary, AB T2N 1N4, Canada
- ⁶ Institute for Quantum Science and Technology, University of Calgary, Calgary, AB T2N 1N4, Canada
- ⁷ Department of Physics, Jadavpur University, Kolkata 700032, India
- ⁸ Department of Physics, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
- ⁹ Department of Applied Physics, Delhi Technological University, Delhi 110042, India
- ¹⁰ Department of Physics, Indian Institute of Technology Jodhpur, Karwar, Jodhpur 342037, India
- ¹¹ National Institute of Technology Karnataka, Mangalore 575025, India
- ¹² Department of Physics, University of Dhaka, Curzon Hall, Dhaka 1000, Bangladesh
- ¹³ Department of Theoretical Physics, University of Madras, Guindy Campus, Chennai 600025, India
- ¹⁴ Department of Physics, Indian Institute of Science and Educational Research, Behrampur 760010, India
- * Correspondence: sayantan_ccsp@sgtuniversity.org or sayanphysicsisi@gmail.com

Abstract: Motivated by recent studies of circuit complexity in weakly interacting scalar field theory, we explore the computation of circuit complexity in Z_2 Even Effective Field Theories (Z_2 EEFTs). We consider a massive free field theory with higher-order Wilsonian operators such as ϕ^4 , ϕ^6 , and ϕ^8 . To facilitate our computation, we regularize the theory by putting it on a lattice. First, we consider a simple case of two oscillators and later generalize the results to N oscillators. This study was carried out for nearly Gaussian states. In our computation, the reference state is an approximately Gaussian unentangled state, and the corresponding target state, calculated from our theory, is an approximately Gaussian entangled state. We compute the complexity using the geometric approach developed by Nielsen, parameterizing the path-ordered unitary transformation and minimizing the geodesic in the space of unitaries. The contribution of higher-order operators to the circuit complexity in our theory is discussed. We also explore the dependency of complexity on other parameters in our theory for various cases.

Keywords: circuit complexity; effective field theory; AdS/CFT correspondence



Citation: Adhikari, K.; Choudhury, S.; Kumar, S.; Mandal, S.; Pandey, N.; Roy, A.; Sarkar, S.; Sarker, P.; Shariff, S.S. Circuit Complexity in Z_2 EEFT. *Symmetry* **2023**, *15*, 31. <https://doi.org/10.3390/sym15010031>

Academic Editor: Abraham A. Ungar

Received: 24 November 2022

Revised: 12 December 2022

Accepted: 14 December 2022

Published: 22 December 2022



Copyright: © 2022 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. Prologue

In recent years, tools and techniques from quantum information have played a vital role in developing new perspectives in areas such as quantum field theory and holography, particularly for AdS/CFT duality. A particular line of study in the context of AdS/CFT correspondence is deciphering the emergence of bulk physics using information from the boundary CFT [1]. It was shown in [2–4] that the codimension-2 extremal surfaces in the AdS are associated with the entanglement entropy (EE) of the boundary CFT. However, in recent years, studies in black hole physics have suggested that EE is not sufficient to capture the complete information, which led Susskind et al. to introduce a new measure known as Quantum Computational Complexity (QCC) [5–13]. In the context of AdS/CFT, the QCC of the dual CFT is proposed to be associated with the properties of the codimension-0 and codimension-1 extremal surfaces. This aroused the study of QCC in QFTs.

The complexity of quantum states has aroused a significant amount of interest not only in the context of holography but across different subfields of physics (from quantum

computing and information to many-body physics), as it appears to be a better measure of information. In [14,15], the notion of circuit complexity was defined and studied for free bosonic field theory, and in [16,17], it was defined and studied for free fermionic field theory. For a weakly interacting field theory, the authors of [18] extended the study to the ϕ^4 theory, where in addition to the study of QCC, its relationship with renormalization group flows was also explored. The growth of complexity in the quantum circuit model was studied in [19]. Circuit complexity was also discussed in the context of chaos, quantum mechanics, and quantum computing in [20–23]. It has been probed in relation to conformal and topological field theories and the Chern–Simmons theory [24–27]. Active study in the context of many-body quantum systems has been also gaining interest in recent years [28]. QCC has been studied in many other contexts. It has been explored extensively in holography [29–57]. The thermodynamic properties of QCC were studied in [58–60]. In addition, various applications and properties of QCC were investigated in [61–84].

In this paper, we extend the work in [18] by including even higher-order Wilsonian operators, which we denote with Z_2 EFT (Even Effective Field Theory). Our theory contains the interaction terms ϕ^4 , ϕ^6 , and ϕ^8 . These are weakly coupled to the free scalar field theory via the coupling constants λ_4 , λ_6 , and λ_8 respectively. The primary motivation for studying QCC in this context is to compute and understand QCC by including higher-order terms. The organization of the paper is as follows. In Section 2, we summarize Nielsen’s method for computing circuit complexity. In Section 3, we briefly discuss the pertinent details of EFT related to our work. In Section 4, we illustrate the computation of QCC for our theory by first giving an example of two coupled oscillators. In Section 5, we generalize the calculation to the N oscillator case. Since we could not observe any analytical expression for the relevant eigenvalues for N oscillators, in Section 6, we resort to numerical computation of the QCC. We plot the corresponding graphs of QCC with the relevant parameters in our theory. We finish up by summarizing and providing possible future prospects for our work.

2. Circuit Complexity and Its Purposes

Computationally, circuit complexity is defined as a measure of the minimum number of elementary operations required by a computer to solve a certain computational problem [85–90]. In quantum computation, a quantum operation is described by a unitary transformation. Therefore, quantum circuit complexity is the length of the optimized circuit that performs this unitary operation. As the size of the input increases, if the complexity grows polynomially, then the problem is called “easy”, but if it grows exponentially, then the problem is called “hard”.

Quantum information-theoretic concepts, such as entanglement, have proven to be helpful in areas other than quantum computing [91–94]. Quantum circuit complexity (QCC) is emerging as one such quantum information-theoretic concept that has the potential to explain phenomena in several areas of quantum physics. However, the lower bounding quantum circuit complexity is an extremely challenging open problem.

For our purpose, we will consider the geometric approach to computing quantum circuit complexity developed by Nielsen et al. [85,87]. The prime reason to consider a geometric approach is that it is much easier to minimize a smooth function in a smooth space than to minimize an arbitrary function in a discrete space. Since the unitaries are continuous, this method of optimization suits our needs well. Interestingly, this approach allows us to formulate the optimal circuit-finding problem in the language of the Hamiltonian control problem, for which a mathematical method called the calculus of variations can be employed to find the minima. Another reason is that this method is similar to the general Lagrangian formalism, where the motion of the test particle is obtained from minimizing a global functional. For example, in general relativity, test particles move along geodesics of spacetime described by the following geodesic equation:

$$\frac{d^2 x^j}{dt^2} + \Gamma_{kl}^j \frac{dx^k}{dt} \frac{dx^l}{dt} = 0$$

where x^j represents the coordinates for the position on the manifold and Γ_{kl}^j represents the Christoffel symbols given by the geometry of the spacetime. Thus, the problem of finding an optimal quantum circuit is related to “freely falling” along the minimal geodesic curve connecting the identity to the desired operation, and the path is given by the “local shape” of the manifold. If we have information about the local velocity and the geometry, then it is possible to predict the rest of the path. In this regard, geometric analysis of quantum computation is quite powerful, as it allows one to design the rest of the shortest quantum circuit with information about only part of it.

2.1. Main Mathematical Ideas

Our goal is to understand how difficult it is to implement an arbitrary unitary operation \mathbb{U} generated by a time-dependent Hamiltonian $H(t)$:

$$\mathbb{U}(s) = \overleftarrow{\mathcal{P}} \exp \left[-i \int_0^s ds' H(s') \right] \quad (1)$$

where $\overleftarrow{\mathcal{P}}$ is the path-ordering operator and the space of the circuits is parameterized by s . The path-ordering operator $\overleftarrow{\mathcal{P}}$ is the same as the time-ordering operator, which indicates that the circuit runs from right to left. We can expand the Hamiltonian $H(s)$ as follows:

$$H(s) = \sum_I Y^I(s) M_I \quad (2)$$

where M_I represents the generalized Pauli matrices and the coefficient $Y^I(s)$ represents the control functions that tell us the gate to be applied at particular values of s .

The Schrödinger equation $d\mathbb{U}/dt = -iH\mathbb{U}$ describes the evolution of the unitary operation:

$$\frac{d\mathbb{U}(s)}{ds} = -iY(s)^I M_I \mathbb{U}(s) \quad (3)$$

where at the final time t_f , $\mathbb{U}(t_f) = \mathbb{U}$.

We can impose a cost function $F(\mathbb{U}, \dot{\mathbb{U}})$ on the Hamiltonian control $H(t)$ which will tell us how difficult it is to apply a specific unitary operation \mathbb{U} . One can then define a Riemannian geometry in the space of the unitary operations with this cost function. Then, the problem of finding an optimal control function is translated to the problem of finding the minimal geodesic in this geometry, and we can define a notion of distance in $SU(2^n)$. For this, we have to define a curve \mathbb{U} between the identity operation I and the desired unitary \mathbb{U} , which is a smooth function $\mathbb{U} : [0, t_f] \rightarrow SU(2^n)$ such that $\mathbb{U}(0) = I$ and $\mathbb{U}(t_f) = \mathbb{U}$. The length of this curve is defined as

$$d([\mathbb{U}]) = \int_0^{t_f} dt F(\mathbb{U}, \dot{\mathbb{U}}) \quad (4)$$

This length $d([\mathbb{U}])$ gives the total cost of synthesizing the Hamiltonian that describes the motion along the curve. In particular, the distance $d(I, \mathbb{U})$ is also a lower bound on the number of one- and two-qubit quantum gates necessary to exactly simulate \mathbb{U} . The proof is available in the original papers of Nielsen [85]. Therefore, one can also consider the distance $d([\mathbb{U}])$ as an alternative description of the complexity.

The cost function F has to satisfy certain properties, such as continuity, positivity, positive homogeneity, and triangle inequality [77]. If we also demand F to be smooth (i.e., $F \in C^\infty$), then the manifold is referred to as the Finsler manifold. Since the field of differential geometry is relatively mature, we hope that borrowing tools from differential geometry can provide a unique perspective on quantum complexity.

In the literature, there are several alternative definitions of the cost function $F(\mathbb{U}, v)$. Some of them are

$$\begin{aligned} F_1(\mathbb{U}, Y) &= \sum_I |Y^I| \\ F_p(\mathbb{U}, Y) &= \sum_I p_I |Y^I| \\ F_2(\mathbb{U}, Y) &= \sqrt{\sum_I |Y^I|^2} \\ F_q(\mathbb{U}, Y) &= \sqrt{\sum_I q_I |Y^I|^2} \end{aligned} \quad (5)$$

where F_1 , the linear cost functional measure, is the concept closest to the classical concept of counting gates, while F_2 , the quadratic cost functional, can be understood as the proper distance in the manifold. F_p is similar to F_1 but with penalty parameters p_I used to favor certain directions over others.

In Figure 1 the left figure represents a unitary transformation from a reference state to a target state using quantum gates, and the right figure represents geometrizing the problem of calculating the minimum number of gates representing the transformation.

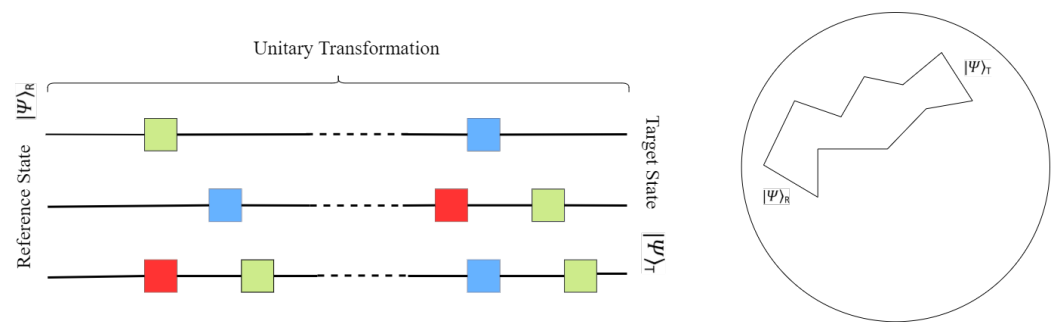


Figure 1. The left figure represents a unitary transformation from a reference state to a target state using quantum gates (square blocks), and the right figure represents geometrizing the problem of calculating the minimum number of gates representing the transformation.

2.2. Geometric Algorithm to Compute Circuit Complexity

We will now describe the algorithm for computing the circuit complexity. These algorithms are not rigorously proven, but from an operational point, these general steps are implemented to calculate the circuit complexity:

1. Give the Hamiltonian corresponding to a particular physical system;
2. Specify the reference state $|\psi\rangle_R$, the target state $|\psi\rangle_T$, and the unitary operation \mathbb{U} that takes the former to the latter, where $|\psi\rangle_T = \mathbb{U}|\psi\rangle_R$;
3. Now, we need to choose some set of elementary gates $Q_{ab} = \exp[\epsilon M_{ab}]$, where M_I represents the generators of the group corresponding to the choice of gates and ϵ is a controllable parameter. For simplicity, we often choose generators satisfying $\text{Tr}[M_I M_J^T] = \delta_{IJ}$.
4. With the basis of generators M_I , we parametrize the unitary operation \mathbb{U} as $\mathbb{U}(s)$;
5. The velocity component $Y^I(s)$ can be explicitly computed using

$$Y^I(s) M_I = i(\partial_s \mathbb{U}(s)) \mathbb{U}^{-1}(s) \rightarrow Y^I(s) = \frac{1}{\text{Tr}[M^I (M^I)^T]} \text{Tr} \left[\partial_s \mathbb{U}(s) \mathbb{U}^{-1}(s) (M^I)^T \right] \quad (6)$$

For generators obeying $\text{Tr}[M_I M_J^T] = \delta_{IJ}$, $Y^I(s)$ reduces to

$$Y^I(s) = \text{Tr}[i(\partial_s \mathbb{U}(s)) \mathbb{U}^{-1}(s) M_I^T] \quad (7)$$

The right invariant metric in the space is given by

$$ds^2 = G_{IJ} Y^I Y^J \quad (8)$$

where G_{IJ} gives the penalty parameters. If $G_{IJ} = \delta_{IJ}$ (i.e., assigning an equal cost to every choice of gate), and having an extra condition $\text{Tr}[M_I M_J^T] = \delta_{IJ}$, we obtain a metric of the reduced simple form

$$ds^2 = \delta_{IJ} \text{Tr}[i(\partial_s U(s)) U^{-1}(s) M_I^T] \text{Tr}[i(\partial_s U(s)) U^{-1}(s) M_J^T] \quad (9)$$

6. The general form of the circuit complexity would be

$$\mathcal{C}[\mathbb{U}] = \int_0^1 ds \sqrt{G_{IJ} Y^I(s) Y^J(s)} \quad (10)$$

The circuit complexity for the F_2 metric (i.e., $G_{IJ} = \delta_{IJ}$) is then

$$\mathcal{C}[\mathbb{U}] = \int_0^1 ds \sqrt{g_{ij} \dot{x}^i \dot{x}^j} \quad (11)$$

7. From the boundary conditions of the evolution of unitary operations, we can compute the geodesic path and geodesic length. This length then gives a measure of circuit complexity.

In the literature, circuit complexity, using this geometric approach, is computed mostly for Gaussian wave functions because of its simpler structure compared with non-Gaussian wave functions. A Gaussian wave function can be represented as follows:

$$\psi \approx \exp \left[-\frac{1}{2} v_a A(s)_{ab} v_b \right], \text{ where } v = \{x_a, x_b\} \quad (12)$$

where x_a and x_b are the bases of vector v . If we can simultaneously diagonalize the reference and target states, then a common pattern observed in the complexity is that it will be given by some function of the ratio of the eigenvalues of $A(s=0)$ and $A(s=1)$. Here, $A(s=0)$ represents the reference state, and $A(s=1)$ represents the target state.

We would like to mention that our approach to computing complexity is based on Nielsen's geometric approach, which suffers from ambiguity in choosing the elementary quantum gates and states. However, these choices of our gates significantly simplify the calculation. Furthermore, the previous works on complexity in QFT and interacting QFT [14,18], using similar quantum gates to ours, have been connected to a holographic proposal, which is the original motivation to study quantum circuit complexity in QFT. Recently, Krylov complexity has been proposed as a tool for studying operator growth and associated quantum chaos [95–103]. Contrary to Nielsen's geometric approach, the Krylov complexity is independent of such arbitrary choices, making it a good candidate for complexity in QFT and holography. However, Krylov complexity does not have a good operational meaning, such as in Nielsen's geometric measure. Nielsen's measure not only gives the state complexity but also gives us a method of constructing an optimal quantum circuit. This feature makes it more appealing than the Krylov complexity. In the future, we would like to study the Krylov complexity for our case too.

3. Effective Field Theory in a Nutshell

An effective field theory (EFT) is a theory corresponding to the dynamics of a physical system at energies that are smaller than the cutoff energy. EFTs have made a significant impact on several areas of theoretical physics, including condensed matter physics [104], cosmology [105–111], particle physics [112,113], gravity [114,115], and hydrodynamics [116,117]. The idea behind an EFT is that we can compute results without knowing the full theory. In the context of quantum field theory, this implies that using the method of

EFTs, one can study the low energy aspect of the theory without having a full theory in the high energy limit. If the high-energy theory is known, then one can obtain an EFT using the “top-down” approach [118], where one has to eliminate high-energy effects. Using the “bottom-up” approach, one can obtain an EFT if the theory for high energy is not available. Here, one has to impose constraints given by symmetry and “naturalness” on suitable Lagrangians.

The Hamiltonian of our theory is

$$H = \frac{1}{2} \int d^{d-1}x \left[\pi(x)^2 + (\nabla\phi(x))^2 + m^2\phi^2(x) + 2 \sum_{n=2}^4 C_{2n}\phi^{2n}(x) \right] \quad (13)$$

where the coefficients $C_{2n} = 2\hat{\lambda}_{2n}/(2n)!$ are called the “Wilson coefficients” for the \mathcal{Z}_2 EFTs in arbitrary dimensions. These coefficients depend on the scaling of the theory. These coefficients are expected to be functions of the λ s, the cutoffs of our theory, and this functional dependence can be found by solving the renormalization group equations or Callan–Symanzik equations. ϕ^{2n} s are called the “Wilson operators” in \mathcal{Z}_2 EFTs. $\phi^2(x)$ and $\phi^4(x)$ are called “relevant operators of EFTs”, and this theory is renormalizable up to $\phi^4(x)$. Beyond that, all the higher-order even terms, which are $\phi^6(x)$ and $\phi^8(x)$ in our case, are called “non-renormalizable irrelevant operators of \mathcal{Z}_2 EFTs”. However, it should be noted that even though this theory goes up in the “Wilson operator” order, the contributions from those terms decrease gradually. Therefore, it is an infinite convergent series. Building upon this, we go on to compute the circuit complexity in the \mathcal{Z}_2 EFT.

4. Circuit Complexity with $(\hat{\lambda}_4\phi^4 + \hat{\lambda}_6\phi^6 + \hat{\lambda}_8\phi^8)$ Interaction for the Case of Two Harmonic Oscillators

We work with massive scalar field theory and with the even interaction terms ϕ^4 , ϕ^6 , and ϕ^8 , which are weakly coupled to the free field theory via the coupling constants $\hat{\lambda}_4$, $\hat{\lambda}_6$, and $\hat{\lambda}_8$, respectively. The inequality between the coupling constants is $\frac{\hat{\lambda}_4}{4!} > \frac{\hat{\lambda}_6}{6!} > \frac{\hat{\lambda}_8}{8!}$. The Hamiltonian for this scalar field in d spacetime dimensions is

$$H = \frac{1}{2} \int d^{d-1}x \left[\pi(x)^2 + (\nabla\phi(x))^2 + m^2\phi(x)^2 + 2 \sum_{n=2}^4 C_{2n}\phi^{2n}(x) \right] \quad (14)$$

where the mass of the scalar field ϕ is m . We work in the weak coupling regime ($\hat{\lambda} \ll 1$) so that perturbative methods can be used to investigate the theory. The system can be reduced to a chain of harmonic oscillators if we regulate the theory by placing it on a $(d-1)$ dimensional square lattice with lattice spacing δ . We are taking the infinite system in Equation (14) and discretizing it to a finite N oscillator system because if we have an infinite convergent theory and an infinite number of terms in the Hamiltonian, then we do not have the finite symmetries that we are interested in. Therefore, the discretized Hamiltonian becomes

$$H = \frac{1}{2} \sum_{\vec{n}} \left\{ \frac{\pi(\vec{n})^2}{\delta^{d-1}} + \delta^{d-1} \left[\frac{1}{\delta^2} \sum_i (\phi(\vec{n}) - \phi(\vec{n} - \hat{x}_i))^2 + m^2\phi(\vec{n})^2 + \frac{2\hat{\lambda}_4}{4!}\phi(\vec{n})^4 + \frac{2\hat{\lambda}_6}{6!}\phi(\vec{n})^6 + \frac{2\hat{\lambda}_8}{8!}\phi(\vec{n})^8 \right] \right\} \quad (15)$$

where \vec{n} denotes the spatial position vectors of the points on the lattice in d dimensions and \hat{x}_i represents the unit vectors along the lattice. We make the following substitutions to simplify the form of the Hamiltonian:

$$\begin{aligned} X(\vec{n}) &= \delta^{d/2}\phi(\vec{n}) & P(\vec{n}) &= \pi(\vec{n})/\delta^{d/2} & M &= \frac{1}{\delta}, \omega = m, \Omega = \frac{1}{\delta} \\ \lambda_4 &= \frac{\hat{\lambda}_4}{4!}\delta^{-d} & \lambda_6 &= \frac{\hat{\lambda}_6}{6!}\delta^{-2d} & \lambda_8 &= \frac{\hat{\lambda}_8}{8!}\delta^{-3d} \end{aligned}$$

After the substitutions, we obtain

$$H = \sum_{\vec{n}} \left\{ \frac{P(\vec{n})^2}{2M} + \frac{1}{2} M \left[\omega^2 X(\vec{n})^2 + \Omega^2 \sum_i (X(\vec{n}) - X(\vec{n} - \hat{x}_i))^2 + 2 \{ \lambda_4 X(\vec{n})^4 + \lambda_6 X(\vec{n})^6 + \lambda_8 X(\vec{n})^8 \} \right] \right\} \quad (16)$$

We observe that the Hamiltonian obtained is identical to that of an infinite family of coupled anharmonic oscillators. The nearest term interaction comes from the kinetic part, and the self-interactions come from the remaining portion of the Hamiltonian. We start with the simple case of two coupled oscillators and generalize it to the case of N oscillators later in this paper. By setting $M = 1$, the Hamiltonian takes the form

$$H = \frac{1}{2} \left[p_1^2 + p_2^2 + \omega^2 (x_1^2 + x_2^2) + \Omega^2 (x_1 - x_2)^2 + 2 \{ \lambda_4 (x_1^4 + x_2^4) + \lambda_6 (x_1^6 + x_2^6) + \lambda_8 (x_1^8 + x_2^8) \} \right] \quad (17)$$

Now, let us consider the normal mode basis:

$$\begin{aligned} \bar{x}_0 &= \frac{1}{\sqrt{2}}(x_1 + x_2), & \bar{x}_1 &= \frac{1}{\sqrt{2}}(x_1 - x_2), \\ \bar{p}_0 &= \frac{1}{\sqrt{2}}(p_1 + p_2), & \bar{p}_1 &= \frac{1}{\sqrt{2}}(p_1 - p_2) \\ \tilde{\omega}_0^2 &= \omega^2, & \tilde{\omega}_1^2 &= \omega^2 + 2\Omega^2 \end{aligned} \quad (18)$$

In the normal mode basis, the unperturbed Hamiltonian becomes decoupled. Then, the eigenfunctions and eigenvalues for the unperturbed Hamiltonian can be easily solved, which is just a product of the ground-state eigenfunctions of the oscillators in the normal basis:

$$\psi_{n_1, n_2}^0(\bar{x}_0, \bar{x}_1) = \frac{1}{\sqrt{2^{n_1+n_2} n_1! n_2!}} \frac{(\tilde{\omega}_0 \tilde{\omega}_1)^{1/4}}{\sqrt{\pi}} e^{-\frac{1}{2} \tilde{\omega}_0 \bar{x}_0^2 - \frac{1}{2} \tilde{\omega}_1 \bar{x}_1^2} H_{n_1}(\sqrt{\tilde{\omega}_0} \bar{x}_0) H_{n_2}(\sqrt{\tilde{\omega}_1} \bar{x}_1) \quad (19)$$

Here, $H_n(x)$ s denote Hermite polynomials of an order n . The ground state wavefunction with first-order perturbative correction in λ_4 , λ_6 , and λ_8 has the following expression:

$$\psi_{0,0}(\bar{x}_0, \bar{x}_1) = \psi_{0,0}^0(\bar{x}_0, \bar{x}_1) + \lambda_4 \psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_4 + \lambda_6 \psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_6 + \lambda_8 \psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_8 \quad (20)$$

Here, $\psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_4$, $\psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_6$, and $\psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_8$ are the terms representing the first-order perturbative corrections to the ground state wavefunction due to the ϕ^4 , ϕ^6 , and ϕ^8 interactions, respectively, which are as follows:

$$\begin{aligned} \psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_4 &= -\frac{3(\tilde{\omega}_0 + \tilde{\omega}_1)}{4\sqrt{2}\tilde{\omega}_0\tilde{\omega}_1^3} \psi_{0,2}^0 - \frac{\sqrt{3}}{8\sqrt{2}\tilde{\omega}_1^3} \psi_{0,4}^0 - \frac{3(\tilde{\omega}_0 + \tilde{\omega}_1)}{4\sqrt{2}\tilde{\omega}_0^3\tilde{\omega}_1} \psi_{2,0}^0 - \frac{3}{4\tilde{\omega}_0(\tilde{\omega}_0 + \tilde{\omega}_1)\tilde{\omega}_1} \psi_{2,2}^0 \\ &\quad - \frac{\sqrt{3}}{8\sqrt{2}\tilde{\omega}_0^3} \psi_{4,0}^0 \\ \psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_6 &= -\frac{45(\tilde{\omega}_0 + \tilde{\omega}_1)^2}{32\sqrt{2}\tilde{\omega}_0^2\tilde{\omega}_1^4} \psi_{0,2}^0 - \frac{15\sqrt{3}(\tilde{\omega}_0 + \tilde{\omega}_1)}{32\sqrt{2}\tilde{\omega}_0\tilde{\omega}_1^4} \psi_{0,4}^0 - \frac{\sqrt{5}}{16\tilde{\omega}_1^4} \psi_{0,6}^0 - \frac{45(\tilde{\omega}_0 + \tilde{\omega}_1)^2}{32\sqrt{2}\tilde{\omega}_0^4\tilde{\omega}_1^2} \psi_{2,0}^0 \\ &\quad - \frac{45(\tilde{\omega}_0 + \tilde{\omega}_1)}{16\tilde{\omega}_0^2(\tilde{\omega}_0 + \tilde{\omega}_1)\tilde{\omega}_1^2} \psi_{2,2}^0 - \frac{15\sqrt{3}}{16\tilde{\omega}_0(\tilde{\omega}_0 + 2\tilde{\omega}_1)\tilde{\omega}_1^2} \psi_{2,4}^0 - \frac{15\sqrt{3/2}(\tilde{\omega}_0 + \tilde{\omega}_1)}{32\tilde{\omega}_0^4\tilde{\omega}_1} \psi_{4,0}^0 \\ &\quad - \frac{15\sqrt{3}}{16\tilde{\omega}_0^2(2\tilde{\omega}_0 + \tilde{\omega}_1)\tilde{\omega}_1} \psi_{4,2}^0 - \frac{\sqrt{5}}{16\tilde{\omega}_0^4} \psi_{6,0}^0 \end{aligned}$$

$$\begin{aligned}
\psi_{0,0}^1(\bar{x}_0, \bar{x}_1)_8 = & \left(\frac{105\sqrt{2}}{8\tilde{\omega}_0^5} + \frac{315\sqrt{2}}{8\tilde{\omega}_0^4\tilde{\omega}_1} + \frac{315\sqrt{2}}{8\tilde{\omega}_0^3\tilde{\omega}_1^2} + \frac{105\sqrt{2}}{8\tilde{\omega}_0^2\tilde{\omega}_1^3} \right) \psi_{2,0}^0 + \left(\frac{105\sqrt{2}}{8\tilde{\omega}_1^5} + \frac{105\sqrt{2}}{8\tilde{\omega}_0^3\tilde{\omega}_1^2} + \frac{315\sqrt{2}}{8\tilde{\omega}_0^3\tilde{\omega}_1^2} \right. \\
& + \left. \frac{315\sqrt{2}}{8\tilde{\omega}_1^4\tilde{\omega}_0} \right) \psi_{0,2}^0 + \left(\frac{315}{4\tilde{\omega}_0^3\tilde{\omega}_1(\tilde{\omega}_0 + \tilde{\omega}_1)} + \frac{315}{2\tilde{\omega}_0^2\tilde{\omega}_1^2(\tilde{\omega}_0 + \tilde{\omega}_1)} + \frac{315}{4\tilde{\omega}_1^3\tilde{\omega}_0(\tilde{\omega}_0 + \tilde{\omega}_1)} \right) \\
& * \psi_{2,2}^0 + \left(\frac{105\sqrt{6}}{16\tilde{\omega}_0^5} + \frac{105\sqrt{6}}{8\tilde{\omega}_0^4\tilde{\omega}_1} + \frac{105\sqrt{6}}{16\tilde{\omega}_0^3\tilde{\omega}_1^2} \right) \psi_{4,0}^0 + \left(\frac{105\sqrt{6}}{16\tilde{\omega}_1^5} + \frac{105\sqrt{6}}{8\tilde{\omega}_1^4\tilde{\omega}_0} + \frac{105\sqrt{6}}{16\tilde{\omega}_0^2\tilde{\omega}_1^3} \right) \\
& * \psi_{0,4}^0 + \left(\frac{105\sqrt{3}}{2\tilde{\omega}_0^3\tilde{\omega}_1(2\tilde{\omega}_0 + \tilde{\omega}_1)} + \frac{105\sqrt{3}}{2\tilde{\omega}_0^2\tilde{\omega}_1^2(2\tilde{\omega}_0 + \tilde{\omega}_1)} \right) \psi_{4,2}^0 + \left(\frac{105\sqrt{3}}{2\tilde{\omega}_1^3\tilde{\omega}_0(2\tilde{\omega}_1 + \tilde{\omega}_0)} \right. \\
& + \left. \frac{105\sqrt{3}}{2\tilde{\omega}_0^2\tilde{\omega}_1^2(\tilde{\omega}_0 + 2\tilde{\omega}_1)} \right) \psi_{2,4}^0 + \frac{105}{4\tilde{\omega}_0^2\tilde{\omega}_1^2(\tilde{\omega}_0 + \tilde{\omega}_1)} \psi_{4,4}^0 + \left(\frac{7\sqrt{5}}{2\tilde{\omega}_0^5} + \frac{7\sqrt{5}}{2\tilde{\omega}_0^4\tilde{\omega}_1} \right) \psi_{6,0}^0 + \\
& \left(\frac{7\sqrt{5}}{2\tilde{\omega}_1^5} + \frac{7\sqrt{5}}{2\tilde{\omega}_1^4\tilde{\omega}_0} \right) \psi_{0,6}^0 + \frac{21\sqrt{10}}{2\tilde{\omega}_1^3\tilde{\omega}_0(3\tilde{\omega}_1 + \tilde{\omega}_0)} \psi_{2,6}^0 + \frac{21\sqrt{10}}{2\tilde{\omega}_1^3\tilde{\omega}_0(3\tilde{\omega}_1 + \tilde{\omega}_0)} \psi_{2,6}^0 \\
& + \frac{3\sqrt{70}}{\tilde{\omega}_0^5} \psi_{8,0}^0 + \frac{3\sqrt{70}}{\tilde{\omega}_1^5} \psi_{0,8}^0
\end{aligned}$$

We can approximate the total ground state wave function in Equation (20) in exponential form as the values of $\lambda_4, \lambda_6, \lambda_8 \ll 1$:

$$\begin{aligned}
\psi_{0,0}(\bar{x}_0, \bar{x}_1) \approx & \frac{(\tilde{\omega}_0\tilde{\omega}_1)^{1/4}}{\sqrt{\pi}} \exp[\alpha_0] \exp \left[-\frac{1}{2} \left(\alpha_1\bar{x}_0^2 + \alpha_2\bar{x}_1^2 + \alpha_3\bar{x}_0^2\bar{x}_1^2 + \alpha_4\bar{x}_0^4 + \alpha_5\bar{x}_1^4 + \alpha_6\bar{x}_0^4\bar{x}_1^2 + \alpha_7\bar{x}_0^2\bar{x}_1^4 \right. \right. \\
& \left. \left. + \alpha_8\bar{x}_0^6 + \alpha_9\bar{x}_1^6 + \alpha_{10}\bar{x}_0^2\bar{x}_1^6 + \alpha_{11}\bar{x}_0^6\bar{x}_1^2 + \alpha_{12}\bar{x}_0^4\bar{x}_1^4 + \alpha_{13}\bar{x}_0^8 + \alpha_{14}\bar{x}_1^8 \right) \right] \quad (21)
\end{aligned}$$

We shall take $\psi_{0,0}(\bar{x}_0, \bar{x}_1)$ as the general target state wavefunction for calculating complexity in the following sections. The coefficients $\alpha_0, \alpha_1, \alpha_2 \dots \alpha_{14}$ involved in the approximate wavefunction Equation (21) are given in the Table 1.

Table 1. Expression for coefficients $\alpha_0, \alpha_1, \alpha_2 \dots \alpha_{14}$, present in the wavefunction.

α_i	Coefficient of α_i
α_0	$ \begin{aligned} & -2 \left[\frac{9\lambda_4}{32\tilde{\omega}_0^3} + \frac{9\lambda_4}{32\tilde{\omega}_1^3} + \frac{3\lambda_4}{8\tilde{\omega}_0\tilde{\omega}_1^2} + \frac{3\lambda_4}{8\tilde{\omega}_0^2\tilde{\omega}_1} + \frac{3\lambda_4}{4\tilde{\omega}_0(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} + \frac{55\lambda_6}{128\tilde{\omega}_0^4} + \frac{55\lambda_6}{128\tilde{\omega}_1^4} + \frac{135\lambda_6}{128\tilde{\omega}_0\tilde{\omega}_1^3} + \frac{45\lambda_6}{32\tilde{\omega}_0^2\tilde{\omega}_1^2} \right. \\ & - \frac{45\lambda_6}{32\tilde{\omega}_0(-2\tilde{\omega}_0-4\tilde{\omega}_1)\tilde{\omega}_1^2} + \frac{45\lambda_6}{16\tilde{\omega}_0(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1^2} + \frac{135\lambda_6}{128\tilde{\omega}_0^3\tilde{\omega}_1} - \frac{45\lambda_6}{32\tilde{\omega}_0^2(-4\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} + \frac{45\lambda_6}{16\tilde{\omega}_0^2(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} \\ & + \frac{875\lambda_8}{1024\tilde{\omega}_0^5} + \frac{875\lambda_8}{1024\tilde{\omega}_1^5} + \frac{385\lambda_8}{128\tilde{\omega}_0\tilde{\omega}_1^4} + \frac{105\lambda_8}{256\tilde{\omega}_0^2\tilde{\omega}_1^3} + \frac{2625\lambda_8}{256\tilde{\omega}_0^3\tilde{\omega}_1^2} + \frac{385\lambda_8}{128\tilde{\omega}_0^4\tilde{\omega}_1} - \frac{315\lambda_8}{64\tilde{\omega}_0\tilde{\omega}_1^3(\tilde{\omega}_0+\tilde{\omega}_1)} \\ & - \frac{2835\lambda_8}{256\tilde{\omega}_0^2\tilde{\omega}_1^2(\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{315\lambda_8}{64\tilde{\omega}_0^3\tilde{\omega}_1(\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{315\lambda_8}{64\tilde{\omega}_0^2\tilde{\omega}_1^2(2\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{315\lambda_8}{64\tilde{\omega}_0^3\tilde{\omega}_1(2\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{105\lambda_8}{64\tilde{\omega}_0^3\tilde{\omega}_1(3\tilde{\omega}_0+\tilde{\omega}_1)} \\ & \left. + \frac{315\lambda_8}{64\tilde{\omega}_0\tilde{\omega}_1^3(\tilde{\omega}_0+2\tilde{\omega}_1)} + \frac{315\lambda_8}{64\tilde{\omega}_0^2\tilde{\omega}_1^2(\tilde{\omega}_0+2\tilde{\omega}_1)} - \frac{105\lambda_8}{64\tilde{\omega}_0\tilde{\omega}_1^3(\tilde{\omega}_0+3\tilde{\omega}_1)} \right] \end{aligned} $
α_1	$ \begin{aligned} & \omega_0 - 2 \left[\frac{-3\lambda_4}{8\tilde{\omega}_0^2} - \frac{3\lambda_4}{4\tilde{\omega}_0\tilde{\omega}_1} - \frac{3\lambda_4}{2(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} - \frac{15\lambda_6}{32\tilde{\omega}_0^3} - \frac{45\lambda_6}{32\tilde{\omega}_0\tilde{\omega}_1^2} + \frac{45\lambda_6}{16(-2\tilde{\omega}_0-4\tilde{\omega}_1)\tilde{\omega}_1^2} - \frac{45\lambda_6}{8(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1^2} \right. \\ & - \frac{45\lambda_6}{32\tilde{\omega}_0^2\tilde{\omega}_1} + \frac{45\lambda_6}{8\tilde{\omega}_0(-4\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} - \frac{45\lambda_6}{8\tilde{\omega}_0(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} - \frac{105\lambda_8}{128\tilde{\omega}_0^4} - \frac{105\lambda_8}{32\tilde{\omega}_0\tilde{\omega}_1^3} - \frac{315\lambda_8}{64\tilde{\omega}_0^2\tilde{\omega}_1^2} - \frac{105\lambda_8}{32\tilde{\omega}_0^3\tilde{\omega}_1} \\ & + \frac{315\lambda_8}{32\tilde{\omega}_1^3(\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{1575\lambda_8}{64\tilde{\omega}_0\tilde{\omega}_1^2(\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{315\lambda_8}{32\tilde{\omega}_0^2\tilde{\omega}_1(\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{315\lambda_8}{16\tilde{\omega}_0\tilde{\omega}_1^2(2\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{315\lambda_8}{16\tilde{\omega}_0^2\tilde{\omega}_1(2\tilde{\omega}_0+\tilde{\omega}_1)} \\ & \left. + \frac{315\lambda_8}{32\tilde{\omega}_0^2\tilde{\omega}_1(3\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{315\lambda_8}{32\tilde{\omega}_1^3(\tilde{\omega}_0+2\tilde{\omega}_1)} - \frac{315\lambda_8}{32\tilde{\omega}_0\tilde{\omega}_1^2(\tilde{\omega}_0+2\tilde{\omega}_1)} + \frac{105\lambda_8}{32\tilde{\omega}_1^3(\tilde{\omega}_0+3\tilde{\omega}_1)} \right] \end{aligned} $
α_2	$ \begin{aligned} & \omega_1 - 2 \left[\frac{-3\lambda_4}{8\omega_1^2} - \frac{3\lambda_4}{4\omega_0\omega_1} - \frac{3\lambda_4}{2\omega_0(-2\omega_0-2\omega_1)} - \frac{15\lambda_6}{32\omega_1^3} - \frac{45\lambda_6}{32\omega_0^2\omega_1} + \frac{45\lambda_6}{16\omega_0^2(-4\omega_0-2\omega_1)} - \frac{45\lambda_6}{8\omega_0^2(-2\omega_0-2\omega_1)} \right. \\ & - \frac{45\lambda_6}{32\omega_0\omega_1^2} + \frac{45\lambda_6}{8\omega_0(-2\omega_0-4\omega_1)\omega_1} - \frac{45\lambda_6}{8\omega_0(-2\omega_0-2\omega_1)\omega_1} - \frac{105\lambda_8}{128\omega_1^4} - \frac{105\lambda_8}{8\omega_0^3\omega_1} + \frac{315\lambda_8}{64\omega_0^2\omega_1^2} - \frac{105\lambda_8}{32\omega_0\omega_1^3} \\ & + \frac{315\lambda_8}{32\omega_0^3(\omega_0+\omega_1)} + \frac{1575\lambda_8}{64\omega_0^2\omega_1(\omega_0+\omega_1)} + \frac{315\lambda_8}{32\omega_0\omega_1^2(\omega_0+\omega_1)} - \frac{315\lambda_8}{32\omega_0^3(2\omega_0+\omega_1)} - \frac{315\lambda_8}{32\omega_0^2\omega_1(2\omega_0+\omega_1)} + \frac{105\lambda_8}{32\omega_0^3(3\omega_0+\omega_1)} \\ & \left. - \frac{315\lambda_8}{16\omega_0\omega_1^2(\omega_0+2\omega_1)} - \frac{315\lambda_8}{16\omega_0^2\omega_1(\omega_0+2\omega_1)} + \frac{315\lambda_8}{32\omega_0\omega_1^2(\omega_0+3\omega_1)} \right] \end{aligned} $
α_3	$ \begin{aligned} & -2 \left[\frac{3\lambda_4}{-2\tilde{\omega}_0-2\tilde{\omega}_1} - \frac{45\lambda_6}{4\tilde{\omega}_0(-4\tilde{\omega}_0-2\tilde{\omega}_1)} + \frac{45\lambda_6}{4\tilde{\omega}_0(-2\tilde{\omega}_0-2\tilde{\omega}_1)} - \frac{45\lambda_6}{4(-2\tilde{\omega}_0-4\tilde{\omega}_1)\tilde{\omega}_1} + \frac{45\lambda_6}{4(-2\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} \right. \\ & - \frac{315\lambda_8}{16\tilde{\omega}_0^2(\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{315\lambda_8}{16\tilde{\omega}_1^2(\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{945\lambda_8}{16\tilde{\omega}_0\tilde{\omega}_1(\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{315\lambda_8}{8\tilde{\omega}_0^2(2\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{315\lambda_8}{8\tilde{\omega}_0\tilde{\omega}_1(2\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{315\lambda_8}{16\tilde{\omega}_0^2(3\tilde{\omega}_0+\tilde{\omega}_1)} \\ & \left. + \frac{315\lambda_8}{8\tilde{\omega}_1^2(\tilde{\omega}_0+2\tilde{\omega}_1)} + \frac{315\lambda_8}{8\tilde{\omega}_0\tilde{\omega}_1(\tilde{\omega}_0+2\tilde{\omega}_1)} - \frac{315\lambda_8}{16\tilde{\omega}_1^2(\tilde{\omega}_0+3\tilde{\omega}_1)} \right] \end{aligned} $
α_4	$ \begin{aligned} & -2 \left[\frac{-\lambda_4}{8\tilde{\omega}_0} - \frac{5\lambda_6}{32\tilde{\omega}_0^2} - \frac{15\lambda_6}{32\tilde{\omega}_0\tilde{\omega}_1} - \frac{15\lambda_6}{8(-4\tilde{\omega}_0-2\tilde{\omega}_1)\tilde{\omega}_1} - \frac{35\lambda_8}{128\tilde{\omega}_0^3} - \frac{105\lambda_8}{64\tilde{\omega}_0\tilde{\omega}_1^2} - \frac{35\lambda_8}{32\tilde{\omega}_0^2\tilde{\omega}_1} - \frac{105\lambda_8}{64\tilde{\omega}_1^2(\tilde{\omega}_0+\tilde{\omega}_1)} \right. \\ & \left. + \frac{105\lambda_8}{16\tilde{\omega}_1^2(2\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{105\lambda_8}{16\tilde{\omega}_0\tilde{\omega}_1(2\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{105\lambda_8}{16\tilde{\omega}_0\tilde{\omega}_1(3\tilde{\omega}_0+\tilde{\omega}_1)} \right] \end{aligned} $
α_5	$ \begin{aligned} & -2 \left[-\frac{\lambda_4}{8\tilde{\omega}_1} - \frac{15\lambda_6}{8\tilde{\omega}_0(-2\tilde{\omega}_0-4\tilde{\omega}_1)} - \frac{5\lambda_6}{32\tilde{\omega}_1^2} - \frac{15\lambda_6}{32\tilde{\omega}_0\tilde{\omega}_1} - \frac{35\lambda_8}{128\tilde{\omega}_1^3} - \frac{35\lambda_8}{32\tilde{\omega}_0\tilde{\omega}_1^2} - \frac{105\lambda_8}{64\tilde{\omega}_0^2\tilde{\omega}_1} - \frac{105\lambda_8}{64\tilde{\omega}_0^2(\tilde{\omega}_0+\tilde{\omega}_1)} + \right. \\ & \left. \frac{105\lambda_8}{16\tilde{\omega}_0^2(\tilde{\omega}_0+2\tilde{\omega}_1)} + \frac{105\lambda_8}{16\tilde{\omega}_0\tilde{\omega}_1(\tilde{\omega}_0+2\tilde{\omega}_1)} - \frac{105\lambda_8}{16\tilde{\omega}_0\tilde{\omega}_1(\tilde{\omega}_0+3\tilde{\omega}_1)} \right] \end{aligned} $

Table 1. Cont.

α_i	Coefficient of α_i
α_6	$-2 \left[\frac{15\lambda_6}{4(-4\tilde{\omega}_0-2\tilde{\omega}_1)} + \frac{105\lambda_8}{16\tilde{\omega}_1(\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{105\lambda_8}{8\tilde{\omega}_0(2\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{105\lambda_8}{8\tilde{\omega}_1(2\tilde{\omega}_0+\tilde{\omega}_1)} + \frac{105\lambda_8}{8\tilde{\omega}_0(3\tilde{\omega}_0+\tilde{\omega}_1)} \right]$
α_7	$-2 \left[\frac{15\lambda_6}{4(-2\tilde{\omega}_0-4\tilde{\omega}_1)} + \frac{105\lambda_8}{16\tilde{\omega}_0(\tilde{\omega}_0+\tilde{\omega}_1)} - \frac{105\lambda_8}{8\tilde{\omega}_0(\tilde{\omega}_0+2\tilde{\omega}_1)} - \frac{105\lambda_8}{8\tilde{\omega}_1(\tilde{\omega}_0+2\tilde{\omega}_1)} + \frac{105\lambda_8}{8\tilde{\omega}_1(\tilde{\omega}_0+3\tilde{\omega}_1)} \right]$
α_8	$-2 \left[\frac{\lambda_6}{24\tilde{\omega}_0} - \frac{7\lambda_8}{96\tilde{\omega}_0^2} - \frac{7\lambda_8}{24\tilde{\omega}_0\tilde{\omega}_1} + \frac{7\lambda_8}{8\tilde{\omega}_1(3\tilde{\omega}_0+\tilde{\omega}_1)} \right]$
α_9	$-2 \left[\frac{-\lambda_6}{24\tilde{\omega}_1} - \frac{7\lambda_8}{96\tilde{\omega}_1^2} - \frac{7\lambda_8}{24\tilde{\omega}_0\tilde{\omega}_1} + \frac{7\lambda_8}{8\tilde{\omega}_0(\tilde{\omega}_0+3\tilde{\omega}_1)} \right]$
α_{10}	$\frac{7\lambda_8}{2(\tilde{\omega}_0+3\tilde{\omega}_1)}$
α_{11}	$\frac{7\lambda_8}{2(3\tilde{\omega}_0+\tilde{\omega}_1)}$
α_{12}	$\frac{35\lambda_8}{8(\tilde{\omega}_0+\tilde{\omega}_1)}$
α_{13}	$\frac{\lambda_8}{32\tilde{\omega}_0}$
α_{14}	$\frac{\lambda_8}{32\tilde{\omega}_1}$

4.1. Circuit Complexity

We will describe complexity in terms of a quantum circuit model. To calculate the circuit complexity for the two-oscillator system with even interactions up to ϕ^8 , we need to fix our reference state, target state, and a set of elementary gates. We will construct the unitary transformation using these gates. This unitary transformation will take the system from the reference state ($|\psi\rangle_R$) to the target state ($|\psi\rangle_T$) (i.e., $|\psi\rangle_T = U |\psi\rangle_R$). The minimum number of gates needed to construct such a unitary transformation is the complexity of the target state. Since our wave functions are nearly Gaussian, we can consider our space of states as the space of positive quadratic forms. This space can be parameterized as a function of a smooth parameter s as follows:

$$\psi^s(\tilde{x}_0, \tilde{x}_1) = \mathcal{N}^s \exp \left[-\frac{1}{2} \left(v_a A(s)_{ab} v_b \right) \right] \quad (22)$$

Here, \mathcal{N}^s is the normalization constant, and the parameter s runs from 0 to 1. If $s = 1$, then the circuit represents the target state in Equation (21) with $\mathcal{N}^{s=1} = \frac{(\tilde{\omega}_0\tilde{\omega}_1)^{1/4}}{\sqrt{\pi}} \exp[\alpha_0]$, and at $s = 0$, the circuit is in the reference state. The continuous unitary transformation, specified by the s parameter, gives us the target state from the reference state. Writing the states in the form of Equation (22) helps us formulate the matrix version of our problem.

Now, we want to represent the exponent of the wavefunction, which is a polynomial in the matrix form $A(s)$:

$$\psi^{s=0}(x_1, x_2) = \mathcal{N}^{s=0} \exp \left[-\frac{\omega_{ref}}{2} (x_1^2 + x_2^2 + \lambda_0^4 (x_1^4 + x_2^4) + \lambda_0^6 (x_1^6 + x_2^6) + \lambda_0^8 (x_1^8 + x_2^8)) \right] \quad (23)$$

Here λ_0^4 , λ_0^6 , and λ_0^8 are the initial coupling constants for ϕ^4 , ϕ^6 , and ϕ^8 respectively. By transforming them into the normal coordinates, we obtain

$$\begin{aligned} \psi^{s=0}(\bar{x}_0, \bar{x}_1) = \mathcal{N}^{s=0} \exp \left[-\frac{\tilde{\omega}_{ref}}{2} (\bar{x}_0^2 + \bar{x}_1^2 + \frac{\lambda_4}{2} (\bar{x}_0^4 + \bar{x}_1^4 + 6\bar{x}_0^2 \bar{x}_1^2) + \frac{\lambda_6}{4} (\bar{x}_0^6 + \bar{x}_1^6 + 15\bar{x}_0^4 \bar{x}_1^2 \right. \\ \left. + 15\bar{x}_1^4 \bar{x}_0^2) + \frac{\lambda_8}{8} (\bar{x}_0^8 + \bar{x}_1^8 + 28\bar{x}_0^6 \bar{x}_1^2 + 28\bar{x}_0^2 \bar{x}_1^6 + 28\bar{x}_0^4 \bar{x}_1^4)) \right] \end{aligned} \quad (24)$$

We represent the exponent of the reference state shown above in a block diagonal matrix form as follows:

$$A(s=0) = \begin{pmatrix} A_1^0 & 0 & 0 & 0 \\ 0 & A_2^0 & 0 & 0 \\ 0 & 0 & A_3^0 & 0 \\ 0 & 0 & 0 & A_4^0 \end{pmatrix}_{14 \times 14} \quad (25)$$

The basis chosen for this representation is

$$\vec{v} = \{ \bar{x}_0, \bar{x}_1, \bar{x}_0 \bar{x}_1, \bar{x}_0^2, \bar{x}_1^2, \bar{x}_0^2 \bar{x}_1, \bar{x}_0 \bar{x}_1^2, \bar{x}_0^3, \bar{x}_1^3, \bar{x}_0 \bar{x}_1^3, \bar{x}_0^3 \bar{x}_1, \bar{x}_0^2 \bar{x}_1^2, \bar{x}_0^4, \bar{x}_1^4 \} \quad (26)$$

We need to ensure that the determinants of the $A(s=0)$ and $A(s=1)$ matrices are positive so that the wavefunction remains square-integrable everywhere. It should be noted that the matrix elements of A (i.e., $A_1^0 - A_4^0$) are matrices themselves as shown below:

$$\begin{aligned} A_1^0 &= \begin{pmatrix} \tilde{\omega}_{ref} & 0 \\ 0 & \tilde{\omega}_{ref} \end{pmatrix} & A_2^0 &= \lambda_0^4 \tilde{\omega}_{ref} \begin{pmatrix} b & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2}(3-b) \\ 0 & \frac{1}{2}(3-b) & \frac{1}{2} \end{pmatrix} \\ A_3^0 &= \tilde{\omega}_{ref} \lambda_0^6 \begin{pmatrix} \frac{p}{2} & 0 & 0 & \frac{1}{8}(15-2k) \\ 0 & k & \frac{1}{8}(15-2p) & 0 \\ 0 & \frac{1}{8}(15-2p) & \frac{1}{4} & 0 \\ \frac{1}{8}(15-2k) & 0 & 0 & \frac{1}{4} \end{pmatrix} \end{aligned}$$

$$A_4^0 = \tilde{\omega}_{ref} \lambda_0^8 \begin{pmatrix} \frac{1}{8} & \frac{1}{4}(\frac{35}{4} - e) & 0 & 0 & 0 \\ \frac{1}{4}(\frac{35}{4} - e) & \frac{1}{8} & 0 & 0 & 0 \\ 0 & 0 & e & \frac{1}{16}(1 - c) & \frac{1}{16}(1 - d) \\ 0 & 0 & \frac{1}{16}(1 - c) & \frac{7}{2} & \frac{1}{4}(\frac{35}{4} - e) \\ 0 & 0 & \frac{1}{16}(1 - d) & \frac{1}{4}(\frac{35}{4} - e) & \frac{7}{2} \end{pmatrix}$$

We have introduced a few parameters (b, p, k, c, d , and e) to ensure that the determinant of each block diagonal matrix is positive definite. Because we are considering higher even interactions, it is necessary to consider various quadratic and other higher-order terms. To find the positive determinant of the A_2^0 block, the value of b must be in the range $2 < b < 4$. To eliminate the off-diagonal components, we set $b = 3$, as it would give the minimum line element. In the A_3^0 block, we fix $k = \frac{15}{2}$, and the determinant becomes

$$\text{Det}(A_3^0) = -\frac{1}{512}p \left(221 + 4(-15 + p)p \omega_{ref}^4 \lambda_6^4 \right)$$

We set p as $15/2$ in the range $\frac{13}{2} < p < \frac{17}{2}$ to satisfy the condition $\text{Det}(A_3^0) > 0$. Similarly, to ensure that the determinant of the A_4^0 block is positive and the line element is at its minimum, we set $c = d = 1$ and $e = 35/4$.

Using the same basis as that mentioned in Equation (26), the target state matrix $A(s = 1)$ can be written as another 14×14 matrix:

$$A(s = 1) = \begin{pmatrix} A_1^1 & 0 & 0 & 0 \\ 0 & A_2^1 & 0 & 0 \\ 0 & 0 & A_3^1 & 0 \\ 0 & 0 & 0 & A_4^1 \end{pmatrix}_{14 \times 14} \quad (27)$$

where we have the following block diagonal entries:

$$A_1^1 = \begin{pmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{pmatrix} \quad A_2^1 = \begin{pmatrix} \tilde{b}\alpha_5 & 0 & 0 \\ 0 & \alpha_3 & \frac{1}{2}(1 - \tilde{b})\alpha_5 \\ 0 & \frac{1}{2}(1 - \tilde{b})\alpha_5 & \alpha_4 \end{pmatrix}$$

$$A_3^1 = \begin{pmatrix} \tilde{p}\alpha_6 & 0 & 0 & \frac{1}{2}(1 - \tilde{k})\alpha_7 \\ 0 & \tilde{k}\alpha_7 & \frac{1}{2}(1 - \tilde{p})\alpha_6 & \\ 0 & \frac{1}{2}(1 - \tilde{p})\alpha_6 & \alpha_8 & 0 \\ \frac{1}{2}(1 - \tilde{k})\alpha_7 & 0 & 0 & \alpha_9 \end{pmatrix}$$

$$A_4^1 = \begin{pmatrix} \tilde{d}\alpha_{10} & \frac{1}{4}(1-\tilde{e})\alpha_{12} & 0 & 0 & 0 \\ \frac{1}{4}(1-\tilde{e})\alpha_{12} & \tilde{c}\alpha_{11} & 0 & 0 & 0 \\ 0 & 0 & \tilde{e}\alpha_{12} & \frac{1}{2}(1-\tilde{c})\alpha_{11} & \frac{1}{2}(1-\tilde{d})\alpha_{10} \\ 0 & 0 & \frac{1}{2}(1-\tilde{c})\alpha_{11} & \alpha_{13} & \frac{1}{4}(1-\tilde{e})\alpha_{12} \\ 0 & 0 & \frac{1}{2}(1-\tilde{d})\alpha_{10} & \frac{1}{4}(1-\tilde{e})\alpha_{12} & \alpha_{14} \end{pmatrix}$$

Here as well, we fix \tilde{k} , \tilde{c} , and \tilde{d} to be one to make the off-diagonal terms zero and keep \tilde{b} , \tilde{p} , and \tilde{e} for the positivity of all the block matrices.

As we are considering a closed quantum system, the reference state evolves into the target state via a certain unitary operator. Now, we represent this as

$$\psi^{s=1}(\tilde{x}_0, \tilde{x}_1) = \mathbb{U}(s=1)\psi^{s=0}(\tilde{x}_0, \tilde{x}_1) \quad (28)$$

We represent the unitary matrix in the following form:

$$\mathbb{U} = \overleftarrow{\mathcal{P}} \exp \left[\int_0^s ds Y^I(s) \mathcal{O}_I \right] \quad (29)$$

We have to enact the operators \mathcal{O}_I in a particular order. The Y^I 's depend on the specific order in which the \mathcal{O}_I 's are acting on the reference state. To find the minimum complexity, we try to have a geometric understanding of this unitary evolution process. Then, we can write the expression in Equation (29) as follows:

$$\mathbb{U} = \overleftarrow{\mathcal{P}} \exp \left[\int_0^s Y^I(s) M_I ds \right] \quad (30)$$

where $(M_I)'_{jk}$ represents the $\text{GL}(14, \mathbb{R})$ generators satisfying

$$\text{Tr} [M_I M_J^T] = \delta_{IJ} \quad (31)$$

Here, I, J runs from 1 to 196. As mentioned above, $A(s=0)$ is the reference state which undergoes a unitary transformation to find the target state $A(s=1)$. It enables us to calculate the boundary conditions that lead us to calculate the complexity functional. Thus, we have

$$A(s=1) = \mathbb{U}(s=1)A(s=0)\mathbb{U}^T(s=1) \quad (32)$$

This leads to the expression

$$Y^I M_I = \partial_s \mathbb{U}(s) \mathbb{U}(s)^{-1} \quad (33)$$

Hence, we obtain

$$Y^I = \frac{1}{\text{Tr} [M^I (M^I)^T]} \text{Tr} \left[\partial_s \mathbb{U}(s) \mathbb{U}^{-1} (M^I)^T \right] \quad (34)$$

Now, the line element can be defined in terms of Y^I 's as follows:

$$\begin{aligned}
 ds^2 &= G_{IJ} dY^I dY^J \\
 &= G_{IJ} \left[\frac{1}{\text{Tr}[M^I (M^I)^T]} \text{Tr} \left[d_s \mathbb{U}(s) \mathbb{U}^{-1} (M^I)^T \right] \right] \left[\frac{1}{\text{Tr}[M^J (M^J)^T]} \text{Tr} \left[d_s \mathbb{U}(s) \mathbb{U}^{-1} (M^J)^T \right] \right]
 \end{aligned} \tag{35}$$

Here, we should mention that dY^I does not denote the total differential for Y^I . When observing the structure of the matrix A , we find that $\mathbb{U}(s)$ can be considered an element of $\text{GL}(14, \mathbb{R})$ with a positive determinant. Now, we will express the \mathbb{U} matrix with a similar structure to that in the target state matrix, and the unitary matrix contains four block diagonal matrices:

$$\mathbb{U} = \begin{pmatrix} \mathbb{U}_1 & 0 & 0 & 0 \\ 0 & \mathbb{U}_2 & 0 & 0 \\ 0 & 0 & \mathbb{U}_3 & 0 \\ 0 & 0 & 0 & \mathbb{U}_4 \end{pmatrix}_{14 \times 14} \tag{36}$$

where

$$\begin{aligned}
 \mathbb{U}_1 &= \begin{pmatrix} x_0 - x_1 & x_3 - x_2 \\ x_3 + x_2 & x_0 + x_1 \end{pmatrix} & \mathbb{U}_2 &= \begin{pmatrix} \tilde{x}_4 & 0 & 0 \\ 0 & \tilde{x}_5 - \tilde{x}_6 & \tilde{x}_8 - \tilde{x}_7 \\ 0 & \tilde{x}_8 + \tilde{x}_7 & \tilde{x}_5 + \tilde{x}_6 \end{pmatrix} \\
 \mathbb{U}_3 &= \begin{pmatrix} \tilde{x}_9 & 0 & 0 & 0 \\ 0 & \tilde{x}_{10} - \tilde{x}_{11} & \tilde{x}_{13} - \tilde{x}_{12} & 0 \\ 0 & \tilde{x}_{13} + \tilde{x}_{12} & \tilde{x}_{10} + \tilde{x}_{11} & 0 \\ 0 & 0 & 0 & \tilde{x}_{14} \end{pmatrix} & \mathbb{U}_4 &= \begin{pmatrix} \tilde{x}_{15} - \tilde{x}_{16} & \tilde{x}_{18} - \tilde{x}_{17} & 0 & 0 & 0 \\ x_{18} + x_{17} & x_{15} + x_{16} & 0 & 0 & 0 \\ 0 & 0 & \tilde{x}_{19} & 0 & 0 \\ 0 & 0 & 0 & \tilde{x}_{20} - \tilde{x}_{21} & \tilde{x}_{23} - \tilde{x}_{22} \\ 0 & 0 & 0 & \tilde{x}_{23} + \tilde{x}_{22} & \tilde{x}_{20} + \tilde{x}_{21} \end{pmatrix}
 \end{aligned}$$

We have decomposed $\mathbb{U}(s)$ in terms of four block diagonal matrices. First, we note that the quadratic part of the first block is always diagonal, which induces a flat space, and thus we have $x_3 = x_2 = 0$. In the unitary operator \mathbb{U} , we do not allow the off-diagonal terms as in the final state, and only the block diagonal form remains. Thus, if we allow off-diagonal terms, we will have an increased line element, which we do not want. Now, $\text{GL}(2, \mathbb{R})$ can be expressed as $\mathbb{R} \times \text{SL}(2, \mathbb{R})$, and so we observe that our \mathbb{U} has an $\mathbb{R}^{10} \times \text{SL}(2, \mathbb{R})^4$ group

structure. We will parameterize each 2×2 block matrix in \mathbb{U} as performed in [18] (i.e., we will parameterize it as an AdS_3 space):

$$\begin{aligned}
 x_0 &= \exp[y_1] \cosh(\rho_1) & x_1 &= \exp[y_1] \sinh(\rho_1) \\
 \tilde{x}_4 &= \exp[y_2] & x_5 &= \exp[y_3] \cos(\tau_3) \cosh(\rho_3) \\
 \tilde{x}_6 &= \exp[y_3] \sin(\theta_3) \cosh(\rho_3) & \tilde{x}_7 &= \exp[y_3] \sin(\tau_3) \cosh(\rho_3) \\
 \tilde{x}_8 &= \exp[y_3] \cos(\theta_3) \sinh(\rho_3) & \tilde{x}_9 &= \exp[y_4] \\
 \tilde{x}_{10} &= \exp[y_5] \cos(\tau_5) \cosh(\rho_5) & \tilde{x}_{11} &= \exp[y_5] \sin(\theta_5) \sinh(\rho_5) \\
 \tilde{x}_{12} &= \exp[y_5] \sin(\tau_5) \cosh(\rho_5) & \tilde{x}_{13} &= \exp[y_5] \cos(\theta_5) \sinh(\rho_5) \\
 \tilde{x}_{14} &= \exp[y_6] & \tilde{x}_{15} &= \exp[y_7] \cos(\tau_7) \cosh(\rho_7) \\
 \tilde{x}_{16} &= \exp[y_7] \sin(\theta_7) \sinh(\rho_7) & \tilde{x}_{17} &= \exp[y_7] \sin(\tau_7) \cosh(\rho_7) \\
 \tilde{x}_{18} &= \exp[y_7] \cos(\theta_7) \sinh(\rho_7) & \tilde{x}_{19} &= \exp[y_8] \\
 \tilde{x}_{20} &= \exp[y_9] \cos(\tau_9) \cosh(\rho_9) & \tilde{x}_{21} &= \exp[y_9] \sin(\theta_9) \sinh(\rho_9) \\
 \tilde{x}_{22} &= \exp[y_9] \sin(\tau_9) \cosh(\rho_9) & \tilde{x}_{23} &= \exp[y_9] \cos(\theta_9) \sinh(\rho_9)
 \end{aligned} \tag{37}$$

Using these parameters for \mathbb{U} , we can then calculate the infinitesimal line element in Equation (35), which now becomes

$$\begin{aligned}
 ds^2 &= \left[2y_1^2 + y_2^2 + 2y_3^2 + y_4^2 + 2y_5^2 + y_6^2 + 2y_7^2 + y_8^2 + 2y_9^2 + 2 \left(\rho_1^2 + \rho_3^2 \right. \right. \\
 &\quad + \rho_5^2 + \rho_7^2 + \rho_9^2 + \cosh(2\rho_3) \left\{ \cosh^2(\rho_3) \tau_3^2 + \sinh^2(\rho_3) \theta_3^2 \right\} - \sinh^2(2\rho_3) \theta_3 \tau_3 \\
 &\quad + \cosh(2\rho_5) \left\{ \cosh^2(\rho_5) \tau_5^2 + \sinh^2(\rho_5) \theta_5^2 \right\} - \sinh^2(2\rho_5) \theta_5 \tau_5 \\
 &\quad + \cosh(2\rho_7) \left\{ \cosh^2(\rho_7) \tau_7^2 + \sinh^2(\rho_7) \theta_7^2 \right\} - \sinh^2(2\rho_7) \theta_7 \tau_7 \\
 &\quad \left. \left. + \cosh(2\rho_9) \left\{ \cosh^2(\rho_9) \tau_9^2 + \sinh^2(\rho_9) \theta_9^2 \right\} - \sinh^2(2\rho_9) \theta_9 \tau_9 \right) \right]
 \end{aligned} \tag{38}$$

We need to find the shortest path between the reference and the target state in this geometry, described by metric expressed in Equation (38). This shortest path will be the circuit complexity for our problem. For this purpose, we also need to calculate the proper boundary conditions denoting the reference and target states.

4.2. Boundary Conditions for the Geodesic

As we mentioned before, the minimal geodesic will be equivalent to finding the geodesic in the $GL(14, R)$ group manifold. The geodesic can be found by minimizing the following equation on the distance functional:

$$\mathcal{D}(U) = \int_0^1 \sqrt{g_{ij} \dot{x}^i \dot{x}^j} ds \tag{39}$$

The boundary conditions from Equation (32) are

$$y_i(0) = \rho_j(0) = 0 \tag{40}$$

where $i = 1, 2, \dots, 9$ and $j = 1, 3, 5, 7, 9$.

For solving the geodesic equations, we have to find conserved charges using the results of [14], as our metric is $\mathbb{R}^{10} \times \text{SL}(2, \mathbb{R})^4$. Using Equations (40) and (42), we obtain

$$y_i(s) = y_i(1)s \quad \rho_j(s) = \rho_j(1)s \tag{41}$$

where, $i = 1, 2, \dots, 9$ and $j = 1, 3, 5, 7, 9$:

$$\begin{aligned}
2(y_1(1) - \rho_1(1)) &= \ln \left[\frac{\alpha_1}{\tilde{\omega}_{ref}} \right] & 2(y_1(1) + \rho_1(1)) &= \ln \left[\frac{\alpha_2}{\tilde{\omega}_{ref}} \right] \\
2y_2(1) &= \ln \left[\frac{\tilde{b}\alpha_5}{3\tilde{\omega}_{ref}\lambda_4} \right] & 2y_3(1) &= \ln \left[\frac{\sqrt{4\alpha_3\alpha_4 - (1 - \tilde{b})^2\alpha_5^2}}{\tilde{\omega}_{ref}\lambda_4} \right] \\
2\rho_3(1) &= \cosh^{-1} \left[\frac{\alpha_3 + \alpha_4}{\sqrt{4\alpha_3\alpha_4 - (1 - \tilde{b})^2\alpha_5^2}} \right] & 2y_4(1) &= \ln \left[\frac{4\tilde{p}\alpha_6}{15\omega_{ref}\lambda_6} \right] \\
2y_5(1) &= \ln \left[\frac{\sqrt{16\alpha_7\alpha_8 - 4(1 - \tilde{p})^2\alpha_6^2}}{\tilde{\omega}_{ref}\lambda_6} \right] & 2y_6(1) &= \ln \left[\frac{4\alpha_9}{\tilde{\omega}_{ref}\lambda_6} \right] \\
2\rho_5(1) &= \cosh^{-1} \left[\frac{2(\alpha_7 + \alpha_8)}{\sqrt{16\alpha_7\alpha_8 - 4(1 - \tilde{p})^2\alpha_6^2}} \right] & 2y_7(1) &= \ln \left[\frac{\sqrt{64\alpha_{10}\alpha_{11} - 4(1 - \tilde{e})^2\alpha_{12}^2}}{\tilde{\omega}_{ref}\lambda_8} \right] \\
2\rho_7(1) &= \cosh^{-1} \left[\frac{\alpha_{10} + \alpha_{11}}{\sqrt{64\alpha_{10}\alpha_{11} - 4(1 - \tilde{e})^2\alpha_{12}^2}} \right] & 2y_8(1) &= \ln \left[\frac{4\tilde{e}\alpha_{12}}{35\tilde{\omega}_{ref}\lambda_6} \right] \\
2\rho_9(1) &= \cosh^{-1} \left[\frac{\alpha_{13} + \alpha_{14}}{\sqrt{4\alpha_{13}\alpha_{14} - ((1 - \tilde{e})^2/4)\alpha_{12}^2}} \right] & 2y_9(1) &= \ln \left[\frac{\sqrt{4\alpha_{13}\alpha_{14} - ((1 - \tilde{e})^2/4)\alpha_{12}^2}}{7\tilde{\omega}_{ref}\lambda_8} \right]
\end{aligned} \tag{42}$$

With the same arguments in [14], we set

$$\tau_j(s) = 0 \quad \theta_j(s) = \theta_{c_j} \tag{43}$$

where $j = 3, 5, 7, 9$ and θ_{c_j} are constants which do not depend on s . Therefore, we have the freedom to choose any constant value of θ_{c_j} here which indicates that it will leave the origin in any direction. (Note: When we are calculating ρ_5 , any arbitrary constant value will not provide us an analytical expression, so we choose θ_5 to be zero to find the simple analytical expression in Equation (42)) By taking into account all of these terms and conditions, we find the complexity functional as follows:

$$\begin{aligned}
\mathcal{D}(U) &= \sqrt{2 \left[\sum_{i=1, \text{odd}}^9 [y_i(1)]^2 + \frac{1}{2} \sum_{i=2, \text{even}}^8 [y_i(1)]^2 + \sum_{j=1, \text{odd}}^9 [\rho_j(1)]^2 \right]} \\
&= \frac{1}{\sqrt{2}} \left(2 \left[\cosh^{-1} \left(\frac{\alpha_3 + \alpha_4}{\sqrt{4\alpha_3\alpha_4 - \alpha_5^2(-1 + \tilde{b})^2}} \right) \right]^2 + 2 \left[\cosh^{-1} \left(\frac{\alpha_{10} + \alpha_{11}}{2\sqrt{16\alpha_{10}\alpha_{11} + (1 - \tilde{e})^2\alpha_{12}^2}} \right) \right]^2 \right. \\
&\quad + 2 \left[\cosh^{-1} \left(\frac{\alpha_{13} + \alpha_{14}}{\sqrt{4\alpha_{13}\alpha_{14} - ((1 - \tilde{e})^2/4)\alpha_{12}^2}} \right) \right]^2 + 2 \left[\cosh^{-1} \left(\frac{2(\alpha_7 + \alpha_8)}{\sqrt{-\alpha_6^2 + 4\alpha_7\alpha_8 + \alpha_6^2\tilde{p}}} \right) \right]^2 \\
&\quad + \frac{1}{2} \left[\ln \frac{\alpha_2}{\alpha_1} \right]^2 + \frac{1}{2} \left[\ln \left(\frac{\alpha_1\alpha_2}{\tilde{\omega}_{ref}^2} \right) \right]^2 + \left[\ln \left(\frac{4\alpha_9}{\lambda_6\tilde{\omega}_{ref}} \right) \right]^2 + 2 \left[\ln \left(\frac{\sqrt{4\alpha_3\alpha_4 - (1 - \tilde{b})^2\alpha_5^2}}{\tilde{\omega}_{ref}\lambda_4} \right) \right]^2 \\
&\quad + 2 \left[\ln \left(\frac{\tilde{b}\alpha_5}{3\lambda_4\tilde{\omega}_{ref}} \right) \right]^2 + 2 \left[\ln \left(\frac{\sqrt{64\alpha_{10}\alpha_{11} - 4(-1 + \tilde{e})^2\alpha_{12}^2}}{\tilde{\omega}_{ref}\lambda_8} \right) \right]^2 + \left[\ln \left(\frac{4\alpha_{12}\tilde{e}}{35\lambda_8\tilde{\omega}_{ref}} \right) \right]^2 \\
&\quad + 2 \left[\ln \left(\frac{\sqrt{4\alpha_{13}\alpha_{14} - ((-1 + \tilde{e})^2/16)\alpha_{12}^2}}{7\tilde{\omega}_{ref}\lambda_8} \right) \right]^2 + 2 \left[\ln \left(\frac{2\sqrt{-\alpha_6^2 + 4\alpha_7\alpha_8 + \alpha_6^2\tilde{p}}}{\tilde{\omega}_{ref}\lambda_6} \right) \right]^2 \\
&\quad \left. + \left[\ln \left(\frac{4\alpha_6\tilde{p}}{15\lambda_6\tilde{\omega}_{ref}} \right) \right]^2 \right)^{\frac{1}{2}}
\end{aligned} \tag{44}$$

which is a straight line, as there is no off-diagonal term for when we set $\tau_i(s)$ to be 0 and $\theta_j(s)$ to be independent of s , according to Equation (41).

For the particular choice of a cost function that we used (i.e., \mathcal{F}_2), the complexity functional is

$$C_2 = \int_{s=0}^1 ds \mathcal{F}_2 \tag{45}$$

As was shown in Equation (44), the complexity functional can be written in terms of some boundary values only. It can also be proven that this functional can just involve the eigenvalues of the reference and target matrix:

$$C_2 = \frac{1}{2} \sqrt{\sum_{i=1}^{14} \log \left[\frac{(\lambda_T)_i}{(\lambda_R)_i} \right]^2} \tag{46}$$

The proof of this expression is explicitly constructed in Appendix B. This result is very crucial, and we exploit this relation to generalize the complexity to N oscillators.

5. Analysis for N Oscillators

To this point, our discussion in this paper has been concerned with two coupled harmonic oscillators involving higher-order interactions. To extend our analysis to effective field theories, we first need to generalize our results to N coupled harmonic oscillators with $(\phi^4 + \phi^6 + \phi^8)$ interaction terms. Then, we will gradually move toward the continuum limit for this problem. With that in mind, we consider the following Hamiltonian:

$$H = \frac{1}{2} \sum_{a=0}^{N-1} [p_a^2 + \omega^2 x_a^2 + \Omega^2 (x_a - x_{a+1})^2 + 2\lambda_4 x_a^4 + 2\lambda_6 x_a^6 + 2\lambda_8 x_a^8] \tag{47}$$

Now, we will assume that the periodic boundary condition is valid on this lattice of N oscillators such that $x_{a+N} = x_a$ (as it allows us to impose translational symmetry and use a Fourier transform for expression in terms of the normal mode coordinates). Then, we perform discrete a Fourier transform for this lattice using

$$x_a = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \exp \left[i \frac{2\pi a}{N} k \right] \tilde{x}_k \quad (48)$$

$$p_a = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \exp \left[i \frac{2\pi a}{N} k \right] \tilde{p}_k \quad (49)$$

Using the above Equations (48) and (49), we can transform the spatial coordinates into normal mode coordinates. The resultant Hamiltonian is then

$$\begin{aligned} H &= \frac{1}{2} \sum_{a=0}^{N-1} [p_a^2 + \omega^2 x_a^2 + \Omega^2 (x_a - x_{a+1})^2 + 2\lambda_4 x_a^4 + 2\lambda_6 x_a^6 + 2\lambda_8 x_a^8] \\ &= \frac{1}{2} \sum_{k=0}^{N-1} \left[|\tilde{p}_k|^2 + \left(\omega^2 + 4\Omega^2 \sin^2 \left(\frac{\pi k}{N} \right) \right) |\tilde{x}_k|^2 \right] + H'_{\phi^4} + H'_{\phi^6} + H'_{\phi^8} \end{aligned} \quad (50)$$

where H'_{ϕ^4} , H'_{ϕ^6} , and H'_{ϕ^8} are the contributions from the ϕ^4 , ϕ^6 , and ϕ^8 interaction terms, respectively. Now, we have

$$H'_{\phi^4} = \frac{\lambda_4}{N} \sum_{k_1, k_2, k_3=0}^{N-1} \tilde{x}_\alpha \tilde{x}_{k_1} \tilde{x}_{k_2} \tilde{x}_{k_3}; \alpha = N - k_1 - k_2 - k_3 \bmod N \quad (51)$$

$$H'_{\phi^6} = \frac{\lambda_6}{N^2} \sum_{k_1, k_2, k_3, k_4, k_5=0}^{N-1} \tilde{x}_\alpha \tilde{x}_{k_1} \tilde{x}_{k_2} \tilde{x}_{k_3} \tilde{x}_{k_4} \tilde{x}_{k_5}; \alpha = \left(N - \sum_{i=1}^5 k_i \right) \bmod N \quad (52)$$

$$H'_{\phi^8} = \frac{\lambda_8}{N^3} \sum_{k_1, k_2, k_3, k_4, k_5, k_6, k_7=0}^{N-1} \tilde{x}_\alpha \tilde{x}_{k_1} \tilde{x}_{k_2} \tilde{x}_{k_3} \tilde{x}_{k_4} \tilde{x}_{k_5} \tilde{x}_{k_6} \tilde{x}_{k_7}; \alpha = \left(N - \sum_{i=1}^7 k_i \right) \bmod N \quad (53)$$

The proof of transformation of the interaction Hamiltonian in a Fourier space is given in Appendix A.

The target state wavefunction is given by

$$\psi_{0,0,\dots,0}(\tilde{x}_0, \dots, \tilde{x}_{N-1}) = \left(\frac{\tilde{\omega}_0 \tilde{\omega}_1 \dots \tilde{\omega}_{N-1}}{\pi^N} \right)^{\frac{1}{4}} \exp \left[-\frac{1}{2} \sum_{k=0}^{N-1} \tilde{\omega}_k \tilde{x}_k^2 + \lambda_4 \psi_4^1 + \lambda_6 \psi_6^1 + \lambda_8 \psi_8^1 \right] \quad (54)$$

where the total perturbation wavefunction ψ^1 is

$$\psi^1 = \lambda_4 \psi_4^1 + \lambda_6 \psi_6^1 + \lambda_8 \psi_8^1 \quad (55)$$

where $\lambda_4 \psi_4^1$, $\lambda_6 \psi_6^1$, and $\lambda_8 \psi_8^1$ are first-order perturbation corrections for the self-interaction terms ϕ^4 , ϕ^6 , and ϕ^8 , respectively.

The expression of ψ_4^1 along with the B terms was taken from [18].

The expression for ψ_4^1 is

$$\begin{aligned} \psi_4^1 = & \sum_{\substack{a=0 \\ 4a \bmod N \equiv 0}}^{N-1} B_1(a) + \sum_{\substack{a,b=0 \\ (2a+2b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} \frac{B_2(a,b)}{2} + \sum_{\substack{a,b=0 \\ (3b+a) \bmod N \equiv 0 \\ a \neq b}}^{N-1} B_3(a,b) \\ & + \sum_{\substack{a,b,c=0 \\ (a+2b+c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} \frac{B_4(a,b,c)}{2} + \sum_{\substack{a,b,c,d=0 \\ (a+b+c+d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{B_5(a,b,c,d)}{24} \end{aligned} \quad (56)$$

The expression for ψ_6^1 is

$$\begin{aligned} \psi_6^1 = & \frac{1}{N^2} \left[\sum_{\substack{a=0 \\ 6a \bmod N \equiv 0}}^{N-1} C_1(a) + \sum_{\substack{a,b=0 \\ (a+5b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} C_2(a,b) \right. \\ & + \sum_{\substack{a,b=0 \\ (3b+3a) \bmod N \equiv 0 \\ a \neq b}}^{N-1} \frac{1}{2} C_3(a,b) + \sum_{\substack{a,b=0 \\ (2a+4b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} C_4(a,b) \\ & + \sum_{\substack{a,b,c=0 \\ (a+b+4c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} \frac{1}{2} C_5(a,b,c) + \sum_{\substack{a,b,c=0 \\ (2a+b+3c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} C_6(a,b,c) \\ & + \sum_{\substack{a,b,c=0 \\ (2a+2b+2c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} \frac{1}{6} C_7(a,b,c) + \sum_{\substack{a,b,c,d=0 \\ (a+b+c+3d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{1}{6} C_8(a,b,c,d) \\ & + \sum_{\substack{a,b,c,d=0 \\ (a+b+2c+2d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{1}{4} C_9(a,b,c,d) + \sum_{\substack{a,b,c,d,e=0 \\ (a+b+c+d+2e) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e}}^{N-1} \frac{1}{4!} C_{10}(a,b,c,d,e) \\ & \left. + \sum_{\substack{a,b,c,d,e,f=0 \\ (a+b+c+d+e+f) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e \neq f}}^{N-1} \frac{1}{6!} C_{11}(a,b,c,d,e,f) \right] \end{aligned} \quad (57)$$

where the terms C_1, C_2, \dots, C_{11} are given according to the Table 2.

Table 2. Expression for the terms $C_1, C_2 \dots C_{11}$.

Expression for C_i Coefficients	
C_1	$\left[\frac{55}{32\tilde{\omega}_a^4} - \frac{15\tilde{x}_a^2}{8\tilde{\omega}_a^3} - \frac{5\tilde{x}_a^4}{8\tilde{\omega}_a^2} - \frac{\tilde{x}_a^6}{6\tilde{\omega}_a} \right]$
C_2	$\left[\frac{-180\tilde{x}_a\tilde{x}_b}{(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+3\tilde{\omega}_b)(\tilde{\omega}_a+5\tilde{\omega}_b)} - \frac{60\tilde{x}_a\tilde{x}_b^3}{(\tilde{\omega}_a+3\tilde{\omega}_b)(\tilde{\omega}_a+5\tilde{\omega}_b)} - \frac{6\tilde{x}_a\tilde{x}_b^5}{\tilde{\omega}_a+5\tilde{\omega}_b} \right]$
C_3	$\left[\frac{-120\tilde{x}_a\tilde{x}_b}{(\tilde{\omega}_a+\tilde{\omega}_b)(3\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+3\tilde{\omega}_b)} - \frac{10\tilde{x}_a^2\tilde{x}_b}{(\tilde{\omega}_a+\tilde{\omega}_b)(3\tilde{\omega}_a+\tilde{\omega}_b)} - \frac{10\tilde{x}_a\tilde{x}_b^3}{(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+3\tilde{\omega}_b)} - \frac{10\tilde{x}_a^3\tilde{x}_b^3}{3(\tilde{\omega}_a+\tilde{\omega}_b)} \right]$
C_4	$\left[\frac{135}{32\tilde{\omega}_a\tilde{\omega}_b^3} + \frac{45}{8\tilde{\omega}_a^2(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+2\tilde{\omega}_b)} - \frac{45\tilde{x}_a^2}{4\tilde{\omega}_a(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+2\tilde{\omega}_b)} - \frac{45(\tilde{\omega}_a+3\tilde{\omega}_b)\tilde{x}_b^2}{8\tilde{\omega}_b^2(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+2\tilde{\omega}_b)} - \frac{45\tilde{x}_a^2\tilde{x}_b^2}{2(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+2\tilde{\omega}_b)} \right. \\ \left. - \frac{15\tilde{x}_b^4}{8\tilde{\omega}_a\tilde{\omega}_b+16\tilde{\omega}_b^2} - \frac{15\tilde{x}_a^2\tilde{x}_b^4}{2\tilde{\omega}_a+4\tilde{\omega}_b} \right]$
C_5	$\left[-\frac{180\tilde{x}_a\tilde{x}_b}{(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+4\tilde{\omega}_c)} - \frac{180\tilde{x}_a\tilde{x}_b\tilde{x}_c^2}{(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+4\tilde{\omega}_c)} - \frac{30\tilde{x}_a\tilde{x}_b\tilde{x}_c^4}{\tilde{\omega}_a+\tilde{\omega}_b+4\tilde{\omega}_c} \right]$
C_6	$\left[\frac{-360(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)\tilde{x}_b\tilde{x}_c}{(\tilde{\omega}_b+\tilde{\omega}_c)(2\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)(\tilde{\omega}_b+3\tilde{\omega}_c)(2\tilde{\omega}_a+\tilde{\omega}_b+3\tilde{\omega}_c)} - \frac{180\tilde{x}_a^2\tilde{x}_b\tilde{x}_c}{(2\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)(2\tilde{\omega}_a+\tilde{\omega}_b+3\tilde{\omega}_c)} - \frac{60\tilde{x}_b\tilde{x}_c^3}{(\tilde{\omega}_b+3\tilde{\omega}_c)(2\tilde{\omega}_a+\tilde{\omega}_b+3\tilde{\omega}_c)} \right. \\ \left. - \frac{60\tilde{x}_a^2\tilde{x}_b\tilde{x}_c^3}{2\tilde{\omega}_a+\tilde{\omega}_b+3\tilde{\omega}_c} \right]$
C_7	$\left[\frac{45}{8\tilde{\omega}_a\tilde{\omega}_b\tilde{\omega}_c^2} + \frac{45}{8\tilde{\omega}_a\tilde{\omega}_b^2\tilde{\omega}_c} + \frac{45}{8\tilde{\omega}_a^2\tilde{\omega}_b\tilde{\omega}_c} - \frac{45}{8\tilde{\omega}_a\tilde{\omega}_b(\tilde{\omega}_a+\tilde{\omega}_b)\tilde{\omega}_c} - \frac{45}{8\tilde{\omega}_a\tilde{\omega}_b\tilde{\omega}_c(\tilde{\omega}_a+\tilde{\omega}_c)} - \frac{45}{8\tilde{\omega}_a\tilde{\omega}_b\tilde{\omega}_c(\tilde{\omega}_b+\tilde{\omega}_c)} \right. \\ + \frac{45}{8\tilde{\omega}_a\tilde{\omega}_b\tilde{\omega}_c(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} - \frac{45(2\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)\tilde{x}_a^2}{4\tilde{\omega}_a(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} - \frac{45\tilde{x}_a^2\tilde{x}_a^2\tilde{x}_c^2}{\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c} - \frac{45\tilde{x}_a^2\tilde{x}_b^2}{2(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} \\ - \frac{45(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)\tilde{x}_c^2}{4\tilde{\omega}_c(\tilde{\omega}_a+\tilde{\omega}_c)(\tilde{\omega}_b+\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} - \frac{45\tilde{x}_a^2\tilde{x}_c^2}{2(\tilde{\omega}_a+\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} - \frac{45\tilde{x}_a^2\tilde{x}_c^2}{2(\tilde{\omega}_b+\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} \\ \left. - \frac{45(\tilde{\omega}_a+2\tilde{\omega}_b+\tilde{\omega}_c)\tilde{x}_a^2}{4\tilde{\omega}_b(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_b+\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c)} \right]$
C_8	$\left[\frac{-360\tilde{x}_a\tilde{x}_b\tilde{x}_c\tilde{x}_d}{(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+3\tilde{\omega}_d)} - \frac{120\tilde{x}_a\tilde{x}_b\tilde{x}_c\tilde{x}_d^3}{\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+3\tilde{\omega}_d} \right]$
C_9	$\left[\frac{-360(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d)\tilde{x}_a\tilde{x}_b}{(\tilde{\omega}_a+\tilde{\omega}_b)(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+2(\tilde{\omega}_c+\tilde{\omega}_d))} - \frac{180\tilde{x}_a\tilde{x}_b((\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_d)\tilde{x}_c^2+(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)\tilde{x}_d^2)}{(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_c)(\tilde{\omega}_a+\tilde{\omega}_b+2\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+2(\tilde{\omega}_c+\tilde{\omega}_d))} \right. \\ \left. - \frac{180\tilde{x}_a\tilde{x}_b\tilde{x}_c^2\tilde{x}_d^2}{\tilde{\omega}_a+\tilde{\omega}_b+2(\tilde{\omega}_c+\tilde{\omega}_d)} \right]$

Table 2. *Cont.*

Expression for C_i Coefficients	
C_{10}	$\left[\frac{-360\tilde{x}_a\tilde{x}_b\tilde{x}_c\tilde{x}_d}{(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d+2\tilde{\omega}_e)} - \frac{360\tilde{x}_a\tilde{x}_b\tilde{x}_c\tilde{x}_d\tilde{x}_e^2}{(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d+2\tilde{\omega}_e)} \right]$
C_{11}	$\left[\frac{-720\tilde{x}_a\tilde{x}_b\tilde{x}_c\tilde{x}_d\tilde{x}_e\tilde{x}_f}{(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d+\tilde{\omega}_e+\tilde{\omega}_f)} \right]$

The expression for ψ_8^1 is

$$\begin{aligned}
\psi_8^1 = & \frac{1}{N^3} \left[\sum_{\substack{a=0 \\ 8a \bmod N \equiv 0}}^{N-1} D_1(a) + \sum_{\substack{a,b=0 \\ (6a+2b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} D_2(a, b) \right. \\
& + \sum_{\substack{a,b=0 \\ (5a+3b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} D_3(a, b) + \sum_{\substack{a,b=0 \\ (4a+4b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} \frac{1}{2} D_4(a, b) \\
& + \sum_{\substack{a,b=0 \\ (a+7b) \bmod N \equiv 0 \\ a \neq b}}^{N-1} D_5(a, b) + \sum_{\substack{a,b,c=0 \\ (a+b+6c) \bmod N \equiv 0}}^{N-1} \frac{1}{2} D_6(a, b, c) \\
& + \sum_{\substack{a,b,c=0 \\ (a+2b+5c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} D_7(a, b, c) + \sum_{\substack{a,b,c=0 \\ (a+4b+3c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} D_8(a, b, c) \\
& + \sum_{\substack{a,b,c=0 \\ (2a+2b+4c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} \frac{D_9(a, b, c)}{2} + \sum_{\substack{a,b,c=0 \\ (3a+2b+3c) \bmod N \equiv 0 \\ a \neq b \neq c}}^{N-1} \frac{D_{10}(a, b, c)}{2} \\
& + \sum_{\substack{a,b,c,d=0 \\ (a+b+2c+4d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{D_{11}(a, b, c, d)}{2} + \sum_{\substack{a,b,c,d=0 \\ 2(a+b+c+d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{D_{12}(a, b, c, d)}{24} \\
& + \sum_{\substack{a,b,c,d=0 \\ (a+2b+2c+3d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{D_{13}(a, b, c, d)}{2} + \sum_{\substack{a,b,c,d=0 \\ (a+b+c+5d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{D_{14}(a, b, c, d)}{6} \\
& + \sum_{\substack{a,b,c,d=0 \\ (a+b+3c+3d) \bmod N \equiv 0 \\ a \neq b \neq c \neq d}}^{N-1} \frac{D_{15}(a, b, c, d)}{4} + \sum_{\substack{a,b,c,d,e=0 \\ a+b+2(c+d+e) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e}}^{N-1} \frac{D_{16}(a, b, c, d, e)}{12} \\
& + \sum_{\substack{a,b,c,d,e=0 \\ (a+b+c+2d+3e) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e}}^{N-1} \frac{D_{17}(a, b, c, d, e)}{6} + \sum_{\substack{a,b,c,d,e=0 \\ (a+b+c+d+4e) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e}}^{N-1} \frac{D_{18}(a, b, c, d, e)}{24} \\
& + \sum_{\substack{a,b,c,d,e,f=0 \\ (a+b+c+d+e+3f) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e \neq f}}^{N-1} \frac{D_{19}(a, b, c, d, e, f)}{5!} + \sum_{\substack{a,b,c,d,e,f=0 \\ (a+b+c+d+2e+2f) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e \neq f}}^{N-1} \frac{D_{20}(a, b, c, d, e, f)}{48} \\
& + \sum_{\substack{a,b,c,d,e,f,g=0 \\ (a+b+c+d+e+f+2g) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e \neq f \neq g}}^{N-1} \frac{D_{21}(a, b, c, d, e, f, g)}{6!} + \sum_{\substack{a,b,c,d,e,f,g,h=0 \\ (a+b+c+d+e+f+g+h) \bmod N \equiv 0 \\ a \neq b \neq c \neq d \neq e \neq f \neq g \neq h}}^{N-1} \frac{D_{22}(a, b, c, d, e, f, g, h)}{8!} \Big]
\end{aligned} \tag{58}$$

The terms $D_1, D_2, D_3 \dots D_{22}$ are given in the Table 3.

Table 3. Expression for the terms $D_1, D_2, D_3 \dots D_{22}$.

Expression for D_i Coefficients	
D_1	$\left[\frac{875}{128\tilde{\omega}_a^5} - \frac{105x_b^2}{16\tilde{\omega}_a^4} - \frac{35x_a^4}{16\tilde{\omega}_a^3} - \frac{7x_a^6}{12\tilde{\omega}_a^2} - \frac{x_a^8}{8\tilde{\omega}_a} \right]$
D_2	$\begin{aligned} & \frac{8!}{2!6!} \left[\frac{5(36\tilde{\omega}_a^4 + 66\tilde{\omega}_a^3\tilde{\omega}_b + 121\tilde{\omega}_a^2\tilde{\omega}_b^2 + 66\tilde{\omega}_a\tilde{\omega}_b^3 + 11\tilde{\omega}_b^4)}{64\tilde{\omega}_a^4\tilde{\omega}_b^2(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)} - \frac{15(11\tilde{\omega}_a^2 + 6\tilde{\omega}_a\tilde{\omega}_b + \tilde{\omega}_b^2)x_a^2}{16\tilde{\omega}_a^3(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)} \right. \\ & - \frac{45x_b^2}{8\tilde{\omega}_b(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)} - \frac{5(5\tilde{\omega}_a + \tilde{\omega}_b)x_a^4}{16\tilde{\omega}_a^2(2\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)} - \frac{45x_a^2x_b^2}{4(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)} \\ & \left. - \frac{x_b^6}{12\tilde{\omega}_a(3\tilde{\omega}_a + \tilde{\omega}_b)} - \frac{15x_a^4x_b^2}{4(2\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)} - \frac{x_a^6x_b^2}{2(3\tilde{\omega}_b + \tilde{\omega}_a)} \right] \end{aligned}$
D_3	$\begin{aligned} & \frac{8!}{3!5!} \left[\frac{-30(23\tilde{\omega}_a + 13\tilde{\omega}_b)x_ax_b}{(\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)(5\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 3\tilde{\omega}_b)(5\tilde{\omega}_a + 3\tilde{\omega}_b)} - \frac{10x_ax_b^3}{(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 3\tilde{\omega}_b)(5\tilde{\omega}_a + 3\tilde{\omega}_b)} \right. \\ & \left. - \frac{40(2\tilde{\omega}_a + \tilde{\omega}_b)x_a^3x_b}{(\tilde{\omega}_a + \tilde{\omega}_b)(3\tilde{\omega}_a + \tilde{\omega}_b)(5\tilde{\omega}_a + \tilde{\omega}_b)(5\tilde{\omega}_a + 3\tilde{\omega}_b)} - \frac{10x_a^3x_b^3}{3(\tilde{\omega}_a + \tilde{\omega}_b)(5\tilde{\omega}_a + 3\tilde{\omega}_b)} - \frac{3x_a^5x_b}{(5\tilde{\omega}_a + \tilde{\omega}_b)(5\tilde{\omega}_a + 3\tilde{\omega}_b)} - \frac{x_a^5x_b^3}{5\tilde{\omega}_a + 3\tilde{\omega}_b} \right] \end{aligned}$
D_4	$\begin{aligned} & \frac{8!}{4!4!} \left[\frac{27(2\tilde{\omega}_a^4 + 7\tilde{\omega}_a^3\tilde{\omega}_b + 7\tilde{\omega}_a^2\tilde{\omega}_b^2 + 7\tilde{\omega}_a\tilde{\omega}_b^3 + 2\tilde{\omega}_b^4)}{64\tilde{\omega}_a^3\tilde{\omega}_b^3(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} - \frac{9(7\tilde{\omega}_a + 2\tilde{\omega}_b)x_a^2}{16\tilde{\omega}_a^2(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} \right. \\ & - \frac{9(2\tilde{\omega}_a + 7\tilde{\omega}_b)x_b^2}{16\tilde{\omega}_b^2(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} - \frac{3x_b^4}{16\tilde{\omega}_b(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} - \frac{3x_b^4}{16\tilde{\omega}_b(\tilde{\omega}_b + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} \\ & \left. - \frac{27x_a^2x_b^2}{4(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} - \frac{3x_a^2x_b^4}{4(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 2\tilde{\omega}_b)} - \frac{3x_a^4x_b^2}{4(\tilde{\omega}_a + \tilde{\omega}_b)(2\tilde{\omega}_a + \tilde{\omega}_b)} - \frac{x_b^4x_b^4}{4(\tilde{\omega}_a + \tilde{\omega}_b)} \right] \end{aligned}$
D_5	$\begin{aligned} & \frac{8!}{7!} \left[\frac{-630x_ax_b}{(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 3\tilde{\omega}_b)(\tilde{\omega}_a + 5\tilde{\omega}_b)(\tilde{\omega}_a + 7\tilde{\omega}_b)} - \frac{210x_ax_b^3}{(\tilde{\omega}_a + 3\tilde{\omega}_b)(\tilde{\omega}_a + 5\tilde{\omega}_b)(\tilde{\omega}_a + 7\tilde{\omega}_b)} - \frac{21x_ax_b^5}{(\tilde{\omega}_a + 5\tilde{\omega}_b)(\tilde{\omega}_a + 7\tilde{\omega}_b)} \right. \\ & \left. - \frac{x_ax_b^7}{\tilde{\omega}_a + 7\tilde{\omega}_b} \right] \end{aligned}$
D_6	$\begin{aligned} & \frac{8!}{6!} \left[\frac{-90x_ax_b}{(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 6\tilde{\omega}_c)} - \frac{90x_ax_bx_c^2}{(\tilde{\omega}_c + \tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 6\tilde{\omega}_c)} \right. \\ & \left. - \frac{15x_ax_bx_c^4}{(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 6\tilde{\omega}_c)} - \frac{x_ax_bx_c^6}{\tilde{\omega}_a + \tilde{\omega}_b + 6\tilde{\omega}_c} \right] \end{aligned}$
D_7	$\begin{aligned} & \frac{8!}{2!5!} \left[\frac{-20x_ax_c^2(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)}{(\tilde{\omega}_a + 3\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)(\tilde{\omega}_a + 5\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 5\tilde{\omega}_c)} - \frac{x_ax_b^2x_c}{(\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 5\tilde{\omega}_c)} \right. \\ & - \frac{x_ax_c^5}{(\tilde{\omega}_a + 5\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 5\tilde{\omega}_c)} - \frac{10x_ax_b^3x_c^3}{(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 5\tilde{\omega}_c)} - \frac{x_ax_b^2x_c^5}{\tilde{\omega}_a + 2\tilde{\omega}_b + 5\tilde{\omega}_c} \\ & \left. - \frac{30x_ax_c(3\tilde{\omega}_a^2 + 6\tilde{\omega}_a\tilde{\omega}_b + 4\tilde{\omega}_b^2 + 18\tilde{\omega}_a\tilde{\omega}_c + 18\tilde{\omega}_b\tilde{\omega}_c + 23\tilde{\omega}_c^2)}{(\tilde{\omega}_a + \tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)(\tilde{\omega}_a + 5\tilde{\omega}_c)(\tilde{\omega}_a + 2\tilde{\omega}_b + 5\tilde{\omega}_c)} \right] \end{aligned}$

Table 3. Cont.

Expression for D_i Coefficients	
D_8	$\frac{8!}{3!4!} \left[\frac{-6x_a x_b^3}{(\tilde{\omega}_a + 3\tilde{\omega}_b)(\tilde{\omega}_a + 3\tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 4\tilde{\omega}_c)} - \frac{36x_a x_b x_c^2 (\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)}{(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 4\tilde{\omega}_c)} \right.$ $- \frac{3x_a x_b x_c^4}{(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 4\tilde{\omega}_c)} - \frac{6x_a x_b^3 x_c^2}{(\tilde{\omega}_a + 3\tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 4\tilde{\omega}_c)} - \frac{x_a x_b^3 x_c^4}{\tilde{\omega}_a + 3\tilde{\omega}_b + 4\tilde{\omega}_c}$ $\left. - \frac{18x_a x_b (3\tilde{\omega}_a^2 + 13\tilde{\omega}_b^2 + 24\tilde{\omega}_b \tilde{\omega}_c + 8\tilde{\omega}_c^2 + 12\tilde{\omega}_a (\tilde{\omega}_b + \tilde{\omega}_c))}{(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + 3\tilde{\omega}_b)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)(\tilde{\omega}_a + 3\tilde{\omega}_b + 4\tilde{\omega}_c)} \right]$
D_9	$\frac{8!}{2!12!4!} \left[\frac{9}{64\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^3} + \frac{3}{32\tilde{\omega}_a \tilde{\omega}_b^2 \tilde{\omega}_c^2} + \frac{3}{32\tilde{\omega}_a^2 \tilde{\omega}_b \tilde{\omega}_c^2} - \frac{3}{32\tilde{\omega}_a \tilde{\omega}_b (\tilde{\omega}_a + \tilde{\omega}_b) \tilde{\omega}_c^2} - \frac{3}{16\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2 (\tilde{\omega}_a + \tilde{\omega}_c)} \right.$ $- \frac{3}{16\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2 (\tilde{\omega}_b + \tilde{\omega}_c)} + \frac{3}{16\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2 (\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c)} + \frac{3}{32\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2 (\tilde{\omega}_a + 2\tilde{\omega}_c)} + \frac{3}{32\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2 (\tilde{\omega}_b + 2\tilde{\omega}_c)}$ $- \frac{3}{32\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2 (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} - \frac{3x_a^2}{16\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2} - \frac{8\tilde{\omega}_a (\tilde{\omega}_a + \tilde{\omega}_b) (\tilde{\omega}_a + \tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_a + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)}{3x_b^2 (\tilde{\omega}_a^2 + 3\tilde{\omega}_b^2 + 6\tilde{\omega}_b \tilde{\omega}_c + 2\tilde{\omega}_c^2 + 3\tilde{\omega}_a (\tilde{\omega}_b + \tilde{\omega}_c))}$ $- \frac{3x_c^2}{8\tilde{\omega}_b (\tilde{\omega}_a + \tilde{\omega}_b) (\tilde{\omega}_a + \tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_b + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} - \frac{3x_c^2}{16\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c^2} + \frac{8\tilde{\omega}_a \tilde{\omega}_b (\tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_b + 2\tilde{\omega}_c)}{3x_c^2}$ $+ \frac{8\tilde{\omega}_a \tilde{\omega}_b (\tilde{\omega}_a + \tilde{\omega}_c) (\tilde{\omega}_a + 2\tilde{\omega}_c)}{x_c^4 (\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_c)} - \frac{16\tilde{\omega}_c (\tilde{\omega}_a + 2\tilde{\omega}_c) (\tilde{\omega}_b + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)}{4(\tilde{\omega}_a + \tilde{\omega}_b) (\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)}$ $- \frac{x_a^2 x_c^2 (2\tilde{\omega}_a + \tilde{\omega}_b + 3\tilde{\omega}_c)}{4(\tilde{\omega}_a + \tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_a + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} - \frac{3x_b^2 x_c^2 (\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)}{4(\tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_b + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)}$ $\left. - \frac{x_a^2 x_c^4}{4(\tilde{\omega}_a + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} - \frac{x_b^2 x_c^4}{4(\tilde{\omega}_b + 2\tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} - \frac{3x_a^2 x_b^2 x_c^2}{2(\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c) (\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} - \frac{x_a^2 x_b^2 x_c^4}{2(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)} \right]$
D_{10}	$\frac{8!}{2!3!3!} \left[\frac{9x_a x_c}{4\tilde{\omega}_b \tilde{\omega}_c (\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c) (3\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c)} - \frac{6x_a x_c}{\tilde{\omega}_b (\tilde{\omega}_a + \tilde{\omega}_c) (3\tilde{\omega}_a + \tilde{\omega}_c) (\tilde{\omega}_a + 3\tilde{\omega}_c)} - \frac{x_a^3 x_b^3 x_c^3}{3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c} \right.$ $+ \frac{9x_a x_c}{4\sqrt{2}\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c (\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)} + \frac{9x_a x_c}{8\tilde{\omega}_a \tilde{\omega}_b \tilde{\omega}_c (3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)} - \frac{x_a^3 x_c}{2\tilde{\omega}_b (\tilde{\omega}_a + \tilde{\omega}_c) (3\tilde{\omega}_c + \tilde{\omega}_c)}$ $+ \frac{3x_a^3 x_c}{2\tilde{\omega}_b (3\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c) (3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)} - \frac{9x_a x_b^2 x_c ((3\tilde{\omega}_a + 2\tilde{\omega}_b)^2 + 10\tilde{\omega}_a \tilde{\omega}_c + 4\tilde{\omega}_b \tilde{\omega}_c + \tilde{\omega}_c^2)}{4\tilde{\omega}_a \tilde{\omega}_c (\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c) (3\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c) (3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)}$ $- \frac{3x_c^3 x_b^2 x_c}{(3\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_c) (3\tilde{\omega}_c + 2\tilde{\omega}_b + 3\tilde{\omega}_c)} - \frac{x_a x_c^3}{2\tilde{\omega}_b (\tilde{\omega}_a + \tilde{\omega}_c) (\tilde{\omega}_a + 3\tilde{\omega}_c)} + \frac{2\tilde{\omega}_a (3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)}{3x_a x_b^2 x_c^3}$ $\left. - \frac{3x_a x_c^3 (\tilde{\omega}_a + 3\sqrt{2}\tilde{\omega}_a + 2\tilde{\omega}_b + 2\sqrt{2}\tilde{\omega}_b + 3\tilde{\omega}_c + 3\sqrt{2}\tilde{\omega}_c)}{4\tilde{\omega}_a \tilde{\omega}_b (\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c) (3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)} - \frac{x_c^3 x_c^3}{3(\tilde{\omega}_a + \tilde{\omega}_c) (3\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_c)} \right]$
D_{11}	$\frac{8!}{2!4!} \left[\frac{6(-3(\tilde{\omega}_a + \tilde{\omega}_b)^2 - 6(\tilde{\omega}_a + \tilde{\omega}_b)\tilde{\omega}_c - 4\tilde{\omega}_c^2 - 12(\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c)\tilde{\omega}_d - 8\tilde{\omega}_d^2)x_a x_b}{(\tilde{\omega}_a + \tilde{\omega}_b)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2(\tilde{\omega}_c + \tilde{\omega}_d))} \right.$ $- \frac{6x_a x_b x_c^2}{(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2(\tilde{\omega}_c + \tilde{\omega}_d))} - \frac{x_a x_b x_d^4}{(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c + 4\tilde{\omega}_d)}$ $- \frac{12(\tilde{\omega}_a + \tilde{\omega}_b + \tilde{\omega}_c + 3\tilde{\omega}_d)x_a x_b x_d^2}{(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2(\tilde{\omega}_c + \tilde{\omega}_d))} - \frac{x_a x_b x_c^2 x_d^4}{\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c + 4\tilde{\omega}_d}$ $\left. - \frac{6x_a x_b x_c^2 x_d^2}{(\tilde{\omega}_a + \tilde{\omega}_b + 2\tilde{\omega}_c + 4\tilde{\omega}_d)(\tilde{\omega}_a + \tilde{\omega}_b + 2(\tilde{\omega}_c + \tilde{\omega}_d))} \right]$

Table 3. *Cont.*

Expression for D_i Coefficients

 D_{12}

[illegible]

 D_{13}

$$\begin{aligned} & \frac{8!}{2!2!3!} \left[\frac{x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + \tilde{\omega}_d)} + \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_d)} + \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 2\tilde{\omega}_c + \tilde{\omega}_d)} \right. \\ & - \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + \tilde{\omega}_d)} + \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 3\tilde{\omega}_d)} - \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_d)} - \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 2\tilde{\omega}_c + 3\tilde{\omega}_d)} \\ & + \frac{3x_a x_d}{8\tilde{\omega}_b \tilde{\omega}_c \tilde{\omega}_d (\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)} - \frac{6(\tilde{\omega}_a + 2\tilde{\omega}_b + \tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + \tilde{\omega}_d)(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)}{6(\tilde{\omega}_a + \tilde{\omega}_b + 2(\tilde{\omega}_c + \tilde{\omega}_d))x_a x_c^2 x_d} \\ & - \frac{x_a x_b^2 x_c^2 x_d^3}{(\tilde{\omega}_a + 2\tilde{\omega}_c + \tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + \tilde{\omega}_d)(\tilde{\omega}_a + 2\tilde{\omega}_c + 3\tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)} - \frac{x_a x_b^2 x_c^2 x_d^3}{\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d} \\ & - \frac{3x_a x_b^2 x_c^2 x_d}{(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + \tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)} - \frac{2(\tilde{\omega}_d + \tilde{\omega}_b + \tilde{\omega}_c + 3\tilde{\omega}_d)x_a x_d^3}{(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + \tilde{\omega}_d)(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_d)(\tilde{\omega}_a + 2\tilde{\omega}_c + 3\tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)} \\ & \left. - \frac{x_a x_b^2 x_c^2 x_d^3}{(\tilde{\omega}_a + 2\tilde{\omega}_b + 3\tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)} - \frac{x_a x_c^2 x_d^3}{(\tilde{\omega}_a + 2\tilde{\omega}_c + 3\tilde{\omega}_d)(\tilde{\omega}_a + 2(\tilde{\omega}_b + \tilde{\omega}_c) + 3\tilde{\omega}_d)} \right] \end{aligned}$$

 D_{14}

$$\frac{8!}{5!} \left[\frac{-30x_ax_bx_cx_d}{(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+3\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+5\tilde{\omega}_d)} - \frac{10x_ax_bx_cx_d^3}{(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+3\tilde{\omega}_d)(\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+5\tilde{\omega}_d)} \right. \\ \left. - \frac{x_ax_bx_cx_d^5}{\tilde{\omega}_a+\tilde{\omega}_b+\tilde{\omega}_c+5\tilde{\omega}_d} \right]$$

Table 3. Cont.

[illegible]

Now, for finding the complexity, we represent the N oscillator wavefunction in the following way:

$$\psi_{0,0,\dots,0}^{s=0}(\tilde{x}_0, \dots, \tilde{x}_{N-1}) \approx \exp \left[-\frac{1}{2} v_a A_{ab}^{s=1} v_b \right] \quad (59)$$

Once again, we have to choose a particular basis. Now, there are many choices for bases, but we consider the choice of bases in the following way:

$$\vec{v} = \{ \tilde{x}_0, \dots, \tilde{x}_{N-1}, \tilde{x}_0^2, \dots, \tilde{x}_{N-1}^2, \dots, \tilde{x}_a \tilde{x}_b, \dots, \tilde{x}_0^3, \dots, \tilde{x}_{N-1}^3, \dots, \tilde{x}_a \tilde{x}_b \tilde{x}_c, \dots, \tilde{x}_0^4, \dots, \tilde{x}_{N-1}^4, \dots, \tilde{x}_a \tilde{x}_b \tilde{x}_c \tilde{x}_d, \dots, \tilde{x}_a^2 \tilde{x}_b^2, \dots, \tilde{x}_0^5, \dots, \tilde{x}_{N-1}^5, \tilde{x}_0^6, \dots, \tilde{x}_{N-1}^6, \dots, \tilde{x}_a \tilde{x}_b \tilde{x}_c \tilde{x}_d \tilde{x}_e \tilde{x}_f, \dots, \tilde{x}_a^3 \tilde{x}_b^3, \dots, \tilde{x}_a \tilde{x}_b \tilde{x}_c \tilde{x}_d \tilde{x}_e \tilde{x}_f \tilde{x}_g \tilde{x}_h, \dots, \tilde{x}_a^{1/2} \tilde{x}_b \tilde{x}_c^{1/2}, \dots \} \quad (60)$$

Here, a, b, c, d, e, f, g , and h are indices that can have any value in the range from 0 to $N - 1$ and must not be equal to each other. In the last term in \vec{v} , we mention a term that can be used to kill off-diagonal entries just as we did for the two-oscillator case. There will be many more terms like this on this basis. Expressing them explicitly is not necessary for our current work, and so we have not mentioned them.

Now, we will represent the matrix $A(s = 1)$ for N oscillators in a block diagonal fashion. In this format, the matrix will look like this:

$$A_{ab}^{s=1} = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix} \quad (61)$$

where A_1 and A_2 are the so-called *unambiguous* and *ambiguous* blocks. Once we fix the target or reference stats, the coefficients in the *unambiguous* blocks are fixed. However, this is not the case for the *ambiguous* block, as it contains numerous parameters which are not fixed beforehand.

In the *unambiguous* block A_1 , we have all of the coefficients of terms such as x_a^2 and $x_a x_b$ in Equation (54) multiplied by -2 . On the other hand, the coefficients (multiplied by -2) for terms such as

$$x_a^2 x_b^2, x_a^2 x_b^2 x_c^2, x_a x_b x_c x_d \quad (62)$$

are there on the A_2 block.

To compute the complexity, we choose a particular non-entangled reference state for arbitrary N oscillators:

$$\psi^{s=0}(x_1, x_2, \dots, x_n) = \mathcal{N}^{s=0} \exp \left[-\sum_{i=0}^{N-1} \frac{\tilde{\omega}_{ref}}{2} (x_i^2 + \lambda_4^0 x_i^4 + \lambda_6^0 x_i^6 + \lambda_8^0 x_i^8) \right] \quad (63)$$

which can be represented as follows: S

$$\psi^{s=0}(\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_n) = \mathcal{N}^{s=0} \exp \left[-\frac{1}{2} (v_a A_{ab}^{s=0} v_b) \right] \quad (64)$$

where the matrix $A_{ab}^{s=0}$ can be written as in the normal mode basis:

$$A_{ab}^{s=0} = \begin{pmatrix} \tilde{\omega}_{ref} \mathbb{I}_{N \times N} & 0 \\ 0 & A_2^{s=0} \end{pmatrix} \quad (65)$$

Here, $\mathbb{I}_{N \times N}$ is the N dimensional unit matrix. We are assuming that all the natural frequencies are (i.e., for all x_i it is true that $\omega_0 = \tilde{\omega}_{ref}$). However, $A_2^{s=0}$ cannot be represented as easily as the first block because there are many undetermined parameters. Nevertheless,

we can choose these parameters in such a way that the $A_2^{s=0}$ block becomes diagonal, just as we did for the two-oscillator case.

The complexity functional depends on the particular cost function that we choose. For the different cost functions mentioned in Equation (5), we find a different expression for the complexity functional. However, we will work with the following cost function for the rest of the paper:

$$\mathcal{F}_\kappa(s) = \sum_I p_I |Y^I|^\kappa \quad (66)$$

With respect to this particular choice for the cost function, the complexity functional becomes

$$\mathcal{C}_\kappa = \int_{s=0}^1 \mathcal{F}_\kappa ds \quad (67)$$

Here, we set all the p_I variables to be one to put all directions in the circuit space on equal footing. Now, if we choose the parameters of $A_2^{s=0}$ such that $A^{s=0}$ is diagonal, then obviously, $A^{s=1}$ and $A^{s=0}$ will commute. If this is the case, then all \mathcal{C}_κ can be written in a single equation as mentioned in [18]:

$$\begin{aligned} \mathcal{C}_\kappa &= \mathcal{C}_\kappa^{(1)} + \mathcal{C}_\kappa^{(2)} \\ &= \frac{1}{2^\kappa} \sum_{i=0}^{N-1} \left| \log \left(\frac{\lambda_i^{(1)}}{\tilde{\omega}_{ref}} \right) \right|^\kappa + \mathcal{C}_\kappa^{(2)} \end{aligned} \quad (68)$$

Here, $\lambda_i^{(1)}$ represents the eigenvalues of the unambiguous block of the $A^{s=1}$ matrix and $\mathcal{C}_\kappa^{(1)}$ and $\mathcal{C}_\kappa^{(2)}$ denote the contributions to the complexity functional for the unambiguous and ambiguous blocks, respectively. From here on, we will use the \mathcal{C}_1 complexity functional.

Commenting on $\mathcal{C}_1^{(2)}$ and the Ambiguous Block

Here, we comment on the difficulties and issues with defining the ambiguous block A_2 , as has also been discussed in [18] for the ϕ^4 interaction theory. One of the reasons for calling the A_2 matrix ambiguous is that there is a lot of arbitrariness in defining this block of the matrix; that is, there are many possible choices for defining the coefficients of the A_2 block, such as some terms which can be defined in the diagonal entries as well as in the off-diagonal entries and several higher-order cross terms, including $\tilde{x}_a \tilde{x}_b \tilde{x}_c \tilde{x}_d \tilde{x}_e \tilde{x}_f \tilde{x}_g \tilde{x}_h$, which can be defined in several forms. One possible solution to this is to try to define the A_2 matrix with the most general entries in which the coefficients are placed among all possible places in the A_2 block so that the determinant of the matrix should be positive definite. For the ambiguous block, the complexity $\mathcal{C}_1^{(2)}$ can be defined with eigenvalues $\lambda_j^{(2)}$, and the total complexity will be given by Equation (68). However, due to the great arbitrariness or ambiguities in defining the A_2 block, we cannot easily define the complexity $\mathcal{C}_1^{(2)}$. One could think of using the renormalization approach to find the general form of $\mathcal{C}_1^{(2)}$, as was performed in [18] for the ϕ^4 interaction, but the theory in our case is non-renormalizable beyond the ϕ^4 term, so it is also not possible to use the standard renormalization procedure for our case.

Here, we calculate the complexity of the unambiguous block, which is easy to analyze. We use this expression to evaluate the complexity functional in the next section.

6. Numerical Evaluation of the Complexity Functional

Up to this point, we have always set the value of $M = 1$ in the two-oscillator Hamiltonian and N oscillator Hamiltonian. However, for a generic analysis and also for the continuum limit, we need to put the M factor back in H . If we reinstate the factor of M in the Hamiltonian, we obtain the following expression for the Hamiltonian:

$$H = \frac{1}{M} \sum_{\vec{n}} \left\{ \frac{P(\vec{n})^2}{2} + \frac{1}{2} M^2 \left[\omega^2 X(\vec{n})^2 + \Omega^2 \sum_i (X(\vec{n}) - X(\vec{n} - \hat{x}_i))^2 + 2 \{ \lambda_4 X(\vec{n})^4 + \lambda_6 X(\vec{n})^6 + \lambda_8 X(\vec{n})^8 \} \right] \right\} \quad (69)$$

The overall factor in front of the Hamiltonian does not have any effect on the structure of the eigenfunctions of this Hamiltonian. However, some of the factors need to be rescaled in presence of the M factor, which are given below:

$$\omega \rightarrow \frac{\omega}{\delta} \quad \Omega \rightarrow \frac{\Omega}{\delta} \quad \lambda_4 \rightarrow \frac{\lambda_4}{\delta^2} \quad \lambda_6 \rightarrow \frac{\lambda_6}{\delta^2} \quad \lambda_8 \rightarrow \frac{\lambda_8}{\delta^2} \quad \tilde{\omega}_{ref} \rightarrow \frac{\tilde{\omega}_{ref}}{\delta} \quad \lambda_4^0 \rightarrow \frac{\lambda_4^0}{\delta}$$

$$\lambda_6^0 \rightarrow \frac{\lambda_6^0}{\delta} \quad \lambda_8^0 \rightarrow \frac{\lambda_8^0}{\delta}$$

Here, we would like to mention again that $M = \frac{1}{\delta}$. Using these rescaled parameters, we assume that the general form of the eigenvalues of A_1 represent the N oscillator Hamiltonian with first-order perturbative correction:

$$\begin{aligned} \Lambda_{i_k} &= \Lambda_{4_{i_k}} + \lambda_6 f_{i_k}(N, \tilde{\omega}_{i_p}) + \lambda_8 g_{i_k}(N, \tilde{\omega}_{i_p}), \quad N: \text{Even} \\ &= \Lambda_{4_{i_k}} + \lambda_6 f'_{i_k}(N, \tilde{\omega}_{i_p}) + \lambda_8 g'_{i_k}(N, \tilde{\omega}_{i_p}), \quad N: \text{Odd} \end{aligned} \quad (70)$$

where N denotes the number of lattice points in each spatial dimension and the i_k indices run from 0 to $N - 1$ for each dimension. Then, the $d - 1$ dimensional spatial volume becomes $L^{d-1} = (N\delta)^{d-1}$.

Here, $\Lambda_{4_{i_k}}$ is the contribution from the ϕ^4 interaction, and f, g, f' , and g' denote the additional contributions to the eigenvalues in the presence of ϕ^6 and ϕ^8 interaction. The form of $\Lambda_{4_{i_k}}$, as mentioned in [18], is

$$\begin{aligned} \Lambda_{4_{i_k}} &= \frac{\tilde{\omega}_{i_k}}{\delta} + \frac{3\lambda_4}{2N} \left(\frac{2}{\tilde{\omega}_{i_k}(\tilde{\omega}_{i_k} + \tilde{\omega}_{N-i_k})} + \frac{2}{\tilde{\omega}_{i_k}(\tilde{\omega}_{i_k} + \tilde{\omega}_{\frac{N}{2}-i_k})} \right), \quad N: \text{Even} \\ &= \frac{\tilde{\omega}_{i_k}}{\delta} + \frac{3\lambda_4}{2N} \left(\frac{2}{\tilde{\omega}_{i_k}(\tilde{\omega}_{i_k} + \tilde{\omega}_{N-i_k})} \right), \quad N: \text{Odd} \end{aligned} \quad (71)$$

These additional terms f, g, f' , and g' cannot be calculated analytically. Therefore, we resort to numerical methods to calculate these.

The work carried out in [18] had a proper analytical expression for the eigenvalues, which made it easier to study the RG flows. However, when we consider higher-order interactions such as ϕ^6 and ϕ^8 , such analytic expressions for the RG flows and complexity cannot be found. This makes it difficult for us to study the RG flows and MERA and is beyond the scope of our model. Instead, we will focus only on complexity. The eigenvalues we obtained were small corrections to the one obtained in [18], so the connection they made will not be affected by the addition of higher interacting terms. Now, we will resort to numerical methods in the next section.

Numerical Analysis of the Complexity Functional

We will calculate the complexity for the unambiguous block first for an increasing number of oscillators. We have found the wavefunction for the Hamiltonian in Equation (47). As we reinserted the M term, we will just update the complexity using the rescaled

parameters mentioned in the previous subsection. We set the following relevant parameter values:

$$\begin{array}{llll} \lambda_4 = 0.5 & \lambda_6 = 0.2 & \lambda_8 = 0.001 & \omega_0 = m = 4.0 \\ \Omega = 0.25 & L = 200 & \tilde{\omega}_{ref} = 1.6 & \end{array}$$

where L is the length of the periodic chain. We chose N and δ so that $N\delta = L$ was always satisfied. We will use the $\mathcal{C}_1^{(1)}$ functional for the unambiguous block.

Case I: Increasing the Interactions

In Figure 2, we have plotted numerically the behavior of the complexity of the unambiguous block as a function of N , which is the number of oscillators in $d = 2$ dimensions. In Figure 2a, we have two complexities, where the points in blue represent the complexity of the theory, which has no interaction term, and this complexity is due to the self-interaction between pairs of oscillators. We also see the points in orange and light orange, which represent the complexity of the theory with $\lambda_4\phi^4$ interaction. We notice that there is a bump initially in the graph for small N values, but in Figure 2a–c, we can observe that the values of the complexity with the free theory and the complexity with interactions became the same as we increased the value of N . We see that $\mathcal{C}_1^{(1)}$ grew linearly with increasing N values, and the contributions to $\mathcal{C}_1^{(1)}$ due to even interaction terms became negligible, while the behavior of the complexity for the unambiguous block would be same as if we were dealing only with the free theory. In Figure 2d, we have plotted $\mathcal{C}_1^{(1)}$ for N being an odd number of oscillators for even interactions of $\lambda_4\phi^4 + \lambda_6\phi^6 + \lambda_8\phi^8$, and we see that the initial values of the complexity increased as we included higher-order terms in the theory, but when we increased N , the contribution from these perturbative terms died out, and the graph followed a ϕ^2 linear pattern of $\mathcal{C}_1^{(1)}$.

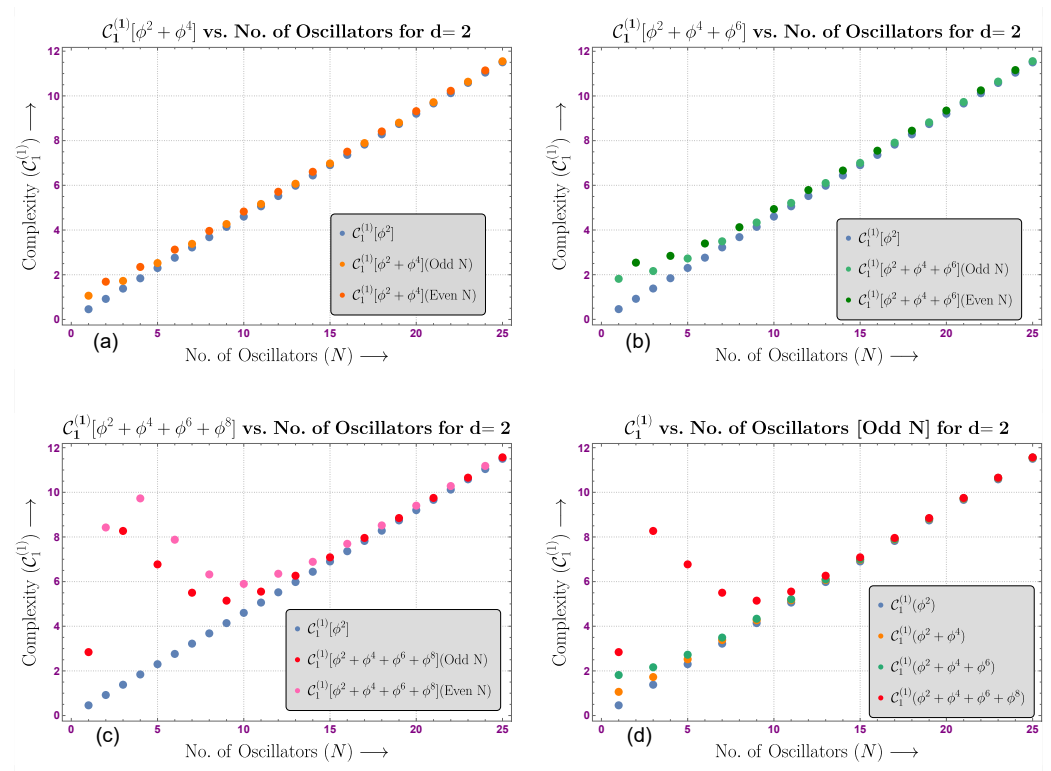


Figure 2. Plot: (a–c) represents the complexity $C_1^{(1)}$ (from the unambiguous block) vs. the number of oscillators (N) for $d = 2$ dimensions with different interactions. In plot (d), complexity $C_1^{(1)}$ vs. an odd number of oscillators (even resembling the same pattern) from all the interactions are placed together in the same plot, showing the contribution from each interaction.

Case II: Increasing the Dimension

In Figure 3, we show six different plots. In the first two plots, the complexity for the unambiguous block (up to ϕ^4 interaction) is plotted with respect to the number of oscillators in dimensions $d = 3$ and 4 . Here, we notice that as we increased the dimensions, the contribution to $C_1^{(1)}$ due to the interaction term increased, and we saw a similar pattern as we included other higher-order even terms (i.e., the third and fourth graphs have $(\lambda_4\phi^4 + \lambda_6\phi^6)$ interactions, and the fifth and sixth graphs contain $(\lambda_4\phi^4 + \lambda_6\phi^6 + \lambda_8\phi^8)$ interactions). However, in higher dimensions, the contributions of these interactions to the complexity $C_1^{(1)}$ also became negligible when we increased the value of N , and the behavior of this complexity became similar to the case where we had only the ϕ^2 term and it grew linearly.

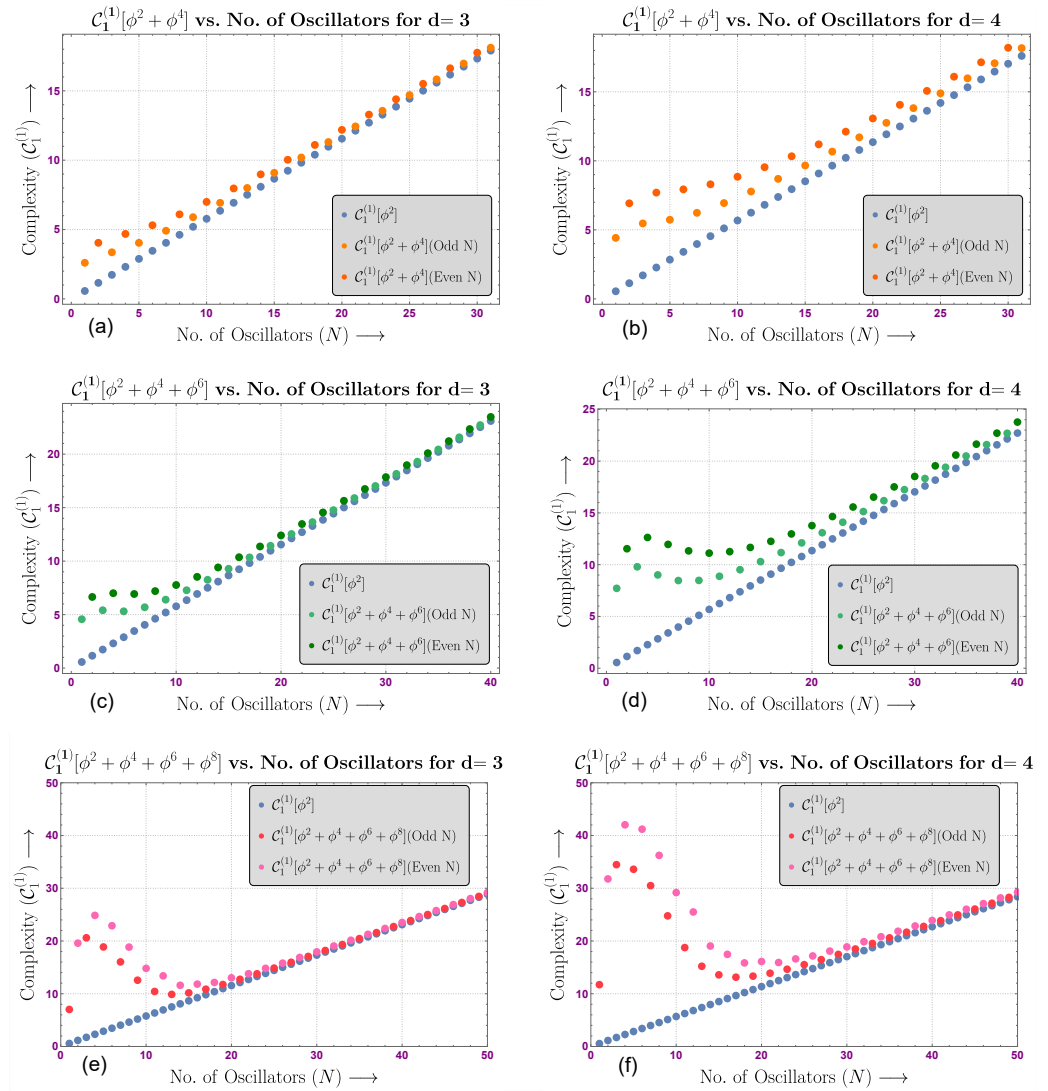


Figure 3. Plot of complexity $\mathcal{C}_1^{(1)}$ vs. number of oscillators in $d = 3$ and $d = 4$, respectively, is shown in (a–f) for $(\lambda_2 \phi^2 + \lambda_4 \phi^4 + \lambda_6 \phi^6 + \lambda_8 \phi^8)$.

Case III: $\mathcal{C}_1^{(1)}$ vs. ω_0

In Figure 4, we have plotted the variation in the complexity $\mathcal{C}_1^{(1)}$ versus ω_0 for a particular value of oscillators $N = 15$, and we also show the variation in the same plot for different dimensions ($d = 2, 3, 4$). As we increased the number of dimensions, the complexity of the unambiguous block $\mathcal{C}_1^{(1)}$ increased, and in a particular dimension, the complexity value increased as we increased the number of interactions, which was noticeable for low values of ω_0 . However, as we increased the value of ω_0 , the behavior became similar to the free scalar theory.

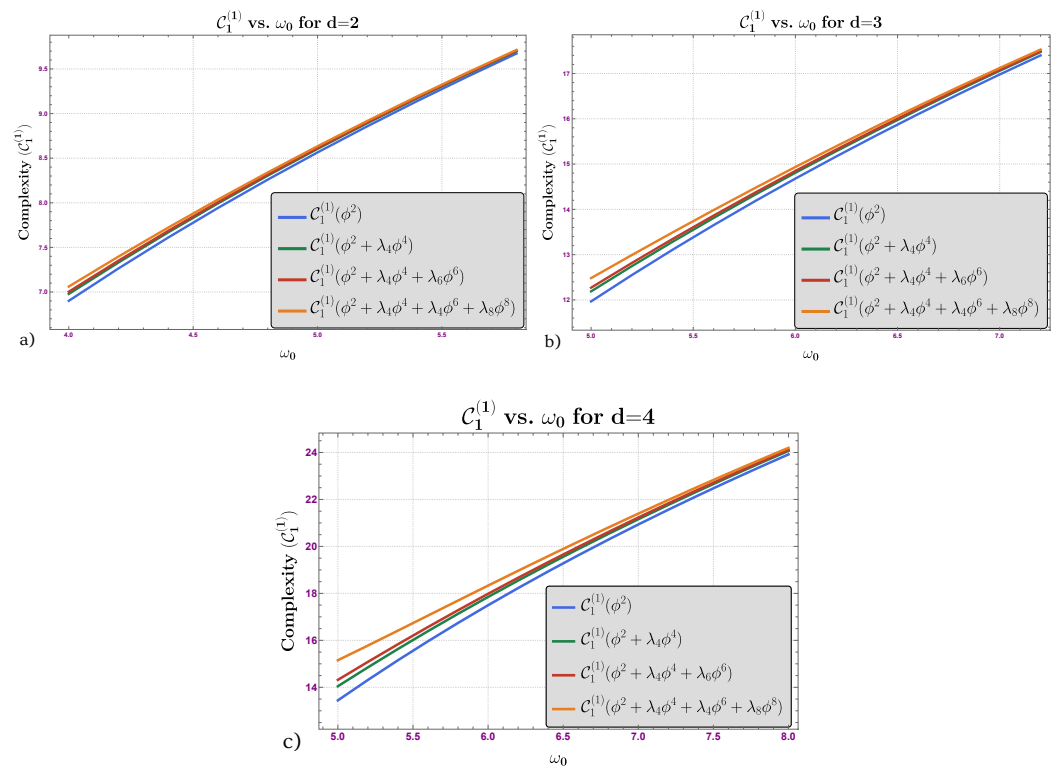


Figure 4. Plot of complexity $C_1^{(1)}$ vs. ω_0 (a) for $d = 2$, (b) for $d = 3$, and (c) for $d = 4$.

Case IV: Fractional Change in $C_1^{(1)}$

We define the fractional change in complexity C_1 for a particular N value as

$$\frac{C_1(N+2) - C_1(N)}{C_1(N)}$$

Here, we have an increment of two in the definition because odd and even branches of N can possibly show different behavior, as was the case for the complexity.

For small values of N , the even and odd complexities were different from each other. This is directly related to the fact that one can distinguish the system with an even or odd number of oscillators, but as we went for a large number of oscillators or in the continuum limit, the distinction between the even and odd numbers of oscillators faded away. In Figure 5, we have plotted the complexity of the unambiguous block, and we find that, initially, the fractional change in complexity was large for small N values, but it decreased continuously as we moved toward a large number of oscillators.

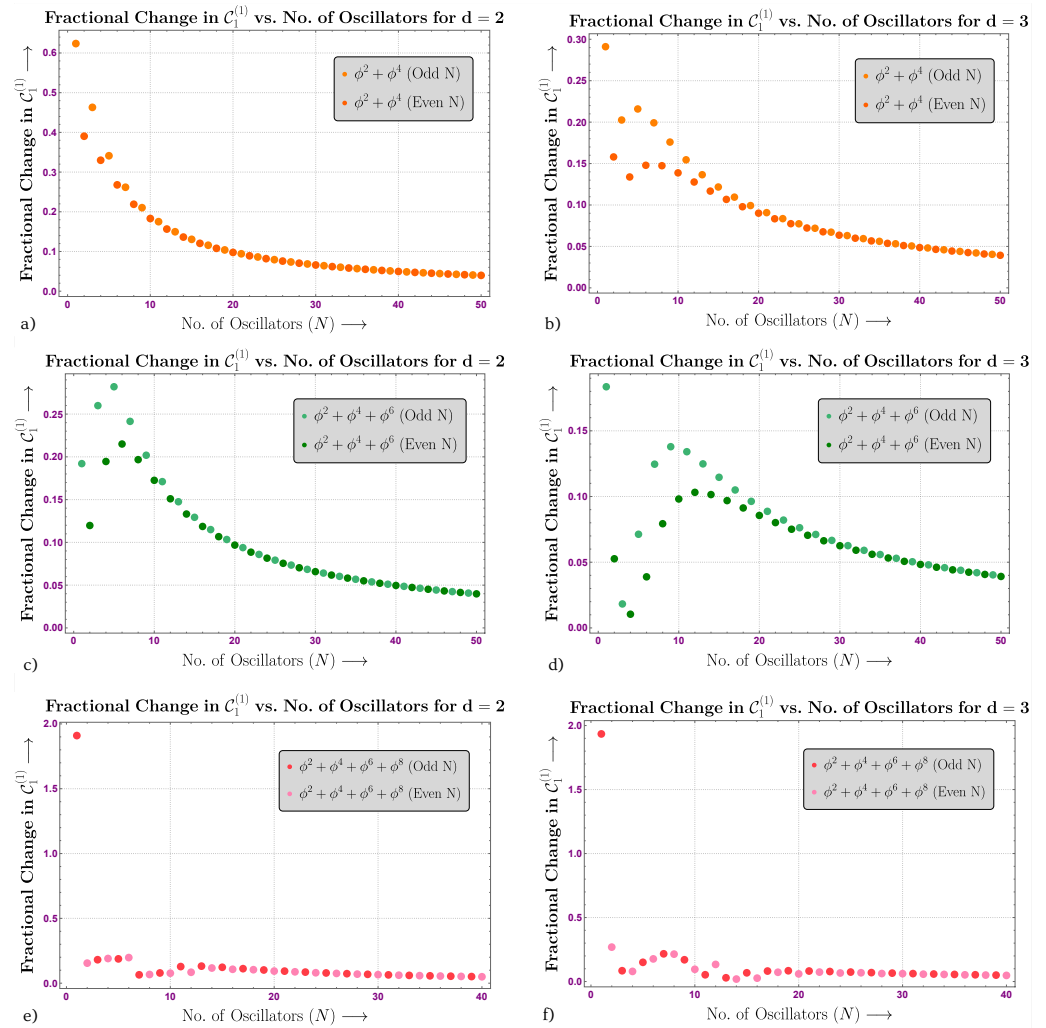


Figure 5. The Plot of fractional change in complexity vs. number of oscillators.

7. Conclusions and Future Prospects

This work studied the circuit complexity for weakly interacting scalar field theory with ϕ^4 , ϕ^6 , and ϕ^8 Wilsonian operators coupled via λ_4 , λ_6 , and λ_8 to a free scalar field theory, respectively. The values of the coupling constants were chosen in the framework of an EFT such that the perturbation analysis was valid. The reference state was an unentangled, nearly Gaussian state, and the target state was an entangled, nearly Gaussian state which was calculated using a first-order perturbation theory. First, we worked with the case of two oscillators, where the unitary evolution \mathbb{U} , which took us from the reference state to the target state, was parameterized using the AdS parameters. With this, we calculated the line element and found the complexity functional by imposing the appropriate boundary conditions. Then, we proceeded to the N oscillator case. Here, the circuit complexity depended on the ratio of the eigenvalues of the target to the reference states of the N oscillators. Since we could not observe any analytical expression of the eigenvalues of the target state of the N oscillators, we resorted to numerical analysis. The target matrix for N oscillators had a part where the bases could be uniquely determined (unambiguous part) and another part where the bases could not be determined (ambiguous part). The contribution to the total complexity came from the ambiguous as well as the unambiguous parts. In our work, we focused mainly on the computation of the complexity for the unambiguous part, denoted by the A_2 matrix. The following are the results that we observed:

1. From our numerical analysis, the QCC with $\kappa = 1$ for the free field theory increased linearly with the number of oscillators. As we included the higher even Wilsonian terms, the growth of the complexity (contribution from the unambiguous part) was no longer linear for a small number of oscillators. For the large N limit, the contribution to the complexity from the interacting part vanished, and the linearity was restored.
2. From the graph of complexity vs. ω_0 , we see that upon fixing the dimensions and the number of oscillators, the complexity from the unambiguous part increased with an increasing value for ω_0 .
3. Another pattern inferred from our analysis is that as the dimension increases, the contribution to $\mathcal{C}_1^{(1)}$ due to the interaction term increases for a fixed number of oscillators. We observed this pattern using degenerate frequencies for higher dimensions. One would expect a similar pattern, even if the frequencies were non-degenerate.

In [18], the eigenvalues had a proper analytical expression, which makes it easier to study RG flows. On the other hand, after adding higher-order corrections, there is no analytical expression of the eigenvalues. This makes it very challenging to study the RG and MERA connection. The eigenvalues we obtained were small corrections to the one obtained in [18], so the connection they made would not be affected by the addition of higher interacting terms. In upcoming works, we will address this issue.

In our analysis, we used $\kappa = 1$ in our complexity functional \mathcal{C}_κ , but there are other different and useful kinds of measures that one can explore to gain new insights into circuit complexity.

Our approach to computing complexity is based on Nielsen's geometric approach, which suffers from ambiguity in choosing the elementary quantum gates and states. Recent works have attempted to develop a new notion of complexity that is independent of these choices. As for our future goals, we have in mind the following:

- We can calculate the circuit complexity for odd Wilsonian terms in the effective theory, such as ϕ^3 , ϕ^4 , and ϕ^7 . We can further generalize the study by adding both even and odd interaction terms together.
- We can study the behavior of circuit complexity in a similar theory when there is a quantum quench in the interaction and mass. We have already performed this for a ϕ^4 interacting theory [119].
- We can further analyze circuit complexity in fermionic field theories and gauge theories.
- We can explore this problem in the context of the Krylov complexity [95,103,120], which is currently a melting pot in this research area.
- We can compare the Krylov complexity and circuit complexity for such theories to know which is a better measure of information for such cases.

Author Contributions: Conceptualization, S.C.; Methodology, S.C., S.M., N.P., A.R., S.S. and S.S.S.; Software, K.A., S.K., S.M., N.P., A.R., S.S., P.S. and S.S.S.; Validation, S.C.; Formal analysis, K.A., S.C., S.K., S.M., N.P., A.R., S.S., P.S. and S.S.S.; Investigation, S.C.; Writing—original draft, K.A., S.C., S.K., S.M., N.P. and A.R.; Writing—review & editing, S.C.; Supervision, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: S.C. would like to thank the work friendly environment and all the members of CCSP, SGT University. The research fellowship of S.C. was supported by the J. C. Bose National Fellowship of Sudhakar Panda. S.C. also would like to thank the School of Physical Sciences at the National Institute for Science Education and Research (NISER) in Bhubaneswar for providing a work-friendly environment. S.C. also thanks all the members of our newly formed virtual international non-profit consortium “Quantum Structures of the Space-Time & Matter”

(QASTM) for their elaborative discussions. S.C. would also like to thank all the speakers of the QASTM zoominar series from different parts of the world (For the uploaded YouTube link, see <https://www.youtube.com/playlist?list=PLzW8AJcryManrTsG-4U4z9ip1J1dWoNgd>) for supporting the research forum by giving outstanding lectures and their valuable time during the COVID pandemic time. Kiran Adhikari would like to thank TTK, RWTH, JARA, and the Institute of Quantum Information for the fellowships. Saptarshi Mandal, Nilesh Pandey, Abhishek Roy, Soumya Sarker, Partha Sarker, and Sadaat Salman Shariff would like to express their heartiest thanks to Jadavpur University, the University of Dhaka, NIT Karnataka, IIT Jodhpur, the University of Madras, and Delhi Technological University, respectively, for imparting knowledge and their enthusiasm for this research. Abhishek Roy would like to thank Sujit Damase for the discussions related to group generators. Partha Sarker would like to thank Syed Hasibul Hasan Chowdhury for the relevant discussions. K.A. would also like to thank David Di Vincenzo for his help in understanding quantum information theoretic concepts such as entanglement entropy and complexity. Finally, we would like to acknowledge our debt to the people belonging to various parts of the world for their generous and steady support for research in the natural sciences.

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

Appendix A. Interacting Part of the Hamiltonian in a Fourier Basis

The interacting part in the N oscillator Hamiltonian is

$$H' = \sum_{a=0}^{N-1} \lambda_4 x_a^4 + \lambda_6 x_a^6 + \lambda_8 x_a^8 = H'_{\phi^4} + H'_{\phi^6} + H'_{\phi^8} \quad (\text{A1})$$

Now, we apply the discrete Fourier transform as in Equation (48) to find the ϕ^4 interaction:

$$H'_{\phi^4} = \sum_{a=0}^{N-1} \frac{\lambda_4}{N^2} \sum_{k', k_1, k_2, k_3=0}^{N-1} \exp \left[i \frac{2\pi a}{N} (k' + k_1 + k_2 + k_3) \right] \tilde{x}_{k'} \tilde{x}_{k_1} \tilde{x}_{k_2} \tilde{x}_{k_3} \quad (\text{A2})$$

We apply the sum over index a and use the relation

$$\sum_{a=0}^{N-1} \exp \left[-i \left(\frac{2\pi a (k - k')}{N} \right) \right] = N \delta_{k, k'} \quad (\text{A3})$$

to obtain

$$H'_{\phi^4} = \frac{\lambda_4}{N} \sum_{k', k_1, k_2, k_3=0}^{N-1} \delta_{k' + k_1 + k_2 + k_3, 0} \tilde{x}_{k'} \tilde{x}_{k_1} \tilde{x}_{k_2} \tilde{x}_{k_3} \quad (\text{A4})$$

Now, the Kronecker delta will reduce one of the indices, such as k' to $-k_1 - k_2 - k_3$. Now, k' only runs from $[0, N-1]$, whereas $-k_1 - k_2 - k_3$ has possible values in the range $[-3N, 0]$. To obtain a valid index value for k' , we use the relation $\tilde{x}_{k+N} = \tilde{x}_k$ and write $k' = N - k_1 - k_2 - k_3 \bmod N$. This will return a valid index value for k' . Then, we have

$$H'_{\phi^4} = \frac{\lambda_4}{N} \sum_{k_1, k_2, k_3=0}^{N-1} \tilde{x}_\alpha \tilde{x}_{k_1} \tilde{x}_{k_2} \tilde{x}_{k_3} \quad (\text{A5})$$

Using similar arguments, we can find H'_{ϕ^6} and H'_{ϕ^8} .

Appendix B. \mathcal{C}_2 in Terms of the Ratio of the Target and Reference Matrix Eigenvalues

We claimed in Equation (46) that \mathcal{C}_2 can be expressed in terms of the ratio of eigenvalues of the target and reference matrix (i.e., $A(s=1)$ and $A(s=0)$, respectively). This was due to the nature of the unitary operator U and the diagonal block structure of $A(s=1)$ and $A(s=0)$.

To prove this, let us look at the complexity functional in Equation (44). The parameters in the 2×2 blocks on the U matrix have AdS parametrization, and they appear in $2[dy_i(1)^2 +$

$d\rho_i(1)^2]$ in \mathcal{C}_2 , where $i = 1, 3, 5, 7, 9$. We can find these values for $y_i(1)$ and $\rho_i(1)$ from the boundary conditions we obtained in Equation (42). These values can be represented by the eigenvalues of $A(s = 0)$ and $A(s = 1)$ in the following way:

$$\begin{aligned} y_i &= \frac{1}{4} \log \left[\frac{\lambda_1 \lambda_2}{\Omega_1 \Omega_2} \right] \\ \rho_i &= \frac{1}{2} \cosh^{-1} \left[\frac{\lambda_1 + \lambda_2}{2\sqrt{\lambda_1 \lambda_2}} \right] \end{aligned} \quad (\text{A6})$$

Here, λ_1 and λ_2 are the eigenvalues of the 2×2 block in the $A(s = 1)$ matrix corresponding to the block in U , whereas Ω_1 and Ω_2 are diagonal elements of the similar block 2×2 in $A(s = 0)$. We can use the relation

$$\cosh^{-1}(x) = \ln(x + \sqrt{x^2 - 1}) \quad (\text{A7})$$

to find the following for ρ_i :

$$\rho_i = \frac{1}{4} \ln \left[\frac{\lambda_2}{\lambda_1} \right] \quad (\text{A8})$$

Then, our desired part in \mathcal{C}_2 will be

$$2(y_i(1))^2 + \rho_i(1)^2 = 2 \left[\ln \left[\frac{\lambda_1}{\Omega_1} \right]^2 + \ln \left[\frac{\lambda_2}{\Omega_2} \right]^2 \right] \quad (\text{A9})$$

Now, $i = 2, 4, 6, 8$, and we have a different scenario. These are the lone diagonal parameters in the U matrix and have boundary conditions such as

$$y_i = \frac{1}{2} \ln \left[\frac{\lambda_T}{\Omega_R} \right] \quad (\text{A10})$$

Here, λ_T and Ω_R denote the particular diagonal elements in $A(s = 0)$ and $A(s = 1)$, respectively, corresponding to the parameter y_i here. With these parameter values in hand, we can find from the complexity functional in Equation (44) the expression for Equation (46).

References

1. Harlow, D. TASI Lectures on the Emergence of Bulk Physics in AdS/CFT. *arXiv* **2018**, arXiv:1802.01040.
2. Ryu, S.; Takayanagi, T. Holographic derivation of entanglement entropy from AdS/CFT. *Phys. Rev. Lett.* **2006**, *96*, 181602. [[CrossRef](#)] [[PubMed](#)]
3. Hubeny, V.E.; Rangamani, M.; Takayanagi, T. A Covariant holographic entanglement entropy proposal. *JHEP* **2007**, *7*, 062. [[CrossRef](#)]
4. Rangamani, M.; Takayanagi, T. *Holographic Entanglement Entropy*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 931. [[CrossRef](#)]
5. Susskind, L. Computational Complexity and Black Hole Horizons. *Fortsch. Phys.* **2016**, *64*, 24–43. Addendum: *Fortsch. Phys.* **2016**, *64*, 44–48. [[CrossRef](#)]
6. Stanford, D.; Susskind, L. Complexity and Shock Wave Geometries. *Phys. Rev. D* **2014**, *90*, 126007. [[CrossRef](#)]
7. Susskind, L.; Zhao, Y. Switchbacks and the Bridge to Nowhere. *arXiv* **2014**, arXiv:1408.2823.
8. Susskind, L. Entanglement is not enough. *Fortsch. Phys.* **2016**, *64*, 49–71. [[CrossRef](#)]
9. Brown, A.R.; Roberts, D.A.; Susskind, L.; Swingle, B.; Zhao, Y. Complexity, action, and black holes. *Phys. Rev. D* **2016**, *93*, 086006. [[CrossRef](#)]
10. Brown, A.R.; Roberts, D.A.; Susskind, L.; Swingle, B.; Zhao, Y. Holographic Complexity Equals Bulk Action? *Phys. Rev. Lett.* **2016**, *116*, 191301. [[CrossRef](#)]
11. Brown, A.R.; Susskind, L.; Zhao, Y. Quantum Complexity and Negative Curvature. *Phys. Rev. D* **2017**, *95*, 045010. [[CrossRef](#)]
12. Couch, J.; Fischler, W.; Nguyen, P.H. Noether charge, black hole volume, and complexity. *JHEP* **2017**, *3*, 119. [[CrossRef](#)]
13. Susskind, L. *Three Lectures on Complexity and Black Holes*; Springer: Berlin/Heidelberg, Germany, 2020.
14. Jefferson, R.; Myers, R.C. Circuit complexity in quantum field theory. *JHEP* **2017**, *10*, 107. [[CrossRef](#)]
15. Chapman, S.; Heller, M.P.; Marrochio, H.; Pastawski, F. Toward a Definition of Complexity for Quantum Field Theory States. *Phys. Rev. Lett.* **2018**, *120*, 121602. [[CrossRef](#)] [[PubMed](#)]

16. Khan, R.; Krishnan, C.; Sharma, S. Circuit Complexity in Fermionic Field Theory. *Phys. Rev. D* **2018**, *98*, 126001. [\[CrossRef\]](#)
17. Hackl, L.; Myers, R.C. Circuit complexity for free fermions. *JHEP* **2018**, *7*, 139. [\[CrossRef\]](#)
18. Bhattacharyya, A.; Shekar, A.; Sinha, A. Circuit complexity in interacting QFTs and RG flows. *JHEP* **2018**, *10*, 140. [\[CrossRef\]](#)
19. Haferkamp, J.; Faist, P.; Kothakonda, N.B.T.; Eisert, J.; Halpern, N.Y. Linear growth of quantum circuit complexity. *Nat. Phys.* **2022**, *18*, 528–532. [\[CrossRef\]](#)
20. Bhattacharyya, A.; Chemsissany, W.; Haque, S.S.; Murugan, J.; Yan, B. The Multi-faceted Inverted Harmonic Oscillator: Chaos and Complexity. *SciPost Phys. Core* **2021**, *4*, 002. [\[CrossRef\]](#)
21. Ali, T.; Bhattacharyya, A.; Haque, S.S.; Kim, E.H.; Moynihan, N.; Murugan, J. Chaos and Complexity in Quantum Mechanics. *Phys. Rev. D* **2020**, *101*, 026021. [\[CrossRef\]](#)
22. Eisert, J. Entangling Power and Quantum Circuit Complexity. *Phys. Rev. Lett.* **2021**, *127*, 020501. [\[CrossRef\]](#)
23. Roberts, D.A.; Yoshida, B. Chaos and complexity by design. *JHEP* **2017**, *4*, 121. [\[CrossRef\]](#)
24. Camilo, G.; Melnikov, D.; Novaes, F.; Prudenziati, A. Circuit Complexity of Knot States in Chern-Simons theory. *JHEP* **2019**, *7*, 163. [\[CrossRef\]](#)
25. Couch, J.; Fan, Y.; Shashi, S. Circuit Complexity in Topological Quantum Field Theory. *Fortsch. Phys.* **2022**, *70*, 9–10. [\[CrossRef\]](#)
26. Chagnet, N.; Chapman, S.; de Boer, J.; Zukowski, C. Complexity for Conformal Field Theories in General Dimensions. *Phys. Rev. Lett.* **2022**, *128*, 051601 [\[CrossRef\]](#)
27. Flory, M.; Heller, M.P. Conformal field theory complexity from Euler-Arnold equations. *JHEP* **2020**, *12*, 091 [\[CrossRef\]](#)
28. Jaiswal, N.; Gautam, M.; Sarkar, T. Complexity and information geometry in the transverse XY model. *Phys. Rev. E* **2021**, *104*, 024127. [\[CrossRef\]](#)
29. Barbon, J.L.F.; Rabinovici, E. Holographic complexity and spacetime singularities. *JHEP* **2016**, *1*, 084. [\[CrossRef\]](#)
30. Alishahiha, M. Holographic Complexity. *Phys. Rev. D* **2015**, *92*, 126009. [\[CrossRef\]](#)
31. Yang, R.Q. Strong energy condition and complexity growth bound in holography. *Phys. Rev. D* **2017**, *95*, 086017. [\[CrossRef\]](#)
32. Chapman, S.; Marrochio, H.; Myers, R.C. Complexity of Formation in Holography. *JHEP* **2017**, *1*, 062. [\[CrossRef\]](#)
33. Carmi, D.; Myers, R.C.; Rath, P. Comments on Holographic Complexity. *JHEP* **2017**, *3*, 118. [\[CrossRef\]](#)
34. Reynolds, A.; Ross, S.F. Divergences in Holographic Complexity. *Class. Quant. Grav.* **2017**, *34*, 105004. [\[CrossRef\]](#)
35. Zhao, Y. Complexity and Boost Symmetry. *Phys. Rev. D* **2018**, *98*, 086011. [\[CrossRef\]](#)
36. Flory, M. A complexity/fidelity susceptibility g -theorem for $AdS_3/BCFT_2$. *JHEP* **2017**, *6*, 131. [\[CrossRef\]](#)
37. Reynolds, A.; Ross, S.F. Complexity in de Sitter Space. *Class. Quant. Grav.* **2017**, *34*, 175013. [\[CrossRef\]](#)
38. Carmi, D.; Chapman, S.; Marrochio, H.; Myers, R.C.; Sugishita, S. On the Time Dependence of Holographic Complexity. *JHEP* **2017**, *11*, 188. [\[CrossRef\]](#)
39. Couch, J.; Eccles, S.; Fischler, W.; Xiao, M.L. Holographic complexity and noncommutative gauge theory. *JHEP* **2018**, *3*, 108. [\[CrossRef\]](#)
40. Yang, R.Q.; Niu, C.; Zhang, C.Y.; Kim, K.Y. Comparison of holographic and field theoretic complexities for time dependent thermofield double states. *JHEP* **2018**, *2*, 082. [\[CrossRef\]](#)
41. Abt, R.; Erdmenger, J.; Hinrichsen, H.; Melby-Thompson, C.M.; Meyer, R.; Northe, C.; Reyes, I.A. Topological Complexity in AdS_3/CFT_2 . *Fortsch. Phys.* **2018**, *66*, 1800034. [\[CrossRef\]](#)
42. Swingle, B.; Wang, Y. Holographic Complexity of Einstein-Maxwell-Dilaton Gravity. *JHEP* **2018**, *09*, 106. [\[CrossRef\]](#)
43. Reynolds, A.P.; Ross, S.F. Complexity of the AdS Soliton. *Class. Quant. Grav.* **2018**, *35*, 095006. [\[CrossRef\]](#)
44. Fu, Z.; Maloney, A.; Marolf, D.; Maxfield, H.; Wang, Z. Holographic complexity is nonlocal. *JHEP* **2018**, *2*, 072. [\[CrossRef\]](#)
45. An, Y.S.; Peng, R.H. Effect of the dilaton on holographic complexity growth. *Phys. Rev. D* **2018**, *97*, 066022. [\[CrossRef\]](#)
46. Bolognesi, S.; Rabinovici, E.; Roy, S.R. On Some Universal Features of the Holographic Quantum Complexity of Bulk Singularities. *JHEP* **2018**, *6*, 016. [\[CrossRef\]](#)
47. Chen, B.; Li, W.M.; Yang, R.Q.; Zhang, C.Y.; Zhang, S.J. Holographic subregion complexity under a thermal quench. *JHEP* **2018**, *7*, 034. [\[CrossRef\]](#)
48. Abt, R.; Erdmenger, J.; Gerbershagen, M.; Melby-Thompson, C.M.; Northe, C. Holographic Subregion Complexity from Kinematic Space. *JHEP* **2019**, *1*, 012. [\[CrossRef\]](#)
49. Hashimoto, K.; Iizuka, N.; Sugishita, S. Thoughts on Holographic Complexity and its Basis-dependence. *Phys. Rev. D* **2018**, *98*, 046002. [\[CrossRef\]](#)
50. Flory, M.; Miekley, N. Complexity change under conformal transformations in AdS_3/CFT_2 . *JHEP* **2019**, *5*, 003. [\[CrossRef\]](#)
51. Couch, J.; Eccles, S.; Jacobson, T.; Nguyen, P. Holographic Complexity and Volume. *JHEP* **2018**, *11*, 044. [\[CrossRef\]](#)
52. Hosseini Mansoori, S.A.; Jahnke, V.; Qaemmaqami, M.M.; Olivas, Y.D. Holographic Complexity of Anisotropic Black Branes. *Phys. Rev. D* **2019**, *100*, 046014. [\[CrossRef\]](#)
53. Chapman, S.; Marrochio, H.; Myers, R.C. Holographic complexity in Vaidya spacetimes. Part I. *JHEP* **2018**, *6*, 046. [\[CrossRef\]](#)
54. Chapman, S.; Marrochio, H.; Myers, R.C. Holographic complexity in Vaidya spacetimes. Part II. *JHEP* **2018**, *6*, 114. [\[CrossRef\]](#)
55. Caceres, E.; Chapman, S.; Couch, J.D.; Hernandez, J.P.; Myers, R.C.; Ruan, S.M. Complexity of Mixed States in QFT and Holography. *JHEP* **2020**, *3*, 012. [\[CrossRef\]](#)
56. Ben-Ami, O.; Carmi, D. On Volumes of Subregions in Holography and Complexity. *JHEP* **2016**, *11*, 129. [\[CrossRef\]](#)
57. Abad, F.J.G.; Kulaxizi, M.; Parnachev, A. On Complexity of Holographic Flavors. *JHEP* **2018**, *1*, 127. [\[CrossRef\]](#)
58. Brown, A.R.; Susskind, L. Second law of quantum complexity. *Phys. Rev. D* **2018**, *97*, 086015. [\[CrossRef\]](#)

59. Bernamonti, A.; Galli, F.; Hernandez, J.; Myers, R.C.; Ruan, S.M.; Simón, J. First Law of Holographic Complexity. *Phys. Rev. Lett.* **2019**, *123*, 081601. [\[CrossRef\]](#)
60. Bernamonti, A.; Galli, F.; Hernandez, J.; Myers, R.C.; Ruan, S.M.; Simón, J. Aspects of The First Law of Complexity. *J. Phys. A Math. Theor.* **2020**, *53*, 294002. [\[CrossRef\]](#)
61. Cai, R.G.; Ruan, S.M.; Wang, S.J.; Yang, R.Q.; Peng, R.H. Action growth for AdS black holes. *JHEP* **2016**, *9*, 161. [\[CrossRef\]](#)
62. Lehner, L.; Myers, R.C.; Poisson, E.; Sorkin, R.D. Gravitational action with null boundaries. *Phys. Rev. D* **2016**, *94*, 084046. [\[CrossRef\]](#)
63. Moosa, M. Evolution of Complexity Following a Global Quench. *JHEP* **2018**, *3*, 031. [\[CrossRef\]](#)
64. Moosa, M. Divergences in the rate of complexification. *Phys. Rev. D* **2018**, *97*, 106016. [\[CrossRef\]](#)
65. Hashimoto, K.; Iizuka, N.; Sugishita, S. Time evolution of complexity in Abelian gauge theories. *Phys. Rev. D* **2017**, *96*, 126001. [\[CrossRef\]](#)
66. Chapman, S.; Eisert, J.; Hackl, L.; Heller, M.P.; Jefferson, R.; Marrochio, H.; Myers, R.C. Complexity and entanglement for thermofield double states. *SciPost Phys.* **2019**, *6*, 034. [\[CrossRef\]](#)
67. Guo, M.; Hernandez, J.; Myers, R.C.; Ruan, S.M. Circuit Complexity for Coherent States. *JHEP* **2018**, *10*, 011. [\[CrossRef\]](#)
68. Camargo, H.A.; Caputa, P.; Das, D.; Heller, M.P.; Jefferson, R. Complexity as a novel probe of quantum quenches: Universal scalings and purifications. *Phys. Rev. Lett.* **2019**, *122*, 081601. [\[CrossRef\]](#)
69. Doroudiani, M.; Naseh, A.; Pirmoradian, R. Complexity for Charged Thermofield Double States. *JHEP* **2020**, *1*, 120. [\[CrossRef\]](#)
70. Chapman, S.; Chen, H.Z. Charged Complexity and the Thermofield Double State. *JHEP* **2021**, *2*, 187. [\[CrossRef\]](#)
71. Bhattacharyya, A.; Nandy, P.; Sinha, A. Renormalized Circuit Complexity. *Phys. Rev. Lett.* **2020**, *124*, 101602. [\[CrossRef\]](#)
72. Bhargava, P.; Choudhury, S.; Chowdhury, S.; Mishara, A.; Selvam, S.P.; Panda, S.; Pasquino, G.D. Quantum aspects of chaos and complexity from bouncing cosmology: A study with two-mode single field squeezed state formalism. *SciPost Phys. Core* **2021**, *4*, 026 [\[CrossRef\]](#)
73. Lehnert, J.L.; Quintin, J. Quantum Circuit Complexity of Primordial Perturbations. *Phys. Rev. D* **2021**, *103*, 063527. [\[CrossRef\]](#)
74. Bhattacharyya, A.; Das, S.; Haque, S.S.; Underwood, B. Rise of cosmological complexity: Saturation of growth and chaos. *Phys. Rev. Res.* **2020**, *2*, 033273. [\[CrossRef\]](#)
75. Choudhury, S.; Dutta, A.; Ray, D. Chaos and Complexity from Quantum Neural Network: A study with Diffusion Metric in Machine Learning. *JHEP* **2021**, *4*, 138. [\[CrossRef\]](#)
76. Choudhury, S.; Chowdhury, S.; Gupta, N.; Mishara, A.; Selvam, S.P.; Panda, S.; Pasquino, G.D.; Singha, C.; Swain, A. Circuit Complexity From Cosmological Islands. *Symmetry* **2021**, *13*, 1301. [\[CrossRef\]](#)
77. Adhikari, K.; Choudhury, S.; Chowdhury, S.; Shirish, K.; Swain, A. Circuit complexity as a novel probe of quantum entanglement: A study with black hole gas in arbitrary dimensions. *Phys. Rev. D* **2021**, *104*, 065002. [\[CrossRef\]](#)
78. Adhikari, K.; Choudhury, S.; Pandya, H.N.; Srivastava, R. *PGW* Circuit Complexity. *arXiv* **2021**, arXiv:2108.10334.
79. Choudhury, S.; Selvam, S.P.; Shirish, K. Circuit Complexity from Supersymmetric Quantum Field Theory with Morse Function. *Symmetry* **2022**, *14*, 1656 [\[CrossRef\]](#)
80. Bai, C.; Li, W.H.; Ge, X.H. Towards the non-equilibrium thermodynamics of the complexity and the Jarzynski identity. *arXiv* **2021**, arXiv:2107.08608.
81. Caputa, P.; Kundu, N.; Miyaji, M.; Takayanagi, T.; Watanabe, K. Liouville Action as Path-Integral Complexity: From Continuous Tensor Networks to AdS/CFT. *JHEP* **2017**, *11*, 097. [\[CrossRef\]](#)
82. Caputa, P.; Magan, J.M. Quantum Computation as Gravity. *Phys. Rev. Lett.* **2019**, *122*, 231302. [\[CrossRef\]](#)
83. Boruch, J.; Caputa, P.; Takayanagi, T. Path-Integral Optimization from Hartle-Hawking Wave Function. *Phys. Rev. D* **2021**, *103*, 046017. [\[CrossRef\]](#)
84. Boruch, J.; Caputa, P.; Ge, D.; Takayanagi, T. Holographic path-integral optimization. *JHEP* **2021**, *7*, 016. [\[CrossRef\]](#)
85. Nielsen, M.A. A Geometric Approach to Quantum Circuit Lower Bounds. *arXiv* **2005**, arXiv:quant-ph/0502070.
86. Nielsen, M.A. Quantum Computation as Geometry. *Science* **2006**, *311*, 1133–1135 [\[CrossRef\]](#)
87. Dowling, M.R.; Nielsen, M.A. The Geometry of Quantum Computation. *Quantum Inf. Comput.* **2008**, *8*, 861–899 [\[CrossRef\]](#)
88. Nielsen, M.A.; Dowling, M.R.; Gu, M.; Doherty, A.C. Optimal control, geometry, and quantum computing. *Phys. Rev. A* **2006**, *73*, 062323 [\[CrossRef\]](#)
89. Watrous, J., Quantum Computational Complexity. In *Encyclopedia of Complexity and Systems Science*; Meyers, R.A., Ed.; Springer: New York, NY, USA, 2009; pp. 7174–7201.
90. Aaronson, S. The Complexity of Quantum States and Transformations: From Quantum Money to Black Holes. *arXiv* **2016**, arXiv:1607.05256.
91. Orús, R. Tensor networks for complex quantum systems. *APS Phys.* **2019**, *1*, 538–550. [\[CrossRef\]](#)
92. Nishioka, T.; Ryu, S.; Takayanagi, T. Holographic Entanglement Entropy: An Overview. *J. Phys. A* **2009**, *42*, 504008. [\[CrossRef\]](#)
93. Almheiri, A.; Dong, X.; Harlow, D. Bulk Locality and Quantum Error Correction in AdS/CFT. *JHEP* **2015**, *4*, 163. [\[CrossRef\]](#)
94. Swingle, B. Entanglement Renormalization and Holography. *Phys. Rev. D* **2012**, *86*, 065007. [\[CrossRef\]](#)
95. Caputa, P.; Magan, J.M.; Patramanis, D. Geometry of Krylov complexity. *Phys. Rev. Res.* **2022**, *4*, 013041. [\[CrossRef\]](#)
96. Parker, D.E.; Cao, X.; Avdoshkin, A.; Scaffidi, T.; Altman, E. A Universal Operator Growth Hypothesis. *Phys. Rev. X* **2019**, *9*, 041017. [\[CrossRef\]](#)
97. Roberts, D.A.; Stanford, D.; Streicher, A. Operator growth in the SYK model. *JHEP* **2018**, *6*, 122. [\[CrossRef\]](#)

98. Rabinovici, E.; Sánchez-Garrido, A.; Shir, R.; Sonner, J. Operator complexity: A journey to the edge of Krylov space. *JHEP* **2021**, *6*, 062. [[CrossRef](#)]
99. Barbón, J.L.F.; Rabinovici, E.; Shir, R.; Sinha, R. On The Evolution of Operator Complexity Beyond Scrambling. *JHEP* **2019**, *10*, 264. [[CrossRef](#)]
100. Jian, S.K.; Swingle, B.; Xian, Z.Y. Complexity growth of operators in the SYK model and in JT gravity. *JHEP* **2021**, *3*, 014. [[CrossRef](#)]
101. Dymarsky, A.; Gorsky, A. Quantum chaos as delocalization in Krylov space. *Phys. Rev. B* **2020**, *102*, 085137. [[CrossRef](#)]
102. Dymarsky, A.; Smolkin, M. Krylov complexity in conformal field theory. *Phys. Rev. D* **2021**, *104*, L081702. [[CrossRef](#)]
103. Balasubramanian, V.; Caputa, P.; Magan, J.; Wu, Q. A new measure of quantum state complexity. *arXiv* **2022**, arXiv:2202.06957.
104. Shankar, R. *Effective Field Theory in Condensed Matter Physics*; Cambridge University Press: Cambridge, MA, USA, 1998.
105. Cheung, C.; Creminelli, P.; Fitzpatrick, A.L.; Kaplan, J.; Senatore, L. The Effective Field Theory of Inflation. *JHEP* **2008**, *3*, 014. [[CrossRef](#)]
106. Weinberg, S. Effective Field Theory for Inflation. *Phys. Rev. D* **2008**, *77*, 123541. [[CrossRef](#)]
107. Agarwal, N.; Holman, R.; Tolley, A.J.; Lin, J. Effective field theory and non-Gaussianity from general inflationary states. *JHEP* **2013**, *5*, 085. [[CrossRef](#)]
108. Burgess, C.P. Intro to Effective Field Theories and Inflation. *arXiv* **2017**, arXiv:1711.10592.
109. Choudhury, S. Field Theoretic Approaches to Early Universe. Ph.D. Thesis, Indian Statistical Institute, Calcutta, India, 2016.
110. Choudhury, S. Can Effective Field Theory of inflation generate large tensor-to-scalar ratio within Randall–Sundrum single braneworld? *Nucl. Phys. B* **2015**, *894*, 29–55. [[CrossRef](#)]
111. Naskar, A.; Choudhury, S.; Banerjee, A.; Pal, S. EFT of Inflation: Reflections on CMB and Forecasts on LSS Surveys. *arXiv* **2017**, arXiv:1706.08051.
112. Pich, A. Effective field theory: Course. Les Houches Summer School in Theoretical Physics, Session 68: Probing the Standard Model of Particle Interactions. *arXiv* **1998**, arXiv:hep-ph/9806303.
113. Burgess, C.P. Introduction to Effective Field Theory. *Ann. Rev. Nucl. Part. Sci.* **2007**, *57*, 329–362. [[CrossRef](#)]
114. Donoghue, J.F. Introduction to the effective field theory description of gravity. *arXiv* **1995**, arXiv:gr-qc/9512024.
115. Donoghue, J.F. The effective field theory treatment of quantum gravity. *AIP Conf. Proc.* **2012**, *1483*, 73–94. [[CrossRef](#)]
116. Dubovsky, S.; Hui, L.; Nicolis, A.; Son, D.T. Effective field theory for hydrodynamics: Thermodynamics, and the derivative expansion. *Phys. Rev. D* **2012**, *85*, 085029. [[CrossRef](#)]
117. Crossley, M.; Glorioso, P.; Liu, H. Effective field theory of dissipative fluids. *JHEP* **2017**, *9*, 095. [[CrossRef](#)]
118. Choudury, S. Cosmic Microwave Background from Effective Field Theory. *Universe* **2019**, *5*, 155 [[CrossRef](#)]
119. Choudhury, S.; Gharat, R.M.; Mandal, S.; Pandey, N. Circuit Complexity in an interacting quenched Quantum Field Theory. *arXiv* **2022**, arXiv:2209.03372.
120. Adhikari, K.; Choudhury, S.; Roy, A. Krylov Complexity in Quantum Field Theory. *arXiv* **2022**, arXiv:2204.02250.

CKS: A Community-based K-shell Decomposition Approach using Community Bridge Nodes for Influence Maximization (Student Abstract)

Inder Khatri^{*1}, Aaryan Gupta^{*2}, Arjun Choudhry^{*1}, Aryan Tyagi^{*2}, Dinesh Kumar Vishwakarma¹, Mukesh Prasad³

¹ Biometric Research Laboratory, Delhi Technological University, New Delhi, India

² Delhi Technological University, New Delhi, India

³ School of Computer Science, FEIT, University of Technology Sydney, Sydney, Australia

{inderkhatri999, aaryan227227, choudhry.arjun, tyagiaryan82}@gmail.com, dinesh@dtu.ac.in, mukesh.prasad@uts.edu.au

Abstract

Social networks have enabled user-specific advertisements and recommendations on their platforms, which puts a significant focus on Influence Maximisation (IM) for target advertising and related tasks. The aim is to identify nodes in the network which can maximize the spread of information through a diffusion cascade. We propose a community structures-based approach that employs K-Shell algorithm with community structures to generate a score for the connections between seed nodes and communities. Further, our approach employs entropy within communities to ensure the proper spread of information within the communities. We validate our approach on four publicly available networks and show its superiority to four state-of-the-art approaches while still being relatively efficient.

Introduction

Online Social Networks are platforms for people to publicize their ideas and products. This approach, called Viral Marketing, raises the problem of Influence Maximisation (IM), which requires selecting k seed nodes to maximize the information spread in the network. Several approaches have been proposed for IM in recent years, based on local, semi-local, and global structures. Global structure-based approaches are efficient due to their consideration of whole network. They find out core nodes with maximum connectivity to the remaining network. However, in real-world situations, where number of seed nodes is very small, influence gets restricted to only a few sub-groups (or communities) in the network containing the selected core seed nodes. To overcome this, we instead consider community bridge nodes as influential seed nodes due to their connections to a larger number of communities, leading to simultaneous information propagation to these communities. We propose CKS centrality measure, which incorporates community structures and K-shell Decomposition to identify influential spreaders in a network. We define three novel measures: Community K-Shells, Community K-Shell Entropy, and CKS-Score, which qualitatively and quantitatively evaluate connections of a node to various communities.

^{*}These authors contributed equally.

Proposed Methodology

Obtaining Community K-Shell (CKS): Community K-Shell concept incorporates knowledge of information flow in a network. Obtaining CKSs comprises the following steps: identifying community structures using Louvain’s algorithm (Blondel et al. 2008); isolating communities by removing connections between different communities; and passing isolated communities through K-Shell algorithm (Kitsak et al. 2010) to obtain K-Shell scores particular to the community of each node (i.e., Community K-shell score). The higher the Community K-shell score of a node, the closer it is to the core of the given community.

Computing K-Shell Entropy (KSE): KSE evaluates connectivity of a node to different regions of a community. We formulate KSE for a node v corresponding to each community c as follows:

$$KSE_{v,c} = - \sum_{s=1}^{shells_c} K_s * \frac{\eta_{v,s}}{\eta_v} * \log\left(\frac{\eta_{v,s}}{\eta_v}\right) \quad (1)$$

$$\eta_v = \sum_{s'=1}^{shells_c} \eta_{v,s'} \quad (2)$$

where, $\eta_{v,s}$ is number of connections of node v with shell s for a given community, $shells_c$ represents the unique CKS in community c , and K_s represents K-value of given shell.

Influence of an activated node lasts only till 2-3 hops. Thus, for maximum influence over a given community, connectivity of a node to the community should be well-distributed across all its regions, which also reduces the chances of overlapping influence in a community. This requires a higher KSE score. Equation 1 represents the entropy submission over shells of the respective community weighted by the K-Value of the respective shell.

CKS-Score: CKS-Score evaluates the overall connectivity of a node to all adjacent communities to which it is connected. It is defined as:

$$CKS-Score(v) = \sum_{c=1}^{comm} NN_c * KSE_{v,c} * \eta_v \quad (3)$$

where, η_v is number of connections of node v with the respective community as shown in Equation 2, $KSE_{v,c}$ represents K-Shell Entropy for node v and community c , and NN_c represents number of nodes in community c . We compute CKS-Score for each node by considering its KSE values corresponding to each community weighted by the community’s size. This ensures a higher score for community bridge nodes connected to the most significant communities.

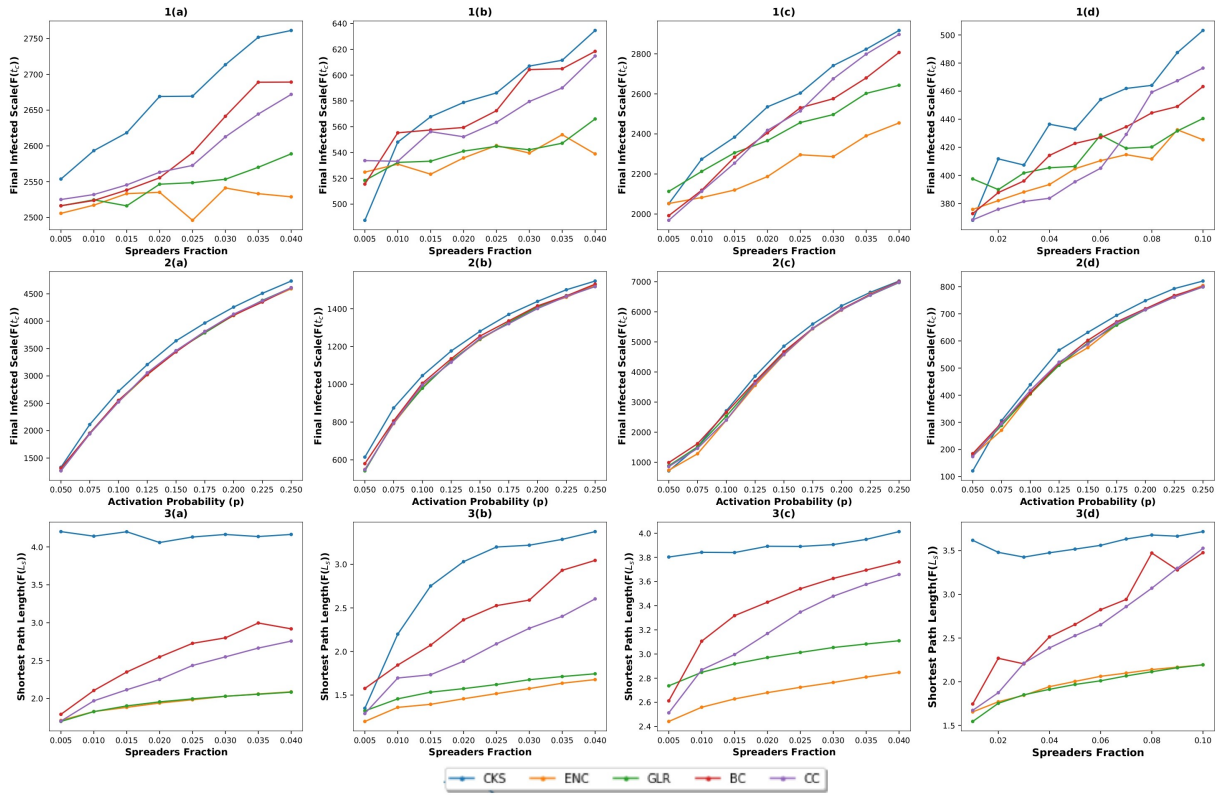


Figure 1: Results of our experiments on datasets (a) Twitch, (b) soc-Hamsterster (c) p2p-Gnutella04 (d) email-univ. (1) FIS after IC model simulation with activation probability 0.5. (2) FIS-v-p after IC model simulation with initial spreaders fraction 0.2. (3) ASPL among initial selected seed nodes vs spreaders fraction.

Dataset	CKS	ENC	GLR	BC	CC
Twitch	28.97	0.176	37.72	533.63	159.07
soc-Hamsterster	49.21	0.04	3.44	40.12	11.84
p2p-Gnutella04	83.21	0.19	54.41	1205.83	352.65
email-univ	0.83	0.01	0.71	10.09	1593.18

Table 1: Execution time for CKS and other approaches.

Experimental Results and Discussion

We evaluated CKS on four metrics: Final Infected Scale (FIS), Average Shortest Path Length (ASPL), Final infected scale vs Activation Probability (FIS-v-p), and Execution Time. We compared CKS with BC (Freeman 1977), CC (Okamoto, Chen, and Li 2008), ENC (Bae and Kim 2014), and GLR (Salavati, Abdollahpouri, and Manbari 2018) on four real-world datasets¹: Twitch, soc-Hamsterster, p2p-Gnutella04, and email-univ. We simulated each model 100 times using Independent Cascade (Goldenberg, Libai, and Muller 2001). Infection probability was set to 0.1. Table 1 and Figure 1 show our findings. We observed:

CKS consistently outperformed other core nodes-based approaches on FIS, ASPL, and FIS-v-p, while being reasonably efficient. This confirms our hypothesis that bridge nodes are more influential core nodes.

¹<https://networkrepository.com> , <https://snap.stanford.edu/snap>

CKS showed a higher inter-spreader distance than competing approaches, while also having a higher infection rate, leading to much lesser overlap between the influence of spreaders (or *rich club effect*).

References

- Bae, J.; and Kim, S. 2014. Identifying and ranking influential spreaders in complex networks by neighborhood coreness. *Physica A Statistical Mechanics and its Applications*, 395: 549–559.
- Blondel, V.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics Theory and Experiment*, 2008.
- Freeman, L. 1977. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40: 35–41.
- Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12: 211–223.
- Kitsak, M.; Gallos, L. K.; Havlin, S.; Liljeros, F.; Muchnik, L.; Stanley, H. E.; and Makse, H. A. 2010. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11): 888–893.
- Okamoto, K.; Chen, W.; and Li, X.-Y. 2008. Ranking of Closeness Centrality for Large-Scale Social Networks. In *Frontiers in Algorithmics*, volume 5059, 186–195. ISBN 978-3-540-69310-9.
- Salavati, C.; Abdollahpouri, A.; and Manbari, Z. 2018. Ranking nodes in complex networks based on local structure and improving closeness centrality. *Neurocomputing*, 336.

Comparative Performance of DVR and STATCOM for Voltage Regulation in Radial Microgrid with High Penetration of RES

Case Study

Ritika Gour

Delhi Technological University,
Electrical Engineering Department, Delhi, India
riti113@gmail.com

Vishal Verma

Delhi Technological University,
Electrical Engineering Department, Delhi, India

Abstract – In recent years, the penetration of renewable energy sources (RES) in microgrids and distribution system feeders has increased manifold. Moreover, the advancement of power electronics-based devices in the distribution system has significantly increased the number of sensitive loads. Variations in the voltage under intermittent RES create functional problems with sensitive loads, necessitating voltage regulators (VR) installation. In this paper, two custom power devices: dynamic voltage restorer (DVR) and static synchronous compensator (STATCOM), used for voltage regulation in a microgrid, are investigated under different operating conditions. The efficacy of DVR and STATCOM for voltage regulation in an 11-node radial microgrid with high penetration of RES is simulated under a MATLAB Simulink environment. Furthermore, the simulated microgrid voltage profile results are analyzed to evaluate the efficacy of both voltage regulators.

Keywords: RES, intermittency, CPL, microgrid, voltage regulation, DVR, STATCOM

1. INTRODUCTION

A microgrid is a local energy grid with a group of connected energy sources and loads that usually operate in synchronization with the conventional grid but can also operate independently in the event of any anomaly. Various renewable energy sources (RES) can be connected to the microgrid such as solar panels, wind turbines, etc. With the advancement of technology, the diminishing supply of conventional power sources and many environmental and socio-economic factors have raised RES penetration in the microgrid [1-4]. Increased renewable energy in the microgrid provides numerous benefits, like increased local power availability, low-cost, clean energy, increased reliability and resilience, reduced grid congestion and peak loads, etc. However, the intermittent nature of RES creates power fluctuations and large variations in the voltage profile. As a result of the unpredictable nature of these RES, the microgrid operator is unable to schedule the load, posing a risk to the entire system. Furthermore, the widespread use of power electronics-based devices in residential and commercial loads has significantly increased the number of sensitive loads. Power and voltage variations due to high penetration of RES may result in the maloperation of these sensi-

tive loads, damage to equipment, and cascading faults. These challenges deteriorate the power quality of the microgrids which causes instability and monetary losses for both the microgrid and the consumer. [5].

The variability of demands and intermittency of renewable sources necessitates the methods to deal with variation in the voltage of the microgrid. Mitigation of the negative effects of intermittent RES is achieved by adopting different methods for regulating the voltage of the microgrid. The static synchronous compensator (STATCOM) and the dynamic voltage restorer (DVR) are two custom power devices that are generally installed in the distribution system to regulate voltage profile [8-10]. DVR regulates the voltage at the point of common coupling (PCC) by supplying/absorbing the voltage in series with the PCC voltage and is connected to the feeder through an injection transformer [5] [9-12]. On the other hand, STATCOM is connected in shunt with the feeder, thereby allowing the supply or absorption of current from the feeder, with an effect on the voltage as well [10] [13].

The efficacy of DVR and STATCOM for voltage regulation in microgrids with significant RES penetration is in-

vestigated in this paper by simulating an 11-node grid-connected radial microgrid in the MATLAB/Simulink environment. Findings of the simulation are then used to draw a microgrid voltage profile from the grid mains to the last node in the presence of the STATCOM and DVR, one by one, and their efficacy is assessed.

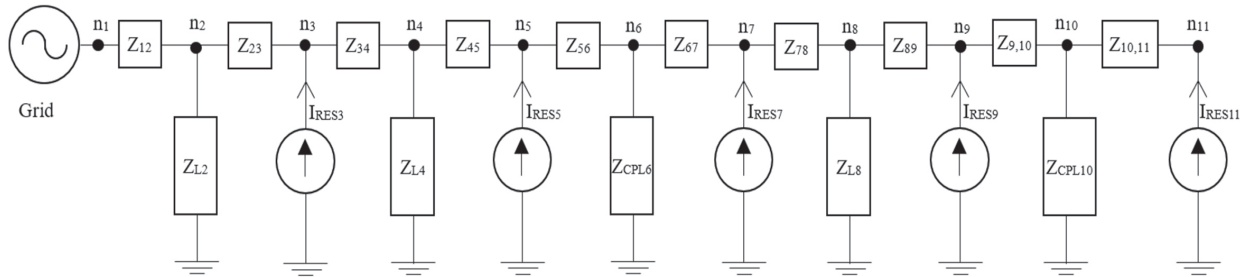


Fig. 1. Single line diagram of system under consideration

2. SYSTEM CONFIGURATION

The system under consideration is shown in Fig. 1. Considered a system an 11-node grid connected radial microgrid. The voltage source/grid is assumed to be a reference node (n_1). RES sources (I_{RES_i}) and loads Z_{L_i} are connected alternatively throughout the radial microgrid (where 'i' is the node number with respect to the reference node). All the RESs are connected on odd nodes (n_3, n_5, n_7, n_9 and n_{11}) and all the loads are connected on even nodes (n_2, n_4, n_6, n_8 and n_{10}) with respect to the reference node (n_1). In the considered system RESs are connected to the microgrid as a current source feeding the microgrid through a current-controlled voltage source converter (VSC). Loads connected to the microgrid are of two types: constant impedance load (CIL) and constant power load (CPL). CILs are connected at node 2, node 4, and node 8. Generalized expression for the impedance Z_{L_i} of the CIL is given as:

$$Z_{L_i} = \frac{(V_{n_i}(\text{rated}))^2}{P_{L_i} + jQ_{L_i}} \quad (1)$$

Where $V_{n_i}(\text{rated})$ is rated voltage at the node and $P_{L_i} + jQ_{L_i}$ is rated power of the load (power consumed by the load at rated voltage).

CPLs are connected at node 6 and node 10. The impedance of CPL is dynamic in nature, it shows a negative impedance characteristic, and power drawn (P_{CPL_i}) remains constant irrespective of the instantaneous voltage (V_{n_i}) at its terminal. CPL changes load current according to the voltage level so that it can draw a constant power from the microgrid. Generalized expression for the dynamic impedance of the CPL is given as [14-15]:

$$Z_{CPL_i} = \frac{(V_{n_i})^2}{P_{CPL_i}} \quad (2)$$

For ease of study, certain assumptions are made which are as follows:

Section 2 describes the block diagram of the considered system, and Section 3 discusses the block diagram and phasor diagram of both voltage regulators DVR and STATCOM. Further Section 4 includes MATLAB Simulink and performance evaluations of both STATCOM and DVR.

- RES and loads are placed alternatively on the microgrid with uniform distancing.
- Microgrid is part of distribution feeder with R/X ratio ≈ 8 such that feeder impedance is $0.642 + j0.0833 \Omega/\text{km}$ [16].
- Length of each section is 0.5 km.

3. OPERATING PRINCIPLE OF DVR AND STATCOM

3.1 OPERATING PRINCIPLE OF DVR

DVR is a custom power device connected in series with the feeder as seen in Fig. 2(a). DVR generally comprises of injection transformer, ripple filter, VSC and DC link. The DC voltage on the DC link is converted to AC by VSC as required for voltage regulation. A further ripple filter is used to filter the AC voltage generated from VSC before the injection transformer injects it into the line.

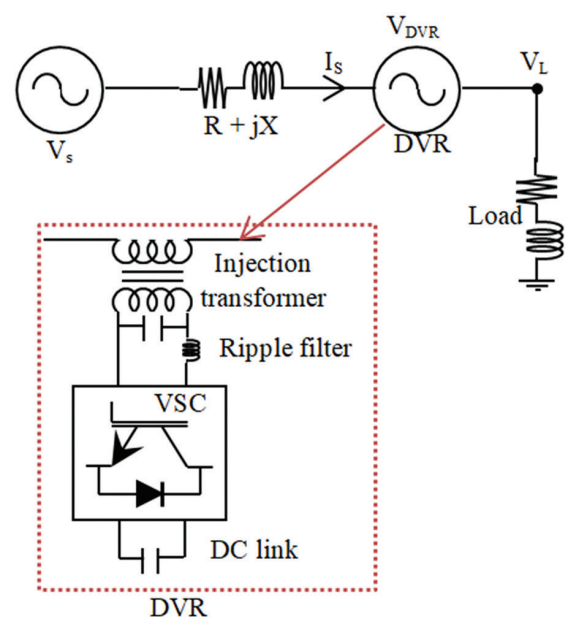


Fig. 2.(a) Block diagram for voltage regulation by DVR

The injection transformer is a low leakage impedance transformer that connects the DVR to the feeder in series. On the DC side of the VSC, a capacitor is typically present to maintain the DC link voltage. To increase the range of regulation of DVR, batteries can also be placed on the DC side. The phasor diagram for the DVR voltage regulation is shown in Fig. 2(b). From the phasor diagram, it can be observed that the net voltage at the terminal of DVR is equal to the phasor sum of the previous terminal voltage and DVR injected/absorbed voltage.

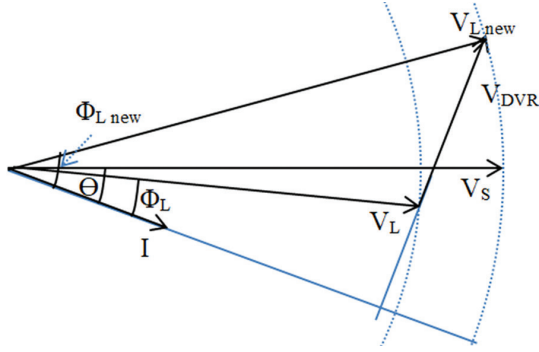


Fig. 2.(b) Phasor diagram of DVR for voltage regulation

Synchronous reference frame theory is used to control the gating pulses of VSC in DVR, which are subsequently used for controlling the injection of requisite voltage from the DVR [9]. Fig. 2(c) shows the block diagram of the control scheme.

The difference between the measured voltage, from the terminals of DVR, and the rated voltage is passed through the proportional-integral (PI) controller for generating the DVR reference (injection/absorption) voltage. Usually, the voltage regulation is done with the reactive component of voltage, which means that the DVR reference voltage is generally the quadrature-axis (q-axis) component (V_q). Nevertheless, if the voltage required to regulate the terminal voltage exceeds the limit of the DVR's reactive capacity, then V_q is reduced with a simultaneous increase of V_d to regulate the terminal voltage of the DVR. Subsequently, reverse park transformation is performed on the reference V_d and V_q through dq to abc transformation. The gating pulses of VSC are generated by passing the resultant abc reference signal through the pulse width modulator (PWM) block. These gating pulses are further used to control the ON/OFF different switches of VSC, for requisite voltage injection, to regulate the voltage at the DVR terminal.

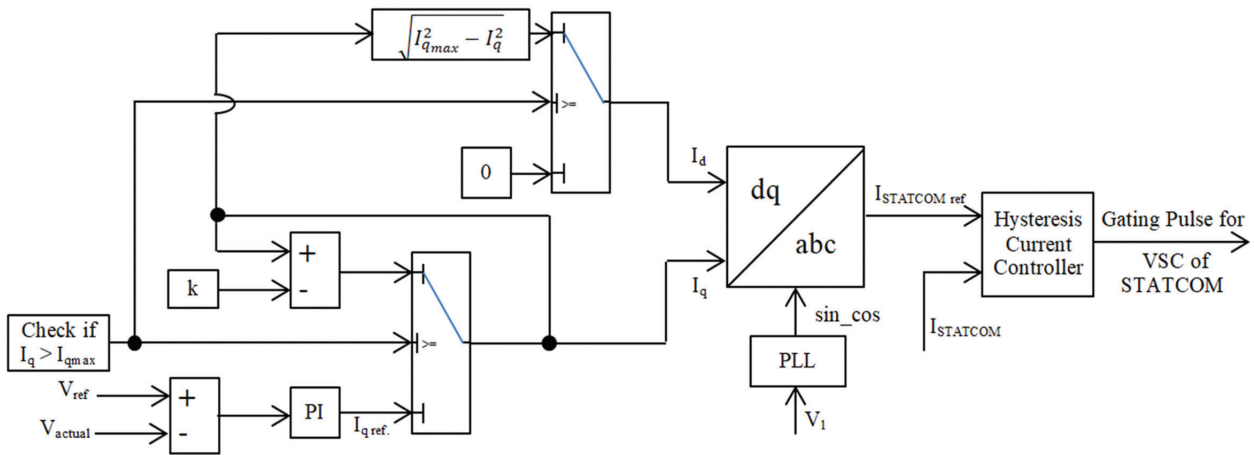


Fig. 2. (c) Block diagram of control of STATCOM

3.2 OPERATING PRINCIPLE OF STATCOM

Another custom power device investigated in the paper is STATCOM, which is in shunt with the feeder, as illustrated in Fig. 3(a). The VSC converts DC to AC, and after passing through the ripple filter, the output current from the VSC is sent to the microgrid. In the design shown in Fig. 3(a), it is connected in parallel to the load connected at nodes 6 and 10. Fig. 3(b) shows a phasor diagram for its voltage regulation.

The synchronous reference frame theory is also used to control the VSC of STATCOM. Fig. 3(c) depicts the control approach for generating the gating pulses for the VSC of STATCOM. The controller senses the volt-

age at its terminal, and then the difference between the sensed voltage and the reference voltage is passed through the PI controller, generating the q-axis component of the current (I_q). As shown in the block diagram, when the STATCOM approaches its reactive power limit, the q-axis component is reduced, and the d-axis component (I_d) is increased, as done in the case of DVR. To generate the reference STATCOM current ($I_{STATCOMref}$), the reference values of I_d and I_q are transformed into reference abc signal through reverse PARKs transformation. Furthermore, the gating pulses for VSC switches are generated by passing STATCOM reference ($I_{STATCOMref}$) current and actual ($I_{STATCOM}$) current through a hysteresis current controller.

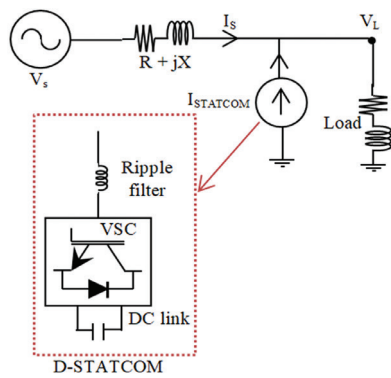


Fig. 3.(a) Block diagram for voltage regulation by STATCOM

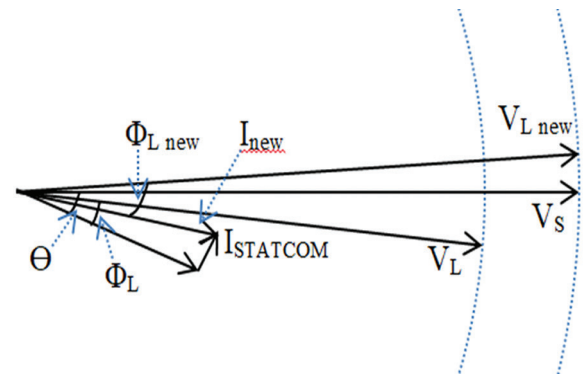


Fig. 3.(b) Phasor diagram of STATCOM for voltage regulation

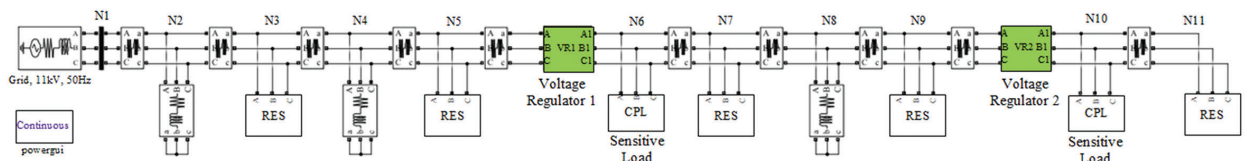


Fig. 4.(a) MATLAB/Simulink diagram for considered radial microgrid

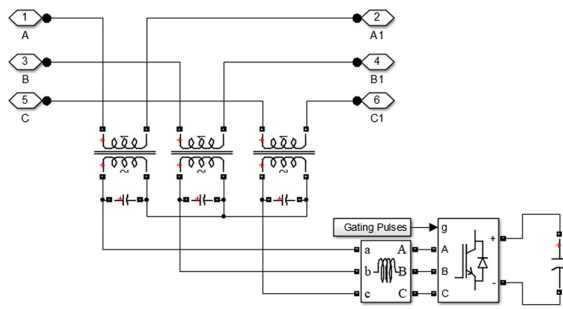


Fig. 4.(b) MATLAB/Simulink diagram for DVR

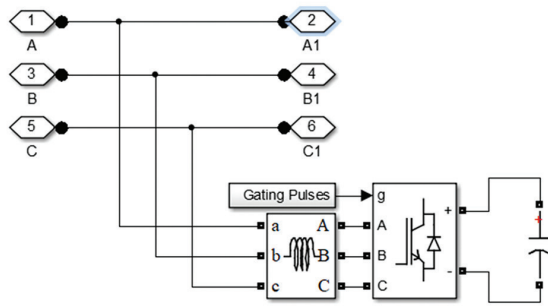


Fig. 4.(c) MATLAB/Simulink diagram for STATCOM

4. PERFORMANCE EVALUATION OF DVR AND STATCOM

Simulations in the MATLAB/Simulink environment are done to investigate the efficacy of DVR and STATCOM in the considered 11 nodes radial microgrid. MATLAB simulation diagram of the microgrid with RES and voltage regulator is shown in Fig. 4(a). The voltage regulator can be either DVR or STATCOM and are connected at node 6 and node 10 and the expanded figures are shown in Fig. 4(b) and Fig. 4(c) for DVR and

STATCOM respectively. Parameters considered for the simulation of the above-mentioned configuration are listed in Table 1.

Table 1. Parameters used for simulation of the considered system

Parameters	Values
Source/Grid Voltage	11 kV, 50 Hz
Feeder impedance	$0.642 + j 0.083 \Omega/\text{km}$, Length of each section = 0.5 km
RES rating	0.75 MW for each unit of RES
CPL rating	1 MW for each CPL load
Constant Impedance load (CIL)	0.5 MVA each load, $R = 193 \Omega$, $L = 462 \text{ mH}$
DVR	0.5 MVA for each unit.
STATCOM	0.5 MVA for each unit.

Both the STATCOM and the DVR are assigned the same rating to evaluate their regulation proficiency. The effect of voltage control is reported for different levels of RES penetration: (i) RES penetration is high with perturbing loads ($I_{\text{RES}} = 60 \text{ A}$). (ii) RES penetration is low with perturbing loads ($I_{\text{RES}} = 25 \text{ A}$).

(i) High penetration of RES

In this case, high penetration of RES is considered, as each RES is generating maximum power as per its rating. In this scenario, the RES current $I_{\text{RES}} = 60 \text{ A}$ (peak), resulting in each unit contributing 0.75 MW of power to the microgrid. The voltage profile is used to assess the responsiveness of DVR and STATCOM for voltage regulation in this scenario for several loading conditions. All the graphs, for each condition, show the voltage profile of the microgrid with considered loading conditions for:

- without any RES and voltage regulator connected to it.
- with RES and without any voltage regulator connected to it.
- with RES and DVR as the voltage regulator.
- with RES and STATCOM as voltage regulators so that the regulation is accomplished by real current injection.
- with RES and STATCOM as voltage regulators so that the regulation is accomplished by reactive current injection.

a. High penetration of RES and rated loading condition:

The voltage profile of the microgrid with high penetration of RES at rated loading conditions, with each CIL drawing 0.5 MVA and each CPL drawing 1 MW power, is shown in Fig. 5(a). The voltage profile reveals that both the STATCOM and the DVR can regulate the voltage at their point of common coupling (PCC), i.e., at node 6 and node 10. DVR can regulate voltage with reactive power injection (V_{qDVR}), whereas STATCOM ($I_{dSTATCOM}$) can do the same with real power injection. Furthermore, Fig. 5(a) depicts the STATCOM's ($I_{qSTATCOM}$) voltage profile while it is regulating through reactive power, yet the voltage does not reach the rated value even at the point of common coupling, but the profile is improved slightly.

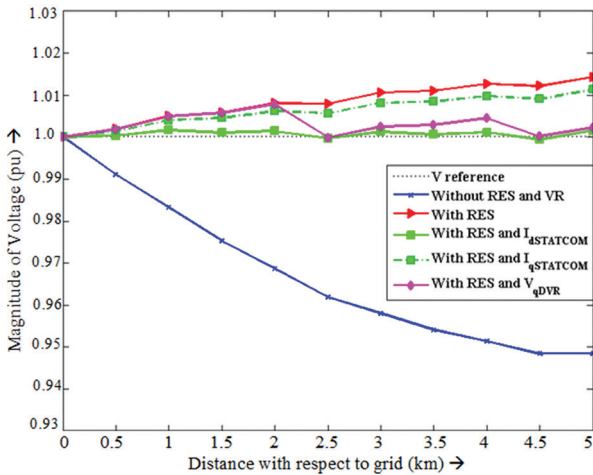


Fig. 5.(a) Voltage profile of microgrid with high RES penetration and rated loading condition

b. High penetration of RES and light loading condition:

The voltage profile of the microgrid with high penetration of RES and light loading conditions, i.e., 20% of the rated value, can be seen in Fig. 5(b). In this instance, the DVR improves the overall voltage profile of the microgrid by regulating the voltage at its two connecting points by absorbing a voltage that is a combination of both real and reactive power. However, since the range

required in this instance is relatively large, the STATCOM appears to be inadequate in regulating the voltage using either real or reactive current transactions.

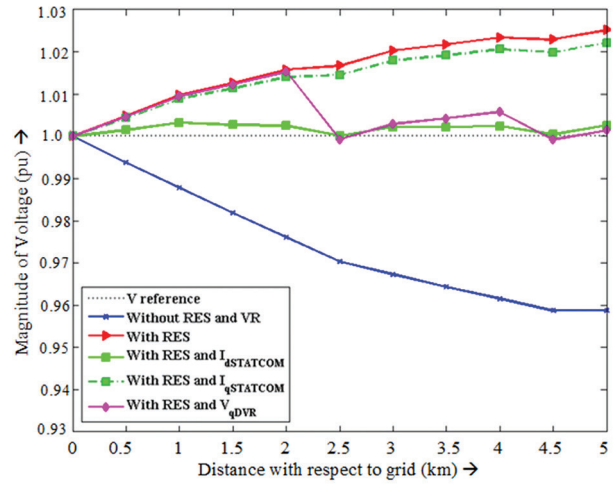


Fig. 5.(b) Voltage profile of microgrid with high RES penetration and light loading condition

c. High penetration of RES and heavy loading conditions:

Figure 5(c) depicts the voltage profile of a microgrid with high-RES penetration and heavy loading conditions of about 150% of the rated value. The voltage profile of the microgrid is enhanced with higher injection from RES because the loads are quite high.

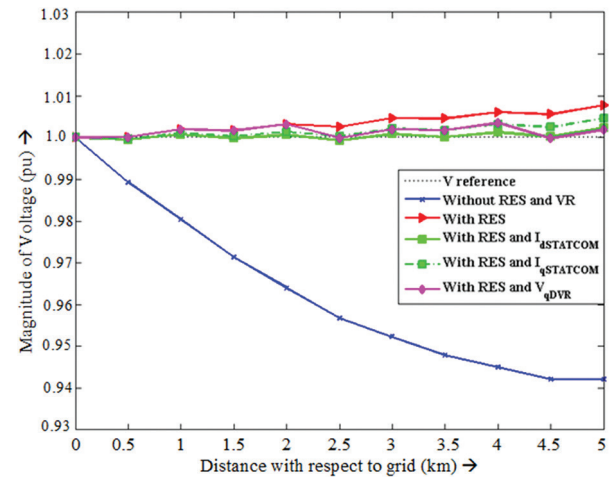


Fig. 5.(c) Voltage profile of microgrid with high RES penetration and heavy loading condition

As a result, both DVR and STATCOM improve the voltage profile even further.

(ii) Low penetration of RES and normal loading condition:

Injection from all the RES is reduced to 40% of the maximum value (given in the preceding scenario), with $I_{RES} = 20$ A (rms), and each unit now supplies 0.30 MW of power to the microgrid.

a. Low penetration of RES and normal loading condition:

Figure 5(d) depicts the voltage profile of the microgrid under low RES penetration and rated load conditions. In this scenario, the DVR utilizes reactive voltage injection to regulate the voltage at its terminals, but the STATCOM employs real current injection.

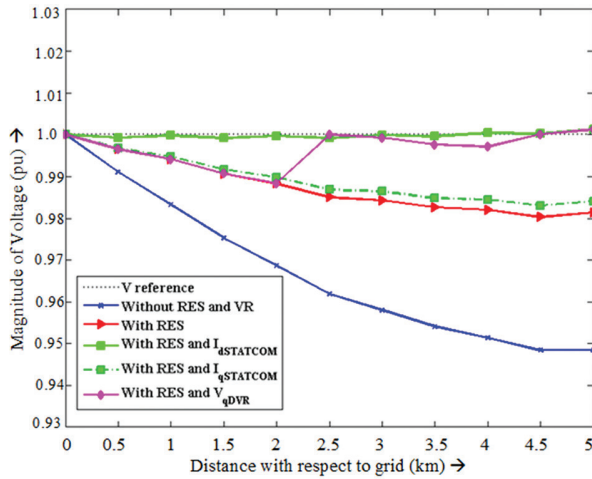


Fig. 5.(d) Voltage profile of microgrid with low RES penetration and rated loading condition

b. Low penetration of RES and light loading condition:

Figure 5(e) depicts the voltage profile of the microgrid at low RES penetration and light loading conditions, i.e. 20% of the rated value. As soon as the RES starts supplying power to the microgrid, the voltage profile improves. The voltage profile is further improved by both DVR and STATCOM.

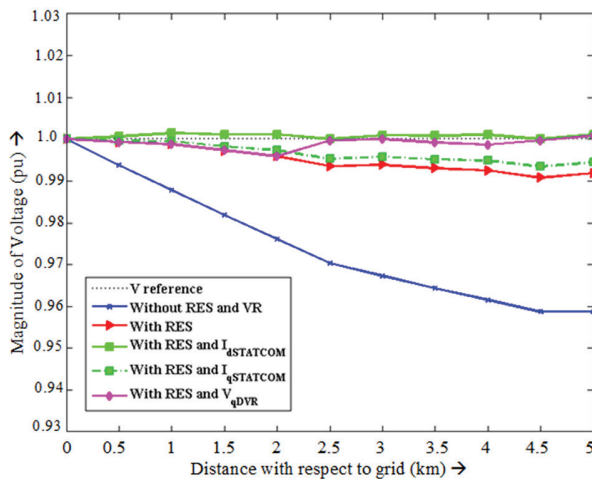


Fig. 5.(e) Voltage profile of microgrid with low RES penetration and light loading condition

c. Low penetration of RES and heavy loading condition:

Figure 5(f) depicts the voltage profile of the microgrid at low RES penetration and heavy loading conditions,

i.e. 150% of the rated value. Because the voltage variation is large in this scenario, the DVR requires both real and reactive power to perform, but it regulates the voltage, whereas the STATCOM is unable to do so even with both the real and reactive current injection.

5. CONCLUSION

The efficacy of DVR and STATCOM has been investigated and presented in this paper for improving the voltage profile of the 11-node radial microgrid with intermittent RES source in the MATLAB simulation environment. The PI controller and synchronous reference frame theory are used to control STATCOM and DVR. Voltage profiles obtained from the simulations demonstrate the comparison of the efficacy of DVR and STATCOM for voltage regulation in the microgrid. Under significant voltage variations, STATCOM can regulate the voltage only with real power support (through battery), while DVR allows voltage regulation with reactive power unless the voltage variations are extremely high. Utilization of STATCOM as a voltage regulator significantly increases the cost of the system as battery cost is also added to the system. In conclusion, the STATCOM is appropriate for voltage regulation if the range of compensation required is small, but as the range of compensation required widens, the DVR appears to be a superior solution.

6. REFERENCES

- [1] D. Ma, M. Liu, H. Zhang, R. Wang, X. Xie, "Accurate Power Sharing and Voltage Regulation for AC Microgrids: An Event-Triggered Coordinated Control Approach", IEEE Transactions on Cybernetics, 2021, pp. 1-11.
- [2] J. von Appen, M. Braun, T. Stetz, K. Diwold, D. Geibel, "Time in the Sun: The Challenge of High PV Penetration in the German Electric Grid", IEEE Power and Energy Magazine, Vol. 11, No. 2, 2013, pp. 55-64.
- [3] N. S. Jayalakshmi, P. B. Nempu, "Performance Enhancement of a Hybrid AC-DC Microgrid Operating with Alternative Energy Sources Using Supercapacitor", International Journal of Electrical and Computer Engineering Systems, Vol. 12 No. 2, 2021.
- [4] T. V. Krishna, M. K. Maharana, C. K. Panigrahi, "Integrated Design and Control of Renewable Energy Sources for Energy Management", Engineering, Technology & Applied Science Research, Vol. 10, No. 3, 2020, pp. 5857-5863.
- [5] V. Verma, R. Gour, "OLTC-DVR hybrid for voltage regulation and averting reverse power flow in the

micro-grid with intermittent renewable energy sources", Proceedings of the IEEE Industrial Electronics and Applications Conference, Kota Kinabalu, Malaysia, 2016, pp. 81-87.

- [6] Y. He, M. Wang, Z. Xu, "Enhanced Voltage Regulation of AC Microgrids with Electric Springs", Proceedings of the IEEE Applied Power Electronics Conference and Exposition, Anaheim, CA, USA, 17-21 March 2019, pp. 534-539.
- [7] A. Pimenta, P. B. C. Costa, G. M. Paraíso, S. F. Pinto, J. F. Silva, "Active Voltage Regulation Transformer for AC Microgrids", Proceedings of the IEEE 9th International Power Electronics and Motion Control Conference, Nanjing, China, 2021, pp. 2012-2017.
- [8] A. Ghosh, G. Ledwich, "Power Quality Enhancement Using Custom Power Devices", The Springer International Series in Engineering and Computer Science book series, 2002, pp 113-136.
- [9] R. Gour, V. Verma, "Voltage Regulation in a Radial Microgrid with High RES Penetration: Approach-Optimum DVR Control", Engineering, Technology & Applied Science Research, Vol. 12, No. 4, 2022, pp. 8796-8802.
- [10] H. M. A. Rashid, S. A. Jumaat, S. H. N. Yusof, S. A. Zulkifli, "Modeling the Grid Connected Solar PV (GCPV) System with D-STATCOM to Improve Stability System", Proceedings of the IEEE International Conference in Power Engineering Application, Shah Alam, Malaysia, 7-8 March 2022, pp. 1-6.
- [11] A. H. Soomro, A. S. Larik, M. A. Mahar, A. A. Sahito, A. M. Soomro, G. S. Kaloi, "Dynamic Voltage Restorer—A comprehensive review", Energy Reports, Vol. 7, 2021, pp. 6786-6805.
- [12] S. F. Al-Gahtani et al. "A New Technique Implemented in Synchronous Reference Frame for DVR Control Under Severe Sag and Swell Conditions", IEEE Access, Vol. 10, 2022, pp. 25565-25579.
- [13] L. E. Christian, L. M. Putranto, S. P. Hadi, "Design of Microgrid with Distribution Static Synchronous Compensator (STATCOM) for Regulating the Voltage Fluctuation", Proceedings of the IEEE 7th International Conference on Smart Energy Grid Engineering, Oshawa, ON, Canada, 12-14 August 2019, pp. 48-52.
- [14] A. P. N. Tahim, D. J. Pagano, M. L. Heldwein, E. Ponce, "Control of interconnected power electronic converters in dc distribution systems", Proceedings of the XI Brazilian Power Electronics Conference, 2011, pp. 269-274.
- [15] N. Ghanbari, S. Bhattacharya, "Constant Power Load Challenges in Droop Controlled DC Microgrids", Proceedings of the 45th Annual Conference of the IEEE Industrial Electronics Society, Lisbon, Portugal, 14-17 October 2019, pp. 3871-3876.
- [16] A. Engler, N. Soultanis, "Droop control in LV-grids", Proceedings of the International Conference on Future Power Systems, Amsterdam, Netherlands, 18 November 2005, pp. 1-6.

Current Limiting Reactors based Time-Domain Fault Location for High Voltage DC Systems with Hybrid Transmission Corridors

Vaibhav Nougain, and Sukumar Mishra, *Senior Member, IEEE*

Abstract—Accurately locating the fault distance helps in the rapid restoration of the isolated line back into the system. This paper proposes a novel time-domain-based algorithm to determine accurate fault location in high voltage direct current (DC) systems with hybrid DC transmission corridors i.e., a combination of underground cables (UGC) and overhead lines (OHL). The work gives a fault location method for a 2-segment and 3-segment hybrid transmission corridor (HTC) and then generalizes the analysis for an n-segment HTC. The algorithm offers flexibility to locate faults ranging from the homogeneous transmission line to n-segment HTC. The algorithm uses a simplified unit resistance-inductance (R-L) representation of transmission lines along with time-domain based measurements i.e., terminal voltage, current and voltage across current limiting reactors (CLRs). Power Systems Computer Aided Design/Electromagnetic Transients including DC (PSCAD/EMTDC) based simulations are used to validate robust performance against variation of key implementation parameters like type of faults, fault resistance, fault location, sampling frequency, and white Gaussian noise (WGN) in measurement. Further, the fault location calculation is analyzed under parameter variation i.e., change in the true value of unit resistance and unit inductance of line or cable and the true value of DC link capacitance.

Index Terms—Power system protection, fault location, CIGRE benchmark, HVDC transmission systems, power system fault diagnosis, transient analysis, power system measurements.

I. INTRODUCTION

HIGH Voltage Direct Current (HVDC) transmission has found its application for long transmission lines over its HVAC counterpart [1]-[6]. The reason is its simple controllability, the capability to transmit more amount of active power due to the absence of reactive current in DC, fewer conversion stages; better scalability, and very low harmonics because of modular multilevel converter (MMC) technology for HVDC systems [7]-[9]. Additionally, multi-terminal HVDC configuration ensures system reliability with continuous operation of power transfer even under DC faults. However, since there are numerous bus terminals finding different paths (with different impedances) to contribute to the fault current, the fault current is relatively higher compared to the conventional point-to-point HVDC system, which has a contribution from only two terminals. Moreover, the combination of overhead lines (OHL) and underground cables (UGC) is used as a segmented HVDC transmission system when two networks separated by water

need to be connected [10]. Underground cables are also utilized to connect offshore wind farms to the existing grid through overhead lines [11]. In general, hybrid DC transmission systems are often used to transmit electricity across water bodies, such as rivers, lakes, and seas where underground, underwater, or submarine cables are used [10], [14]. As the converter stations may not be always located close to the shore of the water body, a combination of OHL and UGC is used as the transmission mediums [14]. A few practical examples of HVDC systems with multiple types of transmission mediums are the Kii Channel HVDC System [32], Anan-Kihoku HVDC System [33], Hokkaido-Honshu HVDC Link [34] of Japan, and the Basslink HVDC interconnector system [35] of Australia. The lengths of the cable sections vary widely, from 44 km in the Hokkaido-Honshu scheme to slightly under 200 km in the Basslink scheme. The lengths of the overhead line sections could vary from 10 km to more than 100 km [14]. This calls for a secure and rapid protection scheme for such segmented multi-terminal HVDC systems.

Identifying and locating DC faults are two challenges on the protection front for such HVDC configurations. Identifying the fault helps in the isolation of the faulty line and accurately locating the DC fault ensures rapid restoration of the isolated faulty line into the transmission system.

Considering fault location using travelling-wave-based methods, the authors in [12] use single-ended and double-ended buffered voltage signal frames around the fault-detection time, segmented via an optimization process. The method is fairly accurate for high impedance faults (HIFs) and white Gaussian noise (WGN) in measurement. The authors in [13] propose a fault section identification method based on energy relative entropy. The difference between the forward and backward current traveling wave is represented by the S-transform energy relative entropy. However, the application of the methods in [12]-[13] is limited to homogeneous transmission corridors only. Considering fault location of hybrid transmission corridor (HTC) as the objective, the literature covers a few travelling-wave-based methods [11], [14]. The authors in [14] use the double-terminal method to accurately detect the surge arrival time. The method is limited to a fault resistance of 100Ω and requires a high sampling frequency. The authors in [11] use the continuous wavelet transform applied to a series of line current measurements. The method addresses the problem of [14]. However, multiple distributed current

Vaibhav Nougain and Sukumar Mishra are with the Department of Electrical Engineering, Indian Institute of Technology, Delhi, New Delhi, 110016 India. e-mail: (nougainvaibhav@gmail.com, sukumariitdelhi@gmail.com).

sensors increase the cost of implementation of the fault location method. Additionally, the insensitivity of travelling-wave-based methods towards high impedance faults (HIFs) and the requirement of high sampling frequency are some general indispensable problems with travelling-wave based methods.

Another way of locating fault is by using a time-domain-based method [15]-[18]. Time-domain methods have been conventionally used in the fault analysis of converter-based alternating current (AC) systems [16]. For a homogeneous DC transmission corridor, authors in [17]-[18] use simplified unit resistance-inductance (R-L) representation of transmission lines to model overhead lines (OHL) along with time-domain-based current and voltage measurements. The methods in [17]-[18] are fairly accurate even for high fault resistances and low sampling frequency, overcoming the inherent problem of travelling-wave-based methods. However, the methods have mathematical limitations and their application is limited to homogeneous DC transmission corridors. This is a result of considering singular respective values of R-L parameters to represent the cable or line. Since, the analysis does not consider the possibility of a segmented corridor, the methods are suited for homogeneous transmission corridors only. The problem of fault location for segmented HTC has not been studied using time-domain-based methods in the literature so far. Such methods can accurately locate HIFs as well using low sampling rate.

The idea of the proposed novel work is to explore the application of the time-domain-based fault location method for segmented HTC i.e., the combination of UGC and OHL. Terminal voltage and current measurements and voltage across current limiting reactors (CLRs) are used with a simplified R-L model representation for transmission lines to propose the fault location algorithm for segmented multi-terminal transmission corridors. The method can be applied to homogeneous transmission corridors as well. The attributes of the proposed fault location algorithm are as follows:

- Unlike existing time-domain-based double terminal methods [17]-[18], the method is flexible for systems with either homogeneous or non-homogeneous combination of cables and lines for transmission.
- Unlike travelling wave-based double-terminal methods [11], [14], the proposed work requires a sampling frequency as low as 5 kHz and is fairly accurate for high resistance faults.
- The proposed method can precisely indicate which segment of the faulty HTC is subjected to a fault. The proposed fault location method is also generalized for an n -segment HTC for wider application.
- The method is robust to variation of key parameters like type of faults, fault resistance, fault location, and white Gaussian noise in the measurement.

The grounding capacitance for OHL is within $0.01\text{--}0.02\mu\text{F}/\text{km}$ while that for UGC is within $0.5\mu\text{F}/\text{km}$ [22]. The equivalent DC link capacitor for an HVDC system

is around $10^3\text{--}10^4\mu\text{F}$ [18]. Hence, the fault contribution from the grounding capacitance of cables and lines can be ignored [18] for a length up to 200km (keeping a margin of 10-100 times considering the current contribution of UGC grounding capacitance in comparison to DC capacitance contribution). This means that a simplified R-L representation can be used for fairly accurate fault location for both UGC and OHL with a length up to 200km. Another intuitive problem with the time-domain-based method for hybrid transmission corridors is joint resistance between segments of UGC and OHL. The value of joint resistance as reported in [19]-[20] is in order of $10\text{--}10^3\mu\Omega$ whereas resistance for OHL and UGC is in the order of $10^2\text{--}10^3\text{m}\Omega/\text{km}$ (see Table II). As a result of which, neglecting joint resistance in the analysis does not affect the accuracy of the fault location method. Additionally, the fault location accuracy, in general, is not sensitive to resistance as discussed in Section IV(E). The rest of the proposed work is organized as follows. Section II introduces the test system. Section III gives the proposed algorithm for a 2-segment, 3-segment and n -segment HTC. Section IV gives the validation of the proposed fault location algorithm. Section V concludes the work.

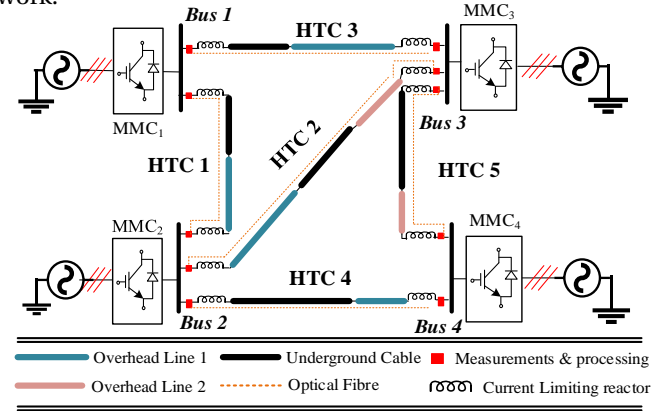


Fig. 1: Four-terminal MMC-HVDC system with HTC

II. TEST SYSTEM CONFIGURATION

International Council on Large Electric Systems (CIGRE) 4-bus multi-terminal benchmark grid [21] is considered for the MMC-HVDC configuration with HTCs [14]-[15] as the test system (shown in Fig. 1). Current limiting reactors are used to limit the rate of rising current in case of a fault contingency in the system. A bipolar line configuration with a frequency-dependent transmission model (FDTL) for OHL and UGC is implemented using Power Systems Computer Aided Design/Electromagnetic Transients including DC (PSCAD/EMTDC). HTC n in Fig. 1 is defined as n^{th} hybrid transmission corridor. The test system considers four 2-segment and one 3-segment HTCs. The parameters for the test system are shown in Table II. Fig. 1 shows the four-terminal MMC-HVDC system with a DC voltage of $\pm 200\text{kV}$. MMC at bus 1 operates with the control objective of DC voltage control and constant reactive power control. MMCs at buses 2-4 operate with constant active and reactive power control [25]-[27]. Considering the typical parameters of

TABLE I: HVDC Fault location methods in literature

Fault Location Methods	Method Type	Applicable to Heterogeneous Line?	Merits	Limitation
[12]	Travelling-wave	No	accurate for high impedance faults (HIFs)	requirement of high sampling rate
[13]	Travelling-wave	No	accurately distinguish between external and internal faults	high sampling rate required, sensitive to parameter variation
[14]	Travelling-wave	Yes	simple and accurate location for heterogeneous lines	application limited to 1000, high sampling rate required
[11]	Travelling-wave	Yes	applicable to HIFs up to 500Ω	high sampling rate required with multiple distributed current sensors
[17]	Time-domain	No	applicable to HIFs with requirement of low sampling rate	sensitive to parameter variation
[18]	Time-domain	No	simple & accurate for HIFs as well with requirement of low sampling rate	sensitive to parameter variation
[Proposed method]	Time-domain	Yes	applicable to heterogeneous lines as well with HIFs using low sampling rate	sensitive to parameter variation

multi-terminal MMC-HVDC systems, the fault interruption needs to be within 3-5ms of fault inception [23]-[26], [29]-[31]. The control algorithm of the MMC-HVDC system typically has a time constant higher than several ms. This means that by the time, the control algorithm identifies the presence of high current due to the fault, the protection algorithm must have already identified and isolated the fault. Therefore, the control system has minimal interaction with the rapid DC fault transients [6], [23]-[26], [29]-[31]. Once the fault is identified, the proposed location method calculates the accurate position of the fault using the data of the initial 2ms of fault inception. The process of fault location is completed before the isolation of the faulty segment. As a result of which, the proposed location method is online in nature and it meets the stringent requirement of protection in DC i.e., fault isolation within 3-5ms of fault inception. Therefore, the other component of the protection system which is fault identification is independent of the process of fault location.

For the initial fault period up to 3ms, Fig. 2 can be used to analyze the faulty section for a much larger system shown in Fig. 1. This period sees fault current contribution from the discharging DC capacitance of MMC. Due to CLRs, the current contribution from other terminals is low. This is a result of the equivalent faulty path having high inductance which makes the rise of current sluggish.

III. PROPOSED FAULT LOCATION ALGORITHM

The proposed fault location method is initially analysed for a smaller 2-segment HTC, then it is explained for a 3-segment and n -segment HTC. A secure and robust fault identification scheme [22]-[23] is a prerequisite for the application of the proposed fault location method. The method is elaborated for pole-to-pole (PTP) fault and briefly discussed for pole-to-ground (PTG) fault. The algorithm uses communication-based double terminal measurements which nearly negates the effect of fault impedance in the analysis. This means that the fault location accuracy remains intact for high resistance faults. The proposed method can also precisely indicate which segment of the faulty HTC is subjected to a fault.

A. Fault Location for a 2-segment HTC

Fig. 2 shows equivalent circuit of faulty section with possibility of a fault at either segment 1 or segment 2. Here $v_1(t)$ and $v_2(t)$ are the terminal DC bus voltage, $i_1(t)$ and $i_2(t)$ are the current through CLRs at each terminal while $v_{dc1}(t)$ and $v_{dc2}(t)$ are voltage after CLRs, L_{m1} and L_{m2} . $v_f(t)$

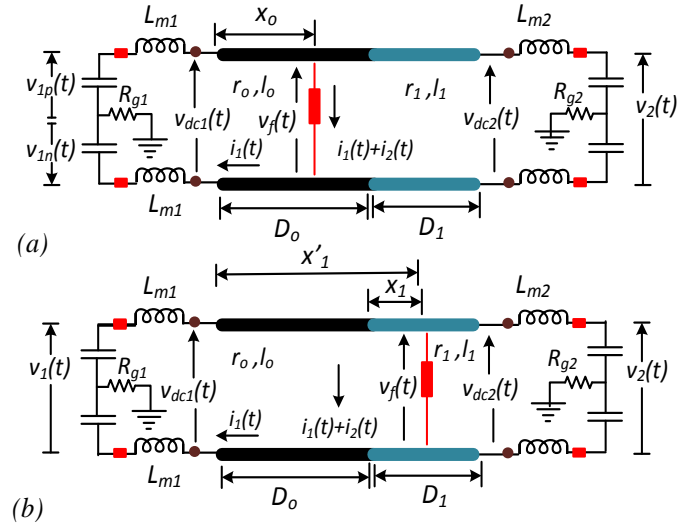


Fig. 2: Equivalent circuit of faulty section for a PTP fault at (a) segment 1, (b) segment 2

is defined as voltage across the fault resistance whereas d is defined as the fault location [30]. Using simplified R-L representation, r_o is the resistance per unit length whereas l_o is the inductance per unit length of segment 1. Similarly, r_1 is the resistance per unit length whereas l_1 is the inductance per unit length of segment 2. The method takes each segment with different values of R-L which gives flexibility to locate fault for any type of n segment non-homogeneous combination of OHL and UGC. As shown in Fig. 2, D_0 & D_1 are the total lengths of segment 1 & 2 respectively. If the fault occurs at segment 1, fault distance is defined as x_0 . Similarly, if the fault occurs at segment 2, fault distance is defined as $x'_1 = x_1 + D_0$. A rolling mean filter [21] is used to filter high frequency components caused by distributed capacitance and white gaussian noise (WGN) in measurements of voltage and current. Further, another rolling mean filter having a moving window of 20 sample steps [16] is used to obtain a conclusive range of calculated location. Applying Kirchhoff's Voltage Law (KVL) to Fig. 2(a) for a PTP fault at segments 1, (1), and (2) are obtained.

$$2r_o x_0 i_1(t) + 2l_o x_0 \frac{di_1(t)}{dt} + v_f(t) = v_{dc1}(t) \quad (1)$$

$$2[r_o(D_0 - x_0) + r_1 D_1] i_2(t) + 2[l_o(D_0 - x_0) + l_1 D_1] \frac{di_2(t)}{dt} + v_f(t) = v_{dc2}(t) \quad (2)$$

The current derivative terms can be obtained using the drop in voltage across the CLR avoiding additional WGN due to differential calculations. These terms are $\frac{di_1(t)}{dt} = \frac{v_1(t) - v_{dc1}(t)}{2L_{m1}} = \frac{u_1(t)}{2L_{m1}}$, $\frac{di_2(t)}{dt} = \frac{v_2(t) - v_{dc2}(t)}{2L_{m2}} = \frac{u_2(t)}{2L_{m2}}$. Subtracting (1) and (2) negates the dependence of fault location on fault resistance. Further, substituting (3) and (4) and rearranging give x_0 as;

$$x_0 = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2(r_o D_o + r_1 D_1) i_2(t) + \frac{l_o D_o + l_1 D_1}{L_{m2}} u_2(t)}{2r_o [i_1(t) + i_2(t)] + \frac{l_o u_1(t)}{L_{m1}} + \frac{l_o u_2(t)}{L_{m2}}} \quad (3)$$

For a positive-pole to ground (P-PTG) fault, the expression for x_0 is given as (4).

$$x_0 = \frac{[v_{dc1p}(t) - v_{dc2p}(t)] + \sum_{j=0}^1 r_j D_j i_{2p}(t) + [R_{g2} i_{2p}(t) - R_{g1} i_{1p}(t)] + \frac{\sum_{j=0}^1 l_j D_j}{L_{m2}} u_{2p}(t)}{r_o [i_{1p}(t) + i_{2p}(t)] + \frac{l_o u_{1p}(t)}{L_{m1}} + \frac{l_o u_{2p}(t)}{L_{m2}}} \quad (4)$$

Here $v_{dc1p}(t)$, $v_{dc2p}(t)$ are the positive pole voltages after CLRs, $i_{1p}(t)$, $i_{2p}(t)$ are the positive pole currents and $u_{1p}(t)$, $u_{2p}(t)$ are positive pole voltages across CLRs. R_{g1} and R_{g2} are the grounding resistances of respective terminal. The rest of the fault location algorithm is elaborated considering a PTP fault. Applying Kirchhoff's Voltage Law (KVL) to Fig. 2(b) for a PTP fault at segment 2, (5) and (6) are obtained.

$$[2r_o D_o + 2r_1 x_1] i_1(t) + [2l_o D_o + 2l_1 x_1] \frac{di_1(t)}{dt} + v_f(t) = v_{dc1}(t) \quad (5)$$

$$2r_1 (D_1 - x_1) i_2(t) + 2l_1 (D_1 - x_1) \frac{di_2(t)}{dt} + v_f(t) = v_{dc2}(t) \quad (6)$$

Subtracting (5) and (6) along with substitution of current derivative terms give x_1 as;

$$x_1 = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2(r_o D_o + r_1 D_1) i_2(t) + \frac{l_o D_o + l_1 D_1}{L_{m2}} u_2(t)}{2r_1 [i_1(t) + i_2(t)] + \frac{l_1 u_1(t)}{L_{m1}} + \frac{l_1 u_2(t)}{L_{m2}}} - \frac{2r_o D_o [i_1(t) + i_2(t)] + \frac{l_o D_o}{L_{m1}} [u_1(t) + u_2(t)]}{2r_1 [i_1(t) + i_2(t)] + \frac{l_1 u_1(t)}{L_{m1}} + \frac{l_1 u_2(t)}{L_{m2}}} \quad (7)$$

where the calculated fault distance from terminal is defined as $x'_1 = x_1 + D_o$. The condition to identify the faulty segment and subsequently, the distance from fault terminal is defined as;

$$d = \begin{cases} x_o, & \forall \mathbf{x} \leq D_o : \mathbf{x} \in \{x'_1, x_0\} \\ x'_1, & \forall \mathbf{x} > D_o : \mathbf{x} \in \{x'_1, x_0\} \end{cases} \quad (8)$$

To precisely know which segment of HTC is subjected to a fault, condition given in (8) is used. If the fault is at segment 1, both x_0 & x'_1 are lesser than total length of segment 1, D_0 . Similarly, if the fault is at segment 2, both x_0 & x'_1 are greater than total length of segment 1, D_0 but lesser than total length of segment 1 & segment 2, $(D_0 + D_1)$. If the transmission line is homogeneous, i.e., $D_1 = 0$, the fault location (d) is defined as (9). This means that the proposed location method is applicable to DC systems with any type of transmission corridor.

$$d = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2r_o D_o i_2(t) + \frac{l_o D_o}{L_{m2}} u_2(t)}{2r_o [i_1(t) + i_2(t)] + \frac{l_o u_1(t)}{L_{m1}} + \frac{l_o u_2(t)}{L_{m2}}} \quad (9)$$

B. Fault location for a 3-segment and n-segment HTC

1) *3-segment HTC*: For a 3-segment HTC, there are 3 possibilities of a fault occurrence, i.e., either at segment 1, segment 2 or segment 3. If the fault occurs at segment 1, applying KVL and rearranging the calculated fault distance, x_0 gives (10).

$$x_0 = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2i_2(t) \sum_{j=0}^2 r_j D_j + \frac{\sum_{j=0}^2 l_j D_j}{L_{m2}} u_2(t)}{2r_o [i_1(t) + i_2(t)] + \frac{l_o u_1(t)}{L_{m1}} + \frac{l_o u_2(t)}{L_{m2}}} \quad (10)$$

If the fault occurs at segment 2, the calculated fault distance at segment 2 gives x_1 , which is defined as (11).

$$x_1 = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2i_2(t) \sum_{j=0}^2 r_j D_j + \frac{\sum_{j=0}^2 l_j D_j}{L_{m2}} u_2(t)}{2r_1 [i_1(t) + i_2(t)] + \frac{l_1 u_1(t)}{L_{m1}} + \frac{l_1 u_2(t)}{L_{m2}}} - \frac{2r_o D_o [i_1(t) + i_2(t)] + \frac{l_o D_o}{L_{m1}} [u_1(t) + u_2(t)]}{2r_1 [i_1(t) + i_2(t)] + \frac{l_1 u_1(t)}{L_{m1}} + \frac{l_1 u_2(t)}{L_{m2}}} \quad (11)$$

where the total calculated fault distance from terminal is defined as $x'_1 = x_1 + D_o$. Similarly, if the fault is at segment 3, the calculated fault distance at segment 3 is defined as (12).

$$x_2 = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2i_2(t) \sum_{j=0}^2 r_j D_j + \frac{\sum_{j=0}^2 l_j D_j}{L_{m2}} u_2(t)}{2r_2 [i_1(t) + i_2(t)] + \frac{l_2 u_1(t)}{L_{m1}} + \frac{l_2 u_2(t)}{L_{m2}}} - \frac{2 \sum_{j=0}^1 r_j D_j [i_1(t) + i_2(t)] + \frac{\sum_{j=0}^1 l_j D_j}{L_{m1}} [u_1(t) + u_2(t)]}{2r_2 [i_1(t) + i_2(t)] + \frac{l_2 u_1(t)}{L_{m1}} + \frac{l_2 u_2(t)}{L_{m2}}} \quad (12)$$

where the calculated fault distance from terminal is defined as $x'_2 = x_2 + D_o + D_1$. A similar condition as (8) can be used to identify the faulty segment and distance from fault terminal.

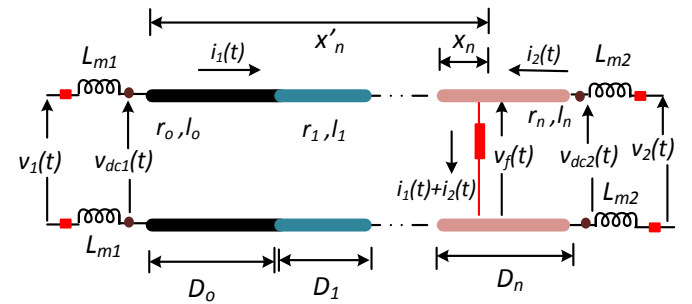


Fig. 3: Equivalent circuit upon inception of a fault contingency in segment n

2) *n-segment HTC*: Generalising the idea of fault location in segmented hybrid transmission corridor, the calculated

fault location at segment n is defined as;

$$x_n = \frac{[v_{dc1}(t) - v_{dc2}(t)] + 2i_2(t) \sum_{j=0}^{n-1} r_j D_j + \frac{\sum_{j=0}^{n-1} l_j D_j}{L_{m2}} u_2(t)}{2r_n[i_1(t) + i_2(t)] + \frac{l_n u_1(t)}{L_{m1}} + \frac{l_n u_2(t)}{L_{m2}}} \quad (13)$$

$$- \frac{2 \sum_{j=0}^{n-2} r_j D_j [i_1(t) + i_2(t)] + \frac{\sum_{j=0}^{n-2} l_j D_j}{L_{m1}} [u_1(t) + u_2(t)]}{2r_n[i_1(t) + i_2(t)] + \frac{l_n u_1(t)}{L_{m1}} + \frac{l_n u_2(t)}{L_{m2}}}$$

where the calculated fault distance from terminal is defined as $x'_n = x_n + \sum_{j=0}^{n-1} D_j$. The condition to identify the faulty segment and subsequently, the distance from fault terminal is defined as;

$$d = \begin{cases} x_0, & \forall \mathbf{x} \leq D_0 : \mathbf{x} \in \{x_0, x'_1, \dots, x'_n\} \\ x'_1, & \forall D_0 < \mathbf{x} \leq D_0 + D_1 : \mathbf{x} \in \{x_0, x'_1, \dots, x'_n\} \\ \vdots \\ x'_n, & \forall \sum_{j=0}^{n-1} D_j < \mathbf{x} \leq \sum_{j=0}^n D_j : \mathbf{x} \in \{x_0, x'_1, \dots, x'_n\} \end{cases} \quad (14)$$

Fig. 4 shows the flowchart for the fault location algorithm. The focus of the work is to locate the fault accurately where DC fault identification is a pre-requisite [22]-[23]. For the purpose of fault identification, initially rolling mean values of voltages are extracted for robustness against WGN in measurements [22]. The line-mode and zero-mode voltages are evaluated using local voltage measurements. As discussed in [16], PTG faults are classified on the basis of polarities of line and zero mode voltages ($u_{12,L}$ & $u_{12,0}$) whereas PTP fault is classified using line mode voltage ($u_{12,L}$). The static mode voltage threshold, $U_{set}=100\text{kV}$ is used as the fault identification threshold for the system. For a PTP fault, $u_{12,L}$ violates U_{set} and $u_{12,0} \approx 0$. For a P-PTG fault, both $u_{12,L}$ & $u_{12,0}$ violate U_{set} . On similar lines, for a N-PTG fault, $u_{12,L}$ violates U_{set} whereas $u_{12,0}$ violates $-U_{set}$. Once line-mode and zero-mode voltage variation identify the type of fault and the faulty segment [22]-[23], [28], terminal current and voltage measurements of the faulty terminals are processed to calculate the fault location. A rolling mean filter with a moving window of 20 samples is used to get a rather conclusive range of fault locations. Using (14), the accurate fault location is calculated as seen in Fig. 4.

IV. RESULTS AND VALIDATION

The test system shown in Fig. 1 is implemented using PSCAD/EMTDC-based electromagnetic transient simulations. A sampling frequency, f_s of 50 kHz is used for conclusive plots. However, a sampling frequency of 5 kHz is rather enough for satisfactory accuracy as shown in section V(D). The detailed models of OHL and UGC are considered in PSCAD/EMTDC. The parameters for the test system are given in Table II. The validation starts with fault identification and eventually the location of the fault in the system shown in Fig. 1. The algorithm is further tested under different fault resistances and fault locations with different sampling frequencies and white Gaussian noise

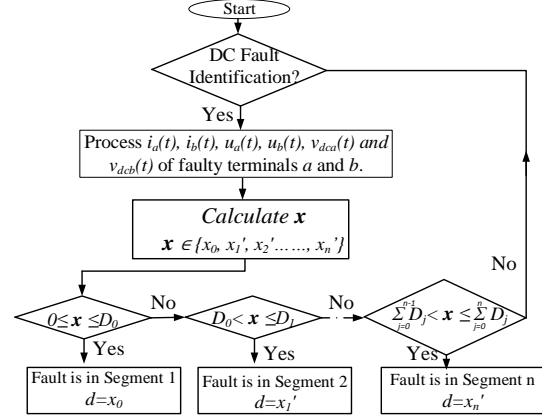


Fig. 4: Layout of the proposed Fault Location Algorithm

(WGN) in measurement. Finally, the effect of parameter variation on the proposed algorithm is discussed which shows the variation of accuracy with the variation of the true value of parameters i.e., unit resistance, unit inductance, and DC link capacitance.

TABLE II: Parameters for test system in PSCAD/EMTDC

DC Voltage (kV)	± 200
Number of Sub-modules (SM)	200
Smoother Reactor (mH)	150
Parameters of MMC ₁ -MMC ₄	$R_{MMC1}/L_{MMC1}/C_{MMC1}$: 192.4mΩ/60mH/300μF $R_{MMC2}/L_{MMC2}/C_{MMC2}$: 243.2mΩ/77.3mH/375μF $R_{MMC3}/L_{MMC3}/C_{MMC3}$: 684.8mΩ/96mH/240μF $R_{MMC4}/L_{MMC4}/C_{MMC4}$: 371.2mΩ/72mH/450μF
Line length (km)	HTC 1: OHL: 150, UGC: 50 HTC 2: OHL ₁ : 50, UGC: 60, OHL ₂ : 90 HTC 3: UGC: 50, OHL: 150 HTC 4: UGC: 100, OHL: 100 HTC 5: OHL: 80, UGC: 120
Parameters of OHL	Resistance [Ω/km]: 0.132 Inductance [mH/km]: 0.201 Capacitance [μF/km]: 0.011
Parameters of UGC	Resistance [Ω/km]: 0.098 Inductance [mH/km]: 0.074 Capacitance [μF/km]: 0.272

A. Validation for fault identification and fault location

Fault identification uses line & zero mode voltages to detect and classify a fault in the system. Once the fault is identified, the algorithm locates the fault using the data for initial 2ms from fault inception. The process of fault location is completed before the isolation of the faulty segment. As a result of which, the proposed location method is online in nature. This section shows the variation of faulty terminal voltages (both positive and negative pole) along with the mode voltage variation for different types of faults i.e., PTP, P-PTG & N-PTG. The validation includes faults at both the segments of *HTC 4* as shown in Fig. 5. Further, the corresponding calculated fault location variation is analyzed in Fig. 6. X_n in Fig. 6 shows the calculated location whereas x_n shows the variation of X_n with a rolling mean filter having a moving window of 20 sample steps [22], [29]. Throughout the analysis, x_n is used to represent the filtered value of the calculated location, X_n . A moving window of greater than 20 sample steps can also be used for practical application of the proposed method. However, the

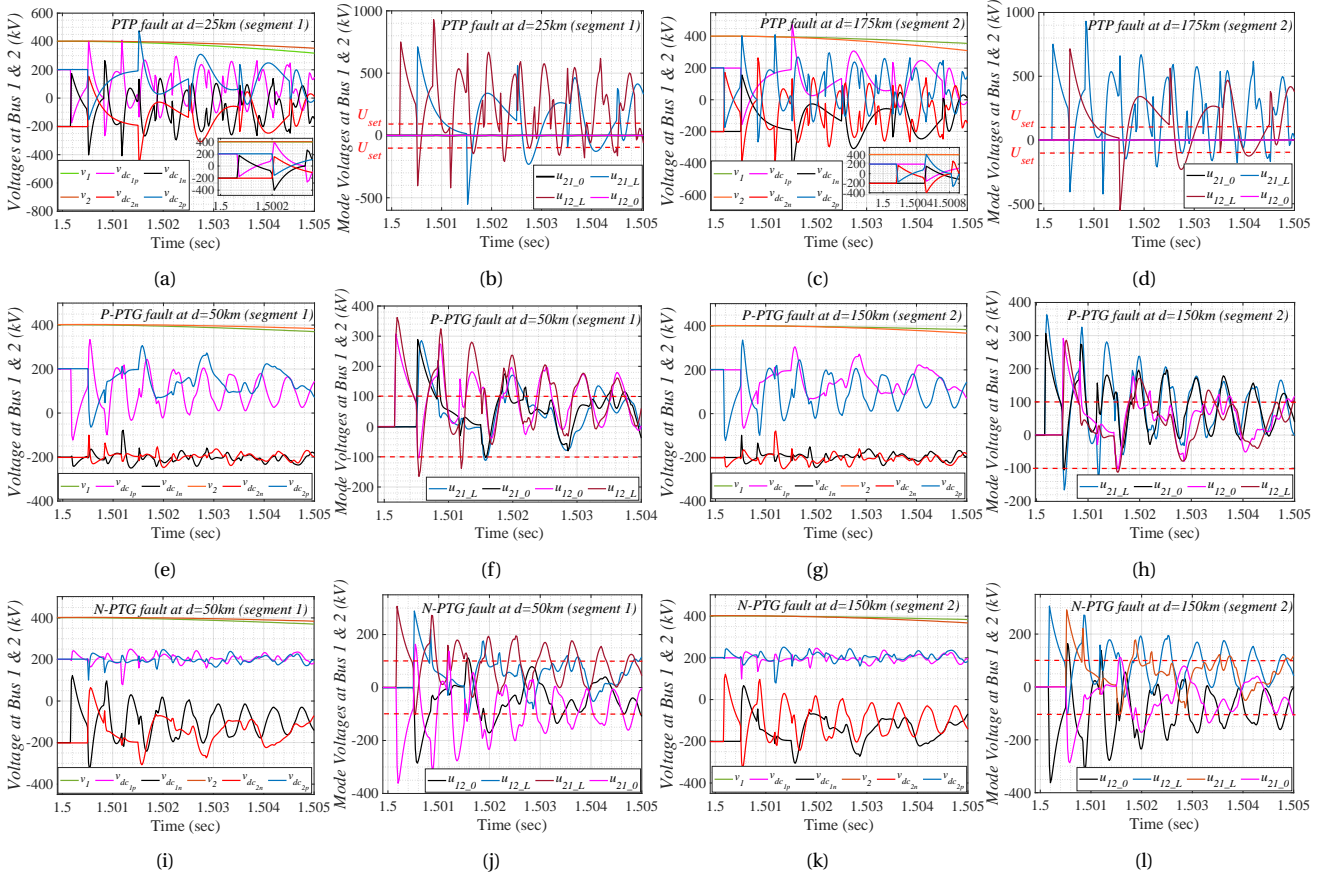


Fig. 5: Variation of bus voltages and mode voltages for (a)-(d) PTP fault, (e)-(h) P-PTG fault, (i)-(l) N-PTG fault at both segments of *HTC 4*.

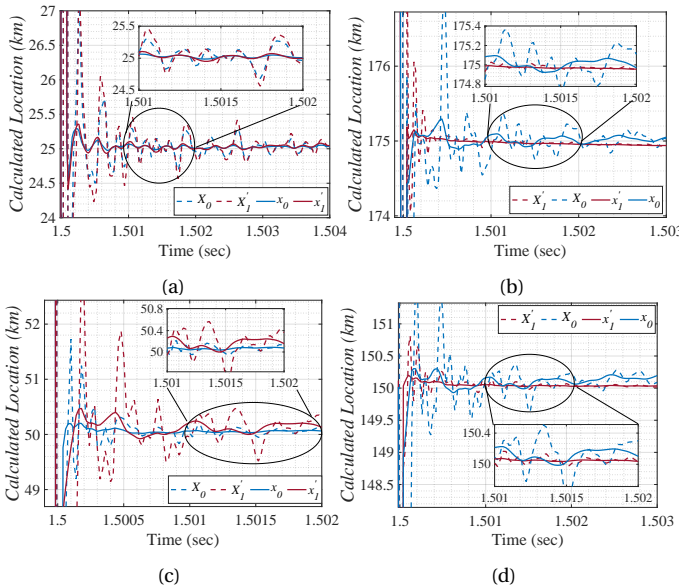


Fig. 6: Calculated fault location for faults at (a) $d=25\text{km}$ (segment 1), (b) $d=175\text{km}$ (segment 2), (c) $d=50\text{km}$ (segment 1), (d) $d=150\text{km}$ (segment 2).

fault location accuracy is improved marginally with further increase in moving window.

1) *PTP fault*: A PTP fault occurs at $d=25\text{km}$ at segment 1 of *HTC 4* with fault resistance, $R_f=5\Omega$. Fig. 5(a) shows both positive (v_{dc1p} & v_{dc2p}) and negative pole (v_{dc1n} & v_{dc2n})

voltages after CLR of bus 1 & 2. Along with it, the terminal voltages of bus 1 & 2 (v_1 & v_2) are also shown. Even though these parameters are not directly used to identify the fault, they are shown to give a better idea of their variation upon a fault inception. As shown in Fig. 5(a), drop in terminal voltages (v_1 & v_2) is sluggish due to DC capacitance whereas the respective pole voltages after CLR (v_{dc1p} , v_{dc2p} , v_{dc1n} & v_{dc2n}) drop almost instantaneously. Fig. 5(b) shows the line (u_{12_L} & u_{21_L}) & zero mode voltages (u_{12_0} & u_{21_0}) for faulty terminal 1 & 2. These mode voltages are used to identify the occurrence of a fault in the system. Line-mode voltages (u_{12_L} & u_{21_L}) violate the fault identification threshold, U_{set} for a PTP fault at segment 1. Similarly, fault identification for a PTP fault at $d=175\text{km}$ at segment 2 of *HTC 4* with fault resistance, $R_f=5\Omega$ is validated using Fig. 5(c)-(d). Once the fault is identified in lesser than 1ms, the fault location algorithm calculates X_0 & X_1 , x_0 & x_1 (rolling mean of X_0 & X_1). For a PTP fault at segment 1 ($d=25\text{km}$), Fig. 6(a) shows $x_0 \approx 25.02\text{km}$ and $x_1 \approx 25.04\text{km}$. The window considered to evaluate the rolling mean is 1-2ms from fault inception. This is a result of fault location transients taking around 1ms to die down. Similarly, for a PTP fault at segment 2 ($d=175\text{km}$), Fig. 6(b) shows $x_0 \approx 175.06\text{km}$ and $x_1 \approx 175\text{km}$ considering the rolling mean between 1.501-1.502ms.

2) *P-PTG faults*: A P-PTG fault occurs at $d=50\text{km}$ at segment 1 of *HTC 4* with fault resistance, $R_f=20\Omega$. Fig. 5(e)

shows large positive pole voltage variation. Fig. 5(f) shows both line & zero mode voltage violating U_{set} . Similarly, fault identification for a P-PTG fault at $d=150\text{km}$ at segment 2 of *HTC 4* with fault resistance, $R_f=20\Omega$ is validated using Fig. 5(g)-(h). Once the fault is identified for a P-PTG fault at segment 1 ($d=50\text{km}$), Fig. 6(c) shows $x_0 \approx 50\text{km}$ and $x'_1 \approx 50.02\text{km}$. Similarly, for a P-PTG fault at segment 2 ($d=150\text{km}$), Fig. 6(d) shows $x_0 \approx 150.03\text{km}$ and $x'_1 \approx 150\text{km}$.

3) *N-PTG fault*: A N-PTG fault occurs at $d=50\text{km}$ at segment 1 of *HTC 4* with fault resistance, $R_f=25\Omega$. Fig. 5(i) shows large negative pole voltage variation. Fig. 5(j) shows both line & zero mode voltage violating U_{set} . However, the zero mode voltage violation is in the negative side. Similarly, fault identification for a N-PTG fault at $d=150\text{km}$ at segment 2 of *HTC 4* with fault resistance, $R_f=25\Omega$ is validated using Fig. 5(k)-(l). Once the fault is identified for a N-PTG fault at segment 1 ($d=50\text{km}$), $x_0 \approx 50\text{km}$ and $x'_1 \approx 50.02\text{km}$ similar to as shown in Fig. 6(c). For a N-PTG fault at segment 2 ($d=150\text{km}$), $x_0 \approx 150.04\text{km}$ and $x'_1 \approx 150\text{km}$ similar to as shown in Fig. 6(d).

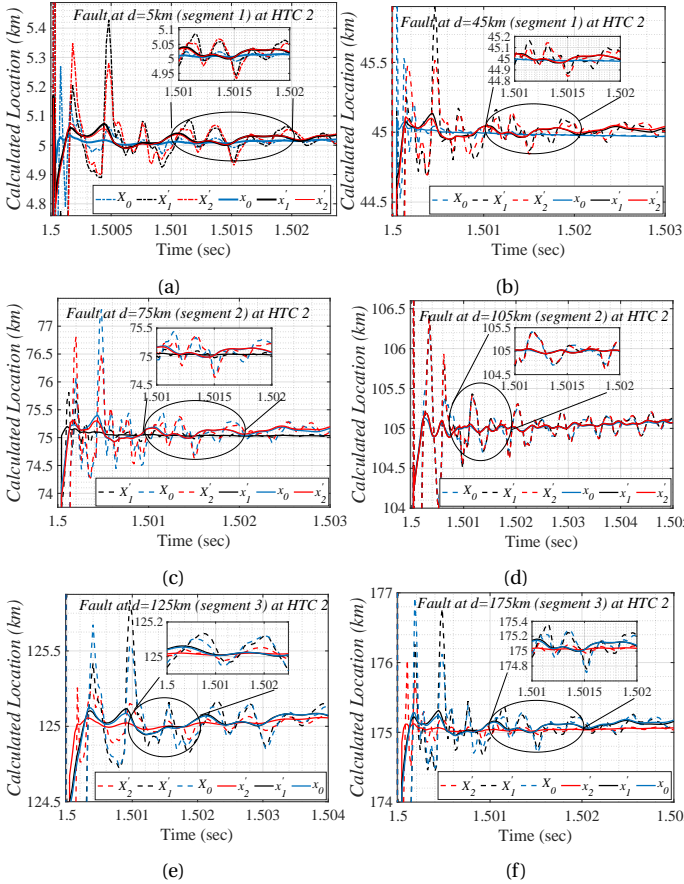


Fig. 7: Calculated fault location for faults at (a) $d=5\text{km}$ (segment 1), (b) $d=45\text{km}$ (segment 1), (c) $d=75\text{km}$ (segment 2), (d) $d=105\text{km}$ (segment 2), (e) $d=125\text{km}$ (segment 3), (f) $d=175\text{km}$ (segment 3).

B. Validation for fault location in a 3 segment HTC

a) *Segment 1*: For the system shown in Fig. 1, faults occur at different segments of *HTC 2*. Fig. 7(a) shows calculated fault location for a fault at $d=5\text{km}$ (segment 1).

The fault location calculated assuming the fault at segment 1, $x_0 \approx 5\text{km}$. Assuming the fault at segment 2, $x'_1 \approx 5.05\text{km}$ whereas assuming the fault at segment 3, $x'_2 \approx 5.04\text{km}$. Similarly, Fig. 7(b) shows calculated fault location for a fault at $d=45\text{km}$ (segment 1). The fault location calculated assuming the fault at segment 1, $x_0 \approx 45\text{km}$. Assuming the fault at segment 2, $x'_1 \approx 45.02\text{km}$ whereas assuming the fault at segment 3, $x'_2 \approx 45.03\text{km}$. For Fig. 7(a)-(b), using (14), $d=x_0$.

b) *Segment 2*: For a fault at $d=75\text{km}$ (segment 2), fault location using segment 1, $x_0 \approx 75.1\text{km}$ as shown in Fig. 7(c). Additionally, fault location using segment 2, $x'_1 \approx 75\text{km}$ whereas fault location using segment 3, $x'_2 \approx 75.06\text{km}$. For a fault at $d=105\text{km}$ (segment 2), fault location using segment 1, $x_0 \approx 105.12\text{km}$ as shown in Fig. 7(d). Additionally, fault location using segment 2, $x'_1 \approx 105\text{km}$ whereas fault location using segment 3, $x'_2 \approx 105.05\text{km}$. For Fig. 7(c)-(d), using (14), $d=x'_1$.

c) *Segment 3*: Similarly, for a fault at $d=125\text{km}$ (segment 3), fault location using segment 1, $x_0 \approx 125.07\text{km}$ as shown in Fig. 7(e). Fault location using segment 2, $x'_1 \approx 125.04\text{km}$ whereas fault location using segment 3, $x'_2 \approx 125\text{km}$. For a fault at $d=175\text{km}$ (segment 3), fault location using segment 1, $x_0 \approx 175.15\text{km}$ as shown in Fig. 7(f). Further, fault location using segment 2, $x'_1 \approx 175.1\text{km}$ whereas fault location using segment 3, $x'_2 \approx 175\text{km}$. For Fig. 7(e)-(f), using (14), $d=x'_2$.

C. Validation for different fault resistances

Table III shows the variation of calculated location with different fault resistances (up to $R_f=200\Omega$). Since the method is double terminal, the accuracy of the algorithm is independent of fault resistance. Table III shows faults at *HTC 1* (2 segments), *HTC 3* (2 segments) and *HTC 2* (3 segments). For high sampling frequencies ($f_s > 20\text{kHz}$), there is no loss of accuracy for high fault resistances. For low sampling frequencies ($f_s < 20\text{kHz}$), there is a marginal loss of accuracy ($< 1\%$) for high fault resistances. This is because as fault resistance increases, the signal damping increases. If the sampling frequency is less, useful samples are passed over. When the fault is at segment 1 of a HTC, the calculated fault location, $d=x_0$. Similarly, when the fault is at segment 2 of a HTC, the calculated fault location, $d=x'_1$ whereas when the fault is at segment 3, the calculated fault location, $d=x'_2$. Table III shows different types of faults i.e., PTP, P-PTG, N-PTG at different locations. The method is fairly accurate for high resistance faults with a maximum recorded error of 30m.

D. Performance under different sampling frequencies

Fault location accuracy with low sampling frequency can help in cost-effective hardware implementation. Table IV shows that a 5kHz sampling frequency gives satisfactory results for the proposed fault location method. However as evident from Table IV, higher f_s means more useful samples which increase the accuracy of the fault location algorithm.

TABLE III: Variation of calculated fault location (km) with fault resistance, R_f (Ω) for different faults (PTG, PTP) at different location

HTC 1 (Total Length: 150+50=200km)						HTC 3 (Total Length: 50+150=200km)						HTC 2 (Total Length: 50+60+90=200km)						
$R_f=0.01\Omega$		$R_f=50\Omega$		$R_f=200\Omega$		$R_f=0.01\Omega$		$R_f=50\Omega$		$R_f=200\Omega$		$R_f=0.01\Omega$		$R_f=50\Omega$		$R_f=200\Omega$		
x_0	x_1'	x_0	x_1'	x_0	x_1'	x_0	x_1'	x_0	x_1'	x_0	x_1'	x_0	x_1'	x_2'	x_0	x_1'	x_2'	
P-PTG																		
10%	20.01	20.04	20.02	20.05	20.02	20.07	20.01	20.04	20.015	20.04	20.02	20.05	20.01	20.02	20.02	20.02	20.05	20.05
25%	50.005	50.02	50.02	50.05	50.03	50.06	50.01	50.02	50.02	50.02	50.03	50.03	50.01	50.01	50.03	50.02	50.02	50.05
50%	100.01	100.02	100.02	100.05	100.02	100.07	100.02	100.01	100.025	100.01	100.04	100.02	100.02	100.01	100.01	100.03	100.01	100.04
75%	150.01	150.01	150.01	150.01	150.02	150.02	150.03	150	150.05	150.008	150.06	150.01	150.04	150.04	150	150.06	150.05	150.01
90%	180.02	180.01	180.04	180.01	180.05	180.02	180.01	180.01	180.04	180.02	180.06	180.02	180.02	180.03	180.01	180.04	180.04	180.02
N-PTG																		
10%	20.01	20.04	20.02	20.05	20.02	20.07	20.01	20.04	20.015	20.04	20.02	20.05	20.01	20.02	20.02	20.02	20.05	20.05
25%	50.005	50.02	50.02	50.05	50.03	50.06	50.01	50.02	50.02	50.02	50.03	50.03	50.01	50.01	50.03	50.02	50.02	50.05
50%	100.01	100.02	100.02	100.05	100.02	100.07	100.02	100.01	100.025	100.01	100.04	100.02	100.02	100.01	100.01	100.03	100.01	100.04
75%	150.01	150.01	150.01	150.01	150.02	150.02	150.03	150	150.05	150.008	150.06	150.01	150.04	150.04	150	150.06	150.05	150.01
90%	180.02	180.01	180.04	180.01	180.05	180.02	180.01	180.01	180.04	180.02	180.06	180.02	180.02	180.03	180.01	180.04	180.04	180.02
PTP																		
10%	20	20.04	20.01	20.04	20.02	20.05	20	20.04	20.01	20.04	20.02	20.05	20.005	20.02	20.02	20.01	20.04	20.04
25%	50.01	50.04	50.02	50.05	50.02	50.06	50.01	50.01	50.02	50.02	50.025	50.03	50	50.01	50.02	50.01	50.01	50.025
50%	100	100.02	100.01	100.025	100.02	100.05	100.02	100	100.025	100.005	100.04	100.01	100.02	100	100.01	100.03	100.01	100.04
75%	150.01	150.01	150.01	150.01	150.02	150.02	150.03	150	150.05	150.008	150.06	150.01	150.04	150.04	150	150.06	150.05	150.01
90%	180.02	180	180.04	180	180.05	180.02	180.01	180	180.04	180.02	180.06	180.02	180.02	180.03	180.005	180.04	180.04	180.02

TABLE IV: Fault location with different sampling frequency

Error (m)=Fault Location-Calculated Fault Location				
Fault Location (%)	50kHz	20kHz	10kHz	5kHz
10%	1.25	3.35	6.56	10.05
30%	1.18	3.32	6.63	11.54
50%	0.75	3.22	6.58	10.45
70%	1.15	3.27	5.77	9.82
90%	1.12	3.32	5.92	9.68

E. Fault location calculation with parameter variation

The sensitivity of fault location algorithm with the variation of system parameters i.e., r (unit resistance) and l (unit inductance) of cables and C (DC link capacitance) of converters is analysed. The true values of parameters are given in Table II. Using different values of r and C up to $\pm 200\%$ do not change the accuracy of the fault location algorithm. This also indicates that joint resistance between segments for OHL and UGC has little effect on the fault accuracy. However, the accuracy of the algorithm varies with variations in unit inductance. Table V shows $\varepsilon(\%) = \varepsilon/d$ with % change in inductance value. The fault location error (%) is measured with respect to the fault distance and not the total length for a fair comparison. For faults closer to the bus terminal, % drop of fault location accuracy is higher. As a result of which, a $\pm 20\%$ change in inductance gives a reduction in accuracy as high as $\pm 18\%$. This suggests that the fault location method is sensitive to the unit inductance of HTC in calculation whereas it is not affected by unit resistance and DC link capacitance value.

TABLE V: Variation of calculated fault location error (%) with variation in unit inductance (%)

Calculated Fault Location error (%)= ε/d						
Fault Distance, d (%)						
Variation in l	10%	30%	50%	70%	90%	100%
$\pm 5\%$	4.2%	1.24%	0.67%	0.45%	0.35%	0.31%
$\pm 10\%$	10.1%	2.35%	1.36%	0.88%	0.71%	0.62%
$\pm 15\%$	13.4%	3.94%	2.05%	1.32%	1.02%	0.97%
$\pm 20\%$	18.1%	4.84%	2.65%	1.81%	1.34%	1.28%

E. Fault detection and location time

The fault identification scheme takes lesser than 1ms which complements the proposed fault location method. For $f_s=20\text{kHz}$, maximum identification time recorded is

0.95ms for faults near the bus terminal. The fault location window is considered between 1 to 2ms from the fault inception time. This makes the maximum detection and location time of the method to be lesser than 2ms.

G. Performance with white gaussian noise in measurement

White Gaussian Noise (WGN) in the measurement induces high-frequency components in the measured data of current and voltage. This can cause inaccuracy in the results for the fault location of the proposed scheme. As a result, a rolling mean filter [22] is used with a moving window of 20 sample steps for voltage and current samples and located fault. WGN is random with the property of zero mean [16]. Using rolling mean samples eliminates the effect of WGN to a great extent without affecting the accuracy of the proposed fault location algorithm. Table VI gives the performance of the algorithm with WGN in measurement.

TABLE VI: Fault location in the presence of WGN

Error (m)				
Fault Location (%)	0dB	10dB	20dB	40dB
10%	1.25	1.42	2.76	4.55
30%	1.18	1.35	2.85	4.72
50%	0.75	1.02	2.52	4.45
70%	1.15	1.34	2.82	3.8
90%	1.12	1.3	2.92	3.85

H. Comparison to relevant literature

The proposed fault location method is compared to the existing time-domain and travelling-wave-based methods [14], [17]-[18]. Table VII shows the variation of calculated fault location error with different fault resistances. The fault occurs either at segment 1 (S1) or segment 2 (S2) and the reported location error is maximum throughout a line or a cable. The method in [14] requires a high sampling rate ($>1\text{MHz}$) for fault location. As a result of which, a sampling rate of 1MHz is considered for the validation of the travelling-wave-based method [14]. Additionally, the time-domain-based methods are validated with a sampling rate of 50kHz . As the fault resistance increases ($>200\Omega$), the accuracy of [14] is reduced in comparison to time-domain-based methods. Additionally, due to the limitation

in mathematical analysis, the existing time-domain-based methods [17], [18] do not apply to HTC. If the fault occurs in segment 1, the accuracy of location is almost the same amongst [17], [18], and the proposed method. However, the methods proposed in [17]-[18] are not applicable if the fault occurs at segment 2.

TABLE VII: Variation of calculated fault location error with different fault resistance (Ω)

Calculated Fault Location error (km) ($\varepsilon = d_{true} - d$)						
Fault Resistance (Ω)	[14]		[18]		[Proposed]	
	S1	S2	S1	S2	S1	S2
0.001 Ω	0.02	0.01	0.02	–	0.02	0.02
10 Ω	0.025	0.02	0.03	–	0.03	0.02
100 Ω	0.14	0.12	0.05	–	0.05	0.036
500 Ω	0.3	0.32	0.08	–	0.08	0.085

V. CONCLUSION

The proposed work gives a novel time-domain-based fault location algorithm for a hybrid transmission corridor (HTC) in a multi-terminal HVDC system. The accuracy of the proposed fault location method is nearly independent of the value of fault resistance. This is a result of double terminal analysis negating the dependence of fault resistance on fault accuracy. The analysis takes each segment with a different value of R-L per km which gives the flexibility to apply the algorithm to different segmented HTCs. The proposed fault location method is sensitive to the use of the true value of unit inductance of HTC, where a 20% deviation in true value can cause a maximum distance error as high as 18% for faults closer to the bus terminal. However, the method is not affected by the change of unit resistance of the cable and the value of DC link capacitance. As a result of which, neglecting joint resistance between HTCs does not affect the accuracy of the location method. The performance of the location method is robust for different sampling frequencies ($f_s=5\text{--}100\text{kHz}$), different WGN in measurement (SNR=0-40dB), for different types of faults at different fault resistances and location. The future scope of work includes analyzing the variation of fault location accuracy with respect to change in unit inductance in detail. Further, proposing a location method that has lesser sensitivity due to its variation.

REFERENCES

- [1] North Sea Wind Power Hub - Consortium Partners. (2019) Power hub as an island. [Online]. <https://northseawindpowerhub.eu/wp-content/uploads/2019/07/NSWPH-Benefit-study-for-potential-locations-of-an-offshore-hub-island-1.pdf>
- [2] C. Liu, F. Zhuo and F. Wang, "Fault Diagnosis of Commutation Failure Using Wavelet Transform and Wavelet Neural Network in HVDC Transmission System," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-8, 2021, Art no. 3525408, doi: 10.1109/TIM.2021.3115574.
- [3] S. Lin, L. Liu, P. Sun, Y. Lei, Y. Teng and X. Li, "Fault Location Algorithm Based on Characteristic Harmonic Measured Impedance for HVdc Grounding Electrode Lines," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 12, pp. 9578-9585, Dec. 2020, doi: 10.1109/TIM.2020.3004682.

- [4] S. Lin, D. Mu, L. Liu, Y. Lei and X. Dong, "A Novel Fault Diagnosis Method for DC Filter in HVDC Systems Based on Parameter Identification," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 9, pp. 5969-5971, Sept. 2020, doi: 10.1109/TIM.2020.3003362.
- [5] A. Orth, A. Hiorns, R. van Houtert, L. Fisher, and C. Fourment, "The European north seas countries' offshore grid initiative-the way forward," in 2012 IEEE Power and Energy Society General Meeting, July 2012, pp. 1-8.
- [6] Y. Li et al., "DC Fault Detection in MTDC Systems Based on Transient High Frequency of Current," in IEEE Transactions on Power Delivery, vol. 34, no. 3, pp. 950-962, June 2019, doi: 10.1109/TPWRD.2018.2882431.
- [7] T. An et al., "A DC grid benchmark model for studies of interconnection of power systems," in CSEE Journal of Power and Energy Systems, vol. 1, no. 4, pp. 101-109, Dec. 2015.
- [8] J. Liu, N. Tai and C. Fan, "Transient-Voltage-Based Protection Scheme for DC Line Faults in the Multiterminal VSC-HVDC System," in IEEE Transactions on Power Delivery, vol. 32, no. 3, pp. 1483-1494.
- [9] W. Xiang, S. Yang, L. Xu, J. Zhang, W. Lin and J. Wen, "A Transient Voltage-Based DC Fault Line Protection Scheme for MMC-Based DC Grid Embedding DC Breakers," in IEEE Transactions on Power Delivery, vol. 34, no. 1, pp. 334-345, Feb. 2019.
- [10] H. Livani and C. Y. Evrenosoglu, "A single-ended fault location method for segmented HVDC transmission line," Elect. Power Syst. Res., vol. 107, pp. 190-198, Feb. 2014.
- [11] D. Tzelepis, G. Fusiek, A. Dyško, P. Niewczas, C. Booth and X. Dong, "Novel Fault Location in MTDC Grids With Non-Homogeneous Transmission Lines Utilizing Distributed Current Sensing Technology," in IEEE Transactions on Smart Grid, vol. 9, no. 5, pp. 5432-5443, Sept. 2018, doi: 10.1109/TSG.2017.2764025.
- [12] M. Farshad and M. Karimi, "A Signal Segmentation Approach to Identify Incident/Reflected Traveling Waves for Fault Location in Half-Bridge MMC-HVdc Grids," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-9, 2022, Art no. 3501209, doi: 10.1109/TIM.2021.3139688.
- [13] D. Mu, S. Lin, H. Zhang and T. Zheng, "A Novel Fault Identification Method for HVDC Converter Station Section Based on Energy Relative Entropy," in IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-10, 2022, Art no. 3507910, doi: 10.1109/TIM.2022.3157374.
- [14] O. M. K. K. Nanayakkara, A. D. Rajapakse and R. Wachal, "Location of DC Line Faults in Conventional HVDC Systems With Segments of Cables and Overhead Lines Using Terminal Measurements," in IEEE Transactions on Power Delivery, vol. 27, no. 1, pp. 279-288, Jan. 2012.
- [15] P. T. Lewis, B. M. Grainger, H. A. Al Hassan, A. Barchowsky and G. F. Reed, "Fault Section Identification Protection Algorithm for Modular Multilevel Converter-Based High Voltage DC With a Hybrid Transmission Corridor," in IEEE Transactions on Industrial Electronics, vol. 63, no. 9, pp. 5652-5662, Sept. 2016.
- [16] J. A. Reyes-Malanche, E. J. Villalobos-Pina, E. Cabal-Yepez, R. Alvarez-Salas and C. Rodriguez-Donate, "Open-Circuit Fault Diagnosis in Power Inverters Through Currents Analysis in Time Domain," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-12, 2021, Art no. 3517512, doi: 10.1109/TIM.2021.3082325.
- [17] C. Li, C. Zhao, J. Xu, Y. Ji, F. Zhang and T. An, "A Pole-to-Pole Short-Circuit Fault Current Calculation Method for DC Grids," in IEEE Transactions on Power Systems, vol. 32, no. 6, pp. 4943-4953.
- [18] J. Xu, Y. Lü, C. Zhao and J. Liang, "A Model-Based DC Fault Location Scheme for Multi-Terminal MMC-HVDC Systems Using a Simplified Transmission Line Representation," in IEEE Transactions on Power Delivery, vol. 35, no. 1, pp. 386-395, Feb. 2020.
- [19] F. Tavano et al. "Diagnostics of compression joints of conductors for HV overhead lines", CIGRE Session 1998, paper 22-206.
- [20] R. Kleveborn et al. "Joints on transmission line conductors: field testing and replacement criteria", CIGRE Session 2002.
- [21] Vrana, T. K., Yang, Y., Jovcic, D., Dennetiere, S., Jardini, J., Saad, H. (2013). The CIGRE B4 DC Grid Test System. Electra, (270), 10-19.
- [22] V. Nougain, S. Mishra, G. S. Misyris and S. Chatzivasileiadis, "Multi-terminal DC Fault Identification for MMC-HVDC Systems Based on Modal Analysis—A Localized Protection Scheme," in IEEE Journal of Emerging and Selected Topics in Power Electronics, vol. 9, no. 6, pp. 6650-6661, Dec. 2021, doi: 10.1109/JESTPE.2021.3068800.
- [23] S. Yang, W. Xiang, R. Li, X. Lu, W. Zuo and J. Wen, "An Improved DC fault Protection Algorithm for MMC HVDC Grids based on Modal Domain Analysis," in IEEE Journal of Emerging and Selected Topics in Power Electronics. doi: 10.1109/JESTPE.2019.2945200.

- [24] X. Li, Q. Song, W. Liu, H. Rao, S. Xu and L. Li, "Protection of Nonpermanent Faults on DC Overhead Lines in MMC-Based HVDC Systems," in IEEE Transactions on Power Delivery, vol. 28, no. 1, pp. 483-490, Jan. 2013.
- [25] M. N. Haleem and A. D. Rajapakse, "Fault Type Discrimination in HVDC Transmission Lines Using Rate of Change of Local Currents," in IEEE Transactions on Power Delivery, doi: 10.1109/TPWRD.2019.2922944
- [26] J. Sneath and A. D. Rajapakse, "Fault Detection and Interruption in an Earthed HVDC Grid Using ROCOV and Hybrid DC Breakers," in IEEE Transactions on Power Delivery, vol. 31, no. 3, pp. 973-981.
- [27] R. Li, L. Xu and L. Yao, "DC Fault Detection and Location in Meshed Multiterminal HVDC Systems Based on DC Reactor Voltage Change Rate," in IEEE Transactions on Power Delivery, vol. 32, no. 3, pp. 1516-1526, June 2017.
- [28] Guobing Song, J. Suonan, Qingqiang Xu, Ping Chen and Yaozhong Ge, "Parallel transmission lines fault location algorithm based on differential component net," in IEEE Transactions on Power Delivery, vol. 20, no. 4, pp. 2396-2406, Oct. 2005.
- [29] S. Jiang, C. Fan, N. Huang, Y. Zhu and M. He, "A Fault Location Method for DC Lines Connected With DAB Terminal in Power Electronic Transformer," in IEEE Transactions on Power Delivery, vol. 34, no. 1, pp. 301-311, Feb. 2019.
- [30] J. Yang, J. E. Fletcher and J. O'Reilly, "Short-Circuit and Ground Fault Analyses and Location in VSC-Based DC Network Cables," in IEEE Transactions on Industrial Electronics, vol. 59, no. 10, pp. 3827-3837, Oct. 2012, doi: 10.1109/TIE.2011.2162712.
- [31] W. Leterme, J. Beerten and D. Van Hertem, "Equivalent circuit for half-bridge MMC dc fault current contribution," 2016 IEEE International Energy Conference (ENERGYCON), 2016, pp. 1-6, doi: 10.1109/ENERGYCON.2016.7513914.
- [32] S. Hara, M. Hirose, M. Hatano, S. Kinoshita, H. Ito and K. Ibuki, "Fault protection of metallic return circuit of Kii channel HVDC system," Seventh International Conference on AC-DC Power Transmission, 2001, pp. 132-137, doi: 10.1049/cp:20010531.
- [33] M. Kamiji, K. Fujii, S. Ogawa and E. Fukuda, "The feature of the Anan-Kihoku direct-current transmission line (overhead line)," IEEE/PES Transmission and Distribution Conference and Exhibition, 2002, pp. 1910-1915 vol.3, doi: 10.1109/TDC.2002.1177750.
- [34] S. Sasaki, "Suppression of Abnormal Overvoltages on a Metallic Return HVDC Overhead Line/Cable Transmission System," in IEEE Transactions on Power Apparatus and Systems, vol. PAS-97, no. 5, pp. 1925-1934, Sept. 1978, doi: 10.1109/TPAS.1978.354689.
- [35] T. Westerweller and J. J. Price, "Basslink HVDC interconnector-system design considerations," The 8th IEE International Conference on AC and DC Power Transmission, 2006, pp. 121-124, doi: 10.1049/cp:20060025.



Vaibhav Nougain received B.Tech degree in electrical engineering in 2017 from Delhi Technological University (Formerly Delhi College of Engineering), New Delhi, India. He is currently working toward the Ph.D. degree at the Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India. He was a Visiting Student

with The Center for Electric Power and Energy (CEE), Technical University of Denmark in 2019. His research interests include protection and control of DC systems.



Sukumar Mishra (Senior Member, IEEE) received his M.Tech and PhD in Electrical Engineering from National Institute of Technology, Rourkela in 1992 and 2000 respectively. Presently, Dr. Mishra is a Professor at the Indian Institute of Technology Delhi and has been its part for the past 17 years. He has won many accolades such as INSA Medal for Young Scientist (2002), INAE Young Engineer Award (2009, 2002), INAE Silver Jubilee Young Engineer Award (2012), Samanta Chandra Shekhar Award (2016), IETE Bimal Bose award (2019), National Mission Innovation Champion award (2019) and NASI-Reliance Industries Platinum Jubilee Award for Application Oriented Innovation in Physical Sciences (2019). He has been granted fellowships from academies like NASI (India), INAE (India), and professional societies like IET (U.K.), IETE (India), IE (India). He has also been recognized as the INAE Industry Academic Distinguished Professor. His research interests lie in the field of Power Systems, Power Quality Studies, Renewable Energy and Smart Grid. Prof. Mishra is currently acting as the ABB Chair professor and has previously delegated as the NTPC, INAE, and Power Grid Chair professor. He has also served as an Independent Director of the Cross Border Power Transmission Company Ltd. and the River Engineering Pvt. Ltd. And has carried out many important industrial consultations with TATA Power, Microtek and others. He is the founder of Silov Solutions Private Limited, a company that specifically deals in products related to renewable energy sources utilizable at household scale as well as at commercial setups. From March 2020, he has also been functioning as the Associate Dean Research and Development of IIT Delhi. Prof. Mishra has been working in close association with IEEE Delhi Section Executive Committee for the past few years and is currently serving as an Editor for the IEEE Transactions on Smart Grid, IEEE Transactions on Sustainable Energy and was an Area Editor for the IET Generation, Transmission Distribution journal.

Design and Implementation of a High-Performance 4-bit Vedic Multiplier Using a Novel 5-bit Adder in 90nm Technology

Hemanshi Chugh
Electronic & Comm. Engineering
Delhi Technological University
Delhi, India
hemanshichugh_2k20phdec508@dtu.ac.in

Sonal Singh
Electronic & Comm. Engineering
Delhi Technological University
Delhi, India
sonalsingh@dtu.ac.in

Abstract— A multiplier is an essential component in high-performance systems including digital signal processors, arithmetic & logical units (ALU), and various other communication systems. The multiplication method essentially needs a lot of hardware resources and more computation time than the other arithmetic operations such as addition and subtraction. In recent years, the development of portable electronics has compelled developers to enhance the existing multiplier designs for improved performance. Vedic mathematics consists of five sutras(formulas) for multiplication; however, the Urdhva Tiryagbhyam sutra is principally utilized as it is a general sutra suited for all types of multiplication providing faster computation with minimum delay time. The proposed architecture in this work uses a novel 5-bit special adder along with conventional full adders and half adders for implementing the 4*4 multiplier using the Urdhva Tiryagbhyam technique of Vedic Mathematic. The multiplier modules are designed in the Cadence Virtuoso System design platform. Subsequently, the conventional and proposed 4-bit multiplier designs are simulated and verified in the Cadence spectre simulation platform in a 90nm CMOS technology library file. The results show an overall improvement for the proposed 4*4 Vedic multiplier with a 52.2% reduction in power, 48.6% reduction in delay, and 75% decrease in the power-delay product (PDP) of the circuit against the conventional CMOS Vedic multiplier.

Keywords— Cadence Virtuoso, Urdhva Tiryagbhyam (UT) Sutra, Vedic Multiplier (VM), 5-bit adder, 90nm CMOS.

I. INTRODUCTION

Multiplication is a significant arithmetic function in digital system designs as multipliers are used extensively in integrated circuits such as FIR filters, communication systems, signal processors etc. There are three major classifications in the multiplier architectures – serial multipliers, parallel multipliers and serial-parallel multipliers (Fig. 1). Several algorithms for implementing fast multipliers have been described in the literature. An efficient multiplier architecture is often designed to offer either of the following design target: low power, high speed and less area in the circuit. However, the combination of these design targets makes them even more suitable for low-power VLSI applications. Multiplication using Vedic mathematics [1] has proved to be advantageous over the other multiplication algorithms due to their architectural regularity, reduced hardware requirements, low power and less delay of the circuit.

Veda is a Sanskrit word signifying ‘Knowledge’. In Vedic mathematics, various principle techniques are put

together to perform multiple calculations in simple and robust way [2]. There are sixteen sutras and thirteen upa

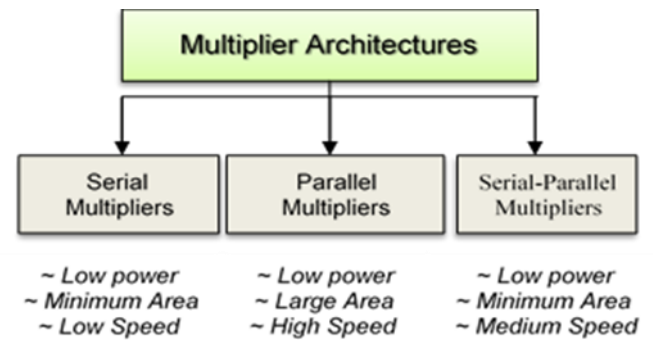


Fig. 1. Classifications of multiplier architectures with their characteristics

sutras (sub sutras) in total that are utilized for various arithmetical operations, geometry, conics, calculus, and algebra. Multiplication can be performed by using five sutras/sub-sutras from these techniques. Fig. 2 illustrates the two approaches of Vedic multiplication. The general approach is suitable for all sets of numbers in any number system whereas a specific approach is suitable only for some particular multiplication sets.

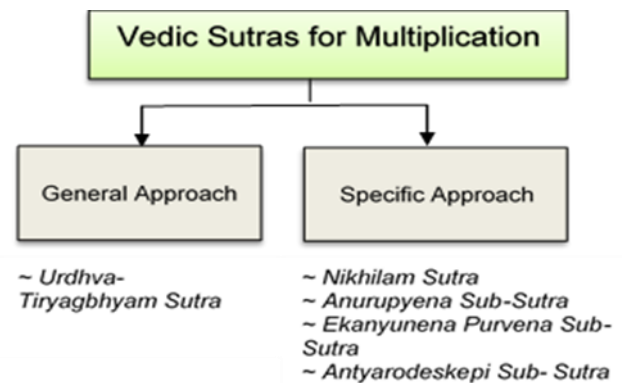


Fig. 2. General and specific approaches of Vedic multiplication sutras

For instance, Nikhilam sutra (all from 9, last from 10) is a Vedic technique for multiplication of numbers closest to any power of 10 [2]. Anurupyena (meaning proportionately) is a sub-sutra to Nikhilam sutra which is used when the operands are not close to the power of 10 [3]. A convenient multiple of the suitable base is taken as the working base to perform the multiplication similar to Nikhilam sutra and the result is then multiplied/divided proportionately. The sutra Ekanyunena Purvena (one less than the one before) is utilized when one

operand is 9 or an array of 9 [4]. Antyayordashake'pi (by completion or non-completion) is a shortcut technique applicable to operands if the addition of last digits of operands is 10 and the remaining digits are same [5]. The UT (vertical & crosswise) technique is the most powerful technique of Vedic multiplication [6] as it can be applied to any set of operands in the number system. Hence, it is widely utilized in digital systems for multiplication.

The paper consists of sections described as follows. Section II describes the UT sutra in detail. Section III provides detailed description of the architecture modules of the 2*2 VM, conventional 4-bit VM and the proposed 4-bit VM. Section IV shows the implementation and simulation results of the architectures in Cadence Virtuoso. Section V gives a performance comparative study of the conventional and proposed 4-bit Vedic multiplier whereas Section VI provides a power comparison analysis of the Vedic multiplier in this work with other related works. Section VII summarizes and concludes the work.

II. URDHVA TIRYAGBHYAM TECHNIQUE

A Urdhva Tiryagbhyam (UT) sutra uses a vertical and crosswise technique for the multiplication of any integral number [7]. This sutra can be used for various bit lengths such as 2, 4, 8, 16.... N for binary multiplication. This paper comprises the design and implementation of the 4-bit multiplier through the UT technique.

The line diagram of 2-bit binary numbers 01 (decimal 1) and 10 (decimal 2) is displayed in Fig. 3 to illustrate the UT algorithm. The LSBs are vertically multiplied in step 1 followed by crosswise multiplication and addition of the two bits in step 2. Further, vertical multiplication of the MSBs is done to get the final product as 010 (decimal 2).

STEP 1	STEP 2	STEP 3	STEP 4
$\begin{array}{r} 01 \\ \\ 10 \\ \hline 0 \end{array}$	$\begin{array}{r} 01 \\ \times \\ 10 \\ \hline 10 \end{array}$	$\begin{array}{r} 01 \\ \\ 10 \\ \hline 010 \end{array}$	$\begin{array}{r} 01 \\ \\ 10 \\ \hline 0010 \end{array}$

Fig. 3. Illustrative depiction of 2-bit UT based Vedic multiplication(adapted from [7])

Similarly, the fundamentals of this sutra can be applied for higher-order binary multiplication [8]. Fig. 4 clearly illustrates a step-wise multiplication of a 4-bit VM.

STEP 1	STEP 2	STEP 3	STEP 4
$\begin{array}{r} 1010 \\ \\ 1011 \\ \hline 0 \end{array}$	$\begin{array}{r} 1010 \\ \times \\ 1011 \\ \hline 10 \end{array}$	$\begin{array}{r} 1010 \\ \times \\ 1011 \\ \hline 110 \end{array}$	$\begin{array}{r} 1010 \\ \times \\ 1011 \\ \hline 1110 \end{array}$
STEP 5	STEP 6	STEP 7	STEP 8
$\begin{array}{r} 1010 \\ \times \\ 1011 \\ \hline 01110 \end{array}$	$\begin{array}{r} 1010 \\ \times \\ 1011 \\ \hline 101110 \end{array}$	$\begin{array}{r} 1010 \\ \\ 1011 \\ \hline 1101110 \end{array}$	$\begin{array}{r} 1010 \\ \\ 1011 \\ \hline 01101110 \end{array}$

Fig. 4. Illustrative depiction of 4-bit Vedic multiplication with UT sutra

The 4-bit binary operands are taken for illustration as 1010(decimal 10) and 1011(decimal 11). The partial products are produced using the vertical and crosswise process in 7 steps and the final product is displayed in Step 8: 01101110(decimal 110).

III. ARCHITECTURE MODULES

The below sections display the conventional CMOS architecture modules of 2-bit and 4-bit VM and the proposed novel architecture for 4-bit binary multiplication using a special 5-bit adder with basic gates, full adder and half adder utilizing the UT algorithm. The sutra is well designed for parallel processing as the partial product generation and additions are carried out simultaneously thereby reducing the overall delay of the circuit and hence making it very efficient for binary multiplications.

A. 2-bit Vedic Multiplier using UT sutra

The multiplication technique for two operands A and B such that $A = A_1A_0$ and $B = B_1B_0$, where A_0 and B_0 signify the LSBs and A_1 and B_1 are the MSBs. The LSB of the product (Q_0) is generated by the vertical multiplication of A_0 and B_0 (LSBs) followed by crosswise multiplication of the bits of A and B (A_1B_0 , A_0B_1). The next bit of the product (Q_1) is generated by the addition of these bits. Thereafter, vertical multiplication of the MSBs (A_1B_1) is done. The third and fourth bit of the final product (Q_2Q_3) are sum and carry of the addition of the previous carry to this partial product respectively(A_1B_1) [9].

The 2-bit architecture of the VM comprises of 4 AND gates with 2 Half Adders; depicted in Fig. 5.

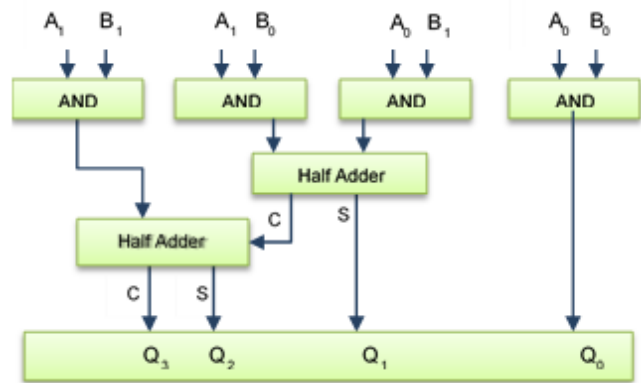


Fig. 5. Conventional CMOS 2-bit Vedic multiplier

B. 4-bit Vedic Multiplier using UT sutra

The implementation of the conventional CMOS 4*4 multiplier module is illustrated in Fig. 5. Let A and B be binary numbers of four bits each such that A_0 and B_0 are the LSBs and A_3 and B_3 are the MSBs. The final 8-bit product is defined as $Q_7Q_6Q_5Q_4Q_3Q_2Q_1Q_0$. To generate the LSB of the product(Q_0), multiply A_0 and B_0 . Now, the operands A and B, split as A_3A_2 ; A_1A_0 , B_3B_2 ; B_1B_0 , are multiplied crosswise and vertically (similar to 2*2 Vedic module) to yield partial products which are added using 4-bit ripple carry adders (RCA), resulting in final product bits, namely, Q_7 to Q_1 .

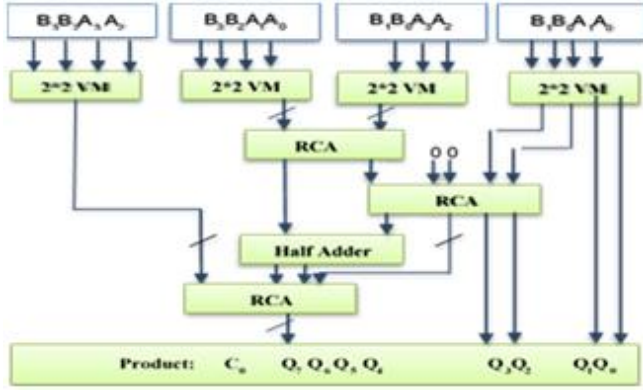


Fig. 6. Conventional CMOS 4-bit Vedic multiplier

C. Proposed architecture of 4-bit Vedic Multiplier

In an effort to enhance the overall working of the multiplier, certain modifications in the initial building blocks of architectural design are proposed. Fig. 7 illustrates the use of sixteen AND gates, six full adders, three half adders, with a special 5-bit adder for making an efficient multiplier architecture. A special 5-bit adder is designed such that it takes 5 bits of binary data and generates a 3-bit output. The internal architecture of a 5-bit adder (Fig. 8) comprises of two full adders to generate the sum and a half adder to generate the carry bits C_0 and C_1 .

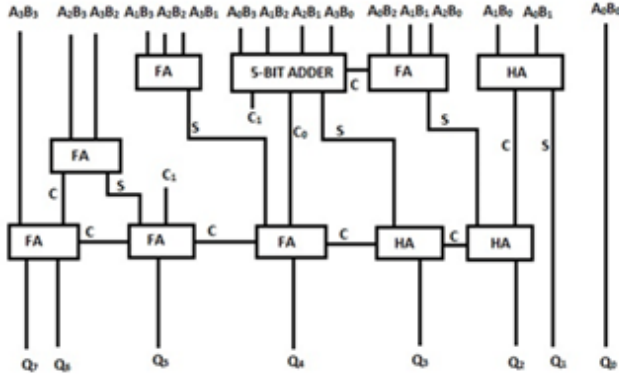


Fig. 7. Architecture of the proposed 4x4 VM module using a 5-bit adder

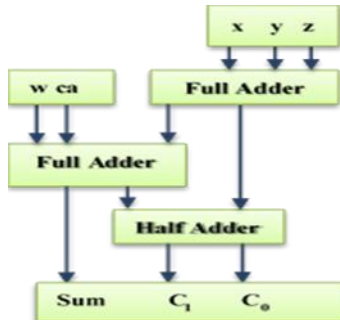


Fig. 8. Internal architecture of the 5-bit special adder

The truth table comprising of all the possible input-output combinations for the novel 5-bit special adder are tabulated. Table I consists of 32 different combinations of inputs (x, y, z, w, ca) represented as a three-bit output (C_1, C_0, Sum).

TABLE I. TRUTH TABLE OF NOVEL 5-BIT SPECIAL ADDER

Sno.	Inputs					Outputs		
	x	y	z	w	Ca	C_1	C_0	Sum
1	0	0	0	0	0	0	0	0
2	0	0	0	0	1	0	0	1
3	0	0	0	1	0	0	0	1
4	0	0	0	1	1	0	1	0
5	0	0	1	0	0	0	0	1
6	0	0	1	0	1	0	1	0
7	0	0	1	1	0	0	1	0
8	0	0	1	1	1	0	1	1
9	0	1	0	0	0	0	0	1
10	0	1	0	0	1	0	1	0
11	0	1	0	1	0	0	1	0
12	0	1	0	1	1	0	1	1
13	0	1	1	0	0	0	1	0
14	0	1	1	0	1	0	1	1
15	0	1	1	1	0	0	1	1
16	0	1	1	1	1	1	0	0
17	1	0	0	0	0	0	0	1
18	1	0	0	0	1	0	1	0
19	1	0	0	1	0	0	1	0
20	1	0	0	1	1	0	1	1
21	1	0	1	0	0	0	1	0
22	1	0	1	0	1	0	1	1
23	1	0	1	1	0	0	1	1
24	1	0	1	1	1	1	0	0
25	1	1	0	0	0	0	1	0
26	1	1	0	0	1	0	1	1
27	1	1	0	1	0	0	1	1
28	1	1	0	1	1	1	0	0
29	1	1	1	0	0	0	1	1
30	1	1	1	0	1	1	0	0
31	1	1	1	1	0	1	0	0
32	1	1	1	1	1	1	0	1

IV. SCHEMATIC DESIGNS & SIMULATION RESULTS

This section presents the schematic designs and simulation results of the architecture modules described in the above sections.

A. Schematic Designs

The schematics designs presented here are implemented in Cadence Virtuoso Design platform v6.1.5.

The CMOS implementation of the schematic design of the novel 5-bit special adder is presented in Fig. 9 which takes 5 bits of data as input and gives a 3-bit output. Further,

Fig. 10 represents the CMOS implementation of the conventional 2-bit VM comprising of AND gates and Half Adder circuit to generate a 4-bit product.

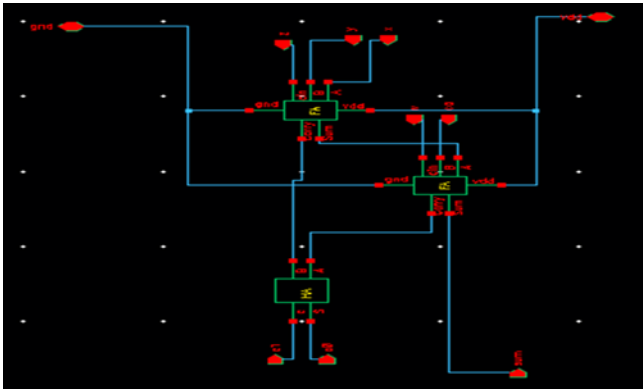


Fig. 9. Schematic of the Novel 5-bit Special Adder

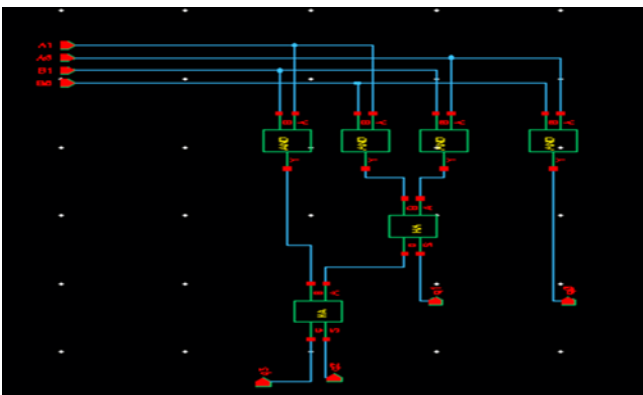


Fig. 10. Schematic of CMOS 2-bit VM

Fig. 11 presents the implementation of the schematic design of the conventional CMOS VM of 4 bits (Fig. 6). The design is made using four blocks of 2-bit Vedic multiplier (Fig.5) together with three four-bit RCAs.

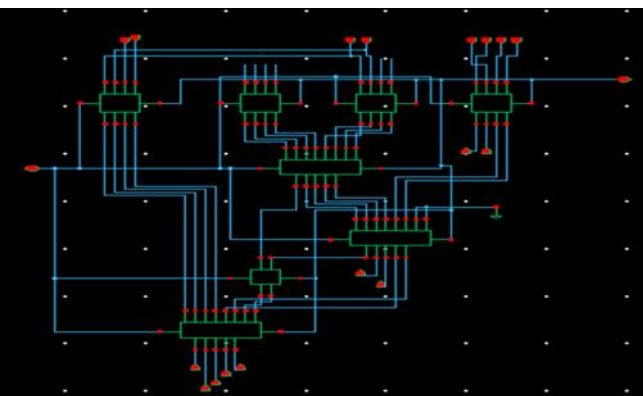


Fig. 11. Schematic of the conventional CMOS 4-bit VM

Fig. 12 presents the schematic design for the novel 4-bit VM. The circuit comprises of 16 AND gates, 6 Full Adders, 3 Half adders and one 5-bit special adder combined together (Fig. 5) to propose a comparatively low power and faster VM.

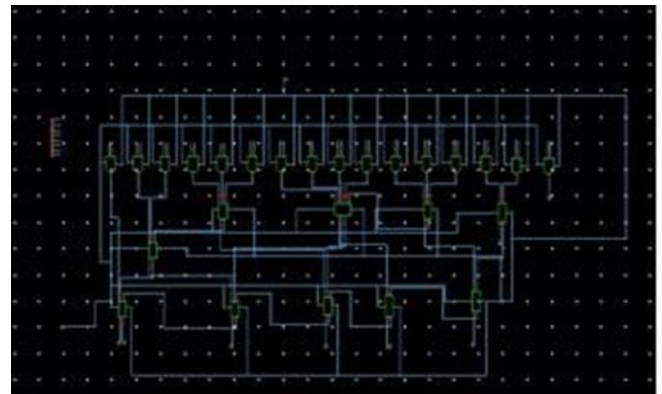


Fig. 12. Schematic of the Novel 4*4 Vedic Multiplier using a 5-bit adder

B. Simulation Waveforms in ADEL

The schematics are verified and simulated in the Analog Design Environment (ADEL) using the spectre simulation platform using 90nm technology at 2 V power supply. The input-output waveforms of the novel 5-bit special adder are shown in Fig. 13. x, y, z, w and ca are the inputs to the adder and C1, C0 and Sum are the outputs of the adder.

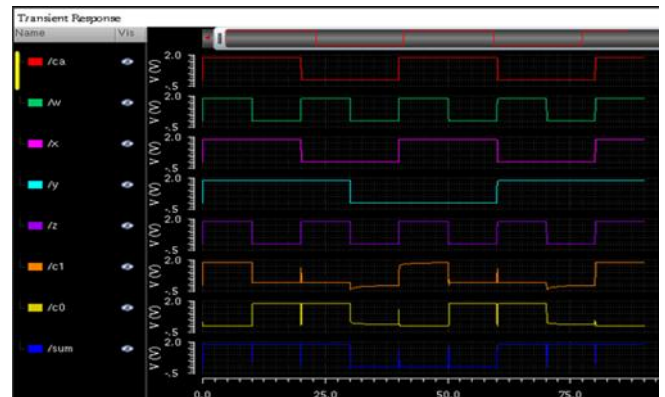


Fig. 13. Input output waveform of 5-bit novel adder

Further, Fig. 14 shows the simulation waveform of the CMOS implementation of the 2-bit VM. The input output waveforms of the conventional 4-bit VM are as shown in Fig. 15 where A3 A2 A1 A0 and B3 B2 B1 B0 are fed as inputs to the multiplier and Q7Q6Q5Q4Q3Q2Q1Q0 display the output of the multiplier and C represents the output carry.

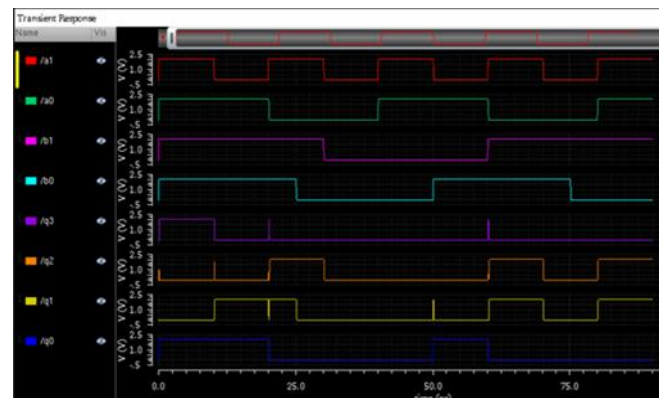


Fig. 14. Input Output Waveform of Conventional 2-bit Vedic multiplier

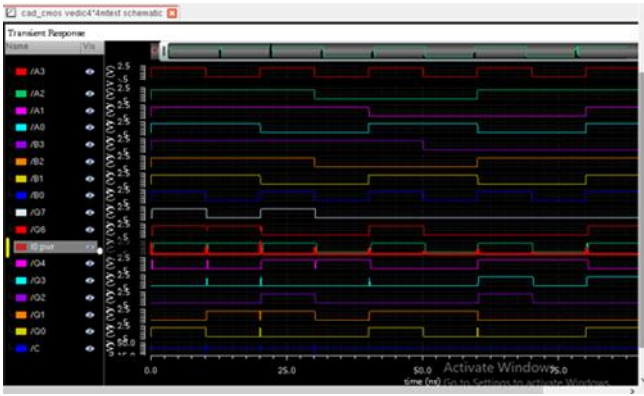


Fig. 15. Input Output Waveform of Conventional 4-bit Vedic multiplier

The simulated waveforms of the novel 4-bit Vedic multiplier are as shown in Fig. 16 where $A_3A_2A_1A_0$ and $B_3B_2B_1B_0$ are inputs to the multiplier and $Q_7Q_6Q_5Q_4Q_3Q_2Q_1Q_0$ display the output(8-bit) of the multiplier. The proposed multiplier is verified for all possible input combinations. Some of the multiplication results (Fig. 16) are also tabulated in Table II for the different input combinations to the multiplier.

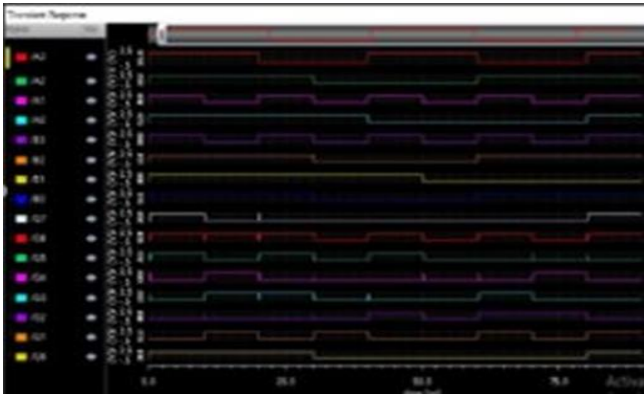


Fig. 16. Input Output Waveform of Proposed 4-bit Vedic multiplier

TABLE II. MULTIPLICATION RESULTS OF THE 4*4 VM

Sno.	A decimal	A binary	B decimal	B binary	Q decimal	Q binary
1	15	1111	15	1111	225	11100001
2	13	1101	07	0111	91	01011011
3	07	0111	15	1111	105	01101001
4	01	0001	02	0010	02	00000010
5	10	1010	10	1010	100	01100100
6	08	1000	00	0000	00	00000000
7	06	0110	13	1101	78	01001110
8	04	0100	05	0101	20	00010100

V. COMPARATIVE ANALYSIS OF THE CONVENTIONAL AND PROPOSED VEDIC MULTIPLIER

The following observations are tabulated in Table III for the conventional and proposed 4-bit multiplier when the conventional and proposed multiplier schematics are simulated for a transition time of 90 ns. The number of

accepted transient steps, peak resident memory used and processor time required is observed from the netlist after simulation and the comparative analysis of the conventional 4*4 Vedic multiplier with the proposed design is shown via the column chart in Fig. 17. Table IV analyses the transistor count, average power, overall propagation delay and power delay product (PDP) for the conventional and proposed 4*4 Vedic multipliers. The compared outcome is also displayed through a column chart (Fig. 18). PDP (also known as switching energy) is associated with the energy efficiency of the system. It is noted that there is a 52.2% reduction in power, 48.6% reduction in delay, and 75% decrease in the PDP of the circuit for the proposed 4*4 Vedic multiplier in contrast to the conventional 4*4 Vedic multiplier.

TABLE III. COMPARATIVE ANALYSIS OF THE ADE NETLIST

4*4 Vedic Multiplier	Number of accepted trans steps	Peak resident memory used (MB)	Processor time (sec)
Conventional	1568	57.5	24.32
Proposed	1329	50.5	13.95

TABLE IV. COMPARATIVE ANALYSIS OF 4*4 VM

4*4 VM	Power (uW)	Delay(ps)	Transistor Count	PDP
Conventional	143.6	341.7	762	49.06
Proposed	68.5	175.4	504	12.01

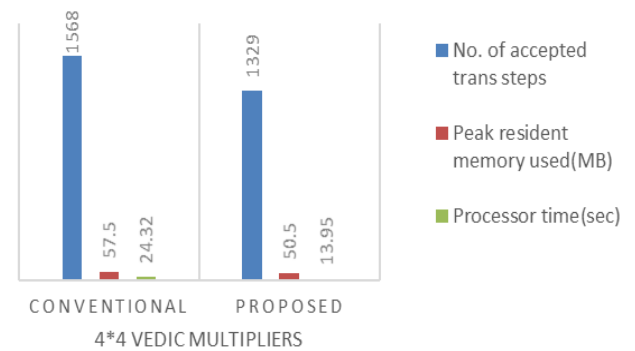


Fig. 17. Graphical representation of netlist analysis for 4-bit VM

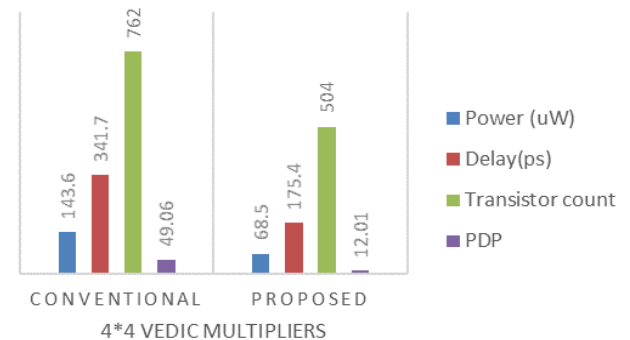


Fig. 18. Graphical representation of performance analysis for 4-bit VM

VI. RELATED WORKS

Table V summarizes the obtained results of power dissipation of the 4-bit Vedic multiplier against the conventional and modified versions of 4-bit Vedic, Array, Wallace tree and Hybrid Dadda multiplier implementations on Cadence Virtuoso in 90 nm technology carried out in the literature. It can be clearly observed that the Vedic mathematics approach in our work is much advantageous as the results of the proposed VM are promising in comparison to other multipliers implemented on the same technology. Fig. 18 depicts the graphical representation of the power analysis of the proposed 4-bit VM design with other 4-bit multiplier designs. It is observed that the proposed design is beneficial concerning the power dissipation in the system.

TABLE V: POWER ANALYSIS OF DIFFERENT 4*4 MULTIPLIERS

Sno.	Multipliers	Power (uW)
1.	Conventional Array [10]	389
2.	Modified Array [10]	170
3.	Conventional Wallace tree [10]	2283
4.	Modified Wallace tree [10]	192
5.	Hybrid Dadda [11]	184.3
6.	Conventional Vedic [12]	361.2
7.	Modified Vedic [12]	290.2
8.	Conventional Vedic (this work)	143.6
9.	Proposed Vedic (this work)	68.5

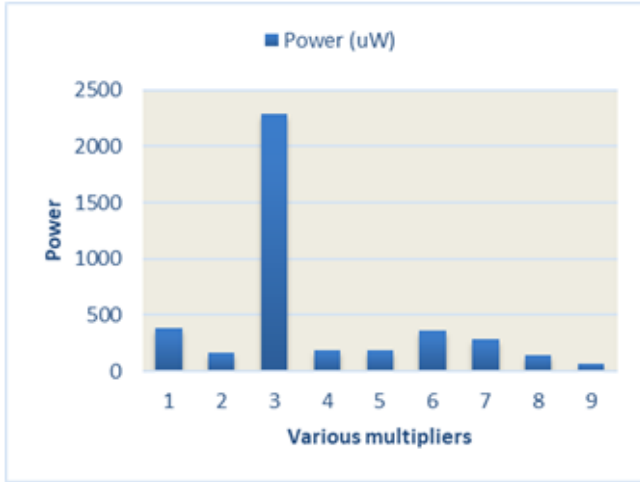


Fig. 19. Graphical Representation of Power Analysis of Different 4*4 Multipliers

VII. CONCLUSION

This work puts forward a proficient 4*4 multiplier utilizing UT Vedic sutra which uses a novel 5-bit special adder in its schematic design. All the schematics have been designed in Cadence Virtuoso v6.1.5 and simulated in the ADEL environment using the spectre simulation platform in a 90nm CMOS technology library file. A comparison of

delay, average power, transistor count, and PDP is carried out in this work. The design implementation and results show 52.2% of the power reduction and 48.6% reduction in delay in the novel design in contrast to the conventional 4-bit Vedic multiplier. The power delay product also decreases to 75% in the novel design. Notably, the time consumption of the processor significantly reduces from 24.32 seconds to 13.95 seconds and the complexity in computation is also reduced as it requires a smaller number of steps in comparison to the conventional VM. Further, it is observed that the VM outperforms other implementations of conventional and modified multipliers namely Array multiplier, Wallace tree multiplier, Hybrid Dadda multiplier, Vedic multiplier in similar technology. Thus, the suggested 4*4 VM will prove beneficial in any ALU, VLSI application, DSP structures or processor [7].

REFERENCES

- [1] Tirtha, S. B. Krishna, and V. S. Agrawala, *Vedic mathematics*. 1992.
- [2] A. Patil, Y. V. Chavan, and S. Wadar, "Performance analysis of multiplication operation based on vedic mathematics," in *2016 International Conference on Control, Computing, Communication and Materials (ICCCCM)*, 2016, no. October, pp. 1–3. doi: 10.1109/ICCCCM.2016.7918246.
- [3] S. K. Parameswaran and G. Chinnusamy, "Design and investigation of low-complexity Anurupya Vedic multiplier for machine learning applications," *Sadhana - Acad. Proc. Eng. Sci.*, vol. 45, no. 1, pp. 1–4, 2020, doi: 10.1007/s12046-020-01500-4.
- [4] M. A. Sayyad and D. N. Kyatanavar, "A Novel Method of Multiplication with Ekanyunena Purvena," in *Advances in VLSI and Embedded Systems*, 2021, pp. 89–96.
- [5] A. Mandloi, "Comparative Analysis of Techniques of Vedic Mathematics," *Int. J. Math. Trends Technol.*, vol. 68, no. 3, pp. 30–32, 2022.
- [6] V. Iyer, V. Sudha, and H. L. Joodith, "Generalised Algorithm for Multiplying Binary Numbers Via Vedic Mathematics," in *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, 2021, pp. 1–5.
- [7] L. S. Jie and S. H. Ruslan, "A 2x2 bit Vedic multiplier with different adders in 90nm CMOS technology," in *AIP Conference Proceedings*, 2017, vol. 1883, no. September 2017. doi: 10.1063/1.5002035.
- [8] K. V. Ramya and S. K. S. Manvi, "Design of a 4-Bit Vedic Multiplier using transistors," *Int. J. Electr. Electron. Eng. Res.*, vol. 4, no. 2, pp. 83–90, 2014.
- [9] C. R. Patel, V. Urankar, B. V. B. Vivek, and V. K. Bharadwaj, "Vedic Multiplier in 45nm Technology," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2020, pp. 21–26. doi: 10.1109/ICCMC48092.2020.ICCMC-0004.
- [10] M. Bansal, V. Bharti, and V. Chander, "Comparison between Conventional Fast Multipliers and Improved Fast Multipliers using PTL Logic," in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1126, no. 1, p. 012041. doi: 10.1088/1757-899x/1126/1/012041.
- [11] M. H. Riaz, S. A. Ahmed, Q. Javaid, and T. Kamal, "Low power 4x4 bit multiplier design using dadda algorithm and optimized full adder," in *15th International Bhurban Conference on Applied Sciences and Technology, IBCAST 2018*, 2018, vol. January, pp. 392–396. doi: 10.1109/IBCAST.2018.8312254.
- [12] C. Rajendra Patel, V. Bettadapura Adishesha, V. Urankar, and K. Vaidyanathan Bharadwaj, "Inverted Gate Vedic Multiplier in 90nm CMOS Technology," *Am. J. Electr. Comput. Eng.*, vol. 4, no. 1, p. 10, 2020, doi: 10.11648/j.ajece.20200401.12.

Development of Novel Model for the Assessment of Dust Accumulation on Solar PV Modules

Astitva Kumar , Member, IEEE, Muhannad Alaraj , Mohammad Rizwan, Senior Member, IEEE, Ibrahim Alsaïdan , and Majid Jamil, Senior Member, IEEE

Abstract—The power generation capability of the photovoltaic (PV) system crucially depends on the soiling of the modules, especially for a hot arid desert in a semi-tropical region. This article focuses to develop a robust and accurate model to estimate PV power loss due to soiling. The proposed article aims to analyze the performance of the newly installed PV system. This involves assessment of reduction in PV power generation due to various derating factors, especially dirt derating factor (K_d). In addition, the impact of dirt, dust, and soiling on the system is studied and a robust empirical model has been developed. Moreover, a simpler model considering module temperature (T_{mod}) has been developed to estimate the power loss due to dirt and dust accumulated on the modules. This article further compares the performance of the two developed models using performance indices. The performance indices, such as root mean squared error and mean absolute percentage error, are admissible with 0.57% and 4.71%, respectively.

Index Terms—Derating factors, power loss, regression analysis, soiling, solar photovoltaic (PV) power.

I. INTRODUCTION

THE global penetration of photovoltaic (PV) is increasing at a fast pace. While utility-scale PV systems have dominated the market, distributed PV systems are becoming more significant in many nations; therefore, the PV landscape is expected to evolve. As a result, PV systems in buildings provide a cost-effective and long-term way to generate renewable energy on-site. Roof surfaces are progressively becoming PV roofs and boosting building energy self-sufficiency, which aids in the reduction of greenhouse gas emissions in cities, particularly in urban regions. Building integrated photovoltaics (BIPV) is the use of solar modules and systems to replace building components, whereas building applied photovoltaics (BAPV) is the use

of PV modules to retrofit existing structures. BIPV is the optimal choice for new buildings and retrofits, where PV modules have a useful role in exterior walls or roofs, while BAPV can be a viable option for existing structures that do not require an entire makeover.

Because the effective incoming irradiance is reduced when dirt, dust, pollen, and other environmental pollutants accumulate on the PV modules' glazing surfaces, the energy conversion efficiency suffers. This effect, known as soiling, is a complex physical-chemical phenomenon impacted by a variety of elements acting on various size and time scales, and multiple models to estimate soiling losses (SLs) have been published. With the addition of other atmospheric particulate matter collecting in the same spot, these particles are deposited and their density will continue to rise over time. As a result, a coating of dust or "soiling" will develop. The absorption of irradiation for energy conversion is reduced in the presence of soiling. The PV system's overall efficiency and energy output will suffer as a result. This issue has to be investigated both before and after the PV system is built since it impacts the return on investment for the owner, investor, or user. As a result, this article is being carried out to solve the serious issue of soiling. The attributes of dust, PV tilt angle, surface material, and geographical climate factors all influence the amount of dust deposited on a PV system. If models are not cleaned, PV power output can be reduced anywhere from 5% to over 50%, depending on the area. The stochasticity of sources in renewable integrated systems creates uncertainty, which is primarily discussed in the functioning of economic systems in [1] and [2]. However, SL is frequently overlooked in field operations.

There are currently tools for estimating SLs and degradation rates directly using solar PV power datasets and weather parameters. However, in systems with a lot of soiling, degradation rates are hard to retrieve due to the combined effect of these processes on PV performance [3]. The improvement in soiling estimation assists in developing a proper maintenance schedule for PV panels. Various literature can be found to predict the soiling on PV system and its impact on power output of PV panels [4]. This power loss in the case of PV modules is also known as SL. The tools for predicting degradation factors and SL using PV production data are available in the literature [5], [6].

In the literature, most of the case studies or test scenarios related to soiling are performed for equatorial, full tropical, and temperate climate regions. Here, the seasonal changes are predictable and seasonal variability impact on soiling can be

Manuscript received 7 July 2022; revised 1 October 2022; accepted 4 November 2022. The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education and, Saudi Arabia for funding this research work through the project number (QU-IF-1-3-3). The authors also thank to the technical support provided by Qassim University. (Muhannad Alaraj and Astitva Kumar are co-first authors.) (Corresponding author: Astitva Kumar.)

Astitva Kumar and Mohammad Rizwan are with the Department of Electrical Engineering, Delhi Technological University, Delhi 110042, India (e-mail: astitvakumar_phd2k16@dtu.ac.in; rizwan@dce.ac.in).

Muhannad Alaraj and Ibrahim Alsaïdan are with the Department of Electrical Engineering, College of Engineering, Qassim University, Buraydah 52571, Saudi Arabia (e-mail: muhannad@qu.edu.sa; alsaidan@qu.edu.sa).

Majid Jamil is with the Department of Electrical Engineering, Jamia Millia Islamia, Delhi 110025, India (e-mail: majidjamil@hotmail.com).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/JPHOTOV.2022.3220923>.

Digital Object Identifier 10.1109/JPHOTOV.2022.3220923

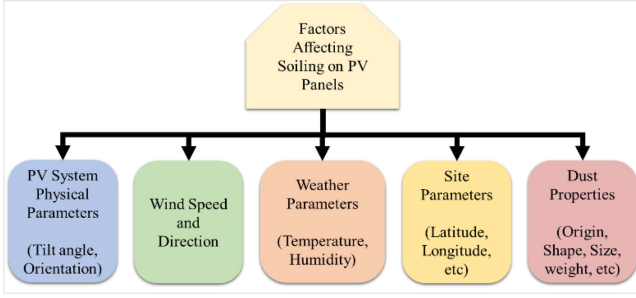


Fig. 1. Parameters impacting soiling of PV panels.

assessed with relative ease [7], [8]. In the previous articles, performance of the PV modules in areas of high dirt accumulation is performed by comparing the power yield from modules during dusty and clean scenarios. Skomedal et al. [9] proposed a model for estimating the losses in PV systems due to degradation and soiling. Here an iterative solution was proposed using the Kalman filter to estimate the losses and assess the performance by comparing it with year-on-year data from the PV system. Research work has been found where image processing has been implemented with deep learning techniques to accurately predict the SL [10], [11]. Yuan et al. [12] proposed a mechanism to predict PV soiling due to evaporation and formation of dew on the glazing surface of modules. The impact of dust and effective cleaning performance of PV panels for maximizing profits from solar plants is presented by Mithu et al. [13]. The various factors affecting the soiling on PV modules are depicted in Fig. 1.

According to the updated world map of the Koppen–Geiger climate classification [14], Saudi Arabia is a hot arid desert in a semi-tropical region. Due to Vision 2050 of Kingdom of Saudi Arabia, the boost in the solar PV sector is massive, but the geographical conditions pose a bigger challenge to the performance of the installed PV system. Thus, this article is focused to study the impact of dust and dirt accumulation on the PV system in a semi-tropical hot arid desert country. Saudi Arabia, being in the early stages of adopting solar PV technology, has a wide scope for application of this article. Moreover, this article attempts to develop a model to estimate the power loss in PV systems due to dirt and dust accumulation on the surface of the modules. The proposed methodology utilizes measurements of rooftop PV power generation as an input and outputs an estimate of the SL over time, expressed as a power loss compared to the unsoiled performance. Secondary data like total loss and monthly loss patterns may be computed using this technique.

This article further discusses different derating factors associated with the PV system in Section II of this article. Section III explains the methodology incorporated for modeling the dirt derating factor. Further, data analysis and results discussion is presented in Section IV of this article. Section V presents the conclusion of the research work done followed by references and author biographies.

II. SOILING AND DERATING FACTOR

Degradation is the gradual deterioration of PV system performance caused by mechanisms that are not reversed in the

appropriate field circumstances. It is widely acknowledged that all PV systems degrade. To decrease the risk associated with investing in PV, PV engineers must be able to detect underperforming systems and the reasons for such performance, while the field as a whole must be able to estimate normal degradation rates using fleet-scale data. Unrecovered soiling appears as deterioration in the current application and is indistinguishable from other degradation modes without the use of extra sensors intended to assess soiling.

For a PV module with rated voltage (V_r) and current (I_r), the generated voltage (V_{pv}) and current (I_{pv}) for incident irradiance (E) is given as follows:

$$V_{pv} = V_r (1 + \alpha (T_m - T_a)) \left[1 + \beta \left(\ln \left(\frac{E}{E_r} \right) \right) \right] \quad (1)$$

$$I_{pv} = I_r (1 + \alpha (T_m - T_a)) \left(\frac{E}{E_r} \right). \quad (2)$$

When dirt accumulated PV module is incident to the irradiance the output of the PV system is reduced and modified as follows:

$$V'_{pv} = V_r (1 + \alpha (T_m - T_a)) \left[1 + \beta \left(\ln \left(\frac{E}{E_r} * \varphi \right) \right) \right] \quad (3)$$

$$I'_{pv} = I_r (1 + \alpha (T_m - T_a)) \left(\frac{E}{E_r} * \varphi \right) \quad (4)$$

$$\varphi = \exp \left(- \frac{3 \cdot \varepsilon \cdot m_{\text{eff}}}{8 \epsilon_{PM} A_{PV} r_d \cos \theta \cos \Phi} \right) \quad (5)$$

where φ is the reduced transmittance of the PV module, and α and β are current and voltage temperature coefficients, respectively. T_m is the module temperature and T_a is the ambient temperature, m_{eff} is the remaining portion of the accumulated dust after rebounding [15], ε is the transmittance of a particle, Φ is the angle of incidence, ϵ_{PM} is PM concentration, and A_{PV} is the area of PV module.

For power engineers with increasing penetration of renewable energy, predicting the output power of a PV system is of key importance. Considering atmospheric conditions for PV system, the power output from PV system can be calculated by multiplying (3) and (4), whereas a simplified method to estimate the power of the PV system is

$$P_{PV} = A_{PV} \left(\frac{E_{\text{sun}}}{t_{\text{psh}}} \right) \times c_d \times K_d \times K_t \quad (6)$$

Here, c_d is the derating coefficient. These derating coefficients signify the PV system efficiency losses. E_{sun} is the daily solar irradiation. t_{psh} is the period of peak sunshine hour. Meanwhile, η_{PV} and η_{INV} are the efficiency of the panel and inverter, respectively. The derating factors involve module power tolerance, module mismatch, wiring, soiling, shading, and downtime [16]. Here, the derating coefficient c_d is defined as

$$c_d = \eta_{PV} \times \eta_{INV} \times \varepsilon_{\text{cable}} \times K_{\text{aging}} \quad (7)$$

Considering the environmental conditions along with physical parameters of the PV system, c_d has been carefully modeled. The reduction in performance of the modules due to absorbing of the

TABLE I
PV SYSTEM DETAILS

Coordinates	26.35° N, 43.76° E
Tilt angle	25.7°
PV modules	450 W_p
Cell type	Monocrystalline
Cell arrangement	144 [$2 \times (12 \times 6)$]
Dimensions	2108 × 1048 × 40 mm
Arrangement	7 (series) × 2 (parallel strings)
Peak power	6300 W_p
Front cover	3.2 mm tempered glass
Module efficiency	20.37%

heat is given by K_t . Effect of wiring and panel age are defined by $\varepsilon_{\text{cable}}$, K_{aging} , respectively. This article is focused on predicting the derating factor due to dirt and soiling, i.e., K_d . In most of the literature, the prediction of K_d is studied by comparing the performance of PV module/system in two conditions: dusty or dirty module and clean module [17], [18], [19], [20], [21]. This data of clean and dusty PV panels were recorded by the following methods:

- 1) Two similar panels at the same location were installed, one regularly cleaned and maintained, whereas the other one was kept dusty and dirty.
- 2) PV panel was kept dirty for a specific period and then was cleaned regularly for a specific period.
- 3) The data were recorded for the nonrainy season and then data were recorded for the rainy season when natural cleaning was done.

All of the above methodologies have their drawbacks, either these methods require relatively higher costs for analysis or they are dependent on environmental factors. In addition, during some methods, the changing meteorological parameters are ignored as the data were recorded on different times or days or months when the panels were cleaned.

III. PROPOSED METHOD

In this article, K_d for a 6300 W_p solar PV system is measured for a location in Saudi Arabia. The PV system installed has 14 panels of 450 W_p in 7×2 configuration, further details about the PV system is presented in Table I. The PV system is installed at the rooftop of the College of Engineering, Qassim University, Saudi Arabia. It has an online energy and meteorological parameters monitoring system for PV system. This monitoring system logs the irradiance (W/m^2), module temperature ($^{\circ}C$), wind speed (m/s), and PV power (W). The following parameters were recorded for every 5 min.

The majority of Saudi Arabia has hot, arid desert weather with extremely high temperatures. Sandstorms strike Saudi Arabia from 12 to 30 times every year, with the likelihood in the east being higher than the rest of the nation. Thus, PV generation is affected by these challenging environmental conditions. This article presents a mathematical model for estimating the losses due to dirt and soiling of PV plants. In this methodology, manual cleaning was not performed since the installation of PV system and the only cleaning of modules due to rain water was considered and analyzed.

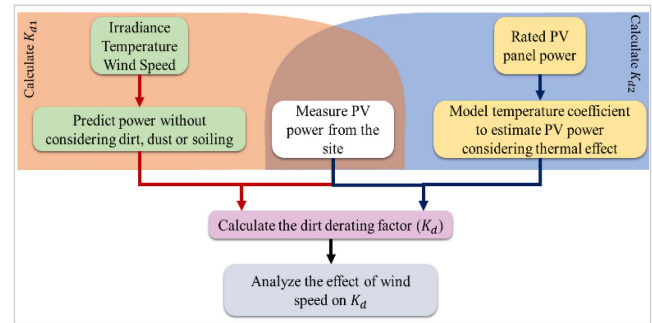


Fig. 2. Methodology adopted for the research work.

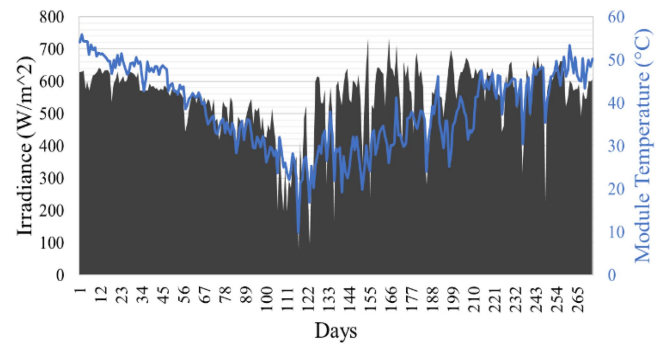


Fig. 3. Irradiance and module temperature of the rooftop site.

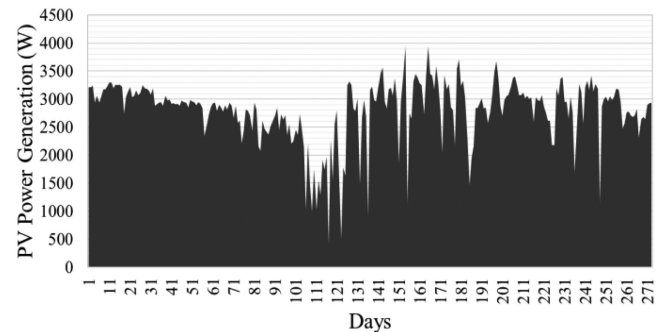


Fig. 4. PV power production at the site.

The proposed article follows the methodology described in Fig. 2. The daily average irradiance incident on the PV system and module temperature of PV modules is presented in Fig. 3. PV power generated at the rooftop site is depicted in Fig. 4. The data represented in Figs. 3 and 4 are from September 2021 to May 2022. To study the impact of dirt and soiling during this period on the power generation of PV system, estimation of the derating factor is of key focus. This approach overcomes the limitations of the previously mentioned methods. In this article, dirt derating factor is modeled considering predicted PV power, PV power (thermal coefficient), and ac power stored by online energy monitoring system for PV system.

To study the impact of soiling or dirt on the performance of the PV panel, following steps have been followed:

Step 1: Using the data of irradiance, temperature and wind speed predict the PV power. The rated power of the PV module

TABLE II
VALUES OF DERATING CONSTANTS

Constants	Description	Value
η_{PV}	Efficiency of PV panels	0.204
η_{INV}	Efficiency of inverter	0.97
ε_{cable}	Cable loss	0.99
K_{aging}	PV panel aging loss	0.99

should be noted and note the ac power of PV system from the monitoring system.

Step 2: Model the temperature coefficient for the PV system to estimate the PV power for the 6300 W_p system.

Step 3: Calculate the dirt derating factors (K_{d1} and K_{d2}) with respect to the ac power from PV system considering predicted power and mathematically estimated power, respectively.

Step 4: K_d and meteorological parameter analysis are performed.

In this method, K_{d1} and K_{d2} are considered because the rooftop PV system does not have a cleaning schedule. Thus, the effect of soiling and dirt on the PV system is estimated using both predictive PV power and regression model. The power prediction method incorporated is presented in [22]. This power is represented by the term P_p . The power from the online energy management system is defined by P_{PV_s} . The rated module power is denoted as P_{PV_r} . The primary objective of the article is to retrieve the data from the data logging device and perform mathematical operations to calculate power and estimate derating factors. Here, two similar approaches of have been incorporated into one method. First, K_{d1} has been calculated using P_{PV_s} and P_p and second K_{d2} has been calculated using P_{PV_s} and PV power of the system considering thermal derating factors (P_{PV_t}).

$$K_d = \text{mean}(K_{d1} + K_{d2}) \quad (8)$$

$$\text{ActualPVPower} = K_d \times \beta \times P_{PV_r} \quad (9)$$

$$\beta = \eta_{PV} \times \eta_{INV} \times \varepsilon_{cable} \times K_{aging} \times K_t \quad (10)$$

Here, β is the derating constant and is defined by (10). η_{PV} , η_{INV} , ε_{cable} are constants with fixed values for a particular PV system. K_{aging} is the factor representing degradation in the performance of the PV panels due to aging. Few authors have proposed different methods to model this factor [23], [24]. The values for each of the factors are defined as in Table II.

A. Calculation for K_{d1}

From [22], the predicted PV power is calculated and (9) is further modified as

$$P_{PV_s} = K_{d1} \times P_p \quad (11)$$

As this article does not incorporate cleaning of the PV system, in this article natural cleaning, i.e., rain water cleaning, is analyzed, and records the data of the grid online rooftop PV system and calculates the derating factor using values from estimated power and recorded data.

TABLE III
REGRESSION ANALYSIS OF β VS T_{mod}

Sum of squared errors (SSE)	1.646×10^{-25}
R-square	1
Adjusted R-square	1
Root mean square error (RMSE)	2.183×10^{-16}

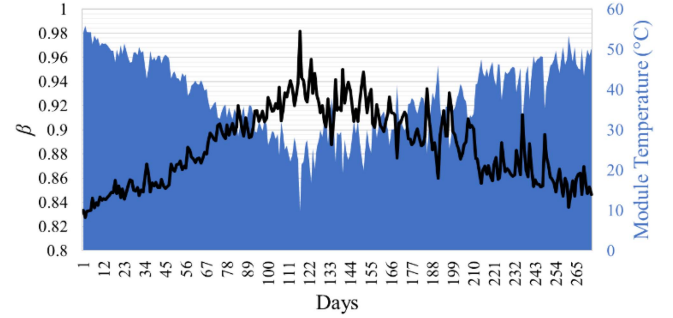


Fig. 5. β and T_{mod} variations for the PV system.

B. Calculation for K_{d2}

From (9) and (10), the thermal coefficient of the PV system and its effect on the power is modeled using regression analysis

$$P_{PV_s} = K_{d2} \times P_{PV_t} \quad (12)$$

$$K_{d2} = \frac{P_{PV_s}}{P_{PV_t}} \quad (13)$$

$$P_{PV_t} = \eta_{PV} \times \eta_{INV} \times \varepsilon_{cable} \times K_{aging} \times K_t \times A_{PV} \times G(t) \quad (14)$$

$$K_t = 1 + K_{tmp}(T_{mod} - T_{stc}) \quad (15)$$

$G(t)$ is the irradiance incident on the module at time t , K_{tmp} is the temperature coefficient of the PV module for the power output at maximum power point, i.e., standard testing conditions. This factor signifies the reduction in peak power of the PV system for every $^{\circ}\text{C}$, T_{mod} is the module temperature in $^{\circ}\text{C}$, and T_{stc} is 25°C . Hence, the variation of $(\eta_{PV} \times \eta_{INV} \times \varepsilon_{cable} \times K_{aging} \times K_t)$ for varying conditions for PV system is modelled using regression analysis. The data of considered for this analysis are available using online data monitoring system, such as power, temperature, and irradiance. Hence, from this regression analysis, β is defined as follows and parameters for this analysis are given in Table III:

$$\beta = -0.003352 T_{mod} + 1.015 \quad (16)$$

Thus, using (15), (13) is modified as

$$K_{d2} = \frac{P_{PV_s}}{(-0.003352 T_{mod} + 1.015) P_{PV_r}} \quad (17)$$

Here, Fig. 5 shows the variation of β and T_{mod} for the effective correlation between module temperature and derating constant as defined in (15). From Fig. 5, it can be observed that during low PV module temperature conditions, β is considerably higher in the range of 0.90–0.98, whereas during hotter conditions with module temperature ranging in between 35°C and 55°C , β is

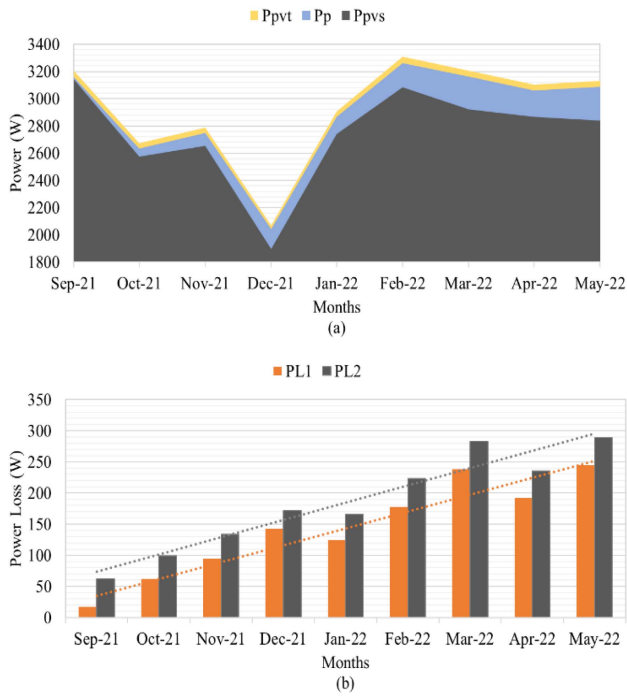


Fig. 6. (a) Variation in P_{PV_t} , P_p , and P_{PV_s} . (b) Monthly power loss.

considerably in lower ranges from 0.83 to 0.89. Thus, showing a negative correlation as modeled in (16).

IV. RESULTS AND DISCUSSION

In this portion of the article, the results and discussion are presented for the dirt-soiling derating factor of the 6.3 kWp PV system. The reduction in power due to the effect of dirt accumulation on the PV modules is analyzed using mathematical modeling and regression analysis. Furthermore, the Pearson correlation coefficient is presented for the relationship between wind speed, dirt-dust accumulation, and reduction in power for the PV panels in dry arid desert type of locations (specifically Saudi Arabia). The results and data analysis of the research article is categorized in two sections. First, the calculation of K_d and its impact on PV power generation capability, followed by developing a relationship between K_d and meteorological parameters for the site.

A. Calculation of K_d

Fig. 6(a) presents the average monthly power for the nine months from September 2021 to May 2022. These values represent the different PV power required for calculation of soiling derating factor K_d . P_{PV_t} is the power calculated using temperature coefficient, P_{PV_s} is the ac power stored by the data logging system for monitoring of PV system and P_p is the predicted PV power using real-time meteorological parameters. From Fig. 6(b), it can be observed that the losses in PV power show a linear increase with every month since the installation.

Since the PV modules of the PV system are not cleaned regularly and are cleaned naturally due to rain, there is an increased accumulation of dust and dirt particles on the module

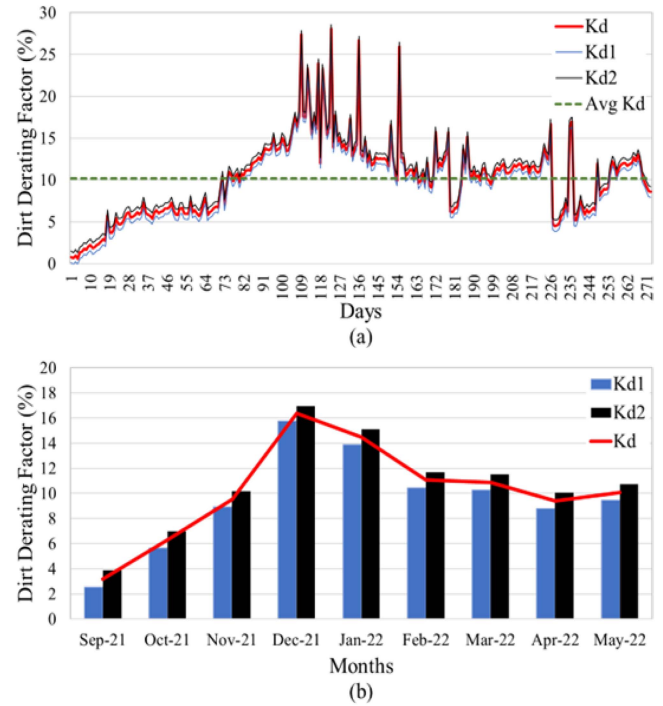


Fig. 7. (a) Daily variation of dirt derating factor. (b) Monthly variation of dirt derating factor.

surface resulting in more power loss every month as shown in Fig. 6(b). These losses in power are defined as follows:

$$P L_1 = P_p - P_{PV_s} \quad (18)$$

$$P L_2 = P_{PV_t} - P_{PV_s} \quad (19)$$

For the installed 6300 kW_p PV system, the two dirt derating factors (K_{d1} and K_{d2}) is shown in Fig. 7 alongside K_d . K_d has been calculated using K_{d1} and K_{d2} , because this method considers modeling of dirt derating factor with mathematical model and also with the artificial neural network prediction model. Thus, the estimated dirt derating factor is more robust and accurate than previous methods. Fig. 7(a) and (b) presents the daily and monthly variations in K_d , K_{d1} , and K_{d2} , respectively. The average value of K_d is 10.15%. The derating factor is very small in September 2021, i.e., approximately 2.5% because the PV system was setup in August 2021 and due to the system being in its early stages the accumulation of dust and dirt on panels is less as compared to in November and December. The maximum value for K_d obtained for December 2021 and January 2022 at 16.40% and 14.50%, respectively.

Meanwhile, in Fig. 8, the graph compares the part of the PV system's output power reduction caused by soiling to the total amount of output power generated in percentage. Both statistics indicate that the impact of soiling on this plant was considerable and that the percentage of the power decrease due to soiling may exceed 15% over an extended period of time. This implies that soiling might occasionally result in a reduction in output power of 20% for the entire system.

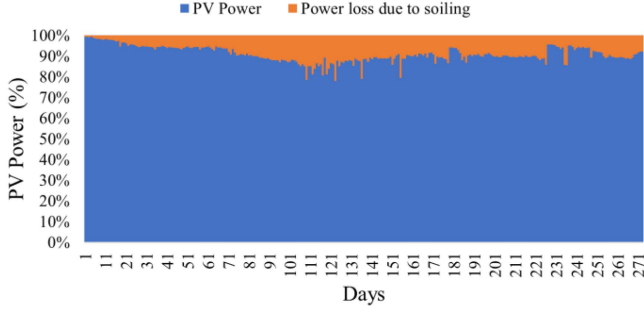


Fig. 8. Percentage of PV power loss due to K_d , with respect to overall generation.

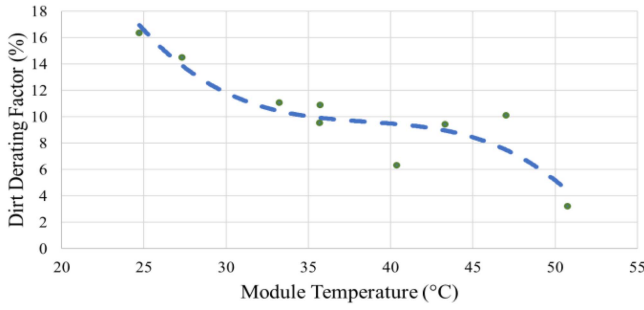


Fig. 9. Scatter plot for K_d vs T_{mod} with best fit line.

B. Relationship Between K_d and T_{mod}

The PV site considered in this article is subjected to varying meteorological parameters, such as wind speed, module temperature (T_{mod}), and irradiance (G) coupled with constant accumulation of dirt without cleaning. Soiling not only provides with soft and partial shading of PV panels, but also impacts the overall temperature of the surface. The regression analysis was performed for finding the relationship of K_d with different parameters, but the best results were obtained for K_d vs T_{mod} . Thus, a correlation factor between dirt derating factor and module temperature has been developed for the PV system. Scatter plot and curve fitting is performed to best suit the relationship between K_d and T_{mod} . The data analysis and regression analysis presented in this article are performed on MATLAB coder and curve fitting application. For reference, the results from MATLAB were compared with IBM SPSS and Microsoft Data Analytics tools and were found to be satisfactory. The scatter plot of K_d vs T_{mod} is given in Fig. 9.

The regression analysis for finding the best-fit curve and relationship is presented in Table IV. Here, R and R^2 are the regression coefficient determining a good relationship between factors namely, K_d and T_{mod} . From Table IV, it is observed that the cubic relationship for K_d and T_{mod} is the best as the values of R and R^2 is nearest to 1. To further analyze this cubic relationship, ANOVA test is performed and Table V represents this. ANOVA is an analysis of variance that implies if the results of the research are significant.

Table IV helps in establishing the type of relationship between K_d and T_{mod} , whereas Table V assists in formulating the mathematical equation with T_{mod} as the independent variable

TABLE IV
REGRESSION ANALYSIS OF K_d vs T_{MOD}

Relationship	R^2	R	Adjusted R^2	Adjusted R
Linear	0.749	0.865	0.72	0.848
log	0.782	0.884	0.742	0.861
Power	0.797	0.892	0.714	0.844
Exponential	0.773	0.879	0.751	0.866
Quad	0.886	0.941	0.852	0.923
Cubic	0.909	0.953	0.899	0.948

TABLE V
ANOVA TEST

	Sum of squares	df	Mean square	F value	Sig
Regression	84.33	2	42.15	8.9	.009
Residual	18.47	6	3.07		
Total	102.8	8			

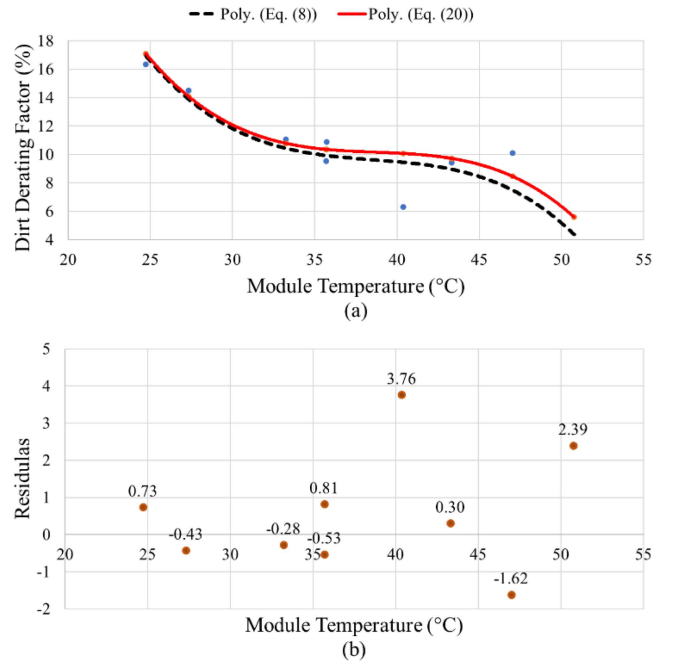


Fig. 10. (a) Comparison of best fit curve from different relationships. (b) Residual plot of from the relationship.

and K_d being the dependent variable. Hence, the regression and curve fitting analysis is obtained as

$$K_d = 145.49 - 10.393 \times T_{mod} + 0.2672 \times (T_{mod})^2 - 0.0023 \times (T_{mod})^3 \quad (20)$$

For effective and robust analysis, the formulated relationship between K_d and T_{mod} in (20) has to be compared with (8). The plot from both the equations is shown in Fig. 10(a). The residual plot is shown in Fig. 10(b).

The performance indices such as root mean squared error (RMSE), mean average percentage error (MAPE), mean average error (MAE), and residual some of squared (RSS) are given

TABLE VI
PERFORMANCE INDICES

Indices	Value
RSS	26.35
MAPE	4.71%
MAE	0.32
RMSE	0.57

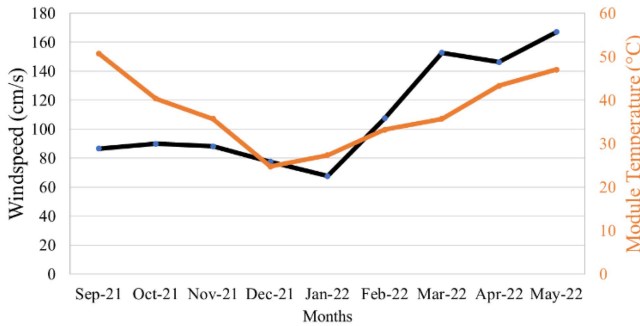


Fig. 11. Monthly wind speed and module temperature.

in Table VI. These values are within admissible limit thus establishing the relationship formulated in (20) is acceptable.

Thus, the relationship formulated between K_d and T_{mod} indicates that the effect of the dirt derating factor decreases with an increase in the module temperature. Moreover, it can be concluded that the effect of K_d on PV power during cold temperatures is higher in comparison to hot conditions. Hence, it is further analyzed that wind speed for the hot temperature months is higher, whereas for December and January with low temperatures the wind speed is low, thus impacting the dispersion of soiling and dirt on the PV panels. The plot for monthly wind speed and T_{mod} is presented in Fig. 11 for better understanding.

The developed relationship in (20) is a much simpler model with an accuracy of more than 95%. The empirical model defined in (8) requires P_p , P_{PV_t} , β , and c_d values, whereas this prediction requires T_{mod} . Hence, it can be utilized for estimating the performance of the PV system due to soiling, dirt, and dust.

V. CONCLUSION

The power generation capability of the PV system crucially depends on the soiling of the modules, especially for a hot arid desert in semi-tropical region. This article had been performed for 6.3kW_p PV system installed in Burydah, Saudi Arabia (i.e., Qassim region) in August 2021. The article focuses on analyzing the performance of the installed PV system considering different meteorological parameters. Furthermore, the impact of dirt, dust, and soiling on the system is studied and a robust empirical model has been developed. It is concluded that from the first month of installation of a PV system about 3% of power is reduced. This reduction in power is increased to 18% within four months of setup. This article also aimed at studying the impact of different meteorological parameters on K_d . This helped in developing a simpler relationship between K_d and T_{mod} using mathematical regression and curve-fitting tools. Hence, the performance of

the empirical model with the regressed model developed using T_{mod} is compared using performance indices. The values of RSSE, RMSE, MAE, and MAPE are 26.35%, 0.57%, 0.32%, and 4.71%, respectively. Moreover, this methodology can be applied to other locations and sites to develop this correlation and regression to estimate the soiling performance of the PV system. This forecasting of K_d is essential for planning a new PV system for prior assessment of the performance, and scheduling cleanup to mitigate its impact.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education, and Saudi Arabia for funding this research work through the project number (QU-IF-1-3-3). The authors also thank to the technical support of Qassim University.

REFERENCES

- [1] J. Wei, Y. Zhang, J. Wang, X. Cao, and M. Khana, "Multi-period planning of multi-energy microgrid with multi-type uncertainties using chance constrained information gap decision method," *Appl. Energy*, vol. 260, 2020, Art. no. 114188.
- [2] J. Wei, Y. Zhang, J. Wang, and L. Wu, "Distribution LMP-Based demand management in industrial park via a bi-level programming approach," *IEEE Trans. Sustain. Energy*, vol. 12, no. 3, pp. 1695–1706, Jul. 2021.
- [3] K. Ilse, B. Figgis, V. Naumann, C. Hagendorf, and J. Bagdahn, "Fundamentals of soiling processes on photovoltaic modules," *Renewable Sustain. Energy Rev.*, vol. 98, pp. 239–254, 2018.
- [4] M. Maghami et al., "Power loss due to soiling on solar panel: A review," *Renewable Sustain. Energy Rev.*, vol. 59, pp. 1307–1316, 2016.
- [5] D. Jordan, C. Deline, S. Kurtz, G. Kimball, and M. Anderson, "Robust PV degradation methodology and application," *IEEE J. Photovolt.*, vol. 8, no. 2, pp. 531–525, Mar. 2018.
- [6] Å. Skomedal, H. Haug, and E. Marstein, "Endogenous soiling rate determination and detection of cleaning events in utility-scale PV plants," *IEEE J. Photovolt.*, vol. 9, no. 3, pp. 858–863, May 2019.
- [7] S. Kalogirou, R. Agathokleous, and G. Panayiotou, "On-site PV characterization and the effect of soiling on their performance," *Energy*, vol. 51, pp. 439–446, 2013.
- [8] J. Lopez-Garcia, A. Pozza, and T. Sample, "Long-term soiling of silicon PV modules in a moderate subtropical climate," *Sol. Energy*, vol. 130, pp. 174–183, 2016.
- [9] A. Skomedal and M. Deceglie, "Combined estimation of degradation and soiling losses in photovoltaic systems," *IEEE J. Photovolt.*, vol. 10, no. 6, pp. 1788–1796, Nov. 2020.
- [10] W. Zhang et al., "Deep-learning-based probabilistic estimation of solar PV soiling loss," *IEEE Trans. Sustain. Energy*, vol. 12, no. 4, pp. 2436–2444, Oct. 2021.
- [11] R. Cavieres, R. Barraza, D. Estay, J. Bilbao, and P. Valdivia-Lefort, "Automatic soiling and partial shading assessment on PV modules through RGB images analysis," *Appl. Energy*, vol. 306, 2022, Art. no. 117964.
- [12] W. Yuan, Z. Liao, K. He, and S. Huang, "Mechanism of photovoltaic module soiling in the presence of dew," *Sol. Energy Mater. Sol. Cells*, vol. 248, 2022, Art. no. 111962.
- [13] M. Mithu, T. Rima, and M. Khan, "Global analysis of optimal cleaning cycle and profit of soiling affected solar panels," *Appl. Energy*, vol. 285, 2021, Art. no. 116436.
- [14] M. Peel, B. Finlayson, and T. McMahon, "Updated world map of the Köppen–Geiger climate classification," *Hydrol. Earth Syst. Sci.*, vol. 11, pp. 1633–1644, 2007.
- [15] S. Sengupta et al., "Model based generation prediction of SPV power plant due to weather stressed soiling," *Energies*, vol. 14, 2021, Art. no. 5305.
- [16] L. Micheli and M. Muller, "An investigation of the key parameters for predicting PV soiling losses," *Prog. Photovolt.: Res. Appl.*, vol. 25, pp. 291–307, 2017.
- [17] D. Thevenard and S. Pelland, "Estimating the uncertainty in long-term photovoltaic yield predictions," *Sol. Energy*, vol. 91, pp. 432–445, 2013.

- [18] W. Jamil, H. Rahman, S. Shaari, and M. Desa, "Modeling of soiling derating factor in determining photovoltaic outputs," *IEEE J. Photovolt.*, vol. 10, no. 5, pp. 1417–1423, Sep. 2020.
- [19] Y. Chanchangi, A. Ghosh, S. Sundaram, and T. Mallick, "An analytical indoor experimental study on the effect of soiling on PV, focusing on dust properties and PV surface material," *Sol. Energy*, vol. 203, pp. 46–68, 2020.
- [20] A. Younis and Y. Alhorr, "Modeling of dust soiling effects on solar photovoltaic performance: A review," *Sol. Energy*, vol. 220, pp. 1074–1088, 2021.
- [21] R. Conceição, I. Vázquez, L. Fialho, and D. García, "Soiling and rainfall effect on PV technology in rural Southern Europe," *Renewable Energy*, vol. 156, pp. 743–747, 2020.
- [22] A. Kumar, M. Rizwan, and U. Nangia, "A hybrid intelligent approach for solar photovoltaic power forecasting: Impact of aerosol data," *Arabian J. Sci. Eng.*, vol. 45, pp. 1715–1732, 2019.
- [23] S. dos Santos, J. Torres, C. Fernandes, and R. Lameirinhas, "The impact of aging of solar cells on the performance of photovoltaic panels," *Energy Convers. Manage.: X*, vol. 10, 2021, Art. no. 100082.
- [24] B. Nehme, N. K. M'Sirdi, T. Akiki, and A. Naamane, "Contribution to the modeling of ageing effects in PV cells and modules," *Energy Procedia*, vol. 62, pp. 565–575, 2014.



Development of water quality management strategies for an urban river reach: A case study of the river Yamuna, Delhi, India

Nibedita Verma¹ · Geeta Singh¹ · Naved Ahsan²

Received: 21 June 2022 / Accepted: 19 November 2022
© Saudi Society for Geosciences 2022

Abstract

The urban reach of River Yamuna at Delhi is contaminated with a high volume of wastewater inflows amounting to a flow of about 3600 MLD consisting of BOD as high as 200 mg/l entering through sixteen outfalling drains within the 22 km of reach. The present study intended to develop water quality management approaches to bring out the BOD and DO concentrations within the prescribed limits. Four different scenarios with twelve cases have been developed to attain this objective. The model QUAL2Kw was calibrated and validated with secondary data. RMSE values were determined to validate the predicted and observed values. Four scenarios, including bottom algal modification, pollutant load modification, pollutant load modification with diversion, and diversion with flow augmentation, were assessed for February 2019 with low flow conditions. The study showed that flow augmentation and diversion of some drains (D1, D11, D12, D13) could ensure the required criteria for BOD < 3 mg/l throughout the reach. At 33 km from upstream, local oxygenation was required to maintain DO above 4 mg/l. Advanced treatment in drain 1 and diversion of a few smaller drains can also maintain the required standards. Results indicated that the reach required a combination of bottom algal reduction, flow augmentation, load modification, and local oxygenation to sustain the river quality. The present study attempts to develop a novel strategy for river water management under different scenarios for the urban reach of the Yamuna River, which is almost anoxic after outfalling drain 1.

Keywords Water quality management · Water quality model · Pollutant load · QUAL2Kw · Diversification · Flow augmentation

Introduction

The rapid population growth is putting high pressure on the water bodies of developing countries like western and southern Africa and south Asia (Turan et al. 2018). The anthropogenic activities release harmful contaminants into the aquatic environment and threaten living organisms

(Abbas et al. 2022). These contaminants lead to destroying the aquatic ecosystem. High oxygen-demanding pollutants discharged from agriculture, municipal, and industrial activities decrease the dissolved oxygen concentration below the extreme level in the receiving water. The aquatic ecosystem of the river becomes unbalanced and causes the mortality of fish and other organisms, produces odors, and becomes unaesthetic (Cox 2003). The wastewater coming from municipal sewage contains organic and inorganic substances, including toxic household chemicals such as insecticides, pesticides, detergents, personal care products, and other nonbiodegradable substances that cause harmful effects to human health when agricultural lands are irrigated with the water from receiving water resources (Khalil et al. 2022). Agricultural runoff contains residual synthetic fertilizer, and pesticides contain micronutrients (Tauqeer et al. 2022a, b). These microcontaminants also accumulate in the aquatic ecosystem and threaten living organisms (Abbas et al. 2022). Industrial wastewater contains inorganic and toxic substances demanding oxygen.

Responsible Editor: Amjad Kallel

✉ Nibedita Verma
nibedita_2k19phd@dtu.ac.in

Geeta Singh
geeta.singh@dce.ac.in

Naved Ahsan
nahsan@jmi.ac.in

¹ Environmental Engineering Department, Delhi Technological University, New Delhi, Delhi 110042, India

² Civil Engineering Department, Jamia Millia Islamia University, Jamia Nagar, New Delhi 110025, India

Water quality management approaches involve a highly multi-disciplinary decision related to the parameter's data input, response, and control (McIntyre 2004). Water quality models can evaluate the river water quality and predict the highly complex relationship between the wastewater and the system's response (Deksissa et al. 2004). These models can predict the water quality after wastewater discharges into the river and, therefore, can decide the degree of treatment required of wastewater for the designated use. A vast amount of municipal and industrial wastewater is generated due to high population growth, non-systematic urbanization, fast industrialization, and irrigation projects. With uncontrolled and accelerated development, wastewater discharged into the river negatively impacts water quality (Singh et al. 2007). Hence, surface water bodies urgently require a sustainable management system. Water quality management aims to reduce pollution's environmental impact (Ghosh and Mujumdar 2010).

The Yamuna River, Delhi segment is the most polluted water reach in India as partially treated, and untreated sewage disposal from wastewater sources cause almost no dissolved oxygen (DO) and high biochemical oxygen demand (BOD) throughout the year (Sharma 2013; Sharma et al. 2009; Yamuna et al. 2020). The deterioration caused due to sixteen main drains outfalling to 22 km of this reach with high oxygen demanding pollutants as point sources in between Wazirabad to Okhla, and a large volume of water is abstracted for the water supply of Delhi at Wazirabad barrage located 24 km from Palla (CPCB 2008). After Wazirabad, the stretch has almost no freshwater (CPCB 2006), and the river becomes a sewage line. Before entering Delhi, at Palla, the DO and BOD concentrations were within the desired limit ($\text{BOD} < 3 \text{ mg/l}$; $\text{DO} > 4 \text{ mg/l}$). These concentrations of BOD and DO at Palla are due to the non-appearance of the wastewater contribution from the enclosing catchment (Jaiswal et al. 2019). The present study analyzed the different scenarios by reducing the BOD from the contributing drains to increase the DO to the acceptable limit throughout the reach. The strategies have been created by varying flow augmentation in the upstream, diversification of different point sources, and increasing the percentage of treatment. QUAL2Kw was used to find a managerial approach for maintaining the Yamuna, Delhi water quality. QUAL2Kw is the enhanced version of QUAL2E, including new elements such as the unequal spacing of reach, anoxic conditions, slow and first biochemical oxygen demand, chemical oxygen demand (COD) as a generic constituent, algal death conversion, and DO interaction with the fixed plant (Pelletier and Chapra 2006). QUAL2Kw analyzes steady and dynamic conditions with multiple (Hobson et al. 2015) loadings and abstractions in any reach. In addition, QUAL2Kw analyzed bottom algae, sediment–water fluxes, nitrification, and pH. The present study intended to develop some

management options using this model QUAL2Kw. This model is suitable for the study reach as this segment is almost anoxic after joining the Najafgarh drain (D1). This model reduces the oxidation reaction to zero when oxygen availability is deficient (Pelletier and Chapra 2008a). This model has been successfully used for water quality management of the Zarjub River, north Iran (Zare Farjoudi et al. 2021), the Mousi River, south Sumatra (Lestari et al. 2019), Wenatchee River, Washington State, USA (Carroll et al. 2006), Bagmati River, Nepal (Kannel et al. 2007), The Yeongsan River, southwestern Korea (Cho and Lee 2019), Bedog River, Indonesia (Setiawan et al. 2018), The Wenatchee River with two tributaries, USA (Cristea and Burges 2010), Silver Creek, Utah (Hobson et al. 2015), Jordan River, Utah, USA (Neilson et al. 2013), Zarjub River, Iran (Nikoo et al. 2016), Cértima River, Portugal (Oliveira et al. 2012), South Umpqua River, Oregon (Turner et al. 2009), and Minho River, Portugal (Santos et al. 2013). Studies show that this model is used worldwide for water quality management and waste load allocation projects. For the study reach, various water quality assessments research has been performed (Bhargava 1985; Kazmi and Hansen 1997); Paliwal and Sharma 2007; Sharma et al. 2009; Parmar and Keshari 2014a; Sharma et al. 1999). The water quality management and waste loading were analyzed using the model QUAL2E (Parmar and Keshari 2014a; Paliwal and Sharma 2007) and generated scenarios with varying load and flow augmentation. Paliwal and Sharma (2007) stated the requirement of advanced treatment with a minimum flow of $10 \text{ m}^3/\text{s}$ to maintain water quality for this reach. Parmar and Keshari (2014a) also used QUAL2E for waste load allocation of Yamuna River reach, Delhi, and concluded that flow augmentation is not feasible for this reach. However, QUAL2E can only simulate equal spacing segments while outfalling drains of this reach are unequally spaced. Moreover, this reach is almost anoxic after massive water withdrawal at Wazirabad, and QUAL2E cannot predict under this circumstance. The enhanced version of QUAL2E, the QUAL2Kw model, is most suitable for this reach as it can accommodate the unequal spacing of input load and anoxic conditions with almost zero oxygen level. The present study's objectives were to arrive at approaches to manage the BOD and DO level of the river reach by bottom algae modification, augmenting flow, diverting the major drains, enhancing treatment, and local oxygenation. The QUAL2Kw model was established with populated and validated to attain these objectives. Statistical analysis was used to check the relationship between predicted and observed values. The present study is a novel approach for the river water management of urban reach, Yamuna River, Delhi, by varying scenarios and using the model QUAL2Kw for obtaining a strategy to retain the oxygen level above 4 mg/l throughout the reach.

Materials and methods

Study area

The river Yamuna originated from Saptrishi Kund, and after traveling 1376 km, it confluences to the river Ganga. After flowing from its source, the river travels through several valleys, approximately 172 km of the Himalayan segment, to meet the upper part. The upper segment started at Tajewala Barrage, and after 224 km, the river reached Delhi at Palla (CPCB 2006). Furthermore, traveling 22 km toward Delhi, the river runs the Wazirabad barrage, where water abstracts the domestic water supply of Delhi. After abstraction, the river contains low flow except during the monsoon season (Yamuna et al. 2020). Sixteen main drains contribute to partially treated and untreated wastewater between Wazirabad and Okhla, making the stretch highly polluted (Fig. 1) (Yamuna et al. 2020). Najafgarh drain, Delhi Gate drain, Sen Nursing Home drain, Barapulla, Tughlakabad, and Shahdara drain contribute approximately 86% of wastewater and 75% of organic matter as BOD (Yamuna et al. 2020). In 2019, the Najafgarh drain contributed 1938.38 MLD wastewater with 133.82 TPD BOD (CPCB, 2020). This point source contributes 64% of the total wastewater and 50% of the total BOD load into the Delhi reach. For the present study, 44 km of the Delhi reach of Yamuna River has been taken and shown in Fig. 1. Study area locations are shown in Table 1. The total catchment area covering the Delhi mega metropolitan city is approximately 1483 km² which is 0.4% of the total catchment area of the river. Four monitoring stations, Palla (M1) at the upstream point; old railway bridge (M2), approximately 27.4 km upstream; Nizamuddin (M3), around 36.4 km upstream; and Okhla (M4), about 44 km upstream, have been taken. Sixteen main drains are shown in Fig. 2 as point sources, and water quality management options have been determined.

Methodology

Modeling framework

The model QUAL2Kw assumes the river to be in a 1-dimensional and steady-state condition. It stimulates the river and does not incorporate the branches and tributaries. QUAL2kw accommodates features such as a software interface, unequally spaced reaches instead of equally spaced reaches with multiple loading and abstractions at any point, anoxia by reducing sediment oxygen demand, nutrient fluxes, bottom algae, pathogenic bacteria, suspended solids, light extinction, pH, alkalinity, total inorganic carbon, hyporheic exchange, and sediment pore water. These features are automatically calibrated using a genetic algorithm ((Hobson et al. 2015; Pelletier and Chapra 2008a, b). This model studies

the interaction of DO, plants, CBOD, and denitrification (Elshorbagy et al. 2005) and includes hyporheic and surface transient storage for each reach (Pelletier and Chapra 2008a, b). The model can also use for kinematic wave flow routing. For QUAL2Kw, a mass balance equation for a substance except for bottom algae can be presented in Eq. 1 (Pelletier and Chapra 2008a, b).

$$\frac{dc_i}{dt} = \frac{Q_{i-1}}{V_i} c_{i-1} - \frac{Q_i}{V_i} c_i - \frac{Q_{ab,i}}{V_i} c_i + \frac{E'_{i-1}}{V_i} (c_{i-1} - c_i) + \frac{E'_i}{V_i} (c_{i+1} - c_i) + \frac{W_i}{V_i} + S_i + \frac{E'_{hyp,i}}{V_i} (c_{2,i} - c_i) \quad (1)$$

where Q_i is the outflow from i into reach $(i+1)$ in m³/day, W_i is the external loading to reach i (mg/day), S_i is the sources and sinks of the constituents (mg/m³/day), $E'_{hyp,i}$ is the mass exchange between the water and the hyporheic sediment zone (m³/day), reach i (in m³/day), Q_{i-1} is the inflow, upstream reach $(i-1)$ (m³/day), $Q_{in,i}$ is the total inflow from point and

Table 1 Study area locations

Name of the stations	Longitude	Latitude
Palla (upstream)	28°49'46" N	77°13'27" E
Old railway bridge (monitoring station)	28°39' 33" N	77°14'42" E
Nizamuddin (monitoring station)	28°35' 31" N	77°18'21" E
Okhla (downstream)	28°32'40" N	77°18'49" E

diffused sources (m³/day), $Q_{ab,i}$ is the total outflow or abstractions (m³/day) to the reach, V is the volume of the reach, and C is the constituent concentration (Fig. 3).

The model has the option of auto-calibration using a genetic algorithm to maximize fitness by adjusting the significant parameters. The fitness is calculated with the following equation (Pelletier and Chapra 2008b)

$$f(x) = \left[\sum_{i=1}^n W_i \right] \left[\sum_{i=1}^n \frac{1}{W_i} \left[\frac{\frac{1}{m} \sum_{j=1}^m O_{i,j}}{[\sum (P_{i,j} - Q_{i,j})^2 / m]} \right]^{1/2} \right] \quad (2)$$

where $O_{i,j}$ is the measured values, $P_{i,j}$ is the predicted values, m is the predicted and measured pairs, W_i is the weighting factors, and n is the different state variables numbers. The model QUAL2Kw (Pelletier and Chapra 2008a, b) was used to evaluate the water quality parameter's fate and develop a management strategy for the Yamuna River Delhi stretch.

Data input

In the present study, the model QUAL2Kw was used to calibrate and validate data for four monitoring stations, M1, M2, M3, and M4 (Fig. 1), sixteen major outfalling drains named

Fig. 1 Yamuna river segment, Delhi

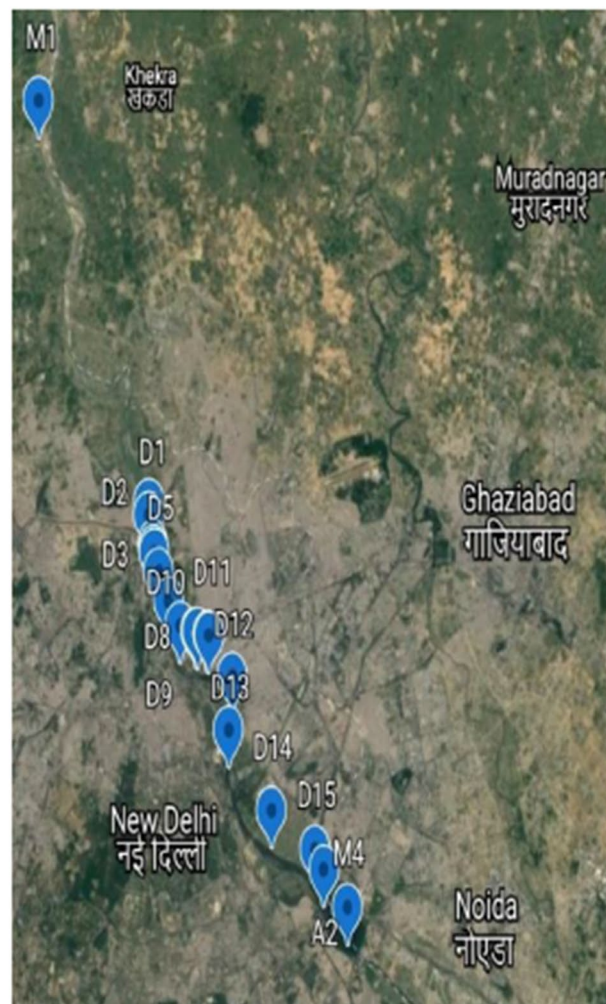
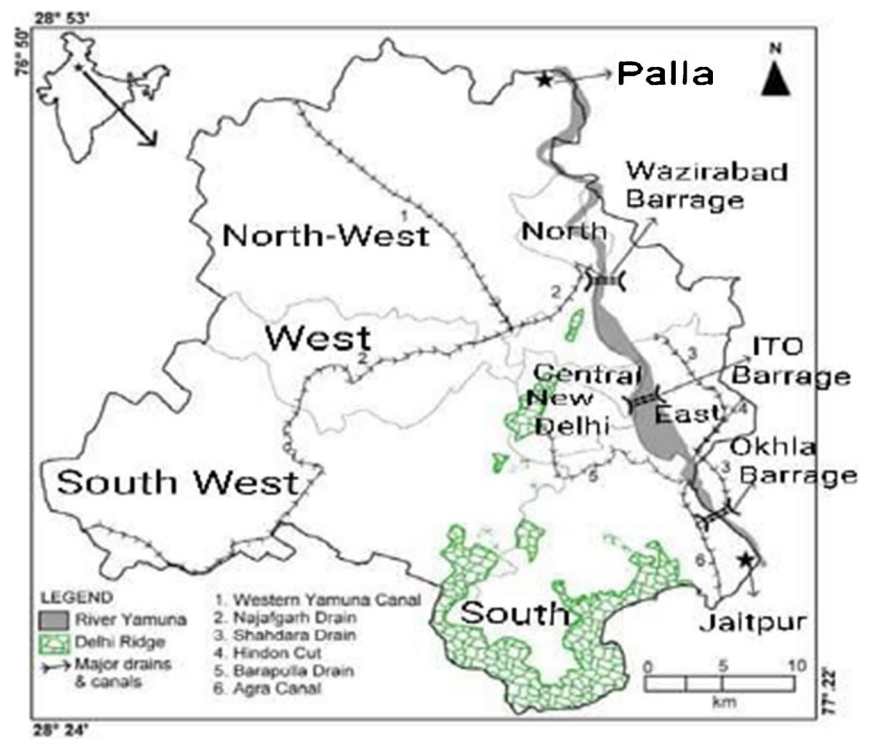
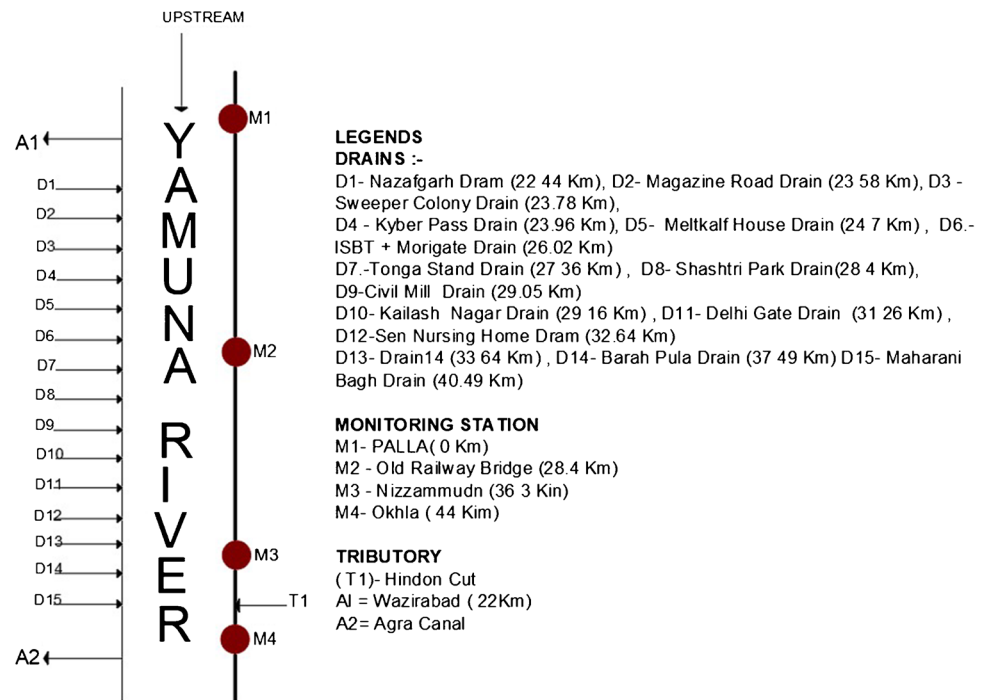


Fig. 2 Monitoring stations, drains, and abstraction locations

D1–D15 (D6 has joined two drains), two abstraction points A1 and A2 (Wazirabad barrage and Agra canal), and one significant tributary T1 (Hindon cut). The river reach has been divided into 21 segments and six sub-segments for the first segment (Palla to Wazirabad). QUAL2Kw can simulate unequal spacing, and the river segmentation is taken based on significant changes such as contributing drains and tributaries or abstraction points. Figure 2 illustrates the segmentation of the river reach with the flow profile. Water quality data for pH, alkalinity, BOD, COD, and conductivity were collected from 1999 to 2008 (CPCB 2006, 2008, 2009), personal communication with the Central Water Commission (CWC), and Delhi Pollution Control Committee (DPCC).

Daily flow and monthly water quality data, including river cross-section, were collected from personal communication with the CWC and DPCC. Table 2 shows the length, width, and depth of the river. Table 2 shows the entire reach is not very deep except for the location where D12 meets. Hence, it satisfies the one-dimensional flow system as the model assumes. Therefore, the O'Connor–Dobbins Eq. (1958) has been used for calculating reaeration coefficients for slow-moving shallow rivers (Chapra 1997). This study has assumed 100% sediment oxygen demand (SOD) coverage and prescribed SOD as 1, 10 cm of hyporheic sediment thickness, sediment porosity of 0.4, and 0% hyporheic exchange flow (Pelletier and Chapra 2008b). Other rate constants were taken as global values (Pelletier and Chapra 2008a, b). The model was used to simulate BOD, DO, pH, alkalinity, and COD. The calibration and validation were performed with the average data from February (1999 to

2005) and (2006 to 2008), respectively. Again, the modeling was performed for February 2019 (DPCC) data, and water quality management options were developed with the recent data. Data for February month have been taken as this is the dry season with low flow conditions. As India is a developing country with an inadequate monitoring system, monthly average data has been accepted for calibration and validation. Air temperature data was collected from IMD (Indian Meteorological Department), and hourly data were used for calibration and validation. Hydraulic characteristics have been calculated from rating curves (Eqs. 3–5) using regression methods

$$V = aQ^b \quad (3)$$

$$D = cQ^d \quad (4)$$

$$W = eQ^f \quad (5)$$

where V , Q , D , and W are the average velocity, discharge, depth, and width, respectively, for each segment, and a , c , and e are the velocity, depth, and width coefficients, as well as b , d , and f are the exponent, respectively. By solving these equations with regression analysis, rate constants were determined. After calibration, the model was verified, and RMSE values were calculated to check the minimum error. Automatic calibration was done after manual calibration. Water quality management options were developed by varying the upstream flow and diversification of the drains. The river segment was divided into twenty-one segments, and segment one was again

divided into six sub-reaches. Palla, old railway bridge, Nizamuddin bridge, and Okhla upstream (M1–M4) were taken as four monitoring stations. The river reaches are divided considering the significant drain discharges with high pollutants. Monthly average data were collected from 1999 to 2008 for pH, DO, BOD, COD, conductivity, alkalinity, temperature, NH_4 , and NO_3 from CPCB and CWC. The model has been calibrated and validated, taking time step 5.625 to avoid instability. The 10-year data were divided into a 7:3 ratio for calibration and validation. In February, air temperature and flow are low due to the late post-monsoon period. Wind speed is low, so aeration and minimum dew point temperature are also low. Therefore, the model was simulated for this month. The fitness was established with root mean squared error variance (RMSEV) for the predicted and observed values of calibration and validation. The model was verified with another dataset, and RMSEV values were calculated for fitness. Table 2 shows the hydraulic characteristics of the river reach, and Table 3 and Table 4 show the water quality parameters used for calibration

and validation. After calibration and verification, RMSEV values were calculated using Eq. 6 and shown in Table 5.

$$\text{RMSE} = 1/\text{mean} \left[\sqrt{\sum_1^i \{ (P_i - O_i)^2 \} / n} \right] \quad (6)$$

where P_i is the predicted values, O_i is the observed values, n is the no. of observation, and mean is the average of observed values.

Water quality management scenarios

After setting the model, different management strategies were analyzed. Four scenarios have been developed varying bottom algal concentrations, modifying pollutant load in the main drains, diversion with load modification, and diversion with flow augmentation. Twelve cases were generated for these scenarios. Four cases were generated for the scenario keeping bottom algal concentrations 100%, 50%, 60%, and

Fig. 3 Schematic Diagram of methodology

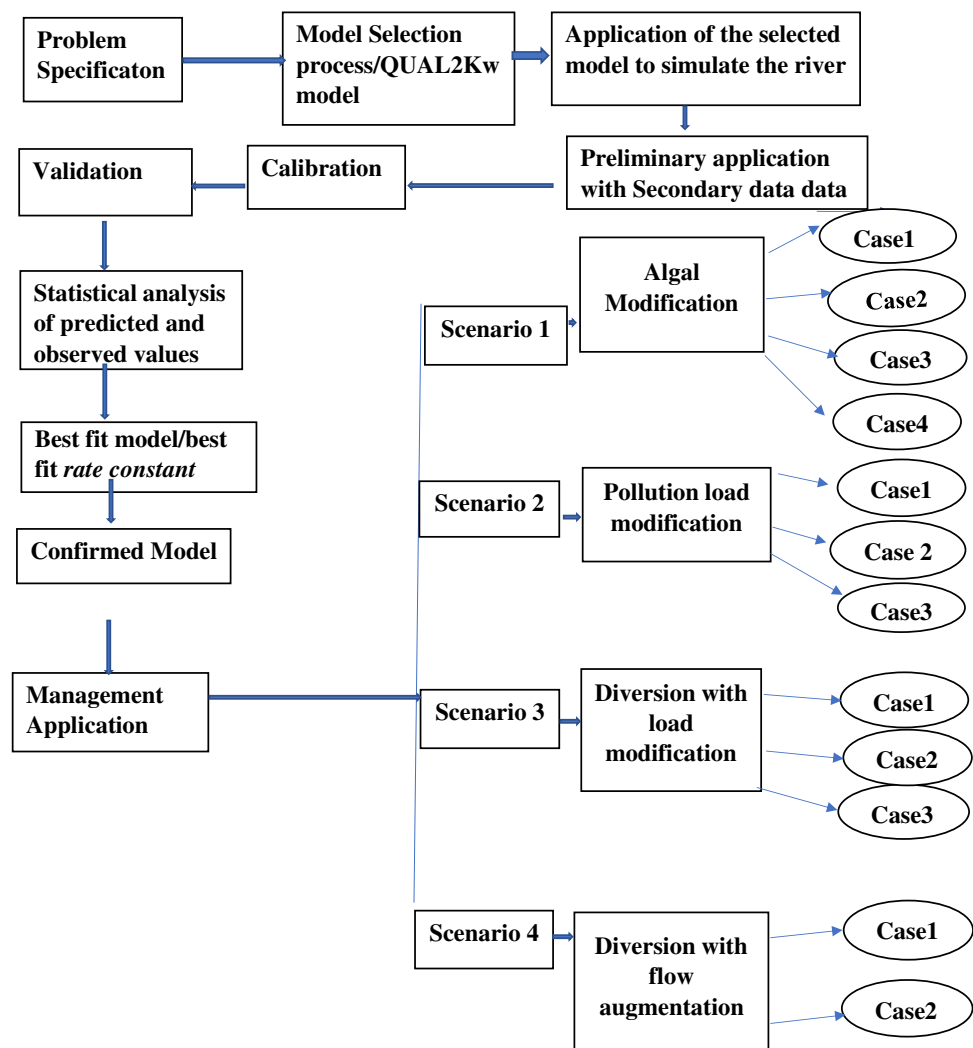


Table 2 Hydraulic characteristics of the river reach

Segment name	Location (km)*	Length (km)*	Depth (m)**	Width (m)**	Velocity (m/s)**
Palla	0	0	0.95	206.30	0.12
Wazirabad	22.00	22.0	0.50	119.70	0.04
Najafgarh drain	22.44	0.44	1.26	158.70	0.13
Magazine road drain	23.58	1.14	1.34	136.50	0.15
Sweeper colony drain	23.78	0.20	1.31	149.30	0.13
Khyber pass drain	23.92	0.14	1.51	115.30	0.15
Metcalf house drain	24.70	0.78	1.50	109.30	0.16
ISBT + Mori gate drain	26.02	1.98	1.45	130.61	0.14
Tonga stand drain	27.38	1.36	1.71	110.90	0.14
Old railway bridge	28.30	0.92	1.71	110.90	0.14
Shastri park drain	28.40	0.10	1.07	175.03	0.12
Civil mill drain	29.05	0.65	1.07	175.03	0.13
Kailash Nagar drain	29.16	0.11	1.28	150.44	0.13
Delhi gate drain	31.26	2.10	1.39	124.20	0.22
Sen nursing home drain	32.64	1.58	4.50	82.90	0.08
Drain 14	33.64	1.00	1.57	122.41	0.16
Nizamuddin bridge	36.34	2.70	1.97	121.60	0.13
Barapulla drain	37.49	1.15	1.97	121.60	0.13
Maharani Bagh drain	40.49	3.00	0.95	105.40	0.08
Agra canal	41.50	1.05	1.19	79.60	0.08
Okhla	44.00	2.50	1.19	79.60	0.08

*Yamuna et al. (2020). **Sharma et al. (2017)

75% (Fig. 5), covering the river bed to study the BOD reductions and increment of DO. For this scenario, there was no external load contributed to the reach. The second scenario was developed by varying the pollutant load concentration in the significant drains D1 and D14. The prescribed BOD load for scenario 2 was varied from 30 mg/l to 0 mg/l for case 1 and case 2, respectively. For case 3, BOD load was kept 0 for D1 and D14, and the rest outfalling sources were

assumed with flow containing 4 mg/l DO and 30 mg/l BOD as prescribed for the effluent wastewater standard by CPCB. Bottom algal conditions were kept 50% covered. In scenario 3, for D1, BOD load has been changed to 0 mg/l, 15 mg/l, and 5 mg/l for case 1, case 2, and case 3, respectively. Also, D13–D15 were diverted to D12, and the load for D12 has been modified for different cases. In scenario 4, D1 has been diverted, and flow augmentation was assumed upstream. In

Table 3 Water quality data for four monitoring stations

Station name	Distance (km)	Con (umhos)	DO (mg/l)	BOD (mg/l)	COD (mg/l)	Alkalinity (mg/l)	pH	NH ₄ ⁴ -N (µg/l)	NO ₃ -N (µg/l)
For calibration									
M1	0	384.14	8.47	1.61	5.51	336.18	7.8	241	576
M2	28.4	918.43	0.09	42.9	65.43	202.81	7.06	10,695	1035.7
M3	37.5	962.86	0.01	23.0	49.67	195.47	6.82	12,980	480.00
M4	44.00	895.29	0.15	22.7	40.43	188.07	6.88	14,257	1064.29
For validation									
M1	0	332	8.65	2.10	5.13	131.65	7.85	213	816
M2	28.4	1192.25	0.08	23	63.33	318.5	7.2	8627.5	810
M3	37.5	1508	0.02	33	39	318	6.93	11,422.5	460
M4	44	1477	0.03	40	29	321.5	6.975	11,422.5	582.5

Data from CPCB (2006), Sharma (2013), and CPCB (2008)

Table 4 Point sources data for calibration and validation

For calibration												
Point sources	Distance (km)	Point input/abs	Temp (°C)	Cond (umhos)	DO (mg/l)	BOD (mg/l)	ON (µg/l)	NH ₄ (µg/l)	NO ₃ + NO ₂ (µg/l)	COD (mg/l)	Alk (mg/l)	pH
Wazirabad	22	– 18.36	19	348.143	8.47	3.29	1870	250	767.14	13.57	193.36	8.4
Najafgarh drain	22.44	25.39	19.43	973.43	0	61.43	22,450	2670	1271.42	111.14	198.62	7.4
Magazine road drain	23.58	0.21	19.43	982.85	0	259.14	36.05	2480	1868.57	274.86	197.67	7.6
Sweeper colony drain	23.78	0.37	19.43	1035.14	0	66.14	36,945	2837.14	1982.86	75.43	206.21	7.4
Metcalf house drain	24.7	0.18	19.71	949.86	0	21.43	31,290	3025.7	2254.29	21.43	177.56	7.7
ISBT + Mori gate drain	26.02	0.19	19.43	1017	0	73.43	29,914.3	275.57	2305.71	91	209.02	6.8
Tonga stand drain	27.38	0.53	19.43	1010.14	0	120.14	32,862.9	2194.29	2267.14	126.51	199.61	7.4
Shastri park drain	28.4	0.09	19.43	931.14	0	196.42	26,033.9	1975.5	2172.9	244.29	158.16	7.6
Civil mill drain	29.05	0.66	19.14	884.57	0	254.57	3335.2	1415.71	2241.43	18.85	161.33	7
Kailash Nagar drain	29.16	0.89	19.14	929.43	0	86	28,177.1	2267.14	2000.8	110.86	169.97	6.4
Delhi gate drain	31.26	0.60	19.14	929.43	0	189	58,042.9	2351.43	2125.17	238.43	164.53	7.4
Sen nursing home drain	32.64	1.10	18.86	901.29	0	212.43	37,044.3	2207.14	2107.14	264.57	165.83	6.74
Drain 14	33.64	1.80	18.86	920.29	0	85.43	21,091.4	1620	1634.29	130.86	192.61	7
Barapulla drain	37.49	0.89	19.1	718.43	0	290.57	21,037.1	1305.7	1701.43	290.86	157.24	7.7
Maharani Bagh drain + Hindon cut	40.49	10.98	19.77	840.39	0.96	43	18,108.6	3554.29	5520	79.42	186.47	7.7
Agra canal	41.5	– 37.6	18.81	917.57	0.16	12.57	11,097.1	9340	126.71	43	210.47	7
For validation												
Point sources/ abstraction	Distance (km)	Point input/abs	Temp (°C)	Cond (umhos)	DO (mg/l)	BOD (mg/l)	ON (µg/l)	NH ₄ (µg/l)	NO ₃ + NO ₂ (µg/l)	COD (mg/l)	Alk (mg/l)	pH
Wazirabad	22.00	–20.75	19.75	351.5	8.78	2.5	1230	315	972.5	11	133.57	7
Najafgarh drain	22.44	23.76	18.75	1569	0	42.5	14,025	1592.5	1495	84.5	331.25	7
Magazine road drain	23.58	0.06	18.75	1494.5	0	187.25	16,860	3420	3377.5	311	334.6	7
Sweeper colony drain	23.78	0.47	18.75	1671.5	0	145	30,730	4152.5	3755	115.75	333.75	7
Metcalf house drain	24.70	0.09	19.00	1583.25	0	11	42,120	3385	4445	40.5	332.5	7

Table 4 (continued)

For calibration												
ISBT+Mori gate drain	26.02	0.10	18.75	1457.25	0	48.75	50,765	3692.5	4265	50.75	333.75	7
Tonga stand drain	27.38	0.28	18.75	1445.5	0	90	52,530	3027.5	5200	117	333.75	7
Shastri park drain	28.40	0.09	18.75	1466.75	0	105	5797.5	5792.5	4792.4	125	332.5	7
Civil mill drain	29.05	0.39	18.50	1568	0	221.5	51,740	4432.5	4727.5	281	332.5	7
Kailash Nagar drain	29.16	0.48	18.50	1560.5	9	11	39,270	3305	3432.5	36.5	333.25	7
Delhi gate drain	31.26	0.48	18.50	1361.5	0	186.25	35,500	3710	3907.5	99	333.75	7
Sen nursing home drain	32.64	0.89	18.25	1596.25	0	160.6	34,000	2887.5	4670	161	334.75	7
Drain 14	33.64	1.91	18.25	1497.25	0	57.5	34,050	2990	3035	190	333.25	7
Barapulla drain	37.49	0.67	18.47	1496.75	0	161	25,210	2397.5	2927.5	60.75	333.75	7
Maharani Bagh drain + Hindon cut	40.49	10.16	19.05	1549.25	1.3	39	18,850	3837.5	3520	51.75	331.25	7
Agra canal	41.50	-36.01	18.35	1547.75	0.08	16	7257.5	13,665	922.5	79.25	344.125	7

case 1 and case 2, 50 cumecs and 60 cumecs flows have been augmented. The model was run for the average values of pollutant load for February 2019. Table 6 shows the pollutant load added to the river segment to develop the scenarios. The upstream water quality parameters used to develop the scenarios are shown in Table 7. Abstraction for Wazirabad has been taken as 1100 MLD (Yamuna et al. 2020), and upstream flow has been taken as the average decadal flow of 21.7 cumecs. In the present study, except for Fig. 9, COD was kept at 100 mg/l. NGT (National Green Tribunal) of India has directed to keep 10 mg/l of BOD and 50 mg/l of COD in the effluent wastewater. Hence, in Fig. 9, outfalling drains containing 50 mg/l of COD and 10 mg/l of BOD concentration have been assumed.

Results and discussion

Calibration and validation of the model

The water quality parameters used for calibration and validation are shown in Tables 3 and 4. Figure 4 shows the predicted and observed values for calibration and validation. Table 5 shows the RMSEV for DO, BOD, COD, and ammonia nitrogen (NH₄-N). The predicted and observed values agreed well, except for some deals. Variations in predicted and observed values might be due to the use of monthly average data. At M1, DO concentration was more than 8 mg/l for both datasets. However, after adding pollutant load from the Najafgarh drain (drain 1), the BOD level and DO were changed to high and 0, respectively. The scarcity of fresh water and high oxygen-demanding wastewater with elevated BOD and COD caused the abrupt degradation of DO at this point (Paliwal and Sharma 2007).

The RMSE values of dissolved oxygen for the predicted and observed values for calibration and validation are 0.021 and 0.0206, respectively. It emphasizes that the model is quite applicable to simulate the dissolved oxygen for this river reach. For BOD, the predicted and observed values in some stations are different. The BOD upstream was 1.81 mg/l, under the required standard, i.e., < 3 mg/l. However, the Najafgarh drain (Table 4) contributed 61.43 mg/l of BOD load with 25.39 cumecs wastewater. This drain contributes approximately 58% of

Table 5 RMSE values for calibration and validation

RMSEV	Calibration	Validation
DO	0.02	0.02
BOD	0.43	0.36
COD	0.38	0.41
NH ₄ -N	0.22	0.24

Table 6 Waste load input/ abstraction for February 2019 (DPCC)

Name	Distance (km)	Flow (cumecs)	Temp (°C)	DO (mg/l)	BOD (mg/l)	COD (mg/l)
A1	22	12.71	19	6.77	1.42	21.32
D1	22.44	24.91	19.43	0	62	152
D2	23.58	0.22	19.43	0	210	520
D3	23.78	0.43	19.43	0	30	92
D4	23.98	0.15	19.71	0	35	80
D5	24.7	0	0	0	0	0
D6	26.02	0.16	19.42	0	62	192
D7	27.38	0.45	19.43	0	60	172
D8	28.40	0.09	19.14	0	54	208
D9	29.05	0.58	19.14	0	130	400
D10	29.16	0.79	19.14	0	93	320
D11	31.26	0.58	18.86	0	52	152
D12	32.64	1.07	18.86	0	180	450
D13	33.64	1.81	19.10	0.96	40	128
D14	37.49	0.81	19.77	0.16	92	272
D15+T1	40.49	10.69	18.81	0	78	196
A2	41.50	37.58	18.81	0	88	336

the total BOD load (Paliwal et al. 2007). Hence, Fig. 4 shows a sudden increment of BOD for both calibration and validation. Discrepancies were between the observed and predicted values at some points, possibly due to the sampling and monitoring system, as India has limited financial resources for frequent monitoring like Nepal (Raj Kannel et al. 2007), and improvement is required to the monitoring system. Additionally, observed values were high as diffused sources due to bathing, washing clothes, cattle wading, and religious activities (Paliwal et al. 2007), which are not included in the present study. However, the RMSE values of BOD for calibration and validation are 0.4167 and 0.3568, respectively. The RMSE values for BOD, COD, and ammonia showed that the model is reasonably fit for this reach except for some variation.

Management approach

Four scenarios were attempted to arrive at management options. The model was run with the monthly average value of

February 2019 (Table 6 and Table 7), taking the decadal average flow of 21.7 cumecs for upstream (Sharma et al. 2017). As flow is varied throughout the year, the decadal average value has been considered for management options. Four scenarios were investigated, and the best combination for effective water quality management was established. Earlier studies stated (Paliwal and Sharma 2007; Bhargava 1985; Parmar and Keshari 2014b) that a combination of wastewater treatment, flow augmentation, and a diversion was necessary for this reach. This study attempted to maintain the water quality standards for this reach stated by the Central Pollution Control Department of India. As D1 (Najafgarh drain) is the highest contributing source, load modifications and diversification were considered for D1.

Bottom algal modification

In scenario 1, the model was run without any load input throughout the stretch and varying bottom algae conditions to check the assimilation capacity of the river bed with existing hydrological and meteorological conditions. For the development of management scenarios, meteorological and hydrological data have been taken for the study period. Cases were analyzed for 100%, 75%, 60%, and 50% bottom algal coverage. Table 8 shows the predicted values for different bottom algal concentrations. From Table 8, without external load, the river has a low assimilation capacity to reduce BOD from 2.4 mg/l to 0.73 mg/l for 100% bottom algae. Figure 5 shows the DO and BOD concentrations of the river for the reach with varying bottom algal conditions; when the bottom algae concentration was reduced to 50%, DO concentration increased by 38% after 22 km and 84% at the end of the reach compared to 100% bottom

Table 7 Headwater input for water quality management scenario

Headwater quality	Values
Temperature	12.80 °C
Conductivity	470.0 umhos
Dissolved oxygen	8.40 mg/l
BOD	2.40 mg/l
COD	12.00 mg/l
Alkalinity	143.00 mg/l
pH	7.9 s.u

Fig. 4 Observed and predicted values for calibration and validation of the model

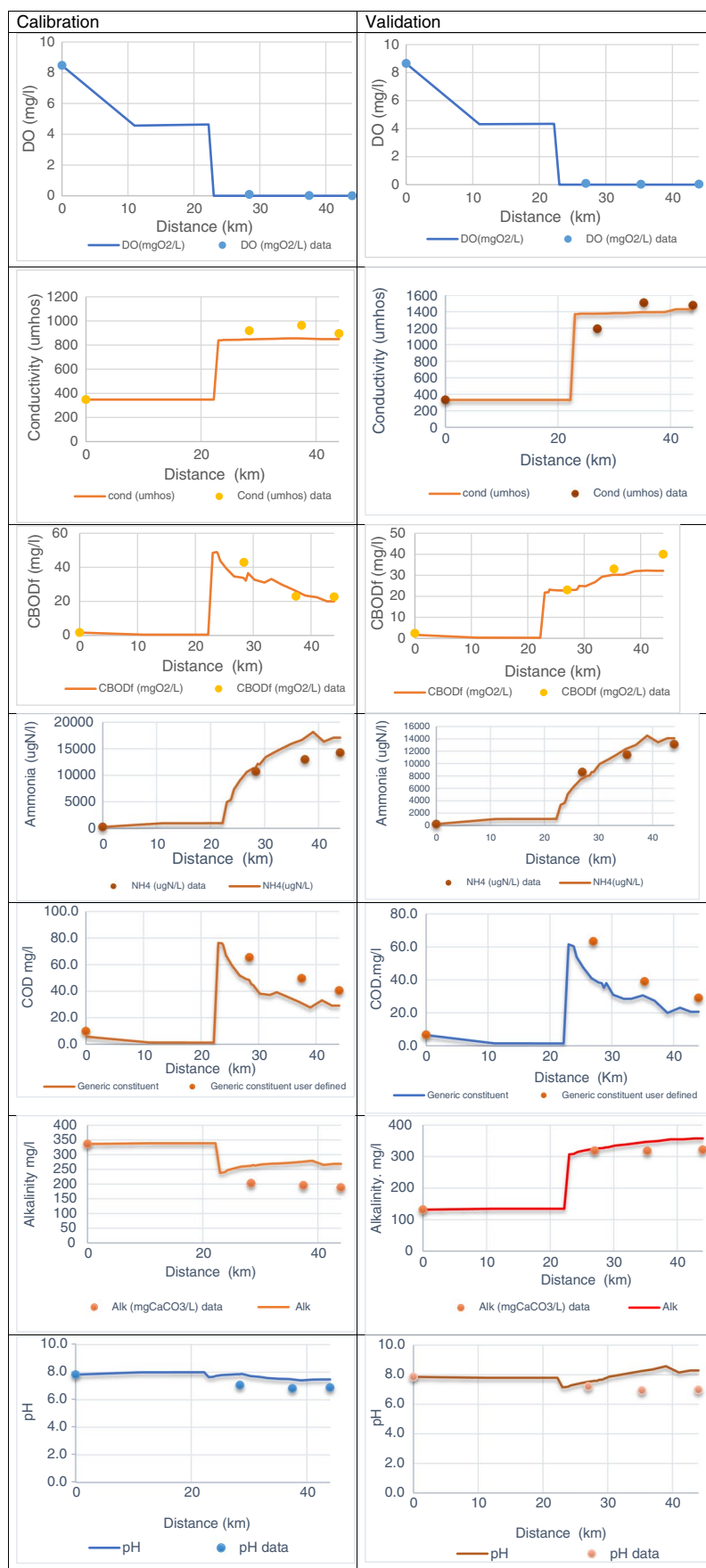


Table 8 Predicted values for scenario 1 changing the bottom algal concentrations (distance in km, BOD and DO in mg/l)

Distance	Case 1		Case 2		Case 3		Case 4	
	DO	BOD	DO	BOD	DO	BOD	DO	BOD
0	8.40	2.40	8.49	2.40	8.40	2.40	8.40	2.40
22.22	4.98	1.03	6.48	0.79	6.44	0.80	6.03	0.87
23.01	4.68	0.98	6.29	0.73	6.21	0.74	5.81	0.81
23.68	4.62	0.97	6.25	0.72	6.20	0.73	5.76	0.81
23.85	4.58	0.97	6.23	0.71	6.18	0.72	5.74	0.80
24.31	4.34	0.93	6.08	0.66	6.03	0.68	5.56	0.76
25.36	4.15	0.89	5.98	0.62	5.94	0.63	5.44	0.72
26.70	3.84	0.86	5.81	0.52	5.76	0.58	5.22	0.67
27.84	3.86	0.85	5.85	0.54	5.80	0.56	5.26	0.65
28.35	3.84	0.85	5.83	0.54	5.79	0.56	5.24	0.65
28.73	3.72	0.85	5.78	0.52	5.73	0.54	5.17	0.63
29.11	3.71	0.84	5.77	0.51	5.73	0.53	5.16	0.63
30.21	3.66	0.82	5.81	0.46	5.77	0.48	5.18	0.59
31.95	3.18	0.81	5.49	0.43	5.43	0.45	4.77	0.56
33.14	2.92	0.81	5.36	0.39	5.27	0.42	4.57	0.55
34.99	2.73	0.82	5.29	0.35	5.16	0.38	4.41	0.52
36.92	2.43	0.84	5.08	0.33	4.94	0.37	4.14	0.51
38.99	2.40	0.81	5.05	0.28	4.68	0.32	3.85	0.48
40.99	2.68	0.79	5.23	0.26	4.76	0.31	3.94	0.47
42.75	2.92	0.73	5.37	0.24	4.63	0.29	3.84	0.44
44.00	2.92	0.73	5.37	0.24	4.63	0.29	3.84	0.44

algal concentration. After 40% reduction of bottom algae, the stretch can maintain a DO concentration of more than 4 mg/l without external pollutant load. Therefore, the percent reduction of bottom algae can also be combined with the water quality management plan to get better results. Kori et al. (2013) also suggested maintaining 75% bottom algal coverage of the Karanja River, India, for getting the prescribed water quality.

Pollution load modification

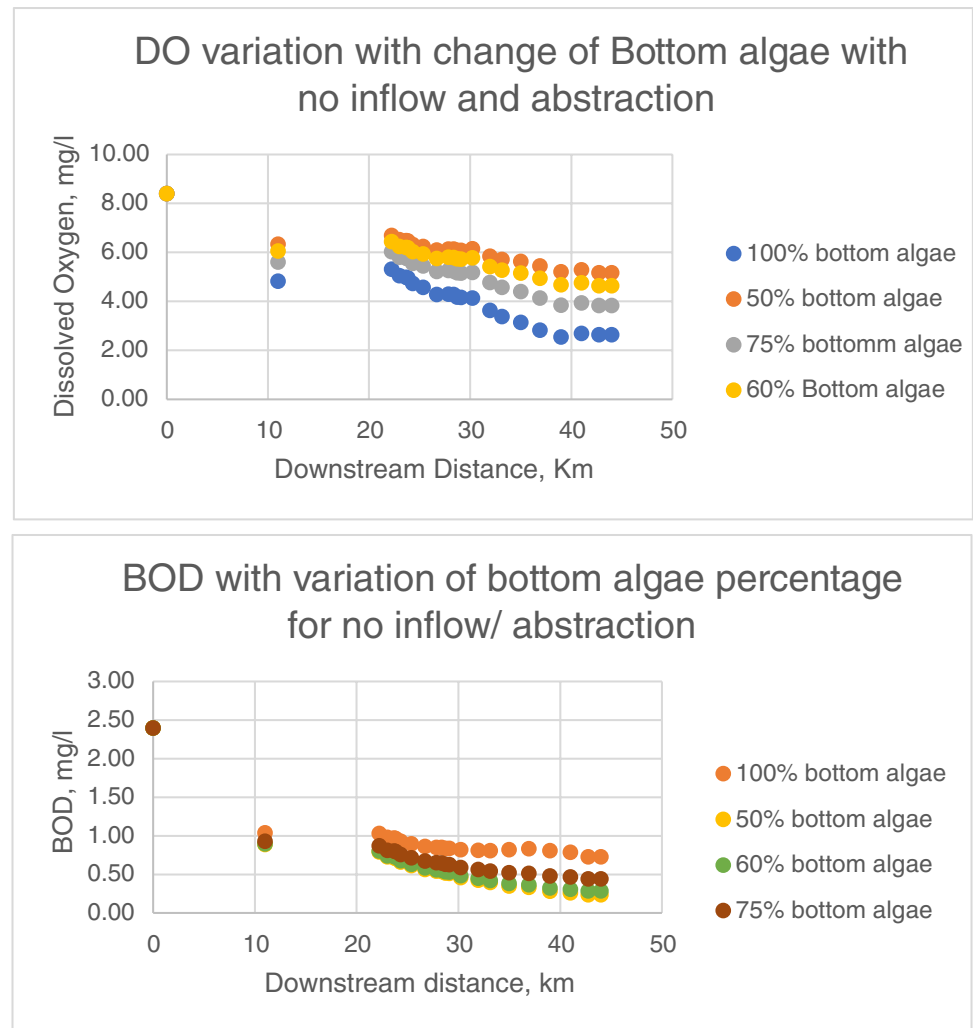
In scenario 2, 50% algal modification and 30 mg/l BOD and 4 mg/l DO in all contributing sources are taken, and modeling has been performed. In case 1, with 30 mg/l of BOD in all outfalling drains, abrupt changes occurred after 0.44 km of the Wazirabad barrage (Table 9) (CPCB 2006). Due to less freshwater availability and high wastewater contribution from D1, the BOD concentration became high, and the DO concentration changed to nil. In case 2, after load modification for D1, 30 km from upstream of the river reach maintained the required BOD value (< 3 mg/l). The reason may be due to the wastewater contribution was much less from other drains than D1 (Table 6). Assuming the BOD concentration of contributing wastewater was 5 mg/l from D1 and D14, all the reach could maintain BOD below 4 mg/l up to 40 km from upstream. DO concentration decreased after 30 km from upstream below the required concentration. Hence, some reaeration was required to increase the

DO concentrations. Thus, case 3 was generated with 100% removal of BOD from the Najafgarh drain and Barapullah drain, as these were the higher contributing sources. A minor improvement was found (Table 9 and Fig. 6). Therefore, multiple treatment options and diversion were also necessary for this highly polluted stretch (Paliwal et al. 2007).

Diversion of drains with load modification

As advanced treatment costs may be higher, the concentration of BOD and DO after increasing the flow at the upstream location and diverting D1 was observed (Fig. 7). The Najafgarh drain, which contributes a considerable amount of wastewater, can be diverted to other sites. This diversion was suggested in 1979 (Bhargava 1985), and it was proposed to shift drain 1 (Najafgarh drain) to the Agra canal. After Wazirabad, the water flow is very little to the river, and wastewater from D1 contributes to the flow of the river. Therefore, in scenario 3, instead of diversion D1, modification of the load for D1 and diversion of D13, D14, and D15 were studied. Case 1, case 2, and case 3 were assessed by taking BOD as 0 mg/l, 15 mg/l, and 5 mg/l, respectively, for the D1 and D12 (after diverting D (13–15), and leaving all drains with 30 mg/l BOD and 4 mg/l DO). Figure 7 and Table 10 show the predicted values for BOD and DO. For case 1, DO concentration was not satisfied after 33 km from

Fig. 5 Scenario 1: BOD and DO profile with bottom algae modification



upstream. Therefore, local oxygenation is needed at 33 km, consistent with earlier studies (Paliwal et al. 2007).

Diversion with flow augmentation

Najafgarh drain (D1), as a primary contributor to wastewater, can be diverted to the Agra canal or other places, as suggested by earlier studies (Bhargava 1985). After the

diversion of D1, the flow of water will be very deficient, so flow augmentation must be required upstream to increase the self-assimilation capacity of the reach. Scenario 4 studied flow augmentation with diversion. Figure 8 shows the variation of increasing flow at the upstream location Palla. When the upstream flow is kept at 50 cumecs at Palla and diverting D1, the DO and BOD are satisfied up to 30 km upstream. Therefore, the last three drains were diverted, and

Fig. 6 DO and BOD profile for the different cases of scenario 2

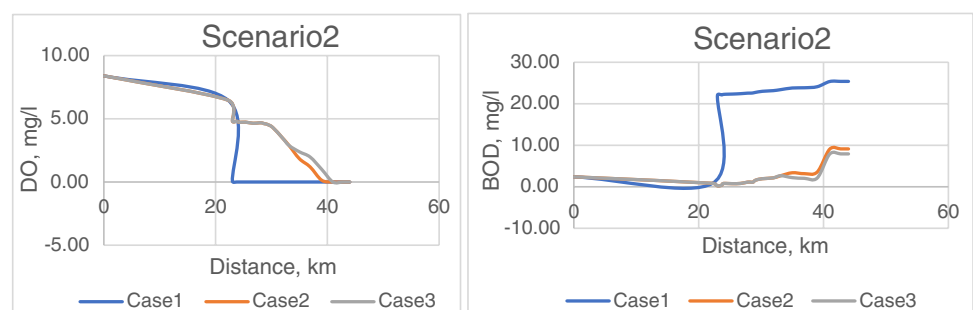


Table 9 Predicted values for scenario 2 (50% bottom algal, BOD 30 mg/l, DO 4 mg/l for all reaches in case 1, in case 2: BOD 0 at D1 and rest same, case 3 BOD 0 mg/l at D1 and D14, and the rest were same) (distance in km, BOD and DO in mg/l)

Distance	Case 1		Case 2		Case 3	
	DO	BOD	DO	BOD	DO	BOD
0	8.40	2.40	8.40	2.40	8.40	2.40
22.22	6.47	0.84	6.47	0.84	6.47	0.84
23.01	0.06	22.07	4.81	0.61	4.81	0.20
23.68	1E-06	22.13	4.80	0.39	4.80	0.39
23.85	1E-06	22.23	4.79	0.75	4.80	0.75
24.31	1E-06	22.30	4.72	0.81	4.72	0.81
25.36	1E-06	22.34	4.74	0.75	4.74	0.75
26.70	1E-06	22.43	4.65	0.81	4.65	0.81
27.84	1E-06	22.56	4.68	1.12	4.68	1.12
28.35	1E-06	22.57	4.67	1.11	4.67	1.12
28.73	1E-06	22.62	4.61	1.13	4.61	1.13
29.11	1E-06	22.75	4.58	1.58	4.58	1.58
30.21	1E-06	23.01	4.32	1.95	4.32	1.95
31.95	1E-06	23.16	3.49	2.14	3.49	2.14
33.14	1E-06	23.41	2.89	2.67	2.89	2.67
34.99	1E-06	23.80	1.84	3.38	2.39	2.20
36.92	1E-06	23.86	1.19	3.16	1.98	2.04
38.99	1E-06	24.11	0.14	3.56	0.99	2.11
40.995	1E-06	25.37	1E-06	9.08	1E-06	7.89
42.75	1E-06	25.39	1E-06	9.13	1E-06	7.92
44.00	1E-06	25.39	1E-06	9.13	1E-06	7.92

the water quality improved. However, the contribution from the tributary Hindon cut caused the DO concentration to be reduced in the last segment. Therefore, higher treatment may be required for T1. As stated in previous studies, it is possible to change the situation by increasing the upstream flow (Parmar and Keshari 2014b). Furthermore, obtaining the designated standard requires more treatment. A recent order from NGT (Deshpande, 2019) suggested effluent standards for BOD as 10 mg/l and COD as 50 mg/l. The model is run with 50 cumecs flow upstream by keeping these values, shown in Fig. 9. However, it was observed that this river segment required advanced treatment. Moreover, after 30 km upstream, aeration was also needed. Although many studies and management plans have been taken to rejuvenate this

reach, the river reach could not attain the desired standard for the complexity of load distribution and proper monitoring systems. This study did not include oxygen demand for the nitrogenous matter, and diffused sources were also not calculated. Hence, the predicted results may differ from the observed results.

Conclusion

A receiving water quality model, QUAL2Kw, was used to calibrate and validate the secondary data for the Yamuna River Delhi segment. The field and predicted data showed a good relationship, except for some variation. The model

Fig. 7 DO and BOD profiles for the different cases of scenario

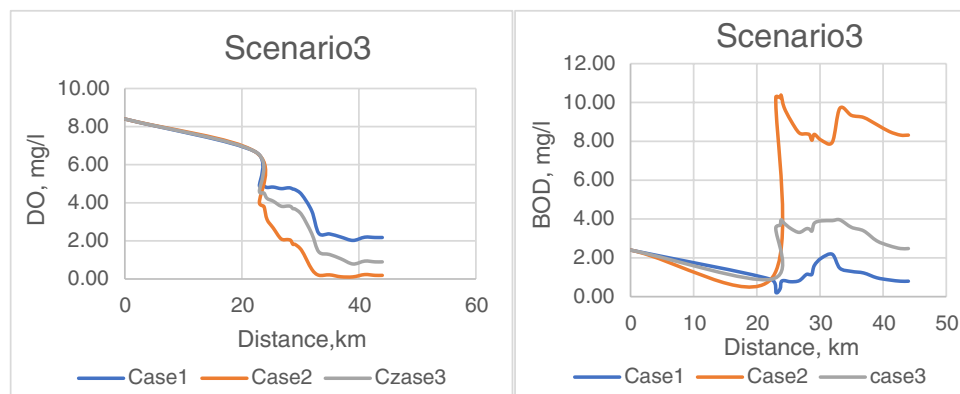


Table 10 Predicted values for scenario 3 (distance in km, BOD and DO in mg/l) (modification of load for D1 and D12; diversion of D(13–15) at D12; case 1: 0 mg/l BOD for D1 and D12; case 2: 15 mg/l BOD for D1 and D12; case 3: 5 mg/l BOD for D1 and D12)

Distance	Case 1		Case 2		Case 3	
	DO	BOD	DO	BOD	DO	BOD
0	8.4	2.4	8.4	2.4	8.4	2.4
22.22	6.69	0.89	6.69	0.89	6.69	0.89
23.01	4.88	0.22	3.98	10.28	4.57	3.56
23.68	4.88	0.40	3.83	10.24	4.52	3.68
23.85	4.87	0.76	3.74	10.37	4.49	3.96
24.31	4.80	0.83	3.12	9.75	4.23	3.79
25.36	4.83	0.77	2.68	9.08	4.08	3.51
26.70	4.74	0.82	2.11	8.43	3.82	3.31
27.84	4.77	1.14	2.07	8.39	3.83	3.50
28.35	4.77	1.13	2.03	8.34	3.81	3.47
28.73	4.72	1.15	1.82	8.06	3.69	3.38
29.11	4.67	1.61	1.81	8.36	3.67	3.78
30.21	4.44	1.98	1.50	8.06	3.38	3.90
31.95	3.57	2.18	0.51	7.95	2.38	3.91
33.14	2.40	1.47	0.20	9.70	1.43	3.96
34.99	2.37	1.30	0.22	9.34	1.27	3.57
36.92	2.20	1.22	0.12	9.23	1.06	3.39
38.99	2.02	0.97	0.11	8.86	0.79	2.88
40.99	2.20	0.86	0.23	8.50	0.94	2.62
42.75	2.18	0.80	0.19	8.32	0.90	2.48
44.00	2.18	0.80	0.19	8.32	0.90	2.48

Fig. 8 Scenario 4 Flow augmentation with diversion

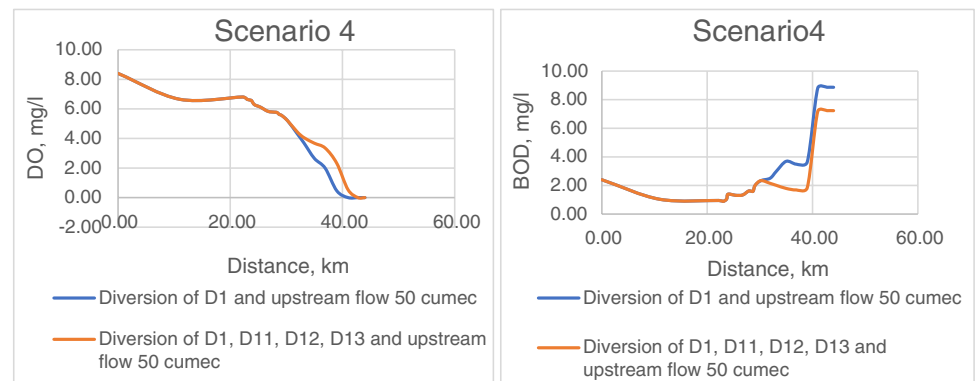
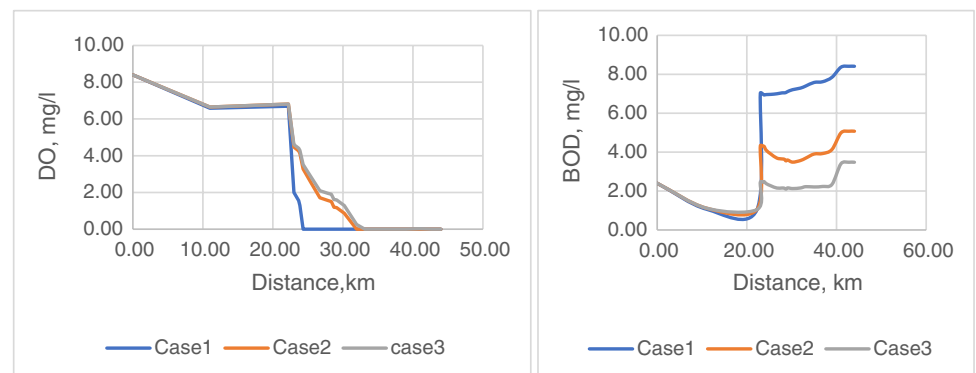


Fig. 9 Flow augmentation with a maximum BOD of 10 mg/l



has been applied for different water quality management approaches during low flow conditions to maintain the DO concentration above 4 mg/l and BOD concentration below 3 mg/l. Scenarios have been established by bottom algal modification, load modification, load modification and diversion, and diversion with flow augmentation. It has been observed that without load input and 40% reduction of bottom algal could attain the desired conditions, and modification of bottom algal condition can be combined with water quality management approaches. Scenario 2 represented that load modification of D1 and D14 (5 mg/l) could maintain BOD level below 3 mg/l throughout the reach. Local oxygenation was required at 30 km downstream to keep the DO concentration above 4 mg/l. Scenario 4 shows that flow augmentation with 60 cumecs at the upstream and diversion of D1 may satisfy the required limit for BOD around 40 km and DO up to 37 km downstream. Hence, the present study shows that this highly polluted reach needs a combination of treatment, including reduction of bottom algae, diversion of major contributing pollutant sources, flow augmentation at the upstream, and local oxygenation in some particular reaches to manage the water quality of the urban reach of Yamuna. The different scenarios and cases studied in this paper can be used to develop a real-time water quality management strategy for various urban reaches of the river. However, diffused sources and the oxygen demand for the nitrogenous matter can also be incorporated into further studies,

Model

The model has been downloaded from Models & Tools for TMDLs—Washington State Department of Ecology.

Acknowledgements The authors are thankful to CWC, DPCC, and IF&C for providing data. The authors are grateful to the reviewers for giving their valuable time and suggestions to improve the quality of the study.

Data availability The dataset analyzed during the present study are available in the CPCB (www.cpcb.nic.in/water/water quality of Yamuna River) DPCC domain (Delhi Pollution Control Committee (delhigovt.nic.in/water/analysis report). Some data were collected by personal communication from DPCC, CWC classified data, and IF&C, which may not be available.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

- Abbas MA, Iqbal M, Tauqeer HM, Turan V, Farhad M (2022) Chapter 16 - Microcontaminants in wastewater. In M. Z. Hashmi, S. Wang, & Z. Ahmed (Eds.), *Environmental Micropollutants* (pp. 315–329). Elsevier. <https://doi.org/10.1016/B978-0-323-90555-8.00018-0>
- Bhargava DS (1985) Water quality variations and control technology of Yamuna River. *Environmental Pollution. Series A, Ecological and Biological*, 37(4), 355–376. [https://doi.org/10.1016/0143-1471\(85\)90124-2](https://doi.org/10.1016/0143-1471(85)90124-2)
- Chapra (1997) *Surface water modelling.*, McGraw Hill Publications
- Cho JH, Lee JH (2019) Automatic calibration and selection of optimal performance criterion of a water quality model for a river controlled by total maximum daily load (TMDL). *Water Sci Technol* 79(12):2260–2270. <https://doi.org/10.2166/wst.2019.222>
- Cox B (2003) NA review of currently available in-stream water-quality models and their applicability for simulating dissolved oxygen in lowland rivers Title. *Sci Total Environ*, 314–316(October 2006), 335–377. [https://doi.org/10.1016/S0048-9697\(03\)00063-9](https://doi.org/10.1016/S0048-9697(03)00063-9)
- CPCB (2006) Water quality status of Yamuna River, assessment and development of river basin. Basin Seri(ADSORBS/41/2006-07). www.cpcb.nic.in
- CPCB (2008) Status of water quality in India- 2007. Monitoring of Indian national aquatic resources, series: MINARS/ 29 /2008–2009, July 2008, 1–247. http://www.cpcb.nic.in/WQ_Status_Report2012.pdf
- CPCB (2009) Annual Report 15(1):1. <https://doi.org/10.1016/j.parkrelidis.2008.12.001>
- Cristea NC, Burges SJ (2010) An assessment of the current and future thermal regimes of three streams located in the Wenatchee River basin, Washington State: some implications for regional river basin systems. *Clim Chang* 102(3):493–520. <https://doi.org/10.1007/s10584-009-9700-5>
- Darajati Setiawan A, Widyastuti M, Pramono Hadi M (2018) Water quality modeling for pollutant carrying capacity assessment using Qual2Kw in Bedog river. *Indonesian J Geogr*, 50(1), 49–56. <https://doi.org/10.22146/ijg.16429>
- Deksisia T, Meirlaen J, Ashton PJ, Vanrolleghem PA (2004) Simplifying dynamic river water quality modelling: a case study of inorganic nitrogen dynamics in the Crocodile River (South Africa). *Water Air Soil Pollut* 155(1):303–320. <https://doi.org/10.1023/B:WATE.0000026548.20608.a0>
- Elshorbagy A, Teegavarapu RSV, Ormsbee L (2005) Total maximum daily load (TMDL) approach to surface water quality management: concepts, issues, and applications. <https://doi.org/10.1139/L04-107>
- Ghosh S, Mujumdar PP (2010) Fuzzy waste load allocation model: a multiobjective approach. *J Hydroinf* 12(1):83–96. <https://doi.org/10.2166/hydro.2010.028>
- Gregory Plettier, Steven Chapra, H. tao. (2006). QUAL2Kw – a framework for modeling water quality in streams and rivers using a genetic algorithm for calibration *Environmental Modelling and Software*, 21(3), 419–425. <https://doi.org/10.1016/j.envsoft.2005.07.002>
- Hobson AJ, Neilson BT, von Stackelberg N, Shupryt M, Ostermiller J, Pelletier G, Chapra SC (2015) Development of a minimalistic data collection strategy for QUAL2Kw. *J Water Resour Plann Manag*, 141(8). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000488](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000488)
- Jaiswal M, Hussain J, Gupta SK, Nasr M, Nema AK (2019) Comprehensive evaluation of water quality status for entire stretch of Yamuna River, India. *Environ Monit Assess*, 191(4). <https://doi.org/10.1007/s10661-019-7312-8>
- Jim Carroll, Sarah O'Neal, and S. G (n.d.) Wenatchee River Basin dissolved oxygen, pH, and phosphorus total maximum daily load study. In Washington State Department of Ecology, April 2006 Publication No. 06–03–018 (Issue 06)
- Kazmi AA, Hansen IS (1997) Numerical models in water quality management: a case study for the Yamuna River (India). *Water Sci Technol* 36(5):193–200. [https://doi.org/10.1016/S0273-1223\(97\)00474-5](https://doi.org/10.1016/S0273-1223(97)00474-5)

- Khalil M, Iqbal M, Turan V, Tauqueer HM, Farhad M, Ahmed A, Yasin S (2022) Chapter 11 - Household chemicals and their impact. In M. Z. Hashmi, S. Wang, & Z. Ahmed (Eds.), *Environmental Micropollutants* (pp. 201–232). Elsevier. <https://doi.org/10.1016/B978-0-323-90555-8.00022-2>
- Kori BB, Shashidhar T, Mise S (2013) Application of automated QUAL2Kw for water quality modeling in the River Karanja. *India* 2(2):193–203
- Lestari H, Haribowo R, Yuliani E (2019) Determination of pollution load capacity using QUAL2Kw program on The Musi River Palembang. *II*(02), 105–116
- McIntyre NRHSW (2004) A tool for risk-based management of surface water quality. *Environ Modell Softw* 19(12):1131–1140. <https://doi.org/10.1016/j.envsoft.2003.12.003>
- Neilson BT, Hobson AJ, VonStackelberg N, Shupryt M, Ostermiller J (2013) Using Qual2K modeling to support nutrient criteria development and wasteload analyses in Utah. 1–49
- Nikoo MR, Beiglou PHB, Mahjouri N (2016) Optimizing multiple-pollutant waste load allocation in rivers: an interval parameter game theoretic model. *Water Resour Manage* 30(12):4201–4220. <https://doi.org/10.1007/s11269-016-1415-6>
- Nitin Shankar Deshpande (2019) Before the National Green Tribunal Principal Bench. New Delhi 04:1–14
- Oliveira B, Bola J, Quinteiro P, Nadais H, Arroja L (2012) Application of Qual2Kw model as a tool for water quality management: Cértima River as a case study. *Environ Monit Assess* 184(10):6197–6210. <https://doi.org/10.1007/s10661-011-2413-z>
- Paliwal R, Sharma P (2007) Application of QUAL2E for the river Yamuna: to assess the impact of point loads and to recommend measures to improve water quality of the river. *Environment* 2702
- Paliwal R, Sharma P, Kansal A (2007) Water quality modeling of the river Yamuna (India) using QUAL2E-UNCAS. *J Environ Manage* 83(2):131–144. <https://doi.org/10.1016/j.jenvman.2006.02.003>
- Parmar DL, Keshari AK (2014a) Wasteload allocation using wastewater treatment and flow augmentation. *Environ Model Assess* 19(1):35–44. <https://doi.org/10.1007/s10666-013-9378-y>
- Parmar DL, Keshari AK (2014b) Wasteload allocation using wastewater treatment and flow augmentation GIS-coupled numerical modeling for sustainable groundwater development: case study of Aynalem Well Field, Ethiopia View project wasteload allocation using wastewater treatment and flow augmentation. Article in *Environmental Modeling and Assessment*. <https://doi.org/10.1007/s10666-013-9378-y>
- Pelletier G, Chapra S (2008a) A modeling framework for simulating river and stream water quality. *Environmental Assessment Program Olympia, Washington* 98504–7710
- Pelletier G, Chapra S (2008b) A modeling framework for simulating river and stream water quality July 2008b Publication No. 08–03-xxx. <http://www.ecy.wa.gov/biblio/04030>
- Raj Kannel P, Lee S, Lee Y, Kanel S, Pelletier G (2007) Application of automated QUAL2Kw for water quality modeling and management in the Bagmati River, Nepal <https://doi.org/10.1016/j.ecolmodel.2006.12.033>
- Santos S, Vilar VJP, Alves P, Boaventura RAR, Botelho C (2013) Water quality in Minho/Miño River (Portugal/Spain). *Environ Monit Assess* 185(4):3269–3281. <https://doi.org/10.1007/s10661-012-2789-4>
- Sharma D (2013) Evaluation of river quality restoration plan and intervention analysis using water quality modeling with focus on the River Yamuna, Delhi (India). Ph.D. Thesis, TERI University, Delhi, India
- Sharma D, Kansal A, Pelletier Greg (1999) Water quality modeling for urban reach of Yamuna river, India (1999–2009), using QUAL2Kw. <https://doi.org/10.1007/s13201-015-0311-1>
- Sharma D, Kansal A, Pelletier G (2017) Water quality modeling for urban reach of Yamuna river, India (1999–2009), using QUAL2Kw. *Appl Water Sci* 7(3):1535–1559. <https://doi.org/10.1007/s13201-015-0311-1>
- Sharma MP, Singal SK, Patra S (2009) Water quality profile of Yamuna River, India. *Hydro Nepal: J Water Energy Environ* 3(3):19–24. <https://doi.org/10.3126/hn.v3i0.1914>
- Singh AP, Ghosh SK, Sharma P (2007) Water quality management of a stretch of river Yamuna: an interactive fuzzy multi-objective approach. *Water Resour Manage* 21(2):515–532. <https://doi.org/10.1007/s11269-006-9028-0>
- Tauqueer HM, Turan V, Farhad M, Iqbal M (2022a) Sustainable agriculture and plant production by virtue of biochar in the era of climate change. In M. Hasanuzzaman, G. J. Ahammed, & K. Nahar (Eds.), *Managing plant production under changing environment* (pp. 21–42). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-5059-8_2
- Tauqueer HM, Turan V, Iqbal M (2022b) Production of safer vegetables from heavy metals contaminated soils: the current situation, concerns associated with human health and novel management strategies. In J. A. Malik (Ed.), *Advances in bioremediation and phytoremediation for sustainable soil management: principles, monitoring and remediation* (pp. 301–312). Springer International Publishing. https://doi.org/10.1007/978-3-030-89984-4_19
- Turan V, Khan SA, Mahmood-ur-Rahman, Iqbal M, Ramzani PMA, Fatima M (2018) Promoting the productivity and quality of brinjal aligned with heavy metals immobilization in a wastewater irrigated heavy metal polluted soil with biochar and chitosan. *Ecotoxicol Environ Saf* 161:409–419. <https://doi.org/10.1016/j.ecoenv.2018.05.082>
- Turner DF, Pelletier GJ, Kasper B (2009) Dissolved oxygen and pH modeling of a periphyton dominated, nutrient enriched river. *J Environ Eng* 135(8):645–652. [https://doi.org/10.1061/\(ASCE\)0733-9372\(2009\)135:8\(645\)](https://doi.org/10.1061/(ASCE)0733-9372(2009)135:8(645))
- Yamuna THE, Monitoring P, Marg MAXM, Estate L (2020) Dated: 21.04.2020. 58, 1–32
- Zare Farjoudi S, Moridi A, Sarang A (2021) Multi-objective waste load allocation in river system under inflow uncertainty. *Int J Environ Sci Technol* 18(6):1549–1560. <https://doi.org/10.1007/s13762-020-02897-5>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

EMOTION-GUIDED CROSS-DOMAIN FAKE NEWS DETECTION USING ADVERSARIAL DOMAIN ADAPTATION

Arjun Choudhry*

Biometric Research Laboratory
Delhi Technological University
New Delhi, India
choudhry.arjun@gmail.com

Inder Khatri*

Biometric Research Laboratory
Delhi Technological University
New Delhi, India
inderkhatri999@gmail.com

Arkajyoti Chakraborty

Biometric Research Laboratory
Delhi Technological University
New Delhi, India
arkajyotichakraborty_2k19ep022@dtu.ac.in

Dinesh Kumar Vishwakarma

Biometric Research Laboratory
Delhi Technological University
New Delhi, India
dinesh@dtu.ac.in

Mukesh Prasad

School of Computer Science
University of Technology Sydney
Ultimo, Australia
mukesh.prasad@uts.edu.au

November 28, 2022

ABSTRACT

Recent works on fake news detection have shown the efficacy of using emotions as a feature or emotions-based features for improved performance. However, the impact of these emotion-guided features for fake news detection in cross-domain settings, where we face the problem of domain shift, is still largely unexplored. In this work, we evaluate the impact of emotion-guided features for cross-domain fake news detection, and further propose an emotion-guided, domain-adaptive approach using adversarial learning. We prove the efficacy of emotion-guided models in cross-domain settings for various combinations of source and target datasets from FakeNewsAMT, Celeb, Politifact and Gossipcop datasets.

Keywords Fake News Detection · Domain Adaptation · Emotion Classification · Adversarial Training · Cross-domain Analysis

1 Introduction

In recent years, our reliance on social media as a source of information has increased multi-fold, bringing along exponential increase in the spread of *fake news*. To counter this, researchers have proposed various approaches for fake news detection (Shu, Cui, Wang, Lee and Liu, 2019; Sheng, Cao, Zhang, Li, Wang and Zhu, 2022). However, models trained on one domain are often brittle and vulnerable to incorrect predictions for the samples of another domain (Saikh, De, Ekbal and Bhattacharyya, 2019; Pérez-Rosas, Kleinberg, Lefevre and Mihalcea, 2018). This is primarily due to the shift between the two domains, as depicted in Figure 1(1). To handle this, some domain-adaptive frameworks (Zhang, Wang, Chen, Zeng, Guo, Miao and Cui, 2020; Huang, Gao, Wang and Shu, 2021; Li, Lee, Kordzadeh, Faber, Fiddes, Chen and Shu, 2021) have been proposed which help align the source and target domains in the feature

*Equal Contribution

space to ameliorate domain shift across different problems. These frameworks guide the feature extractors to extract domain-invariant features by aligning the source and target domains in the feature space, thus generalizing well across domains. However, due to the absence of labels in the target-domain data, the adaptation is often prone to negative transfer, which can disturb the class-wise distribution and affect the discriminability of the final model, as shown in Figure 1(2).

Some recent studies have observed a correlation between the veracity of a text and its emotions. There exists a prominent affiliation for certain emotions with fake news, and for other emotions with real news (Vosoughi, Roy and Aral, 2018), as illustrated in Figure 1(3). Further, some works have successfully utilized emotions as features, or emotion-guided features to aid in fake news detection (Guo, Cao, Zhang, Shu and Yu, 2019; Zhang, Cao, Li, Sheng, Zhong and Shu, 2021; Choudhry, Khatri and Jain, 2022). However, we observe that these works only consider the in-domain setting for evaluation, without analyzing the robustness of these frameworks to domain shift in cross-domain settings. This is another important direction that needs to be explored.

In this paper, we study the efficacy of emotion-aided models in capturing better generalizable features for cross-domain fake news detection. Table 1 shows the improvements observed in various cross-domain settings when our emotion-guided models were evaluated in cross-domain settings. We observe that emotion-guided frameworks show improved performance in cross-domain settings, as compared to their baseline models without the said emotion-aided features, thus underscoring the generalized feature extraction in emotion-aided models. We further propose an emotion-guided unsupervised domain adaptation framework, which utilizes emotion labels in a multi-task adversarial setting for better feature alignment across domains. The emotion labels for emotion classification, trained parallel to the fake news detection branch in the multi-task learning setup, help provide additional supervision for improved alignment during domain adaptation, mitigating the issue of incorrect alignment of domains. This is illustrated in Figure 1(4)). This leads to better discriminability. We experimentally prove the efficacy of our approach across a variety of datasets

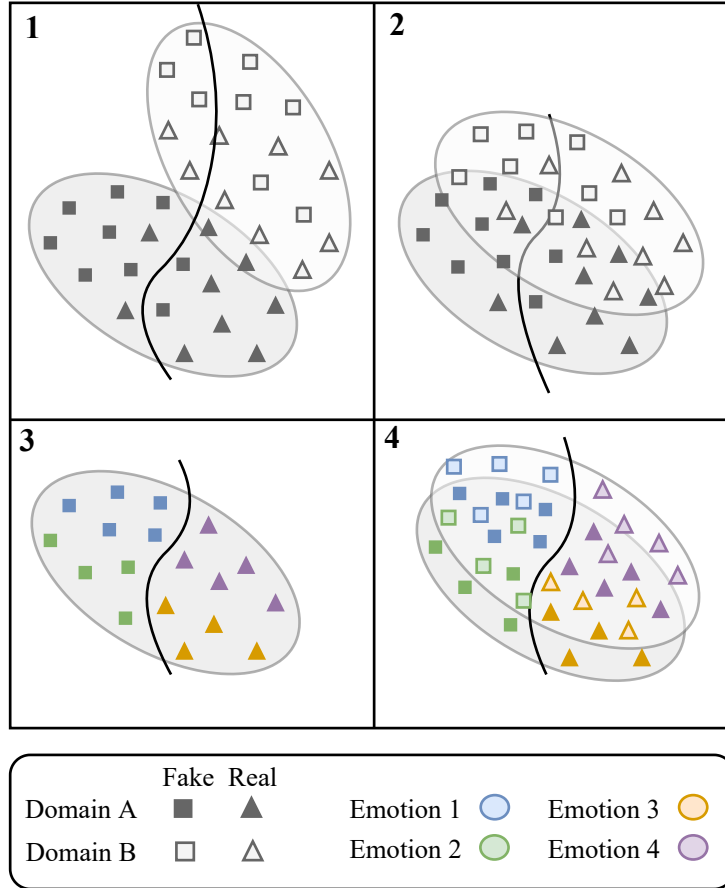


Figure 1: (1) Cross-domain texts not aligned. (2) Domain adaptation leads to some alignment. (3) Emotion-guided classification in one domain. (4) Emotion-guided domain adaptation leads to improved alignment of the two domains.

in cross-domain settings for various combinations of single-task or multi-task, domain-adaptive or non-adaptive, emotion-guided or unguided settings on the accuracy of the models.

Our contributions can be summarized as follows:

- We suggest the use of emotion classification as an auxiliary task for improved fake news detection in cross-domain settings, indicating the applicability of emotion-guided features across domains.
- We compare how Ekman’s and Plutchik’s base emotion classes individually affect the performance of our multi-task domain-adaptive framework, and if there are meaningful differences between them.
- We propose an emotion-guided domain-adaptive framework for fake news detection across domains. We show that domain-adaptive fake news detection models better align the two domains with the help of supervised learning using emotion-aided features.
- We evaluate our approach on a variety of source and target combinations from four datasets. Our results prove the efficacy of our approach.

2 Related Works

Several studies over the last few years have explored the correlation of fake news detection with emotions. K, P and L (2020) *emotionized* text representations using explicit emotion intensity lexicons. Guo et al. (2019) utilized the discrepancies between publisher’s emotion and the thread’s comments’ emotions to detect fake news. However, most of these methods relied upon some additional inputs during evaluation. Choudhry et al. (2022) proposed an emotion-aided multi-task learning approach, where emotion classification was the auxiliary task implicitly aligning fake news features according to emotion labels.

Inspired by Ganin, Ustinova, Ajakan, Germain, Larochelle, Laviolette, Marchand and Lempitsky (2015), Zhang et al. (2020) proposed the first fake news detection work to tackle domain shifts between different datasets. They proposed a multi-modal framework with a Gradient Reversal Layer (GRL) to learn domain-invariant features across different domains and used a joint fake news detector on the extracted features. Huang et al. (2021) proposed a robust and generalized fake news detection framework adaptable to a new target domain using adversarial training to make the model robust to outliers and Maximum Mean Difference (MMD)-based loss to align the features of source and target. Li et al. (2021) extended the problem by treating it as a multi-source domain adaptation task, using the labeled samples from multiple source domains to improve the performance on unlabeled target domains. They also utilized weak labels for weak supervision on target samples to improve performance.

However, no previous work has aligned features between different domains using emotion-guided features and domain adaptation using adversarial training. We show that applying both of these approaches leads to improved performance due to better alignment of inter-domain features.

3 Proposed Methodology

3.1 Datasets, Emotion Annotation & Preprocessing

We use the FakeNewsAMT (Pérez-Rosas et al., 2018), Celeb (Pérez-Rosas et al., 2018), Politifact¹, and Gossipcop² datasets for cross-domain fake news detection. FakeNewsAMT is a multi-domain dataset containing samples from technology, education, business, sports, politics, and entertainment domains. The Celeb dataset has been derived from the web, and contains news about celebrities. Politifact is a web-scraped dataset containing political news, while Gossipcop contains news extracted from the web, manually annotated via crowd-sourcing and by experts.

We use the Unison model (Colnerič and Demšar, 2020) to annotate all datasets with the core emotions from Ekman’s (Ekman, 1992) (6 emotions: *Joy, Surprise, Anger, Sadness, Disgust, Fear*) and Plutchik’s (Plutchik, 1982) (8 emotions: *Joy, Surprise, Trust, Anger, Anticipation, Sadness, Disgust, Fear*) emotion theories. During preprocessing, we convert text to lowercase, remove punctuation, and de-contract verb forms (eg. “I’d” to “I would”).

3.2 Multi-task Learning

We use multi-task learning (MTL) to incorporate emotion classification as an auxiliary task to our fake news detection branch. Multi-task learning enables a model to learn the shared features between two or more correlated tasks for

¹<https://www.politifact.com>

²<https://www.gossipcop.com>

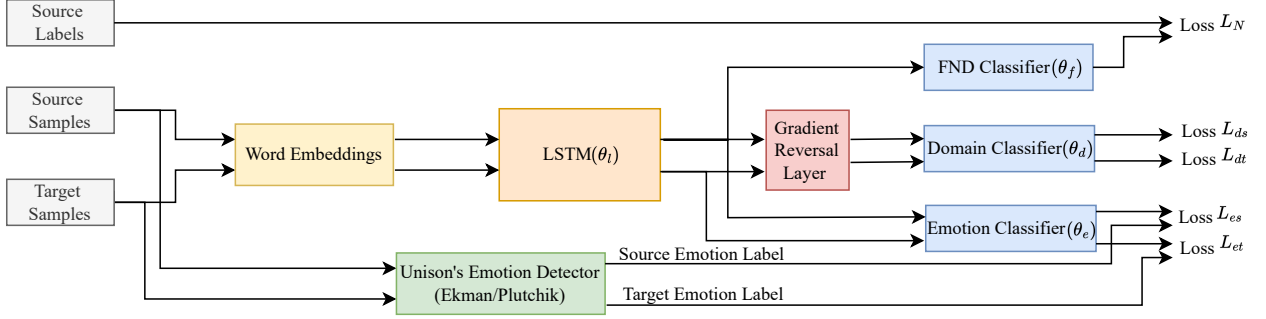


Figure 2: Pictorial depiction of our emotion-guided domain-adaptive approach for cross-domain fake news detection.

improved feature extraction and performance. We use Ekman’s or Plutchik’s emotions labels for emotion classification branch in our MTL models to see which performs better, and compare the performance with the corresponding single-task (STL) models in domain-adaptive and non-adaptive settings.

3.3 Emotion-guided Domain-adaptive Framework

We propose the cumulative use of domain adaptation and emotion-guided feature extraction for cross-domain fake news detection. Our approach aims to improve the feature alignment between different domains using adversarial domain adaptation, by leveraging the correlation between the emotion and the veracity of a text (as shown in Figure 1(4)). Figure 2 shows our proposed framework. We use an LSTM-based (Hochreiter and Schmidhuber, 1997) feature extractor, which is trained using the accumulated loss from fake news classifier, emotion classifier and the discriminator (aids in learning domain-invariant features). LSTM can be replaced with better feature extractors. We used it specifically for easier comparison to non-adapted emotion-guided and non-adapted single-task models. The domain classifier acts as the discriminator. In our proposed framework, we do not use the truth labels for the target dataset for domain adaptation. However, we utilize the target domain emotion labels in our approach to better align the two domains using the emotion labels for supervised learning. The fake news classification loss, emotion classification loss, adversarial loss, and total loss are defined as in Equations 1, 2, 3, and 4:

$$L_{FND} = \min_{\theta_l, \theta_f} \sum_{i=1}^{n_s} L_f^i \quad (1)$$

$$L_{emo} = \min_{\theta_l, \theta_e} \left(\sum_{i=1}^{n_s} L_{es}^i + \sum_{j=1}^{n_t} L_{et}^j \right) \quad (2)$$

$$L_{adv} = \min_{\theta_d} \left(\max_{\theta_l} \left(\sum_{i=1}^{n_s} L_{ds}^i + \sum_{j=1}^{n_t} L_{dt}^j \right) \right) \quad (3)$$

$$L_{Total} = (1 - \alpha - \beta) * L_{FND} + \alpha * (L_{adv}) + \beta * (L_{emo}) \quad (4)$$

where n_s and n_t are number of samples in source and target sets; θ_d , θ_f , θ_e and θ_l are parameters for discriminator, fake news classifier, emotion classifier and LSTM feature extractor; L_{ds} and L_{dt} are binary crossentropy loss for source and target classification; L_{es} and L_{et} are crossentropy loss for emotion classification; L_f is binary crossentropy loss for Fake News Classifier; α and β are weight parameters in L_{Total} . We optimised α and β for each setting for optimal performance.

We used 300 dimension GloVe (Pennington, Socher and Manning, 2014) embeddings. All models were trained for up to 50 epochs, stopped when the peak validation accuracy for the in-domain validation set was achieved. We used a batch size of 25 for every experiment. Each model used the Adam optimizer with learning rate 0.0025. We used an LSTM layer with 256 units for feature extraction, while both fake news detection and emotion classification branches consisted of two dense layers each.

Source	Target	Setting	Accuracy
FakeNewsAMT	Celeb	STL	0.420
		MTL(E)	0.520
		MTL(P)	0.530
		DA STL	0.560
		DA MTL(E)	0.540
		DA MTL(P)	0.600
Celeb	FakeNewsAMT	STL	0.432
		MTL(E)	0.471
		MTL(P)	0.476
		DA STL	0.395
		DA MTL(E)	0.501
		DA MTL(P)	0.551
Politifact	Gossipcop	STL	0.527
		MTL(E)	0.555
		MTL(P)	0.603
		DA STL	0.585
		DA MTL(E)	0.698
		DA MTL(P)	0.671
Celeb	Gossipcop	STL	0.488
		MTL(E)	0.501
		MTL(P)	0.490
		DA STL	0.525
		DA MTL(E)	0.555
		DA MTL(P)	0.587
FakeNewsAMT	Gossipcop	STL	0.451
		MTL(E)	0.652
		MTL(P)	0.620
		DA STL	0.790
		DA MTL(E)	0.805
		DA MTL(P)	0.795
FakeNewsAMT	Politifact	STL	0.363
		MTL(E)	0.450
		MTL(P)	0.530
		DA STL	0.621
		DA MTL(E)	0.704
		DA MTL(P)	0.621

Table 1: Cross-domain evaluation of non-adaptive and adaptive models on FakeNewsAMT, Celeb, Politifact and Gossipcop datasets. Emotion-guided domain-adaptive models (DA MTL(E) and DA MTL(P)) outperform their corresponding STL models in cross-domain settings. Domain-adaptive MTL models consistently outperform baseline STL, non-adaptive MTL and domain-adaptive STL models.

4 Experimental Analysis & Results

We evaluated our proposed approach on various combinations of source and target datasets. Each model was optimized on an in-domain validation set from the source set. Table 1 illustrates our results proving the efficacy of using emotion-guided models in non-adapted and domain-adapted cross-domain settings. It compares non-adaptive models, domain-adaptive models, and our emotion-guided domain-adaptive models in various settings. MTL(E) and MTL(P) refer to emotion-guided multi-task frameworks using Ekman’s and Plutchik’s emotions respectively. STL refers to the single-task framework. DA refers to the use of the domain-adaptive framework, containing a discriminator. Some findings observed are:

MTL(E) and MTL(P) models outperform their STL counterparts in cross-domain settings, as seen in Table 1. This indicates improved extraction of generalizable features by the emotion-guided models, which aids in improved fake news detection across datasets from different domains. MTL(E) and MTL(P) further perform comparably for most settings, and each outperforms the other in three settings respectively.

DA STL models generally outperform STL models in cross-domain settings across multiple combinations of datasets. However, we see the STL model outperformed the DA STL model for Celeb dataset as the source dataset and FakeNewsAMT dataset as target, confirming that unguided adaptation can sometimes lead to negative alignment, reducing the performance of the model.

DA MTL(E) and DA MTL(P) models improve performance in cross-domain settings. Table 1 shows improved results obtained using the emotion-guided adversarial DA models over their non-adaptive counterparts. This shows the scope for improved feature extraction even after using DA, and emotion-guided models act as a solution aiding in correct alignment of the samples and features extracted by the adaptive framework from different domains. Emotion-guided DA models mitigated the issue of negative alignment when Celeb dataset was the source and FakeNewsAMT dataset the target, where STL model outperformed the DA STL model. The emotion-guided DA models helped correctly align the two domains, leading to significantly improved performance.

5 Conclusion

In this work, we showed the efficacy of emotion-guided models for improved cross-domain fake news detection and further presented an emotion-guided domain-adaptive fake news detection approach. We evaluated our proposed framework against baseline STL, emotion-guided MTL, DA STL and emotion-guided DA MTL models for various source and target combinations from four datasets. Our proposed approach led to improved cross-domain fake news detection accuracy, indicating that emotions are generalizable across domains and aid in better alignment of different domains during domain adaptation.

References

- Choudhry, A., Khatri, I., Jain, M., 2022. An emotion-based multi-task approach to fake news detection (student abstract). Proceedings of the AAAI Conference on Artificial Intelligence 36, 12929–12930. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/21601>, doi:10.1609/aaai.v36i11.21601.
- Colnerič, N., Demšar, J., 2020. Emotion recognition on twitter: Comparative study and training a unison model. IEEE Transactions on Affective Computing 11, 433–446. doi:10.1109/TAFFC.2018.2807817.
- Ekman, P., 1992. An argument for basic emotions. Cognition & Emotion 6.
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2015. Domain-adversarial training of neural networks. URL: <https://arxiv.org/abs/1505.07818>, doi:10.48550/ARXIV.1505.07818.
- Guo, C., Cao, J., Zhang, X., Shu, K., Yu, M., 2019. Exploiting emotions for fake news detection on social media. ArXiv abs/1903.01728.
- Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. URL: <https://doi.org/10.1162/neco.1997.9.8.1735>, doi:10.1162/neco.1997.9.8.1735, arXiv:<https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf>.
- Huang, Y., Gao, M., Wang, J., Shu, K., 2021. DAFD: domain adaptation framework for fake news detection, in: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (Eds.), Neural Information Processing - 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8-12, 2021, Proceedings, Part I, Springer. pp. 305–316. URL: https://doi.org/10.1007/978-3-030-92185-9_25, doi:10.1007/978-3-030-92185-9_25.
- K, A., P, D., L, L.V., 2020. Emotion cognizance improves health fake news identification, in: Proceedings of the 24th Symposium on International Database Engineering & Applications, Association for Computing Machinery. doi:10.1145/3410566.3410595.
- Li, Y., Lee, K., Kordzadeh, N., Faber, B., Fiddes, C., Chen, E., Shu, K., 2021. Multi-source domain adaptation with weak supervision for early fake news detection, in: 2021 IEEE International Conference on Big Data (Big Data), pp. 668–676. doi:10.1109/BigData52589.2021.9671592.
- Pennington, J., Socher, R., Manning, C., 2014. GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 1532–1543. URL: <https://aclanthology.org/D14-1162>, doi:10.3115/v1/D14-1162.
- Pérez-Rosas, V., Kleinberg, B., Lefevre, A., Mihalcea, R., 2018. Automatic detection of fake news, in: Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics. pp. 3391–3401.

- Plutchik, R., 1982. A psychoevolutionary theory of emotions. *Social Science Information* 21.
- Saikh, T., De, A., Ekbal, A., Bhattacharyya, P., 2019. A deep learning approach for automatic detection of fake news, in: *Proceedings of the 16th International Conference on Natural Language Processing*, NLP Association of India, International Institute of Information Technology, Hyderabad, India. pp. 230–238.
- Sheng, Q., Cao, J., Zhang, X., Li, R., Wang, D., Zhu, Y., 2022. Zoom out and observe: News environment perception for fake news detection. URL: <https://arxiv.org/abs/2203.10885>, doi:10.48550/ARXIV.2203.10885.
- Shu, K., Cui, L., Wang, S., Lee, D., Liu, H., 2019. Defend: Explainable fake news detection, in: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery, New York, NY, USA. p. 395–405. URL: <https://doi.org/10.1145/3292500.3330935>, doi:10.1145/3292500.3330935.
- Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. *Science* 359, 1146–1151. doi:10.1126/science.aap9559.
- Zhang, T., Wang, D., Chen, H., Zeng, Z., Guo, W., Miao, C., Cui, L., 2020. Bdann: Bert-based domain adaptation neural network for multi-modal fake news detection, in: *IJCNN*. doi:10.1109/IJCNN48605.2020.9206973.
- Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., Shu, K., 2021. Mining dual emotion for fake news detection, in: *Proceedings of the Web Conference 2021*, Association for Computing Machinery, New York, NY, USA. p. 3465–3476. doi:10.1145/3442381.3450004.

EV Control in G2V and V2G modes using SOGI Controller

Sudhanshu Mittal
Electrical Engineering
Delhi Technological University
Delhi, India
sudhanshu15mittal@gmail.com

Alka Singh
Electrical Engineering
Delhi Technological University
Delhi, India
alkasingh.eed@gmail.com

Prakash Chittora
Electrical Engineering
Delhi Technological University
Delhi, India
prakashchittora@gmail.com

Abstract— Electric vehicle (EV) technology is developing at a very fast pace. The Vehicle to grid (V2G) and Grid to vehicle (G2V) technology enables bidirectional power transfer between an electric vehicle and the grid. It is an emergent area of research. In the G2V operation mode EV, batteries are charged from the grid. The energy stored in the batteries may also be supplied back to the power grid during the V2G operation mode, which helps in maximum demand saving, load balancing voltage management and improved system reliability. An onboard bilateral battery charger for EV is proposed in this research paper, with G2V and V2G applications. Bi-directional power electronics-based converter is interfaced between EV and the grid to provide G2V and V2G modes. A Second Order Generalized Integrator (SOGI) control technique is used to control the H-bridge inverter which shows stable steady state and good dynamic performance. The battery charging/discharging current and voltage are controlled using a PWM controller further effect of nonlinear load dynamics on the system performance is also studied. Exhaustive simulation study is performed in MATLAB/Simulink environment which is also reflected in this paper.

Keywords— G2V (Grid to Vehicle), V2G (Vehicle to Grid), GSC (Grid side converter), BSC (Battery side converter), EV (Electric Vehicle), OBC (On Board Charger).

I. INTRODUCTION:

Moreover, the world population is growing at a very fast rate as compared to limited availability of natural resources. The rising prices of fossil fuels such as diesel and gasoline has boosted the demand of EVs. EVs have the capability to reduce our dependency on fossil fuels. Moreover, India has targeted to develop, battery-interfaced vehicles at a CAGR (Compounded Annual Growth Rate) of 90%, reaching \$150 billion by 2030.

In EV, an onboard battery charger connects the battery to the grid. The On-Board Charger (OBC) allows electric vehicles to get charged from the AC grid, it is commonly utilized in the automobile industry due to its simplicity, especially when compared to off-board charging systems, which are expensive and need a huge amount of space. [1-4]

Unidirectional OBCs are widely utilized due to its low battery deterioration and simplicity but they do not support V2G technology [5]. So, bidirectional OBCs which can achieve V2G capability by returning electrical energy to the grid are very beneficial during peak power demand [6-7]. Single phase bidirectional IGBT switches are used for both G2V and V2G mode of operation. In paper [8], a bidirectional EV charger with an adjustable DC link voltage is proposed. It focuses on minimizing the ripple of higher frequency in the Igrid. Paper [9] discusses a bidirectional EV

charger employing the SOGI approach to improve overall system performance and reliability and PLL may be used in the control loop. Paper [10] proposed, an onboard bidirectional battery charger for EVs having G2V, V2G and Vehicle-to-Home (V2H) technologies combined. In G2V mode of operation, the charging the battery of EV must be controlled to conserve the quality of power in grids.

It is envisioned that with the increase in number of EVs, a significant quantity of energy will be stored in their batteries and this will create future possibility of an energy flow in the reverse direction during V2G mode. This can be helpful in managing energy crisis in future.

Paper [11] lists various battery charger topologies and charging power levels are presented in accordance with industry standards. The issues of EV/ Plug-in hybrid EV charging infrastructure are also highlighted and the state of the art industrial electronics solutions are presented. In [12], the role of power electronics in onboard energy storage and battery management in EV is examined. This paper also proposed methods to improve productivity and effectiveness by minimizing charger size, charging time, quantity and cost of power extracted from the grid.

Paper [13] focusses on design and modeling of bidirectional battery charger and control algorithm for EV has been investigated. Two converters configuration is used in which a common DC link is present. One of the converter is connected to the grid and the another is connected to the battery. The bi-directional AC-DC converter on grid side converter (GSC) that is responsible for maintaining UPF operation of grid and the second converter is responsible for maintaining charge on the battery.

Paper [14] discuss the design and modelling of converter using a front end AC-DC PFC converter and a bidirectional DC-DC converter is investigated. Paper[15] discusses a bidirectional isolated DAB (Dual Active Bridge) DC-DC converter. A single phase full-bridge 53.2-V, 40-A·h Lithium ion battery bank is integrated with a bidirectional DC-DC converter. The bidirectional converter shows high efficiency in low-voltage and high-current mode of operation. Although bidirectional OBCs face numerous challenges [16] increased system such as cost, low power density, high weight etc. However, due to V2G technology power transfer from vehicle to grid will enable a remarkable improvement in system reliability [17-18]. In future it is widely expected that bidirectional OBCs will become the primary charging solution [19].

In this paper an onboard bidirectional EV charger is proposed with additional power quality improvement features. SOGI technique is used for fundamental current

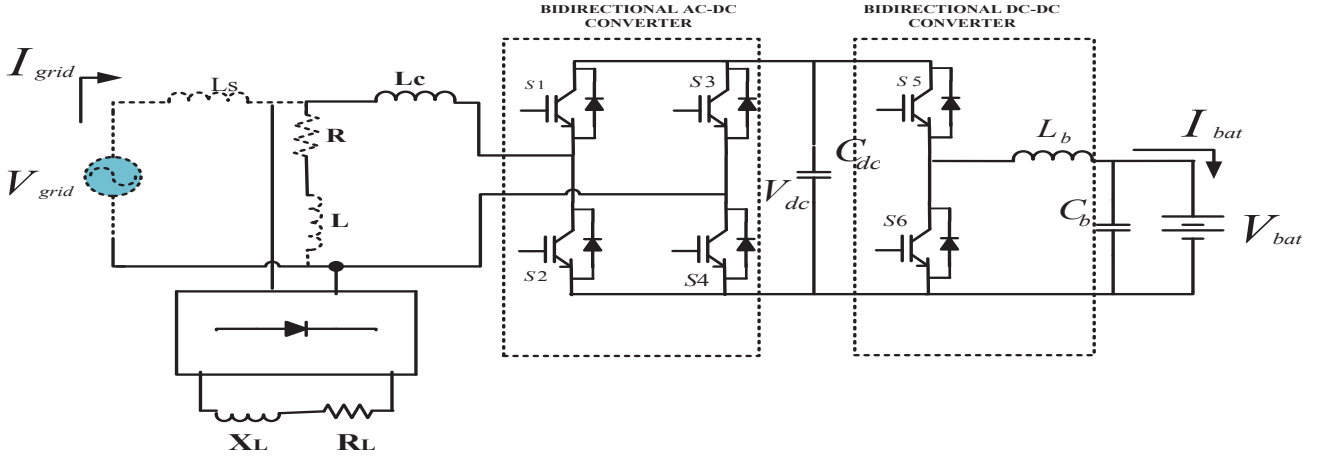


Fig.1 Circuit diagram of complete system with H-bridge inverter and bidirectional battery charger.

extraction, which is necessary for the generation of reference current for the H-bridge inverter controlled as shunt compensator. The proposed system offer G2V, V2G modes of operation along with harmonic reduction, power factor improvement on the source side.

II. CONFIGURATION OF PROPOSED SYSTEM

The proposed system is depicted in Fig.1. The system has two converters, one is Grid Side Converter (GSC) and other is Battery Side Converter (BSC). Both GSC and BSC show a common DC link.

The GSC is a bidirectional AC-DC converter, In G2V mode it operates as a rectifier to supply to the battery via dc link capacitor and in V2G mode it feeds energy into grid from batteries which acts as filter to high frequency harmonic converters current. The battery is connected via bidirectional buck boost converter to DC link. The buck converter is used to manage the charging current and voltage during G2V mode. The boost converter elevates the battery voltage; allowing the battery power to be fed into grid. The GSC is connected to grid via a coupling inductor L_c utilize.

III. CONTROLLER DESIGN

G2V side AC-DC converter is delineate for 2.5 KW. Where its minimal dc link voltage vital for power transfer is 260 volts. Therefore, dc link capacitor is delineate for 400 volts.

A. SECOND ORDER GENERALIZED INTEGRATOR METHOD: [SOGI]

The SOGI controller is used to create two orthogonal components in alpha-beta reference frame from the input.

The fundamental component of load current as well as unit template is generated using SOGI controller and it is described in following section.

1) Generation of gating pulses for GSC:

In the SOGI controller developed for extraction of load fundamental current as shown in Fig. 2, the error between input and alpha component of the output signal is amplified by a factor of 'K' and it is compared to the beta component signal. The total performance of SOGI is influenced by the parameter 'K'. To make the system flexible to frequency changes, SOGI controller employs the angular frequency

block created by the second half of the circuit as shown in Fig 2. The higher order frequency component occurs in the measured sample can be removed using SOGI block. In the frequency domain, SOGI has the following structure:

$$TF = \frac{ws}{w^2 + s^2} \quad (1)$$

where, TF is the transfer function of SOGI and ' ω ' depicts the resonance frequency, often known as the angular frequency of fundamental. The transfer function of closed loop for α and β current is shown in equation (2) and (3), respectively.

$$H_\alpha(s) = \frac{i_\alpha(s)}{i(s)} = \frac{Kws}{s^2 + Kws + w^2} \quad (2)$$

$$H_\beta(s) = \frac{i_\beta(s)}{i(s)} = \frac{Kw^2}{s^2 + Kws + w^2} \quad (3)$$

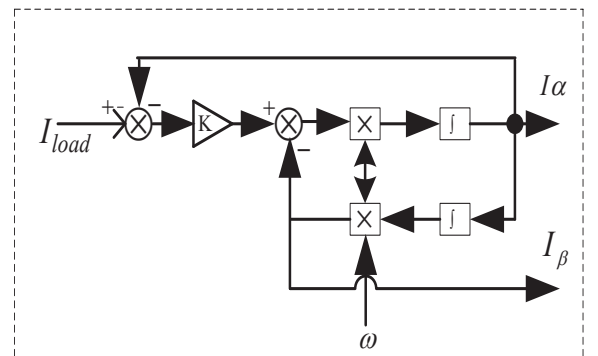


Fig 2: Block diagram of the SOGI controller.

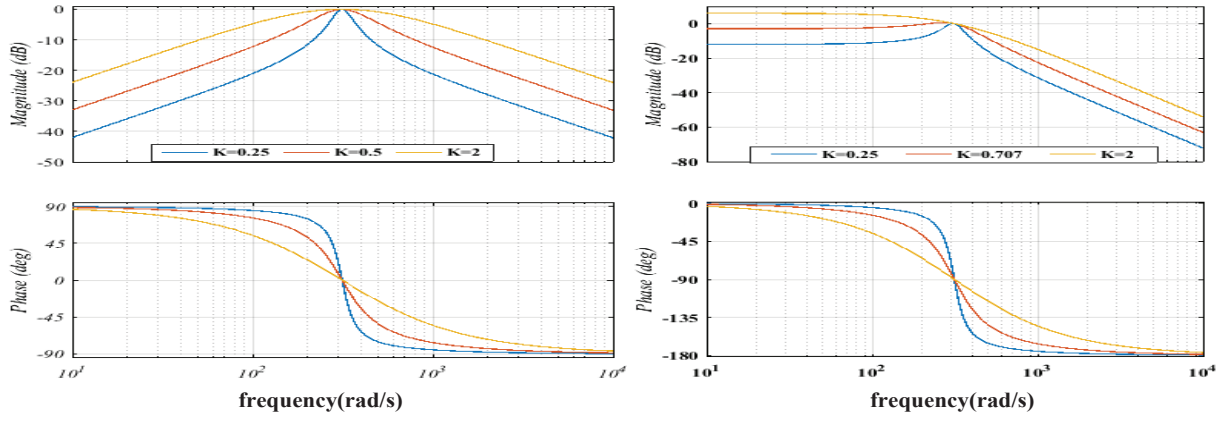


Fig.3 Bode plot of SOGI filter (a). $H_\alpha(s)$ for different values of K. (b). $H_\beta(s)$ for different values of K.

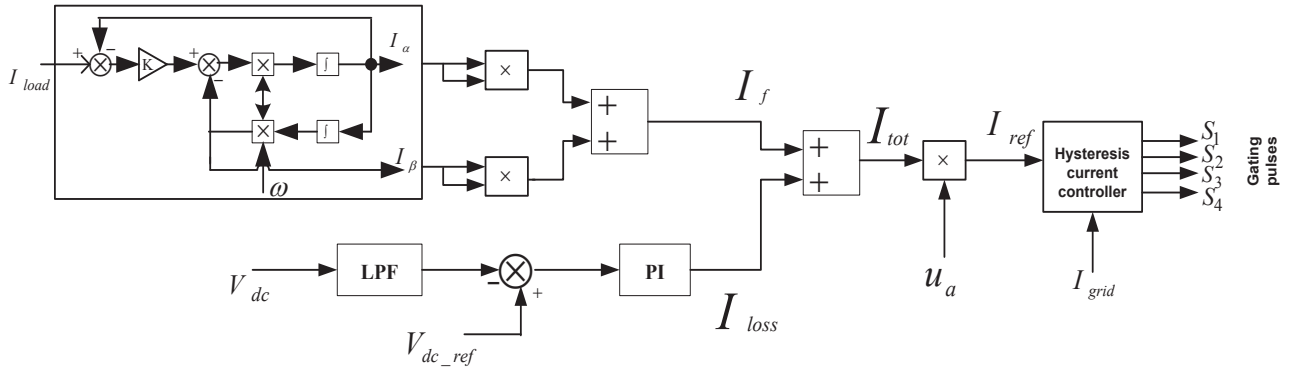


Fig 4: Proposed control scheme.

Fig.3 shows the bode plot of three different values of ‘K’ (0.25, 0.707, 2) for alpha and beta current transfer function depicted by equation (2) and (3). $H_\alpha(s)$ is defined as a 2nd order band-pass filter having zero phase shift and $H_\beta(s)$ is defined as a 2nd order low-pass filter having a unity gain with 90° phase deviation at ‘ ω ’ taken as 314 rad/s.

It is observed that bandwidth of the system is directly proportional to gain K. For $H_\alpha(s)$, if the value of ‘K’ is less, the bandwidth is less and filtering response may not be good. At K=0.707 as compared to other values, bandwidth is optimized so the system has good filtering performance. Similarly, in $H_\beta(s)$, smaller the value of ‘K’, the bandwidth lowers which effects the filtering effect.

The fundamental component of load current is extracted by the use of SOGI filter. The input signal (I_{load}) is passed through the SOGI block and in-phase, quadrature phase signals, given as I_α and I_β respectively are obtained at the output. The fundamental current (I_f) thus can be obtained using

$$I_f = \sqrt{I_\alpha^2 + I_\beta^2} \quad (4)$$

For the calculation of switching losses in the shunt compensator, the DC-link voltage (V_{dc}) is passed through low pass filter and it is computed with the reference DC-link voltage ($V_{dc_ref}^*$). The error of dc link voltage is given to a PI (Proportional Integral) controller, which calculates the

switching loss (I_{loss}) of the proposed system. The DC link voltage is maintained at 400V. The equation is shown below,

$$I_{loss}(k) = [(k_p + k_i/s) (V_{dc_ref}^*(k) - V_{dc}(k))] \quad (5)$$

where, k_p and k_i are the gain coefficients of PI controller.

The total reference current magnitude (I_{tot}) is calculated by adding fundamental load current component and DC link loss component which is provided in equation (6).

$$I_{tot} = I_{loss} + I_f \quad (6)$$

Now, for the generation of unit template, first V_α and V_β , is generated by the SOGI block and provided in equation (7) and (8).

$$H_\alpha(s) = \frac{v_\alpha(s)}{v(s)} = \frac{Kws}{s^2 + Kws + w^2} \quad (7)$$

$$H_\beta(s) = \frac{v_\beta(s)}{v(s)} = \frac{Kw^2}{s^2 + Kws + w^2} \quad (8)$$

Now, the unit template (u_a), is generated using following equation

$$u_a = \frac{v_\alpha}{\sqrt{v_\alpha^2 + v_\beta^2}} \quad (9)$$

The reference current is generated by multiplying total reference current and unit template.

$$I_{ref} = I_{tot} * u_a \quad (10)$$

Finally a hysteresis current controller (HCC) is used to generate gating pulses for IGBT switches (insulated gate bipolar transistor) of GSC. The entire control scheme is depicted in Fig.4.

b). Generation of gating pulses for BSC:

The reference battery current (I_{bref}) and actual battery current (I_{bat}) are compared to get error signal. As illustrated in Fig.5, the error signal is sent into the PI controller, and the output of the PI controller is fed into the PWM generator, which provides the pulses for the bi-directional DC-DC converter. When switch S_5 is high and S_6 is low, it is discharging operation, similarly charging operation results when S_5 is low and S_6 high.

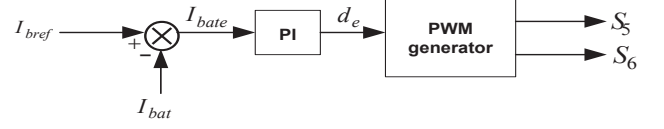


Fig 5: Bidirectional DC-DC converter.

Where, I_{bate} is the error between I_{bat} and I_{bref} .

IV. RESULT ANALYSIS:

A Matlab/Simulink model is developed and analysis under G2V, V2G mode is performed along with PQ improvement. A detailed description is shown in the following section.

A). Operation in G2V mode:

The matlab/simulink result of the grid voltage (V_{grid}), grid current (I_{grid}), battery voltage (V_{bat}), battery current (I_{bat}), % SOC and DC link voltage (V_{dc}) for G2V mode is depicted in Fig.6. The proposed charger is design to operate with 230 V, 50 Hz, single-phase supply and the battery rating selected is 240V, 35Ah. The battery charging current is negative indicating it is providing power to grid. In G2V mode (charging condition) the signals V_{grid} and I_{grid} are in same phase.

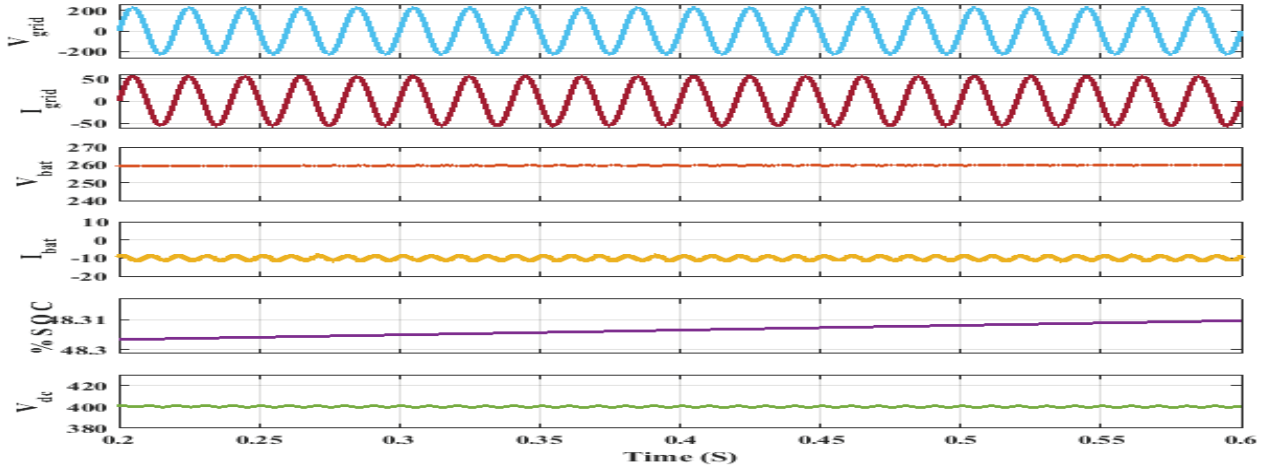


Fig 6: G2V charging, (a) V_{grid} (b) I_{grid} (c) V_{bat} (d) I_{bat} (e) % SOC (f) V_{dc}

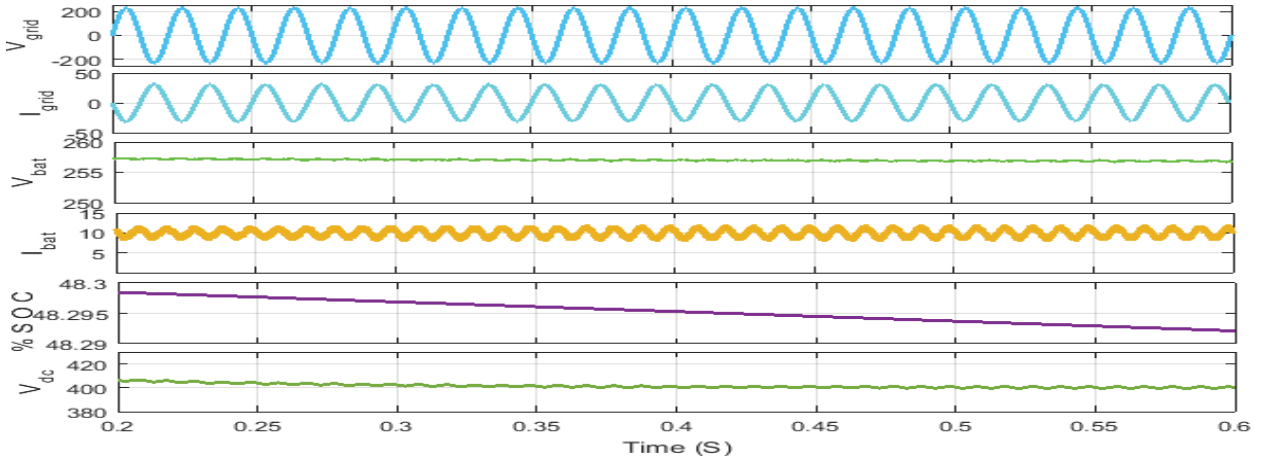


Fig 7: V2G discharging, (a) V_{grid} (b) I_{grid} (c) V_{bat} (d) I_{bat} (e) % SOC (f) V_{dc}

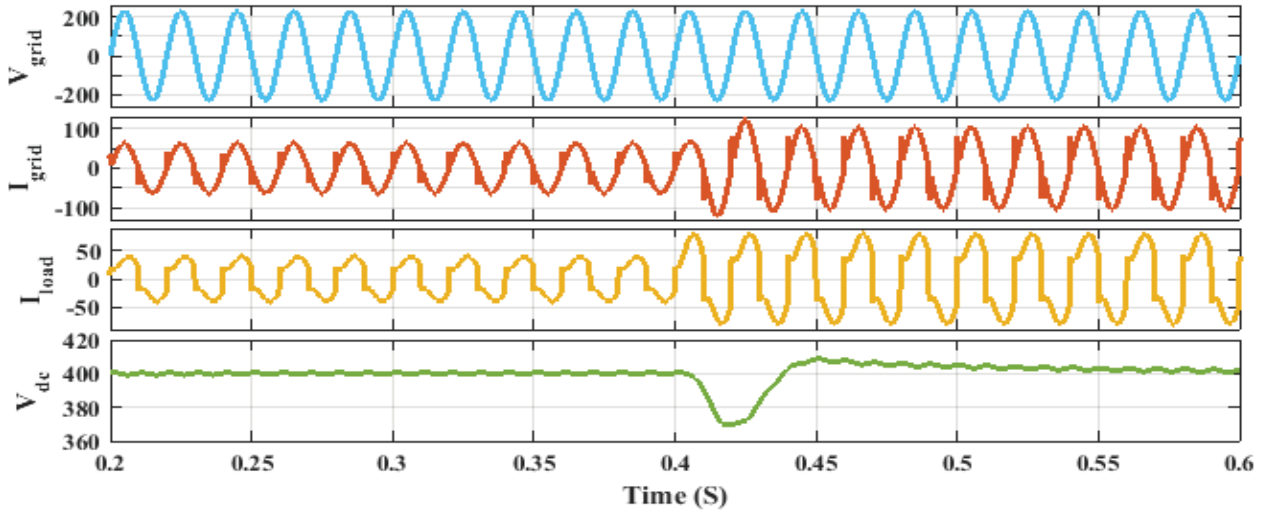


Fig 8. Load Increase, (a) V_{grid} (b) I_{grid} (c) I_{load} (d) V_{dc}

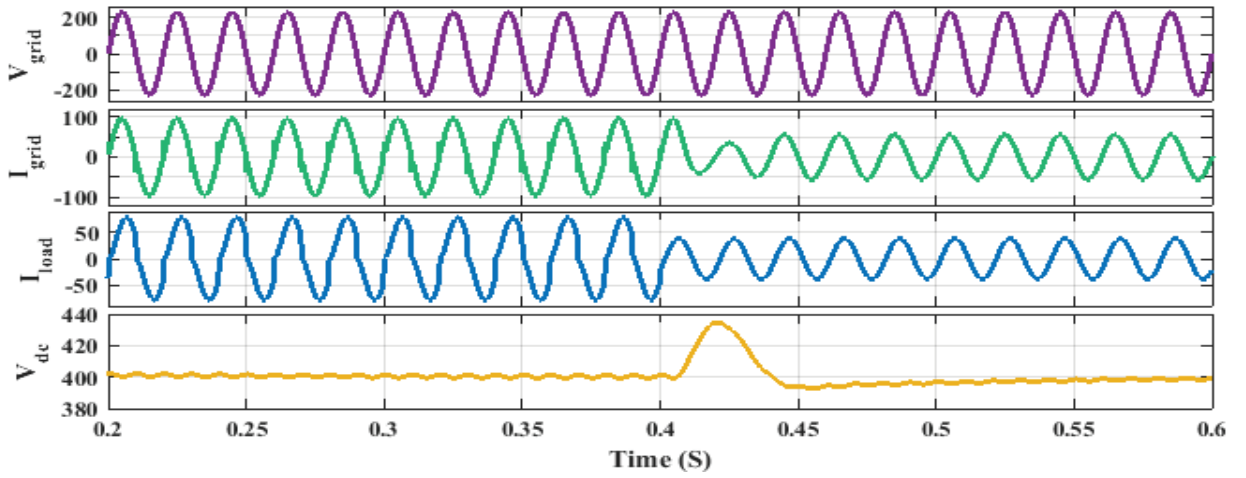


Fig 9. Load decrease, (a) V_{grid} (b) I_{grid} (c) I_{load} (d) V_{dc}

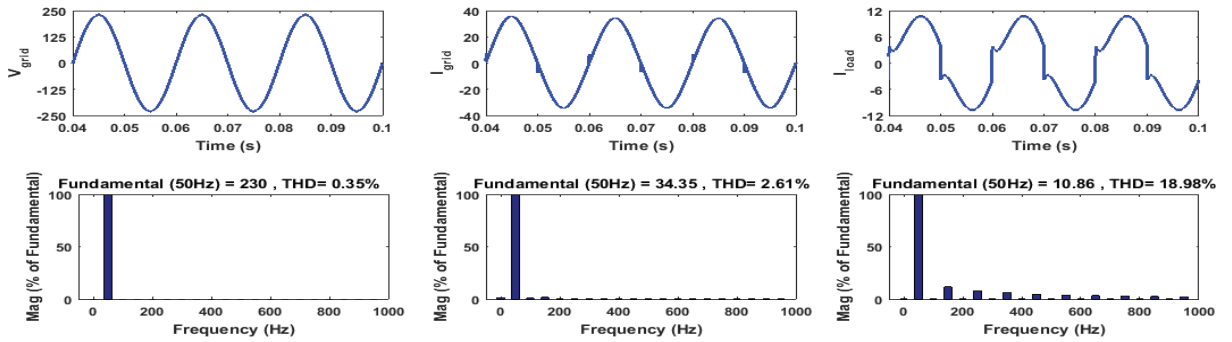


Fig 10. Harmonic spectrum of (a) V_{grid} (b) I_{grid} (c) I_{load}

B. Operation in V2G Mode:

In V2G mode (discharging condition), the matlab/simulink results of the grid voltage, grid current, battery voltage, battery current, % SOC and DC link voltage are shown in Fig.7.

The grid voltage and current are observed to be in phase opposition to each other. For charging and discharging mode, the DC link voltage (260V) is used for minimizing the high frequency ripple in the grid current.

The performance under load change is observed in Fig.8 and Fig.9. Under the load increase and decrease conditions, the grid voltage, grid current, load current and DC link voltage are shown in Fig.8(a)-(d) and Fig.9(a)-(d) respectively. In these results, as the load increases/decreases at 0.4 S, the V_{grid} remains constant but the I_{grid} and I_{load} increases/decreases respectively. In both conditions, the DC link voltage has small transients around $V_{dc} = 400$ volts which is the reference value. The total harmonic distortion (THD) of the V_{grid} , I_{grid} are 0.35%, 2.61% respectively

as depicted in Fig.10 (a)-(b). In Fig. 10(c), a non-linear load having Iload's THD is 18.98%.

IV. CONCLUSION:

In this paper, modelling and simulation of single-phase on-board bi-directional battery charger for an electric vehicle have been presented in detail. Two converters in the form of a AC-DC bidirectional converter and a bidirectional DC-DC converter have been considered. The grid side converter operation is controlled with a SOGI controller. Matlab/Simulink results shows the operability of the proposed method. The bidirectional controller has been designed for the control of EV in G2V and V2G modes. Additionally power quality improvement using SOGI controller for GSC is also achieved and demonstrated. Simulation study in steady state as well as dynamic condition is also carried out to show the effectiveness of the SOGI controller.

REFERENCES

- [1] I.-O. Lee, "Hybrid PWM-resonant converter for electric vehicle onboard battery chargers," *IEEE Trans. Power Electron.*, vol. 31, no. 5, pp. 3639–3649, May 2016.
- [2] D. C. Erb, O. C. Onar, and A. Khaligh, "Bi-directional charging topologies for plug-in hybrid electric vehicles," in *Proc. 28th IEEE Appl. Power Electron. Conf. Expo. (APEC)*, Palm Springs, CA, USA, Feb. 2010, pp. 2066–2072.
- [3] B.-K. Lee, J.-P. Kim, S.-G. Kim, and J.-Y. Lee, "An isolated/bidirectional PWM resonant converter for V2G (H) EV on-board charger," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 7741–7750, Sep. 2017.
- [4] Y. Xiao, C. Liu, and F. Yu, "An effective charging-torque elimination method for six-phase integrated on-board EV chargers," *IEEE Trans. Power Electron.*, vol. 35, no. 3, pp. 2776–2786, Mar. 2020.
- [5] K. Uddin, M. Dubarry, and M. B. Glick, "The viability of vehicle-to-grid operations from a battery technology and policy perspective," *Energy Policy*, vol. 113, pp. 342–347, Feb. 2018.
- [6] Z. Liu, B. Li, C. F. C. Lee, and Q. Li, "Design of CRM AC/DC converter for very high-frequency high-density WBG-based 6.6 kW bidirectional on-board battery charger," in *Proc. IEEE Energy Convers. Congr. Expo. (ECCE)*, Milwaukee, WI, USA, Sep. 2016, pp. 1–8.
- [7] S. Semsar, T. Soong, and P. W. Lehn, "On-board single-phase integrated electric vehicle charger with V2G functionality," *IEEE Trans. Power Electron.*, vol. 35, no. 11, pp. 12072–12084, Nov. 2020.
- [8] B. Singh and A. Verma, "Adaptive DC-link Voltage Based Bi-Directional Charger for Electric Vehicles," 2019 IEEE International Conference on Environment and Electrical Engineering and 2019 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), 2019, pp. 1–6, doi: 10.1109/EEEIC.2019.8783982.
- [9] N. Z. Kashani, M. A. Parazdeh, M. Eldoromi, A. A. M. Birjandi and P. Amiri, "Grid Synchronization of Bidirectional Electric Vehicle Chargers Using Second Order Generalized Integrator based Phase Lock Loop," 2021 12th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC), 2021, pp. 1–5, doi: 10.1109/PEDSTC52094.2021.9405867.
- [10] Kotla Aswini, Jilidimudi Kamala, Lanka Sriram, Bhasuru Kowshik, Bugatha Ram Vara Prasad, Damaraju Venkata Sai Bharani, 2021, Design and Analysis of Bidirectional Battery Charger for Electric Vehicle, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 07 (July 2021).
- [11] S. S. Williamson, A. K. Rathore and F. Musavi, "Industrial Electron. for Electric Transportation: Current State-of-the-Art and Future Challenges," *IEEE Trans. Ind. Electron.*, vol. 62, no. 5, pp. 3021–3032, May 2015.
- [12] M. Yilmaz and P. T. Krein, "Review of Battery Charger Topologies, Charging Power Levels, and Infrastructure for Plug-In Electric and Hybrid Vehicles," *IEEE Trans. Power Electron.* vol. 28, no. 5, pp. 2151–2169, May 2013..
- [13] A. Verma and B. Singh, "Bi-directional charger for electric vehicle with four quadrant capabilities," 2016 *IEEE 7th Power India International Conference (PIICON)*, 2016, pp. 1–6, doi: 10.1109/POWERI.2016.8077451.
- [14] M. Restrepo, J. Morris, M. Kazerani and C. Canizares, "Modeling and Testing of a Bidirectional Smart Charger for Distribution System EV Integration," *IEEE Trans. Smart Grid*, Early Access.
- [15] N. M. L. Tan, T. Abe and H. Akagi, "Design and Performance of a Bidirectional Isolated DC–DC Converter for a Battery Energy Storage System," *IEEE Trans. Power Electron.*, vol. 27, no. 3, pp. 1237–1248, March 2012.
- [16] C. Gould, K. Colombage, J. Wang, D. Stone, and M. Foster, "A comparative study of on-board bidirectional chargers for electric vehicles to support vehicle-to-grid power transfer," in *Proc. IEEE 10th Int. Conf. Power Electron. Drive Syst. (PEDS)*, Kitakyushu, Japan, Apr. 2013, pp. 639–644.
- [17] J. Escoda, J. Fontanilles, D. Biel, V. Repecho, R. Cardoner, and R. Griño, "G2 V and V2G operation 20 kW battery charger," *World Electr. Vehicle J.*, vol. 6, no. 4, pp. 839–843, Dec. 2013.
- [18] J.S. Lai, L. Zhang, Z. Zahid, N.-H. Tseng, C.-S. Lee, and C.-H. Lin, "A high-efficiency 3.3-kW bidirectional on-board charger," in *Proc. IEEE 2nd Int. Future Energy Electron. Conf. (IFEEC)*, Taipei, Taiwan, Nov. 2015, pp. 1–5.
- [19] K. Fahem, D. E. Chariag, and L. Sbita, "On-board bidirectional battery chargers topologies for plug-in hybrid electric vehicles," in *Proc. Int. Conf. Green Energy Convers. Syst. (GECS)*, Hammamet, Tunisia, Mar. 2017, pp. 1–6.

Experimental Simulation of Hydraulic Jump for the Study of Sequent Depth Using an Obstruction

Daisy Singh^a, Abhishek Prakash Paswan^a, S. Anbukumar^a and Rahul Kumar Meena^{a,1}

^aDepartment of Civil Engineering, Delhi Technological University, Delhi

Abstract. Hydrological jump characteristics in a rectangular open channel flume with sluice gates at both ends are the focus of this article. Among the many aspects of hydraulic jump that have been examined analytically and experimentally are: (i) Sequent depth relation ($\frac{y_2}{y_1}$), (ii) Length of the jump, (iii) Relative loss of energy ($\frac{E_L}{E_1}$), (iv) Water profile of the jump. There was a total of 15 inflow tests conducted. With values between 1.5 and 1.7, Froude's number displays considerable variation. The results show that when the slope increases, the sequent depth ratio, leap length, and Relative energy loss all decrease. The fluctuation of the relative energy loss ($\frac{E_L}{E_1}$) with Pre-jump Froude's number (F_1) for different slopes is also indicated by the graph. The graph also depicts the water profile of the leap. Continuous flow is shown by the relatively uniform discharge at each location.

Keywords. Hydraulic jump, froude number, rectangular flume, sequent depth, jump length

1. Introduction

The hydraulic jump is a well-known hydraulics performance that often occurs in open channel flow (such rivers and spillways). As a high-velocity supercritical flow decelerates to a subcritical flow, the fast-following flow gradually slows and rises, transforming part of the flow's kinetic energy into potential energy. This phenomenon is known as a hydraulic jump. Hydraulic jumps are typical in open channels as flows move from supercritical to subcritical. When the depth or width of a channel changes, a flow transition occurs.

1.1. Types of Jumps

The jump on a horizontal floor might be categorized as follows, depending on the incoming Froude No. F_1 :

- The flow is essential when $F_1 = 1$, thus no jump is conceivable.
- The jump is known as an undular jump when F_1 is between 1 and 1.7, when the water surface undulates.

¹ Rahul Kumar Meena, Corresponding author, Department of Civil Engineering, Delhi Technological University, Delhi; E-mail: rahul.08dtu@gmail.com.

- A succession of little rollers forms on the surface of the jump at $F_1 = 1.7$ to 2.5 , while the water downstream remains smooth. This is referred to as a faint jump.
- For $F_1 = 2.5$ to 4.5 , an oscillating jet dips into the bottom of the jump and rises to the surface without a noticeable effect. This kind of jump is known as an oscillating jump. The entering Froude number for barrages and canal regulators is often in this zone or the preceding zone, i.e. the weak jump zone.
- The jump is nicely balanced and performs at its optimum for $F_1 = 4.5$ to 9.0 . The amount of energy dissipated varies between 45 and 70%. A steady jump is a name for this type of jump.
- For $F_1 = 9.0$ and larger, the jump is a strong jump, with energy dissipation reaching up to 85%.
- Studying the feasibility of a hydraulic jump has been studied for close to two centuries. Bidone was the one who initially looked into the situation (1819). After then, researchers paid a lot more attention to the issue, and several well-known hydraulicians conducted extensive experiments and theoretical analyses of free hydraulic jump on horizontal beds. As a common method of energy dissipation, hydraulic jumps are frequently used in hydraulic structures.

In the smooth rectangular open channel, Belanger asserted that a hydraulic jump will occur when the depth ratio or velocity ratio is:

$$y_2 = \frac{1}{2} y_1 \left(\sqrt{1 + 8F_1^2} - 1 \right) \quad (1)$$

It is performed in the interval between the start of the project and the completion of the hydraulic jump. This indicates that hydraulic jump will occur in a smooth rectangular channel if the initial depth y_1 , the subsequent depth y_2 , and the inflow Froude number F_2 all fulfil the aforementioned momentum equation. The flow regime is specified in terms of the Froude number, which is calculated by dividing the unit inertial force by the unit gravitational force (F_2).

For inflow, the Froude number F_2 is calculated as follows:

$$F_2 = \frac{v_2}{\sqrt{gy_2}} \quad \text{where} \quad v_2 = \frac{q}{y_2} \quad \text{where} \quad q = \frac{Q}{b} \quad (2)$$

Supercritical flow occurs when $F_2 \geq 1$; critical flow occurs when $F_2 = 1$; and subcritical flow occurs when $F_2 \leq 1$.

Subcritical flow is caused by $F_1 \leq 1$, critical flow is caused by $F_1 = 1$, and supercritical flow is caused by $F_1 \geq 1$.

Typically, the following eight factors are involved in any hydraulic jump: F_1 , V_1 , y_1 , F_2 , V_2 , q and H_L . Four independent equations link these variables, as shown below [1]:

$$y_1 y_2 (y_1 + y_2) = \frac{2q^2}{g} \quad (3)$$

$$H_L = \frac{(y_2 - y_1)^3}{4y_1 y_2} \quad (4)$$

$$V_1 = \frac{q}{y_1} \quad (5)$$

$$V_2 = \frac{q}{y_2} \quad (6)$$

Many Researchers as Elnikhely (2014), Neluwat et al. (2013), Imran and Akib (2013), Chern and Syamsuri (2013), Habibzadeh et al. (2012), Habibzadeh et al. (2011), Elsebaie and Shabayek (2010), Carallo et al. (2007), Ead and Rajaratnam (2002) Ali Mohamed (1991), Hughes and Flack (1984) etc..have Performed Experiments for Analysis of the Hydraulic Jump;

Elnikhely studied the "Effect of staggered roughness element on flow characteristics in rectangular channel," in which an artificial staggered roughness was established at the flume's base and the profile of the water's surface was measured at various points. The experimental data matches the estimates for relative depth and energy loss very well [2,3]. The following experiment was carried out by Imran and Akib: A evaluation of hydraulic jump qualities in various channel bed conditions. When compared to a smooth bed, the jump and subsequent depth were reduced. It was discovered that the performance of a corrugated or rough bed outperformed that of a smooth bed [4]. The "Effect of corrugated bed on hydraulic jump characteristic using a S.P.H (smoothed particle hydrodynamic) approach" was investigated by Chern and Syamsuri, the hydraulic jump is common in open channel flows [5]. The following experiment was carried out by Habibzadeh et al. on "Baffle Block Performance in Submerged Hydraulic Jumps. For a single one flow region, empirical equation was produced, for critical value prediction of submergence factor [6]. Habibzadeh et al. performed an experiment on "Exploratory study of submerged hydraulic jump with blocks. It was found that efficiency of energy dissipation was justification or role of submergence function with efficiency maximum excess efficiency parallel to Froude jump [7]. The following experiment was carried out by Elsebaie and Shabayek on the "Formation of hydraulic jump on a corrugated bed", to create the roughness, a variety of corrugated beds were used. The jump is shorter on different corrugated beds than it is on a smooth bed. The shear stress produced by supercritical flow is reduced as a result of interaction with the corrugated bed [8]. Carallo et al. performed the following "Hydraulic jump on bumpy bed". As bed roughness grew, the subsequent depth ratio, as well as Froude number decreased [9]. A "Hydraulic jump on corrugated bed" experiment was carried out by Ead and Rajaratnam. With a small departure from the plane wall jet profile, the velocity profiles at various sections were found to be equal [10]. Ali Mohamed performed an experiment on the Effect of roughened bed stilling Basin on length of rectangular hydraulic jump. For the optimum length of roughness, a render practical equation is required to express the hydraulic jump characteristics by the use of roughness [11]. Hughes and Flack tested "Hydraulic jump properties over rocky bed" as follows, to produce roughness on a smooth bed, a strip roughness bed and densely packed gravels were used [12]. The experiment was conducted by Yu Zhou, Jianhua Wu, Fei Ma, and Jianyong Hu, and it was titled "Uniform flow and energy dissipation of hydraulic-jump-stepped spillways". [13]. R. E. E. Antigha, E. Nyah, and J. G. Egbe, in this paper, researchers conduct experimental studies of energy loss during hydraulic jump When compared to a level-bed confined flume, the amount of energy lost during a weir-created hydraulic jump is greater. [14]. Sonia Cherhabil and Mahmoud Debabeche conducted "Experimental Study of Sequent Depths Ratios of Hydraulic Jump in Sloped Trapezoidal Channels" [15]. The experiment "Reproduction

of Water Surface Profile Using ANSYS – CFD" was conducted out by Prasanna S V S N D L and Suresh Kumar N: Numerical analysis is a valuable tool for acquiring a comprehensive understanding of the flow field's characteristics, which is difficult to achieve using traditional methods. The purpose of this inquiry is to draw the water surface profile for varied discharges. Both experimental and analytical calculations agree with the water surface profile for the selected discharges [16-19].

In this paper Laboratory experiment was conducted in the hydraulics lab on a rectangular open channel flume which has a width of 0.30 m and length of 10 m. After calculating pre- and post- jump depth of the rectangular flume, various results were calculated in which the type of water profile which is shown by the graph of Length of jump vs. Height of the jump is calculated, type of jump and graph of sequent depths vs. Froude no. (F_1) is also calculated. After calculating these parameters, the results were also verified with the previous studies of these parameters.

2. Methodology

2.1. Experimental Set-up

i) Open Channel Rectangular Flume: Flumes are used in open channels to assess flow rate (discharge). They usually range in width from a few centimetres to over 15 metres. The flume used in the experiment is represented in figure 1. The water depth in the approach portion of flumes can range from a few centimetres to over two metres.



Figure 1. Rectangular flume used in this project placed in the Laboratory.

Rectangular flumes have a constriction at the throat and/or a raised invert (bottom) at the throat to work. In a properly operated flume, either characteristic can create critical flow near the throat. These flumes are easier to build and may be readily integrated into an existing channel.

ii) Brick: The figure 2 shows the brick used as the obstruction in the experiment. As the end point goes upward, the roller downstream of the hydraulic leap moves upward to the point where it almost overflows.



Figure 2. The brick.

iii) Point Gauge: The sequent depths y_1 and y_2 is measured in the experiment with the help of Point gauge. The instrument represented in figure 3 is used to measure the flow depth. It is generally fitted on rails over the flume so that it can slide from one point to another for easy measurement.



Figure 3. Point Gauge.

iv) Storage Tank: The Tank shown in figure 4 was used for transferring the water and measuring the discharge.



Figure 4. The Storage Tank.

v) Inlet valve: It used to control the rate of flow during the experiment, increase or decrease in the flow is controlled by the inlet valve and the same is presented in figure 5 and the sluice gate is used to control the inlet and outlet boundary condition during the experiment.



Figure 5. The inlet valve.

vi) Water pump: It is necessary for the flow of water through the flume and it is working for the lift of water through the source to the flume.

3. Experimental Process

- As a means of controlling hydraulic jump, the schematic diagram is presented in the figure 6, and maximizing hydraulic jump efficiency, a brick has been installed in the center of the flume as presented in figure 8. The force acting on such an impediment in a hydraulic jump rapidly decreases to a minimum as depicted in figure 7 when the end point of the roller downstream of the hydraulic jump rises up to the point where it is almost overflowing as shown in side view of the hydraulic jump in figure 9.
- The scour danger zone may be less than in a regular hydraulic jump because of the obstruction. In this case, the force on the barrier increases as the hydraulic leap decreases in length and moves upstream. Such a rapidly changing flow is characterized by a dispersed dispersion of velocities. Therefore, in a cross-section with different speeds, momentum is easily gained.
- To explore the characteristics of hydraulic leap and to analyse the water profile distribution, a physical experiment modelling the behaviour of a sluice gate and an associated hydraulic jump experiment were carried out in various settings.
- The physical experiment collects data, which is then utilised to identify relationships that are then used as design requirements. The open-channel Rectangular Flume system used in this experiment. For flow monitoring, the 10.0 m long, 0.30 m wide, and 0.8 m tall straight-line pipe has reinforced glass sides. To induce steady-state input, it has three distribution plates at the inlet.
- Although the inflow flows through the distribution plates and becomes stable, a sluice gate was constructed at places 2.0 m out from the intake to reduce the inlet contraction impact. The cross section of a channel is shown in figure 5. The installed pump's maximum water delivery was $0.056 \text{ m}^3/\text{s}$.

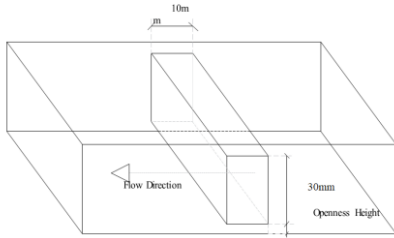


Figure 6. Schematic diagram of brick placed in the rectangular flume.



Figure 7. Occurrence of hydraulic jump in the rectangular flume.

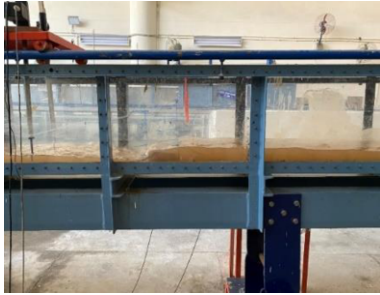


Figure 8. Brick placed in the flume and used as an obstruction to regulate the flow.



Figure 9. Side view of the hydraulic jump.

4. General Experimental Procedure

1. Conduct a general examination of the channel's condition.
2. Adjust the bed to a slope of 1° , 2° , 3° by turning the screw-jack at the channel outlet nine times.
3. Fill the reservoir with water with the help of water pump.
4. Disconnect the sluice gate and provide a consistent water flow to the channel, so that water can flow in the rectangular flume.
5. Keep track of the water flowing and measure the necessary metrics such as pre-jump depth (y_1), post-jump depth (y_2), length of the jump (L), velocity at int inlet (v_1) and velocity at the outlet (v_2).
6. Step 5 should be repeated for 15-20 times to measure the different parameters at different slopes.
7. Increase or decrease flow by adjusting the flow valve to make the jump in the rectangular flume.
8. Carry out the experiment with at least five distinct flow patterns that are y_1 , y_2 , L , v_1 and v_2 .
 - b = channel width = 30 mm = 0.030 m,
 - v_1 = measured pre-hydraulic flow velocity,
 - v_2 = measured post-hydraulic flow velocity,
 - y_1 = measured pre-hydraulic jump depth,

y_2 = measured post-hydraulic jump depth,

Q = discharge = $0.012 \text{ m}^3 \text{ sec}$,

$g = 9.806 \text{ m/s}^2$,

$y' =$ sequent depth ratio, i.e., $\frac{y_2}{y_1}$, where y_2 is equal to

$$y_2 = \frac{1}{2} y_1 \left(\sqrt{1 + 8F_1^2} - 1 \right)$$

5. Results and Discussions

The position and kind of hydraulic jumps were found to be affected by the sluice gate through extensive physical laboratory studies (table 1). As the opening of the sluice gate increases, the experimental findings reveal that the hydraulic jump moves closer to the gate, until the gate is eventually submerged.

According to the research, the type of hydraulic leap is linked to both the gates and the discharge. The Froude number (Fr_1) rises from 1.55 to 1.73 when discharge (Q) rises from 20 to 30 l/min in supercritical flow with constant y_2 and gate opening.

However, increasing post-jump depth (y_2) from 0.0755 m to 0.1038 m increases Fr_1 from 1.55 to 1.73 for continuous discharge and gate opening. The simulated post-jump depth (y_2) is roughly 5 cm bigger than the measured data when pre-jump depth (y_1) is 0.01 m and discharge (Q) is 50 l/sec.

Furthermore, a 0.45 m shift in energy is caused by a change in upstream and downstream depth.

Table 1. The table represents the pre jump depth (y_1), post jump depth (y_2), Froude numbers F_1 and F_2 , Energy loss (E_L) and Length of the jump.

S No.	Pre-Jump (y_1) in (m)	Post-Jump (y_2) in (m)	Sequent Depth y' in (m)	Froude Number (F_1)	Froude Number (F_2)	Length of the Jump (L_j) in (m)	Height of the jump $H=y_2-y_1$ (m)	Energy Loss (E_L)	E_1	E_L/E_1
1.	0.0375	0.0755	2.01	1.7345	0.04158	4.51	0.038	0.514	0.0947	5.92
2.	0.0380	0.0760	2	1.7330	0.04165	4.65	0.038	0.494	0.0952	5.84
3.	0.0387	0.0770	1.98	1.7324	0.04186	4.71	0.0383	0.507	0.0952	5.75
4.	0.0392	0.0775	1.977	1.7265	0.04900	4.85	0.0383	0.472	0.0962	5.55
5.	0.0400	0.079	1.975	1.7233	0.04042	4.92	0.0390	0.469	0.0972	5.30
6.	0.0412	0.0815	1.97	1.7161	0.0398	5.010	0.0403	0.497	0.0982	5.18
7.	0.0425	0.0828	1.94	1.7087	0.0392	5.062	0.0403	0.574	0.0992	5.02
8.	0.0432	0.0841	1.94	1.6954	0.0386	5.15	0.0409	0.478	0.1002	4.92
9.	0.0462	0.0873	1.88	1.6827	0.03702	5.26	0.0411	0.440	0.1032	4.85
10.	0.0485	0.0898	1.85	1.6545	0.03765	5.354	0.0413	0.424	0.1052	4.72
11.	0.0497	0.0922	1.855	1.6339	0.03637	5.467	0.0425	0.447	0.1062	4.55
12.	0.05102	0.0940	1.842	1.6015	0.03658	5.556	0.04298	0.414	0.1082	4.30
13.	0.05380	0.0970	1.80	1.5876	0.03685	5.742	0.0432	0.415	0.1102	4.18
14.	0.0548	0.0987	1.80	1.5624	0.03549	5.80	0.0439	0.419	0.1112	4.07
15.	0.0587	0.1038	1.76	1.5549	0.03570	5.852	0.0451	0.400	0.1152	3.92

(1) Froude no. vs sequent depth ratio: The Froude number was calculated using the ratio of the sequent depth ($\frac{y_2}{y_1}$). When the approaching Froude number and channel slope change, as seen in the figure 10, the ratios of the sequent depth vary. As the channel bed slope and approaching Froude number increase, the subsequent depth ratio appears to grow.

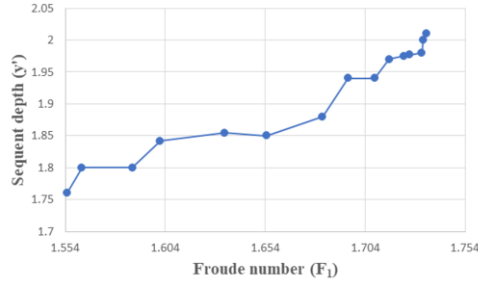


Figure 10. Experimental results of present study are plotted between sequent depth and the Froude Number (F_1).

(2) Length of jump vs height of jump: The Height of jump varies with Length of jump as shown in figure 11. The height of jump increases with the increase in the length of the jump.

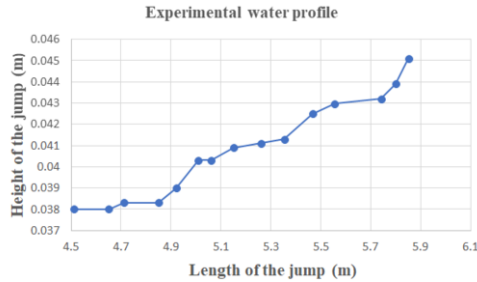


Figure 11. Experimental results of the water profile show the M_3 water profile.

(3) Energy Dissipation: Figure 12 demonstrates the relationship between the Froude number and the relative energy loss ($\frac{E_L}{E_1}$) for various kinds of slope. It is noticed that the relative energy loss non-linearly increases with the increase in the approaching Froude number from 1.5 to 1.7.

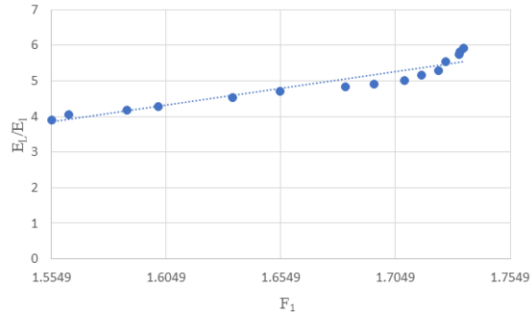


Figure 12. Present study of variation between the percentage of dissipation of energy.

(4) Hydraulic Jump Length: Figure 13 represents the relation between hydraulic jump length (L_j) and the sequent depth (y_2) for different slopes. It clears that the hydraulic jump length increases by 70% and the sequent depth increases by 100%.

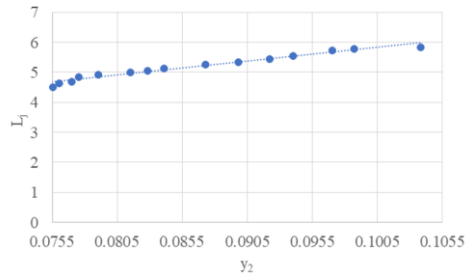


Figure 13. Present study variation of the hydraulic jump length (L_j) and the post Jump depth (y_2).

(6) Type of jump: Subcritical fluxes are indicated by Froude values ranging from 0.035 to 0.041 ($0.035 < Fr_2 < 0.041$). The Froude values range from 1.5549 to 1.7309 ($1.5549 < Fr_1 < 1.7309$) at the pre-hydraulic jump phase, indicating that the fluxes were supercritical and the jump in the hydraulic jump shown in the figure 14 acquired was an undular jump.

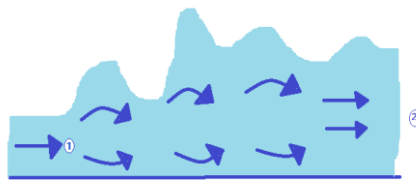


Figure 14. Shows the undular jump formed in the hydraulic jump.

6. Conclusion

A flow is performed in this study employing a waterway with a movable sluice gate to analyses the hydraulic jump. The following are some of the most important interpretations drawn from the study of hydraulic jump properties:

- In the results, several graphs show the change in the pattern of the properties of the hydraulic jump. These results have been concluded from the experiment performed in the laboratory.
- The Froude number has a great impact in the reduction of the jump length and subsequent depth. The amount of reduction was minimal for small Froude numbers, but it was higher for large Froude numbers. The hydraulic jump experiment was performed in an open channel flume which was rectangular in shape.
- Subcritical flows are denoted by Froude number values varying from 0.035 to 0.041 ($0.035 < Fr_2 < 0.041$). At the pre-hydraulic jump phase, the Froude number values vary from 1.5549 to 1.7309 ($1.5549 < Fr_1 < 1.7309$), indicating that the flow was supercritical and the jump obtained was an undular jump.
- For the level-bedded restricted flume, the energy loss due to hydraulic jump ranged from 0.37 to 0.48, denoting some energy gain with a faster rate of discharge through the flume. Because it indicates a supercritical flow area, the movable sluice gate may cause hydraulic problems.
- To address the constraints of this hydraulic jump experiment, subsequent hydraulic and numerical experiments will most likely incorporate more diverse settings. Overall, the experiment demonstrated that in an open channel flow, when the flow continuity exists, then the hydraulic jump takes place and the flowing liquid transition takes place which converts the flow into the subcritical flow from the supercritical flow.

Acknowledgement

Thanks to Delhi Technological University for providing us with all the resources.

References

- [1] Nallanathel M and Reddy VK. Study on hydraulic jump a review. Available: <http://www.acadpubl.eu/hub/>
- [2] Abdel-Mageed N. Effect of channel slope on hydraulic jump characteristics. *Physical Science International Journal*. 2015; 7(4): 223–233. doi: 10.9734/psij/2015/18527.
- [3] Elnikhely EA. Effect of staggered roughness elements on flow characteristics in rectangular channel. [Online]. Available: <http://www.ijret.org>
- [4] Imran HM and Akib S. A review of hydraulic jump properties in different channel bed conditions. 2013. <http://www.lifesciencesite.com><http://www.lifesciencesite.com><http://www.lifesciencesite.com>
- [5] Chern MJ and Syamsuri S. Effect of corrugated bed on hydraulic jump characteristic using SPH method. *Journal of Hydraulic Engineering*. 2013; 139(2): 221–232. doi: 10.1061/(asce)hy.1943-7900.0000618.
- [6] Habibzadeh A, Loewen MR and Rajaratnam N. Performance of Baffle blocks in submerged hydraulic jumps. *Journal of Hydraulic Engineering*. 2012; 138(10): 902–908. doi: 10.1061/(asce)hy.1943-7900.0000587.

- [7] Habibzadeh AS, Wu F, Rajaratnam N and Loewen MR. Exploratory study of submerged hydraulic jumps with blocks. *Journal of Hydraulic Engineering*. 2011; 137(6): 706–710. doi: 10.1061/(asce)hy.1943-7900.0000347. “10.1.1.369.945(1)”.
- [8] Carollo FG, Ferro V and Pampalone V. Hydraulic jumps on rough beds. doi: 061/ASCE0733-94292007133:9989.
- [9] Ead SA, Asce M, Rajaratnam N and Asce F. Hydraulic jumps on corrugated beds. doi: 10.1061/ASCE0733-94292002128:7656.
- [10] Ah HSM. Effect of roughened-bed stilling basin on length of rectangular hydraulic jump.
- [11] Hughes WC and Flack JE. Hydraulic jump properties over a rough bed.
- [12] Zhou Y, Wu J, Ma F and Hu J. Uniform flow and energy dissipation of hydraulic-jump-stepped spillways. *Water Science and Technology: Water Supply*. 2020; 20(4): 1546–1553. doi: 10.2166/ws.2020.056.
- [13] Ewah EG, Nyah EE, Antigha REE and Egbe JG. Experimental investigation of energy dissipation in hydraulic jump: A comparison of weir and level bedded constricted flume. 2018. [Online]. Available: <http://www.ijettjournal.org>
- [14] Cherhabil S and Debabeche M. Experimental study of sequent depths ratios of hydraulic jump in sloped trapezoidal channels. In 6th International Symposium on Hydraulic Structures: Hydraulic Structures and Water System Management, ISHS 2016. 2016; pp. 336–341. doi: 10.15142/T3610628160853.
- [15] Kumar S. Reproduction of water surface profile using ANSYS-CFD. 2019. [Online]. Available: www.ijesi.org/Volume8www.ijesi.org
- [16] Rajaratnam N. Hydraulic jumps. *Advances in Hydroscience*. 1967.
- [17] KUMAR GARG, SANTOSH Water Resources Engineering (Vol. II), Irrigation Engineering, and Hydraulic Structures, Khanna Publishers.
- [18] Chow, V. Te. McGraw-Hill New York, open-channel hydraulics (Vol. 1). 1959.
- [19] John D Jr. Computational fluid dynamics - The basics with applications. Anderson, McGraw Hill Education (India) Pvt. Ltd., New Delhi, 2012.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/357190395>

Exploiting Linguistic Information from Nepali Transcripts for Early Detection of Alzheimer's Disease using Natural Language Processing and Machine Learning Techniques

Article in *International Journal of Human-Computer Studies* · December 2021

DOI: 10.1016/j.ijhcs.2021.102761

CITATIONS

6

READS

65

7 authors, including:



Surendrabikram Thapa

Virginia Tech (Virginia Polytechnic Institute and State University)

21 PUBLICATIONS 206 CITATIONS

[SEE PROFILE](#)



Gnana K Bharathy

University of Technology Sydney

51 PUBLICATIONS 637 CITATIONS

[SEE PROFILE](#)



Mukesh Prasad

University of Technology Sydney

152 PUBLICATIONS 3,368 CITATIONS

[SEE PROFILE](#)

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



Intelligent Techniques for Cyber-Physical Systems [View project](#)



CEC-08 Special Session on "Evolutionary Computation and Neural Network for Combating Cybercrime" [View project](#)

Exploiting Linguistic Information from Nepali Transcripts for Early Detection of Alzheimer's Disease using the State-of-the-art Techniques of Machine Learning and Natural Language Processing

Surabhi Adhikari¹, Surendrabikram Thapa^{1,2}, Usman Naseem³, Priyanka Singh⁴, Angela Huo⁴, Gnana Bharathy⁵, Mukesh Prasad⁴

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

² Department of Computer Science, Virginia Tech, Blacksburg, Virginia

³ School of Computer Science, The University of Sydney, Sydney, Australia

⁴ School of Computer Science, FEIT, University of Technology Sydney, Sydney, Australia

⁵ School of Information, Systems and Modelling, University of Technology Sydney, Sydney, Australia

Abstract: Alzheimer's disease (AD) is considered as progressing brain disease, which can be slowed down with the early detection and proper treatment by identifying the early symptoms. Language change serves as an early sign that a patient's cognitive functions have been impacted, potentially leading to early detection. The effects of language changes are being studied thoroughly in the English language to analyze the linguistic patterns in AD patients using Natural Language Processing (NLP). However, it has not been much explored in local languages and low-resourced languages like Nepali. In this paper, we have created a novel dataset on low resources language, i.e., Nepali, consisting of transcripts of the AD patients and control normal subjects. We have also presented baselines by applying various machine learning (ML) and deep learning (DL) algorithms on a novel dataset for the early detection of AD. The proposed work incorporates the speech decline of AD patients in order to classify them as control subjects or AD patients. This study makes an effective conclusion that the difficulty in processing information of AD patients reflects in their speech narratives of patients while describing a picture. The dataset is made publicly available.

Keywords: Alzheimer's Disease, Deep Learning, Natural Language Processing, Machine Learning, Nepali Language, Low Resourced Language

1. Introduction

Alzheimer's disease (AD) is a progressive neurodegenerative condition affecting more than 50 million population across the globe. With someone developing the disease every three seconds, AD renders to be the most common form of dementia [1]. According to the 2018 World Alzheimer Report [2], the number of patients suffering from AD will cross the mark of 150 million by 2050, and the cost of treatment of Alzheimer's Disease is expected to cross 2 trillion US dollars by 2030. Currently, there aren't any approved drugs that can cure or completely stop how AD progresses [3]. However, there are some drugs and medications that can aid patients who are diagnosed in the earlier stages of AD. The early diagnosis of AD thus also helps in better management of the disease for both patients and caretakers. Hence, it is extremely necessary to find out the methods for the early diagnosis of AD for our aging society. AD patients show a wide range of symptoms due to the changes in the cortical anatomy [4]. One of the essential early indications of AD is cognitive impairment. Such cognitive impairments are mostly due to biological factors like atrophies in the various regions of the brain [5]. For example, atrophies in the left anterior temporal lobe impair naming tasks, such as picture description problems [6]. Such atrophies in the brain regions can be detected only by imaging techniques such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT) scans of the brain. Analyzing such imaging modalities would help us to classify the AD patients from the CN subjects, but analyzing them should be highly mediated by medical personnel. On the other hand, the patients with cognitive impairments show some visible symptoms like aphasia or limited ability in producing and understanding speech even for day-to-day tasks [7]. Such cognitive impairment is also often characterized by semantic memory deficits

and is mostly evidenced by naming impairment and the use of substitution words [8]. AD patients tend to reduce the amount of information, and such impaired subjects tend to use reduced working vocabulary. These impairments become noticeably evident with the progression of AD. Faber-Langendoen et al. [9], in the study of aphasia in AD patients, found out that 100% of the AD patients and 36% of the patients with mild cognitive impairment (MCI) had problems aphasia whose severity increased with increased severity of dementia. Such anomalies in linguistic features of speech produced by AD patients can be leveraged in building intelligent predictive systems for the diagnosis of AD in earlier stages.

Ahmed et al. [10] found that more than two-thirds of the participants showed significant changes in speech production way earlier before the medical diagnosis of AD. The speech patterns were significant as early as one year before the diagnosis of AD. Thus, speech can be a simple yet most prominent feature that can be used to build a powerful model for AD diagnosis. Kirshner et al. [11] found that all the participants had naming impairments despite absolutely normal speech in other respects. So, picture description tasks that heavily involve naming and identifying objects can be useful for learning the problems in speech. Also, thematic coherence, the ability of the speaker to maintain flow or theme in their speech, is heavily impaired in AD patients [12]. Their discourse lacked coherence as compared to CN (Control Normal) individuals. Currently, there are various neuropsychological tests available to assess the cognitive abilities of patients with AD. Some of the most widely used tests are Mini-mental State Examination (MMSE) [13], Rowland Universal Dementia Assessment Scale (RUDAS) [14], Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) [15], etc. The neuropsychological tests are mostly general, and since the memory impairment cannot be assessed with narrow criteria, the questions in such tests cannot assess cognitive abilities effectively. The same set of questions does not fit all the patients because the questions in which one patient may excel can be found difficult by other patients [16]. Also, most neuropsychological tests are used for an extended period and require psychologists or trained personnel to intervene throughout the assessment process. Similarly, ethics over the collection of personal information in neuropsychological assessment is also a problem to be looked upon. Moreover, AD diagnosis becomes difficult in times when the patients keep the symptoms to themselves only [2]. In such scenarios, the assessment of cognitive impairment can be done using self-generated speech on problems like picture description tasks [6].

When psychologists try to use naturally spoken language for the analysis of dementia or AD, it takes much time because of different linguistic patterns for different individuals. When computational linguistics is used for this purpose, the learning models trained on a large corpus of speech transcripts can show promising results as such models can be instrumental in learning the pattern in speech narratives of the subjects. Natural Language Processing (NLP) can hence be an alternative as well as a more appropriate technique for analyzing and interpreting the AD patient's speech. With computers becoming faster and faster, the speech narratives of subjects under study can be processed using NLP in real-time for the detection of AD. With the prospects of NLP being explored for mental illnesses like depression or schizophrenia, NLP can thus be significantly useful in improving the care delivery system for AD as well [17]. Apart from the difficulties in carrying out daily activities due to the severe symptoms of AD, it also poses an unprecedented burden and stigma upon those diagnosed with the disease. This study is also in the direction of lessening the stigma around AD by leveraging NLP tools. Due to the very limited amount of work done in the South Asian region and especially in low resources languages, the study employs a work-around for procuring data and underlying NLP experiments for the purpose. The dataset we have created is a translation from the pre-existing DementiaBank dataset in the English language. The

motivation behind this work is majorly the use of automation and mainly NLP for detection of AD in the low-resource language, as the advanced computational tools are still lagging in this region. This way of detecting AD is fast, cost-effective, and very accurate. If this approach can be demonstrated, it would provide an economic augmentation to both traditional assessments and primary data collection in AD detection on several under-resourced languages in various regions of the world.

The main contributions of the paper are:

- A novel manually annotated Alzheimer’s disease dataset for low resource language, i.e., Nepalese, consisting of 168 Alzheimer’s disease patients and 98 Control normal subjects, is presented. The dataset is made publicly available to the research community.
- An NLP-based framework is presented for the early detection of AD patients using Nepali transcripts and developed a visualization of content present in textual data. In addition to this, a word cloud of the most common words is presented to give qualitative analysis.
- The performance of different state-of-the-art machine learning-based textual classification mechanisms are presented with the baseline results.

Section 2 of the paper describes the works that have been done to detect AD from the linguistic features of the speech. The literature includes the work done using speech and transcripts for early detection of AD. Section 3 describes the methodology that has been used in this paper. The experimental results are discussed in section 4, and section 5 is the conclusion section that summarizes the findings of the paper, along with the future works that need to be done.

2. Related Works

In recent times, there have been various research going around in the task of the early diagnosis of AD using speech narratives of the subjects under study. In the last decade or so, much research is being conducted to figure out ways for the detection of AD using speech and linguistic features as the impairment of speech is one of the earliest symptoms of AD or Mild Cognitive Impairment (MCI). Thus, various machine learning (ML) methods are being used to detect anomalies in the speech narratives of subjects under study. Orimaye et al. [18] took syntactic, lexical, and n-gram based features for building the diagnostic model. The n-gram models had improved performance as compared to those models which used syntactic and lexical features alone. Using the top 1000 n-gram features, the model gave the Area Under Curve (AUC) value of 0.930, which was estimated using the Leave-Pair-Out Cross-Validation (LPOCV) technique. Also, Vincze et al. [19] used transcripts to classify patients with MCI and AD. The importance of morphological and speech-based features was highlighted in the research. Using only statistically significant features, Support Vector Machines (SVM) provided accuracy as high as 75%. These previously mentioned works used machine learning models. ML techniques require hand-crafted features. Such hand-crafted features vary extensively because of the different levels of expertise of the researchers in the diagnosis of AD. Also, hand-picked features are very easily outdated as the culture and language keep evolving continuously.

To overcome this drawback of using ML methods for the diagnosis of AD using transcripts of speech, some of the recent works have used intelligent deep learning models which can learn the intrinsic complexities of speech transcripts to automatically identify the linguistic features that reflect in narratives of AD patients with multiple levels of abstraction. Fritsch et al. [20] used a

neural network language model (NNLM) with Long Short-Term Memory (LSTM) cells to enhance the statistical approach of n-gram language models. The model was evaluated by measuring its perplexity. The scripts were evaluated by the model in a Leave-One-Out Cross-Validation (LOOCV) scheme. The perplexity values showed that the model could classify the AD and CN subjects with an accuracy of 85.6%. This suggests that the AD patients described the picture in an unexpected manner leading to unpredictable language structures that resulted in higher perplexity values. Chen et al. [21] proposed an attention-based hybrid network for automatic detection of AD. The hybrid model of attention-based Convolutional Neural Networks (CNN) and attention-based Bidirectional Gated Recurrent Unit (BiGRU) categorized the transcripts with an accuracy of 97.4%. The paper suggests that including attention mechanisms allowed the network to emphasize the decisive features of the subjects. Much work has been done in the English language, and some of the experiments have resulted in state-of-the-art (SOTA) models. The linguistic components of the English language, which affect the classification of CN vs. MCI vs. AD, are well explored. On the other hand, the research for the early detection of AD using linguistic features in languages other than English is not well explored. According to WHO, among the total number of dementia patients worldwide, 58% of the patients are from low and middle-income generating countries [22]. Building models only in the English language would leave a considerable fraction of the population without diagnostic tools that use NLP.

Much work has been done in major languages like Mandarin Chinese, German [23], Hungarian [24], etc. For instance, Liu et al. [25] used a dependency network approach to examine syntactic impairments of Chinese AD patients. Most of the AD patients showed regular syntactic impairments, which is evidence that there is language deterioration. Apart from Chinese, linguistic and acoustic features have been explored in various other languages like German [23], Hungarian [24], etc. There have also been researches on how spontaneous speech in various languages can be used in the analysis of AD through speech. Weiner et al. [23] used spontaneous conversational speeches in the German language to build models. Using Linear Discriminant Analysis (LDA) classifier with singular value decomposition, the researchers could get an F1-score of 0.800. They, however, used models for three classes classification viz. Control Normal (CN), Aging-associated Cognitive Decline (AACD), and Alzheimer's Disease (AD). Similarly, low resource language researchers have researched this domain using telephonic conversations also. Khodabakhsh et al. [26] used 10 minutes of telephonic conversations recorded using microphones for Turkish speakers. The texts were manually transcribed, and the learning algorithms were used. With conversational recording transcripts of 20 AD patients and 20 healthy individuals, the models were built. The features like hesitation and puzzlement features, Part of Speech (POS) based features, unintelligible word rate, complexity features like phonemes per word, etc., were used. They used algorithms like Support Vector Machine (SVM), LDA, and decision trees. With the LOOCV scheme, the researchers were able to get accuracy as high as 90%. It can be seen that there have been many initiatives to build models in multiple languages. However, there has not been any research in this domain in the South Asian regional languages. For a low-resource language like Nepali, where there is very limited research in NLP, this work in the detection of AD using speech narratives by exploiting linguistic cues is the first of its kind.

3. Methodology

The proposed framework for the experiment is as shown in Fig. 1. The process starts with data collection, which involves extracting the transcripts from the dementia bank and translating them into the Nepali language. The text is further pre-processed, and features are extracted. Similarly,

the models are trained and tested using a 10-fold stratified cross-validation scheme. After that, the various performance measures like precision, recall, accuracy, and F1-score are calculated. The components of the framework are explained below in great detail.

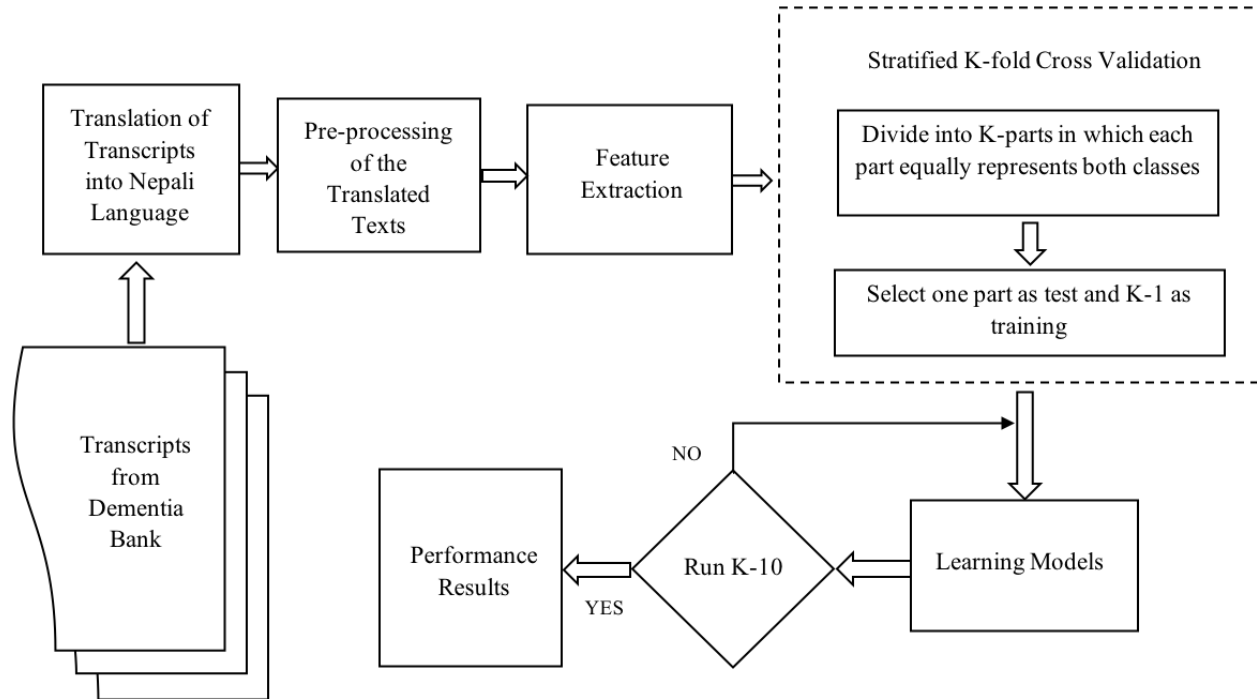


Fig. 1 Flow diagram of the Framework Used in Experiment

3.1 Data Collection

In this study, DementiaBank's Pitt Corpus has been utilized. The DementiaBank is one of the largest available datasets of audio recordings and transcripts of subjects who participated in the research conducted by Becker et al. [27] of the University of Pittsburgh School of Medicine. The recordings were manually transcribed by the CHAT protocol [28]. CHAT stands for (Codes for the Human Analysis of Transcripts), which was the format used for the CHILDES (Child Language Data Exchange System).

This study uses the transcripts of audio recordings for the Cookie Theft picture description task explicitly. The cookie theft description task was first used by the Boston Diagnostic Aphasia Examination protocol and hence mentioned as the Boston Cookie Theft picture description task in literature [29]. In this task, participants were asked to describe the kitchen scene, as shown in Fig. 2. There were 292 participants in this study conducted by the University of Pittsburgh School of Medicine. Among 292 participants, 194 had at least one sort of dementia. Some of the participants had several recording sessions. Thus, the dementia category consists of 309 transcripts. Since this study deals with AD diagnosis, only the transcripts of patients with AD are taken. So, 255 CHAT transcripts belonging to 168 AD patients were taken. Similarly, there were 244 transcripts from 98 CN participants used in this experiment. Furthermore, all the participants of Becker's [27] study were over the age of 44 and had a minimum MMSE score of 10. The demographic information about the subjects under study can be shown in Table 1. The mean of attributes, along with their

standard deviation, can also be found in Table 1.

Table 1. Demographic information of subjects from Dementia Bank

Attributes	Control Normal (CN)	Alzheimer's Disease (AD)
No. of participants	98 CN subjects	168 AD patients
Gender	31M / 67F	55M / 113F
Age	64.7 (7.6)	71.2 (8.4)
Education	14.0 (2.3)	12.2 (2.6)
MMSE	29.1 (1.1)	19.9 (4.2)

Originally, the available transcripts of the recordings are available in the English language, which was translated into the Nepali language for this study. The translations were done by two native Nepali language speakers who had at least 13 years of formal education in the Nepali language. Translating the entire dataset took around eighty-four hours. After the entire translation, they were again sent for verification to an independent linguistic expert, who assessed and verified the adequacy of the translations. We chose manual translation, as this has a better chance of capturing cultural nuances required for the target language. Justification in the literature will be shown in the Discussion section. Here, we will urge the readers to take this at face value and move forward.



Fig. 2 Cookie Theft Picture

An example of the manual translation of each category is shown in Table 2(a). Words such as uhm, uhh, and other pause words were not removed and translated as it is. Such words are not removed from translation to retain more accurate transcripts of the actual recordings. Khodabakhsh [30]

suggested that such filter sounds like uhm, uhh, etc., are used more often by AD subjects and form a significant feature in their speech. AD patients tend to use longer pauses than cognitively normal individuals. The repetitions, linguistics, and syntactic errors and the words that depict confusion of the AD participants have been translated as they are to retain the originality. The transcripts have been translated in a way that maximum linguistic characteristics are preserved. However, the annotations in the transcripts, such as clears throats, laughs, etc., have not been included as they are not a part of the linguistic feature used by the subjects. The ten most frequently used words are given in Table 3, with their number of occurrences.

We have also developed a dataset by machine translation (google translate) for comparison. The manual and machine translations are being compared in Tables 2(a) and Tables 2(b). The translations from google translate did not show the accurate translation of the text. Moreover, in many of the NLP tasks that require annotations, machine translation fails to show accurate translation at par with human translation. Even though the translations involving deep learning methods provide substantial advantages, they still lack human performance on data that require cultural nuances to be preserved [31]. Hence, manual translation was incorporated into the study for better perseverance and assertion of the general tone of the texts. The word clouds of the English and Nepali texts are shown in Fig 3. Fig. 3 (a) shows the word cloud of the transcripts of CN individuals, and Fig. 3 (b) shows the word cloud of the transcripts of AD patients in the English language. Similarly, Fig. 3 (c) and Fig. 3 (d) represent the transcripts in the Nepali language by CN and AD subjects, respectively.

Table 2 (a). Examples of the Manual Translation of the DementiaBank

CHAT ID	English Sentence	Nepali Sentence
015-2.cha	you have two children and the boy is on a stool getting to the cookie jar. and the stool is tilting over and he's probably going to fall. his mother in the meantime is wiping dishes, looking out what is obviously the kitchen window. she has the water on in the sink and the sink is overflowing. there are two cups and one plate sitting on the sink. the little girl is laughing at the little boy who's getting into the cookie jar and is going to fall.	दुईजना बच्चाहरू छन् र केटो कुर्सीमा चढेर कुकी जारबाट कुकी लिन खोजिरहेको छ कुर्सी बाझिएको छ र ऊ सायद लड्ने वाला छ उसकी आमा यो समयमा भाँडा पुस्टै छे ऊ झ्याल बाहिर हेरिरहेकी छे सिंकमा पानी छ र पानी भरिएको छ सिंकमा दुईवटा कप र एउटा प्लेट छ सानी केटी सानो केटो माथि हाँसिरहेकी छे ऊ कुकी लिँदै छ र लड्ने वाला छ
472-0.cha	the boy and the girl are playing and he's gonna fall down off the ladder. and the mother's washing the dishes and it's flying out over the sink down to the floor. what else do you want me to tell you whatever you see happening. yeah that's it.	केटो र केटी खेलिरहेका छन् अनि ऊ भन्याङ बाट लड्नेवाला छ अनि आमा भाँडा माझिरहेकी छे अनि त्यो सिंकबाट माथी उडिरहेको छ, भुईँमा पुगिरहेको छ अरू के चाहन्छौ कि म भनौं भनेर.. यति हो

Table 2 (b). Examples of the Google Translate Translations of the DementiaBank

CHAT ID	English Sentence	Nepali Sentence
015-2.cha	you have two children and the boy is on a stool getting to the cookie jar. and the stool is tilting over and he's probably going to fall. his mother in the meantime is wiping dishes, looking out what is obviously the kitchen window. she has the water on in the sink and the sink is overflowing. there are two cups and one plate sitting on the sink. the little girl is laughing at the little boy who's getting into the cookie jar and is going to fall.	तपाईंका दुई बच्चाहरू छन् र केटा स्टूलमा कुकी जारमा पुगिरहेको छ। र मल माथि झुकेको छ र ऊ सायद खस्दैछ। यस बीचमा उनकी आमा भान्साकोठा पुछिरहेकी छिन्, स्पष्ट रूपमा भान्साको झ्याल हेर्दै। उसको सिङ्कमा पानी छ र सिङ्क भरिएको छ। त्यहाँ सिङ्कमा दुई कप र एउटा प्लेट छ। सानो केटी कुकीको भाँडोमा पसेको सानो केटालाई देखेर हाँस्दै छ।
472-0.cha	the boy and the girl are playing and he's gonna fall down off the ladder. and the mother's washing the dishes and it's flying out over the sink down to the floor. what else do you want me to tell you whatever you see happening. yeah that's it.	केटा र केटी खेलिरहेका छन् र ऊ भर्षाडबाट तल खस्नेछ। र आमाले भाँडा धुँदै हुनुहुन्छ र यो सिङ्क माथि भुईँमा उडिरहेको छ। अरु के चाहन्छौ म तिमीलाई जे भइरहेछ देख्छु। हो त्यो हो।

Table 3: top 10 most used words in the transcript

Words	Number of Appearances
कुकी (cookie)	1092
भाँडा (utensil)	588
पानी (water)	580
केटो (boy)	448
आमा (mother)	384
केटी (girl)	326
कुर्सी (chair)	276
बाहिर (outside)	252
सानी (small)	222
सायद (maybe)	213



Fig 3. The word clouds of the English and Nepali texts

3.2 Data Preprocessing

The preprocessing step usually includes the removal of filter words, unnecessary noise, and unwanted information that does not add any value to the true meaning of the text [32]. In the English language, a major preprocessing step would be to make all the words uppercase or lowercase. The Nepali language is a case insensitive language. Hence, it does not require any such conversion. In this study, as a preprocessing step, the punctuation marks like commas, semicolons, etc., that do not add any semantic meaning to the text are removed. Another general practice in classification tasks using NLP involves removing stop words, which usually helps improve performance metrics. Since AD vs. CN is also a text classification task, anyone would think of proceeding with preprocessing the text by removing the stop words. The domain knowledge of AD helps tackle the ways to preprocess the CHAT transcripts used in this study. The AD patients tend to repeat stop words like ‘and,’ ‘therefore,’ etc. more often, and in this experiment, stop words are not removed since they preserve the linguistic characteristics of AD patients [30].

3.3 Feature Extraction

Text feature extraction is the process of extracting a list of words and creating a vocabulary from the text data [33]. These words are transformed into a feature set that a classifier can use. In this experiment, word statistics-based feature extraction techniques have been used. Vectorization techniques are used to transform the words into vectors. They give positional weights to the words used in the text data. Similarly, word embeddings are a way of transforming words into vectors by capturing the similarity between words. Words with similar meanings appear in the same feature space. The various feature extraction techniques used in this experiment have been discussed below:

3.3.1 Vectorization Methods

The experiment uses two popular vectorization methods, namely CountVectorizer and Term Frequency Inverse Document Frequency (TF-IDF). CountVectorizer is used to build a dictionary of known words from the test dataset. It is also used to encode the new documents using the vocabulary [34]. CountVectorizer tokenizes and creates a respective vector representation of each word fed to a machine learning model. TF-IDF is another popular vectorization technique for generating vector representations of the text [35]. TF-IDF represents the importance of a word to a document. It does so by being able to count the number of occurrences. TF-IDF punishes the words that are used very often in the documents, hence being able to give more weightage to the words that are more relevant and important to a particular document.

3.3.2 Word embeddings

After the text is preprocessed, real-valued vectors are assigned to words or phrases using word embeddings. Word embeddings are based on the idea that if features have similar meanings, it is useful to represent the features to depict this similarity [36]. Bengio et al. [37] proposed a probabilistic neural model where the words in the vocabulary were mapped to a distributed word feature vector. The feature vector represents several aspects of the word. These features are smaller than the size of the vocabulary. This study makes use of the two most efficient word embeddings viz. Word2Vec and fastText. Both the pre-trained and domain-specific Word2Vec [38] and fastText [39] models are trained to produce embeddings.

Domain-specific Embeddings: Domain-specific embeddings are trained on the dataset being used. It has been found that the pre-trained word embeddings perform very well in a large text corpus, but in sparse and specialized texts, the pre-trained word embeddings generally fail to produce appropriate vectors [40]. In this study, 300-dimensional embeddings have been used for both Word2Vec and fastText embeddings, and the maximum length is set to 270. Gensim [41] library is used to generate Word2Vec and fastText models from the text used in the study.

Pre-trained Embeddings: The pre-trained Nepali Word2Vec model created by Lamsal [42] is used in the study. This pre-trained Word2Vec model has 300-dimensional vectors for more than 0.5 million Nepali words and phrases. The embedding dimension is 300, and continuous bag-of-words (CBOW) architecture was used to create the given Word2Vec model. Similarly, for the pre-trained fastText embeddings, the pre-trained word vectors trained on Common Crawl and Wikipedia using fastText were used [39]. The model was trained by using CBOW with position-weights, in dimension 300, with character n-grams of length 5, a window of size 5, and 10 negatives.

3.3 Learning Models (Classifiers)

For the classification of the transcripts of CN and AD patients, some learning models should be

used. In this paper, both machine learning models and deep learning models were used to find the better model that would classify the transcripts with greater accuracy.

3.3.1 Machine Learning Baselines

Since the work of such classification in the Nepali language is the first of its kind, machine learning baselines are taken to evaluate the performance of machine learning algorithms in the delineation of the transcripts of AD patients from that of CN subjects. The machine learning algorithms like Decision Tree (DT) [43], K-Nearest Neighbors (KNN) [44], Support Vector Machines (SVM) [45], and Naïve Bayes (NB) [46] were used. Also, ensemble learners like Random Forest (RF) [43], AdaBoost [47], and XGBoost (XGB) [48] were used. Apart from the traditional vectorization techniques, such as CountVectorizer and TF-IDF vectorizer, word embeddings were also used for vectorizing the input text to feed them to the machine learning model.

3.3.2 Deep Learning Models

The deep learning models have recently shown very promising results in text classification, especially when the classification tasks deal with intrinsic and complex details of the linguistic features in the text. In our experimentation, three deep learning models have been used, viz, Convolutional Neural Network (CNN) [49], Bidirectional Long Short-Term Memory (BiLSTM) [50], and a combination of CNN and BiLSTM [51].

Convolutional Neural Network: Convolutional Neural Network (CNN) is a deep neural network architecture that uses layers with convolving filters. CNNs have been traditionally used for computer vision for identifying images. However, CNNs have also proven to be significantly useful for NLP. They have been used for various NLP tasks such as semantic parsing, search query retrieval, sentence modeling, and other traditional NLP tasks. The convolving filters, as well as applying max-pooling extract relevant n-gram features of the texts used. The input of the convolutional layer is the vector produced by word embeddings. A one-dimensional Convolutional Neural Network has been used in this study. As a 300-dimensional embedding with a maximum length of 270 has been used in this study, the input size is a matrix of size 270x300. Four convolutional layers with ReLu as the activation function have been used, and after every two layers, a max-pooling [52] of size three was done. For kernel regularization, L2 regularizers have been used. The optimizer used is Adam [53]. After flattening the convolutional layers, the output is connected to a fully connected dense layer. The softmax function is used in the output layer to predict the probabilities of the CN and AD categories. The number of epochs and batch size has been fixed to 20 and 50, respectively, for all embeddings.

Kim's Architecture: Apart from the CNN model mentioned above, the famous Kim's CNN architecture [54] has also been used. In Kim's architecture, after every convolutional layer, max-pooling is applied. In this experiment, three convolutional layers with tanh as the activation function have been used, and after each layer, a max pool of filter size three has been applied. After the last max pool filter, the flatten layer reshapes the input size, followed by the dropout layer with a rate of 0.5. The dropout layer randomly sets inputs to 0 and prevents overfitting. The output layer has softmax as the activation function that transforms the results into probabilities of each class. The number of epochs and batch size has been fixed to 20 and 50, respectively.

BiDirectional Long Short-Term Memory (BiLSTM): Unlike Long Short-Term Memory (LSTM) [55], in BiLSTMs, the signal propagates in both directions, i.e., backward and forward. BiLSTMs train first on the input sequence and then on the reversed input sequence. The forget, input, and

output gates and the cell states decide what information to throw away, update the cells, and then produce the output by carrying only the relevant information. In this work, four BiLSTM cells with 16, 8, 4, and 2 nodes subsequently and tanh as the activation function have been used. The input is the same as the convolutional layer, i.e., 270x300 dimensional vector of word embeddings. After the first BiLSTM layer, a dropout with a 0.5 rate has been used for regularization. After the three BiLSTM layers, again, a dropout of 0.25 has been used. The output of the BiLSTM cell has been connected to a dense layer with four nodes and ReLU as the activation function. The output layer has softmax as the activation function in order to predict probabilities for the two categories. To prevent overfitting, L2 regularizers have been used. Similarly, for optimization, an Adam optimizer has been used. The number of epochs and batch size has been fixed to 20 and 50, respectively, for all embeddings.

CNN with BiLSTM cells: CNNs learn the local features of the text, and RNNs learn long-term dependencies. Combining these architectures can better perform in various NLP tasks such as sentiment analysis and text classification [56]. In this experiment, four convolutional layers and two BiLSTM cells have been used. The word embeddings are fed to the convolutional layer. After every two convolutional layers, a max-pooling of size three has been applied. To prevent overfitting, L2 regularizers have been used in both networks. Tanh has been used as the activation function for the BiLSTM cells. After the first BiLSTM cell, batch normalization [57] has been done. Adam optimizer has been used. The output of the BiLSTM cell has been connected to a fully dense layer with ReLU as the activation function and twenty nodes. One more dense layer has been added with ReLU as the activation function with ten nodes. The softmax function in the output layer transforms the vectors to predict the category of the transcripts. The number of epochs and batch size has been fixed to 20 and 50, respectively.

Deep Learning Models with Attention Mechanisms: Attention mechanisms are used in encoder-decoder architectures to attend to the encoder and previous hidden states. With an input sentence and all the associated hidden states, attention layers decide what part of the input was most relevant and useful with each output instance. Attention preserves the context from beginning to end hence achieving great results on various NLP tasks such as machine translation [58], text summarization [59], text classification, etc. All the deep learning models used in this study have also been trained with an attention layer [60]. Apart from attending to the encoder and previous hidden states, attention can also be used to get a distribution over features, such as the word embeddings of a text [61]. The attention used in this study is the multiplicative self-attention layer because of its space efficiency and less operation time. Self-attention [62] is used to extract the relevant features by enabling it to attend to itself. The architecture of the models is the same as described above, with only an attention layer after the first layers in every model.

Deep Learning Models with Vectorization Techniques: Apart from pre-trained and domain-specific word embeddings, deep learning classifiers were also fed with the vectorized texts done by CountVectorizer and TF-IDF vectorizer as inputs. The input was a matrix of dimensions (499, 270). The remaining layers of the architecture of the deep learning models were kept the same as described above.

3.5 Performance Measures

In all the architectures afore-mentioned, binary cross-entropy has been used as the loss function. The validation has been done using stratified k-fold cross-validation. The stratified K-fold cross-validation is a variation of k-fold cross validation that returns stratified folds in which the

percentage of samples of each class is preserved. After stratified 10-fold cross-validation, the performance of the proposed architectures has been measured using four evaluation metrics viz. accuracy (acc), precision (pre), recall (rec), and F1-score as shown in equations (1)-(4).

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

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

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

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

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative respectively.

4. Results and Discussion

The baseline is established with various machine learning algorithms. TF-IDF, CountVectorizer (CV), Word2Vec, and FastText were used to convert the text document into vectors for the experiment. The results of the machine learning baselines with the TF-IDF and CountVectorizer are shown in Table 4. For machine learning models, the Naive Bayes classifier performed the best for both the vectorization techniques. With CountVectorizer, the model had an F1-score of 0.940. With TF-IDF vectorization as well, the model was able to achieve an F1-score of 0.940. TF-IDF seemed to perform slightly better when vectorization methods are compared than CountVectorizer for different machine learning models. A possible explanation for this is that TF-IDF, instead of just representing words with vectors in terms of their number of appearances, balances the most frequent words by giving them less weightage. Rarer words common in a particular class would be scored higher, eventually leading to better performance of models.

Similarly, with domain-specific Word2Vec word embeddings, the decision tree performed the best with an F1-score of 0.937. As far as pre-trained word embeddings are concerned, pre-trained Word2Vec performed the best with the XGBoost algorithm giving an F1-score of 0.828. On the other hand, with pre-trained fastText, the SVM classifier outperformed other models with an F1-score of 0.934. It can be inferred from the comparison of vectorization techniques with word embeddings that vectorization techniques performed better than word embeddings with machine learning models. The reason behind this is that the data corpus was small to train the word embeddings. Hence, the similarity between words is not captured well. The results with Word2Vec and FastText embeddings with the machine learning models are shown in Table 5 and Table 6, respectively.

Table 4. ML Classifiers with CV and TF-IDF

ML Classifier	TF-IDF				Count Vectorizer (CV)			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
DT	0.900	0.903	0.900	0.900	0.854	0.855	0.854	0.854
KNN	0.910	0.918	0.910	0.909	0.870	0.892	0.870	0.868
SVM	0.936	0.939	0.936	0.936	0.902	0.907	0.902	0.902
NB	0.940	0.944	0.940	0.940	0.940	0.945	0.940	0.940

RF	0.934	0.937	0.934	0.934	0.936	0.939	0.936	0.936
ADB	0.886	0.889	0.886	0.886	0.904	0.907	0.904	0.904
XGB	0.926	0.929	0.926	0.926	0.920	0.924	0.920	0.920

Table 5. ML Classifiers with Word2Vec

ML Classifiers	Domain-Specific Word2Vec				Pre-trained Word2Vec			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Decision Tree	0.938	0.949	0.938	0.937	0.695	0.698	0.695	0.694
KNN	0.718	0.780	0.716	0.694	0.547	0.564	0.547	0.464
SVM	0.940	0.947	0.940	0.931	0.764	0.771	0.764	0.760
Naïve Bayes (Gaussian)	0.891	0.930	0.892	0.890	0.529	0.582	0.529	0.414
Random Forest	0.902	0.912	0.902	0.901	0.788	0.797	0.782	0.785
AdaBoost	0.890	0.898	0.890	0.889	0.754	0.766	0.754	0.751
XGBoost	0.918	0.924	0.918	0.917	0.826	0.836	0.826	0.828

As far as the deep learning models are concerned, they seem to have performed better than the machine learning models in the experiments. The initial experiments with deep learning models showed that with domain-specific Word2Vec, Kim’s Architecture had the best F1-score of 0.964. Similarly, with pre-trained word embeddings, pre-trained fastText with the BiLSTM model outperformed other models with pre-trained embeddings with an F1-score of 0.887. The deep learning models performed slightly better with domain-specific word embeddings than pre-trained embeddings. As domain-specific word embeddings are formed from the data corpus, it can capture the domain words well and perform better. The dataset is based on the cookie theft description task, and hence it contains words related explicitly to the problem than general words. The results of the deep learning models with domain-specific and pre-trained word embeddings are shown in Table 7.

Table 6. ML Classifiers with fastText

ML Classifiers	Domain-Specific fastText				Pre-trained fastText			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Decision Tree	0.674	0.683	0.673	0.668	0.794	0.799	0.794	0.792
KNN	0.599	0.639	0.599	0.56	0.908	0.913	0.908	0.907
SVM	0.739	0.747	0.739	0.737	0.934	0.944	0.934	0.932
Naïve Bayes (Gaussian)	0.531	0.575	0.531	0.419	0.523	0.56	0.523	0.400
Random Forest	0.7876	0.797	0.788	0.7855	0.914	0.92	0.914	0.913
AdaBoost	0.755	0.762	0.755	0.753	0.918	0.925	0.918	0.917
XGBoost	0.826	0.837	0.826	0.824	0.914	0.924	0.914	0.912

Table 7. Deep Learning Models with word embeddings

Deep Learning Models		Word2Vec				fastText			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Domain-Specific	Kim's CNN	0.964	0.965	0.964	0.964	0.936	0.937	0.936	0.936
	CNN	0.950	0.952	0.950	0.950	0.928	0.930	0.928	0.928
	BiLSTM	0.946	0.948	0.946	0.946	0.900	0.910	0.900	0.897
	CNN + BiLSTM	0.962	0.964	0.962	0.962	0.928	0.931	0.928	0.927
Pretrained	Kim's CNN	0.828	0.834	0.828	0.827	0.870	0.880	0.870	0.867
	CNN	0.861	0.865	0.861	0.861	0.866	0.877	0.866	0.864
	BiLSTM	0.872	0.884	0.872	0.869	0.888	0.892	0.888	0.887
	CNN + BiLSTM	0.872	0.880	0.872	0.871	0.756	0.839	0.756	0.727

Table 8. Deep Learning Models with vectorizers

Deep Learning Models	CountVectorizer				TF-IDF Vectorizer			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Kim's Architecture	0.897	0.900	0.897	0.897	0.847	0.852	0.847	0.846
CNN	0.793	0.798	0.793	0.773	0.731	0.755	0.731	0.727
BiLSTM	0.735	0.738	0.735	0.733	0.738	0.743	0.738	0.726
CNN+BiLSTM	0.83	0.849	0.832	0.831	0.732	0.796	0.732	0.717

Apart from the initial experiments with word embeddings for the deep learning models, they were also trained with vectorization techniques. As done with machine learning models, CountVectorizer and TF-IDF were used to vectorize the words for deep learning models. In this case, CountVectorizer outperformed TF-IDF with Kim's CNN architecture. The model was able to achieve an F1-score of 0.897. Kim's CNN contains max-pool filters after each convolution operation. This potentially extracts just the relevant features with reducing dimensionality simultaneously. The performance of the models with vectorizers is shown in Table 8.

Attention mechanisms were also applied to deep learning models in the experiments. With attention mechanisms, CNN with Word2Vec showed the best performance with an F1-score of 0.968. From the results obtained, it can be seen that the attention mechanism gave the best results, implying that more weightage was given to those words which carried more importance in the sentence. Also, CNN outperformed the other models with attention giving an idea that the features captured by the model were just the relevant ones, and only they were attended to. With attention as well, domain-specific Word2Vec performed better than pre-trained word embeddings. The results of the deep learning models trained with word embeddings and attention are shown in Table 9.

Table 9. Attention with DL models and word embeddings

Deep Learning Models		Word2Vec				fastText			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Domain specific	Kim's Architecture	0.954	0.955	0.954	0.954	0.924	0.927	0.923	0.923
	CNN	0.968	0.969	0.968	0.968	0.932	0.933	0.932	0.932
	BiLSTM	0.956	0.959	0.956	0.956	0.923	0.934	0.924	0.929
	CNN+BiLSTM	0.962	0.962	0.962	0.962	0.922	0.928	0.922	0.921
Pre-Trained	Kim's Architecture	0.890	0.907	0.900	0.900	0.922	0.924	0.922	0.922
	CNN	0.902	0.901	0.902	0.901	0.902	0.904	0.907	0.902
	BiLSTM	0.913	0.918	0.913	0.913	0.764	0.825	0.784	0.767
	CNN+BiLSTM	0.890	0.904	0.890	0.887	0.660	0.711	0.660	0.572

Table 10. Attention with vectorizer in DL models

Deep Learning Models		CountVectorizer				TF-IDF			
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Kim's Architecture		0.701	0.712	0.701	0.702	0.627	0.627	0.620	0.623
CNN		0.765	0.763	0.765	0.766	0.699	0.699	0.689	0.632
BiLSTM		0.629	0.634	0.629	0.616	0.723	0.723	0.717	0.729
CNN+BiLSTM		0.678	0.682	0.678	0.685	0.713	0.713	0.710	0.721

The word embeddings and deep learning models used with vectorization techniques were also trained by using an attention layer. The CNN model with CountVectorizer had the highest F1-score of 0.766 in this experiment. From the results obtained, it can be inferred that attention methods with vectorization did not perform well as they did with word embeddings. When attention layers are applied, attending to features with vectors as the representations of the number of appearances of words, does not perform well. Hence, with attention, the vectorization techniques had lower F1-scores overall. The results with attention-based deep learning models with vectorization techniques are shown in Table 10.

It can be seen from the results that the best performing model is the attention-based CNN with domain-specific Word2Vec. This model can be utilized in making a clinician-friendly application for helping them with identifying Alzheimer's disease in its earliest stages. A pipeline with a mechanism for speech synthesis could also be developed with the given methodologies for better detection of the disease.

5. Discussions and Future Works

5.1 Discussions

Computational methods are very significant for clinical research and enable healthcare professionals to make clinical decisions about disease identification. There are many emerging

success stories in NLP applications in the English language. This motivates researchers to bring about results in the clinical domain, other than the English language. Hence, to put another dimension, the dataset in the clinical research in the domain of Alzheimer's disease in the Nepali language is substantial for further research. The work presented in this paper contributes to the growing body of research in the area of clinical applications in AD, albeit a preliminary one.

The model provides a methodology for early detection of AD using translated corpus in a low resource language, which in this case is Nepali. With such a model, we would be able to label, identify, and provide an early detection system in low-resourced languages. For this project, we were able to reach a reasonable accuracy even without specifically collecting data in the local language. The approach makes use of an AD corpus in English and translates the corpus into Nepali using human translators. We chose manual translation, as this has a better chance of capturing cultural nuances required for the target language. As there is very little amount of text data in languages such as Nepali for NLP tasks, manual translation was an effective method for collecting data. Even in the best of times, manual data collection in the field requires a lot of resources. Now, it is even more challenging to manually collect the dataset from patients in Nepal, given the ongoing pandemic, and hence the available data were used. This might seem like a contradiction, and also raises the question of why not use machine learning to do the translation at all. After all, using machine translations would be a full machine learning approach to detecting Alzheimer's using natural language processing. However, the literature search did not support this approach with the amount of data available.

Earlier, Guzman et.al. [63] experimented with machine translations of the FLORES dataset, and the results were poor, indicating more annotations and preparatory work might be needed before embarking on machine translations. On the other hand, there have been hundreds of years of successful manual translations [64][65] even to low-resource languages such as Nepali. There are also previous studies that supported manual translation as a better option at least in the early stages of working with low resource languages. Mohammad et al. [31] did a sentiment analysis in the Arabic language utilizing the available English texts and showed competitive results with the manually translated Arabic data. Similarly, Balahur et al. [66] did a sentiment analysis on four different languages, namely, Italian, Spanish, German, and French, by translating English data. They considered the manual translation as the gold standard of their datasets. Some machine translations have been attempted between English and Nepali [67]. In terms of low-resource datasets, DARPA programs like LORELEI [68] and the Asian Language Treebank project [69] have collected and introduced translations on several low-resource languages. However, these are still in the early stages, the coverage is still low, and do not include Nepali. Also, such machine learning translation systems need improvements before they can be employed on their own. There are reports of about 68% accuracy on TDIL (Technology Development for Indian Languages). Some studies employed English-Nepali parallel corpus for machine translation. Therefore, it would be some time before we can use the machine-based approach entirely and eliminate manual translations.

The Nepali dataset thus derived by manual translation works well with early detection of Alzheimer, and is a good candidate for creating a baseline for detecting Alzheimer's disease in the Nepali language since there is no available data for use for the purpose. As mentioned earlier, the translations were later verified by a linguistic expert who is currently working at the University of Auckland. The expert corroborated that the translations preserved the emotions of the participants. Moreover, the overall intonation of the text has been maintained, and hence there were no cultural

inconsistencies. Therefore, the experiments were finally conducted with the assurance from the expert about the dataset. The expert review should not be surprising, as manual translators could address both the message as well as cultural meaning [70]. The reasons for better performance through translations could be postulated based on a number of factors. Firstly, English has come to be used and studied worldwide, and Nepali speech has a significant degree of code-switching that occurs. Code-mixing [71] is common in Nepal, and code-switching between English and Nepali is fairly common among urban and educated Nepali speakers [72]. Medical professionals tend to fall in these categories. So much so, that the Gurung [72] reports extensive code-switching and code-mixing, argues that Nepali-English mixed language has emerged as a dialect in the Nepali speech community through the recurrent use of the English elements in the Nepali conversation.

In addition, Nepali is one of the several languages spoken in Nepal and is *lingua franca*, i.e., a common language [73][74][75]. The multi-lingual nature of Nepal's landscape, along with code-mixing make the speakers familiar with or have evolved, cultural insertions from English. The machine translation of the AD corpus has an inherent limitation, as the cultural nuances are harder to replicate algorithmically. Without a significantly annotated corpus, the machine translations will not capture cultural and linguistic nuances native to the target language. This will lower the accuracy of AD detection. However, in this particular case, the translations were carried out by native Nepali speakers with 13 years of formal education in that language and verified by a linguistic expert. This is the strength of the research. There are some syntactical differences between English and Nepali, especially in the order and placement of elements. For example, English follows the default word order of subject-verb-object (SVO). Whereas in Nepali the default word order is subject-object-verb (SOV). That is, in Nepali, the verb occurs at the end of a sentence. In English, the object complements the verb and occurs after the verb (to the right), while in Nepali, the object occurs to the left of the verb. Nepali nouns following numerals will be marked for plurality. While translating, considerations have been given to such structural differences between English and Nepali grammar. This can be relatively easily codified as described by researchers [76]. The features like hesitation and puzzlement, Part of Speech (POS) based features, unintelligible word rate, complexity features like phonemes per word, etc. were taken care of. The performance has been improved through human translators. However, using human translators and linguists does take time, although considerably lower than collecting primary data in Nepali and annotating them.

5.2 Future Works

The future work could include actually getting medical practitioners to verify and validate the translated corpus as representative of actual patient's language usage. The exercise would provide validation as well as promote understanding of the richness and appropriateness of the translation-based approach. In addition to medical experts, it is also possible to combine alternative approaches such as MR-based image recognition of neuroanatomy to build multimodal systems. The NLP-based model may still benefit from improvements in the form of injection of native features to strengthen the translated corpus. In the future, we will also assess, if any cultural or linguistic features are missing in translation, and accordingly, inject language and culturally specific features of Nepali (low-resource) language into the translated corpus (as part of translation and processing).

In addition, in the long term, primary data of corpus can be developed in the low resource language, in this case, Nepali. This native corpus can be compared with the translated corpus for similarities.

Also, developing a speech recognition system that helps to analyze the speech of people can be a direct method for early detection of Alzheimer's disease. Also, the work can be extended to other forms of dementia, such as Parkinson's among Nepalese patients. When collecting data as the primary source, the program should be well designed otherwise it would amplify the advantage of one sub-groups over others. Through this approach, we can plan where the gaps are and compensate for collecting data. This way, it would nullify the enforcement of social disadvantages caused by any normative biases and finally expand a language to improve the ML technology in the domain. While these improvements will make the solution more efficient, it is also advisable that in order to improve efficiency, a combined human-machine translation be explored.

6. Conclusion

Detecting Alzheimer's disease at its earliest stage is still a challenging task. Speech degeneration, being one of the most common and earliest symptoms in AD patients, should be leveraged to identify the disease. Since there is no clinical medicine or method to cure the disease completely, the only practical way would be to identify it in its early stage to stop the progression of the disease. Hence, the study aims to detect AD early for the people who speak the Nepali language. This is a step towards solving problems in identifying the disease and motivation for further researchers working in this field. The significant advantage of this automated system is that it takes significantly less time to predict the presence of AD. Also, the treatment costs are highly reduced and can be used over a large number of cycles for many people. The further improvement in the study can include acoustic features such as the duration of pause a person takes while speaking, how confused his words sound, etc. Also, developing a speech recognition system that helps to analyze the speech of people can be a direct method for early detection of Alzheimer's disease. Also, the work can be extended to other forms of dementia, such as Parkinson's among Nepalese patients. It is especially vital because healthcare service is not very useful in the country. Thus, it can help the health specialist in their decision-making and reduce the time and cost associated with the identification of the disease.

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

References

- [1] X. Zhou and J. W. Ashford, "Advances in screening instruments for Alzheimer's disease," *Aging Med.*, vol. 2, no. 2, pp. 88–93, 2019, doi: 10.1002/agm2.12069.
- [2] P. Benefits, "2018 ALZHEIMER'S DISEASE FACTS AND FIGURES Includes a Special Report on the Financial and Personal Benefits of Early Diagnosis," 2018.
- [3] H. Liu-Seifert *et al.*, "Disease Modification in Alzheimer's Disease: Current Thinking," *Ther. Innov. Regul. Sci.*, vol. 54, no. 2, pp. 396–403, 2020, doi: 10.1007/s43441-019-00068-4.
- [4] 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," *Cereb. Cortex*, vol. 19, no. 3, pp. 497–510, 2009, doi: 10.1093/cercor/bhn113.
- [5] S. Thapa, P. Singh, D. K. Jain, N. Bharill, A. Gupta, and M. Prasad, "Data-Driven Approach based on Feature Selection Technique for Early Diagnosis of Alzheimer's Disease," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9207359.
- [6] K. Domoto-Reilly, D. Sapolsky, M. Brickhouse, and B. C. Dickerson, "Naming impairment in Alzheimer's disease is associated with left anterior temporal lobe atrophy," *Neuroimage*, vol. 63, no. 1, pp. 348–355, 2012, doi: 10.1016/j.neuroimage.2012.06.018.

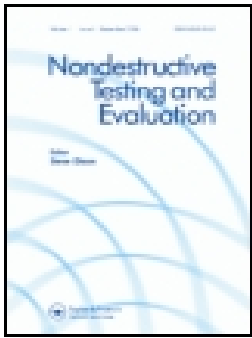
- [7] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. M. Stadlan, "Clinical diagnosis of alzheimer's disease: Report of the NINCDS-ADRDA work group* under the auspices of department of health and human services task force on alzheimer's disease," *Neurology*, vol. 34, no. 7, pp. 939–944, 1984, doi: 10.1212/wnl.34.7.939.
- [8] K. Fraser and G. Hirst, "Detecting semantic changes in Alzheimer's disease with vector space models," *Proc. Lr. 2016 Work. Resour. Process. Linguist. Extra-Linguistic Data from People with Var. Forms Cogn. Impair.*, no. May, pp. 1–8, 2016.
- [9] K. Faber-Langendoen, J. C. Morris, J. W. Knesevich, E. LaBarge, J. P. Miller, and L. Berg, "Aphasia in senile dementia of the alzheimer type," *Ann. Neurol.*, vol. 23, no. 4, pp. 365–370, 1988, doi: 10.1002/ana.410230409.
- [10] S. Ahmed, A. M. F. Haigh, C. A. de Jager, and P. Garrard, "Connected speech as a marker of disease progression in autopsy-proven Alzheimer's disease.," *Brain*, vol. 136, no. Pt 12, pp. 3727–3737, 2013, doi: 10.1093/brain/awt269.
- [11] H. S. Kirshner, W. G. Webb, and M. P. Kelly, "The naming disorder of dementia," *Neuropsychologia*, vol. 22, no. 1, pp. 23–30, 1984, doi: 10.1016/0028-3932(84)90004-6.
- [12] G. Glosser, "Patterns of Discourse Production among Neurological Patients with Fluent Language Disorders," *Brain Lang.*, vol. 40, pp. 67–88, 1990.
- [13] J. P. R. Dick, R. J. Guilloff, and A. Stewart, "Mini-mental state examination in neurological patients," *J. Neurol. Neurosurg. Psychiatry*, vol. 47, no. 5, pp. 496–499, 1984, doi: 10.1136/jnnp.47.5.496.
- [14] J. E. Storey, J. T. J. Rowland, D. A. Conforti, and H. G. Dickson, "The Rowland Universal Dementia Assessment Scale (RUDAS): A multicultural cognitive assessment scale," *Int. Psychogeriatrics*, vol. 16, no. 1, pp. 13–31, 2004, doi: 10.1017/S1041610204000043.
- [15] S. J. Cano *et al.*, "The ADAS-cog in Alzheimer ' s Disease clinical trials : Psychometric evaluation of the sum and its parts To cite this version : HAL Id : hal-00580696," 2011.
- [16] R. W. Heinrichs, "Current and Emergent Applications of Neuropsychological Assessment: Problems of Validity and Utility," *Prof. Psychol. Res. Pract.*, vol. 21, no. 3, pp. 171–176, 1990, doi: 10.1037/0735-7028.21.3.171.
- [17] S. Velupillai, H. Suominen, M. Liakata, A. Roberts, and D. Anoop, "Europe PMC Funders Group Using clinical Natural Language Processing for health outcomes research : Overview and actionable suggestions for future advances," pp. 11–19, 2020, doi: 10.1016/j.jbi.2018.10.005.Using.
- [18] S. O. Orimaye, J. S. M. Wong, K. J. Golden, C. P. Wong, and I. N. Soyiri, "Predicting probable Alzheimer's disease using linguistic deficits and biomarkers," *BMC Bioinformatics*, vol. 18, no. 1, pp. 1–13, 2017, doi: 10.1186/s12859-016-1456-0.
- [19] V. Vincze *et al.*, "Detecting mild cognitive impairment by exploiting linguistic information from transcripts," *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Short Pap.*, no. August, pp. 181–187, 2016, doi: 10.18653/v1/p16-2030.
- [20] J. Fritsch, S. Wankerl, N. Elmar, E. Polytechnique, and F. De Lausanne, "AUTOMATIC DIAGNOSIS OF ALZHEIMER ' S DISEASE USING NEURAL NETWORK LANGUAGE MODELS Friedrich-Alexander-University Erlangen-Nuremberg , Germany," pp. 5841–5845, 2019.
- [21] J. Chen, J. Zhu, and J. Ye, "An attention-based hybrid network for automatic detection of Alzheimer's disease from narrative speech," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2019-Sept, pp. 4085–4089, 2019, doi: 10.21437/Interspeech.2019-2872.
- [22] J. H. Chen *et al.*, "Dementia-related functional disability in moderate to advanced parkinson's disease: Assessment using the world health organization disability assessment schedule 2.0," *Int. J. Environ. Res. Public Health*, vol. 16, no. 12, 2019, doi: 10.3390/ijerph16122230.

- [23] J. Weiner, C. Herff, and T. Schultz, "Speech-based detection of Alzheimer's disease in conversational German," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, pp. 1938–1942, 2016, doi: 10.21437/Interspeech.2016-100.
- [24] F. Rudzicz, L. C. Currie, A. Danks, T. Mehta, and S. Zhao, "Automatically identifying trouble-indicating speech behaviors in Alzheimer's disease," *ASSETS14 - Proc. 16th Int. ACM SIGACCESS Conf. Comput. Access.*, pp. 241–242, 2014, doi: 10.1145/2661334.2661382.
- [25] J. Liu, J. Zhao, and X. Bai, "Syntactic Impairments of Chinese Alzheimer's Disease Patients from a Language Dependency Network Perspective," *J. Quant. Linguist.*, vol. 28, no. 3, pp. 253–281, 2021, doi: 10.1080/09296174.2019.1703485.
- [26] A. Khodabakhsh, F. Yesil, E. Guner, and C. Demiroglu, "Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech," *Eurasip J. Audio, Speech, Music Process.*, vol. 2015, no. 1, 2015, doi: 10.1186/s13636-015-0052-y.
- [27] K. I. James T. Becker, Francois Boller, Oscar L. Lopez, Judith Saxton, "The natural History of Alzheimer's Disease," *Sight and Sound*, vol. 26, no. 12, pp. 16–19, 2016, doi: 10.1177/1097184x09352181.
- [28] B. MacWhinney and Carnegie Mellon University, "The CHILDES Project: Tools for Analyzing Talk. Part 1: The CHAT Transcription Format," no. 2000, 2000.
- [29] 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, doi: 10.1080/02687039608248419.
- [30] A. Khodabakhsh, S. Kusxuoglu, and C. Demiroglu, "Natural language features for detection of Alzheimer's disease in conversational speech," *2014 IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2014*, pp. 581–584, 2014, doi: 10.1109/BHI.2014.6864431.
- [31] S. M. Mohammad, M. Salameh, and S. Kiritchenko, "How translation alters sentiment," *J. Artif. Intell. Res.*, vol. 55, pp. 95–130, 2016, doi: 10.1613/jair.4787.
- [32] U. Naseem and K. Musial, "DICE: Deep intelligent contextual embedding for twitter sentiment analysis," *Proc. Int. Conf. Doc. Anal. Recognition, ICDAR*, pp. 953–958, 2019, doi: 10.1109/ICDAR.2019.00157.
- [33] U. Naseem, I. Razzak, and I. A. Hameed, "Deep Context-Aware Embedding for Abusive and Hate Speech detection on Twitter," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.
- [34] A. Kulkarni and A. Shivananda, "Natural Language Processing Recipes," *Nat. Lang. Process. Recipes*, pp. 67–96, 2019, doi: 10.1007/978-1-4842-4267-4.
- [35] D. Isa, L. H. Lee, V. P. Kallimani, and R. Rajkumar, "Text document preprocessing with the bayes formula for classification using the support vector machine," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1264–1272, 2008, doi: 10.1109/TKDE.2008.76.
- [36] U. Naseem, K. Musial, P. Eklund, and M. Prasad, "Biomedical Named-Entity Recognition by Hierarchically Fusing BioBERT Representations and Deep Contextual-Level Word-Embedding," *Proc. Int. Jt. Conf. Neural Networks*, 2020, doi: 10.1109/IJCNN48605.2020.9206808.
- [37] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Adv. Neural Inf. Process. Syst.*, no. July, 2001.
- [38] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Adv. Neural Inf. Process. Syst.*, no. October, 2013.
- [39] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, pp. 3483–3487, 2019.
- [40] A. Roy, Y. Park, and Sh. Pan, "Learning Domain-Specific Word Embeddings from Sparse Cybersecurity Texts," 2017, [Online]. Available: <http://arxiv.org/abs/1709.07470>.

- [41] B. Srinivasa-Desikan, "Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras," *Packt Publ.*, 2018.
- [42] J. Bhatta, D. Shrestha, S. Nepal, S. Pandey, and S. Koirala, "Efficient Estimation of Nepali Word Representations in Vector Space," *J. Innov. Eng. Educ.*, vol. 3, no. 1, pp. 71–77, 2020, doi: 10.3126/jiee.v3i1.34327.
- [43] S. R. Safavian and D. Landgrebe, "A Survey of Decision Tree Classifier Methodology," *IEEE Trans. Syst. Man Cybern.*, vol. 21, no. 3, pp. 660–674, 1991, doi: 10.1109/21.97458.
- [44] S. Rajora *et al.*, "A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance," *Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018*, pp. 1958–1963, 2019, doi: 10.1109/SSCI.2018.8628930.
- [45] C. CORTES and V. VAPNIK, "Support-Vector Networks," *Mach. Lang.*, vol. 7, no. 2, pp. 142–147, 1995, doi: 10.1111/j.1747-0285.2009.00840.x.
- [46] I. Rish, "An Empirical Study of the Naïve Bayes Classifier An empirical study of the naive Bayes classifier," *Cc.Gatech.Edu*, no. January 2001, pp. 41–46, 2014, [Online]. Available: <https://www.cc.gatech.edu/~isbell/reading/papers/Rish.pdf>.
- [47] S. Adhikari, S. Thapa, and B. K. Shah, "Oversampling based Classifiers for Categorization of Radar Returns from the Ionosphere," *Proc. Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2020*, no. Icesc, pp. 975–978, 2020, doi: 10.1109/ICESC48915.2020.9155833.
- [48] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
- [49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2012.
- [50] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.
- [51] M. Rhanoui, M. Mikram, S. Yousfi, and S. Barzali, "A CNN-BiLSTM Model for Document-Level Sentiment Analysis," *Mach. Learn. Knowl. Extr.*, vol. 1, no. 3, pp. 832–847, 2019, doi: 10.3390/make1030048.
- [52] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6354 LNCS, no. PART 3, pp. 92–101, 2010, doi: 10.1007/978-3-642-15825-4_10.
- [53] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–15, 2015.
- [54] Y. Kim, "Convolutional neural networks for sentence classification," *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1746–1751, 2014, doi: 10.3115/v1/d14-1181.
- [55] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [56] X. Wang, W. Jiang, and Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts," *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Tech. Pap.*, pp. 2428–2437, 2016.
- [57] S. Ioffe, "Batch Renormalization: Towards reducing minibatch dependence in batch-normalized models," *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 1946–1954, 2017.
- [58] H. Choi, K. Cho, and Y. Bengio, "Fine-grained attention mechanism for neural machine translation," *Neurocomputing*, vol. 284, pp. 171–176, 2018, doi: 10.1016/j.neucom.2018.01.007.

- [59] Y. Diao *et al.*, “CRHASum: extractive text summarization with contextualized-representation hierarchical-attention summarization network,” *Neural Comput. Appl.*, vol. 32, no. 15, pp. 11491–11503, 2020, doi: 10.1007/s00521-019-04638-3.
- [60] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
- [61] R. Kadlec, M. Schmid, O. Bajgar, and J. Kleindienst, “Text understanding with the attention sum reader network,” *54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap.*, vol. 2, pp. 908–918, 2016, doi: 10.18653/v1/p16-1086.
- [62] J. Salazar, K. Kirchhoff, and Z. Huang, “Self-attention Networks for Connectionist Temporal Classification in Speech Recognition,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May, pp. 7115–7119, 2019, doi: 10.1109/ICASSP.2019.8682539.
- [63] F. Guzmán *et al.*, “The Flores evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English,” *EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf.*, pp. 6098–6111, 2020, doi: 10.18653/v1/d19-1632.
- [64] D. Wild, A. Grove, S. Eremenco, S. McElroy, A. Verjee-Lorenz, and P. Erikson, “Wild2005_Value in Health 8(2)_94-104.pdf,” *Value Heal.*, vol. 8, no. 2, pp. 94–104, 2005.
- [65] S. McKown *et al.*, “Good practices for the translation, cultural adaptation, and linguistic validation of clinician-reported outcome, observer-reported outcome, and performance outcome measures,” *J. Patient-Reported Outcomes*, vol. 4, no. 1, 2020, doi: 10.1186/s41687-020-00248-z.
- [66] A. Balahur and M. Turchi, “Improving sentiment analysis in twitter using multilingual machine translated data,” *Int. Conf. Recent Adv. Nat. Lang. Process. RANLP*, no. September, pp. 49–55, 2013.
- [67] A. Paul and B. S. Purkayastha, “English to Nepali Statistical Machine Translation System,” *Lect. Notes Networks Syst.*, vol. 24, pp. 423–431, 2018, doi: 10.1007/978-981-10-6890-4_41.
- [68] S. Strassel and J. Tracey, “LORELEI language packs: Data, tools, and resources for technology development in low resource languages,” *Proc. 10th Int. Conf. Lang. Resour. Eval. Lr. 2016*, pp. 3273–3280, 2016.
- [69] H. Riza *et al.*, “Introduction of the Asian Language Treebank,” *2016 Conf. Orient. Chapter Int. Comm. Coord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2016*, no. October, pp. 1–6, 2017, doi: 10.1109/ICSDA.2016.7918974.
- [70] A. Riccardi, “Translation Studies: Perspectives on an Emerging Discipline,” *South. African Linguist. Appl. Lang. Stud.*, vol. 24, no. 1, pp. 129–132, 2006, doi: 10.2989/16073610609486411.
- [71] B. C. Myers-scotton, “SIL Electronic Book Reviews 2006-006 Contact Linguistics : Bilingual encounters and grammatical outcomes,” *Oxford Univ. Press. 2002. Pp. 356. Pap.*, 2006, doi: 10.1093/acprof:oso/9780198299530.001.0001.
- [72] D. Gurung, “Nepali-English code-switching in the conversations of Nepalese people Nepali-English Code-switching in the Conversations of Nepalese People: A Sociolinguistic Study,” 2018, [Online]. Available: <https://pure.roehampton.ac.uk/ws/portalfiles/portal/1284973/>.
- [73] P. Trudgill, “A Glossary of Sociolinguistics,” *Edinburgh Univ. Press*, pp. 234–235, 2003, doi: 10.1590/S0102-44502003000100014.
- [74] R. K. Dahal, “Language Politics in Nepal,” *J. Polit. Sci.*, vol. 1, no. 1, 1998, doi: <https://doi.org/10.3126/jps.v1i1.1685>.
- [75] C. Genetti, *How languages work: An introduction to language and linguistics*. Cambridge University Press and Assessment, 2014.
- [76] L. Wei and M. G. Moyer, “The Blackwell Guide to Research Methods in Bilingualism and

Multilingualism,” *Blackwell Publ. Ltd.*, pp. 1–403, 2009, doi: 10.1002/9781444301120.



Failure analysis of a low-pressure stage steam turbine blade

Pooja Rani & Atul K. Agrawal

To cite this article: Pooja Rani & Atul K. Agrawal (2022): Failure analysis of a low-pressure stage steam turbine blade, Nondestructive Testing and Evaluation, DOI: [10.1080/10589759.2022.2156503](https://doi.org/10.1080/10589759.2022.2156503)

To link to this article: <https://doi.org/10.1080/10589759.2022.2156503>



Published online: 15 Dec 2022.



Submit your article to this journal [↗](#)



Article views: 66



View related articles [↗](#)



View Crossmark data [↗](#)



Failure analysis of a low-pressure stage steam turbine blade

Pooja Rani and Atul K. Agrawal

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

ABSTRACT

Steam turbine blades are regularly damaged because of their harsh working conditions, which include elevated temperatures and fluctuating loads. Most investigations of blade failures end with a metallurgical analysis, which does not provide sufficient positive identification of the mechanisms involved. Hence, in the current research work a mechanical analysis is performed in conjunction with the metallurgical analysis for competent analysis of blade failure. For the purpose of evaluating the damage, non-destructive testing (NDT) was carried out. The purpose of this examination is to qualitatively examine the blade of a 210 MW low-pressure steam turbine after 1,52, 241 h of working to identify the critical locations of damage and the reason behind it. Visual examination, chemical analysis, dye penetration testing, and metallurgical testing are all part of this examination. In addition, mechanical properties were evaluated using hardness and tensile testing. The findings revealed that water droplet erosion accelerated blade failure, preferentially attacking the blade's edges. These erosion pits act as stress concentrators and serve as a potential crack propagator if neglected, which can lead to catastrophic failure of the system. Hence, to increase reliability and to avoid such failures in future, this type of failure analysis is highly recommended.

ARTICLE HISTORY

Received 17 October 2022

Accepted 2 December 2022

KEYWORDS

Steam turbine blade;
damage causes; erosion
pitting; NDT; X20cr13
stainless steel

1. Introduction

Steam turbines are very significant and crucial components of thermal power plants, as they characterise the entire unit's lifetime and efficiency, and thus the entire power plant. As a result, the steam turbine reliability is critical [1]. Steam turbine blades are a critical component and play a critical role in the turbine's reliability. If turbine blades fail, it will lead to more failures and significant financial losses. So, detailed research into the causes of turbine blade failure is critical in order to improve turbine system reliability [2,3].

Low-pressure (LP) turbine blades have a higher failure rate than high-pressure (HP) and intermediate-pressure (IP) turbine blades, according to statistics. Generally, LP blades are expected to last 30 years, although there are several incidents of blades failing prematurely in practice [4]. A low-pressure blade failure occurred prematurely at a 110 MW fossil fuel power plant. The fracture occurs in the profile region near the root, and the causes are investigated using various techniques. They found that corrosion fatigue was the reason for the failure. There was no evidence of blade material deterioration [5].

Visual inspection, microstructure analysis, chemical analysis, microhardness, and tensile testing were used to investigate the untimely demise of steam turbine rotor blades. The low-pressure side of the blade's lower trailing edge had erosion caused by foreign particles and water droplet erosion on the upper leading edge [6]. Due to improper filler attachment, microcracks in the brazing metal and high cyclic fatigue, which was identified by fractographic observations as the main cause of the blade's failure in a 17th stage steam turbine blade. The fatigue crack was also shown to be initiated and propagated adjacent to the lacing hole from the brazing interface between the rod and the blade [7].

There can be various mechanisms involved responsible for the failure of blades in a turbine.

LP turbine blades are prone to premature failure, and corrosion, erosion, fatigue, and their interactions are the principal causes [8]. Many researchers have done work in this area to find the root cause of failure of turbine components. They carried out the failure analysis in the form of chemical composition, microstructural degradation and mechanical tests. A low-pressure steam turbine blade failed after 13,200 service hours because of an environment-assisted fatigue fracture [2]. In a 210 MW plant, Corrosion fatigue initiated from pits was the reason for blade's failure in LP steam turbine [9]. Foreign particle erosion-corrosion was the reason for fatigue failure of a steam turbine blade after 72,000 h of working [10]. Damage to the steam turbine blade occurred after approximately 165,000 h of operation. Fatigue was found to be the primary cause of failure, followed by erosion and the development of notches [11]. A case study concluded that high cycle fatigue and impact of the brass ring was the reason for low-pressure turbine blade failure [1]. After 8000 h of operation at high temperatures, a failed investigation of 40 MW gas turbine blades was conducted. Because of the transformation, a continuous film of carbides was found in the base material's grain boundaries. Intergranular cracks that developed as a result of high-temperature exposure are identified as the failure's primary cause. When the cracks reached a threshold length after starting at the grain boundaries, catastrophic fracture resulted [12].

Cano et al. predicted the damage to last-stage blades in a 110 MW steam turbine. The findings revealed that centrifugal force causes damage to the blades, which most likely leads to crack initiation due to low cycle fatigue [13]. According to Banaszekiewicz et al., stress corrosion cracking was the primary cause of a 60 MW turbine rotor failure after 197,654 hours of service and 488 starts [14]. A 210 MW thermal power plant's LP turbine blade failed, and the failure analysis presented by Mukhopadhyay et al. and found that the blade failed due to high cyclic fatigue caused due to excessive vibrations generated as a result of grid frequency fluctuations [4]. Kubiak et al. [15] looked into what causes damage to the steam turbine blade's trailing edge. The corrosion fatigue of the blade was determined to be the cause of the damage at the bottom. In reality, due to the presence of the flaw in the arrangement, cracks propagated in the area where the greatest amount of force was applied. The occurrence of vibration and the flaw between the base platform and the tree-like design of the blades were both identified as contributing factors in the spread of the cracks. Kubiak et al. [16] evaluated the role that environmental factors have in turbine blade failure in another study. Upon investigation, it was determined that erosion in the blade led to the emergence of a void, ultimately leading to the failure. Note that blade edge degradation is not unusual. Erosion happens when foreign particles and pressures are present for a long time. Fei Xue et al. [17] studied the reason of a nuclear

power plant's fractured turbine blade. SEM studied the blade's fracture surface. Experiments comparing fretting-induced fatigue to high-cycle fatigue were designed. Finite Element Analysis was used to study the blade's operational stress. The fourth-stage blade crack of the low-pressure steam turbine showed fretting-induced high cycle fatigue with trans-granular fracture morphology. The combination of fretting wear and a rather high stress level led to fretting fatigue. Also, in order to prove that failure study is the key to enhancing turbine efficiency, Poppy Puspitasari et al. discussed various case studies of failure in turbine hot section components such as blades [18]. Pitting, fatigue, corrosion, and creep are a few of the different types of failure that can occur with blades. These failures can be repaired, but doing so would not be safe and replacing the blade would be expensive, so it is necessary to conduct a critical analysis of failure to extend the life of blades. Once a crack appears in the blade, it will spread and eventually fail the turbine blades, forcing the plant to shut down [19].

Understanding how a material degrades is therefore critical not only in the event of failure but also when deciding on a new part, a repair, or a modification. Having a clear understanding of the deterioration mode allows us to replace and repair turbine parts at the optimal time, avoiding unnecessary replacements and avoiding forced shut-downs [20].

This work presents an experimental case study of a low-pressure stage blade after 396 start-ups of the power plant. Generally, the blades are recommended to be checked after 100,000 h of operation. This blade has completed over 100,000 h of operation. This LP blade has been in use for more than 1,52,241 h, according to the information available. Since commissioning, there have been an estimated total of 379 turbine starts till Dec. 2017. The details are as follows:

No. of cold starts: 5, No. of warm starts: 205, No. of hot starts: 118

Total no. of starts: 379

During inspection after 152,241 h of service, dent marks and pits are observed on the surface of the LP stage-7 blade. In order to assess the material's state of health and the dominant damage mechanism to decide on its further operation, detailed analysis results, such as visual examination, chemical analysis, dye penetration testing, metallographic study and mechanical testing, are presented. A pictorial view of the studied blade is shown in Figure 1.

Pits observed on the blade are due to erosion caused by water droplets. The steam received from the (high-pressure) HP turbine is low-quality steam, and further, the steam continuously expands in low-pressure turbine. So, when condensed water is sprayed against the blades, they can be quickly impinged and eroded by the high volume of water droplets [21]. Droplet impact erosion results in blade material loss, making a significant change to the aerodynamically optimal blade geometry and causing a significant flow disruption. This negatively impacts the machine's performance, eventually necessitating turbine blade replacement [22].

The thinnest regions of the aerofoil are the trailing edges, where material removal could modify the stress level. Erosion damage caused by water droplets at high-speed swirling from tip to the base of the aerofoil, scarring the trailing edge and forming macroscopic notches and 'worm holes' at the base of these scars. These notches act as stress concentrators, and in a high-stress region become detrimental to the integrity of the blade. If unattended, the notches act as crack initiators developing into propagating



Figure 1. Pictorial view of a steam turbine blade.

cracks. In combination with dynamic stresses, this could quickly develop into a blade failure. This type of failure analysis plays a very important role in improving the reliability of turbine systems and also prevents such failure incidents in the future.

2. Material and experimental procedures

The turbine's operating conditions are extremely complicated. In the wet steam area, the turbine is subjected to centrifugal force, steam power, exciting steam force, corrosion and vibration, and high-speed erosion [23]. When choosing a blade material, high stress and erosion should be considered [24]. The last stage blades, which are up to about 40 inches in height, have been made of 12Cr martensitic stainless steel for many years for 3600-rpm designs because of its excellent properties like supreme toughness and resistance to corrosion [24]. The chemical composition and mechanical properties of X20Cr13 steel are presented in Tables 1 and 2 [25].

2.1. Non-destructive testing (NDT)

Damage to steam turbine blades can occur at any step of the design, production, or operation. To avoid a major accident and significant economic losses, it is necessary to inspect the blades periodically during planned outages and repair or replace damaged or deformed blades as soon as they are found to be faulty [23].

Non-destructive evaluation (NDE) is becoming increasingly important in assuring pre-service quality and monitoring in-service degradation to avoid premature component or structural failure. After visual inspection, Dye Penetration Testing (DPT) was carried out on the Steam Turbine Blade according to ASTM E165 [26] to find out the presence of any

Table 1. Chemical composition (wt.%).

C	Cr	Si	Mn	P	S	Ni
0.17–0.22	12.50–14.0	0.10–0.60	0.30–0.80	≤0.03	≤0.02	0.30–0.80

Table 2. Mechanical properties.

Tensile strength MPa	Hardness BHN	Proof stress MPa	Elongation Min %	Reduction in Area Min %
800–950	280	≥600	15	50

defect or any crack-like defect over the surface and subsurface. The chemical analysis was carried out on LP blade by Spectro analysis according to ASTM E-A448 M [27] to determine the chemical composition of the blade material. Metallurgical testing was performed for microstructural studies on LP blades by polishing, and chemical etching grain size was measured after surface preparation and etching with Kalling II solution as per ASM handbook for metallography and microstructure [28]. Mechanical tests such as hardness, tensile strength, and proof stress were performed on tube samples. The sample was tested in accordance with ASTM-370 [29]. All the tests for the present analysis were performed at the AVIS (Aequitas Veritas Industrial Services) laboratory in Vadodara, Gujarat.

3. Result and discussion

3.1. Visual examination

A visual assessment was carried out to look for any markings or pitting on the blade surface that would indicate erosion or mechanical failure. Some dent marks were observed on the profile and root of the blade. According to visual checks, the erosion pits occurred on the edges of the turbine blade, which can increase the stress concentration. However, significant erosion damage was discovered on Blade's edge parts, which is shown in Figure 2. On the surface, there was no sign of a crack or any other fault.

3.2. Dye penetration test (DPT)

In dye penetration testing, the developer either bleeds penetrants out from defects onto the surface, resulting in a visible signal known as bleed-out, or pulls dye penetrant-containing flaws back to the surface, resulting in a red mark on a white backdrop, as shown in Figure 3. Any bleed-out area on the surface can reveal the location, orientation, and type of faults present.

It can be seen that there is no defect or crack found on the blade root and only erosion pits were present on the edges of the blade. The erosion pits are formed due to striking of fine moist droplets formed by condensation of steam. Due to the loss of aerodynamic efficiency caused by leading-edge erosion, annual energy production drops dramatically. Table 3 shows the detailed results of dye penetration testing.

3.3. Chemical analysis

The chemical composition of the blade material was found to be as shown in Table 4. It was found to be consistent with AISI 420 grade martensitic stainless steel.

3.4. Metallurgical testing

The damaged blade was also subjected to microstructural examinations. Standard metallographic sample preparation was followed by optical microscopy for this objective. Microstructures show uniform and homogeneous structures in the matrix. The hardened and tempered martensitic structure can be seen in the micrographs. Figure 4 shows the

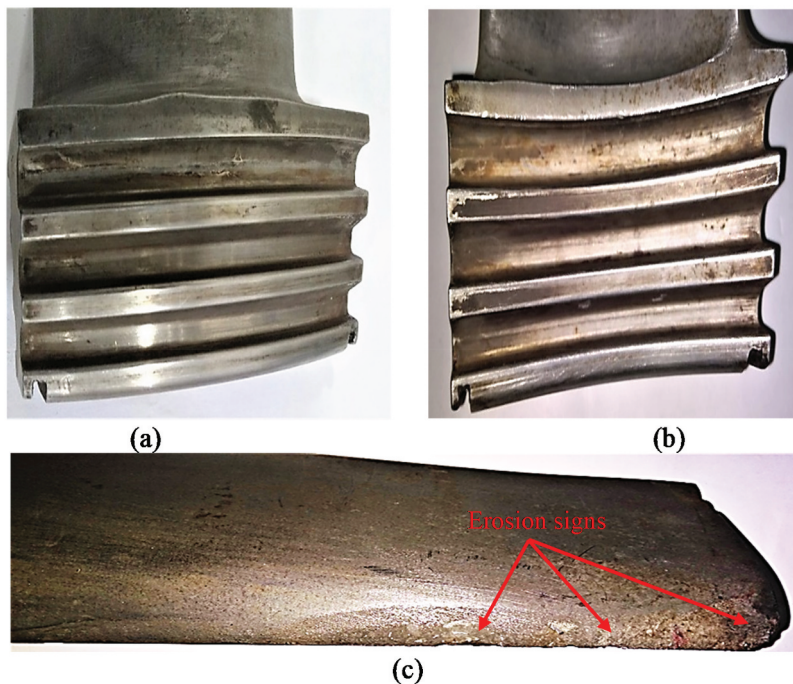


Figure 2. Blade root (a) back (b) front view with some dent marks (c) Erosion signs on blade's edges.

microstructure of the blade consisting of tempered martensitic and delta ferrite phases in the matrix.

Figure 5 illustrates the microstructures of the sample containing very fine tempered martensite and ferritic pool areas in the matrix. Optical micrographs of blade root and airfoil region clearly demonstrate uniformly distributed hardened and tempered martensite. There was no indication of microstructural degradation in the airfoil and root regions of the damaged blade, which is in compliance with the X20Cr13 specification (AISI 420).

The eroded portion is further studied to understand the changes in microstructure of the blade due to erosion pits and it the microstructure shows the bainitic structure and retained austenite as shown in Figure 6. Delta ferrite pools have also been observed on the top eroded portion of the blade. The bright white phase shows the presence of ferritic pools in microstructure shown in Figure 6. The presence of grain boundary carbides as shown in Figure 7 was also observed. These grain boundaries provide a path to crack propagation resulting in a faster rate of failure.

For the analysis of the failure behaviour of the material surfaces, it was necessary to ascertain the grain sizes of the blades under consideration; hence, grain size was measured after surface preparation and etching with Kalling II solution, as per the ASM handbook for metallography and microstructure [30]. Though grain boundaries are not identified in this structure, the grain size of this blade material was measured by grain size number-10 as per ASTM, as shown in Figure 8.

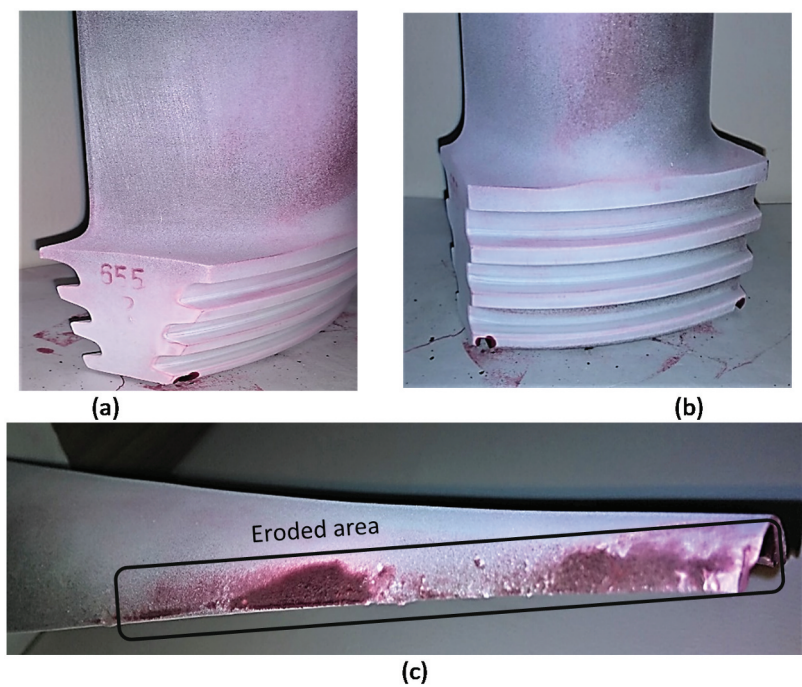


Figure 3. Blade root (a) side view (b) front view with no sign of a crack(c) Erosion pits on blade’s trailing edge.

Table 3. Dye penetration test results.

Details of parts	Qty./No. tested	Finding
Steam turbine blade-front	1 no.	No defect/crack was observed on the surface.
Steam turbine blade-back	1 no.	No defect/crack was observed on the surface.
Steam turbine blade-LHS	1 no.	No defect/crack was observed except erosion on edges.
Steam turbine blade-RHS	1 no.	No defect/crack was observed except erosion on edges.

Remarks: No crack observed except pitting damage corrosion.

Table 4. Chemical Analysis Results.

Element	C	Cr	Si	Mn	P	S	Ni
Observed value in %	0.20	13.3	0.43	0.62	0.029	0.019	0.58

3.5. Mechanical testing

To determine the change in material properties, the mechanical properties of the fractured blade were also investigated by measuring the material’s hardness and subjecting the material specimen to a tensile test.

3.5.1. Hardness testing

The hardness was carried out on a tube sample. Brinell hardness measurements were performed on blade materials using a universal hardness tester shown in Figure 9(a). The hardness values of the blades are found to be in the standard range, indicating that they

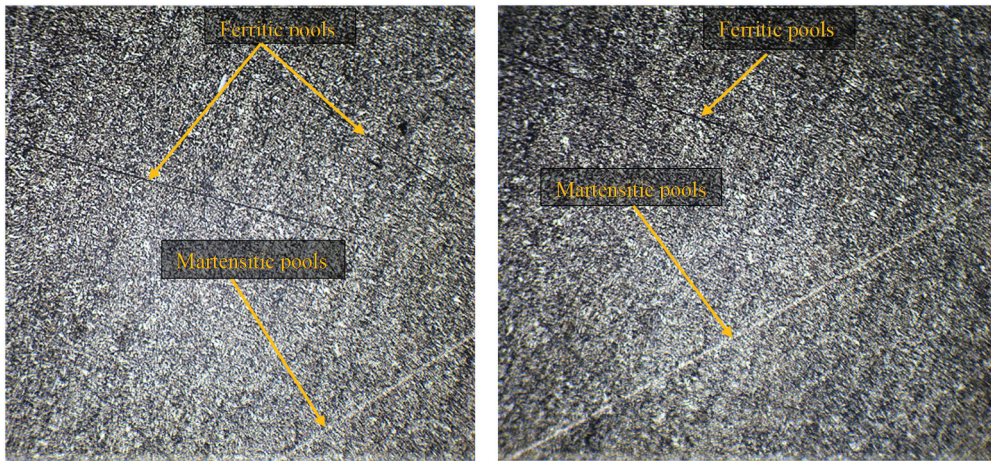


Figure 4. Tempered martensite and ferritic pools can be seen in the microstructure (100x).

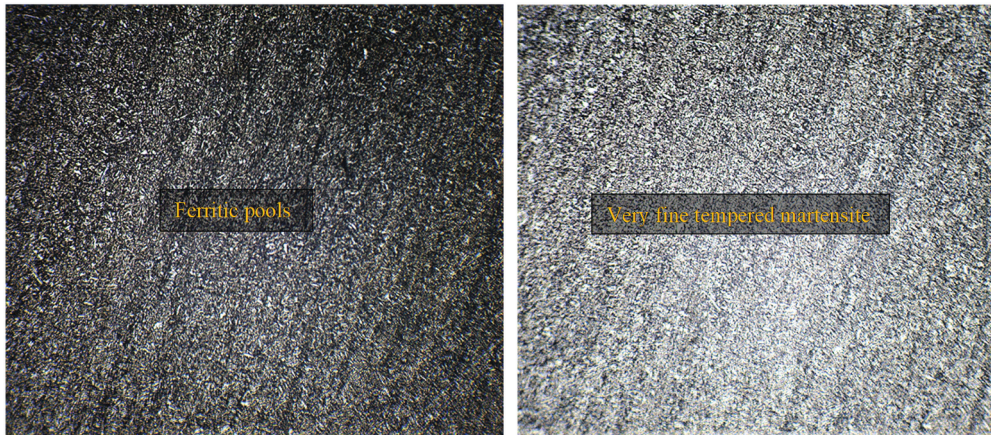


Figure 5. Optical micrographs at 100x of the root of damaged blade showing Very fine tempered martensite and ferritic pools.

have been properly tempered. Table 5 shows the observed values for hardness at different positions of turbine blade profile and root as shown in Figure 9(b). Figure 10 shows a comparison of hardness at different locations of LP blade with the standard hardness required for LP turbine blade.

The hardness results were compared with standard requirements, i.e. 280 BHN, and the obtained values are slightly lower. We show that no appreciable decrease in hardness or creep initiation was observed in blade root and blade profile areas.

However, it was observed that the value of hardness dropped significantly from 500 BHN to 261 BHN near the tip of turbine blade (approx. 200 mm from the tip) where the hardness was increased initially by flame hardening. The hardening was done to improve the erosion resistance of the blade tip where pitting occurs due to moisture content of steam.

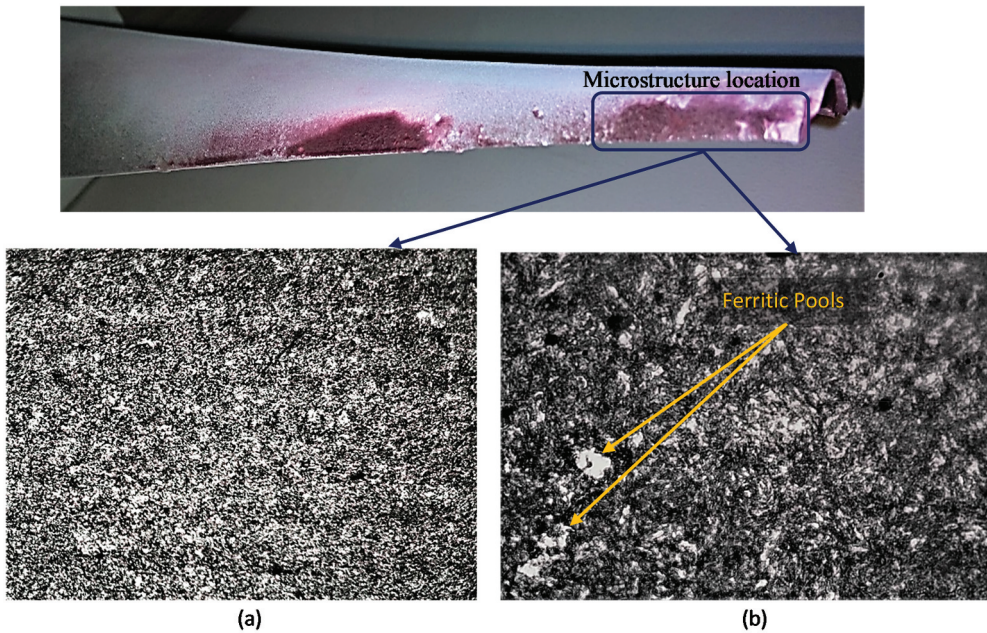


Figure 6. Microstructure shows bainitic and structure retained austenite (40x).

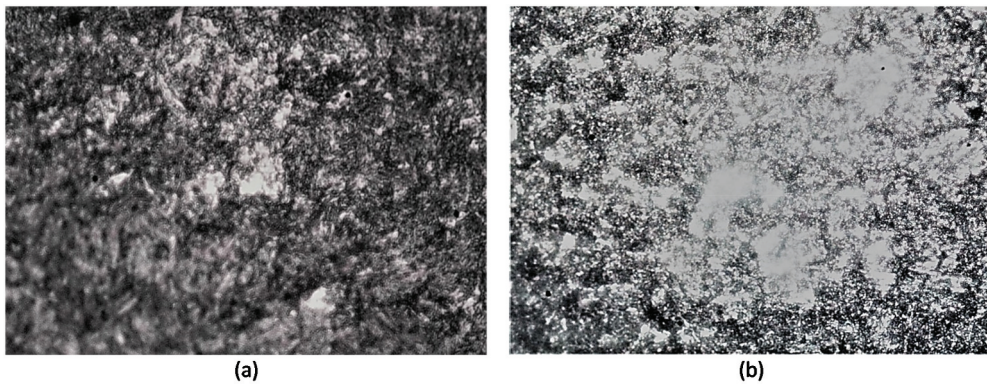
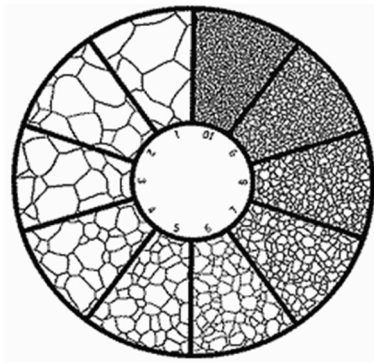
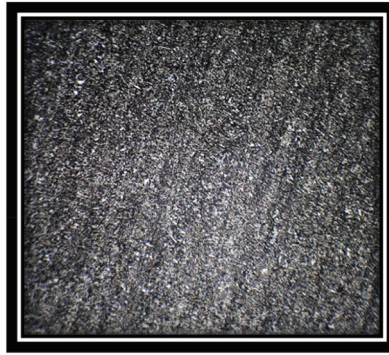


Figure 7. Microstructure shows delta ferrite pool (100x).

Flame Hardening: Flame hardening is used to give the blades more durability against this erosion. An improper application of this procedure can make the blade material, X 20Cr 13, more vulnerable to stress corrosion cracking. Stress corrosion cracking requires the presence of sufficiently high tensile stresses. The amount of total surface stress at the flame-hardened leading edge of the blade is mostly influenced by the residual stresses that remain after the process. Stress corrosion cracking is to be expected after a given amount of time in service if there are high tensile residual stresses. Therefore, it is very important to estimate the residual stresses to prevent structural failure in industrial turbines. For improved resistance to water droplet erosion, Siemens invented x-ray diffraction residual stress control of the flame hardened edge in the 1980s. The XRD

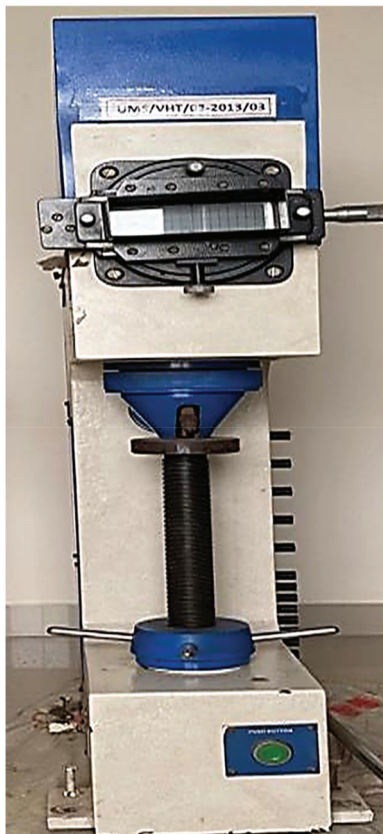


ASTM Grain Size No.



Blade Grain Size Measured- 10

Figure 8. Microstructure showing grain boundary carbides (a) 200x (b) 400x.



(a)



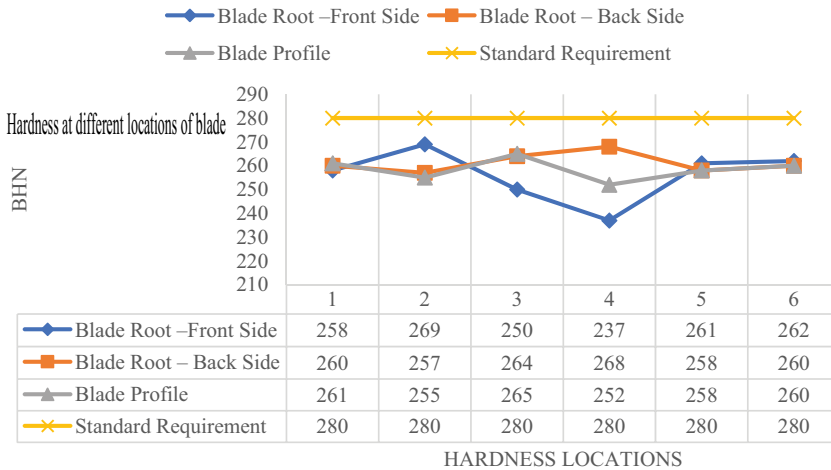
(b)

Figure 9. Grain size in accordance with ASTM.

Table 5. Hardness Testing results.

Sr. No.	Steam Turbine Blade (BHN)		
	Blade Root –Front Side	Blade Root – Back Side	Blade Profile
1	258	260	261
2	269	257	255
3	250	264	265
4	237	268	252
5	261	258	258
6	262	260	260

The average hardness value for blade root front side, blade root back side and blade profile are 256, 259, and 261 respectively.

**Figure 10.** (a) Hardness testing machine (b) blade profile and root hardness locations.

method measures strain by utilising the d-spacing between crystallographic planes. The d-spacing of a material will change depending on the direction of stress: it will widen when the material is under tension and narrow when it is under compression. A change in the angular position of the XRD peak caused by the presence of residual stresses can be directly detected by the instrument. Using X-ray or micromagnetic testing tools, the crucial residual surface stresses can be precisely identified [31–36]. Pineault et al. discussed how XRD can be used to characterise residual stresses in a component or assembly in order to prevent, minimise, or get rid of the impact of residual stress on premature failures, and then evaluate corrective measures that alter the residual-stress condition of a component [35]. Later on, magnetic Barkhausen noise residual stress measurement supplanted the X-ray method. Using a noise-like signal produced by applying a magnetic field to a ferromagnetic sample, the Barkhausen noise analysis method measures residual stress. Domains, collections of magnetic dipoles that can be thought of as little magnets, are found in ferromagnetic materials. Heinrich Barkhausen originally noticed that domain walls move back-and-forth under alternating magnetic fields in 1919 [37].

3.5.2. Tensile testing

The tensile test was performed on steam turbine blade specimen with specification detailed in Table 6.

Table 7 summarises the important tensile property data obtained by tensile testing specimens prepared from blade material. It can be argued that the blade did not significantly lose strength values while it was in use, disproving the idea that a decline in mechanical qualities would cause the blade to fail. It can be seen from the table that X20Cr13 blade material qualifies the requirement of mechanical properties.

Summarising the test results, it can be seen that all of the determined parameters, like chemical composition, tensile strength, hardness and yield strength, meet the standard requirements for the given material.

4. Discussions

- Damage mechanisms affecting components within complex machines are typically difficult to detect and diagnose, especially if they are difficult to reach, examine, and are in constant use, jeopardising system reliability and performance.
- The blades shall be free from folds due to forging, cracks, tearing, material defects, elongated non-metallic inclusions, seams, etc. Any blade containing such defects shall be rejected for further use.
- The failed blade's material was discovered to have the same chemical composition and hardness as standard martensitic stainless-steel grade AISI 420. There was no sign of microstructural deterioration. As a result, the blade material is found to be in compliance with the standard, indicating that material is not a factor in this failure.
- Distribution and pattern of erosion pits on the edges of the final stage low-pressure blade in a steam turbine shows that quenching cannot eliminate the erosion of the surface completely.
- Erosion scars and notches create stress concentrations that can lead to crack initiation and growth.

Table 6. Tensile Testing specimen and loading details.

Diameter (mm)	Area (mm ²)	Initial Gauge Length (mm)	Yield Load	Tensile load	Final Gauge Length (mm)
12.54	123.44	50.0	83.94	108.38	590.0

Table 7. Tensile Testing results.

Sr No.	Details	Yield Strength (MPa)	Elongation %	Tensile Strength (MPa)
1.	Required	≥600 MPa	≥15 Min.	800-950
2.	Obtained	680	18	878

5. Conclusions

- Modern turbine blade's last two low-pressure (LP) stages are expected to operate in a wet steam medium in steam turbines. Condensation during steam expansion typically produces fine mist droplets.
- This study analysed stainless steel ex-service steam turbine blades. (Water droplet erosion) WDE on different portions of the blade was identified.
- No crack or any other defect-like appearance was noticed on the surface during visual inspection of the turbine blade. No sign of corrosion or thermal fatigue was found on the surface. However, considerable erosion damage was observed in the edge sections of Blade.
- It was determined that the blade material was not faulty based on chemical analysis and mechanical testing.
- The turbine blade under inquiry did not fail due to a material flaw, according to the dye penetration testing and microstructure study.
- The defective sections were carefully analysed in order to determine the causes of failure, and it was found that water droplet erosion was the cause of blade damage. Identifying the deteriorating mode in advance allows us to replace and repair turbine parts at the best possible moment, reducing the need for unneeded replacement and preventing unscheduled outages.
- The airfoil's edges are scarred by water droplets leaving behind microscopic notches. As stress concentrators, these notches compromise the integrity of the blade in areas of high stress. If left ignored, the notches serve as crack initiators that eventually lead to spreading cracks. This could soon progress to a blade failure if combined with dynamic loads. Each facet of a failure inquiry contributes to figuring out what went wrong and how to prevent such incidents in the future. Failure analysis is thus crucial for enhancing turbine system reliability and avoiding similar failure incidences.

6. Future scope

The study of erosion over blades in the low-pressure stage steam turbine can be done using numerical simulations. This will give a good idea about the effect of factors like humidity levels, droplet diameter, mass flow rate, forces exerted on blade surfaces on aerodynamic behaviour and energy conversion efficiency. This study can also be further extended to calculate the stress generated on the turbine blade and to evaluate the effect of stresses on the life of the blade.

Acknowledgement

The authors are grateful to Dr M. K. Sharma, Technical Director: AEQUITAS VERITAS INDUSTRIAL SERVICES (AVIS) laboratory, for helping us in conducting the experiments and Mr D. C. Nirmal, Sr. DGM (STE-BHEL Bhopal), Mr Manoj Yadav, Manager (COE-BHEL BHOPAL) for technical guidance on steam turbines.

Disclosure statement

No potential conflict of interest was reported by the authors.

Statements and declarations

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: POOJA RANI reports equipment, data, or supplies was provided by AVISLABORATORY, VADODARA, INDIA.

References

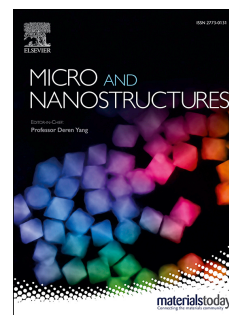
- [1] Molodtsov A, Dedov A, Klevtsov I, et al. Investigation of steam turbine blades damage and reliability in a power plant. *Key Eng Mater.* **2019**;799:89–94.
- [2] Wang WZ, Xuan FZ, Zhu KL, et al. Failure analysis of the final stage blade in steam turbine. *Eng Fail Anal.* **2007**;14(4):632–641.
- [3] Ilieva GI. Erosion failure mechanisms in turbine stage with twisted rotor blade. *Eng Fail Anal.* **2016**;70:90–104.
- [4] MUKHOPADHYAY NK, CHOWDHURY SG, DAS G, et al. An investigation of the failure of low pressure steam turbine blades. *Fail Anal Case Stud II.* **2001**;5(3):211–223.
- [5] Antony Harison MC, Swamy M, Pavan AHV, et al. Root cause analysis of steam turbine blade failure. *Trans Indian Inst Met.* **2016**;69(2):659–663.
- [6] Azevedo CRF, Sinátora A. Erosion-fatigue of steam turbine blades. *Eng Fail Anal.* **2009**;16(7):2290–2303.
- [7] Sameezadeh M, Hasanlou S, Zafari H, et al. Numerical simulation and experimental investigation on a steam turbine blade fractured from the lacing hole. *Eng Fail Anal.* **2020**;117(July):104809.
- [8] Kirols HS, Kevorkov D, Uihlein A, et al. Water droplet erosion of stainless steel steam turbine blades. *Mater Res Express.* **2017**;4(8):086510.
- [9] Bhagi LK, Gupta P, Rastogi V. Fractographic investigations of the failure of L-1 low pressure steam turbine blade. *Case Stud Eng Fail Anal.* **2013**;1(2):72–78.
- [10] Rivaz, Anijdan SHM, Moazami-Goudarzi M, et al. Damage causes and failure analysis of a steam turbine blade made of martensitic stainless steel after 72,000 h of working. *Eng Fail Anal.* **2022**;131(October):105801.2021. [10.1016/j.engfailanal.2021.105801](https://doi.org/10.1016/j.engfailanal.2021.105801).
- [11] Rivaz A, Mousavi Anijdan SH, Moazami-Goudarzi M. Failure analysis and damage causes of a steam turbine blade of 410 martensitic stainless steel after 165,000 h of working. *Eng Fail Anal.* **2020**;113(April):104557.
- [12] Vardar N, Ekerim A. Failure analysis of gas turbine blades in a thermal power plant. *Eng Fail Anal.* **2007**;14(4):743–749.
- [13] Cano S, Rodríguez JA, Rodríguez JM, et al. Detection of damage in steam turbine blades caused by low cycle and strain cycling fatigue. *Eng Fail Anal.* **2019**;97(August):579–588. 2018. DOI:.
- [14] Banaszkievicz M, Rehmus-Forc A. Stress corrosion cracking of a 60MW steam turbine rotor. *Eng Fail Anal.* **2015**;51:55–68.
- [15] Sz JK, Segura JA, Gonzalez R G, et al. Failure analysis of the 350 MW steam turbine blade root. *Eng Fail Anal.* **2009**;16(4):1270–1281. DOI:[10.1016/j.engfailanal.2008.08.015](https://doi.org/10.1016/j.engfailanal.2008.08.015)
- [16] Kubiak SJ, Gonzalez GR, Juarez DR, et al. An investigation on the failure of an L-O steam turbine blade. *J Fail Anal Prev.* **2004**;4(3):47–51.
- [17] Xue F, Wang ZX, Zhao WS, et al. Fretting fatigue crack analysis of the turbine blade from nuclear power plant. *Eng Fail Anal.* **2014**;44:299–305.
- [18] Puspitasari P, Andoko A, Kurniawan P. Failure analysis of a gas turbine blade: a review. *IOP Conf Ser Mater Sci Eng.* **2021**;1034(1):012156.

- [19] Rajendra KD, Arakerimath R. ICRRM 2019 – system reliability, quality control, safety, maintenance and management,” ICRRM 2019 – syst. Reliab Qual Control Safety, Maint Manag. 2020;46–52. DOI:10.1007/978-981-13-8507-0
- [20] Enomoto Y. Steam turbine retrofitting for the life extension of power plants. Japan: Elsevier Ltd; 2017.
- [21] Connor N, “What is LP turbine - low-pressure steam turbine - definition.” 2019, [Online]. Available: <https://www.thermal-engineering.org/what-is-lp-turbine-low-pressure-steam-turbine-definition/> .
- [22] Ahmad M, Casey M, Sürken N. Experimental assessment of droplet impact erosion resistance of steam turbine blade materials. Wear. 2009;267(9–10):1605–1618.
- [23] Tian L, Hai Y, Qingyue Z, et al. Non-destructive testing techniques based on failure analysis of steam turbine blade. IOP Conf Ser Mater Sci Eng. 2019;576(1):012038.
- [24] Tanuma T. Development of last-stage long blades for steam turbines. Japan: Elsevier Ltd; 2017.
- [25] “En_10088-3-stainless-steel.”.
- [26] A. Specification. Standard test method for liquid penetrant examination. Se-165. 2001;(165): 464–488.
- [27] A. A.-751 ASTM A-751 E-23. ASTM A-751. standard test methods, practices, and terminology for chemical analysis of steel products. ASTM E-23. 2011;ASTM A-751:1–6.
- [28] Vander Voort GF, Lucas GM, Manilova EP. Metallography and microstructures of stainless steels and maraging steels. Metallogr Microstruct. 2018;9(c):670–700.
- [29] ASTM A370. Standard test methods and definitions for mechanical testing of steel products. ASTM Int. 2004;01.03(Rapproved): 1–48. DOI:10.1520/A0370-16.2.
- [30] Vander Voort GF, Lucas GM. Metallography and microstructures of stainless steels and maraging steels. 2004;9(c). DOI:10.1361/asmhba0003767.
- [31] Pineault JA, Belassel M, Brauss ME. X-ray diffraction residual stress measurement in failure analysis. Fail Anal Prev. 11. ASM International, Jan. 01, 2002, 10.31399/asm.hb.v11.a0003528.
- [32] Theiner WA. “Micromagnetic techniques,” struct. residual stress anal. by nondestruct. Methods. 1997;564–589. DOI:10.1016/b978-044482476-9/50019-0
- [33] Noyan Ismail Cevdet, Cohen Jerome Bernard. Residual stress: measurement by diffraction and interpretation. New York: Springer; 2013.
- [34] Cullity BD. Elements of X-ray diffraction. University of Georgia: Addison-Wesley Publishing; 1956.
- [35] Pineault PMLJA, Belassel M, Brauss ME. X-ray diffraction residual-stress.Pdf. ASM Handbook. 2021;11. DOI:10.31399/asm.hb.v11.a0006768.
- [36] Stress AR, Technique M, Turbine FOR, et al., “A residual stress measurement technique for turbine blade dovetails,” Proc. ASME Turbo Expo 2011 GT2011 June 6-10, 2011, Vancouver, Br. Columbia, Canada, pp. 1–8, 2017.
- [37] “Barkhausen noise analysis,” Cond-M. Hill Engineering, 2004, [Online]. Available: <https://hill-engineering.com/barkhausen-noiseanalysis/>.

Journal Pre-proof

Hydrogenic impurity effect on the optical properties of $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire under terahertz field

Priyanka, Rinku Sharma, Manoj Kumar, Pradumn Kumar



PII: S2773-0123(22)00264-3

DOI: <https://doi.org/10.1016/j.micrna.2022.207451>

Reference: MICRNA 207451

To appear in: *Micro and Nanostructures*

Received Date: 26 August 2022

Revised Date: 9 November 2022

Accepted Date: 17 November 2022

Please cite this article as: Priyanka, R. Sharma, M. Kumar, P. Kumar, Hydrogenic impurity effect on the optical properties of $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire under terahertz field, *Micro and Nanostructures* (2022), doi: <https://doi.org/10.1016/j.micrna.2022.207451>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2022 Published by Elsevier Ltd.

Hydrogenic impurity effect on the optical properties of $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire under terahertz field.

Priyanka¹, Rinku Sharma^{1,*}, Manoj Kumar^{2,*} and Pradumn Kumar³

¹Department of Applied Physics, Delhi Technological University, Delhi-110042, India

²Department of Physics, Govt. College for Women, Jind, 126102, India

³Department of Physics, University of Delhi, Delhi-110021, India

**Email-rinkusharma@dtu.ac.in, manojmalikdu@gmail.com*

Abstract

In this study, the effect of impurity factor on the optical absorption coefficients, refractive index changes, second harmonic generation, and third-harmonic generation for the intersubband transitions is explored between the electronic states of $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire initiated by the symmetric parabolic potential. The system is conquered by the presence of an intense electric field, magnetic field, and Rashba spin-orbit interaction. For the linear and non-linear optical absorption coefficients, refractive index, second harmonic generation, and third harmonic generation coefficient, the analytical expressions are obtained with the assistance of the compact density-matrix approach. The arithmetical outcomes illustrate the optical properties are significantly intuitive to the concentration of impurity and can be controlled by this parameter. The variation in the magnitude and position of peaks via impurity factor indicates the opportunity in the mechanism of optical non-linearity in the quantum wire and also, helps in the optical non-linearity tuning which has device application.

Keywords: impurity factor, Quantum wire, optical absorption coefficients, second harmonic generation, and third harmonic generation.

1. Introduction

In the few last years, low-dimensional nanostructures like: quantum dots, quantum wires, and quantum well have considerable devotion for their alluring potential applications in optical and laser technology [1–5]. In these nanostructures, the charge carrier having quantum confinement accompany the formation of energy states in the discrete form and enhance the density of states at definite energies which leads to variation in the optical spectra and helps in the evolution of novel properties. The optical properties viz. absorption spectra and refractive index changes (RICs) in the low-dimensional nanostructures take fascinated courtesy due to their high-level performance. One of the most arduously explored low-dimension nanostructures is quantum

wire, especially in the theoretical and experimental research of the impurity effect on their optical properties[6–13]. Moreover, the tunability of the energy dispersion by the intense magnetic field, electric field, impurity factor, and Rashba spin-orbit interaction in the quantum wire has made a fruitful role in examining non-linear and linear properties for applications of novel devices[14,15].

A region of large potency is the spin-based phenomena in quantum wire for its profusion of the physically observable phenomena and has an encouraging future for spin-related electronic devices with a high degree of functionality, fast speed, and low power consumption [16–18]. Fortunately, spin physics in low-dimension nanostructures utilizes electron spin degree of freedom as a chunk of information instead of an electron-charge and assurance potential for imminent spin-based application devices that are smaller, faster, and more influential ergo those that are presently available. Specifically, the spin-orbit interaction (SOI) has fascinated enthusiasm as it permits optical spin detection and spin orientation. Moreover, SOI is assumed as an opportunity for controlling and manipulating the state of an electron via gate voltage. The type of SOI, eminent in a certain quantum wire heterostructure is the Rashba SOI. The Rashba SOI comes out in the picture due to the confinement potential which explains the quantum wire is a function perpendicular to the 2-dimension electron gas (2 DEG). This conjectures a structural inversion symmetry that easily can be tweaked by the external gate potential and also control the spin-related phenomena [19–23].

Although more study has been admiring to research the effect of Rashba SOI on the optical and physical properties of the quantum wire. The optical properties have been explored by lots of investigators in both theoretical and experimental [24–34]. Some external constraints like the intense electric field, magnetic field, Rashba SOI, impurity factor (x), etc play a significant role to influence the optical properties of quantum wire [35–38]. M. Santhi et al. [39] have vastly explored the effects of hydrogenic impurity on the linear and third-order nonlinear optical absorption coefficients (ACs), and third harmonic generation (THG) optical properties of GaAs/GaAlAs quantum wire. However, when the electron is bound with the impurity atom within the presence of external perturbations shows rudimental character in grasping the electro-optical properties of hydrogenic impurities in nanostructures. The inclusion of hydrogenic impurities in nanostructures will excessively change the electrical and optical properties and affect the quantum device concert.

Martinez et al.[40] have explored the hydrostatic pressure and temperature effects on the hydrogenic impurity which relates cross-section of photoionization and impurity binding

energy in $\text{Ga}_{1-x}\text{Al}_x\text{As}/\text{GaAs}$ quantum wire. Zeiri et al. [41] discussed the linear and nonlinear susceptibility of self-organized in $\text{GaN}/\text{Ga}_{1-x}\text{Al}_x\text{N}$ quantum wire. Embroidering the energy dispersion is enthusiastic to yield the expedient optoelectronic device from the belief of confinement in the low-dimension nanostructures, transmission properties, and tunable emission are assumed to be significant features for them. Moreover, the optical properties are studied by many researchers.

With this motivation, the current work aims to represent the behavior of linear and third-order nonlinear optical (ACs), an absolute change in refractive index coefficients (RICs), second harmonic generation (SHG), and THG associated with the $\text{Ga}_{1-x}\text{Al}_x\text{As}$ Quantum wire. The following is the work's alignment: In a nutshell, Section 2 represents the theoretical model of our system. The corresponding outcomes and discussion are noted in Section 3. Belatedly, Section 4 covers the study's foremost conclusions.

2. Theory

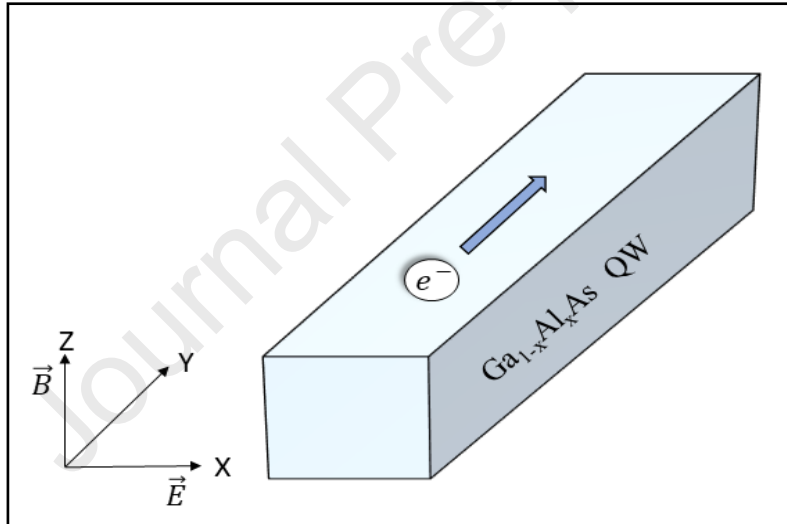


Figure 1. The diagram of $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire within the existence of an external magnetic electric field, and Rashba SOI.

The Hamiltonian equation for $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire under the influence of electric field (\mathbf{E}) along X-direction, intense magnetic field (\mathbf{B}) along Z-direction, and Rashba SOI, respectively is written by

$$\hat{H}_{oe} = -\frac{1}{2m_e^*(x,P,T)} (p_X + (p_Y + eBX)^2 + \frac{1}{2}m_e^*(x,P,T)\omega_{1e}^2 X^2 + eEX + \frac{1}{2}g^*\mu_B B\sigma_Z + \frac{\alpha}{\hbar} [\sigma_X(p_Y + eBX) - \sigma_Y p_X]), \quad (1)$$

Where p_x and p_y represents the momentum component of an electron in X and Y direction. ω_{1e} is called effective cyclotron frequency. α , g^* , μ_B , σ_x , σ_y , and σ_z are known as Rashba SOI factor, Landé factor, Bohr magneton and Pauli matrices along x, y and z direction, respectively. $m_e^*(x, P, T)$ is known effective mass represented by [22]

$$m_e^*(x, P, T) = m_0 \left[\left(\frac{1}{\Delta_{oj}(x) + E_g^j(x, P, T)} + \frac{2}{E_g^j(x, P, T)} \right) \frac{\pi_j^2(x)}{3} + \delta_j(x) + 1 \right]^{-1}, \quad (2)$$

Here, m_0 is known as free electron mass, x is used for impurity factor, $\Delta_{oj}(x)$ and $\pi_j(x)$ are known as valence band spin-orbit coupling and inter-band matrix element, respectively ($\pi_j^2(x) = (-6290x + 28900)\text{meV}$ and $\Delta_{oj}(x) = (-66x + 341)\text{meV}$). When the remote-band effects are examined over $\delta_j(x)$ the parameter is given by [42]:

$$\delta_j(x) = 4.938x^2 + 0.488x - 3.935, \quad (3)$$

In eq. (2), $E_g^j(x, P, T)$ is called conduction band's an energy gap and is expressed via

$$E_g^j(x, P, T) = p_j + q_j x + r_j x^2 + s_j P - \frac{\beta_j T^2}{\gamma_j + T}. \quad (4)$$

where the parameters' values p_j , q_j , r_j , s_j , γ_j and β_j are given by 1519.4 meV, 1360 meV, 220meV, 10.7 meV/kbar, 204 K and 0.5405 meV/K, respectively. And the values of these parameters are determined with the assistance of photoluminescence.

Eq. (1) can be rewritten as write $\hat{H}_{oe} = \hat{H}_i + \hat{H}_R$, where

$$\begin{aligned} \hat{H}_i = & -\frac{p_x^2}{2m_e^*(x, P, T)} + \frac{1}{2}m_e^*(x, P, T)\omega_{1e}^2(X - X_{oe})^2 - \frac{e^2 E^2}{2m_e^*(x, P, T)\omega_{1e}^2} + \frac{\omega_0^2 \hbar^2 k_y^2}{\omega_{1e}^2 2m_e^*(x, P, T)} - \\ & \frac{e^2 E B \hbar k_y}{m_e^*(x, P, T)\omega_{1e}^2} + \frac{1}{2}g^* \mu_B B \sigma_z, \end{aligned} \quad (5)$$

And

$$\hat{H}_R = \alpha(\sigma_x \left(k_y + \frac{eBX}{\hbar} \right) - i\sigma_y \frac{d}{dX}). \quad (6)$$

Where $X_{oe} = -\left(\frac{eE}{m_e^*(x, P, T)\omega_{1e}^2} + \frac{eB\hbar k_y}{m_e^{*2}(x, P, T)\omega_{1e}^2} \right)$ is recognized as guiding centre coordinate.

For the complete solution within the presence of external fields and Rashba SOI. As a consequence of complex coupling in the \hat{H}_{oe} , we assume that there is no analytic solution of the Schrödinger can be comes out, aside from the some trival limits. Consequently, we must solve the Schrödinger equation numerically to achieve an insight about the interplay of SOI.

So, expanding the $\varphi(x) = \sum_{n\sigma} a_{n\sigma} \Psi_{n\sigma}(x)$, the Hamiltonian ' H_{oe} ' eigenvalue equation can be written as: -

$$\sum_{n\sigma} a_{n\sigma} (E_{n\sigma} - E) \Psi_{n\sigma}(x) + \sum_{n\sigma} a_{n\sigma} \Psi_{n\sigma}(x) = 0, \quad (7)$$

$$(E_{n\sigma} - E) a_{n\sigma} + \sum_{n'\sigma'} \langle \Psi_{n\sigma} | \hat{H}_R | \Psi_{n'\sigma'} \rangle = 0, \quad (8)$$

Where the matrix elements' 2nd term of Eq. (8) is calculated as:

$$\begin{aligned} \langle n\sigma | \hat{H}_R | n'\sigma' \rangle = & \alpha \left[\left(1 - \frac{\omega_c^2}{\omega_{1e}^2} \right) k_y - \frac{\omega_c e E}{\hbar \omega_{1e}^2} \right] \delta_{n,n'} \delta_{\sigma,\sigma'} + \frac{\alpha}{c_i} \left[\left(\frac{\omega_c}{\omega_{1e}} + \sigma \right) \sqrt{\frac{n+1}{2}} \delta_{n,n'-1} + \right. \\ & \left. \left(\frac{\omega_c}{\omega_{1e}} - \sigma \right) \sqrt{\frac{n}{2}} \delta_{n,n'+1} \right] \delta_{\sigma,-\sigma'}, \end{aligned} \quad (9)$$

Now, our problem reduces to finding an appropriate numerical procedure for finding the various quantum states. The energies eigenvalues and their corresponding eigenvectors are obtained, and we can apply the analytical expressions via the perturbation method and density matrix approach for optical ACs, RICs and THG. For calculating these optical properties, we assume a circularly polarised electromagnetic (EM) field having incident photon frequency (ω) along the Z-direction, then the interaction within the system is given by[43,44]

$$\mathbf{E}(t') = \frac{E_0(t')}{\sqrt{2}} (\hat{e}_X + \hat{e}_Y), \quad (10)$$

Here, the terms \hat{e}_X and \hat{e}_Y represents unit vector for X and Y directions respectively, and $E_0(t')$ is

$$E_0(t') = E_0 \cos(\omega t') = \tilde{E} e^{-i\omega t'} + \tilde{E}^* e^{i\omega t'}, \quad (11)$$

Therefore, the system is excited through an electromagnetic field. By applying, the Bloch theorem for symmetry, for $\Psi_{nm,k}$ and $\Psi_{n'm',k'}$ states the dipole transition moment can be written $\langle \Psi_{nm,k} | qX | \Psi_{n'm',k'} \rangle = \delta_{k,k'} \langle \varphi_{n,m} | qX | \varphi_{n',m'} \rangle$,

Where δ is known as the Kronecker delta function.

With the help of a compact iterative procedure and density-matrix method, we will drive the mien of THG susceptibility for the 2D model of the isotropic harmonic oscillator. The term $\bar{\rho}$ represents the electron density matrix. And the time-dependent Liouville-equation is

$$\frac{\partial \bar{\rho}_{ij}}{\partial t} = \frac{1}{i\hbar} [H_o - qXE(t'), \bar{\rho}]_{ij} - [i_{ij}(\bar{\rho} - \bar{\rho}^{(o)})]_{ij} \quad (13)$$

$\bar{\rho}^{(o)}$ is known as a density-matrix operator for an unperturbed system, $[ij]$ called as phenomenological operator. It is hypothesized that $[ij]$ is a diagonal matrix known as relaxation rate. Equation (13) can be resolved with the help of the standard iterative method [45] then

$$\bar{\rho}(t') = \sum_n \bar{\rho}^{(n)}(t'), \quad (14)$$

With

$$\frac{\partial \bar{\rho}^{(n+1)}}{\partial t'} = \frac{1}{i\hbar} \{ [H_o, \bar{\rho}^{(n+1)}]_{ij} - i\hbar [ij] \bar{\rho}_{ij}^{(n+1)} \} - \frac{1}{i\hbar} [qr, \bar{\rho}^{(n)}]_{ij} E(t'). \quad (15)$$

The system's electronic polarization can also be expanded phenomenologically as an electric field series. Therefore, the three orders of electronic polarization $P(t')$ are expressed by

$$P(t') = \left(\epsilon_o \chi_{\omega}^{(1)} \tilde{E} e^{-i\omega t'} + \epsilon_o \chi_o^{(2)} |\tilde{E}|^2 + \epsilon_o \chi_{2\omega}^{(2)} \tilde{E}^2 e^{-2i\omega t'} + \epsilon_o \chi_{3\omega}^{(3)} |\tilde{E}|^2 \tilde{E} e^{-i\omega t'} + \epsilon_o \chi_{3\omega}^{(3)} \tilde{E}^3 e^{-3i\omega t'} \right) + c.c., \quad (16)$$

Where $\chi_{\omega}^{(1)}$, $\chi_{2\omega}^{(2)}$, $\chi_o^{(2)}$ and $\chi_{3\omega}^{(3)}$ are the linear susceptibility, SHG, optical rectification, and THG, respectively. The ϵ_o is known as Vacuum dielectric constant. For n^{th} -order electronic polarization is written by

$$P^{(n)}(t') = \frac{1}{V} \text{Tr}(\bar{\rho}^{(n)} qr), \quad (17)$$

The term V is used for volume of interaction and Tr (trace) represents the summation over diagonal elements of the matrix $\bar{\rho} qX$. The trace (Tr) and susceptibility $\chi(\omega)$ are correlated to absorption coefficient $\alpha(\omega)$:

$$\alpha(\omega) = \omega \sqrt{\frac{\mu}{\epsilon_r}} \text{Im}[\epsilon_o \chi(\omega)], \quad (18)$$

Where ϵ_o and μ is known as the real part of the relative permittivity (ϵ_r) and permeability, respectively. Now, the linear and non-linear optical AC expression can be written as [46–49];

$$\alpha^{(1)}(\omega) = \hbar \omega \sqrt{\frac{\mu}{\epsilon_r}} \frac{N_c \Gamma_{if} |M_{if}|^2}{\hbar^2 \{ (\omega_{fi} - \omega)^2 + \Gamma_{if}^2 \}}, \quad (19)$$

And

$$\alpha^{(3)}(\omega, I) = -\hbar \omega \sqrt{\frac{\mu}{\epsilon_r}} \frac{I}{2\epsilon_o n_r c} \times \frac{4N_c \Gamma_{if} |M_{if}|^2}{\hbar^4 \{ (\omega_{fi} - \omega)^2 + \Gamma_{if}^2 \}} \times \left[\frac{|M_{if}|^2}{(\omega_{fi} - \omega)^2 + \Gamma_{if}^2} + \frac{(M_{ff} - M_{ii})^2 (3\omega_{fi}^2 - 4\omega_{fi}\omega) (\omega_{fi}^2 - \Gamma_{if}^2)}{4(\omega_{fi}^2 + \Gamma_{if}^2) \{ (\omega_{fi} - \omega)^2 + \Gamma_{if}^2 \}} \right] \quad (20)$$

In eq. (11), $\hbar\omega$ and I are the energy of incident photon and intensity of incident light respectively, N_c is the carrier density and n_r represents the refractive index of quantum wire material, Γ_{if} is called relaxation time & the subscript i and f are used for initial and final states. $\omega_{if} = (E_f - E_i)/\hbar$ is known as transition frequency (where, E_i and E_f is the energy for the initial and final state). M_{if} is known as transition matrix element and determined by $M_{if} = |\langle\phi_i|X|\phi_f\rangle|$.

The total optical absorption coefficient is written as[50–52]

$$\alpha_T(\omega, I) = \alpha^{(1)}(\omega) + \alpha^{(3)}(\omega, I). \quad (21)$$

The linear and third-order nonlinear refractive index for optical transitions are obtained by

$$\frac{\Delta n_r^{(1)}}{n_r} = \frac{N_c}{2n_r^2\epsilon_0} \frac{|M_{if}|^2(\omega_{fi}-\omega)}{\hbar\{(\omega_{fi}-\omega)^2 + \Gamma_{if}^2\}}, \quad (22)$$

$$\begin{aligned} \frac{\Delta n_r^{(3)}(\omega, I)}{n_r} = & \frac{N_c}{2n_r^3\epsilon_0} \frac{I\mu c|M_{if}|^2(\omega_{fi}-\omega)}{\hbar^3(\omega_{fi}-\omega)^2 + \Gamma_{if}^2} \times \\ & \left[\frac{|M_{if}|^2}{(\omega_{fi}-\omega)^2 + \Gamma_{if}^2} + \frac{2\Gamma_{if}(M_{ff}-M_{ii})}{(\omega_{fi}-\omega)^2 + \Gamma_{if}^2} \times \{(\omega_{fi}-\omega)[\omega_{fi}(\omega_{fi}-\omega) + \Gamma_{if}^2] - \Gamma_{if}^2(2\omega_{fi}-\omega)\} \right] \end{aligned} \quad (23)$$

The total refractive index is

$$\frac{\Delta n_r(\omega, I)}{n_r} = \frac{\Delta n_r^{(1)}}{n_r} + \frac{\Delta n_r^{(3)}(\omega, I)}{n_r}. \quad (24)$$

And now, we focus on the SHG and THG (i.e., from the term of oscillating with 4ω , we only cogitate the third-order contribution). With the help of the iterative method and compact density-matrix approach, the SHG and THG per unit volume are given as[12,53,54];

$$\chi_{2\omega}^{(2)} = -\frac{e^3 n_0}{\epsilon_0 \hbar^3} \frac{M_{12}M_{23}M_{31}}{(\omega - \omega_{21} + i\Gamma_{21})(2\omega - \omega_{32} + i\Gamma_{32})}, \quad (25)$$

$$\chi_{3\omega}^{(3)} = -\frac{e^4 n_0}{\epsilon_0 \hbar^3} \frac{M_{12}M_{23}M_{34}M_{41}}{(\omega - \omega_{21} + i\Gamma_{21})(2\omega - \omega_{31} + i\Gamma_{31})(3\omega - \omega_{41} + i\Gamma_{41})}. \quad (26)$$

The results and their implications are presented in the next section.

3. Result and discussion

In this investigation, the effects of hydrogenic impurity on the linear and the third-order nonlinear optical ACs, RICs, SHG, and THG in a typical $\text{Ga}_{1-x}\text{Al}_x\text{As}$ within the presence 2-DEG have been studied. The physical constraints used for arithmetical computation are [12,55]: $n_0 = 10^{16}\text{cm}^{-3}$, $n_r = 3.2$, $\Gamma_{12} = 1/T_{12}$ where $T_{12} = 0.2\text{ps}$ and $\mu = 4\pi \times 10^{-7}\text{Hm}^{-1}$.

In fig. 2(a-b), the changes in the effective mass with the variation in impurity factor (x) for different values of temperature and pressure viz. 100K, 200K, and 300K and 10 kbar, 15 kbar

and 20 kbar respectively, are presented. As the impurity factor (x) increases, the effective mass is also increased for each value of temperature and pressure. This is due to the strong dependence of impurity on the effective mass. When the impurity factor (x) is changed from 0 to 0.3, the behavior of the quantum wire in terms of effective mass is counter-intuitive when the rise in impurity factor probably enhances the effective mass. This nature helps in the study of the influences of hydrogenic impurity (x) on optical properties.

In Fig. 3 (a), energy gap variation with impurity factor (x) has been shown. As the hydrogenic impurity increases, the energy gap between the two subsequent levels is also increased. This enhancement in the energy gap due to impurity yields the change in optical properties of the $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire. Fig. 3 (b) shows the matrix elements variation with the impurity factor (x). The matrix elements reduce with an increment in hydrogenic impurity factor (x). From the equations (19-25), we can easily see the dependence of matrix elements on the linear and third-order nonlinear ACs, RICs and THG therefore, when the matrix element changes these parameters also changes.

Fig. 4 (a-c), demonstrates the linear $\alpha^{(1)}$, the third-order nonlinear ACs $\alpha^{(3)}$ and the total optical ACs $\alpha^{(T)}$ as a function of the incident photon energy for the various value of the hydrogenic impurity factor (x). The magnetic field, electric field, Rashba SOI, and incident laser intensity are fixed at 1T, 50kV/m, 35meV and $2.04 \times 10^7 \text{W/m}^{-2}$, respectively. Generally, the $\alpha^{(1)}$ illustrates resonance at the zero thwarts which denotes when the energy difference between the subsequent states becomes cognate with an incident photon. Similarly, in the case of $\alpha^{(3)}$ with a negative sign. It is initiated that when the x is raised from 0.1 to 0.3, the maxims of $\alpha^{(1)}$ and $\alpha^{(3)}$ shows blue shifting. However, this effect is abetted via the decrement in the value of $\alpha^{(1)}$ and $\alpha^{(3)}$, with $\alpha^{(3)}$ entity effected firmly as compared to linear optical AC ($\alpha^{(1)}$). Whereas, the blue shifting takes place due to the enrichment in the energy gap between the subsequent states of energy which can easily be understood by fig. 3 (a). This in turn consequences in further energy spacing between the ensuing states. Moreover, the impurity factor (x) effect on the $\alpha^{(3)}$ strongly as compared to $\alpha^{(1)}$ whose results illustrate the rare shape in the $\alpha^{(T)}$ curve. It can be noted that with an increment in the value of x there is a decrement in the height of peak maxims. This happened as a consequence of the small value of dipole matrix elements (M_{ij}) when the x rise. In accordance, with the eq. (19) and eq. (20) the $\alpha^{(1)}$ and $\alpha^{(3)}$ resonant peaks values are proportional to $M_{12}^2 E_{12}$ and $M_{12}^4 E_{12}$, respectively. Therefore, the consequence of M_{12} is dominant, resulting in a decrease in the magnitude of the

$\alpha^{(3)}$ term more than the $\alpha^{(1)}$. It is manifest that the $\alpha^{(1)}$, $\alpha^{(3)}$ and $\alpha^{(T)}$ peaks diminution and shifts in the direction of the higher energies as the impurity boosts.

In the optical studies of quantum wire, RICs play a significant role. In Fig. 5 (a-c), the linear, third order non-linear, and total RICs are plotted as a function of the incident photon energy for various hydrogenic impurity factor (x). As demonstrated from these plots, the linear RICs increase gradually with the incident photon energy and come to an extreme value. This brings out to the normal dispersion for any frequency of incident photo where $\frac{dn}{d\omega} > 0$. However, the energy of the photon approaches threshold energy, the dispersion $\frac{dn}{d\omega}$ in the RICs change its sign. At every resonant frequency of a quantum wire, this anomalous dispersion is defined by $\frac{dn}{d\omega} < 0$. In this area, photons are sturdily absorbed and behave like an absorption band. In Fig. 5(a-c), the anomalous dispersion region moves towards large photon energies (i.e; blueshift) as a contrast to the low value of impurity factor cases. The root cause for the resonance shifting is an enhancement in the energy gap between the two electronic states among which an optical transition takes place. As the third-order nonlinear RIC has a negative sign therefore it takes an enfeebling effect on the total RIC. Therefore, it has been discovered that the impurity, as a tunable parameter plays a significant role in supervising the optical properties of quantum wire. Fig. 6, represents the SHG as a function of incident photon energy for the various value of impurity ($x= 0.1, 0.2$, and 0.3). It can be perceived from the plot that the resonant peaks of the SHG decrease and shift towards large value of photon energy (blue shift) when the impurity factor (x) increases. SHG figure has only one resonant peak due to the equally spaced energy spectrum. Additionally, the intensity in Fig.6 is much larger with the comparison of resonant peaks in the situation of unequal spacing of the energy spectrum. The changes in the amplitude and position of the peaks are due to the change in the geometric factor of the quantum wire such as energy gap and matrix elements.

Fig. 7 shows that, for various values of the hydrogenic impurity factor $x=0, x=0.1$, and $x=0.3$, the THG is plotted as a function of incident photon energy. For the low energy region, the triple resonance condition is not perfectly achieved due to the weaker confinement cause the energy separation between the states unequally and additionally, for high energy the transition energy between subbands decreases, and the overlaps of states will rise. When the hydrogenic impurity (x) is considered, the peaks of the THG coefficients show a blue shift. The magnitude of the peak for each curve reduces with the rise in impurity factor (x), due to allied with the product of matrix elements. This is due to the energy gap between two states containing hydrogenic

impurity being larger than the energy gap between two states containing no hydrogenic impurity. This means that the THG drops and moves towards higher energies when the effect of impurity is taken into account.

The THG coefficient as a function of hydrogenic temperature and pressure when $x=0.1$, 0.2 , and $x=0.3$ are illustrated in Fig. 8(a) and (b), respectively. From fig. 8(a), it can be observed that as the temperature enhances there is an increment in height of the resonant peaks for $x=0.1$, 0.2 , and 0.3 . When the hydrogenic impurity(x) rises from 0.1 to 0.3 the resonant peak amplitude reduces. In fig. 8(b), as the pressure enhances then the THG coefficient value decreases, unlike in the previous case. The enhancement in the impurity factor (x) parameter for supplementing values of P is reflected in the fall of the resonance peak amplitude. The amplitude of the resonant peak is calculated by the product of the transition matrix elements. These optical effects can be exploited as a probe for the precise mechanism such as optical nonlinearity in quantum wires and also, used in optical-magneto instruments.

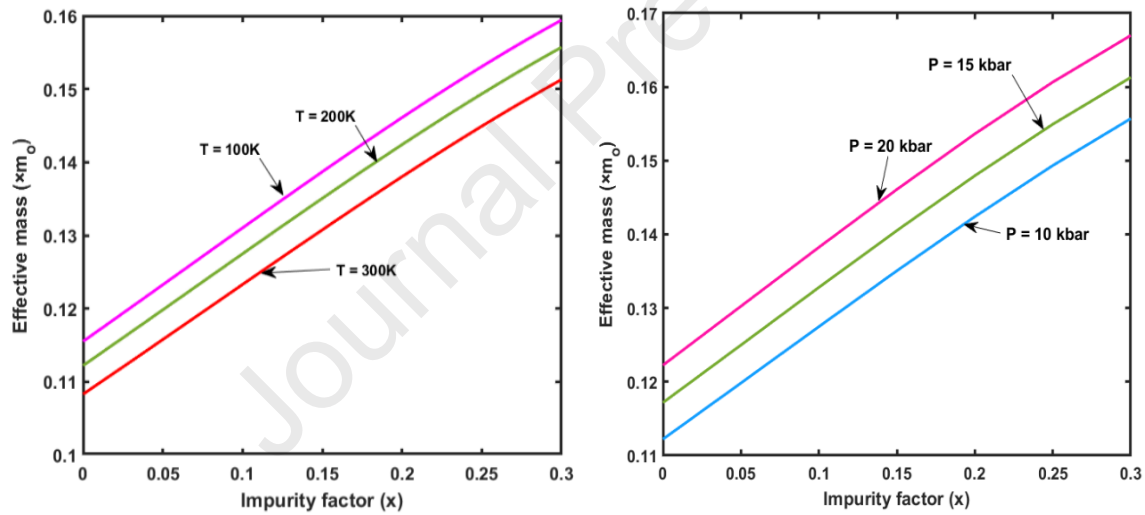


Figure 2. Effective mass variation with the hydrogenic impurity (x) for the $Ga_{1-x}Al_xAs$ quantum wire at $T= 100K$, $200K$ and $300K$.

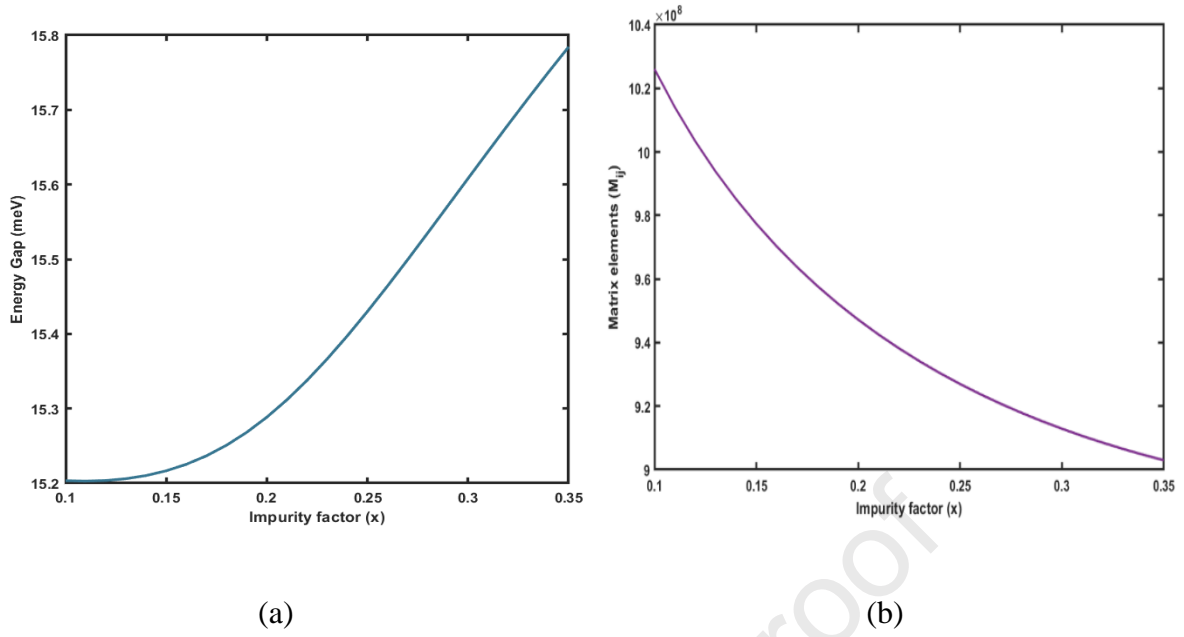
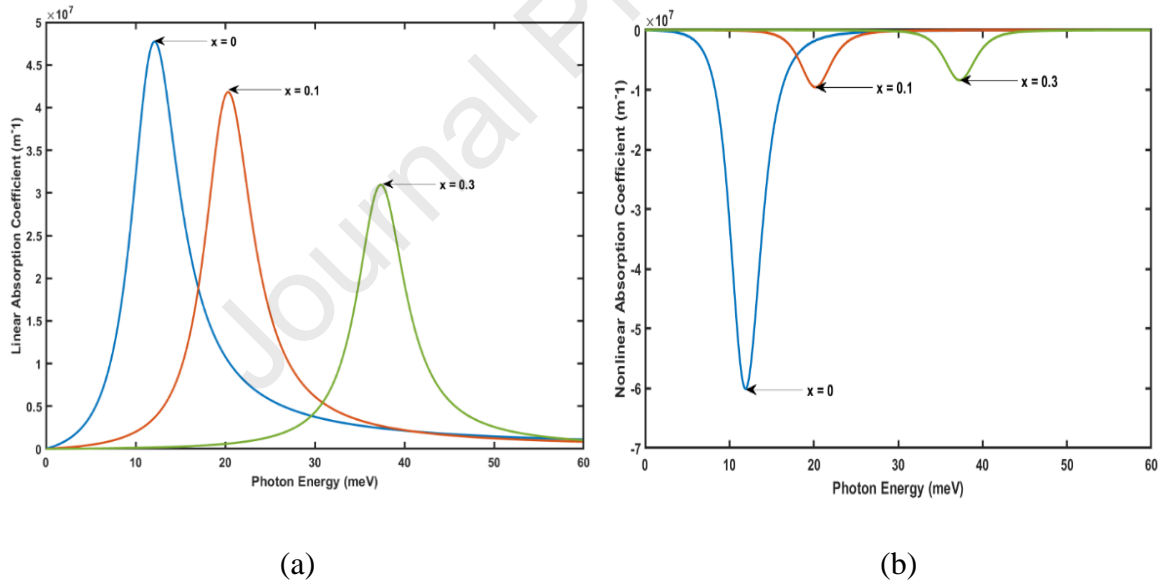
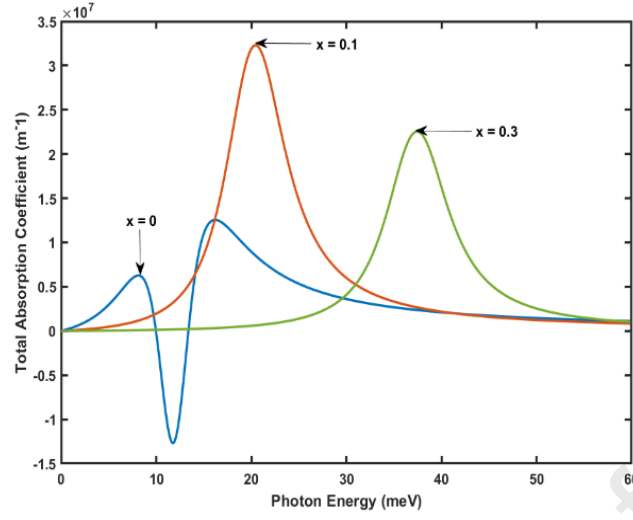


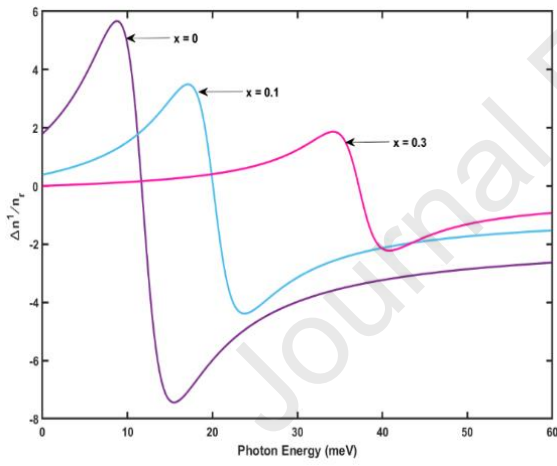
Figure 3. (a) Energy gap between the two subsequent levels as a function of impurity factor (x), and (b) Matrix element as a function of impurity factor (x) for the $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire.



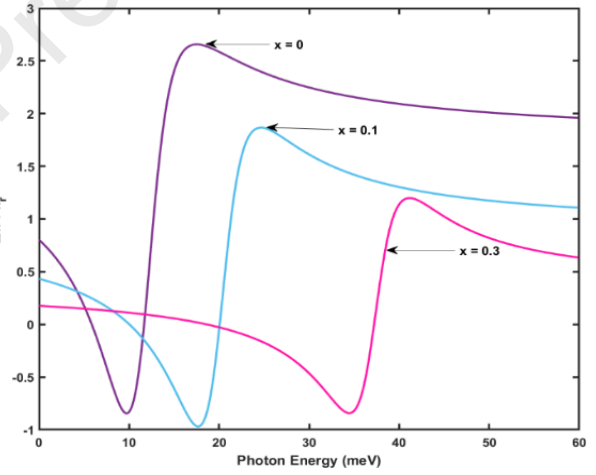


(c)

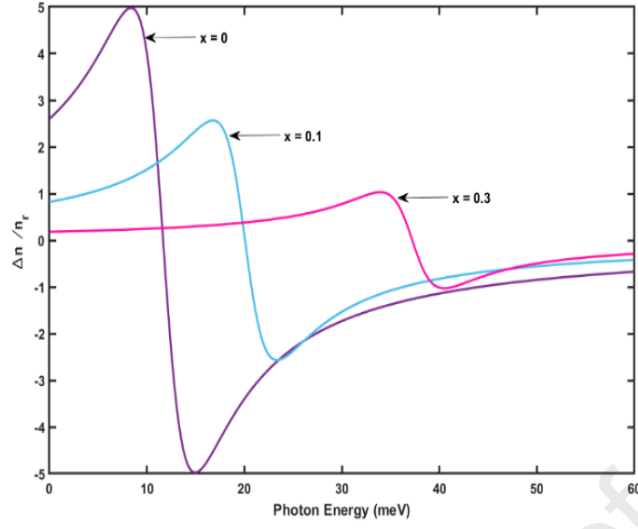
Figure 4. (a) Linear optical ACs, (b) Third-order nonlinear optical ACs, and (c) total optical ACs as a function of Photon energy with the various values of the impurity.



(a)



(b)



(c)

Figure 5. (a) The linear RIC, non-linear RIC, and total RIC as a function of the incident photon energy for the various value of hydrogenic impurity.

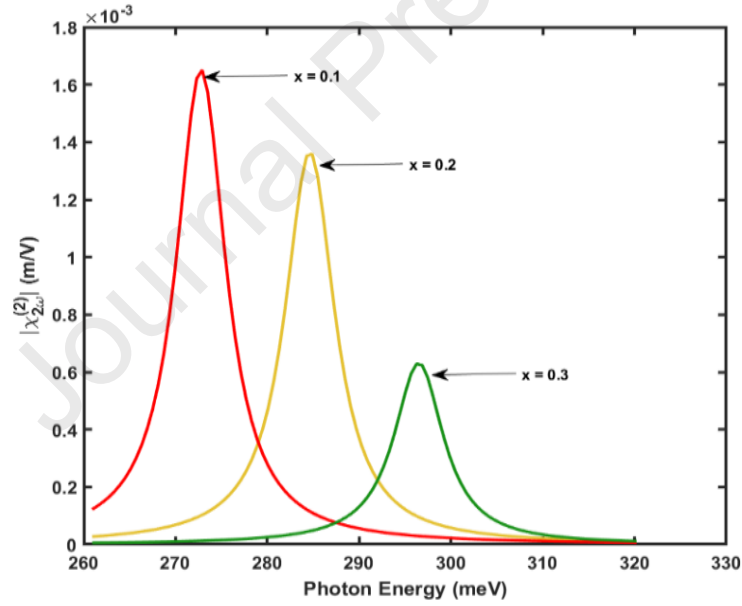


Figure 6. SHG as a function of photon energy when $x=0.1$, 0.2 , and 0.3 .

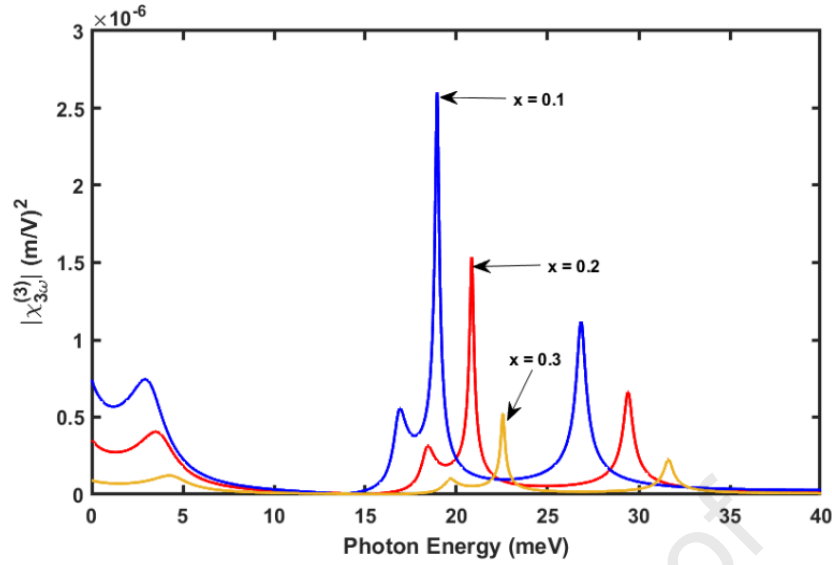


Figure 7. THG as a function of incident photon energy for the various value of hydrogenic impurity.

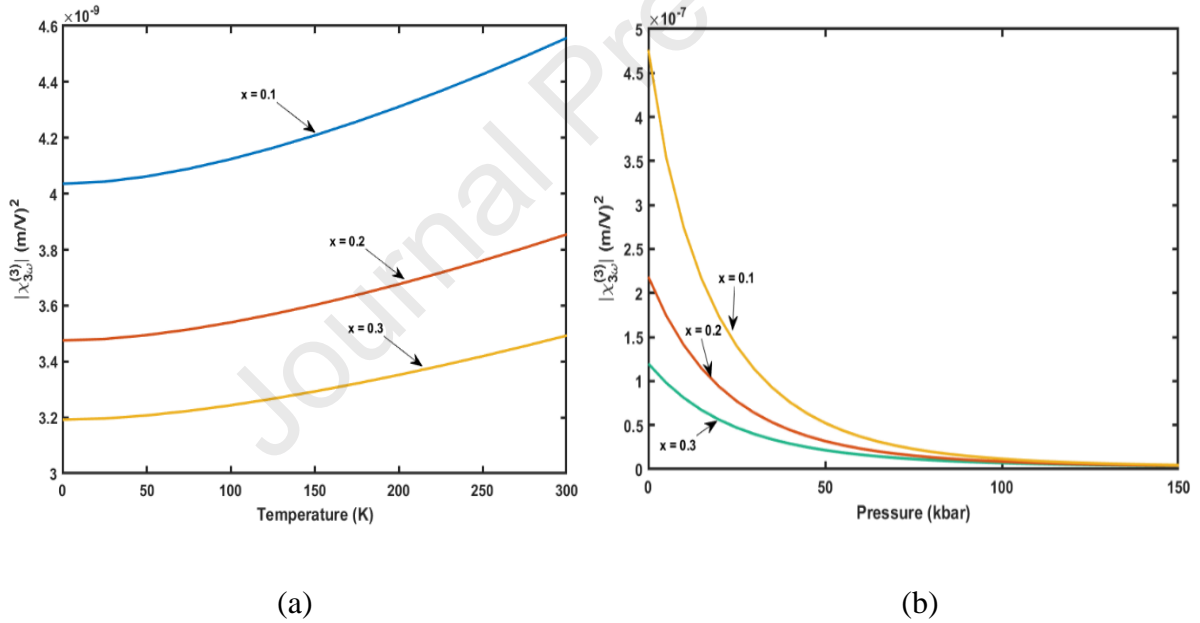


Figure 8. (a) THG coefficient for the $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire as a function of hydrogenic temperature with $x=0.1, 0.2$ and 0.3 at $P=15$ kbar, and (b) THG coefficient as a function of pressure with $x=0.1, 0.2$ and 0.3 at $T=300$ K.

4. Conclusion

In this paper, we attention to the behavior of optical properties for the $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire and how its changes with the various value of hydrogenic impurity. The alteration in linear

and third-order nonlinear optical ACs, RICs, SHG, and THG by the impurity factor under the presence of the intense magnetic field, electric field, and Rashba SOI are demonstrated. We experiential the energy gap and the dipole moment matrix element (M_{12}), are strongly affected by the hydrogenic impurity. Due to this optical ACs, RICs, SHG, and THG of the system change. It was noticed that the enhancement in hydrogenic impurity produces blue shifts in optical properties. However, hydrogenic impurity increases the energy gap and decreases the matrix elements, therefore as a result, the peaks of optical ACs, RICs, SHG, and third harmonic generation shift to higher photon energy. It is also concluded that the impurity factor can be explored as tuning tool for varying the optical properties via the introduction of spacing between the confined energy levels. It is presumed that impurity can show a significant role in the optical properties of semiconductor quantum wire. Therefore, the tunability and detuning of the terahertz laser in low-dimension nanostructure takes place due to the variation in hydrogenic impurity. These varieties of aspects would be beneficial for the evolution of tunable optoelectronic devices. ssss

Acknowledgment

Priyanka acknowledges the financial support from University Grants Commission and R. Sharma is thankful to the Delhi Technological University for the research facilities.

References

- [1] E. Kasapoglu, F. Ungan, H. Sari, I. Sökmen, Binding energies of donor impurities in modulation-doped GaAs/Al_xGa_{1-x}As double quantum wells under an electric field, *Superlattices Microstruct.* 45 (2009) 618–623. <https://doi.org/10.1016/J.SPMI.2009.02.011>.
- [2] R. Khordad, S.K. Khaneghah, Intersubband optical absorption coefficients and refractive index changes in a V-groove quantum wire, *Phys Status Solidi B Basic Res.* 248 (2011) 243–249. <https://doi.org/10.1002/pssb.201046348>.
- [3] Y. v. Pershin, J.A. Nesteroff, V. Privman, Effect of spin-orbit interaction and in-plane magnetic field on the conductance of a quasi-one-dimensional system, *Phys Rev B Condens Matter Mater Phys.* 69 (2004). <https://doi.org/10.1103/PhysRevB.69.121306>.
- [4] U. Yesilgul, F. Ungan, E. Kasapoglu, H. Sari, I. Sökmen, The linear and nonlinear intersubband optical absorption coefficients and refractive index changes in a V-shaped quantum well under the applied electric and magnetic fields, *Superlattices Microstruct.* 50 (2011) 400–410. <https://doi.org/10.1016/j.spmi.2011.08.002>.
- [5] S. Zhang, R. Liang, E. Zhang, L. Zhang, Y. Liu, Magnetosubbands of semiconductor quantum wires with Rashba and Dresselhaus spin-orbit coupling, *Phys Rev B Condens Matter Mater Phys.* 73 (2006). <https://doi.org/10.1103/PhysRevB.73.155316>.

- [6] S. Zhang, R. Liang, E. Zhang, L. Zhang, Y. Liu, Magnetosubbands of semiconductor quantum wires with Rashba and Dresselhaus spin-orbit coupling, *Phys Rev B Condens Matter Mater Phys.* 73 (2006). <https://doi.org/10.1103/PhysRevB.73.155316>.
- [7] W. Xie, The nonlinear optical rectification of a confined exciton in a quantum dot, *J Lumin.* 131 (2011) 943–946. <https://doi.org/10.1016/j.jlumin.2010.12.028>.
- [8] W. Xie, The nonlinear optical rectification coefficient of quantum dots and rings with a repulsive scattering center, *J Lumin.* 143 (2013) 27–30. <https://doi.org/10.1016/j.jlumin.2013.04.041>.
- [9] S. Lahon, M. Kumar, P.K. Jha, M. Mohan, Spin-orbit interaction effect on the linear and nonlinear properties of quantum wire in the presence of electric and magnetic fields, *J Lumin.* 144 (2013) 149–153. <https://doi.org/10.1016/j.jlumin.2013.06.054>.
- [10] R. Khordad, Optical properties of quantum wires: Rashba effect and external magnetic field, *J Lumin.* 134 (2013) 201–207. <https://doi.org/10.1016/j.jlumin.2012.08.047>.
- [11] R. Khordad, Optical properties of quantum wires: Rashba effect and external magnetic field, *J Lumin.* 134 (2013) 201–207. <https://doi.org/10.1016/j.jlumin.2012.08.047>.
- [12] R. Khordad, S. Tafaraji, Third-harmonic generation in a quantum wire with triangle cross section, *Physica E Low Dimens Syst Nanostruct.* 46 (2012) 84–88. <https://doi.org/10.1016/j.physe.2012.07.025>.
- [13] P.K. Jha, M. Kumar, S. Lahon, S. Gumber, M. Mohan, Rashba spin orbit interaction effect on nonlinear optical properties of quantum dot with magnetic field, *Superlattices Microstruct.* 65 (2014) 71–78. <https://doi.org/10.1016/j.spmi.2013.10.025>.
- [14] Y.Y. Zhang, G.R. Yao, Performance enhancement of blue light-emitting diodes with AlGaN barriers and a special designed electron-blocking layer, *J Appl Phys.* 110 (2011). <https://doi.org/10.1063/1.3651393>.
- [15] T. Sugaya, K.Y. Jang, C.K. Hahn, M. Ogura, K. Komori, A. Shinoda, K. Yonei, Enhanced peak-to-valley current ratio in InGaAs/InAlAs trench-type quantum-wire negative differential resistance field-effect transistors, *J Appl Phys.* 97 (2005). <https://doi.org/10.1063/1.1851595>.
- [16] I.Z. Utic'utic', J. Fabian, S. das Sarma, *Spintronics: Fundamentals and applications*, n.d.
- [17] S. Datta, B. Das, Electronic analog of the electro-optic modulator, *Appl Phys Lett.* 56 (1990) 665–667. <https://doi.org/10.1063/1.102730>.
- [18] M. Kumar, S. Lahon, P.K. Jha, M. Mohan, Energy dispersion and electron g-factor of quantum wire in external electric and magnetic fields with Rashba spin orbit interaction, *Superlattices Microstruct.* 57 (2013) 11–18. <https://doi.org/10.1016/j.spmi.2013.01.007>.
- [19] I.Z. Utic'utic', J. Fabian, S. das Sarma, *Spintronics: Fundamentals and applications*, n.d.
- [20] D. Ahn, S.L. Chuang, Calculation of linear and nonlinear intersubband optical absorptions in a quantum well model with an applied electric field, *IEEE J Quantum Electron.* 23 (1987) 2196–2204. <https://doi.org/10.1109/JQE.1987.1073280>.

- [21] M. Ahin, Third-order nonlinear optical properties of a one- and two-electron spherical quantum dot with and without a hydrogenic impurity, *J Appl Phys.* 106 (2009). <https://doi.org/10.1063/1.3225100>.
- [22] K.J. Kuhn, G.U. Iyengar, S. Yee, Free carrier induced changes in the absorption and refractive index for intersubband optical transitions in $\text{Al}_x\text{Ga}_{1-x}\text{As}/\text{GaAs}/\text{Al}_x\text{Ga}_{1-x}\text{As}$ quantum wells, *J Appl Phys.* 70 (1991) 5010–5017. <https://doi.org/10.1063/1.349005>.
- [23] Y. bin Yu, K.X. Guo, Exciton effects on nonlinear electro-optic effects in semi-parabolic quantum wires, *Physica E Low Dimens Syst Nanostruct.* 18 (2003) 492–497. [https://doi.org/10.1016/S1386-9477\(03\)00190-5](https://doi.org/10.1016/S1386-9477(03)00190-5).
- [24] J. Ganguly, S. Saha, S. Pal, M. Ghosh, Noise-driven optical absorption coefficients of impurity doped quantum dots, *Physica E Low Dimens Syst Nanostruct.* 75 (2016) 246–256. <https://doi.org/10.1016/j.physe.2015.09.027>.
- [25] S. Saha, S. Pal, J. Ganguly, M. Ghosh, Exploring optical refractive index change of impurity doped quantum dots driven by white noise, *Superlattices Microstruct.* 88 (2015) 620–633. <https://doi.org/10.1016/j.spmi.2015.10.021>.
- [26] I. Karabulut, S. Baskoutas, Linear and nonlinear optical absorption coefficients and refractive index changes in spherical quantum dots: Effects of impurities, electric field, size, and optical intensity, *J Appl Phys.* 103 (2008). <https://doi.org/10.1063/1.2904860>.
- [27] G. Safarpour, A. Zamani, M.A. Izadi, H. Ganjipour, Laser radiation effect on the optical properties of a spherical quantum dot confined in a cylindrical nanowire, *J Lumin.* 147 (2014) 295–303. <https://doi.org/10.1016/j.jlumin.2013.11.053>.
- [28] B. Gisi, S. Sakiroglu, E. Kasapoglu, H. Sari, I. Sokmen, Spin-orbit interaction effects on the optical properties of quantum wires under the influence of in-plane magnetic fields, *Superlattices Microstruct.* 86 (2015) 166–172. <https://doi.org/10.1016/j.spmi.2015.06.046>.
- [29] A. Bouazra, S.A. ben Nasrallah, M. Said, Theory of electronic and optical properties for different shapes of $\text{InAs}/\text{In}_{0.52}\text{Al}_{0.48}\text{As}$ quantum wires, *Physica E Low Dimens Syst Nanostruct.* 75 (2016) 272–279. <https://doi.org/10.1016/j.physe.2015.09.039>.
- [30] S. Sakiroglu, B. Gisi, Y. Karaaslan, E. Kasapoglu, H. Sari, I. Sokmen, Optical properties of double quantum wires under the combined effect of spin-orbit interaction and in-plane magnetic field, *Physica E Low Dimens Syst Nanostruct.* 81 (2016) 59–65. <https://doi.org/10.1016/j.physe.2016.02.048>.
- [31] N. Arunachalam, A. John Peter, C. Woo Lee, Pressure induced optical absorption and refractive index changes of a shallow hydrogenic impurity in a quantum wire, *Physica E Low Dimens Syst Nanostruct.* 44 (2011) 222–228. <https://doi.org/10.1016/j.physe.2011.08.019>.
- [32] C. v. Nguyen, N. Ngoc Hieu, C.A. Duque, D. Quoc Khoa, N. van Hieu, L. van Tung, H. Vinh Phuc, Linear and nonlinear magneto-optical properties of monolayer phosphorene, *J Appl Phys.* 121 (2017). <https://doi.org/10.1063/1.4974951>.
- [33] E. Kasapoglu, F. Ungan, C.A. Duque, U. Yesilgul, M.E. Mora-Ramos, H. Sari, I. Sökmen, The effects of the electric and magnetic fields on the nonlinear optical properties in the step-like asymmetric quantum well, *Physica E Low Dimens Syst Nanostruct.* 61 (2014) 107–110. <https://doi.org/10.1016/j.physe.2014.03.024>.

- [34] M.G. Barseghyan, A.A. Kirakosyan, C.A. Duque, Hydrostatic pressure, electric and magnetic field effects on shallow donor impurity states and photoionization cross section in cylindrical GaAs-Ga_{1-x}Al_xAs quantum dots, *Phys Status Solidi B Basic Res.* 246 (2009) 626–629. <https://doi.org/10.1002/pssb.200880516>.
- [35] L.E. Oliveira', L.M. Falicov, Energy spectra of donors and acceptors in quantum-well structures: Effect of spatially dependent screening, 1986.
- [36] N. Porras-Montenegro, S.T. Pérez-Merchancano, A. Latgé, Binding energies and density of impurity states in spherical GaAs-(Ga,Al)As quantum dots, *J Appl Phys.* 74 (1993) 7624–7626. <https://doi.org/10.1063/1.354943>.
- [37] A. Montes, C.A. Duque, N. Porras-Montenegro, Density of shallow-donor impurity states in rectangular cross section GaAs quantum-well wires under applied electric field, 1998. <http://iopscience.iop.org/0953-8984/10/24/012>.
- [38] M.G. Barseghyan, M.E. Mora-Ramos, C.A. Duque, Hydrostatic pressure, impurity position and electric and magnetic field effects on the binding energy and photo-ionization cross section of a hydrogenic donor impurity in an InAs Pöschl-Teller quantum ring, *European Physical Journal B.* 84 (2011) 265–271. <https://doi.org/10.1140/epjb/e2011-20650-7>.
- [39] M. Santhi, A. John Peter, C. Yoo, Hydrostatic pressure on optical absorption and refractive index changes of a shallow hydrogenic impurity in a GaAs/GaAlAs quantum wire, *Superlattices Microstruct.* 52 (2012) 234–244. <https://doi.org/10.1016/j.spmi.2012.04.020>.
- [40] J.C. Martínez-Orozco, M.E. Mora-Ramos, C.A. Duque, Electron-related optical properties in T-shaped Al_xGa_{1-x}As/GaAs quantum wires and dots, *European Physical Journal B.* 88 (2015). <https://doi.org/10.1140/epjb/e2015-60021-x>.
- [41] N. Zeiri, N. Sfina, S.A. ben Nasrallah, M. Said, Linear and non-linear optical properties in symmetric and asymmetric double quantum wells, *Optik (Stuttg).* 124 (2013) 7044–7048. <https://doi.org/10.1016/j.ijleo.2013.05.169>.
- [42] Priyanka, R. Sharma, M. Kumar, Effects of impurity factor on the physical and transport properties for Ga_{1-x}Al_xAs quantum wire in the presence of Rashba spin-orbit interaction, *Physica B Condens Matter.* 629 (2022). <https://doi.org/10.1016/j.physb.2021.413649>.
- [43] R. Khordad, B. Mirhosseini, Optical properties of GaAs/Ga_{1-x}Al_xAs ridge quantum wire: Third-harmonic generation, *Opt Commun.* 285 (2012) 1233–1237. <https://doi.org/10.1016/j.optcom.2011.11.070>.
- [44] R. Khordad, Second and third-harmonic generation of parallelogram quantum wires: Electric field, *Indian Journal of Physics.* 88 (2014) 275–281. <https://doi.org/10.1007/s12648-013-0414-1>.
- [45] G. Wang, Third-harmonic generation in cylindrical parabolic quantum wires with an applied electric field, *Phys Rev B Condens Matter Mater Phys.* 72 (2005). <https://doi.org/10.1103/PhysRevB.72.155329>.
- [46] F. Rossi, E. Molinari, Coulomb-Induced Suppression of Band-Edge Singularities in the Optical Spectra of Realistic Quantum-Wire Structures, 1996.

- [47] F. Zaouali, A. Bouazra, M. Said, A theoretical evaluation of optical properties of InAs/InP quantum wire with a dome cross-section, *Optik (Stuttg)*. 174 (2018) 513–520. <https://doi.org/10.1016/j.ijleo.2018.08.101>.
- [48] S. Antil, M. Kumar, S. Lahon, S. Dahiya, A. Ohlan, R. Punia, A.S. Maan, Influence of hydrostatic pressure and spin orbit interaction on optical properties in quantum wire, *Physica B Condens Matter*. 552 (2019) 202–208. <https://doi.org/10.1016/j.physb.2018.10.006>.
- [49] R. Khordad, Optical properties of quantum wires: Rashba effect and external magnetic field, *J Lumin*. 134 (2013) 201–207. <https://doi.org/10.1016/j.jlumin.2012.08.047>.
- [50] M.J. Karimi, M. Hosseini, Electric and magnetic field effects on the optical absorption of elliptical quantum wire, *Superlattices Microstruct*. 111 (2017) 96–102. <https://doi.org/10.1016/j.spmi.2017.06.019>.
- [51] K. Chernoutsan, V. Dneprovskii, S. Gavrilov, V. Gusev, E. Muljarov, S. Romanov, A. Syrniov, O. Shaligina, E. Zhukov, Linear and nonlinear optical properties of excitons in semiconductor-dielectric quantum wires, 2002. www.elsevier.com/locate/physe.
- [52] V. Dneprovskii, E. Zhukov, V. Karavanskii, V. Poborchii, I. Salamatina, Nonlinear optical properties of semiconductor quantum wires, 1998.
- [53] R. Khordad, Second and third-harmonic generation of parallelogram quantum wires: Electric field, *Indian Journal of Physics*. 88 (2014) 275–281. <https://doi.org/10.1007/s12648-013-0414-1>.
- [54] G. Wang, Third-harmonic generation in cylindrical parabolic quantum wires with an applied electric field, *Phys Rev B Condens Matter Mater Phys*. 72 (2005). <https://doi.org/10.1103/PhysRevB.72.155329>.
- [55] N. Zeiri, A. Bouazra, S.A. ben Nasrallah, M. Said, Linear and nonlinear susceptibilities in GaN/Al_xGa_{1-x}N quantum wire, *Phys Scr*. 95 (2020). <https://doi.org/10.1088/1402-4896/ab5b45>.

Highlights

- We investigate the optical properties such as linear and non-linear absorption coefficients, second harmonic generation, and third harmonic generation.
- We focus on the impurity influence on the optical properties in $\text{Ga}_{1-x}\text{Al}_x\text{As}$ quantum wire.
- We have used the density matrix theory approach for numeric calculation.
- We find that the absorption coefficients, refractive changes, second harmonic generation, and third harmonic generation are shifted.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

--



Impact of sustainability reporting and performance on organization legitimacy

Varsha Sehgal¹ · Naval Garg¹ · Jagvinder Singh²

Received: 4 August 2022 / Revised: 8 October 2022 / Accepted: 11 December 2022

© The Author(s) under exclusive licence to The Society for Reliability Engineering, Quality and Operations Management (SREQOM), India and The Division of Operation and Maintenance, Lulea University of Technology, Sweden 2022

Abstract There is an on-going academic debate on why firms participate in sustainability initiatives. The two arguments often tossed around are either because it supports profit maximization or because it enhances organization legitimacy. Profit maximization and organization legitimacy are not mutually exclusive as past literature states that sustainability initiatives enable an organization to gain legitimacy which enhances reputation and could translate to higher financial returns. This paper explores if the sustainability reporting type impacts the relationship between sustainability performance and organization legitimacy. We study the impact of sustainability reporting on organization legitimacy through two effects. First is the compliance mechanism based on the normative perspective, where a firm complies with the values and norms prevailing in the industry. Second, we look at the disclosure mechanism where the firm can choose what to make public i.e., it chooses what to disclose to maintain stakeholder interest and legitimacy. We extensively study the sustainability disclosure and sustainability performance of 60 Indian firms listed on NSE and who have published sustainability reports for the year 2021. We study how these firms respond to the compliance and disclosure effects over a period of a year. We analyze the impact of sustainability reporting and performance on firm's media legitimacy. Analyzing a sample of 60 Indian companies over 2021, we demonstrate that sustainability reporting positively influences the firm legitimacy. This is the compliance mechanism at work. We also find that in case of a comprehensive disclosure sustainability performance

has a positive impact on firm legitimacy. However, in case of restricted disclosure, sustainability performance has no impact on firm legitimacy. This is the disclosure mechanism at work. Firms following a comprehensive disclosure provide stakeholders with a true picture while firms following a restricted disclosure leave their stakeholders confused and stakeholders are forced to look at other constructs like age, size, reputation to proxy performance. Companies with high performance are better off following comprehensive reporting while firms with low sustainability performance are safer, giving restricted disclosures in the interim for a short period of time.

Keywords Sustainability reporting · Sustainability performance · Legitimacy

Abbreviations

WCED	World Commission on Environment and Development
SEBI	Securities and Exchange Board of India
GRI	Global reporting initiative
NSE	National stock exchange
PAT	Profit after tax
SDD	Sustainability disclosure dummy
SP	Sustainability performance
CD	Comprehensive disclosure
RD	Restricted disclosure
BRSR	Business responsibility and sustainability report

1 Introduction

In recent times, sustainability initiatives have gained a lot of importance. Sustainable development is defined as “development that meets the needs of the present without

✉ Naval Garg
naval.garg@dtu.ac.in

¹ Delhi Technological University, New Delhi, India

² University of Delhi, New Delhi, India

compromising the ability of future generations to meet their own needs” (WCED 1987, p. 43). The perception of what constitutes a firm’s role in the society has undergone a sea of change. The early economists put forward the statement that making money for its shareholders was the only motive of a firm (Friedman 1970). This was in stark contrast to the stakeholder theory which puts forward that a business should strive to maximize value for all stakeholders. The benefit should not be limited to shareholders alone. A lot of importance is given to the network of relationships in a firm’s task environment. This includes the firm’s customers, suppliers, employers, investors, community and others with a stake in the organization (Freeman & Reed 1983). Researchers have stated that it’s beneficial to the firm if it gives stakeholders attention. Therefore, a firm should be associated with actions supporting all stakeholders and not restrict or limit itself to just shareholders (Gelb and Strawser 2001).

It is worth noting that if a firm can influence stakeholders, stakeholders can also influence firms. One reason why firms participate in sustainability activities or indulge in sustainability practices are the benefits associated with the same. These activities positive impact or influence it has on stakeholders. Further, it assures the firm of stakeholder support. (Waddock and Graves 1997; Surroca et al. 2010). Positives also include an increase in legitimacy, reputation ultimately translating to financial performance (Philippe and Durand 2011; Touboul 2013). It’s not surprising that all firms would want to reap the benefits that come from high environmental or high social performance even if their environmental or social performance is not up-to the mark. The possibility of “greenwashing” always exists (Cho et al. 2018).

2 Related work

In this paper, we study two contrasting theoretical frameworks, i.e., the signaling perspective and the normative perspective. The normative perspective stresses upon the pressure to comply with the norms of the particular area (Reid and Toffel 2009; Philippe and Durand 2011). This basically treats the disclosure as means to increase legitimacy which contributes to the firm’s legitimacy and reputation and may increase financial returns also. On the other hand, the signaling perspective interprets sustainability reporting as a signal or as a means to selectively disclose or hide/manipulate some information (Golub et al. 2013). This paper contributes to past literature comparing these two perspectives to find common ground in the relationship between sustainability reporting and sustainability performance (Hummel and Schlick 2016; Touboul 2013; Golub et al. 2013). This paper then attempts to analyze the effect of sustainability reporting and sustainability performance on firm legitimacy.

Legitimacy can be termed as perception that there exist certain norms, values or procedures that are highly necessary to operate in a social environment. Failure to meet this norm can also lead to failure of an organization. In this study we explore how sustainability reporting impacts the relationship between sustainability performance and legitimacy in India. Our study is based on a future research opportunity in past literature (Touboul 2013) where the researcher analyzed the consequences of a restricted disclosure on financial performance. We attempt to analyze the consequences of restricted disclosure on legitimacy over the period of a year. As a further differentiation, we derive the sustainability disclosure score based on content analysis of sustainability reports. The disclosure index is based on GRI criteria and has been used in previous studies (Hummel and Schlick 2016). An important point to note is that we only consider those sustainability performance values which corresponded to the respective sustainability reporting points included in our index. The remaining paper has the following sections. We have a literature review section followed by the theoretical framework. This is followed by research methodology, results and finally discussion, and conclusion.

2.1 Motivation

Very few empirical studies have been conducted examining the relationship between sustainability reporting and organization legitimacy (Aerts and Cormier 2009). There has been related research highlighting how legitimacy impacts other parameters. For example—Research has demonstrated that sustainability reporting indirectly affects other parameters like analyst earnings forecasts through legitimacy (Cormier and Magnan 2015). In this paper we attempt to go one step further. We attempt to empirically analyze the relationship between sustainability performance, sustainability reporting and organization legitimacy. We further seek to study if type of reporting (comprehensive or restricted) impacts the relationship between sustainability performance and organization legitimacy.

2.2 Literature review

2.2.1 Sustainability performance

Sustainable responsibility” encompasses the “economic, legal, ethical and discretionary expectation society has of organizations” (Carroll 1979). A firm is sustainably responsible if it acts based upon stakeholder expectations. A stakeholder is any group impacted or who impacts the fulfillment of organizational goals (Freeman and Reed 1983). Any action taken to fulfill sustainability responsibility can be sustainable action. Sustainability performance is the end result of the sustainability actions or the benefits the

stakeholders receive from the actions (Barnett 2007; Waddock and Graves 1997; Zeadally et al. 2013). However, it is not necessary that sustainable responsibility will lead to sustainable actions or that sustainable actions will lead to sustainable performance. Stakeholder interest areas differ, and they expect the firm management to disclose the sustainability performance. (Herzig and Schaltegger 2006). It is important to acknowledge that sustainability performance is not easily observable (King and Toffel 2009; Touboul 2013). Therefore, stakeholders are forced to rely on signals to form opinions about the firm's performance. Sustainability disclosure can be said to be one of those signals which contribute to the sustainability performance construct. Sustainability disclosure can lead to stakeholder support or non-support, leading to increased reputation, legitimacy, and, finally, higher financial returns (Daub 2007; Herzig and Schaltegger 2006).

2.2.2 Sustainability disclosure

Sustainability reporting discloses the firm's environment and social impact due to its daily activities. To gain more legitimacy and validation, firms report their adherence to specific governed standards like (Global Reporting Initiatives) GRI Or United Nations Global Impact (UNGC) principles. For this paper, we will be using the GRI standards as a reference point for measuring firm disclosure. Sustainability disclosure is never an "all or none" scenario. Sustainability Disclosure often lies in a range. Previous research suggests that depending on normative pressure, firms may choose to release a disclosure (Gray et al. 1995) either to inform the public or try to divert attention to other areas or may try to change the relevant stakeholder opinion (Lindbloom 1994). Further, firms may strategically reveal information comprehensively or may restrict disclosure to a select number of indicators. Past researchers have suggested that firms may manipulate the information made public (Ullmann 1985); They could limit the information being revealed, avoid incriminating information, or make public private information (O'Donovan 2002). Basically, a firm may public various versions of information to ensure it appears to be legitimate to stakeholders. In this paper, we consider firms that make public a selective amount of information as following a restricted disclosure, while firms that make public extensive information about social and environmental indicators adopt a comprehensive disclosure.

2.2.3 Sustainability disclosure in India

In India, Sustainability reporting has recently gained importance. India legislated mandatory CSR spending by enacting the new Companies Act, 2013 (India, 2013). The Securities and Exchange Board of India (SEBI) in 2012 made it

mandatory for the top 100 listed companies by market capitalization to file a business responsibility report. This was later changed to be the top 500 listed companies by market capitalization in 2015. In May 2021, the SEBI introduced a new ESG reporting structure named Business Responsibility and Sustainability Report (BRSR). Under BRSR, listed entities (top 1000) need to provide an overview of the entity's material ESG risks and opportunities, an approach to mitigate or adapt to the risks, and financial parameters. Very few studies have been done on the sustainability reporting of Indian companies. The early studies mainly dealt with the CSR practices of the firms (Sen et al. 2011; Tewari and Dave 2012). Apart from these initial studies, there have been studies investigating the increasing pace of sustainability reporting in India (Laskar and Maji 2016; Jain and Winner 2016).

2.3 Legitimacy

The most commonly used definition of legitimacy broadly conceptualizes legitimacy as "the generalized perception or assumption that an entity's activities are desirable, right, or appropriate within a socially constructed system of norms, values, beliefs, and definitions" (Suchman 1995). Broadly legitimacy can be said to be the social approval of a firm's activities (Deephhouse et al. 2017). Social approval provides firms a way to win over stakeholders (Lamin and Zaheer 2012). It has been observed that many market entities will only do business with legitimate firms (Deephhouse et al. 2016). For example—Nestle India manufactured "Maggi" noodles were found to contain high levels of monosodium glutamate in May 2015. Subsequent tests found high levels of lead in June 2015 (IndiaToday 2021). This led to a widespread recall of Maggi Products, a nationwide ban in India and other countries like Nepal, Kenya, Uganda, Tanzania, Zimbabwe and South Sudan (BBC 2015). Therefore, legitimacy is a source of competitive advantage. Hence, if a firm is able to make the relevant stakeholders believe that the competition lacks legitimacy, it can win over a larger market share (Deephhouse and Carter 2005).

Legitimacy is a complex construct composed of multiple dimensions. Legitimacy is made up of the subjective perception of individuals or at the micro level (Tost 2011). It is important to note that legitimacy is clustered or aggregated together at the macro level (Berger and Luckmann 1966).

2.4 Sustainability disclosure and sustainability performance based on voluntary disclosure theory, legitimacy theory

Various scholars have studied the linkage between sustainability disclosure and sustainability performance. Early studies mostly did not find a significant relationship between sustainability disclosure and sustainability performance (Wiseman

1982; Fekrat et al. 1996; Ingram and Frazier 1980). Later studies show a mixed bag. A number of scholars cite a positive relationship based on the voluntary disclosure theory, i.e., firms with high sustainability performance have high disclosure (Al-Tuwaijri et al. 2004; Clarkson et al. 2008; Nin and Tomas 2019). Interestingly, A significant number of scholars who base their results on the legitimacy theory and cite a neutral or negative relationship, i.e., firms with poor sustainability performance disclose more (Cho and Patten 2007; Deegan 2002). Recent research (Hummel and Schlick 2016) specifies that these results are mutually not exclusive and are different facets of the same problem i.e., firms with a high sustainability performance provide a high-quality disclosure. This is in line with Voluntary Disclosure theory or economic theory. Further, firms with a poor sustainability performance provide a low-quality disclosure. This behavior is again in line with legitimacy theory.

Research also puts forward those different report formats target different stakeholder groups (De Villiers and Van Staden 2011). Firms with low sustainability performance publish more information in their annual report publications while firms facing an ongoing sticky situation with regards to the environment prefer putting information on their website. This could also be to avoid additional cost.

We have now looked at the past literature between the nature of the relationship between sustainability disclosure and sustainability performance and the attempts to explain the same using voluntary disclosure theory and legitimacy theory.

We attempt to study how various scenarios about the sustainability disclosure plan concerning two theoretical perspectives, i.e., the signaling framework and the normative framework, play out.

The normative perspective details how the firm is under pressure to comply with the norms in that field (DiMaggio and Powell 1983; Philippe and Durand 2011; Reid and Toffel 2009). A norm of transparency is highly valued. A firm that discloses precise, comprehensive information is more transparent than the one providing qualitative data. The other perspective is the signaling framework (Spence 1973), i.e., firms utilize sustainability disclosure as a signal. It may manipulate or selectively disclose certain items to ensure stakeholder support (Mahoney et al. 2013). This study attempts to show that both these effects exist. We empirically study the relationship between sustainability disclosure, sustainability performance and organization legitimacy over the period of 1 year to arrive at this conclusion.

2.5 Theoretical framework and hypotheses development

The main purpose of implementing sustainability reporting is to comply with the transparency norm. Stakeholders

attribute a lot of value to transparency (Michelon 2011). Transparency further implies the exact and comprehensive availability of requisite information. For example: A firm giving discharged wastewater's water quality parameters (high-quality quantitative data) is definitely more unambiguous than just saying that the wastewater complies with the laid-out government norms (low-quality qualitative data). Firms with a comprehensive disclosure basically signal their compliance to unambiguity and may thus lead to stakeholders attributing higher legitimacy or reputation to them. These will then lead to higher financial returns. The compliance effect (compliance with the transparency norm) as well as the compliance with stakeholder expectations is basically derived from the normative perspective, i.e., basically a firm has to comply with the norms and procedures in its field. So, if 4 out of 6 firms in a particular industry are conforming with a norm, the remaining two firms would be pressured to put comply. Previous researchers have found that adherence to sustainability reporting standards generate a positive market response (Guidry and Patten 2010). This enables us to hypothesize that stakeholders are able to estimate if the report complies with their expectations.

Therefore, we hypothesize.

H1 Organization sustainability reporting positively influences organization's legitimacy.

The disclosure mechanism implies that the firm may make public or hide the actual sustainability performance of the firm. Basically, a firm chooses what to disclose. It can manipulate its disclosure based on its actual environmental or social performance. Suppose a firm invests heavily in "green" initiatives. Then it would reap the benefits of this association only if it would publicize the same. By this logic, then, it makes sense for firms with not-so-good sustainability performance to follow a restricted disclosure. The disclosure mechanism is derived from the signaling framework. The sustainability disclosure is basically a signal from the firm. A firm decides on its disclosure based on its sustainability performance. High sustainability firms disclose more in-line with the economic theory. They signal their compliance to the unambiguity norm and also comply with the norms, values and procedures. On the other hand, low sustainability performance firms choose to put out a restricted disclosure to ensure stakeholders don't withdraw their support. They also attempt to signal compliance to the norms but a restricted disclosure ensures stakeholders don't form immediate judgements on legitimacy. The stakeholders are forced to look at other avenues like financial performance, size, age, reputation to proxy the firm's legitimacy.

Therefore, we hypothesize.

H2 Organization sustainability performance influences organization legitimacy for firms with comprehensive disclosure.

H3 Organization sustainability performance does not influence firm legitimacy for firms with restricted disclosure.

3 Research methods

3.1 Sample

We use a sample of 60 Indian firms spread across nine different industry segments in the year 2021. The 60 sampled firms are listed on the NSE/BSE and have published GRI-based externally assured sustainability reports for 2021 as well as have corresponding environmental and social performance data in the Refinitive Eikon database.

We conducted a regression analysis using firms listed in the NSE. (National Stock Exchange). The firm number is based on firms that have published GRI-based externally assured sustainability reports and have corresponding environmental and social performance data in the Refinitive Eikon database.

3.2 Variables

3.2.1 Dependent variable

Organization Legitimacy is taken as the dependent variable. It has proxied through media content analysis. This has been adopted by past literature (Bansal and Clelland 2004; Deephouse and Carter 2005; Lopez Balboa et al 2021; Yazdi et al 2017). We have basically classified whether the media article generates positive or negative influence. News was obtained through Google Search sub-tab “News”. The following sequence of steps was followed.

- (i) Navigate to Google Search.
- (ii) Enter firm name.
- (iii) Choose tools submenu and enter date range. We chose date range from 1st April 2021 to 31st March 2022. The search result headlines and linked articles were studied and classified as positive, negative or neutral. As an article may mention more than one firm, one newspaper article was utilized for computing legitimacy of more than one firm. In all 5996 news articles were studied in which 2434 articles mentioned more than one firm.

Each article was taken to be either supporting a firm or challenging the firm. The legitimacy was computed using

the Janis–Fadner coefficient of imbalance. The formula for the same is

$$\text{Legitimacy} = (p^2 - p * n) / (k * t) \rightarrow \text{if } p > n$$

$$0 \rightarrow \text{if } p = n$$

$$(p * n - n^2) / (k * t) \rightarrow \text{if } p < n$$

where p is the number of articles with a positive sentiment n is the number of articles with a negative sentiment; t is the total number of articles; k is the number of articles with some sentiment.

The final legitimacy value is between -1 and $+1$. Table 1 and Fig. 1 highlight the news sample analyzed per company.

3.2.2 Independent variables

Sustainability performance is taken as one of the independent variables. Its values are obtained from the Refinitive Eikon database. Eikon has been extensively used in previous studies (Garcia et al. 2017; Jitmaneeroj 2017). The social performance score is an amalgamation of scores in four categories (workforce, human rights, community, and product responsibility). The environment performance score is further split into three categories (resource use, emissions, and innovation). The scores range from 0 to 100%. An important point to note is that based on past literature (Hummel and Schlick 2016), we only took those environment/social performance values which corresponded to the respective disclosure points we included in our index. For example—We obtained the values of the Resource Use and Emission categories from Environment Pillar in the Refinitive Eikon database. The corresponding disclosures included Materials used, Energy Consumption, Water Consumption, Supply Chain Screening, Waste Management, and Emissions. Similarly, we obtained the value of human rights and workforce categories from the social pillar in the Refinitive Eikon. The corresponding disclosures included human rights Assessment, workforce characteristics, health and safety, training and career development. It ensured our disclosure and performance values correspond to common parameters.

Sustainability disclosure is an independent variable. The Sustainability disclosure is based on content analysis of the sustainability reports based on an index designed by previous researchers (Hummel and Schlick 2016). The disclosure index is based on the GRI sustainability reporting guidelines. The GRI guidelines give exact and precise descriptions of all information that a company must provide for each item. The “Environment” sustainability category consisted of seven disclosure indicators while the “Social” sustainability category consisted of eight disclosure indicators. All indicators in the index are core fundamental items applicable to all companies. The sustainability disclosure index measure focuses on primary disclosure items. These

Table 1 News sample analyzed per company

Sl. no	Firm	Number of news samples
1	ACC Ltd	90
2	Adani Green Energy Ltd	65
3	Adani Ports and Special Economic Zone Ltd	76
4	Aditya Birla Fashion and Retail Ltd	79
5	Ambuja Cements Ltd	95
6	Ashok Leyland Ltd	101
7	Asian Paints Ltd	105
8	Bajaj Auto Ltd	104
9	Berger Paints commodity	98
10	Bharat Heavy Electricals Ltd	70
11	Bharat Petroleum Corporation Ltd	94
12	Bharti Airtel Ltd	120
13	Coal India Ltd	116
14	Dabur India NC	97
15	Dalmia Bharat	100
16	DLF Ltd	115
17	Eicher Motors Ltd	91
18	Exide Industries Ltd	90
19	GAIL (India) Ltd	88
20	Godrej Consumer Products Ltd	92
21	Grasim Industries Ltd	84
22	Havells India Ltd	85
23	HCL Technologies Ltd	104
24	Hero MotoCorp Ltd	129
25	Hindalco Industries Ltd	93
26	Hindustan Petroleum Corp Ltd	85
27	Indian Oil Corporation Ltd	76
28	Indus Towers Ltd	60
29	Infosys Ltd	121
30	ITC Ltd	133
31	J K Cement Ltd	89
32	Jindal Steel And Power Ltd	102
33	JSW Steel Ltd	98
34	Larsen and Toubro Infotech Ltd	111
35	Larsen and Toubro Ltd	130
36	Mahindra and Mahindra Ltd	112
37	Marico Ltd	78
38	Maruti Suzuki India Ltd	123
39	National Aluminium CoLtd	92
40	NTPC Ltd	80
41	Oil and Natural Gas Corporation Ltd	135
42	Oil India Ltd	82
43	Page Industries Ltd	56
44	Ramco Cements Ltd	85
45	Reliance Industries Ltd	120
46	Shree Cement Ltd	95
47	Steel Authority of India Ltd	120
48	Tata Chemicals Ltd	95
49	Tata Communications Ltd	92

Table 1 (continued)

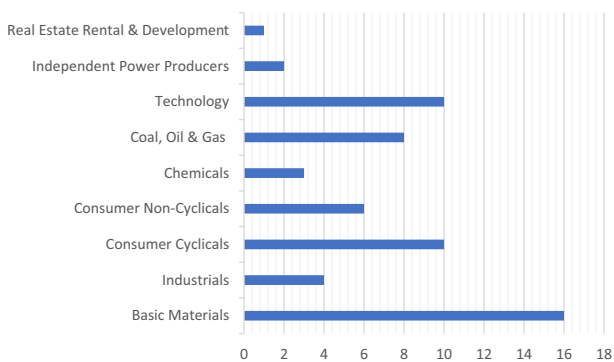
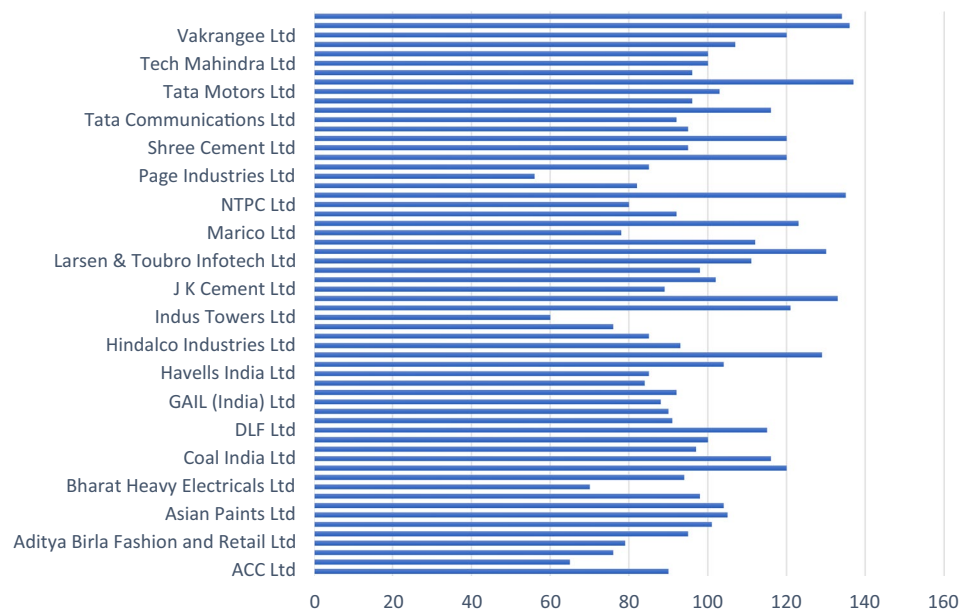
Sl. no	Firm	Number of news samples
50	Tata Consultancy Services Ltd	116
51	Tata Consumer Products Ltd	96
52	Tata Motors Ltd	103
53	Tata Power Company Ltd	137
54	Tata Steel Ltd	96
55	Tech Mahindra Ltd	100
56	UltraTech Cement Ltd	100
57	UPL Ltd	107
58	Vakrangee Ltd	120
59	Vedanta Ltd	136
60	Wipro	134

are usually the items which form the crux or the nerve center of a firm's sustainability performance. As these are highly vital parameters, we believe that a precise disclosure of these parameters is a good measure of the overall disclosure of the firm in the environment or social dimension.

To test our hypotheses, we needed to split our sample set into those with restricted disclosure and comprehensive disclosure. We calculate a separate parameter—a disclosure dummy. The dummy variable was hard coded as 1 when the firm's disclosure level was higher or equal to the median of all firm's environmental or social disclosure level for that year. Otherwise, the dummy is marked 0. For the purposes of this study, if the firm's disclosure value was greater than the median of all firms, we took it that the firm implemented a comprehensive environment or social disclosure for that year. Conversely, any firm which had the dummy disclosure variable marked as 0 was treated as following a restricted disclosure. The GRI index used for evaluating firm sustainability disclosure is given in Table 2. Figure 2 details the firms by industry grouping.

For each disclosure, 4 points were given for high-quality disclosure, 2 points for low-quality disclosure, and 0 point for non-disclosure. High quality disclosure is basically numerical data disclosure across the firm or at least above the minimum requirements specified in the GRI guidelines. The minimum requirements based on GRI are outlined in Table 2. In case these guidelines are not fulfilled or some other information is given, then two points are given for low quality data. Interestingly, even if comprehensive information is provided but the precise required data is missing, the disclosure will be marked as low quality and will be given two points. If there is no information provided at all, then the disclosure is treated absent and no point is awarded for non-disclosure.

We make certain sector-specific changes in the minimum requirements in the environmental category. This is

Fig. 1 News sample analyzed per company**Fig. 2** Firm distribution by industry. *Source:* Based on Refinitive Eikon distribution of firms and firms which have published sustainability reports for year 2021

mainly done keeping in mind that not all disclosures apply to all industry sectors. For example: Technology offices mostly obtain water supply from municipal sources. They do not rely on surface water for their water needs. Further, effluents are not generated as a by-product of their activities and hence not discharged into water sources). The minimum possible score is 0 and maximum possible score is 64.

Corporate sustainability has multiple facets (Elkington 1997). Hence having disclosure items based on environmental and social criteria strengthens the index validity. Both Sustainability Performance and Sustainability Scores are standardized based on past literature (Aiken et al. 1991; Pérez-Cornejo et al. 2020; Ramiya and Suresh 2021; Ubaid et al. 2020).

3.2.3 Control variables

Based on past literature, we control for the major factors that contribute to Organization Legitimacy such as size and age (Deephhouse and Carter 2005). We also control for financial performance using PAT. This has also been adopted by past literature (Diez Martin et al. 2021). All control variable data was obtained from the CMIE Prowess database. Descriptive statistics are provided in Table 3. Figure 1 describes the industry wise break-up in our sample.

4 Methods

To test hypothesis 1, we used the sustainability disclosure dummy as the independent variable and analyzed its impact on firm legitimacy. To test hypothesis 2 and 3 we used the sustainability disclosure dummy to split our sample between firms having a comprehensive disclosure (Sustainability Disclosure Dummy = 1) and firms with restricted disclosure (Sustainability Disclosure Dummy = 0). We test a total of three models.

Table 5 highlights the various regression models. In model 1, sustainability disclosure dummy is the independent variable and firm legitimacy is the dependent variable. In models 2 and 3, sustainability performance is considered as the independent variable and firm legitimacy is the dependent variable. The only difference between model 2 and model 3 is the reporting type. Using the dummy variable, model 2 has firms with comprehensive reporting while model 3 has firms with restricted reporting.

Model 1, 2 and 3 are simple linear regression models and were estimated using linear regression, control

Table 2 Environment & social disclosure index. *Source:* Hummel and Schlick (2016)

Code	Disclosure item	Minimum requirements	GRI linkage
<i>Environmental dimensions</i>			
E1/301	Materials used	All substantial input materials by weight or volume, Percentage of recyclable material used	EN1/301
E2/302	Energy consumption and renewables	direct and indirect energy consumption, share of renewable energy sources. (Includes energy consumption within and outside organization) ^a	EN3/4/302
E3/303a	Water withdrawal	Water Withdrawal by source ^b	EN8/303a
E4/305a	GHG emissions	GHG Scope 1, Scope 2 and Scope 3 emissions	EN16/17/305a
E5/305b	Ozone-depleting and other emissions	total emissions of ozone-depleting substances; other significant air emissions by type and weight for at least one substance; alternatively, an explicit statement of irrelevance for both ^c	EN19/20/305b
E6/306a	Water discharge	total discharge by quality (emissions to water by type and weight for at least one substance; alternatively, an explicit statement of irrelevance ^d and destination ^d)	EN22/306a
E7/306b	Waste	Total waste by type and disposal method	EN23/306b
E8/308	New supplier assessment	Percentage of new suppliers screened using environmental criteria	308-1/EN-32
<i>Social dimensions</i>			
S1	Employment	Total workforce based on at least three criteria (division, region, employment type, employment contract, qualification, age or gender)	LA1/102-7/102-8
S2	Turnover	Total number of employees leaving by any reason	LA2/401
S3	Labour management	Minimum number of weeks' notice typically provided to employees and their representatives prior to the implementation of significant operational changes	402
S4	Collective bargaining	Percentage of total workforce covered by collective bargaining agreements	LA4/407
S5	Safety and health	work safety and health based on following criteria (rates of injury, occupational diseases, lost days, absenteeism, fatalities)	LA7/403
S6	Training	total training time	LA10/404
S7	Discrimination	total number of incidents or explicit statement that no incidents occurred	HR4/406
S8	Child, forced and compulsory labour	scope and numerical results of audits (within company or supply chain) regarding at least one aspect	HR6/7/408/409

^aFor industry groups 8: share of renewable energy produced

^bFor industry groups 7: by source is excluded

^cFor industry groups 7: ozone-depleting substances or other significant air emissions

^dFor industry groups 7: by quality and destination is excluded

variables, dummy variables and robust estimations to counter heteroscedasticity.

5 Results

Table 4 defines the correlation between the different models of study. Table 5 highlights the various regression models. Model 1 demonstrates that the sustainability disclosure positively and significantly impacts legitimacy (Coefficient 0.059 is significant at the 0.1% level). This supports hypothesis 1 that sustainability disclosure positively influences legitimacy. This is the compliance effect in action. We can deduce

that the compliance effect is active in the 1 year time period. Further, Model 2 shows that in the case of comprehensive disclosure, sustainability performance has a positive impact on legitimacy (0.035, which is significant at a 0.1% level). This supports hypothesis 2, i.e., sustainability performance positively influences the legitimacy of firms with comprehensive disclosure or firms with comprehensive disclosure make public their actual sustainability performance level and the stakeholders are able to gauge their worth. This positively influences firm legitimacy (The coefficient is positive and significant). On the other hand, in the case of Model 3, the impact is insignificant. This supports hypothesis 3, i.e., sustainability performance has no impact on the legitimacy

Table 3 Firm distribution by industry. *Source:* Based on Refinitive Eikon distribution

SI	Industry group	Number of firms
1	Basic Materials	16
2	Industrials	4
3	Consumer Cyclical	10
4	Consumer Non-cyclical	6
5	Chemicals	3
6	Coal, Oil and Gas	8
7	Technology	10
8	Independent Power Producers	2
9	Real Estate Rental and Development	1

takeaway here is that in case of restricted reporting if the firm is not able to take benefit of its sustainability initiatives, it is not penalized also for lack of the same. In short, restrictive reporting at least in the considered time horizon gives the firm some slack to get its processes in order.

6 Discussion and conclusion

This paper attempts to study the relationship between sustainability disclosure, sustainability performance, and firm legitimacy, specifically in the Indian Context. We try to basically lend credence to the statement that firms have the option to “manipulate” their environmental or social disclosure. We actually compare two theoretical approaches:

Table 4 Pearson correlations (DV-organization legitimacy)

S. no	Variable	Mean	S. D	1	2	3	4	5	6
1	Legitimacy	0.5121	0.133	1					
2	SDD	0.53	0.50	0.3173**	1				
3	SP	0	1	0.3849**	0.4253*	1			
4	Age	50.27	24.48	0.0829	0.2441*	0.2675*	1		
5	Size	12.89	1.26	0.1437	0.166*	0.3582*	0.1675*	1	
6	ROA	7.86	8.44	0.2283 ⁺	−0.1381***	0.0020	−0.1083***	−0.3039*	1

SDD sustainability disclosure dummy, *SP* sustainability performance

* $p < 0.001$, ** $p < 0.01$, *** $p < 0.05$, ⁺ $p < 0.10$

of firms with restrictive reporting. This demonstrates that the restricted reporting prevents the stakeholders from forming opinions about the firm. The stakeholders are unable to take a call on the same. Although legitimacy does not see any benefit from their sustainability performance (insignificant coefficient), it is not penalized for the same also. Hypothesis 2 and 3 basically show the disclosure effect in action. A key

(a) The pressure to comply (The normative perspective);
(b) The Signaling theory through which a firm tries to hint at or hide certain underlying characteristics (The signaling perspective).

Firms with high sustainability performance issue a comprehensive disclosure and are assured of stakeholder

Table 5 Sustainability performance and sustainability disclosure impact on organization legitimacy (media legitimacy as a proxy for firm legitimacy). *Source:* Refinitive Eikon distribution

Variables	Model 1	Model 2	Model 3
DV	Organization legitimacy		
Screening criteria	None	CD	RD
SDD	0.059* (0.00)	NA	NA
SP	NA	0.035* (0.00)	0.013 (0.24)
Age	0.003*** (0.05)	0.0652*** (0.031)	0.001 (0.339)
Size	0.0429* (0.00)	0.044** (0.01)	0.019 ⁺ (0.299)
ROA	0.003*** (0.02)	0.087* (0.00)	0.0043 ⁺ (0.103)
Ind. dummy	Yes	Yes	Yes
Constant	0.072 (0.00)	0.065 (0.22)	0.111*** (0.04)
R-square	0.40	0.44	0.41
F-statistics	3.14** (0.00)	2.63** (0.00)	5.12** (0.00)

SDD sustainability disclosure dummy, *SP* sustainability performance, *CD* comprehensive disclosure, *RD* restricted disclosure

* $p < 0.001$, ** $p < 0.01$, *** $p < 0.05$, ⁺ $p < 0.10$

support. This increases their legitimacy, reputation and can translate to higher financial returns over time. Firms with a low level of sustainability performance are better off issuing a restricted disclosure. This confuses the stakeholders and they are not able to form an opinion on the firm's sustainability performance. They are forced to rely on other constructs like reputation or financial performance for taking a judgement call.

Sustainability disclosure positively impacts organization legitimacy. (Hypothesis 1).

Sustainability performance impacts the organization legitimacy only if the disclosure is comprehensive. (Hypothesis 2). Sustainability performance has no impact on firm legitimacy if the firm puts out a restricted disclosure. (Hypothesis 3) This also implies that companies which have high environmental or social performance are more likely to put out information in the public domain while firms with weak performance are safer, giving restricted disclosures in the interim for a short period of time. It further implies that in case of a restricted disclosure, if the firm does not benefit from the high sustainability performance, then it is not penalized for low sustainability performance either.

The basic premise here is that firms manipulate sustainability reporting. The reporting is comprehensive when firms perform superlatively. The reporting is restrictive when the firm performs below the stakeholder expectations. By this logic, firms wait for a certain period of time after adopting sustainability measures before putting them out comprehensively in the public domain. While analyzing past literature for this paper we have seen a number of examples where when a firm starts publishing a sustainability report, it is not externally assured. Gradually, as it gains exposure to various sustainability standards and its sustainability performance improves, it starts publishing extensive, externally assured sustainability reports. An interesting point to note is that with globalization and social media shrinking the world, it has become very difficult for firms to hide their true sustainability performance of firms. Therefore, it is best if managers include sustainability planning in the organization at the grass-root level.

6.1 Limitations/future scope

One major limitation is the lack of externally assured standards-based sustainability reports. Very few Indian firms publish externally assured sustainability reports. It was one of the reasons why we could not increase the number of firms. This study can be extended to include other countries or over an extended time duration to gain more understanding of how the normative and signaling frameworks function in different regulatory conditions or over a larger time period.

Funding No funds, grants or other support was received.

Declarations

Conflict of interest No potential conflict of interest. No funds, grants or other support was received.

Ethical approval The present research follows ethical guidelines as prescribed by Helisinki Declaration.

Informed consent Informed consent of participants was taken.

References

- Aerts W, Cormier D (2009) Media legitimacy and corporate environmental communication. *Account Organ Soc* 34(1):1–27
- Aiken LS, West SG, Reno RR (1991) Multiple regression: testing and interpreting interactions. Sage, London
- Al-Tuwaijri SA, Christensen TE, Hughes II KE (2004) The relations among environmental disclosure, environmental performance, and economic performance: a simultaneous equations approach. *Account Organ Soc* 29(5–6):447–471
- Bansal P, Clelland I (2004) Talking trash: legitimacy, impression management, and unsystematic risk in the context of the natural environment. *Acad Manag J* 47(1):93–103
- Barnett ML (2007) Stakeholder influence capacity and the variability of financial returns to corporate social responsibility. *Acad Manag Rev* 32(3):794–816
- BBC (2015) <https://www.bbc.com/news/world-africa-33053683>
- Berger PL, Luckmann T (1966) The social construction of reality: a treatise in the social construction of reality. Anchor Books, Garden City
- Carroll AB (1979) A three-dimensional conceptual model of corporate performance. *Acad Manag Rev* 4(4):497–505
- Cho CH, Patten DM (2007) The role of environmental disclosures as tools of legitimacy: a research note. *Account Organ Soc* 32(7–8):639–647
- Cho CH, Laine M, Roberts RW, Rodrigue M (2018) The frontstage and backstage of corporate sustainability reporting: evidence from the Arctic National Wildlife Refuge Bill. *J Bus Ethics* 152(3):865–886
- Clarkson PM, Li Y, Richardson GD, Vasvari FP (2008) Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Account Organ Soc* 33(4–5):303–327
- Cormier D, Magnan M (2015) The economic relevance of environmental disclosure and its impact on corporate legitimacy: an empirical investigation. *Bus Strategy Environ* 24(6):431–450
- Daub CH (2007) Assessing the quality of sustainability reporting: an alternative methodological approach. *J Clean Prod* 15(1):75–85
- De Villiers C, Van Staden CJ (2011) Where firms choose to disclose voluntary environmental information. *J Account Public Policy* 30(6):504–525
- Deegan C (2002) Introduction: the legitimising effect of social and environmental disclosures—a theoretical foundation. *Account Audit Account J* 15:282–311
- Deephhouse DL, Carter SM (2005) An examination of differences between organizational legitimacy and organizational reputation. *J Manag Stud* 42(2):329–360
- Deephhouse DL, Newbury W, Soleimani A (2016) The effects of institutional development and national culture on cross-national differences in corporate reputation. *J World Bus* 51(3):463–473

- Deephhouse DL, Bundy J, Tost LP, Suchman MC (2017) Organizational legitimacy: six key questions. *SAGE Handb Organ Inst* 4(2):27–54
- Díez-Martín F, Blanco-González A, Díez-de-Castro E (2021) Measuring a scientifically multifaceted concept. The jungle of organizational legitimacy. *Eur Res Manag Bus Econ* 27(1):100131
- DiMaggio PJ, Powell WW (1983) The iron cage revisited: institutional isomorphism and collective rationality in organizational fields. *Am Sociol Rev* 48:147–160
- Elkington J (1997) The triple bottom line. *Environ Manag Read Cases* 2:49–66
- Fekrat MA, Inclan C, Petroni D (1996) Corporate environmental disclosures: competitive disclosure hypothesis using 1991 annual report data. *Int J Account* 31(2):175–195
- Freeman RE, Reed DL (1983) Stockholders and stakeholders: a new perspective on corporate governance. *Calif Manag Rev* 25(3):88–106
- Friedman M (1970) A theoretical framework for monetary analysis. *J Polit Econ* 78(2):193–238
- Garcia AS, Mendes-Da-Silva W, Orsato RJ (2017) Sensitive industries produce better ESG performance: evidence from emerging markets. *J Clean Prod* 150:135–147
- Gelb DS, Strawser JA (2001) Corporate social responsibility and financial disclosures: an alternative explanation for increased disclosure. *J Bus Ethics* 33(1):1–13
- Golub A, Mahoney M, Harlow J (2013) Sustainability and intergenerational equity: Do past injustices matter? *Sustain Sci* 8(2):269–277
- Gray R, Kouhy R, Lavers S (1995) Constructing a research database of social and environmental reporting by UK companies. *Account Audit Account J* 8:78–101
- Guidry RP, Patten DM (2010) Market reactions to the first-time issuance of corporate sustainability reports: Evidence that quality matters. *Sust Account Manag Policy J* 1(1):33–50. <https://doi.org/10.1108/20408021011059214>
- Herzig C, Schaltegger S (2006) Corporate sustainability reporting. An overview. *Sust Account Report*, 301–324
- Hummel K, Schlick C (2016) The relationship between sustainability performance and sustainability disclosure—reconciling voluntary disclosure theory and legitimacy theory. *J Account Public Policy* 35(5):455–476
- Indiatoday (2021) <https://www.indiatoday.in/business/story/nestle-unhealthy-food-controversy-looking-back-at-the-maggi-noodle-crisis-in-india-1810003-2021-06-02>
- Ingram RW, Frazier KB (1980) Environmental performance and corporate disclosure. *J Account Res* 18:614–622
- Jain R, Winner LH (2016) CSR and sustainability reporting practices of top companies in India. *Corp Commun Int J* 21(1):36–55
- Jitmaneeroj B (2017) Does investor sentiment affect price-earnings ratios? *Stud Econ Finance* 34:183–193
- King AA, Toffel MW (2009) Self-regulatory institutions for solving environmental problems: perspectives and contributions from the management literature. *Governance for the environment: New perspectives*, pp 98–115
- Lamin A, Zaheer S (2012) Wall street vs. main street: firm strategies for defending legitimacy and their impact on different stakeholders. *Organ Sci* 23(1):47–66
- Laskar N, Maji SG (2016) Corporate sustainability reporting practices in India: Myth or reality? *Soc Responsib J* 12:625–641
- Lindbloom G (1994) Learning about organizational cultures and professional competence. In: *Ethical and social issues in professional education*, p 219
- López-Balboa A, Blanco-González A, Díez-Martín F, Prado-Román C (2021) Macro level measuring of organization legitimacy: its implication for open innovation. *J Open Innov Technol Mark Complex* 7(1):53
- Mahoney LS, Thorne L, Cecil L, LaGore W (2013) A research note on standalone corporate social responsibility reports: signaling or greenwashing? *Critical Perspect Account* 24(4–5):350–359. <https://doi.org/10.1016/j.cpa.2012.09.008>
- Michelon G (2011) Sustainability disclosure and reputation: a comparative study. *Corp Reput Rev* 14(2):79–96
- Nin J, Tomás E (2019) Default propagation in customer-supplier networks. *J Ambient Intell Humaniz Comput*. <https://doi.org/10.1007/s12652-019-01370-7>
- O'donovan G (2002) Environmental disclosures in the annual report: extending the applicability and predictive power of legitimacy theory. *Account Audit Account J* 15(3):344–371
- Pérez-Cornejo C, de Quevedo-Puente E, Delgado-García JB (2020) Reporting as a booster of the corporate social performance effect on corporate reputation. *Corp Soc Responsib Environ Manag* 27(3):1252–1263
- Philippe D, Durand R (2011) The impact of norm-conforming behaviors on firm reputation. *Strateg Manag J* 32(9):969–993
- Ramiya S, Suresh M (2021) Factors influencing lean-sustainable maintenance using TISM approach. *Int J Syst Assur Eng Manag* 12(6):1117–1131
- Reid EM, Toffel MW (2009) Responding to public and private politics: corporate disclosure of climate change strategies. *Strateg Manag J* 30(11):1157–1178
- Sen M, Mukherjee K, Pattanayak JK (2011) Corporate environmental disclosure practices in India. *J Appl Account Res* 12:139–156
- Spence M (1973) Job market signaling. *Q J Econ* 87(3):355–374
- Suchman MC (1995) Managing legitimacy: strategic and institutional approaches. *Acad Manag Rev* 20(3):571–610
- Surroca J, Tribó JA, Waddock S (2010) Corporate responsibility and financial performance: the role of intangible resources. *Strateg Manag J* 31(5):463–490
- Tewari R, Dave D (2012) Corporate social responsibility: communication through sustainability reports by Indian and multinational companies. *Glob Bus Rev* 13(3):393–405
- Tost LP (2011) An integrative model of legitimacy judgments. *Acad Manag Rev* 36(4):686–710
- Touboul S (2013) The strategic value of sustainability and its disclosure: three essays on the impact of sustainability performance, disclosure and reputation on firms' financial performance. Doctoral dissertation, Jouy-en-Josas, HEC
- Ubaid AM, Dweiri FT, Ojiako U (2020) Organizational excellence methodologies (OEMs): a systematic literature review. *Int J Syst Assur Eng Manag* 11(6):1395–1432
- Ullmann AA (1985) Data in search of a theory: a critical examination of the relationships among social performance, social disclosure, and economic performance of US firms. *Acad Manag Rev* 10(3):540–557
- Waddock SA, Graves SB (1997) The corporate social performance–financial performance link. *Strateg Manag J* 18(4):303–319
- WCED SWS (1987) World commission on environment and development. *Our Common Future* 17(1):1–91
- Wiseman J (1982) An evaluation of environmental disclosures made in corporate annual reports. *Account Organ Soc* 7(1):53–63
- Yazdi M, Nikfar F, Nasrabadi M (2017) Failure probability analysis by employing fuzzy fault tree analysis. *Int J Syst Assur Eng Manag* 8(2):1177–1193
- Zeadally S, Pathan ASK, Alcaraz C, Badra M (2013) Towards privacy protection in smart grid. *Wirel Pers Commun* 73(1):23–50

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Investigation of combustion and emission characteristics of an SI engine operated with compressed biomethane gas, and alcohols

Pradeep Kumar Meena¹ · Amit Pal¹ · Samsheer Gautam²

Received: 8 September 2022 / Accepted: 7 December 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

Alternative fuels in spark-ignition engines significantly reduce engine exhaust emissions and improve fuel efficiency. This research investigates the performance of a multicylinder SI engine using 10%, 20% (ethanol, methanol, methyl acetate), and 100% compressed biomethane gas (CBG) as alternative fuels. Engine performance parameters (BTE, ITE, ME, BP), BSFC, ISFC, FF, combustion phenomenon (cylinder pressure, crank angle, cylinder volume, mass fraction burned, net heat release, mean gas temperature, cumulative heat release, rate of pressure rise), and emission characteristics (HC, CO, CO₂, NO_x) are measured. CBG achieved a maximum BTE of 23.33% compared to all other fuels. Minimum fuel consumption rate of 1.72 kg/h at maximum rpm achieved BSFC value of 0.44 kg/kWh and ISFC value of 0.261 kg/kWh. The highest cylinder pressure of 6.79 bar was achieved in the G90M10 with a cylinder volume of 48.58 cc. NHR of 3.08 J/deg was found in the G80M20 at a crank angle of 376°, and the maximum MGT was 390.20 °C in the G80E20. The highest CHR values of 0.12 kJ at crank angles of 432°, 420°, 422°, and 427° were achieved in the G100, CBG, G80E20, and G90E10. G90M10 reached a maximum value of 0.14 bar/degree of rate of pressure rise at a crank angle of 374°. Average minimum emission gas was found in CBG at a minimum and maximum RPM, indicating that CBG gives the best emission result with engine performance compared to all alternative fuels.

Keywords Alcohols · Engine performance · CBG, Emission · Biogas

Abbreviations

BTE	Brake thermal efficiency
ITE	Indicated thermal efficiency
ME	Mechanical efficiency
BG	Biogas
CBG	Compressed biomethane gas
MGT	Mean gas temperature
MFB	Mass fraction burned
BSFC	Brake specific fuel consumption
ISFC	Indicated specific fuel consumption
FF	Fluid flow
CR	Compression ratio

SOB	Start of burning
EOB	End of burning
NHR	Net heat release
CHR	Cumulative heat release
RPR	Rate of pressure rise
TDC	Top dead center
G100	Pure gasoline fuel
G90E10	90% Gasoline 10% ethanol
G80E20	80% Gasoline 20% ethanol
G90M10	90% Gasoline 10% methanol
G80M20	80% Gasoline 20% methanol
G90MA10	90% Gasoline 10% methyl acetate
G80MA20	80% Gasoline 20% methyl acetate

Responsible Editor: Philippe Garrigues

✉ Pradeep Kumar Meena
pradeep_2k18phdme08@dtu.ac.in

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

² Harcourt Butler Technical University, Kanpur, India

Introduction

The global economy has weakened following COVID-19. To compensate, gasoline and diesel prices steadily rise in practically all emerging countries. Due to the energy crisis, global warming, high fossil fuel costs, and rigorous emission rules,

renewable oxygenated fuels have received greater attention in recent decades (Awad et al. 2018a), (Gülüm and Bilgin 2018). So, the globe is transitioning to a sustainable energy period, focusing on energy efficiency and renewable energy sources (Chauhan et al. 2010). In reality, fossil fuels remain the primary source of global energy, with global energy consumption expected to climb by around 33% by 2050 (Hosseini and Wahid 2013), (Saidur et al. 2011). In recent years, the hunt for alternative fuels that offer a harmonious relationship with sustainable development, energy-saving, efficiency, and environmental protection has intensified. Bio-fuels have the potential to provide a viable solution to the global petroleum dilemma. Automobiles that run on gasoline or diesel also substantially contribute to greenhouse gas emissions. Furthermore, the increasing number of circulating diesel and petrol cars accounts for roughly 20% of global greenhouse gas (GHG) emissions (Iodice et al. 2016), (Rajesh Kumar and Saravanan 2016). Energy policy, planning, and associated issues have become a significant public agenda item in most industrialized and developing countries in recent years. As a result, governments support using alternative fuels in automobile engines. Several alternative fuels, such as gasoline and diesel with natural gas (CNG/CBG), ethanol, methanol, methyl acetate, butanol, and hexanol, have been judged acceptable and cost-effective alternatives for conventional fuels based on these criteria. Because of their excellent physicochemical qualities, ethanol, methanol, butanol, methyl acetate, and other alcohols are essential renewable fuels when blended with pure gasoline among the renewable energies available for spark-ignition (SI) engines (Awad et al. 2018b).

“Some of the experimental studies are as follows: Four-stroke spark-ignition engine, the effects of ethyl alcohol blended fuel with different blending ratios (10, 20, and 30% by volume) on engine performance and exhaust emissions were explored, and the results showed that combining ethanol with gasoline improves BTE and BSFC and lowers exhaust gas temperature, as well as lower CO and HC exhaust emissions, while NO_x emissions are higher” (Vivek Pandey and Gupta 2016). As ethanol blends were utilized in lower quantities, engine torque increased by 2.31–4.16%, and BP increased by 0.29–4.77%, while BSFC increased when the ethanol percentage grew from 5.17 to 56.0% (Thakur et al. 2017). “Methanol (M5, M7.5, M10, M12.5, M15) was tested for the performance and combustion characteristics of a four-cylinder, four-stroke, spark-ignition engine (SI). According to the experiments’ results, adding methanol enhanced the engine’s performance. It was also discovered that increasing methanol concentration lowered CO and HC emissions while increasing CO₂ and NO_x emissions” (Shayan et al. 2011). In terms of methanol mixtures (0–15%), there has been a rise in gasoline octane rating, an increase in BTE and ITE,

and a drop in knocking (Mallikarjun and Mamilla 2009). “When the compression ratio for the methanol/gasoline blend was increased from CR8 to CR10, the peak pressure and NHR value increased by 27.5% and 30%, respectively, at a speed of 1600 rpm. At a compression ratio of 10:1, the performance results demonstrate a good agreement of improvisation with a 25% rise in BTE and a 19% reduction in BSFC. CO and HC emissions were reduced by 30–40% at a more excellent compression ratio of 10:1, and the same trend was detected at all speeds; however, NO_x emissions rose with increasing CR” (Nuthan Prasad et al. 2020), (Jhalani et al. 2021). “At varied loads of 104, 207, 311, and 414 kPa, methyl acetate is used in a single-cylinder spark-ignition engine, which is fueled with base gasoline, M5 (95% base gasoline + 5% methyl acetate), and M10 (90% base gasoline + 10% methyl acetate). According to these findings, adding methyl acetate to base gasoline boosts BSFC while lowering the engine’s BTE. Additionally, it was discovered that while methyl acetate does not significantly influence HC emissions, it did reduce CO and increase CO₂. Adding methyl acetate to the NO_x data showed a significant increase in NO_x emissions” (Cakmak et al. 2018).

Biogas (BG), also known as an alternative or renewable fuel, has been recommended to solve the problem since it has numerous advantages over natural gas, often utilized as a car fuel. BG is mainly a mixture of CH₄ and CO₂ and other gases formed in anaerobic conditions. Both agricultural and industrial wastes can be used to make BG (Holm-Nielsen et al. 2009), (Pradeep Kumar Meena and Sumit Sharma 2022). Removing CO₂ and H₂S from raw biogas and compressing pure biogas at high pressure can be used in the automobile sector and power generation (Larsson et al. 2016). Because CBG possesses qualities similar to CNG, biogas has a great potential to replace natural gas (Subramanian et al. 2013). Furthermore, biomethane might be compressed into a fuel tank as CBG for transportation fuel in a CNG vehicle, which is easy to store and reduces transportation expenses (D. Deublein 2008). “CBG was utilized in a multicylinder engine compared to CNG at 50% maximum load and engine speed (1500–3500 rpm). Results suggest that the engine run with CBG has higher thermal efficiency and reduced NO_x and HC emissions. As a result, CBG fuel can replace CNG in spark-ignition engines as an alternate fuel” (Limpachoti and Theinnoi 2021).

Many researchers have worked on alcohol fuels, but very little research has shown the effects of ethanol, methanol, and methyl acetate alcohols on engine performance and emissions. This research has used three types of alcohol fuels: ethanol, methanol, and methyl acetate mixed with 10% and 20% gasoline fuels and 100% CBG. So, here is a comparison of the performance and emission parameters of the multicylinder SI engine using four alternate fuels.

Material and method

Fuels properties

In this experimental process, three types of alcohol blends of ethanol, methanol, and methyl acetate have been used with gasoline fuel, which is pure fuel up to 98–99%. It is a volatile, colorless liquid with a distinct aroma and flavor resembling alcohol. And by making BG from solid organic waste (fruit, vegetable wastes) and removing CO₂ and H₂S composition, pure biogas is compressed at 200 bar pressure and filled in a high-pressure bar cylinder. Different fuel properties of these fuels are given in Table 1.

Experimental setup

Experimental setup used is a Maruti Wagon R with a maximum power of 47.70 kW @ 6200 rpm. It is a four-cylinder, four-stroke, variable speed, water-cooled, and petrol engine, whose details are given in Table 2. Various alcohols such as ethanol, methanol, and methyl acetate have been tested by mixing 10% and 20% (G90E10, G80E20, G90M10, G80M20, G90MA10, G80MA80) with gasoline. Gasoline and alcohol blend readings were taken in a burette tube at an interval of 60 s. Experiment data has been taken by setting the load from the dynamometer to 4 kg and varying the speed from 2000 to 4500 rpm. And pure CBG has also been studied on the same parameters. Using CO₂ and H₂S scrubbers to purify the raw biogas, pure biogas, i.e., up to 96.6% CH₄, is obtained, whose composition is checked with a biogas analyzer.

For use, the CBG is fed into a high-pressure cylinder using a compressor. For safety features, a gas stop valve, pressure gauge, gas conversion kit, and gas filter have also

Table 2 Details of experimental setup

Engine specification	Details
Stroke length	72.00 (mm)
Cylinder bore	68.50 (mm)
Connecting rod length	112.50 (mm)
Compression ratio	9.2:1
Swept volume	265.34 (cc)
Engine type	Maruti Wagon R 4 strokes 4 cylinders
No. of cylinders	4
Maximum power output at 6200 rpm	47.70 kW
Cooling system	Water cooling close system
Orifice diameter	40 mm
Dynamometer arm length	210 mm
Fuel pipe diameter	33.90 mm
Number of cycles	10

been installed, which are shown in Fig. 1b. During the experiment, water is supplied from cooling waters used to cool the engine setup, whose flow is adjusted by rotameters. Compression studies of gasoline, alcohol blends, and CBG fuels have been performed. The resulting combustion parameters include cylinder pressure, rate of pressure rise, mass fraction burned, pressure volume, net heat release, mean gas temperature, and cumulative heat release. Thermal efficiencies, BSFC, ISFC, etc., have been studied in performance parameters. All experimental data was saved from the NI unit to the computer with the help of IC Engine software, shown in Fig. 1a. Apparatus used for studying these fuel blends and the different properties of CBG are given in Table 1. CO₂, CO, HC, and NO_x gases from the AVL emission apparatus were also checked (Table 3).

Table 1 Fuel properties (ethanol, methanol, methyl acetate, and CBG)

Fuel properties	Unit	G100%	E10%	E20%	M10%	M20%	MA10%	MA20%	CBG
Chemical formula	-	C ₅ -C ₁₂	C ₂ H ₅ OH	C ₂ H ₅ OH	CH ₃ OH	CH ₃ OH	C ₃ H ₆ O ₂	C ₃ H ₆ O ₂	CH ₄
Density at 40 °C	kg/m ³	721	734	735	723	736	737	757	0.90
Lower heating value	MJ/kg	44	42.38	40.76	41.59	39.18	41.75	39.5	48.5
RON	-	94.5	96.3	98.5	97.1	98.8	99.5	107.5	127
MON	-	84.3	84.5	86.2	84.2	86.1	97.3	104.8	119
Stoichiometric air/fuel ratio	-	14.8	14.3	13.5	14.1	13.3	13.9	13.1	17.2
Reid vapor pressure at 38 °C	-	55.7	56.3	57.1	94.5	95.8	54.2	52.7	-
Flash point °C	-	26.3	30.5	29.8	28.2	27.5	21.5	18.1	-
Fire point °C	-	25.1	28.9	29.7	29.9	31.5	28.3	31.48	-
Methane (CH ₄)	%	-	-	-	-	-	-	-	96.6
Hydrogen sulfide (H ₂ S)	%	-	-	-	-	-	-	-	0.0 ppm
O ₂	%	-	-	-	-	-	-	-	0.4
CO ₂	%	-	-	-	-	-	-	-	3.0

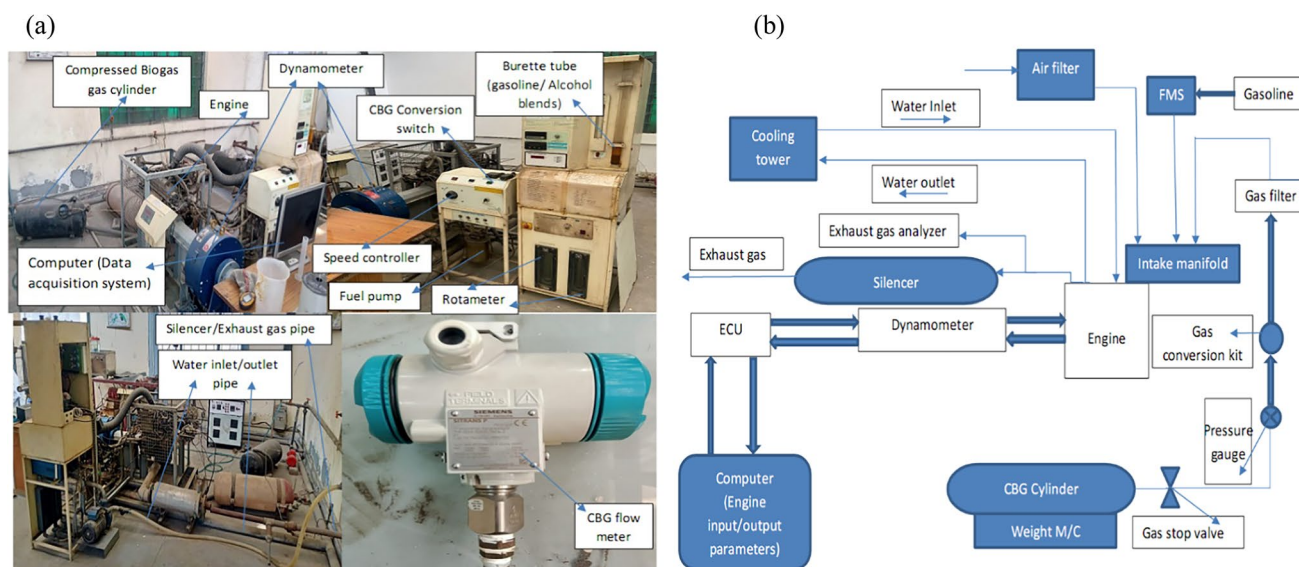


Fig. 1 **a** Experimental setup with parameters measuring instruments. **b** Schematic diagram of experimental setup

Table 3 Apparatus used during experiment

Apparatus	Name of the company
Biogas analyzer	OX-300B, Nunes Instruments
Biogas compressor	Italy tech
Viscometer	Anton Paar
Junkers calorimeter	H. L. Scientific Industries
Emission gas analyzer	AVL

Results and discussion

Engine parameters have been studied at 4 kg constant load and different speeds. Given below are various parameters such as engine performance (BP, BTE, ITE, ME), BSFC, ISFC, FF, and combustion phenomena (cylinder pressure, crank angle, crank angle, cylinder volume, mass fraction burned, NHR, mean gas temperature, cumulative heat release, rate of pressure rise) and emission parameters (HC, CO, CO₂, NO_x) have been studied.

Engine performance

Figure 2a shows that at a constant load of 4 kg, the engine speed was 2000 rpm, and the brake power value was 1.73 kW; at that time, the highest brake thermal efficiency of 23.33% CBG was achieved. Compared to gasoline, CBG has a higher octane rating and more excellent knock resistance. CBG burns more efficiently than gasoline or diesel, and very little of it remains unburned. As a result, engines designed explicitly for CBG have more excellent compression ratios and hence higher stated efficiency. BTE of CBG

from a minimum rpm of 2000 to a maximum rpm of 4500 was superior than the other fuels. G100 fuel had a BTE value of 21.76% at 2000 rpm, and the highest BTE value of 17.28% in the alcohol fuel was obtained in the G80E20, and the lowest value was 13.25% in the G90M10. At a maximum of 4500 rpm, the BTE value of CBG was 16.76%, 15.63% for G100, and 14.69% for G90M20. Alternative fuels G90M20, G80MA20, and CBG have BTE values higher than the G100 at 2500 and 3000 rpm, meaning all these alternative fuels have the potential to replace gasoline fuels. Similarly, in a study, BTE values of G90E10 and G80E20 and G70E30 blends in a four-stroke engine at 2000 to 3000 rpm were found to be 16.2%, 18.9%, and 21.2% (Vivek Pandey and Gupta 2016). The BTE value of methanol blend G88M12% is achieved at 18.5% at 2000 rpm, 21.5% at 2500 rpm, and 23.5% at 3000 rpm (Mohammed Kamil and Ibrahim Thamer Nazzal 2016). G90MA10 blend at constant 1500 rpm has achieved BTE values ranging from 10 to 28% at effective pressure (104 to 414 kPa) (Cakmak et al. 2018).

Figure 2b shows that at 2000 rpm, the maximum value of ITE was achieved at 48.52% in G100, 38.65% in alcoholic fuel (G80E20), and 31.09% in CBG. CBG has a higher calorific value than gasoline and alcohol fuel, and the fuel flow rate is also higher at minimum rpm and constant load. Hence, the value of ITE at low speed was lower in CBG. At a maximum of 4500 rpm, the highest ITE value was obtained in CBG at 28.35%, G100 gained 17.8%, and the ITE value in the alcohol fuel (G90M20) was reached at 16.64%. As the rpm increases from 2000 to 4500, the value of ITE decreases in G100 and other alcohol blends, but the value of ITE in CBG has increased compared to other fuels. At higher speeds, CBG consumes

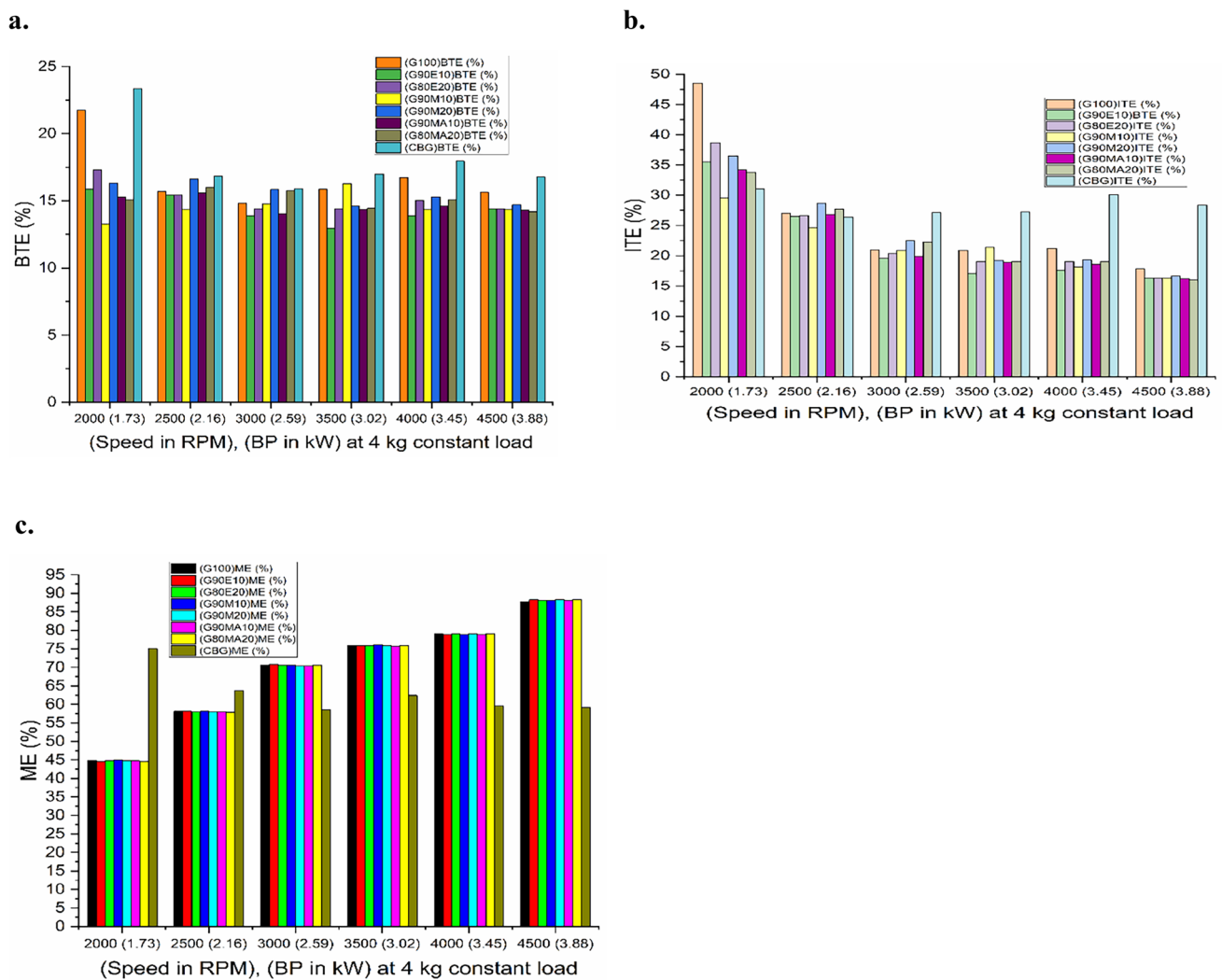


Fig. 2 a Brake thermal efficiency, brake power varies w.r.t Speed. b Indicated thermal efficiency, brake power varies w.r.t speed. c Mechanical efficiency, brake power varies w.r.t speed

less fuel rate than other fuels, due to which the value of ITE was found to be higher in CBG at higher speed. In Fig. 2c, CBG has less friction loss at low rpm than other fuels, and the difference between indicated power and brake power is less. Hence, the value of ME (75.04%) at low rpm was found to be higher in CBG. And as the speed increases, the friction loss also increases in CBG, so the ME value is found to be less at higher rpm than in other fuels. In contrast, the friction loss in gasoline and alcohol blends decreases, so the ME value was lower in CBG and higher in gasoline and alcohol blends.

Brake specific fuel consumption (BSFC) and indicated specific fuel consumption (ISFC)

Figure 3a shows that in the alcohol G90M10, the highest FF value was obtained at 1.13 kg/h at 2000 rpm, and

the G100 value was 0.65 kg/h. The lowest FF value at the lowest rpm was 0.88 kg/h and 0.55 kg/h in G80E20 blends and CBG, respectively. Fuel ITE with a higher flow rate will have higher BSFC and lower BTE value. In BSFC at 2000 rpm, G100 found 0.38 kg/kWh; the lowest BSFC value in the alcohol blend was 0.51 kg/kWh in the G80E20 and the highest at 0.97 kg/kWh in the G90MA10. A value of 0.32 kg/kWh was achieved in CBG, the lowest value among all the fuels overall, due to which the BTE value of CBG was achieved the highest. At the maximum rpm, i.e., at 4500 rpm, the value of FF in the G100 is 2.03 kg/h. The lowest value of 2.29 kg/h in alcohol blends is found in G90E10, and the highest is 2.5 kg/h in G80MA20. FF in the CBG value is obtained at 1.72 kg/h, which is the lowest compared to other fuels. At same rpm, the BSFC in CBG was 0.44 kg/kWh, while the G100 got 0.52 kg/kWh and G90M10 and G90MA10 got 0.6 kg/kWh. CBG consumes

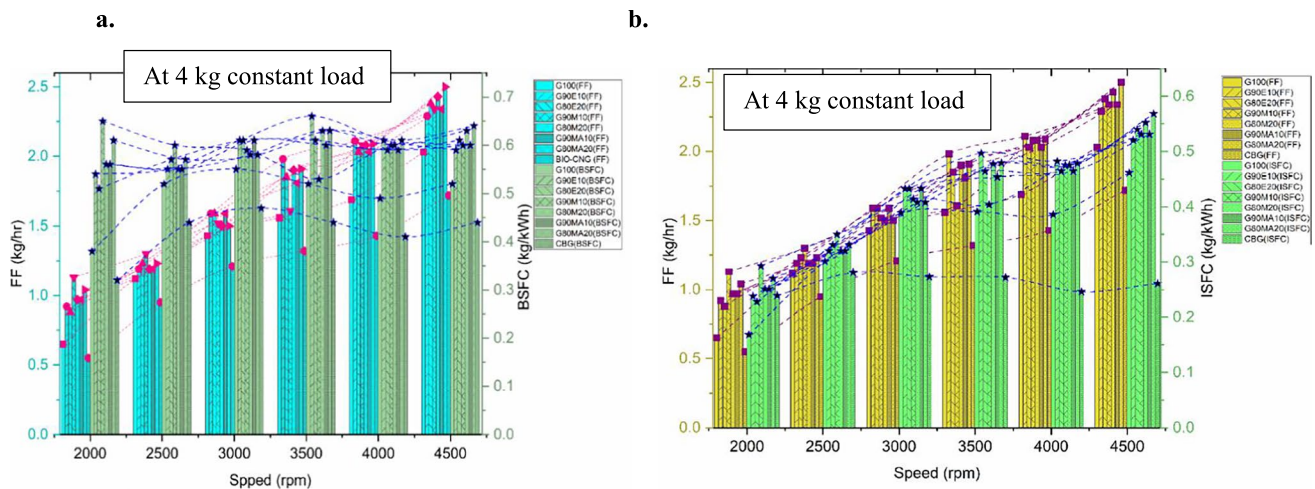


Fig. 3 a Fluid flow, brake specific fuel consumption varies w.r.t speed. b Fluid flow, indicated specific fuel consumption varies w.r.t speed

less fuel than other fuels at higher engine speeds, thereby increasing the engine's efficiency. Gasoline, G90E10, and G80E20 at 2000 to 2500 rpm have BSFC values in the range of 0.375 to 0.4 kg/kWh. As the RPM increases, the value of BSFC will also increase to a limit (Vivek Pandey and Gupta 2016). The BSFC value in G88M12 blends from 2000 to 3000 rpm has been found in the range of 0.42 to 0.4 kg/kWh (Mohammed Kamil and Ibrahim Thamer Nazzal 2016). At constant 1500 rpm and brake mean effective pressure (104 to 414 kPa), the MA5 and MA10 have obtained BSFC values between 0.9 and 0.3 kg/kWh (Cakmak et al. 2018).

In Fig. 3b, the ISFC value in the G100 was achieved at the minimum speed, i.e., 0.168 kg/kWh at 2000 rpm. Among alcoholic fuels, the IFSC was found to be 0.293 kg/kWh at the highest FF value in G90M10. The IFSC value of 0.228 kg/kWh was obtained in G80E20 at the lowest value of FF. CBG had the lowest value of FF compared to other fuels, while IFSC had a value of 0.239 kg/kWh. IFSC value at 4500 rpm in G100 was found to be 0.461 kg/kWh. In alcohol fuel G80MA20, the IFSC value was found to be 0.568 kg/kWh at the maximum FF value. IFSC value of 0.261 kg/kWh in CBG at maximum rpm was obtained, which was the lowest fuel consumption among all the fuels.

Combustion phenomenon

Figure 4a shows the start of burning (SOB) fuel in G100 and alcohol fuel when cylinder pressure is between 3 and 4 bar, and the crank angle is 335° before TDC. In CBG, SOB starts when cylinder pressure is 4.25 bar, and crank angle is 335° before TDC.

Experimental setup for CBG testing is started on gasoline fuel; when the engine cylinder pressure reaches 4 bar, SOB is started on CBG fuel. So, in gasoline and alcohol blends,

the SOB starts above 3 bar pressure, while in CBG, the SOB starts above 4 bar pressure. The SOB of a 100% gasoline and all alcohol mixture is started between 3 and 4 bar/335°. Whereas in the case of CBG, it began at 4.25 bar/335°, as the engine has to run at a higher speed than pure gasoline before running on CBG fuel, the cylinder pressure value also increased in the case of CBG. Cylinder pressure is calculated by taking an average of 10 cycles for each fuel. Ten percent fuel burn in all fuels starts just after TDC when cylinder pressure is 6 to 6.5 bar at a crank angle of 375°, and 90% fuel burn occurs in all fuels when cylinder pressure is 5 to 5.75 bar, and the crank angle is 415°. Maximum cylinder pressure was up to 6.79 and 6.76 bar, respectively, in the G90M10 and G90M20, and the lowest cylinder pressure achieved was 5.54 bar in the G80MA20 fuel when the crank angle was 385° after TDC. Maximum cylinder pressure in CBG is 6.06 bar at a 377° of crank angle after TDC, and its end-of-burning (EOB) fuel starts when cylinder pressure reaches 2.75 bar at a crank angle of 415° after TDC. In G100 and other alcohol fuels, when the cylinder range gets 1.25 bar at a crank angle of 450° after TDC, EOB starts in these fuels. CBG completes the EOB cycle earlier than gasoline, and alcohol blends because unburned particles are negligible in CBG, and the combustion cycle ends earlier. Whereas gasoline and alcohol blends contain more unburned particles, their EOB cycle is longer than CBG.

In Fig. 4b, the highest cylinder pressure value was found at 6.79 bar in the G90M10 when the cylinder volume was 48.58 cc, and in the G90M20, with a cylinder volume of 49.86 cc, the pressure value was 6.76 bar. Maximum cylinder pressure in G100 was 5.84 bar when the cylinder volume was 49.86 cc, and in CBG, the maximum pressure was 6.06 bar at a cylinder volume of 39.97 cc. Among all the fuels, the G80MA20 raised the lowest cylinder pressure to 5.54 bar when the cylinder volume value was 46.15 cc.

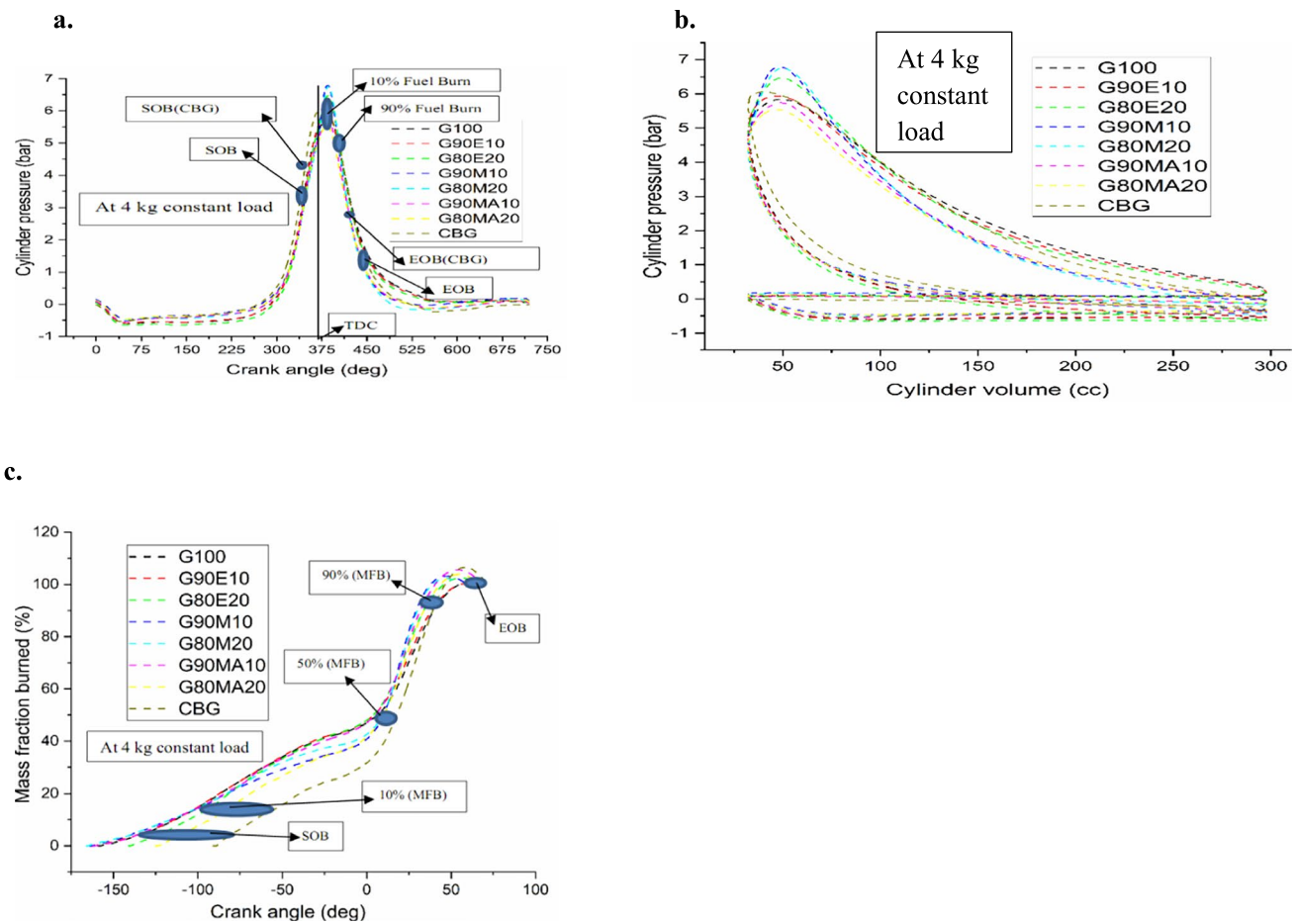


Fig. 4 **a** Cylinder pressure vs crank angle, **b** cylinder pressure vs cylinder volume, **c** mass fraction burned vs crank angle

Piston advances from TDC to BDC with the intake valve already open. As the piston completes its stroke, the volume keeps growing. When the piston is at BDC, the maximum volume is attained. Because the piston action creates volume and the vacuum effect draws air into the cylinder, the pressure is below atmospheric pressure throughout the stroke. Compression stroke starts once the piston has passed BDC. Volume begins to fall, and the pressure rises during this phase. Intake valve is still open even after the piston has passed BDC because it takes some time for the pressure inside the cylinder to exceed the pressure outside. Pressure progressively rises as the piston approaches TDC. When the ignition is started, the pressure increases until it reaches its peak. Since the cylinder's high pressure pushes the piston, the volume increases, and the pressure gradually decreases. Piston is back at the BDC after the power stroke. Once more, the cylinder's volume is at its maximum value, and its pressure is similar to the atmosphere. Cumulative heat release to total heat release ratio is known as MFB. Apparent heat release can be roughly calculated if the MFB is known as a function of crank angle. Value of MFB in CBG was lower

than in gasoline and alcohol, as there is complete combustion in CBG.

In Fig. 4c, before TDC, at a crank angle of 165 to 124°, G100 and alcohol fuel are just fuel-burning, whereas, in CBG, combustion starts when the crank angle is 89°. G100 and alcohol blends have a 5% MFB crank angle at 138.2 to 108.82° before TDC, while the CBG has this value at 79.55°. And when the crank angle is 138 to 93.76° before TDC, the G100 and alcohol blends burn 10% of the fuel, while the CBG burns when the crank angle is 67.27°. Fifty percent of MFB was found in G100 and rest alcohols at 9.08 to 2.91° after TDC, whereas in CBG, it was located at 16.95°. After TDC, 90% of MFB was detected in G100 and the rest in alcohols at 38.85 to 28.26°, while CBG was found at 38.16°. EOB in G100, alcohol blends, and CBG were located at 71 to 28.26° after TDC.

Conversion of chemical energy from the reactants in the charge into thermal energy is measured by the NHR profile, which is estimated from the cylinder pressure trace. Heat and mass transfer are not taken into account by the NHR profile. As shown in Fig. 5a, the maximum NHR

value of the average ten cycles in the G100 was 2.47 J/deg at a crank angle of 387°; similarly, the CBG averaged an NHR value of 2.41 J/deg at a crank angle of 388°. And the highest NHR value among alcohol blends was 3.08 J/deg at a crank angle of 376° in the G80M20 mixture. NHR value of the average cycle across all fuels was the highest at a crank angle of 376 to 388°. Figure 5b shows that the maximum mean gas temperatures in G100, G80E20 alcohol blends, and CBG with crank angles of 412°, 406°, and 411° were 384.2 °C, 390.20 °C, and 388.17 °C, respectively. The lowest MGT, 324.97 °C, was achieved in the G80MA20 at a 406° of crank angle. CBG and alcohol fuels are highly flammable as compared to gasoline fuels.

In addition to raising exhaust gas temperature and having a slower flame propagation speed than gasoline, CBG also has a higher auto-ignition temperature than other fuels. Therefore, CBG and alcohols G100E80 were found to have higher MGT values. In Fig. 5c, the highest CHR values of 0.12 kJ were found in the G100, CBG, G80E20, and G90E10 at crank angles of 432°, 420°, 422°, and 427°, respectively. And the lowest CHR value of 0.10 kJ is found in G80MA20 and G90MA10 at the crank angles of 419° and 418°. Due to CBG and ethanol blends are highly inflammable, the flame consumes the unburned mass. Hence, the maximum value of CHR was found in

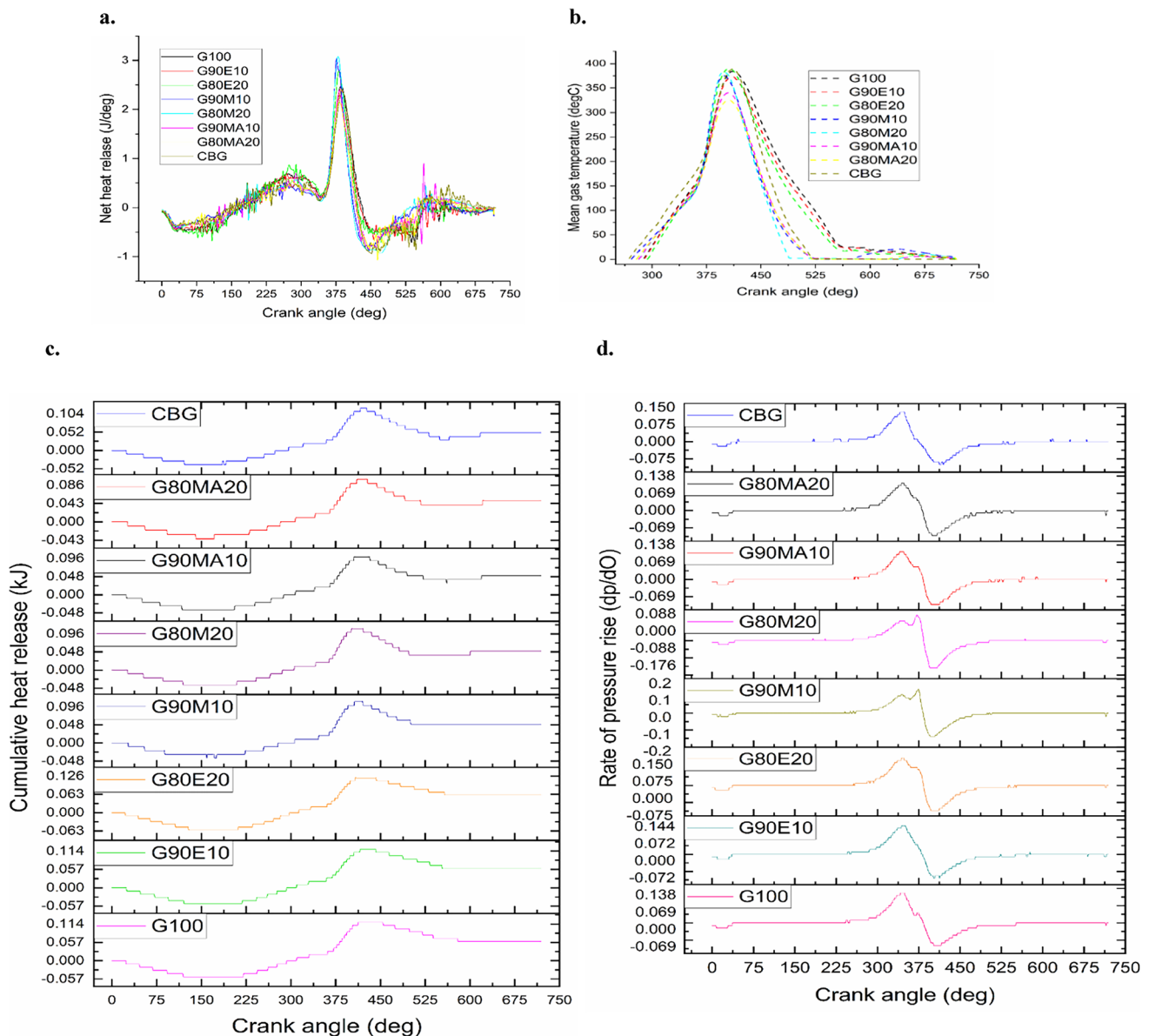


Fig. 5 **a** Net heat release vs crank angle. **b** Mean gas temperature vs crank angle. **c** Cumulative heat release vs crank angle. **d** Rate of pressure rise vs crank angle

these fuels, whereas in methyl acetate, it got a minimum value of CHR due to low flame.

Figure 5d shows that the G100, CBG, and G90M10 had maximum RPRs of 0.12, 0.13, and 0.14 bar/degree at 344°, 348°, and 374° of crank angles, respectively. At 344° and 346° of crank angles, the lowest RPR value of 0.11 bar/degree was achieved in G90MA10 and G80MA20. Gasoline, CBG, and methanol blends found the most significant increase in gas pressure during combustion, due to which the RPR value was higher in these fuels. Methyl acetate was found to have the lowest pressure increase during combustion, due to which the value of RPR was found to be the lowest in these blends.

Emission characteristics

Figure 6a presents that at 2000 rpm, the highest 20% and 22% CO₂ were obtained in the blends G90M10 and G90M20, respectively, and the lowest 3% was obtained in

CBG. And at the highest 4500 rpm, G100 and G80M20, CO₂ yielded were 13% and 21%, respectively, while CBG produced 6% CO₂, which means CBG green energy is considered the best alternative fuel of all fuels. Atoms of carbon and hydrogen constitute gasoline. CO₂ is created during combustion when oxygen is from the air and carbon from the fuel mix (CO₂). Similarly, in methanol blends and gasoline, the value of CO₂ has increased from the minimum speed to the maximum speed, which means that the CO₂ emission from methanol blends increases. Due to its low carbon content, CBG burns more cleanly than petroleum-based products. In addition, compared to gasoline and alcohol fuels, CBG emits 10 to 15% less CO₂. Maximum amount of CO is due to the burning of G100 fuel, which causes environmental pollution. In Fig. 6b, from the lowest speed to the highest speed, the maximum amount of CO was found in G100, from 1.64 to 2.63%.

The highest CO content of 1.45 to 1.08% was found in G90E10 among alcohol blends, and the lowest CO content

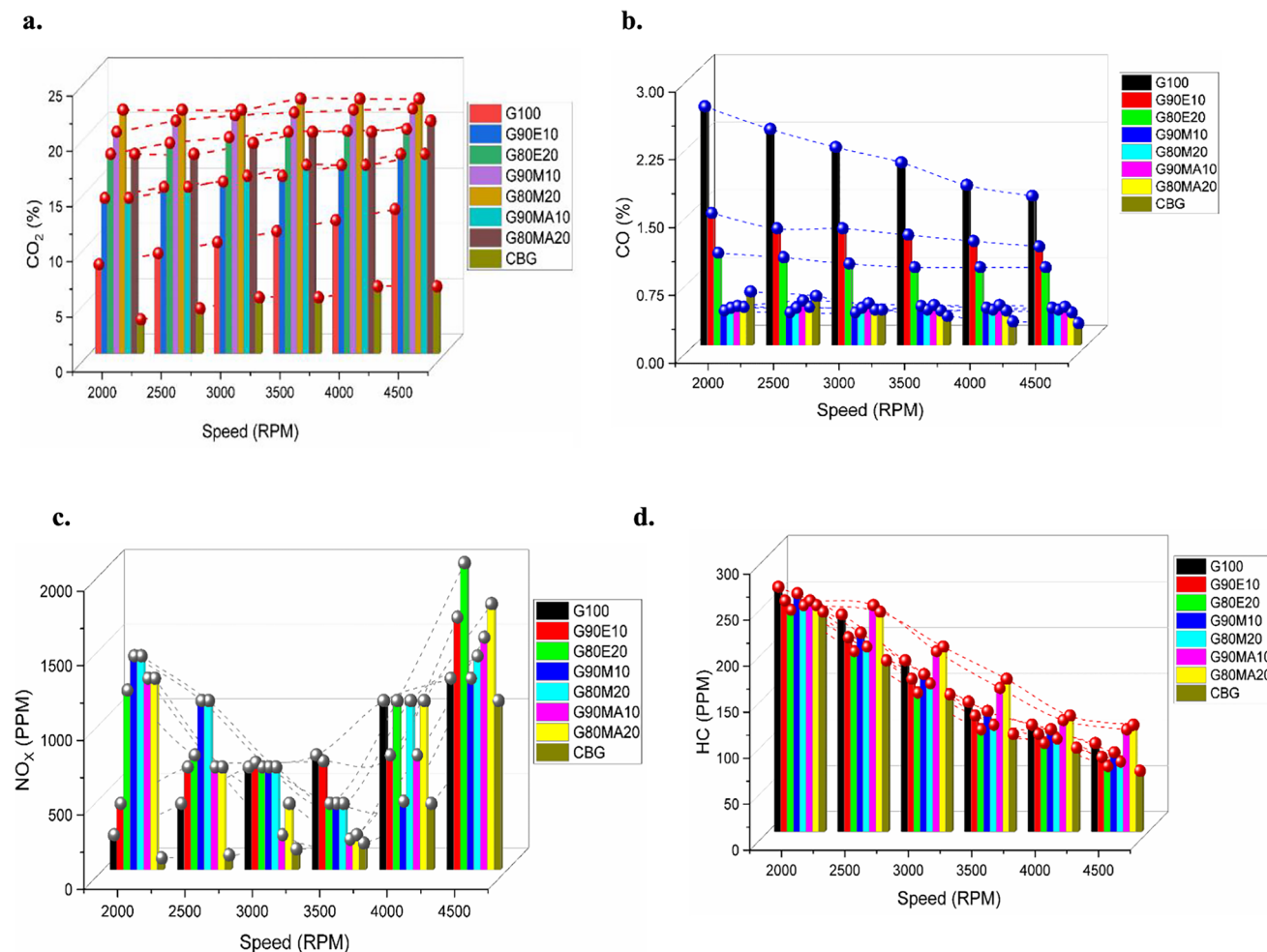


Fig. 6 a Carbon dioxide vary w.r.t speed. b Carbon monoxide vary w.r.t speed. c Nitrogen oxide vary w.r.t speed. d Hydrocarbon vary w.r.t to speed

at the highest speed was 0.232% in CBG. Due to incomplete combustion, a lack of oxygen, inadequate mixing, or all three, gasoline fuel was discovered to have a high CO content. Alcohol benefits engine performance and lowers exhaust since it has a high vaporization heat, octane number, and flammability temperature. Because alcohol is an oxygenate, meaning its molecules include oxygen, it burns efficiently and CO emissions are thus decreased. To assist the alcohol burn thoroughly, the oxygen atoms within it interact with the oxygen molecules in the surrounding air. When combined with alcohol, this extra oxygen makes gasoline burn more efficiently. Due to the low oxygen gas concentration in CBG, relatively little CO gas is generated.

As shown in Fig. 6c, the NO_x value in fuel G100 and alcohol blends G90M10 and G80M20 was found to be 225 and 1425 PPM at a minimum of 2000 rpm, while in CBG, its value was found to be 70 PPM. At maximum rpm, G80E20, G80MA20, and G100 have NO_x values of 2050, 1775, and 1275 ppm, respectively, while CBG has achieved 1125 ppm at the highest RPM, which means CBG emits the lowest NO_x from gasoline and other alcohol fuels and pollutes the environment less. Because engine speed affects NO_x emissions, when engine speed increases, more fuel is used, temperatures rise, and NO_x emissions increase. During combustion, nitrogen is oxidized to NO_x . Fuel burns more in the gasoline and alcohol band, which increases combustion temperature, cylinder pressure, and heat release, due to which these fuels were found to have higher NO_x values. In contrast, CBG had lower fuel consumption, allowing the engine performance increases, and NO_x is also emitted less.

Figure 6d shows that the HC values at minimum speed were 265, 258, and 238 PPM, respectively, in G100, G90M10, and CBG. And the HC values at maximum speed were 110, 115, and 65 PPM in the G90MA10, G80MA20, and CBG, respectively. Gasoline and alcohol blends have higher hydrocarbon emissions because the fuel does not burn entirely at low speeds. As the speed of the engine increases, the fuel starts burning well, so the value of HC is obtained less in all the fuels at higher rpm. CBG fuel burns well at minimum RPM to maximum RPM, due to which the HC value in CBG is rarely achieved at all RPMs.

In a study found, CO_2 values are ranging from 11 to 13% in gasoline at 2000 to 5000 rpm, CO values are ranging from 1.5 to 4.5%, and HC values are ranging from 180 to 450 ppm (Geok et al. 2009). Blends G85M15 and G70M30 at 2000 to 4000 rpm yielded CO values ranging from 0.14 to 0.06%, CO_2 in the range of 13.5 to 14.8%, and HC values ranging from 150 to 90 ppm (Shayan et al. 2011). The CO values ranged from 0.5 to 0.75% in blends G90E10 and G80E20 at 2000 to 4500 rpm, and HC values ranged from 145 to 65 ppm (Iodice and Cardone 2021). The CO_2 values ranged from 12.5 to 13.75% in blend G75E25 at 2000 to 4500 rpm, and the NO_x values

ranged from 800 to 600 ppm (Thangavelu et al. 2015). In methyl acetate blends G95MA5 and G90MA10, CO values ranged from 0.3 to 3.8% at constant 1500 rpm, while HC values ranged from 80 to 170 ppm and CO_2 values ranged from 10.5 to 13% (Cakmak et al. 2018).

Conclusion

At a constant load of 4 kg, from a minimum speed of 2000 rpm to a maximum speed of 4500 rpm, the FF rate (0.55–1.72 kg/h and BSFC 0.32–0.44 kg/kWh) in CBG fuel has been achieved, which is the lowest compared to gasoline and alcohol fuel blends, resulting in the highest BTE value in CBG at 23.33%. At a cylinder volume of 39.97 cc, the CBG achieved the highest cylinder pressure of 6.06 bar, and the G80MA20 achieved the lowest cylinder pressure of 5.54 bar among all fuels when the cylinder volume was 46.15 cc. At lower rpm, friction loss is higher in G100 and alcohol blends and lower in CBG, resulting in higher ME (75.05%) in CBG at lower rpm. SOB started at all fuels when the crank angle was 335°, and the cylinder pressure was between 3 and 4.50 bar. Its end-of-burning (EOB) began when the crank angle was 415° after TDC. Ninety percent mass of fraction burned in G100, alcohol blends, and CBG fuel after TDC was found at 38.85 to 28.26° of crank angle. In contrast, the EOB mass fraction was between 71 and 28.26° after TDC.

All alcohol blends have different properties due to their various characteristics, resulting in the G80M20 having an NHR value of 3.08 j/deg at a crank angle of 376°, which was higher than the NHR values for all fuels. The maximum mean gas temperature value in the G80E20 blends was achieved at 390.20 °C at a crank angle of 406 °C. At 432°, 420°, 422°, and 427° of crank angles, the G100, CBG, G80E20, and G90E10 achieved the highest CHR values of 0.12 kJ. The value of CO_2 , CO, HC, and NO_x emission gases in CBG at minimum speed to maximum speed is deficient compared to other fuels. Due to the low carbon content in CBG, it less pollutes the environment than gasoline and alcohol fuels. And it burns cleaner than petroleum-based products. Therefore, CBG fuel is also the best solution for solid organic waste, is the best alternative to gasoline fuel, and is eco-friendly. Our results suggest that CBG has the best results among all fuels in terms of engine performance, combustion, and emissions.

Acknowledgements The authors would like to thank Dr. Anil Kumar for his support in editing this research and the Delhi Technical University Administration for their valuable support.

Author contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were

performed by Pradeep Kumar Meena, Amit Pal, and Samsher Gautam. The first draft of the manuscript was written by Pradeep Kumar Meena, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data Availability Not applicable.

Declarations

Ethical approval As authors, we would like to tell you that this is our original work, and this paper has not been submitted anywhere except in this journal.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interest The authors declare no competing interests.

References

- Awad OI, Mamat R, Ali OM et al (2018a) Alcohol and ether as alternative fuels in spark ignition engine: a review. *Renew Sustain Energy Rev* 82:2586–2605. <https://doi.org/10.1016/j.rser.2017.09.074>
- Awad OI, Mamat R, Ibrahim TK et al (2018b) Overview of the oxygenated fuels in spark ignition engine: environmental and performance. *Renew Sustain Energy Rev* 91:394–408. <https://doi.org/10.1016/j.rser.2018.03.107>
- Cakmak A, Kapusuz M, Ganiyev O, Ozcan H (2018) Effects of methyl acetate as oxygenated fuel blending on performance and emissions of SI engine. *Environ Clim Technol* 22:55–68. <https://doi.org/10.2478/rtruet-2018-0004>
- Chauhan BS, Kumar N, Du Jun Y, Lee KB (2010) Performance and emission study of preheated Jatropa oil on medium capacity diesel engine. *Energy* 35:2484–2492. <https://doi.org/10.1016/j.energy.2010.02.043>
- D Deublein AS (2008) *Biogas from Waste and Renewable Resources*. Wiley-VCH Verlag GmbH & Co KGaA, Weinheim, Germany
- Geok HH, Mohamad TI, Abdullah S et al (2009) Experimental investigation of performance and emissions of a sequential port injection compressed natural gas converted engine. *SAE Tech Pap*
- Gülüm M, Bilgin A (2018) A comprehensive study on measurement and prediction of viscosity of biodiesel-diesel-alcohol ternary blends. *Energy* 148:341–361. <https://doi.org/10.1016/j.energy.2018.01.123>
- Holm-Nielsen JB, Al Seadi T, Oleskowicz-Popiel P (2009) The future of anaerobic digestion and biogas utilization. *Bioresour Technol* 100:5478–5484. <https://doi.org/10.1016/j.biortech.2008.12.046>
- Hosseini SE, Wahid MA (2013) Feasibility study of biogas production and utilization as a source of renewable energy in Malaysia. *Renew Sustain Energy Rev* 19:454–462. <https://doi.org/10.1016/j.rser.2012.11.008>
- Iodice P, Cardone M (2021) Ethanol/gasoline blends as alternative fuel in last generation spark-ignition engines: a review on co and hc engine out emissions. *Energies* 14:4034. <https://doi.org/10.3390/en14134034>
- Iodice P, Senatore A, Langella G, Amoresano A (2016) Effect of ethanol–gasoline blends on CO and HC emissions in last generation SI engines within the cold-start transient: an experimental investigation. *Appl Energy* 179:182–190. <https://doi.org/10.1016/j.apenergy.2016.06.144>
- Jhalani A, Sharma D, Soni S et al (2021) Feasibility assessment of a newly prepared cow-urine emulsified diesel fuel for CI engine application. *Fuel* 288:119713. <https://doi.org/10.1016/j.fuel.2020.119713>
- Larsson M, Grönkvist S, Alvfors P (2016) Upgraded biogas for transport in Sweden - effects of policy instruments on production, infrastructure deployment and vehicle sales. *J Clean Prod* 112:3774–3784. <https://doi.org/10.1016/j.jclepro.2015.08.056>
- Limpachoti T, Theinnoi K (2021) The comparative study on compressed natural gas (CNG) and compressed biomethane gas (CBG) fueled in a spark ignition engine. *E3S Web Conf* 302:01005. <https://doi.org/10.1051/e3sconf/202130201005>
- Mallikarjun MV, Mamilla VR (2009) Experimental study of exhaust emissions & performance analysis of multi cylinder S.I. engine when methanol used as an additive. *Int J Electron Eng Res* 1:201–212
- Mohammed Kamil, Ibrahim Thamer Nazzal (2016) Performance evaluation of spark ignited engine fueled with gasoline-ethanol-methanol blends. *J Energy Power Eng* 10:. <https://doi.org/10.17265/1934-8975/2016.06.002>
- Nuthan Prasad BS, Pandey JK, Kumar GN (2020) Impact of changing compression ratio on engine characteristics of an SI engine fueled with equi-volume blend of methanol and gasoline. *Energy* 191:116605. <https://doi.org/10.1016/j.energy.2019.116605>
- Pradeep Kumar Meena, Sumit Sharma AP (2022) Evaluation of in-house compact biogas plant thereby testing four-stroke single-cylinder diesel engine. In: *Introduction to Artificial Intelligence for Renewable Energy and Climate*. Scrivener Publishing, pp 277–343
- Rajesh Kumar B, Saravanan S (2016) Use of higher alcohol biofuels in diesel engines: a review. *Renew Sustain Energy Rev* 60:84–115. <https://doi.org/10.1016/j.rser.2016.01.085>
- Saidur R, Abdelaziz EA, Demirbas A et al (2011) A review on biomass as a fuel for boilers. *Renew Sustain Energy Rev* 15:2262–2289. <https://doi.org/10.1016/j.rser.2011.02.015>
- Shayan SB, Seyedpour SM, Ommi F et al (2011) Impact of methanol – gasoline fuel blends on the performance and exhaust emissions of a SI engine. *Int J Automot Eng* 1:219–227
- Subramanian KA, Mathad VC, Vijay VK, Subbarao PMV (2013) Comparative evaluation of emission and fuel economy of an automotive spark ignition vehicle fuelled with methane enriched biogas and CNG using chassis dynamometer. *Appl Energy* 105:17–29. <https://doi.org/10.1016/j.apenergy.2012.12.011>
- Thakur AK, Kaviti AK, Mehra R, Mer KKS (2017) Performance analysis of ethanol–gasoline blends on a spark ignition engine: a review. *Biofuels* 8:91–112. <https://doi.org/10.1080/17597269.2016.1204586>
- Thangavelu SK, Chelladorai P, Ani FN (2015) Emissions from petrol engine fueled gasoline-ethanol-methanol (GEM) ternary mixture as alternative fuel. *MATEC Web Conf* 27:2–5. <https://doi.org/10.1051/mateconf/20152701010>
- Vivek Pandey, Gupta VK (2016) Technical assessment of performance emission characteristics of an SI engine using ethanol-gasoline blended fuel. *Int J Eng Res* V5:422–426. <https://doi.org/10.17577/ijertv5is070431>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Summer 2022

Knowledge-Infused Learning

Manas Gaur

Follow this and additional works at: <https://scholarcommons.sc.edu/etd>



Part of the [Computer Sciences Commons](#), and the [Engineering Commons](#)

Recommended Citation

Gaur, M.(2022). *Knowledge-Infused Learning*. (Doctoral dissertation). Retrieved from <https://scholarcommons.sc.edu/etd/6914>

This Open Access Dissertation is brought to you by Scholar Commons. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact digres@mailbox.sc.edu.

KNOWLEDGE-INFUSED LEARNING

by

Manas Gaur

Bachelor of Technology
Guru Gobind Singh Indraprastha University, 2013
Master of Technology
Delhi Technological University, 2015

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Computer Science and Engineering

College of Engineering and Computing

University of South Carolina

2022

Accepted by:

Amit P. Sheth, Major Professor

Krishnaprasad Thirunarayan, Committee Member

Jyotishman Pathak, Committee Member

Biplav Srivastava, Committee Member

Valerie L. Shalin, Committee Member

Pooyan Jamshidi, Committee Member

Lorne Hofseth, Committee Member

Vignesh Narayanan, Committee Member

Tracey L. Weldon, Interim Vice Provost and Dean of the Graduate School

© Copyright by Manas Gaur, 2022
All Rights Reserved.

DEDICATION

To my sister, Chetanya, and my Mother,
I wasn't sure what I would do without you both.

ACKNOWLEDGMENTS

Frankly, I have never expected my Ph.D. to be such a joyful, exciting, and transformative experience of myself from a student, researcher, mentor, teacher to a human being. Quite oxymoronic to a very known fact of Ph.D.s to be the most challenging ultra marathon, I realized it to my breath as a journey that made me a calm, thoughtful and organized person. There is no way I would ever be able to cherish it without giving heartfelt gratitude to my family, fantastic professors, researchers, and friends who constitute my lifelong academic family.

The perseverance and determination to accomplish a Ph.D. came from my mother, Lata Gaur, who served a credible academic life as the University of Delhi's librarian. Seeing her fascination with reading and exploring new facets of education made me an out-of-the-box thinker. During my school days, I spent hours sitting in University Library and reading about computer science, mathematics, and competition review. My proclivity toward computer science came from my time at University, and today, I realize that every penny of the time spent has its worth. Further, having a younger and aspiring sibling is icing on the cake. My sister, Chetanya Gaur, developed a fondness for mathematics and research after seeing my Ph.D. life. I am amazed to see her growing into an economics researcher and would like to see her hooded with a doctorate.

Indeed, family members are your backbone and inspiration. Still, your Ph.D. colleagues and friends form an entirely new support system that brings in necessary distraction with entertainment and patiently listens to your anxious tirades.

My doctoral research started in Fall 2016 under the tutelage of Dr. Amit P. Sheth, then LexisNexis Ohio Eminent Scholar and Director of Kno.e.sis Center at Wright State Uni-

versity, and present, Founding Director of Artificial Intelligence Institute at University of South Carolina (AIISC). Throughout my stay at Kno.e.sis Center and AIISC, I was inspired to strive for the best in everything I do. Apart from astute research skills, a set of distinctive traits that I cultivated from my advisor and Ph.D. committee members includes (a) excellence in an industry-academic proposal and patent writing, (b) organization of conferences, workshops, and tutorials, (c) serving as guest editors in prestigious organizations like IEEE, (d) serving as session chairs and track chairs in ACM/IEEE conferences, and (e) mentoring diverse students from high school, undergraduate, and graduates. Finally, I want to take this opportunity to express my heartfelt gratitude for all who constantly and untiringly helped me during this journey.

First, I would like to thank my advisor Dr. Amit Sheth for his mentorship, motivation, and support and for considering me as their Ph.D. student. From the first day of my Ph.D., Dr. Sheth has been very energetic, disciplined, punctual, and hardworking towards his students' research, meetings, and individualistic growth. Being his Ph.D. student means adhering to Gurukul principles, a set of core principles that a researcher should follow to succeed. He acted as a catalyst that enforced those principles throughout my journey and made me work harder to compete with researchers in top schools and win prestigious awards/fellowships. He was never disappointed if I failed; rather, he gave me other opportunities to excel. He made me think of the famous "So What" question in any research, developing collaboration, working efficiently in a team, and presenting and communicating yourself in front of a diverse audience effectively. For him, non-technical skills, such as effective communication subtlety in writing, were equally important to technical skills. I must admire his proclivity in selecting the best research problem and motivating his students to draft a competitive NIH/NSF/DoD/AFRL proposal, which has improved my skills in writing quality research.

Further, my repetitive inclusion in proposal group meetings and contributing to some fantastic interdisciplinary research enhanced my research collaboration network. It allowed

me to co-author research with eminent researchers in cognitive science, medicine, neuropsychiatry, healthcare informatics, communication science, etc. Such encouragement towards internal and external collaboration has helped me develop the skill set required to collaborate in diverse environments and aim for small, medium, and large (NSF Institute Planning) grants. I credit all my accomplishments and accolades to Dr. Sheth, who has been a pillar of continuous support and guidance. I owe him a huge debt of gratitude for all he's done for me, and I consider myself extremely fortunate to be working under his supervision. "Knowledge-infused Learning" is a success of uncountable discussions through long campus walks, video calls, and corridor chats with Dr. Sheth on how to incorporate knowledge graphs in making explainable AI. I feel proud to say that Dr. Sheth is known to be among few CS scientists in AI and Semantic Web, and designing this dissertation with such an expert in my life's honor.

Obviously, none of the Ph.D. research is complete without a theoretical contribution in an area of specialization. I am grateful to be co-advised by Dr. Krishnaprasad Thirunarayan (a.k.a Dr. T.K Prasad), who has helped me shape the theoretical underpinnings of Knowledge-infused Learning. Under his supervision, I learned how to formulate a problem mathematically and learned pragmatic programming skills required to make your research reproducible. My first interaction with Dr. Prasad happened through the NSF-funded Hazard SEES project, in which I contributed to ontology creation and building language models to study Twitter streams. Every meeting of Hazard SEES would involve researchers from diverse backgrounds, and I was amazed to see Dr. Prasad's to-the-point clarifying questions in simplifying the problem and responses when one of us gets stuck while presenting. In such meetings, I realized the diverse knowledge he possesses and the quality of questions he asked that shape a student's research and win NSF/NIH proposals. I am always eager to share my first draft of research, proposal, or even dissertation with him before Dr. Sheth. Since Dr. Prasad and Dr. Sheth's minds resonate, I always benefited from the multi-rounds of reviews I got in shaping my research. I want to acknowledge here that my

first co-authored paper in IEEE data science and advanced analytics came from the theory on Trust Networks that Dr. Prasad presented in Kno.e.sis Center. I am sure, as long as I aspire to be in research and academia, I will always look up to Dr. Prasad for an extra pair of eyes.

When I was not talking about research with Dr. Prasad, we shared the achievements of my sibling and his kids (Neeti and Vidur). Both Neeti and Vidur have astonished me at a very young age with their accomplishments. He would also check on my father's health when my father was undergoing a chronic phase in life. He listened to me and conversed during my hard times, irrespective of time and place. I am very fortunate to work under his guidance. Both Dr. Prasad and Dr. Sheth inculcate in me a habit of constant reading and continual learning, as these are the source of constant growth as an academician and researcher.

Alongside Dr. Prasad, I was fortunate for my acquaintance with Dr. Valerie Shalin, a member of the NSF HazardSEES project. I first collaborated with Dr. Valerie Shalin on the Wisdom of Crowd project, winning the runner-up award in the Web Intelligence Conference in 2018. One characteristic of Dr. Shalin that intrigued me was the predisposition to learn and read about current research in psychology and computer science, especially natural language understanding (NLU) and AI, which is my forte. During group meetings, paper discussions, or proposal writing, I have seen how her thoughts shape the research directions by highlighting pressing problems, intuitiveness of the solution, and impact of the outcome. Apart from being well-read, she brings along with her years of experience in research, having knowledge of AI research from the 1960s to 2000s which helped me craft the design of experiments. Under her supervision, I successfully put my first ever research in ACM CSCW, a purely interdisciplinary computer science conference. I would like to acknowledge that her "diamond style of writing" theory is very effective and has recently given me back-to-back acceptance in the top-most CS/AI/NLU conferences. Also, in my mind, I used to consider Dr. Shalin as a "sentence/word cutter" as she is very ver-

satire in shortening 25 pages long draft of a proposal to 15 pages. Because of her skill, we were able to timely submit some competitive NIH/NSF proposals. Along with her and Dr. Sheth, I learn about framing the problem, and together with Dr. Prasad, the conceptualization happens with necessary mathematical groundings. A person similar in work ethics and beautiful craftsmanship in proposal writing like Dr. Sheth, Dr. Prasad, and Dr. Shalin is Dr. Jyotishman Pathak. I am delighted to have collaborated with him on multiple projects at the intersection of social computing in healthcare and AI. A trait in me that I acquired after collaborating with him is multitasking with time management. He has always been a PI/Co-PI on some phenomenally large NIH grants, which he has written or collaborated with other esteemed researchers. While virtually attending his group meeting at Weill Cornell Medicine, I recognized how he understands the conceptual problems faced by students/researchers while working on these grants and suggests feasible alternatives to reach the outcome. After Dr. Sheth, Dr. Prasad, and Dr. Shalin, Dr. Pathak has excellent skills in effective scientific communication with fabulous command of the choice of words. Though I am still learning from these maestros of impactful presentations, I credit my achievements of research acceptance in non-technical journals to his skill set to revamp the technical and complex draft into a more subtle and precise research article understandable to non-technical audiences. Finally, I want to acknowledge that he is the fittest faculty I have noticed in my network and I cherish sharing my marathon and ultra-marathon training goals with him.

My personal and professional connection with Dr. Prasad, Dr. Shalin, and Dr. Pathak has been persistent despite my move with Dr. Sheth to the newly founded Artificial Intelligence Institute at The University of South Carolina (AIISC). Particularly, I admire their support, irrespective of authorship, in assisting me while submitting conference/journal rebuttals, which are scary to convince the reviewers. Diversification in my Ph.D. organization of workshops, conferences, and doing tutorials came after meeting with Dr. Biplav Srivastava. He is known as the master inventor in IBM with a record of 50+ patents and

organizers of conferences and workshops. I remember him guiding me through the organization process of my first ACM SIGKDD workshop on Knowledge-infused Mining and Learning. With his support, I put together a stellar and well-attended workshop program at SIGKDD, which resulted in me winning the high-ranking social influencer award at the conference. Also, since 2020, I have been co-organizing CASY, the university-wide conference on Collaborative Assistants for the Society, with support from Dr. Srivastava and AIISC colleagues. I am delighted to gain experience in patent writing from him. Furthermore, I'm extremely thankful to him for his time and effort in improving my social presentation skills.

I extend my deepest gratitude to Dr. Lorne Hofseth and Dr. Pooyan Jamshidi for being on my Ph.D. dissertation committee and for the helpful and insightful conversations during large-scale research proposals. I also want to thank other mentors that I've had throughout my career, from Dr. Carlos Castillo, Dr. Dilshod Achilov, Dr. Hemant Purohit, Dr. Joseph Reagle, Dr. Ke Zhang, Dr. Meera Narasimhan, Dr. Michael Huhns, Dr. Pavan Kapanipathi, Dr. Raminta Daniulaityte, Dr. Randy Welton, Dr. Saeedeh Shekarpour, Dr. Sam Anzaroot, Dr. Shayak Bhattacharya, Dr. Srinivasan Parthasarathy, Dr. Sriraam Natarajan, and Dr. William Groves. Also, thanks to all the diverse high school, undergraduate, masters', and Ph.D. students around the globe whom I got to mentor and all the expert crowd workers that worked on my datasets. Special thanks to Amanuel Alambo and Swati Padhee, whom I mentored during my time in Kno.e.sis Center. They have proved to be amazing independent researchers.

I want to thank my colleagues in Kno.e.sis Center and present AIISC, who patiently tolerated my mischievous behavior in different time brackets over six years: Kalpa Gunaratna (my mentor in Samsung Research America and hitting AAAI), Sanjaya Wijeratne (a friend in need is a friend indeed), Sarasi Lalithsena (learned how to do independent research), Shreyansh Bhatt (first co-authored award-winning paper from Kno.e.sis Center), Sujana Perera (my first mentor in Kno.e.sis Center), Ugur Kurşuncu (more than a friend, a

brother with whom I could seek advice on personal issues and an accomplished collaborator), Deepa Tilwani (an incoming Ph.D. student who exactly knows what she wants to achieve in Ph.D. From her time as an intern to being officially accepted as a Ph.D. student with marvelous recommendations, I know she will rise), Utkarshani Jamini (a.k.a. ninni, whom I consider as my sister), Ruwan Wickramarachchi (a friend from whom you can expect wise and unbiased advice; I will always remember him saying: “Regarding marriage, Manas, you got all your concepts wrong!!”), Joey Yip (was amazed by his creativity in research), Thilini Wijesiriwardene, Revathy Venkataramanan (love to see her cat on her Instagram), Kaushik Roy (a coffee-food addict like me and my younger brother), and Amelie Gyrard (the post-doc from whom I learnt efficient multi-tasking).

Special thanks to Dr. Christian O’Reilly for being my partner in half-marathon, marathon, and ultra-marathon training. It was super fun running with him over long distances as he is such a kind of person with whom you won’t feel bored as there is always a topic for conversation. I also extend my thanks to Dr. Vignesh Narayanan for allowing me to teach a few classes in the Neural Network and Optimization course. Further, I would like to thank Dr. Qi Zhang for assisting Kaushik Roy and crafting a powerful paradigm within Reinforcement Learning: Knowledge-infused Reinforcement Learning. Also, it was a pleasure collaborating with you on the NSF SCH proposal.

I was extremely fortunate to be at Stanford Research Institute (presently SRI International) with Dr. Natarajan Shankar, Dataminr Inc. as AI for Social Good Fellow with Dr. Alejandro (Alex) Jaimes, Data Science for Social Good Fellow with Dr. Rayid Ghani, Dr. Leid Zejnilovic and Dr. Qiwei Han, Samsung Research America with Dr. Hongxia Jin, Dr. Vijay Srinivasan, and Dr. Kalpa Gunaratna, and Weill Cornell Medicine with Dr. Jyotishman Pathak. I want to extend my gratitude and appreciation to these amazing people who have introduced me to powerful methods to conduct impactful research. I greatly appreciate my fortune to have met these amazing people in my journey to achieve a doctorate, and it would not have been possible without the prayers and support of my mother and father.

Thanks are even less to appreciate the enthusiasm, interest, and support for my research endeavors from my parents and sister. In addition, my sister, Chetanya Gaur, has always been there for me despite being on the other side of the globe. My Ph.D. journey from State University of New York to Wright State University to the University of South Carolina has drawn a memorable chapter in my life, which will be a living, breathing, and ever-changing blueprint of dedication and perseverance.

GRANT ACKNOWLEDGEMENT

I acknowledge partial support from the National Science Foundation (NSF) award CNS-1513721: “Context-Aware Harassment Detection on Social Media”, NSF award 2133842 “EAGER: Advancing Neuro-symbolic AI with Deep Knowledge-infused Learning”, NSF award 1761931: “Spokes: MEDIUM: MIDWEST: Collaborative: Community-Driven Data Engineering for Substance Abuse Prevention in the Rural Midwest”, National Institutes of Health (NIH) award: MH105384-01A1: “Modeling Social Behavior for Healthcare Utilization in Depression”, and National Institute on Drug Abuse (NIDA) Grant No.

5R01DA039454-02 “Trending: Social media analysis to monitor cannabis and synthetic cannabinoid use”. I also acknowledge generous support through EPSRC-UKRI Alan Turing Institute Fellowship for Mental Health Research, travel award as a part of University of Chicago Data Science for Social Good Fellowship, Microsoft Research Travel Award, Dataminr Ph.D. Fellowship, and Samsung Research America Ph.D. Research on Knowledge-infused Learning. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, NIH, NIDA, Microsoft Research, Dataminr, Samsung Research America, or Alan Turing Institute.

ABSTRACT

In DARPA's view of the three waves of AI, the first wave of AI, symbolic AI, focused on explicit knowledge. The second and current wave of AI is termed statistical AI. Deep learning techniques have been able to exploit large amounts of data and massive computational power to improve human levels of performance in narrowly defined tasks. Separately, knowledge graphs have emerged as a powerful tool to capture and exploit a variety of explicit knowledge to make algorithms better apprehend the content and enable the next generation of data processing, such as semantic search. After initial hesitancy about the scalability of the knowledge creation process, the last decade has seen significant growth in developing and applying knowledge, usually in the form of knowledge graphs. Examples range from the use of DBPedia in IBM's Watson to Google Knowledge Graph in Google Semantic Search to the application of ProteinBank in AlphaFold, recognized by many as the most significant AI breakthrough. Furthermore, numerous domain-specific knowledge graphs/sources have been applied to improve AI methods in diverse domains such as medicine, healthcare, finance, manufacturing, and defense.

Now, we move towards the third wave of AI built on the "Neuro-Symbolic" approach that combines the strengths of statistical and symbolic AI. Combining the respective powers and benefits of using knowledge graphs and deep learning is particularly attractive. This has led to the development of an approach and practice in computer science termed "knowledge-infused (deep) learning" (KiL). This dissertation will serve as a primer on methods that use diverse forms of knowledge: linguistic, commonsense, broad-based, and domain-specific and provide novel evaluation metrics to assess knowledge-infusion algorithms on various datasets, like social media, clinical interviews, electronic health records,

information-seeking dialogues, and others. Specifically, this dissertation will provide necessary grounding in shallow infusion, semi-deep infusion, and a more advanced form called deep infusion to alleviate five bottlenecks in statistical AI: (1) Context Sensitivity, (2) Handling Uncertainty and Risk, (3) Interpretability, (4) User-level Explainability, and (5) Task Transferability. Further, the dissertation will introduce a new theoretical and conceptual approach called Process Knowledge Infusion, which enforces semantic flow in AI algorithms by altering their learning behavior with procedural knowledge. Such knowledge is manifested in questionnaires and guidelines that are usable by AI (or KiL) systems for sensible and safety-constrained response generation.

The hurdle to prove the acceptability of KiL in AI and natural language understanding community lies in the absence of realistic datasets that can demonstrate five bottlenecks in statistical AI. The dissertation describes the process involved in constructing a wide variety of gold-standard datasets using expert knowledge, questionnaires, guidelines, and knowledge graphs. These datasets challenge statistical AI on explainability, interpretability, uncertainty, and context-sensitivity and showcase remarkable performance gains obtained by KiL-based algorithms. This dissertation termed these gold-standard datasets as Knowledge-intensive Language Understanding (KILU) tasks and considered them complementary to well-adopted General Language Understanding and Evaluation (GLUE) benchmarks. On KILU and GLUE datasets, KiL-based algorithms outperformed existing state-of-the-arts in natural language generation and classification problems. Furthermore, KiL-based algorithms provided user-understandable explanations in sensitive problems like Mental Health by highlighting concepts that depicts the reason behind model’s prediction or generation. Mapping of these concepts to entities in external knowledge source can support experts with user-level explanations and reasoning. A cohort-based qualitative evaluation informed that KiL should support stronger interleaving of a greater variety of knowledge at different levels of abstraction with layers in a deep learning architecture. This would enforce controlled knowledge infusion and prevent model from extrapolating

or overgeneralization. This dissertation open future research questions on neural models within the domain of natural language understanding. For instance, (a) Which layer within a deep neural language model (NLMs) require knowledge? (b) It is known that NLMs learn by abstraction. How to leverage external knowledge’s inherent abstraction in enhancing the context of learned statistical representation? (c) Layered knowledge infusion might result in high-energy nodes contributing to the outcome. This is counter to the current softmax-based predictions. How to pick the most probable outcome? and others. This dissertation provide a firsthand towards addressing these questions; however, much efficient methods are needed that provide user-level explanations, be interpretable, and propel safe AI.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGMENTS	iv
ABSTRACT	xii
LIST OF TABLES	xvii
LIST OF FIGURES	xxii
CHAPTER 1 INTRODUCTION	1
CHAPTER 2 FROM BLACK BOX TO GREY BOX: IMPROVING INTERPRETABILITY AND EXPLAINABILITY OF DEEP LEARNING SYSTEMS	19
2.1 Context Sensitive Capture	21
2.2 Explaining DL Models	25
2.3 Interpretable Models	28
2.4 From GLUE to KILU	30
2.5 Summary	32
CHAPTER 3 SHALLOW INFUSION	35
3.1 Benefits of Shallow Infusion	37
3.2 Method under Shallow Infusion	41

3.3	Shallow Infusion For Suicide Risk Severity Detection	45
CHAPTER 4 SEMI-DEEP INFUSION		66
4.1	Benefits of Semi-Deep Infusion	67
4.2	Semi-Deep Infusion	70
4.3	Semantic Encoding and Decoding Optimization (SEDO)	77
4.4	ISEEQ for Conversational Information Seeking	81
4.5	ISEEQ Architecture and Evaluation	87
4.6	Summary	99
CHAPTER 5 PROCESS KNOWLEDGE INFUSION		102
5.1	Process Knowledge and Its Infusion into Statistical AI	104
5.2	Benefits of Process Knowledge Infusion	106
5.3	C-SSRS 2.0	111
5.4	PRIMATE	119
5.5	Summary	127
CHAPTER 6 DEEP KNOWLEDGE INFUSION		129
6.1	Deep Infusion Module	131
6.2	Differential Knowledge Engine	136
6.3	Deep Infusion in Neural Language Models	138
6.4	Summary	141
BIBLIOGRAPHY		142

LIST OF TABLES

Table 1.1	Reflection of pioneer in AI towards knowledge infusion and Neuro-symbolic AI in general.	4
Table 1.2	Summary of Contributions made by KiL. Each of the task and application comprises of either dataset construction, knowledge source construction or adapting existing datasets for the task/application. Further, outcomes from KiL-based algorithms on different testsets is evaluated by subject matter experts by conducting blind human evaluation.	18
Table 2.1	Benefits of context capture by KiL.	24
Table 2.2	Comparison between different methods that make DL algorithms interpretable. The complexity classification is discussed in detail in Chapter 3.	30
Table 2.3	GLUE tasks are classification or prediction tasks taking a sentence or pair of sentences as input. It is not meant for generation or structured prediction. On the other hand KILU tasks subsumes GLUE Tasks and challenges DL models on user-level explainability and interpretability. To provide explanations to KILU tasks, the model should leverage variety of explicit knowledge to capture context and learn necessary abstraction for human comprehension. EM: Evaluation Metrics used in GLUE and KILU.	33
Table 3.1	Other methods that are classified under shallow infusion. However, not all of them are supportive of system-level explainability (SysEx). Methods which are SysEx are also capable of user-level explainability with manual effort comprising of search and retrieval over related knowledge sources [1].	42
Table 3.2	Suicide Risk Severity Lexicon. It can be downloaded from here	48

Table 3.3	Shallow infusion improves recall when the model is tasked to predict suicide risk severity of a user. In such scenario Recall is the judge of model’s performance as high false negatives would result in wrong care plan for a patient with high-levels of suicide risk tendencies.	53
Table 3.4	Example posts from a user ordered by timestamp (TS) and prediction from LSTM with semantic embedding loss. These examples illustrates the longitudinal efficiency brought into statistical LSTMs through shallow infusion.	54
Table 3.5	Example posts from a user(u_i) and prediction from TinvM. The italicized text are phrases which contributed to the representation of the post. These phrases had similarity to the concepts in medical knowledge bases	55
Table 3.6	Qualitative comparison of TinvM and TvarM models representative posts from users who are either supportive or showing signs of suicide ideations, behaviors or attempt. Pred.: Predictions, SW: r/SuicideWatch	58
Table 3.7	Paraphrased posts from candidate suicidal redditors and associated suicide risk severity level.SU: Supportive users or no-risk users.	60
Table 3.8	(left). Pairwise annotator agreement, (right). Group wise annotator agreement. A, B, C,and D are annotators. Agreement scores are for Time-invariant modeling of suicide risk severity dataset.	61
Table 3.9	Inter-rater reliability agreement using Krippendorff metric. A,B,C,and D are mental healthcare providers as annotators. The annotations provided by MHP “B” showed the highest pairwise agreement and were used to measure incremental groupwise agreement for the robustness in the annotation task. Agreement scores are for Time-variant modeling of suicide-risk severity dataset.	61
Table 4.1	Existing methods and approach that are classified based on whether they provide user-level explainable and knowledge-based interpretability. DLMS: Deep Language Models.	71
Table 4.2	Evaluating retrievers. ECE: Electra Cross Encoder, (*): variant of (Clark et al. 2019), DPR: Dense Passage Retrieval.	91
Table 4.3	An ablation study showing improvement in the quality of ISQs after encodings of retrieved passages ($\mathbf{P}_{1:K}$) are concatenated with knowledge-augmented query (k_d) after SQE. The concatenation is performed for each $p \in \mathbf{P}_{1:K}$	93

Table 4.4	Scores on test set of datasets. In comparison to T5-FT CANARD, a competitive baseline, ISEEQ-ERL generated better questions across three datasets (30%↑ in QADiscourse, 7%↑ in QAMR, and 5%↑ in FB Curiosity). For fine-tuning we used SQUADv2.0.	95
Table 4.5	Performance of KPR on MS-MARCO passages while retrieving atleast one passage per IS query in CAsT-19. 269 is the size of CAST-19 train set. KPR covered the train set but left 16% of the IS queries in test set. Ret.Pass. : Retrieved Passages.	96
Table 4.6	Transferability test scores using ISEEQ-ERL to answer RQ3. gray cell: ISEEQ-ERL trained and tested on same dataset. dark gray cell: shows acceptable cross-domain {Train-Test} pairs, where train size is smaller than test size.	98
Table 4.7	Assessment of human evaluation. G1: ISQs are diverse in context and non-redundant. G2: ISQs are logically coherent and share semantic relations. >: difference is statistically significant. SD: Standard Deviation. S1, S2, and S3 are ground truth, ISEEQ-ERL, and T5-FT CANARD, respectively.	99
Table 5.1	This is an example of how a process knowledge-integrated dataset is constructed in collaboration with mental healthcare providers. The leftmost column presents example questions mental healthcare providers (MHPs) asked. The MHPs provided Tag and Rank shown in the rightmost columns representing process knowledge. The middle column provides a series of questions gathered using Google SERP API (https://tinyurl.com/G-SERP-api) and Bing Search API (https://tinyurl.com/bing-search-api) logically ordered by MHPs.	108
Table 5.2	Attention visualization based explanations in C-SSRS 1.0	110
Table 5.3	A process knowledge-guided improved explanations in C-SSRS 2.0 . . .	110
Table 5.4	Explanations provided from W2V on C-SSRS 2.0.	118

Table 5.5	Examples of questions generated by T5 when tasked to generate FQs when the user query for the post in Figure 5.3 was provided as input. Model 1 , which is a pre-trained T5 [2], often generates questions which are irrelevant, unsafe, incoherent, and redundant. Model 2 , which is T5 fine-tuned on r/depression_help seems to be relatively coherent and inquisitive compared to Model 1 . However, both models generate questions about the topic that user has discussed in their query. As a result, we see that pre-trained and fine-tuned DLMs fail to generate FQs. By enforcing FQ generation using using a dataset curated using extended PHQ-9, generated questions have been mostly inquisitive. This is shown by Model 3 . Still, a lot of generations are around the problem the user mentioned.	121
Table 5.6	In this example, the generated questions from both Model 2 and Model 3 seem to be relevant FQs, but they are not assessing the severity of the mental health condition, despite Model 3 being fine-tuned on a dataset filtered by PHQ-9 questions. In comparison to the qualitative outcome in Table 5.5, this showcases the inability of T5 to support mental health triage.	122
Table 5.7	Experimental results comparing different models in generating questions that match the sub-questions in PHQ-9. \hat{Q} is the set of generated questions in each chunk. The performance is recorded over all the generated questions (\hat{Q}). δ was used as the threshold on the similarity between generated question and PHQ-9 sub-questions while calculating hit rate. BLEURT records semantic similarity, whereas Rouge-L records the longest common subsequence exact match between generated question and PHQ-9 sub-questions. The highest performance on semantic and string similarity is bolded. Acceptable performance in Model 3 achieved using PHQ-9 motivated us to prepare PRIMATE	124
Table 5.8	Distribution of 2003 posts in PRIMATE according to whether the text in the post answers a particular PHQ-9 question. Through this imbalance, PRIMATE presents its importance in training DLM(s) to identify potential FQs in PHQ-9 that would guide a generative DLM(s) to conduct a discourse with a patient with a vision to assist MHPs in triage. Q1-Q9 are described in Figure 5.7	125

Table 5.9	The MCC score for all 9 questions across different thresholds is in the range 0 to +1 (low to high positive relationships). The MCC for some configurations runs into a divide by zero error, and we replace this value with 0.0. Unable : model is unable to learn cues to determine answerability in a post. Maybe : model is uncertain whether a particular PHQ-9 question is answerable or not. Certain : answerability can be determined by the model with high reliability. Class-Type: Classification Type when $\delta = 0.9$	126
-----------	--	-----

LIST OF FIGURES

Figure 1.1	<p>(left) Neural Networks and Deep Learning Models (e.g classification/discriminative models, generative models) modeled as black box because the decision/actions do not support user-level explainability. Further, there is no other means to interpret the internal mechanics of black box other than studying the low-level data. This black box is considered as system 1. (right) System 2 represents symbolic knowledge represented in the form of a knowledge graph (KG; a graph), semantic lexicons (a dictionary), rules (a set of constraints) that can be made in machine understandable form and infused in System 1 for model interpretations. To study the outcome of System 1, then System 2 supports user-level explainability.</p>	2
Figure 1.2	<p>An illustration of user-level explainability using the important conceptual phrases identified by a deep learning model trained using the method described in the Presentation. Highlighted phrases in (A) are queried in SNOMED-CT, thus forming a contextual tree. Formation of this tree is stopped when a node is hit that has high similarity to either leaf nodes or one hop parent nodes. The resulting tree is shown in (B). The numbers in the boxes are SNOMED-CT IDs.</p>	9
Figure 1.3	<p>An illustration of the process followed by human annotators(or subject matter experts) in creating datasets of high quality. (A) It shows that annotators (or subject matter experts) used the external pieces of information but were not provided to ML/DL models. (B) Suppose these external sources of knowledge for annotators or experts are infused into ML/DL models and used to evaluate model outcomes. In that case, we achieve model interpretability and user-level explainability. (C) Since these diverse sources of knowledge are voluminous, their infusion can be a resource and computation-heavy. We believe information graphs can replace these sources, and being machine-readable, their infusion would not increase computation costs drastically.</p>	10

Figure 1.4	Example KG constructed either from manual effort (A, B, C), automatically (D, E), or semi-automatically (F). (A) is empathi ontology designed to identify concepts in disaster scenarios [3]. (B) Chem2Bio2RDF [4]. (C) ATOMIC [5]. (D) Education Knowledge Graph by Embibe [6]. (E) Event Cascade Graph in WildFire [7]. (F) Opioid Drug Knowledge Graph [8]	12
Figure 1.5	Technical Contributions (Y) made by KiL to address limitations of current data driven ML/DL algorithms (X).	17
Figure 2.1	Infuse knowledge context to capture conceptual (left) and ambiguous entities (right) for correct classification in QQP dataset. Picture credit to [9].	22
Figure 2.2	(A) System-level Explainability showing the overlap (a proxy of matching a support seeker (SS) with its nearest support providers (SPs) using T-SNE visualization. (B) User-level Explainability showing the reason behind mapping semantically-related SPs to a SS through the use of phrases that are semantically similar to concepts in Patient Health Questionnaire Lexicon (PHQ-9).	26
Figure 2.3	An illustration of context modeling in language model using external domain-specific corpus. Subsequent clustering manifest pairing of concepts that co-occur in a domain-specific corpus. The clusters are explainable using the relationships between these concepts. This figure illustrate deeper semantics in a computational social science problem of detecting radicalization behaviors in dynamic stream of tweets. The word <i>jihad</i> occur in two connotations and is clearly separable using domain-specific knowledge.	31
Figure 3.1	An Illustration of ordered knowledge (right; also called process knowledge) constructed from the Columbia Suicide Severity Rating Scale (left), one of questionnaire used by MHPs for suicidality detection. . . .	36
Figure 3.2	An illustration that associates system-level explainability with user-level explainability. The highlight phrases in the left-side of the figure is obtained from a DL model trained using the method in Gaur et al. [10]. This is a manifestation of system-level explanations. Highlighted phrases in the input text are queried in SNOMED-CT, thus forming a contextual tree (right-side of the figure). This is manifestation of user-level explanations. Formation of this tree is stopped when a node is hit that has high similarity to either leaf nodes or one hop parent nodes. The numbers in the boxes are SNOMED-CT IDs. . .	37

Figure 3.3	An illustration of concept classes to assess suicide risk. These concept classes are obtained from Columbia Suicide Severity Rating Scale [11]. Dotted arrow from a “not-so-well-defined” label to well-defined concept class shows that the label can resembles this class if predicted probability for solid arrow is lower than dotted arrow. Solid arrow from “not-so-well-defined” labels to well-defined concept classes shows that these labels certainly resembles this class if predicted probability for solid arrow is higher than dotted arrow. This dichotomy on the part of “not-so-well-defined” labels is removed using concept classes.	38
Figure 3.4	A generic architecture of Shallow Infusion	40
Figure 3.5	A shallow infusion process using contextual dimensions from the radicalization literature. The visualization is performed using T-SNE method [12]. Explainable view of the clustering is provided in Figure 2.3	41
Figure 3.6	A view of input raw text being annotated by the expert and a model, respectively. It illustrates the gap between the “what a model understands as important features” compared with “how an annotator sees the text”.	46
Figure 3.7	A snapshot illustrates that the discrepancy persists even with an increase in model complexity (from SVM to CNN). As a result, there is a misclassification. Since it is a case of suicide risk <i>severity</i> detection, a prediction of a low severity label can impact the quality of care a patient would receive.	47
Figure 3.8	The text that contains some bracketed tokens is the transformed input text. The bracketed tokens are either similar to the concepts in the lexicon or definitions of the concept classes or present within them. Thus, we call them concept phrases. The elliptical shapes are an illustration of concept classes. Suicide Indication and Suicide Ideation are highlighted because the bracketed concept phrases are significantly similar to these classes. This transformed input text is input a model described in 3.10.	48
Figure 3.9	An example of constituency parse tree for the first sentence in figure 3.8. The image is created using Berkely Neural Constituency Parser, available online here.	49

Figure 3.10	The transformed input text is the input to the CNN model that learns by computing semantic embedding loss (\mathcal{L}_{se}). This loss is defined because the concept classes have a representation form as vectors. \vec{v}_{SIn} represents the vectorized form of the definition and concepts that describe suicide indication. Likewise, \vec{v}_{SId} , \vec{v}_{SB} , and \vec{v}_{SA} represents the vectorized form of the suicide ideation, suicide behavior, and suicide attempt, respectively. \mathcal{L}_{se} compute the Euclidean distance between the representation of the input text and vectorized form of the concept classes. The output shows that by identifying concept phrases, the model learns their combined representation, resulting in an increase in their importance scores.	52
Figure 3.11	The ROC plots show the capability of either approach in detecting users with different levels of suicide risk severity based on their behavior over time on the SW subreddit. We notice that TvarM (right) effectively detects supportive and ideation users. TinvM (left) is capable of detecting behavior and attempts users. We also record that a hybrid of TinvM and TvarM is required for detecting users with suicidal behaviors.	56
Figure 3.12	Distribution of 500 annotated users in different mental health subreddits. ADD: Addiction, DPR: Depression, SLF: Self Harm, BPD: Borderline Personality Disorder, BPL: Bipolar Disorder, SCZ: Schizophrenia, and ANX: Anxiety	59
Figure 3.13	Results showing reduction in Perceived Risk Measure through an ablation of Concept Phrase (CP), Supportive Label (SL), and Semantic Embedding Loss (SE).	63
Figure 3.14	The transient posting of potential suicidal users in other subreddits, requires careful consideration to appropriately predict their suicidality. Hence, we analyze their content by harnessing their network and bringing their content if it overlaps with other users within r/SuicideWatch (SW). We found, Stop Self Harm (SSH) > Self Harm (SLH) > Bipolar (BPR) > Borderline Personality Disorder (BPD) > Schizophrenia (SCZ) > Depression (DPR) > Addiction (ADD) > Anxiety (ANX) to be most active subreddits for suicidal users. After aggregating their content, we perform MedNorm using Lexicons to generate clinically abstracted content for effective assessment.	64

Figure 4.1	(A) & (B): An illustration of self-attention matrices computed in current attention-based transformer models and autoencoders. (C) The cross-attention matrix is what we desire and seek to achieve using autoencoders. We mainly use autoencoders as they are proven to be good representation generators and modulators. Credit: Image adapted from a Presentation	67
Figure 4.2	An overall pipeline illustrating the benefit of Semi-Deep Infusion in making ML/DL explainable and interpretable.	69
Figure 4.3	A general architecture of Semi-Deep Infusion . KS_c^t : c^{th} concept in a knowledge source (KS) that is similar to a t^{th} topic or phrase extracted from free form input text. Semi-Deep Infusion concerns with making AI model that learns a weight matrix which intersects with input observational data and expert knowledge.	77
Figure 4.4	Proposed approach to DSM-5 classification using SEDO based word-vector modulation together with Horizontal Linguistic Features (HLF), Vertical Linguistic Features (VLF) and Fine-grained features (FGF). HLF includes, <i>number of definite articles, number of words per Reddit post, first person pronouns, number of pronouns, and subordinate conjunction</i> . VLF includes, <i>number of POS tags, similarity between Reddit posts made by a user, intra-subreddit similarity, and inter-subreddit similarity</i> . FGF includes, <i>sentiment scores, emotion scores, and readability scores</i> . These Linguistic Features are specific to mental health for which SEDO was used. Details of these features are presented here [10].	78
Figure 4.5	δ controls the amount of knowledge infusion in SEDO for acceptable classification mental health disorder given a user's profile in the form of posts. Upon 34% knowledge infusion the model's recommendations matched five MHPs provided labels 84% of the times [10]. .	80
Figure 4.6	Results showing reduction in False Alarms by replacing statistical features with knowledge and its subsequent ablations of various form of knowledge. CC: Concept Classes, DSM-5: Diagnostic Statistical Manual for Mental Health Disorders, a knowledge source for mental healthcare practitioners, K_{onto} : Drug Abuse Ontology, a domain-specific ontology for substance use and addictive disorders, RF: Random Forest, CNN: Convolutional Neural Network. Model(Features or Knowledge) : It represents that either statistical features or concepts from knowledge sources are given as input to the model.	81
Figure 4.7	ISEEQ's one-shot procedural question generation	82

Figure 4.8	ISEEQ’s generation of information seeking questions reduces the number of turns involved in providing the response needed by the end-user. Thus improving user engagement.	82
Figure 4.9	Minimize Annotation Effort in Conversational Information seeking . . .	85
Figure 4.10	Overview of our approach. ISEEQ combines a BERT-based constituency parser, Semantic Query Expander (SQE), and Knowledge-aware Passage Retriever (KPR) to provide relevant context to a QG model for ISQ generations. The QG Model illustrates a structure of ISEEQ variants: ISEEQ-RL and ISEEQ-ERL. We train ISEEQ in generative-adversarial reinforcement learning setting that maximizes semantic relations and coherence while generating ISQs. (Patented with Samsung Research America)	88
Figure 4.11	Dark Blue chatbot is a pictograph of ISEEQ-RL. With KG and minimal set of passages, ISEEQ-RL generated 27% more questions that are semantically similar to ground truth compared. Without entailment constraints, questions generated by KG triples never made to top-K. Hence performance of ISEEQ-RL with or without KG triples is same.	94
Figure 4.12	Light blue chatbot is a pictograph of ISEEQ-ERL. Forcing the entailment constraints in ISEEQ-ERL yielded high scores on Semantic Relation and Logical Agreement – Conceptual Flow. This performance is seen uniformly from minimal set of 5K passages to 50K passages. Though in initial epochs, “entity only” generated question takes precedence over questions generated using triples from KG. In higher epochs, questions generated from passages retrieved from ConceptNet information took precedence. Longer training cycles were there for ISEEQ-ERL over ISEEQ-RL (2 hours long).	96
Figure 4.13	Improvement in performance of ISEEQ-ERL over ISEEQ-RL and Baseline: T5-FT CANARD concerning SR and LC in generated ISQs. This experiment was performed on CAsT-19 with <i>unannotated</i> passages.	97

Figure 5.1	An illustration of a classification task that benefits from process knowledge. Here, an AI model using a process knowledge structure would consume the user’s input, extract conceptual cues that can answer questions in process knowledge, and provide a classification label. The figure illustrates this process in assessing suicide risk severity using a partial sequence of questions from the C-SSRS. The highlighted text on the left is concept phrases that contribute to the yes/no in the C-SSRS questions.	103
Figure 5.2	Illustration of process knowledge for different purposes. (Left) A process knowledge to assess anxiety disorder in an individual using GAD-7 questionnaire. (Right) A process knowledge to assess severity of suicide risk in an individual	106
Figure 5.3	Reddit is a rich source for bringing crowd perspective in training DLMs over conversational data. On the left is a sample post from r/depression_help which sees inquisitive interaction from other Reddit users. At the top-right are the FQs asked by the Reddit users in the comments. These FQs are aimed at understanding the severity of the mental health situation of the user and are hence, diagnostically relevant. At the bottom-right are the questions generated by DLMs. It can be seen that these are not suitable FQs.	107
Figure 5.4	Inclusion in sanity checks in the neural response generation or question generation model. The sanity checks entails a check on the quality of generation by computing either jaccard index between the tokens in the generation and medical lexicons or compute cosine similarity between the generated sentence and medical questionnaire (Med_Q). \hat{Q}_{k+1} : Generated Question. $p_\theta(\hat{Q}_{k+1})$: Encoding of the generated question and $p(Med_Q)$: Encoding of a question in the list of questions in medical questionnaire. δ : a threshold, above which the generated question is accepted.	109
Figure 5.5	An overview diagram of our proposed PK-iL approach using C-SSRS 2.0 Dataset. The process structure of C-SSRS is stacked over a fine-tuned or end-to-end pre-trained LM to provide yes/no responses to questions in C-SSRS. Highlighted parts of the posts contribute to yes in C-SSRS.	113
Figure 5.6	Mean Accuracy and AUC-ROC scores, rounded up, for all LMs used in PK-iL algorithm over C-SSRS 2.0 dataset. There are two variants of PK-iL that was evaluated: (a) PK-iL with Cosine Similarity (CS) and (b) PK-iL with Gaussian Kernel (GK) for Kernel choice for each language model of representation. V: The LMs in their vanilla state.	117

Figure 5.7	A post in PRIMATE which is annotated with PHQ-9. The questions marked “YES” are answerable by DLMs using the mental health specific cues from user text. The questions marked “NO” are the questions a DLM should consider asking as FQs. Sentences within [] were taken as signals that the “YES” marked questions had already been answered in the post	125
Figure 6.1	An illustration of deep knowledge infusion. The procedure provides an improvement over existing DL architectures by including (a) layer-wise knowledge augmentation(\mathcal{K}) and (b) monitoring correct infusion through knowledge attention matrix. The later component controls the information flow between the previous layer ($x_{layer_{prev}}$) and the next layer ($x_{layer_{next}}$).	130
Figure 6.2	Overall Architecture: Contextual representations of data are generated, and domain knowledge amplifies the significance of specific important concepts that are missed in the learning model. Classification error determines the need for updating a Seeded SubKG with more relevant knowledge, resulting in a Seeded SubKG that is more refined and informative to our model.	131
Figure 6.3	Inner Mechanism of the Knowledge Infusion Layer in an LSTM Network	134

CHAPTER 1

INTRODUCTION

For many people, the purpose of Artificial Intelligence (AI) has been to achieve human-level intelligence. In that direction, recent years have seen data-driven machine learning (ML) or deep learning (DL) models, specifically neural networks, acquiring remarkable success in many tasks such as object detection in images and speech recognition. On the other hand, these approaches prove limited in their ability to perform the tasks with generality, adaptability, and explainability toward accomplishing “machine intelligence”. As the dependence on large labeled datasets for learning continues to be critical, the challenge is to obtain adequate and high-quality labeled data or provide some means to overcome that gap. Moreover, such a dataset may not cover all possibilities concerning the task in question, including those likely to arise in the future. For example, in natural language understanding (NLU), algorithms have not yet progressed adequately to capture the implicit contextual meaning of the content. One approach to address such limitations and make intrinsically more intelligent systems is to combine the bottom-up data-dependent processing with a top-down goal-driven or plan-based processing, as observed by cognitive scientists (e.g. symbolic reasoning, expert knowledge) and to a lesser extent by computer scientists [13] [14]. Figure 1.1 provides an illustrative comparison between blackbox (left) and greybox (right). The blending of ML/DL with structured knowledge (e.g., knowledge graphs (KG)) to achieve greybox AI is what we call “**Knowledge-Infused Learning**” (KiL) [15] [16], is an approach to address significant limitations of statistical methods in AI: (a) **Context Sensitivity:** A statistical AI model is opinionated based on the input it sees which is a partial representation of the world. (b) **Uncertainty and Risk:** There exists a knowl-

edge gap between what annotators use to create and curate datasets and the models that consume labeled datasets. (c) **Model Interpretability:** An interpretation of the internal mechanics of AI models is sought to study anomalies learned while training over diverse datasets. (d) **User level Explainability:** A human can endorse a statistical AI model if the outcome is traceable to some external source of information with which a human can relate the findings deduced from the model. Synonymously, it is considered post-hoc explainability, which is still a manual effort from end-users, whereas KiL seeks its automation using KG. (e) **Task Transferability:** Any AI model should learn to perform a task and utilize the known to perform acceptably well on similar tasks. KiL sees an opportunity to exploit the semantic similarity across tasks using external sources of information (e.g., KG, lexicons, documents, etc.).

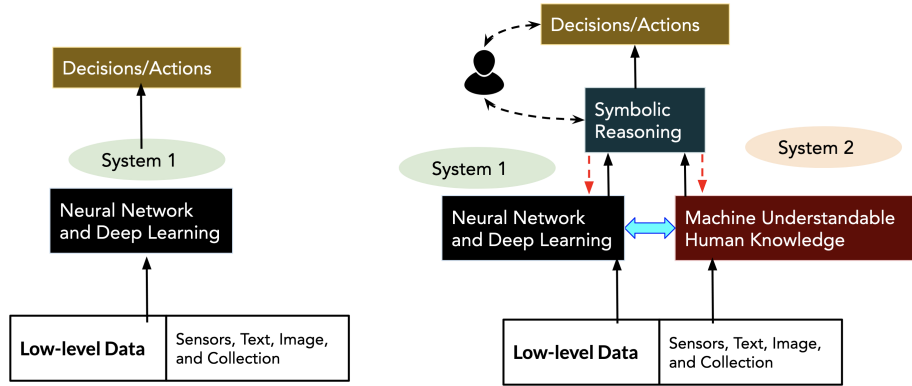


Figure 1.1 **(left)** Neural Networks and Deep Learning Models (e.g classification/discriminative models, generative models) modeled as black box because the decision/actions do not support user-level explainability. Further, there is no other means to interpret the internal mechanics of black box other than studying the low-level data. This black box is considered as system 1. **(right)** System 2 represents symbolic knowledge represented in the form of a knowledge graph (KG; a graph), semantic lexicons (a dictionary), rules (a set of constraints) that can be made in machine understandable form and infused in System 1 for model interpretations. To study the outcome of System 1, then System 2 supports user-level explainability.

ML/DL models have shown significant advances in improving natural language processing (NLP) by probabilistically learning latent patterns in the data using topic models (e.g. LDA), decision trees [17], Bayesian networks [18], Markovian processes [19], and

multi-layered network of computational nodes arranged sequentially, or parallelly with attention and positional encodings [20]. These methods consume a tremendous amount of training data with increased scaling in parameters for better performance. Further, researchers have started to recognize the lack of uncertainty handling in these models when used in closed-domain tasks (or domain-specific tasks). In addition, the desire to increase the model complexity by an order of 10 for $1/10^{th}$ of the improvement in safety, explainability, and model interpretability [21] raises further concerns in their utility. The use of relevant background knowledge will aid ML/DL models in coping with uncertainty and reduce model complexity. The background knowledge contains the representations of real-world that would facilitate explainability. These features would contain explicit representation of entities, their synonyms, and a variety of typed relationships. Further, if these features allow any ML/DL models to learn input and output mapping and tune its attention, then we would attain a matrix that should contribute to model interpretability. Learning over conceptual features will also benefit ML/DL model to showcase handling of uncertainty and risk by introducing expert-defined rules as constraints. Pioneers in AI are manipulating the structured KGs for ML/DL machine with relational inductive biases¹, transfer learning (cross-domain knowledge sharing), and other new methods of infusing KG into ML/DL (see Table 1.1). Let us explore how past research in ML/DL, its associated challenges, and impact areas inform the need to infuse KG into ML/DL.

1.0.1 PAST, CHALLENGES, AND IMPACT AREAS

Hand-crafted methods in AI, such as the UCB Hearst Pattern [22] and the NYU Proteus (1997), were effective in learning underlying patterns. With supervised learning methods, AI drifted towards large-scale data acquisition and annotations, ignoring the explicit knowledge implicitly embedded in data preparation. Weakly and distantly supervised learning methods showed the importance of including explicit knowledge, but still relying

¹zd.net/2Jb1g2A

Table 1.1 Reflection of pioneer in AI towards knowledge infusion and Neuro-symbolic AI in general.

Knowledge-infused Learning is a class of Neuro-Symbolic AI techniques that incorporate broader forms of knowledge (lexical, domain-specific, common-sense, and constraint-based) into addressing limitations of either symbolic or statistical AI approaches, such as model interpretations and user-level explanations. Compared to powerful statistical AI that exploits data, KiL benefits from data as well as knowledge.

Leslie Valiant's vision: "The aim here is to identify a way of looking at and manipulating broader and other richer forms of knowledge that is consistent with and can support what we consider to be the two most fundamental aspects of intelligent cognitive behavior: the ability to learn from experience, and the ability to reason from what has been learned. We are therefore seeking a semantics of knowledge that can computationally support the basic phenomena of intelligent behavior." Further, Knowledge-infused Learning aims to incorporate broader forms of knowledge that are linguistic, lexical, domain-specific, rule-based, and word sense-based

Douglas Hoftstader mentioned the reasons for the question "why AI is far from being intelligent?" by pressing on human thinking being artistic and beautiful. Later you will see Knowledge-infused learning brings the concept of entity normalization, semantic query expansion, zero shot learning, and contextual bandits using knowledge graphs to generate probable outcomes that intuitively lies in human understanding of the problem.

Gary Marcus reflected on the lack of common-sense reasoning in deep language models, much like Leslie Valiant. Considering Marcus's example: "What happens when you stack kindling and logs in a fireplace and then drop some matches is that you typically start a ____", a KiL approach will look for a connection between "kindling", "log", "fireplace", "matches", giving "fire" as the concept with most closely related representations.

on knowledge's statistical representation. Efforts such as recommender systems, learning to rank, summarization, and conversational artificial intelligence, revealed drawbacks of existing statistical AI methods in achieving acceptability and adoption by communities of experts [23–25]. For instance, Han et al. enumerated the reasons for under-performance of content+collaborative recommender systems in healthcare, mostly directed to the ignorance of ground-truth guidelines (or rules), such as International Classification of Diseases 10th Edition (ICD-10) and Unified Medical Language System (UMLS) hierarchy, and conceptual features explaining one's health conditions [26].

ML/DL algorithms excel in automatically but opaquely uncovering semantic equivalence and subsumption relationships based on the similarity of usage contexts detected and

encoded during training. This does provide a limited means to impose hierarchical organization (e.g., IS-A, HAS-A). However, this is not robust with respect to reliably uncovering synonymy, polysemy, part-of/part-whole/has-a, and other labeled relationships in general, which are required in upcoming challenging tasks in natural language processing (e.g. KILU [27], KILT [28], GEM [29], etc.). These tasks require ML/DL to be supported with curated resources such as WORDNET for formal arbitration of linguistic knowledge or UMLS for biomedical knowledge [30] [9] [31].

In **learning to rank**, state-of-the-art statistical AI methods heavily rely on the co-occurrences of pairs of words. As a consequence, it is difficult to rank documents/contents when the query is about an emerging topic with minimal co-occurrences (e.g. long-tail entities [27] [9]). Moreover any statistical AI algorithm functions on latent dimensions making the ranking of documents/contents hard to explain. Algorithms in **summarization** have a hard time in modeling knowledge constraints causing the end result to significantly differ from useful and actionable summaries [23]. In **conversational artificial intelligence**, an intrinsic task of any ML/DL algorithm is to understand user behavior during an interactive search and later improve accuracy during search sessions. Research on conversational artificial intelligence has been hampered by a lack of datasets that involve process knowledge, an approach experts follow during any formal conversational setting. We noticed that while deep learning enables one to perform empirically defined tasks well in the aggregate, it is not conducive to accomplishing general or human-like intelligence tasks accurately. It is because we are unable to scrutinize and programmatically exploit the learned representations to provide formal guarantees in the conclusions derived. For instance, algorithms trained on some standard benchmark datasets, such as General Language Understanding and Evaluation (GLUE) become rigid and lack reusability across other domains [32] [33]. Furthermore, using neural networks and deep learning include the difficulty in characterizing hidden biases and quality issues in data, making it vulnerable to spurious correlations. This lack of representative and unbiased data can hamper the adoption of deep learning al-

gorithms in critical applications that require guarantees, transparency, and accountability. This black-box nature of neural networks can hamper its broader adoption in some critical application domains where a human-in-the-loop is necessary to rationalize actionable decisions to inspire confidence.

To the extent that knowledge-based systems can declaratively specify and exploit "causative" features characterizing classes, in preference to "correlated" features implicitly learned from the training samples, their integration to get the best of both worlds will create a powerful and reliable system. An attractive option is to use a two-stage representation and reasoning system that uses neural networks and deep learning algorithms for low-level perceptual tasks while using a knowledge-based system built on top for high-level reasoning and decision making [13].

By developing the KiL paradigm, the dissertation answers the question:
Can the incorporation of various forms of domain knowledge enhance the performance and explainability of data-intensive learning models? And how effectively can we do it?

1.0.2 IMPACT AREAS AFTER INFUSION OF KG INTO ML/DL

KiL is a foundational technology for the third wave of AI. It seeks to achieve following goals in open domain and domain-specific natural language processing tasks:

Context Sensitive Capture: Current data-driven ML/DL algorithms are opinionated based on the input they see. The input is a partial representation of the world. For instance, consider a pair of sentences from the Quora Question Pairs (QQP) dataset, where the task of an ML/DL model is to predict whether sentences are similar or different; (*Sentence A*): What would have happened if Facebook was present in World War I? and (*Sentence B*): What would have happened if Facebook was present in World War II? A state-of-the-art transformer model yielded “similar” as the predicted outcome for an actual outcome “different.” Essentially, the model focused its attention on following words: what, happened, facebook, world, and war and gave low attention

to “I” and “II. Whereas the KiL-based BERT discussed in Faldu et al. yielded the correct outcome [9]. This is because it used a data augmentation scheme, wherein the input representations of *Sentence A* and *Sentence B* was altered using triples from **KG**: (*Sentence A*): World War I < fought_with > Trenches; World War I < fought_with > Posionous Gas; World War I < fought_with > Guns and (*Sentence B*): World War II < fought_with > Ships; World War II < fought_with > Fighter Planes; World War II < fought_with > Tanks. Further, scaling across all the samples in the QQP dataset, KiL-based BERT yielded 3% improvement over simply BERT-based model.

Uncertainty and Risk: Current data-driven ML/DL algorithms fail to establish the connection between input data and outcome, resulting in black box approximation. This is because it is hard to answer these questions: “How do you know that a training set has a good domain coverage?”, “How many samples are needed to achieve desired confidence for end-user?” The paradigm on Probably Approximately Correct (PAC) learning establishes a theory to address these questions and achieve robust and consistent classification [34] [35]. PAC learning can be formalized as:

$$Prob(Test_{err} > \epsilon \mid Train_{err} \approx 0) < |H|e^{-\epsilon m}; |H|e^{-\epsilon m} < \delta$$

where $|H|$ represents all possible hypotheses for classification, ϵ is the minimal misclassification error, and δ is an empirical threshold (e.g. human annotation error). This formulation can be interpreted as: To achieve a testing error in an acceptable range ϵ , the number of training samples (m) needed are exponential and require an ensemble of ML/DL algorithms ($|H|$). Keeping ϵ and δ constant, KiL aims to reduce the size of $m := m/|KG|$; and $|H| := (|H|)^{1/|KG|}$ and enable the model benefit from human knowledge [36].

Another utility of KiL is in the domain of conversational systems, which tend to hallucinate while generating a response or a question. Separate from the classification context, this domain sees the use of KiL in making conversational systems

safe. Consider a simple transformer model (e.g. BERT) used to conduct a discourse with a user in the domain of mental healthcare. The BERT-based agent asks the question² "Do you feel nervous?" and user answers "More than half the day." The agent then asks following two subsequent questions: (a) "Do you feel irritated or self destructive?" and (b) "Do you feel something extreme might happen to you?" These generated questions are not only irrelevant but also risky. A mental healthcare provider would not ask such question. Under the hood of the BERT-based model lies statistical methods that exploit co-occurrence: "nervousness co-occur with irritation and irritation co-occur with self-destructive", which yields utterances that might have severe consequences. In KiL-based agent, the generation is checked for safety by using semantic lexicons and questionnaire often used mental healthcare [17]. Further a risk evaluation metric for ML/DL algorithms would bolster the confidence in the agent before it can be deployed [37].

Interpretability: We need to be able to extract information about the inner functioning of a model. It helps to make users aware of the model’s decision making capabilities, and potentially avoid any ambiguity. Users can compare and verify the relevance of information leveraged by the model or can help determine what information should be extracted [9]. In KiL-based algorithms, the interpretability is achieved by conditioning either the loss function or attention function on the concepts in KGs.

User-Level Explainability: It is defined as post-hoc explainability using general-purpose or domain-specific KGs (or lexicons, any structured knowledge sources comprehensible to experts). Data-driven ML/DL algorithms derive system-oriented explanations and are not rich enough for user-level understanding. Figure 1.2 shows user-level explainability by mapping the attention-defined features from a trans-

²assume that a casual starting conversation has already begun

former model onto a medical KG in SNOMED-CT (Systematized Nomenclature of Medicine- Clinical Terms).

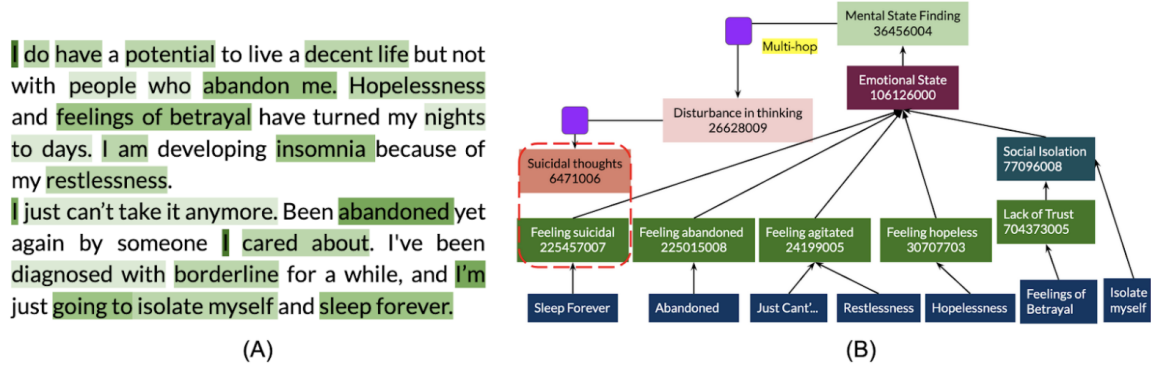


Figure 1.2 An illustration of user-level explainability using the important conceptual phrases identified by a deep learning model trained using the method described in the Presentation. Highlighted phrases in (A) are queried in SNOMED-CT, thus forming a contextual tree. Formation of this tree is stopped when a node is hit that has high similarity to either leaf nodes or one hop parent nodes. The resulting tree is shown in (B). The numbers in the boxes are SNOMED-CT IDs.

Task Transferability: It is considered an ML/DL algorithm's ability to efficiently learn patterns from a task so that it can be utilized in another similar or same task. Analogously, it is termed generalizability in AI. There have been methods in zero-shot learning to achieve task transferability, but their opaqueness has always been a concern [38]. Further, statistical AI in general learns efficiently on the data and not the task. This dissertation will discuss KiL-based algorithms that can be transferred across various-sized tasks.

1.0.3 KNOWLEDGE GRAPHS AND THEIR ESSENTIAL ROLE IN KiL

These goals are laid out after realizing the central problem in the current data-driven AI model: the mismatch between the input given to the model and the output expected from the model. Early attempts at using external knowledge in machine learning to address these challenges are long from achieving their true potential. It is because choosing an appropriate source of superficial knowledge and defining a method for its inclusion depends

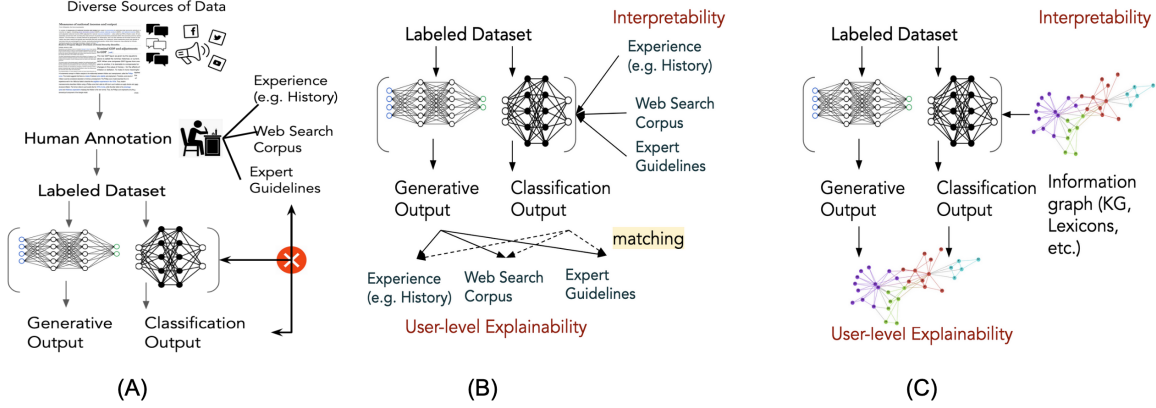


Figure 1.3 An illustration of the process followed by human annotators(or subject matter experts) in creating datasets of high quality. (A) It shows that annotators (or subject matter experts) used the external pieces of information but were not provided to ML/DL models. (B) Suppose these external sources of knowledge for annotators or experts are infused into ML/DL models and used to evaluate model outcomes. In that case, we achieve model interpretability and user-level explainability. (C) Since these diverse sources of knowledge are voluminous, their infusion can be a resource and computation-heavy. We believe information graphs can replace these sources, and being machine-readable, their infusion would not increase computation costs drastically.

on the process of creating the dataset. A wrong choice might affect the ML/DL algorithm leading to over-generalized, factually incorrect, or inaccurate predictions. Creating the dataset can be characterized as a procedure for annotating the dataset to achieve the gold or silver standard. For instance, making the task of human annotation is one such process (see Figure 1.3).

The labeled datasets that serve as the training resource for various ML/DL algorithms are created by human annotators (e.g., Amazon Mechanical Turks, ParIAI, CrowdTruth, or Subject Matter Experts) after doing an extensive set of exercises: (1) Going over the guidelines for annotations, (2) Exploring resources on the web, using their experience or hunches, or leverage expert guidelines for more information, and (3) Evaluate the annotation for quality check and assurance. The external information used for annotators is not available for ML/DL models to train [39].

Alternatively, one can utilize weakly supervised learning (e.g. SNORKEL [40]), knowledge distillation [41], label propagation [42] and others to compensate for process knowl-

edge. But it would not alleviate the statistical bottlenecks mentioned above. Majumder et al. explored the similarity between the ROC Stories and the dialogues in PersonaChat [43] [44]. Hence, to personalize natural language generation by a conversational agent, sentences from ROC Stories were given as context to the DL model inside the agent.

Understanding the processes behind creating a dataset and leveraging it to choose the proper external knowledge can create ML/DL models that are user-level explainable. Analogously, the **processes** are prior knowledge, and its inclusion into ML/DL models is what we term “Process Knowledge Infusion”. Process knowledge manifests the human decision-making process and thus needs to be integrated. It can help answer a user-level explainable question; “How does the data apply to an ML/DL algorithm yield this kind of results?”. Further, a process knowledge infusion can help achieve **Safety**, criteria well discussed in Google’s recent effort to develop safe chatbots by using their long list of safety guidelines [21].

KiL sees the incorporation of these information sources into the ML/DL models as *interpretability* and confirming the model’s output against these sources as *user-level explainability*. In this dissertation, along with KiL, we will look specifically into mental health process knowledge in clinical questionnaire and design computational methods under KiL for safe and explainable chatbots in mental healthcare (more details in Chapter 5). Process knowledge sees wider applicability in autonomous driving, precision nutrition, and improving sales engagement platforms for enhancing the productivity of sales representatives. This dissertation will focus on diverse forms of knowledge obtained from either general-purpose (e.g. Wikipedia, ConceptNet [45], WordNet [46]) or domain-specific (e.g. SNOMED-CT [47], ICD-10 [26], UMLS [48] [49]) KGs.

A Knowledge Graph (KG) is a machine readable structured representation of knowledge consisting of entities (entity and entity type) and relationships in various forms (e.g., labeled property graphs and RDFs) [50]. KiL-based ML/DL seamlessly integrates external knowledge to address challenging problems in open-domain and domain-specific low

resource natural language processing problems. Domain-specific problems are defined by their need to apply task-specific knowledge (implicit/explicit) to generic AI models. For instance, to detect emerging events in a stream of crisis-related tweets (e.g. Hurricane, COVID-19 Pandemic), a generic language model (e.g. Word2Vec [51], BERT [52]) can be fine-tuned using the concepts and relationship in disaster ontology (e.g. empathi [3]). Low resource problems have few labeled samples and further labeling is difficult in terms of effort, quality, and time. Consider a case of annotating millions of posts from users in various mental health communities on Reddit that would require (a) establishing guidelines for annotation, (b) training of annotators, (c) resolving conflicts in annotation, and (d) enriching quality over multiple iterations to achieve high annotator agreement. A study by Gaur et al. defined a KiL pipeline to annotate such big social data at scale and moved humans from the role of annotators to evaluators [31].

Figure 1.4 Example KG constructed either from manual effort (A, B, C), automatically (D, E), or semi-automatically (F). (A) is empathi ontology designed to identify concepts in disaster scenarios [3]. (B) Chem2Bio2RDF [4]. (C) ATOMIC [5]. (D) Education Knowledge Graph by Embibe [6]. (E) Event Cascade Graph in WildFire [7]. (F) Opioid Drug Knowledge Graph [8]

There are various forms of KG that are constructed either through manual effort, automatically, or semi-automatically, as illustrated in Figure 1.4. KG constructed with manual effort and following expert-defined guidelines are called Ontologies. For instance, in Figure 1.4 (A), The empathi ontology³ is constructed from archives of the Federal Emergency Management Agency (FEMA), disaster ontology [53], geonames , and others. The structure of the ontology is laid out based on the process in which an event is described in FEMA archives. Figure 1.4 (D) illustrates an Educational Knowledge graph⁴ constructed from epub's of Amazon books and other course textbooks to assess a student's learning outcome and suggest ways to intervene. These domain-specific KGs are at the core of KiL to provide necessary information aid for machine learning/deep learning algorithms for domain adaptation and reasoning over the outcomes. There are various ways to incorporate external knowledge that which my dissertation categorizes into (a) Shallow KiL, (b) Semi-Deep KiL, and (c) Deep KiL. Shallow KiL contextualizes the training examples with expert knowledge to capture meaningful patterns. Some of the shallow infusion examples include contextual modeling [54], entity normalization [37] and explainable clustering [55]. Semi-deepKiL guides the model's attention in the learning process. It utilizes expert knowledge concepts as weights or constraints to guide an explainable learning process. This strategy falls short in assisting deep learning models to adjust the high-level abstractions learnt through multiple layers. Deep KiL, combines the stratified representation of knowledge at varying abstraction levels to be transferred in different layers of deep learning models [15].

1.0.4 WHO SHOULD READ THE DISSERTATION?

This dissertation is easily accessible to readers with a computer science background, specifically in artificial intelligence, data mining, natural language processing, and information retrieval. A preliminary understanding of linear algebra, probability, and statistics would

³<https://shekarpour.github.io/empathi.io/>

⁴<https://www.embibe.com/ai-in-education/articles>

benefit a reader in appreciating the results discussed in the dissertation. This dissertation is designed to serve as a primer on Knowledge-infused Learning (KiL) which is analogous to Neuro-symbolic AI⁵. The target audience for this dissertation is students and faculty in computer science and interdisciplinary centers on data science. The theory, explanations, experiment design, and evaluation strategies discussed in the dissertation would benefit students and faculties in psychology, social science, linguistics, information systems, mathematics, and computing; the KG is a structural resource of expert knowledge which comes from research in non-computer science disciplines.

This dissertation can independently serve as a seminar course on Knowledge-infused Learning, part of Trusted AI, Data Science for Social Good or AI for Social Good. Industry researchers and Practitioners who are interested in exploring knowledge graphs and ways to infuse it in artificial intelligence can look and borrow lessons from tangible use-cases, theory, related research, and evaluation strategies to address issues in their respective fields. The reader may consider this dissertation as a tutorial that provides a detailed walk-through on KG and its utility in developing knowledge-infusion techniques for interpretable and explainable learning from text, video, images, and graphical data on the web. The dissertation will motivate the novel paradigm of Knowledge-infused Learning using computation and cognitive theories. It describes different forms of knowledge, methods for automatic construction and modeling of KG, and its infusion in current methods and state-of-the-art techniques in machine or deep learning. Further, it discusses application-specific evaluation methods for explainability and reasoning using benchmark datasets, real-world datasets, and knowledge resources that show promise in advancing the capabilities of AI. In the future directions, the dissertation provides sufficient grounding on KG and robust learning for the Web and Society.

⁵https://www.nsf.gov/awardsearch/showAward?AWD_ID=2133842&HistoricalAwards=false

1.0.5 DISSERTATION CONTRIBUTIONS AND SCOPE

This dissertation illustrates contribution in creating natural datasets that challenges data-driven ML/DL algorithms for achieving user-level explainability and interpretability. In process, the dissertation develop novel computational methods: (a) loss functions, (b) optimization functions, (c) retrieval and ranking methods, (d) entity normalization technique, (e) new domain-specific ontologies and KGs, and (f) evaluation metrics that examine algorithm’s capability to be user-level explainable, interpretable, handle uncertainty and risk, and context sensitivity. Table 1.2 provide a short summary of the contributions that will be explained later in the dissertation. The data scope of the dissertation is as follows:

- **Reddit Dataset:** In the light of recent pandemic, a nascent community of Reddit, also called subreddit had < 2000 subscribers as reported in December 2019. Starting January 2020, the community showed a startling rise reaching to > 2 *Million* subscribers in less than two months. This shows the timeliness, engagement and outreach that subreddits on Reddit provides to people across the globe. From the perspective of social good and social impact, the content on such communities⁷ can improve development of data-driven AI for proactive decision making. Studies described in this dissertation are structured on Reddit. It explored 15 prominent mental health subreddits, identifying and exploring the support seeking and support providing roles that users take in such communities, understanding how informative the conversations between people in a community are, capturing the movement of the users between communities, how to develop clinical context from noisy conversations, exploring and exploiting medical knowledge graphs and databases, and how to effectively utilize clinical questionnaires in social media for reliable decision making. The ground-truth Reddit datasets developed and used in this dissertation can be obtained from following links:

⁷Talklife, Talkcampus, Twitter, Reachout, etc.

- All Mental Health Communities: It covers 13 *Million* posts and > 52 *Million* comments from 2005 to 2018.
 - Suicide Risk Dataset: It covers mental health professionals’ annotated posts of 500 users who were judged as people with suicidal tendencies based on their posts on r/SuicideWatch. The annotation was performed using C-SSRS, a Columbia Suicide Severity Rating Scale, making it a special and unique dataset for explainable ML/DL algorithms to detect suicide risk severity.
 - Time-Variant Suicide Risk Dataset: This is another format of C-SSRS-based suicide risk dataset where users’ posts are ordered and annotated by time.
- Twitter Dataset: Along with Reddit, Twitter is another social media platform known for event-specific tweets. Crisis events are first to be reported on Twitter; thus, crawling Twitter data for deriving insights and alerting emergency responders is another impactful application of AI. A study with Dataminr Inc. demonstrates the applicability of domain-specific knowledge in unsupervised event detection in > 80 *Million* tweets [24].
 - Clinical Diagnostic Interviews: Thiruvalluru et al. identified that there is a discrepancy between what patient report to clinicians and what patient post on social media [55]. Hence a study on both the communication platforms is required. We utilized 60 minute long clinical interviews of 180 patients created through a Wizard-of-Oz procedure to extract PHQ-9-related cues for summarizing interviews⁸.
 - Heterogeneous Dialog Datasets: The aforementioned datasets were used to design KiL-based algorithms for summarization, recommendations, and learning to rank. These dialog datasets were used to broaden the scope on KiL in conversational AI. These dataset span various domains where explainability of conversational AI is important. These are politics and policies, travel, news, mental health, and geography.

⁸<https://dcapswoz.ict.usc.edu/>

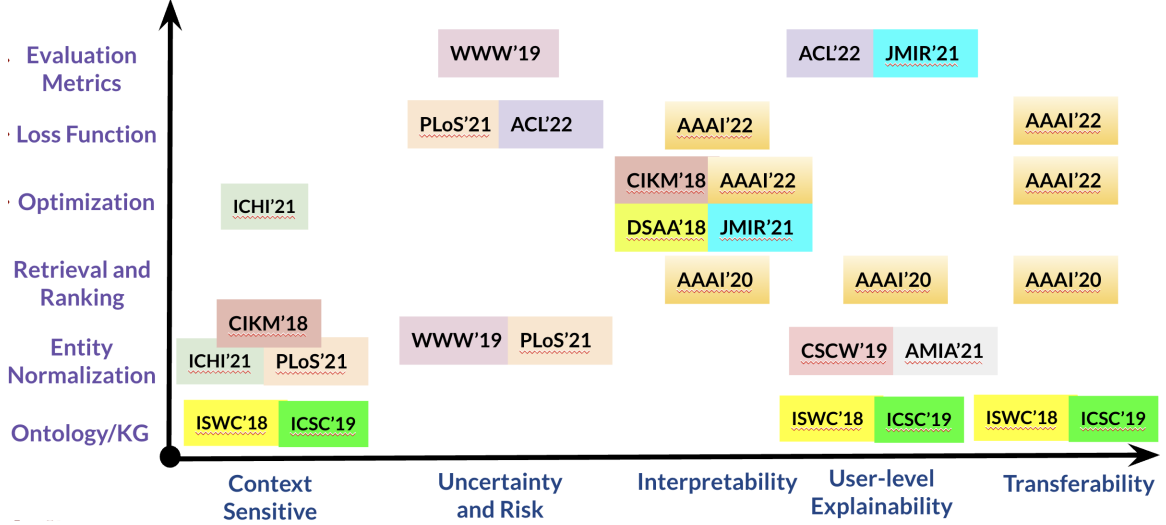


Figure 1.5 Technical Contributions (Y) made by KiL to address limitations of current data driven ML/DL algorithms (X).

Figure 1.5 illustrates the core technical contributions made by KiL in advancing AI, out of which this dissertation will focus on two specific areas:

- **Classification Tasks:** It is defined as a class of problems wherein machine learning or deep learning models are tasked with assigning class labels to unlabeled inputs in the domain [56]. This dissertation will focus on multi-label and multi-class classification in problems specific to recommender systems.
- **Generative Tasks:** It is defined as a problem setting in machine learning or deep learning where the model is tasked with text generation based on specific constraints, preventing irrelevant or incoherent generation. This dissertation will focus on neural text generation for question generation.

This work will primarily look into methods and develop metrics that exercise knowledge infusion, providing user-level explanations and making models interpretable to end-user.

Table 1.2 Summary of Contributions made by KiL. Each of the task and application comprises of either dataset construction, knowledge source construction or adapting existing datasets for the task/application. Further, outcomes from KiL-based algorithms on different testsets is evaluated by subject matter experts by conducting blind human evaluation.

Outcome	Task	Application; Knowledge Source	Achievement; Papers
<i>Outcome:</i> Theoretical framework of Shallow, Semi-Deep, and Deep Infusion	Methods of Knowledge Infusion and Knowledge-intensive language understanding	Social Media; ConceptNet, BabelNet, WordNet, Semantic Lexicons	Theoretical and Conceptual enhancement to state-of-the-art machine learning (incl. deep learning) models; 4 Papers in IEEE Internet Computing
<i>Novel Method Outcome:</i> Semantic Encoding and Decoding Optimization (SEDO) and Knowledge-infused Siamese Network. <i>Novel Metric:</i> Perceived Risk Measure	(a) Social Media Informed identification of Diagnosis Disorder, (b) Severity of Risk Levels, (c) Measuring uncertainty in recommendation, and (d) Recommending support providers on Social Media for Support Seeker.	Explainable AI in Mental healthcare; Wide variety of socio-clinical knowledge sources: DSM-5, Drug Abuse Ontology, SNOMED-CT, ICD-10, PHQ-9, and C-SSRS	(a) Semi-Deep KiL showed 92% gains over state-of-the-art in mental health disorder classification; CIKM . (b) Shallow KiL showed 12.5% reduction in model's uncertainty and risk; WWW . (c) KiL explains why in mental healthcare assessment both time-variant and time-invariant models are required; PLoS One . (d) Making ML/DL model risk-averse resulting in 83% certain and safe predictions; ACL . (e) 8 out of 10 Domain Experts picked KiL's recommended support providers for support seekers over GPT-2; ICHI
<i>Novel Method Outcome:</i> Process KiL	Assessment of Suicide Risk, Severity of Depression, and Severity of Anxiety using C-SSRS, PHQ-9, and GAD-7 respectively	User-level Explainable and Interpretable Mental Healthcare	(a) 7 out of 10 domain experts picked our recommendations over XLNET. (b) Simple language models outperform large-scale deep language models; ACL
<i>Novel Method Outcome:</i> Algorithms that utilize knowledge as constraints to mandate conceptual flow, informativeness, and linguistic quality in conversation systems. <i>Novel Metrics:</i> are introduced to assess logical agreement, semantic relations, and legibility of generated questions and summaries.	Curiosity-aware conversational information seeking and clinical interview summarization	Explainable and Engaging Conversational Systems for general knowledge queries and mental healthcare. Learning to Rank	(a) Conceptual-flow based question generation using semi-deep infused knowledge outperformed Google T5 and human created questions; AAAI . (b) Semi-Deep KiL for Abstractive Summarization of long clinical interviews outperformed state-of-the-arts in interview summarization by 49% on Rouge-L and 61% on Rouge-2. On quality of summaries, 23.3%, 4.4%, 2.5%, and 2.2% gains in thematic overlap, Flesch Reading Ease, contextual similarity, and information entropy compared to state-of-the-art; JMIR .
<i>Novel method</i> to retrieve and rank using ontology	Sub-Event Detection	Information Retrieval and Ranking	Unsupervised Retrieval of Tweets with Emerging Sub-event using Crisis Ontology outperformed state-of-the-art by 89%. Shallow knowledge infusion explained retrieval and ranking of tweets reducing annotation effort; AAAI
<i>Future Directions</i>	Explainable solving of Math Word Problem and paraphrasing for information disguise	Education and Digital Security	Investigatory study with propositions for Tractable solutions for solving math word problems ⁶ . Qualitative study and empirical evidence on effective, ethical, and explainable methods of paraphrasing for information disguise; First Monday .

CHAPTER 2

FROM BLACK BOX TO GREY BOX: IMPROVING INTERPRETABILITY AND EXPLAINABILITY OF DEEP LEARNING SYSTEMS

Artificial Intelligence (AI)¹ has demonstrated rapid growth in various classification and generation tasks of varied complexity. Sometimes even surpassing human-level performance on narrowly and well-defined data analysis tasks in domains such as healthcare (e.g., radiology image inspection), education (e.g., estimating user’s concept mastery), and social good (e.g., patient and primary-care matching based on trust and ICD-10 information). However, Deep Learning models within AI are complex and opaque. The cascading sequences of linear and nonlinear mathematical transformations learned by models comprising millions of parameters are beyond human comprehension and reasoning. This renders them as “black-box” models for decision-making. DL’s black-box nature and over-reliance on massive amounts of labeled data condensed into labels and dense latent representations pose challenges for user-level explainability and model’s interpretability. DL models seldom capture context defined by the end-user, resulting in an approximate response that can be inferred true with justification provided by humans and not the model. For example, consider this trivial case of question answering where DL’s capability to capture token’s co-occurrence patterns is acceptable:

¹A term that covers all machine learning and deep learning algorithms

- **Context:** I sometimes wonder how many alcoholics are relapsing under the lockdowns.
- **Question:** Does the person have an addiction?
- **DL's Response:** Yes

Comparing this to a nontrivial case of question answering:

- **Context:** Then others insisted that what I have is depression even though *manic episodes aren't characteristic to depression*. I dread having to retread all this again because the clinic where I get my mental health addressed is closing down due to lost business caused by the pandemic.
- **Question:** Does the person suffer from depression?
- **DL's Response:** Yes (**the correct answer is No**)

This illustrates that DL model can approximate well in gathering the context if the entities are manifested explicitly. However, when the context is either implicit or convoluted with negation-type attributes, DL hallucinates in providing a response. In such context-sensitive scenarios, the need to interpret internal mechanics of the model and matching its learning representation with concepts in the real world for user-level explanations becomes pivotal [57]. This need may be addressed by infusing domain knowledge, yielding both better performance and explainability. This chapter motivates and demonstrates how knowledge, provided as a knowledge graph, is incorporated into DL using Knowledge-infused Learning (KiL). Through examples from natural language processing applications in healthcare and education, we discuss the utility of KiL towards interpretability and explainability. This chapter, through the examples, will introduce Knowledge-Intensive Language Understanding (KILU), a novel set of real-world datasets that necessitates the use of KiL.


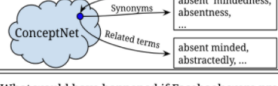


Significance

Incorporating knowledge through Knowledge-infused Learning addresses key limitations of statistical AI: (a) reliance on data alone, (b) user-level explainability and (c) interpretability.

2.1 CONTEXT SENSITIVE CAPTURE

The main idea behind the development of CYC Ontology, Freebase, DBPedia, and other knowledge resources was the realization that systems cannot be truly intelligent if they do not understand the underlying concepts and links to the semantics that they are recognizing to yield a classification or generation. Making machines context sensitive is like giving them the power to make connections between facts and observation to enhance the learning process. Recent research in the NLP community has begun to inspect the generalizability of the DL models by using them to perform simple tasks. Ribeiro et al. and Bowman concluded in making a statement that DL needs support from external support to contextualize over input data [58] [59]. I want to discuss two prominent cases where DL is certain to miss context.

Conceptual v/s Ambiguous Entities: In our previous example from the QQP dataset, where we asked a DL model to classify whether two sentences: “What would have happened if Facebook was present in *World War I*?” and “What would have happened if Facebook was present in *World War II*?” and the model’s classification was incorrect. It was because DL model tokenized *conceptual entities*: World War I and World War II, giving attention to phrase “World War” and ignored “I” and “II.” Then the question arises: “how to bracket conceptual entities so that DL model generates their representation together rather token-wise.” One way is to leverage a Concept-Net knowledge graph to create a knowledge context surrounding these conceptual

Task: Duplicate detection in Quora Question Pairs	Ground Truth	KI-BERT Prediction	BERT Prediction
<p>What are great examples of absent mindedness?</p>  <p>What are some of the great examples of absence of mind?</p> 	Duplicate	Duplicate	Non Duplicate
<p>What would have happened if Facebook were present at the time of World War I?</p>  <p>What would have happened if Facebook were present at the time of World War II?</p> 	Non Duplicate	Non Duplicate	Duplicate

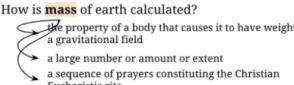
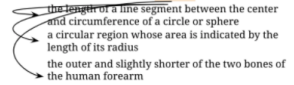
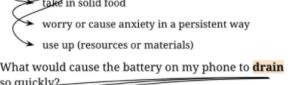
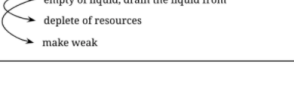
Task: Duplicate detection in Quora Question Pairs	Ground Truth	KI-BERT Prediction	BERT Prediction
<p>How is mass of earth calculated?</p>  <p>How can we calculate the radius of the earth?</p> 	Non Duplicate	Non Duplicate	Duplicate
<p>What does eat the phone battery quickly?</p>  <p>What would cause the battery on my phone to drain so quickly?</p> 	Duplicate	Duplicate	Non Duplicate

Figure 2.1 Infuse knowledge context to capture conceptual (left) and ambiguous entities (right) for correct classification in QQP dataset. Picture credit to [9].

entities independently and generate better representation by concatenating representations of neighboring contexts.

Ambiguous entities (e.g. polysemic words) are shown in the following two sentences: "What *eats* the phone battery quickly?" and "What would cause the battery on my phone to *drain* so quickly?" In these two sentences, "eat" and "drain" are polysemic words as they carry similar word senses in these two sentences. KGs like BabelNet² or WordNet³ can provide senses for these words, along with definitions and relationships through synonyms, which can help DL create concatenated representations of these words independently, resulting in high similarity scores compared to representation without KG infusion. Other examples of ambiguous and conceptual entities are provided in Figure 2.1 showing cases where BERT needs external knowledge to capture the context for correct classification on QQP dataset.

Long Tail Entities: DL models in NLP provide representations after learning a large volume of raw text. Essentially, it creates and stores an index of words and word-word co-occurrences which is considered distributional semantics to generate numerical

²<https://babelscape.com/doc/pythondoc/pybabelnet.html>

³<https://github.com/goodmami/wn>

representation [51]. Majority of time, the entity representing the theme of the document is sparsely present. As a result its representation is not as rich as other words occurring frequently. These sparsely distributed entities are called long tail entities and affect any DL model by missing context. This is often the case in multi-hop question answering problems [60]. Consider an example [6]:

Question: Sodium azide is used in air bags to rapidly produce gas to inflate the bag. The products of the decomposition reaction are:

1. Na and water
2. Ammonia and Sodium Metal
3. N_2 and O_2
4. Sodium and Nitrogen
5. Sodium Oxide and Nitrogen Gas (Correct Answer)

The entities in the correct answer are not present in the question. Further, retrieval of passages that can fine-tune a DL model to generate correct answers is hard as it would require passages to be semantically related and logically ordered for correct deduction [31]. One can utilize RAG Model [61] or REALM [62] for apropos passage retrieval and extend their capabilities using KGs. To correctly answer the question, we retrieve conjunctive or disjunctive sets of passages using keywords: {sodium azide, air bags, gas, and decomposition}. (*Passage 1*) Sodium azide (NaN_3) reacts in heat and decomposes to Na and N . (*Passage 2*) Oxidation-Reduction decomposition reactions are redox reactions wherein electrons are transferred from the atom that is oxidized to the atom that is reduced. (*Passage 3*) Ionic-Compound decomposition, like in NaN_3 occurs when a binary ionic compound is heated. (*Passage 4*) Air bags contains sodium azide and other gas to prevent sodium hyperoxide. Passage 2 & 3 are semantically related by the term "decomposition," and Passage 3 directly informs

Passage 1 using "decompose" and "heat" as the concepts. Since Nitrogen (N) undergoes oxidation or reduction, it is related to Passage 2. Finally, Passage 4 logically follow Passage 1 with the term "sodium hyperoxide." This yields Sodium Oxide and Nitrogen Gas as the correct answer. The order {Passage 2 & 3} → Passage 1 → Passage 4 is possible by exploring the relationships between passages, for which KGs are required [17]. Table 2.1 illustrate the importance of context sensitivity in other domains of social impact.

Table 2.1 Benefits of context capture by KiL.

Domain	Post	Outcome from DL	Outcome from KiL
Mental Health	Really struggling with my bisexuality which is causing chaos in my relationship with a girl. Being a fan of LGBTQ community, I am equal to worthless for her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts, and need to get it out of my head.	<struggling, worthless, drunk> Prediction: Depression (True: 0.71) (✗)	<struggling, bisexuality, chaos, relationship, worthless, drunk, intrusive thoughts> Explanations (high-level concepts): <health-related behavior, level of mood, drinking, obsessive compulsive personality disorder, disturbance in thinking> Prediction: Obsessive Compulsive Disorder (True: 0.96) (✓)
Radicalization	Here is the fragrance of Paradise. Here is the field of Jihad. Here is the land of #Islam. Here is the land of the Paradise.	<Jihad, Islam> Prediction: Extremist (True: 0.90) (✗)	<Paradise, Jihad, Land, Islam> Explanations: <Paradise_Land, Jihad_Islam> Prediction: Non-Extremist (True: 0.87) (✓)
COVID-19	#Flu and #Pneumonia killed six times more people as #Covid19	<kill, more people, covid19> Prediction: Fact (True: 0.64) (✓)	<affected population, communicable diseases> Prediction: Fact (True: 0.865) (✓✓)

2.2 EXPLAINING DL MODELS

For broader assimilation of DL models in a variety of domains, their black box nature needs to be addressed. In healthcare, clinicians routinely choose methods that allow them to understand how an outcome was derived compared to an objectively superior method that cannot be explained. In education, tracing students' learning outcomes with attribution to weak academic and behavioral areas is a better tool for teachers compared with the ability to predict only a student's performance [27] [63]. Making explanations of model behavior is subjective from the stakeholder perspective. A set of privileged knowledge (e.g., domain expertise, advice specific to the situation) must be infused to comprehend the model outcomes and interpret its functioning. Thus, we define two forms of explainability: System-level Explainability and User-level Explainability. An illustration of these two forms of explainabilities is shown in Figures 2.2(A) and 2.2(B) [64].

System-Level Explainability (SysEx): Generating explanations after the analysis of word and token level feature importances through a suitable visualization mechanism, such as a saliency map. For instance, first derivative Saliency based methods explain the decision of an algorithm by assigning values that reflect the importance of input features in their contribution to that decision in the form of a gradient map (heat map) [65] [66] [67]. Another method for SysEx is Layer-wise relevance propagation, that decomposes the prediction of a deep neural network for a specific example into individual contributions from sub-parts of the text [68] [69] [70]. Input perturbations and Attention Models are other methods for SysEx [71] [72] [73].

User-level Explainability (UseEx): is the ability to generate human-comprehensible explanations around the decision-making process. Explanations would generally be in natural language, or with a visual depiction with trace over a generic or domain specific knowledge.

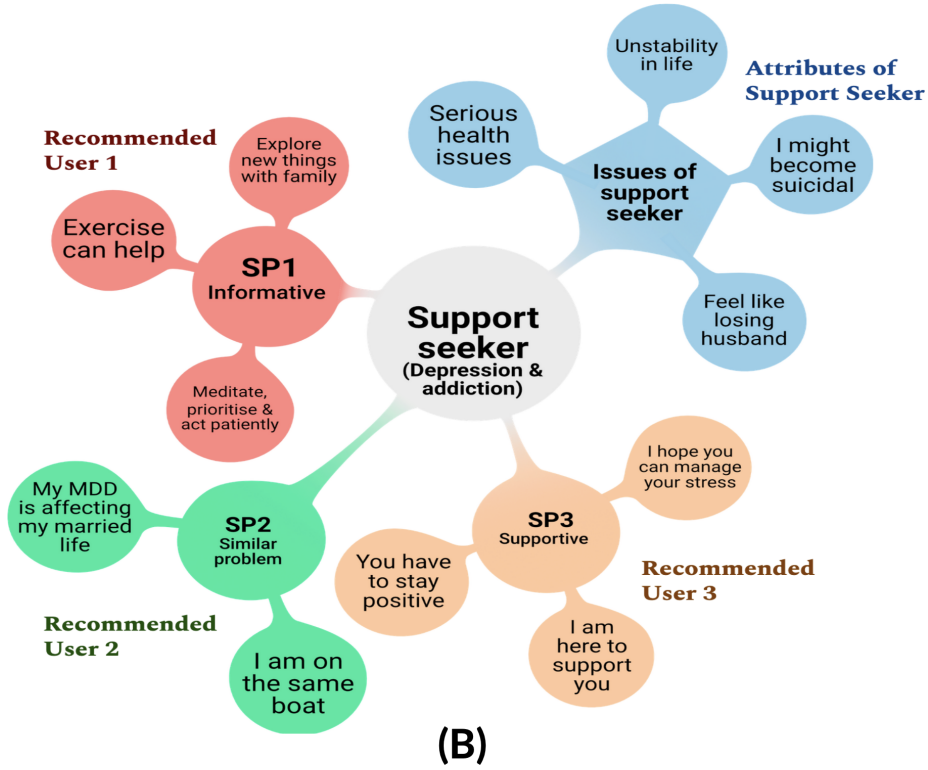
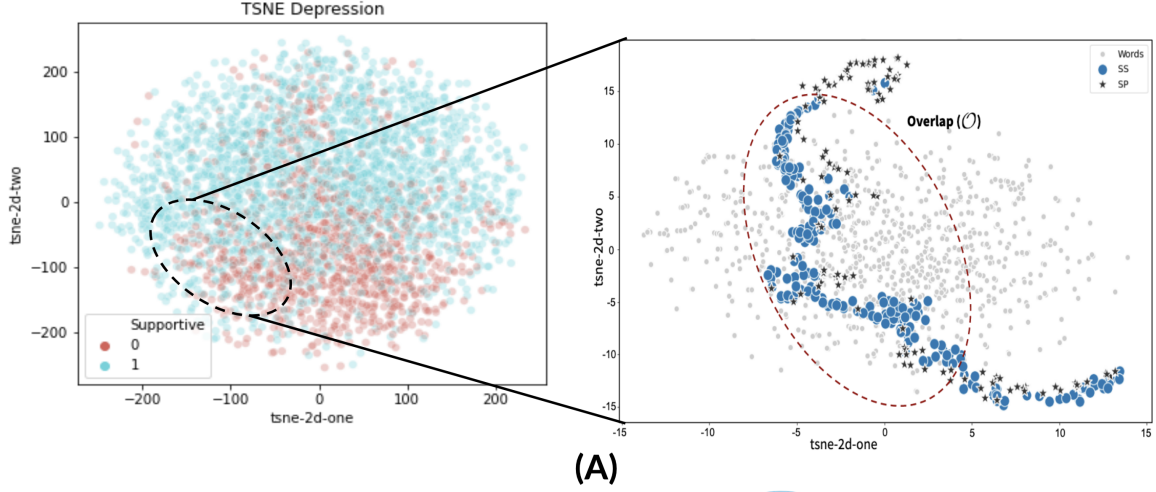


Figure 2.2 (A) System-level Explainability showing the overlap (a proxy of matching a support seeker (SS) with its nearest support providers (SPs) using T-SNE visualization. (B) User-level Explainability showing the reason behind mapping semantically-related SPs to a SS through the use of phrases that are semantically similar to concepts in Patient Health Questionnaire Lexicon (PHQ-9).

There is a growing trend of fine tuning the pre-trained models with limited labeled data. These have succeeded when (a) the distribution of the labeled dataset is similar to unlabeled data used for pretraining, and (b) tasks are relatively straightforward like natural

language entailment and span extractive question answering. However, real-world scenarios are often more complex, which poses the following challenges: (a) Fine-tuned models for domain-specific tasks with limited labeled data may not be sufficient to capture domain knowledge [74]. (b) Self-supervised training objectives over unlabeled data are not attempting to learn/acquire the domain knowledge required for real-world.

Methods for Explainable AI (XAI): Recent research in XAI has attempted to address several aspects of *opening this black box* to help humans, both the system users and domain experts, understand such models' functioning and decision-making process [75]. Adoption of AI systems occurs in two stages:

Model Building Phase: This includes model features, algorithmic development and error analysis and refinement of model. Explicit knowledge as abstract concepts, processes, policy/guidelines, and regulations are essential to infuse into the AI system for sensible explanations comprehensible to humans.

Explaining Phase: This includes decision-making, knowledge capture, and trust and bias analysis. This phase includes user-in-the-loop (e.g. stakeholder) to assess consistency in the model and match user expectations [76].

Developing a good quality XAI system requires domain experts in the annotation, supervision, and evaluation phases [77] [64]. For this purpose, domain experts require explanations that are in the form of an expert working in that domain or that application would give, using the language and concepts normally employed by a person working in that field. For example, in the medical domain, the outcome of a model needs to be explained by positioning against conceptual knowledge contained in clinical guidelines. Analysis of word-level and token level features is of little to no use to a domain expert during evaluation [61]. Methods that incorporate KGs to provide a conceptual level explanation of the model outcome could improve explanations and ease of evaluating AI systems. Popular metrics to

assess model’s language understanding such as, BLEU, ROUGE-L [78], QBLEU4 [79], BLEURT [80], and MAUVE [81] should be improved by included a score computing component that measures closeness of predicted outcome with concepts in KG [17] [31]. This will lead to trust in the systems by end-users and speedy adoption into the real world.

2.3 INTERPRETABLE MODELS

Interpretability is the ability to discern the internal mechanisms of an optimization module within an AI or Data mining framework. For example, consider a transformer model whose key modules are: (a) input embeddings, (b) positional encodings, (c) attention layer, (d) loss function, and (e) batch normalization with or without dropout. We can call a transformer model interpretable if we can meaningfully interpret the functioning of each internal component ((a)-to-(e)) and can affirm that model is functioning in our intended way. There are four methods to construct an interpretable model:

Probing: Probes are shallow neural networks (e.g., 2-layer Neural Network, Restricted Boltzmann Machines) placed over intermediate layers of a larger neural network, whose functioning needs to be interpreted ⁴. They help investigate what information is captured by different layers or attention heads. Probes are trained and validated using auxiliary tasks to discover if such auxiliary information is captured. Through probing, it is fairly interpretable to see how input tokens are contextualized in successive layers using attention mechanisms and how the model performs in sub-tasks that are a decomposition of the major task. A KG can help in probing by computing the distance between the intermediate hidden representations of a DL model and concepts in KG [27].

Fine-Tuning: A pre-trained model that is not fine-tuned comprises learned parameters supporting global parametric knowledge. Fine-tuning allows the model to respond

⁴<https://tinyurl.com/AI-probes>

sensibly to a given task. It is a process of precisely adjusting the model's parameters to observations that are related or similar to the observations on which the model was trained. Problems that require fine-tuning would require a known mechanism to explain the model's behavior and support reasoning. This is seen in the form of human evaluation tasks, visual inspection of the model's output or qualitative error analysis [82]. There are various fine-tuned transformer models on Huggingface⁵, but for an interpretable fine-tuning a KG component is required [83]. It has been shown in K-BERT, where a KG is augmented to a data representation. For example, in K-BERT representation of the term "cholesterol" is enhanced by augmenting the representation of the triple "<cholesterol> <causes> <heartattack>." For the downstream task, the relationships and entities captured in the KG can help in improved prediction.

Multi-Task Learning: is a popular phenomenon to train the same model for multiple tasks. It enriches the semantic representations of models and avoids them getting overfitted. Auxiliary tasks could also be part of such a setup. For instance, sentiments associated with a medical text can be well studied automatically through DL if the algorithm can master the identification of medical conditions, treatment, and medication. This forms a multi-task learning problem solvable through a suitable DL algorithm [84].

Autoencoders: are interpretable models as they are weighting functions and contextual representation learners because of the optimization function, which is a reconstruction loss. The encoder-decoder architecture is a container that can accept the DL model (e.g., sequence-to-sequence long short term memory (LSTM), graph neural networks (GNN), convolutional neural networks (CNN)) and trains it to learn representation by mapping input to output. An amazing utility of autoencoder comes from

⁵<https://huggingface.co/docs/transformers/training>

replacing the decoder end with a knowledge source. It can be a knowledge graph if the internal component is a GNN, a document if the internal component is an LSTM model, a lexicon if the internal component is the simple continuous bag of words embedding model, and many others [10].

Table 2.2 Comparison between different methods that make DL algorithms interpretable. The complexity classification is discussed in detail in Chapter 3.

Interpretability Methods	Probes	Fine Tuning	Multi-task Learning	Autoencoder
Goal	Auxiliary Task	Primary Task	Primary Task	Optimize Input with Knowledge for Primary Task
Update Model Parameters	No	Yes	Yes	Yes
Access Model Internals	Yes	No	No	Yes
Complexity	Shallow	Shallow or Semi-Deep	Shallow or Semi-Deep	Semi-Deep

It is important to note that model interpretability achieved by probing and fine-tuning provides system-level explainability and not user-level explainability. Autoencoders differ from fine-tuning and probing by making the model interpretable and explainable through user-level knowledge, which is introduced by calculating conceptual information loss and proportionately propagating it in the neural network by modulating hidden representations (see Table 2.2 for comparison). A model’s capability to be interpretable and user-level explainable also lies in the type of dataset it is trained on.

2.4 FROM GLUE TO KILU

The NLP community has set up a set of tasks across various benchmark datasets called General Language Understanding Evaluation (GLUE) tasks [32]. They test a variety of natural language tasks such as textual entailment, textual similarity, and duplicate detec-

tion. However, recent research has shown that such tasks do not require external knowledge, which is often the requirement in real-world problems concerning natural language understanding [28]. GLUE tasks do not test if the model can leverage knowledge, the explanations generated are of limited utility to humans [33]. Recently developed benchmarks under the name “Knowledge Intensive Language Tasks” (KILT) has focused on building retrieval-augmented AI models to better understand of natural language with support from passages that can capture context in user’s input. Parallely, we also introduced “Knowledge Intensive Language Understanding”(KILU) tasks that are as of now focus in making AI model usable in mental healthcare setting. Table 2.3 enumerate the tasks in KILU that require external knowledge to match with human-level performance. Further, there are task-specific metrics to evaluate the performance of models built to solve KILU tasks.

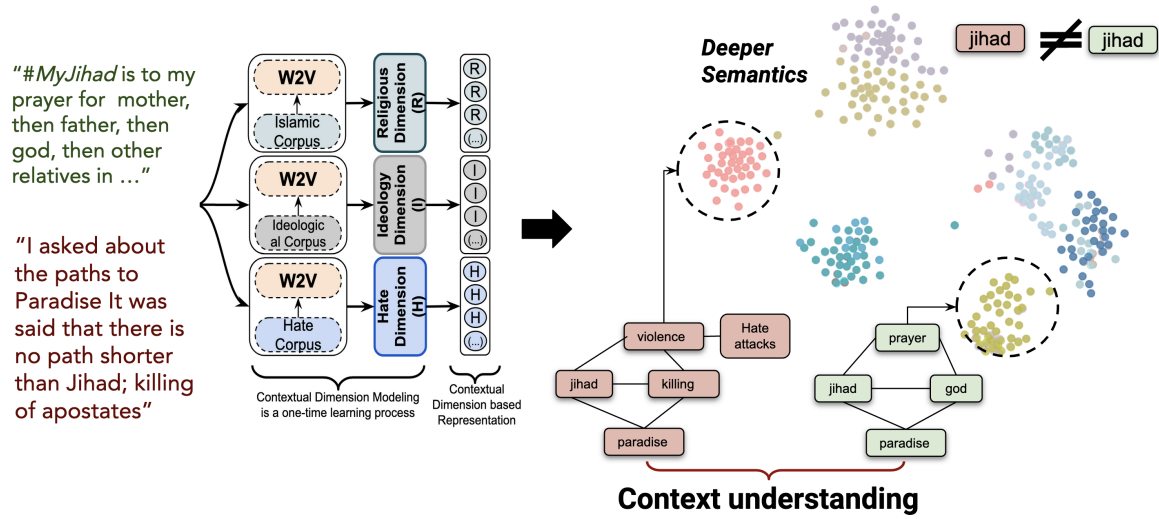


Figure 2.3 An illustration of context modeling in language model using external domain-specific corpus. Subsequent clustering manifest pairing of concepts that co-occur in a domain-specific corpus. The clusters are explainable using the relationships between these concepts. This figure illustrate deeper semantics in a computational social science problem of detecting radicalization behaviors in dynamic stream of tweets. The word *jihad* occur in two connotations and is clearly separable using domain-specific knowledge.

Essentially, KILT or KILU induce another set of capabilities in AI models to capture information similar to how a human does. These are:

Abstraction: The task of mapping low-level features to higher-level human-understandable abstract concepts is known as abstraction. Humans often speak in terms of higher-level abstract concepts when explaining their decision to a user. AI systems also need to explain decisions to the end users using abstract domain-relevant concepts constructed from low-level features and external knowledge in a KG.

Contextualization⁶: is defined as interpreting a concept with reference to relevant use or application. Human contextualize by processing the information through various knowledge sources (e.g. syntactic, structural, linguistic, common-sense, and domain-specific). Contextualization is necessity in a domain of social good wherein a mis-classification can have severe consequences. For example, to classify if tweets are from *Extremists* or *Non-Extremists*, it requires various forms of contextual knowledge to precise classification [85]. Tweets in the domain of radicalization represent a mixed context of religion, ideology, and violence/hate. Thus, modeling of user content independently from these domain contexts is important for better clustering and classification thereof (see Figure 2.3).

Personalization: Identifying data point-specific information and integrating it with external knowledge to construct a personalized knowledge source is known as personalization. For example, a person's depressive disorder can be due to family issues, relationship issues, and clinical factors. All of these affect the context specific to the individual and consequently affect his symptoms and medications differently than that for another person.

2.5 SUMMARY

In this chapter, we discussed, (i) why AI models should be context sensitive, (ii) be capable of providing user-level explanations, (iii) be transparent by being interpretable, and (iv) how such capabilities be achieved in improving benchmarking datasets. Essentially,

Table 2.3 GLUE tasks are classification or prediction tasks taking a sentence or pair of sentences as input. It is not meant for generation or structured prediction. On the other hand KILU tasks subsumes GLUE Tasks and challenges DL models on user-level explainability and interpretability. To provide explanations to KILU tasks, the model should leverage variety of explicit knowledge to capture context and learn necessary abstraction for human comprehension. EM: Evaluation Metrics used in GLUE and KILU.

GLUE Tasks	EM-GLUE	KILU	EM-KILU	Knowledge Source
Corpus of Linguistic Acceptability (CoLA) [86]	Matthew's Correlation	Summarization of Conversational Data [23]	Thematic Overlap, Flesch Reading Scale, Jensen Shannon Divergence, and Rouge-L	Structured Clinical Interviews, PHQ-9
Stanford Sentiment Treebank (STB) [87]	Accuracy	Predicting severity class of suicide on Reddit [37]	Precision, Recall, Ordinal Error, and Perceived Risk Measure	DSM-5 and Drug Abuse Ontology [88]
Microsoft Research Paraphrase Corpus [89]	F1-Score and Accuracy	Information Disguise [90] using The user-language Paraphrase Corpus and Reddit data	Word Mover Distance [91], BLEURT [80]	ConceptNet [45], WordNet [46]
Semantic Textual Similarity Benchmark [92]	Pearson Spearman Correlation	Text-based Emoji Sense Disambiguation ⁷	Average Accuracy	EmojiNet [93]
Quora Question Pairs [94]	F1-Score and Accuracy	-	-	-
MultiNLI Matched/ Mismatched [95]	Accuracy	Mediator to link User with diverse roles (Need-Resource) [64]	Time-to-good match, Precision, Recall, F1-Score, and Human Evaluation	Psycholinguistics, Mental Health Lexicon (for a use-case), domain-specific, & Event-specific Features
Question NLI [96]	Accuracy	Information Seeking Question Generation for Conversational Assistance [31]	BLEURT, Semantic Relations, Logical Coherence, Rouge-L	Wikipedia [97], WikiNews [98], MS-MARCO [99]
Recognizing Textual Entailment [100]	Accuracy	ProKnow: Dataset and Method for Process-guided, Safety-Constrained, and Explainable Mental Health Diagnostic Assistance [17]	BLEU, Rouge-L, Avg. num. of unsafe matches (AUM), Average Knowledge Base Concept Matches (AKCM), Average squared rank error (ASRE)	PHQ-9 and GAD-7
Winograd NLI [101]	Accuracy	-	-	-

this chapter motivates, why KiL is needed and pressing on the points (i)-(iii), how knowledge infusion will take place in AI. We reviewed existing statistical methods and metrics

devised to assess the explainability of the model and interpretability of its mechanisms quantitatively. Existing frameworks categorized as post-hoc interpretability, counterfactual explanations, and rule-based explanations fall short in providing answers to the following open questions: (a) Can the model mine (varied) relationships from the existing text? (b) Can the model reliably classify entities into known ontology? (c) Can the model answer the question with trust and transparency? (d) Is it possible to measure the model’s “reasonability” and “meaningfulness” of the response to a question? (e) How much context is needed for the model to provide a precise response?

An emerging trend to fine-tune a pre-trained model on limited labeled data for a downstream task and the inability of distributional semantics learning to capture domain-specific knowledge pose limitations in addressing the above questions. We noted the necessity of KG as an integral component in neuro-symbolic AI systems with capabilities to generate explainable outcomes. System oriented explanations do little for a domain-expert or an end user who need to be able to trust the AI system’s decision making process, and its adherence to the real-world processes, rules and guidelines. For this, the XAI needs to offer explanations that the end-user or domain expert can easily comprehend. Users do not think in terms of low-level features, nor do they seek to understand the inner workings of an AI system. The user thinks in terms of abstract, conceptual, process-oriented, and task-oriented knowledge external to the AI system. Such external knowledge also needs to be explicit (e.g., as modeled by a knowledge graph), not implicit (i.e., implied by statistics or a vector representation). This chapter also shows the need to develop better NLU benchmarks beyond GLUE that can effectively test the ability of the AI system to explain decisions in a human-understandable manner. In the subsequent chapter, we will delve into methodological and application details to retrieve answers for the aforementioned questions (a) to (e).

CHAPTER 3

SHALLOW INFUSION

KiL is a continuum that comprises three stages for infusion of knowledge into the machine/deep learning architectures. As this continuum progresses across these three stages, it starts with a **Shallow Infusion** in the form of embeddings, and attention and knowledge-based constraints improve with a **Semi-Deep Infusion**. For deeper incorporation of knowledge, we articulate the value of incorporating knowledge at different levels of abstractions in the latent layers of neural networks. We consider it to be a **Deep Infusion** of Knowledge as a new paradigm that will significantly advance the capabilities and promises of deep learning.

Significance

Shallow infusion is about converting the knowledge into the same form as data used by current data-driven statistical AI. The numerical representation (or vector) of the knowledge is used to enhance the representation of data in statistical AI algorithms.

When we talk about knowledge infusion, we consider two forms of knowledge:

Unordered Knowledge: It is defined as any structural information that **does not** enforce logical ordering in the outcome. Examples include all the existing knowledge graphs (KGs), such as DBPedia is the unordered knowledge of Wikipedia, UMLS (Unified Medical Language System) is the unordered knowledge of medical information (disease, symptoms, treatment, medication, etc.), and many others (see Figure 1.4). Semantic Lexicons are another form of unordered knowledge. In comparison with KGs, lexicons are driven by a purpose and can be considered a subset of

KGs. For instance, Linguistic Inquiry Word Count (LIWC) is a competitive lexicon to capture psycho-linguistic information [102]. ANEW and GoEmotions are example lexicons to capture emotions [103] [104]. The severity of Suicide Risk and Depression are specialized use-cases under mental healthcare that require dedicated lexicons. Recent studies have developed lexicons to capture entities that contribute to the assessment of suicide risk or depression from noisy social media communications [37] [105]. Unordered knowledge infusion is helpful in classification tasks and generative tasks as long as the task does not seek logical ordering in the outcome.

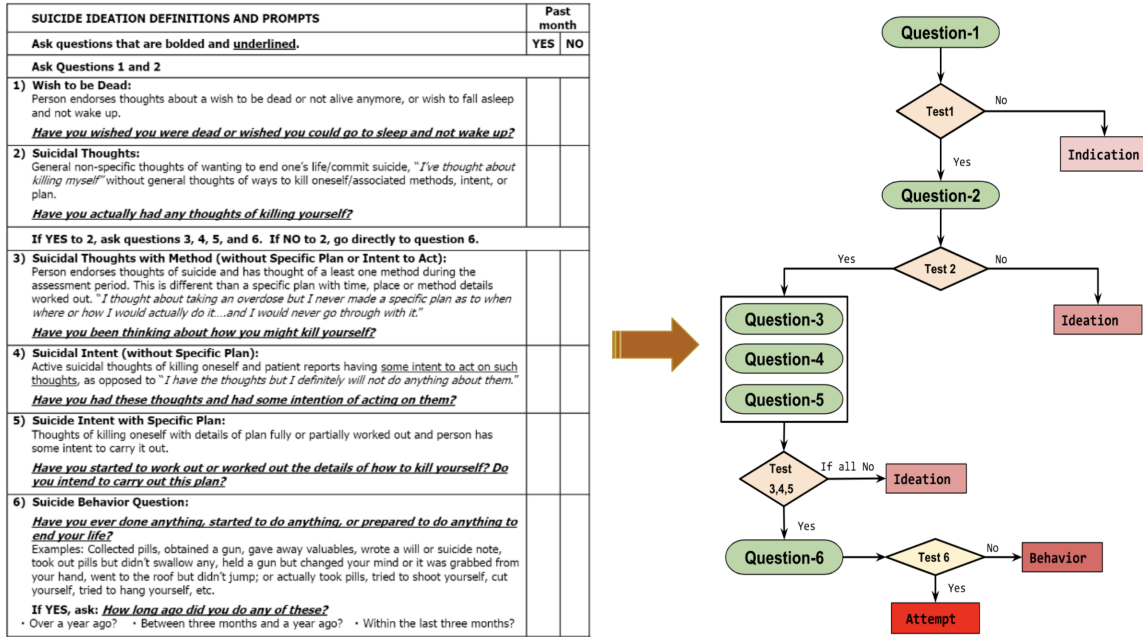


Figure 3.1 An Illustration of ordered knowledge (right; also called process knowledge) constructed from the Columbia Suicide Severity Rating Scale (left), one of questionnaire used by MHPs for suicidality detection.

Ordered Knowledge: It is defined as any structural information that enforces logical ordering manifested in the form of conceptual flow in the output of an AI model. This form of knowledge is required for generative tasks, such as question generation, response generation, or response shaping, wherein information is desired in a particular way. An example illustration of ordered knowledge is shown in Figure 3.1. It sees wide application in current conversational AI research, wherein the task is

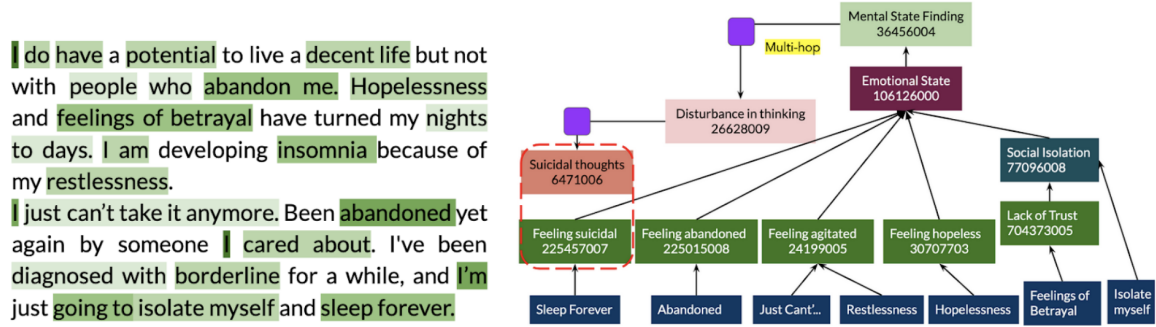


Figure 3.2 An illustration that associates system-level explainability with user-level explainability. The highlight phrases in the left-side of the figure is obtained from a DL model trained using the method in Gaur et al. [10]. This is a manifestation of system-level explanations. Highlighted phrases in the input text are queried in SNOMED-CT, thus forming a contextual tree (right-side of the figure). This is manifestation of user-level explanations. Formation of this tree is stopped when a node is hit that has high similarity to either leaf nodes or one hop parent nodes. The numbers in the boxes are SNOMED-CT IDs.

to engage with the user in a meaningful manner. For instance, the task of conversational information-seeking requires the agent to either ask questions to the user or provide a response to the user in a particular order. At a broad level, this order can be seen as categories: <Definition> is followed by <Method> is followed by <Application/Use-Case>. Consider an example utterance from a user, “How to prepare Hibiscus tea?” a convincing response would have the following order: <Ingredients> is followed by <Method> is followed by <Use>. If an AI model is able to formulate the sequential nature of the knowledge, it can generalize over a set of similar tasks. So far, there has been one study that utilizes such ordered knowledge (a.k.a procedural knowledge or process knowledge) in generating sentences that describe the severity of suicide risk of an individual [17].

3.1 BENEFITS OF **Shallow Infusion**

Major focus in this chapter would be to learn various ways in which the datasets can be transformed using external knowledge. Shallow infusion concerns with semantic data transformation and provides following benefits:

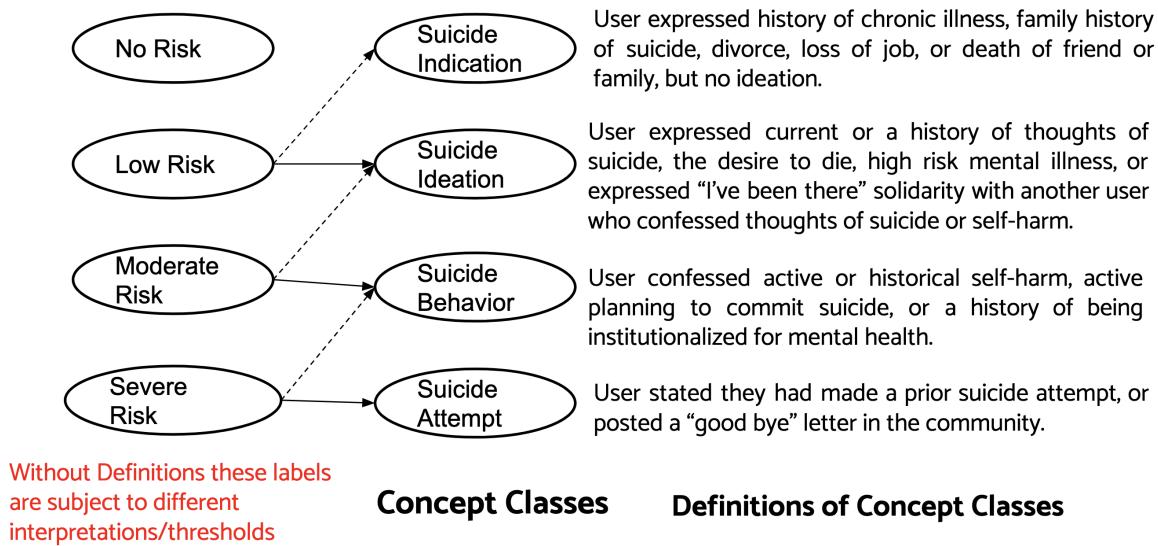


Figure 3.3 An illustration of concept classes to assess suicide risk. These concept classes are obtained from Columbia Suicide Severity Rating Scale [11]. Dotted arrow from a “not-so-well-defined” label to well-defined concept class shows that the label **can** resembles this class if predicted probability for solid arrow is lower than dotted arrow. Solid arrow from “not-so-well-defined” labels to well-defined concept classes shows that these labels certainly resembles this class if predicted probability for solid arrow is higher than dotted arrow. This dichotomy on the part of “not-so-well-defined” labels is removed using concept classes.

- **Concept Classes:** Suppose the outcomes labels predicted by an AI model lack concrete definitions that distinguish one label from another; then, the classification is subject to varied interpretations. Furthermore, such labels are created based on an empirically defined threshold that is inappropriate for high-stakes decision-making problems. It is acceptable in the general-purpose domain; however, it is not affordable in a healthcare setting, where a subsequent decision has to be made upon the predicted label. The *Shallow Infusion* brings in the concept of *concept classes* which are labels with definitions. Figure 3.3 illustrates a map between a not-so-well-defined set of labels and concept classes. These classes are domain-specific and can make AI systems capable of capturing context [106], handling uncertainty and risk associated with ambiguity [37], and providing system-level explainability and user-level explainability by mapping the model-defined important features to concepts in KGs.

- Entity Normalization (EN): The linguistic variations in online communication raise challenges for the supervised learning algorithm in determining discriminative patterns. For example, consider the following two posts: **(P1)** “I am *sick of loss* and *need a way out*” ; **(P2)** “*No way out*, I am *tired of my losses*”; (P3) “Losses, losses, I want to die”. The italicized and underlined phrases in P1 and P2 are a predictor of suicidal tendencies but are expressed differently [106]. **Shallow Infusion** of knowledge remove these variations through a process called entity normalization that calculates the semantic similarity between n-gram phrases and concepts in a knowledge source (e.g., KGs, Lexicons). To perform EN, we generate a vectors of words in the input using an embedding model (e.g., Word2Vec [51], ConceptNet Numberbatch [45]) and computer similarity (e.g. Word Mover Distance [91], Cosine Similarity, BERTScore) with concepts in various knowledge sources. If we perform EN, then P1, P2, and P3 transform to “depress, suicide ideation”, “suicide ideation, depress”, and “depress, suicide attempt” respectively. This clearly shows that P1 and P2 are related and distinct from P3.
- System-level Explainability: please refer to Section 2.2
- User-level Explainability (UseEx): please refer to Section 2.2. Figure 3.2 show the difference between system-level explainability and user-level explainability in natural language processing applications involving neural attention models [73].

3.1.1 WHAT IS **Shallow Infusion**?

We define shallow infusion, the first category of knowledge infusion, as any attempt that compresses or transforms knowledge into a flattened intermediate form for use with DL models. Specifically, shallow infusion does not require the learning model to be significantly changed to ingest the external information. The shallow infusion encapsulates external knowledge and enriches the deep network representation as either word embeddings

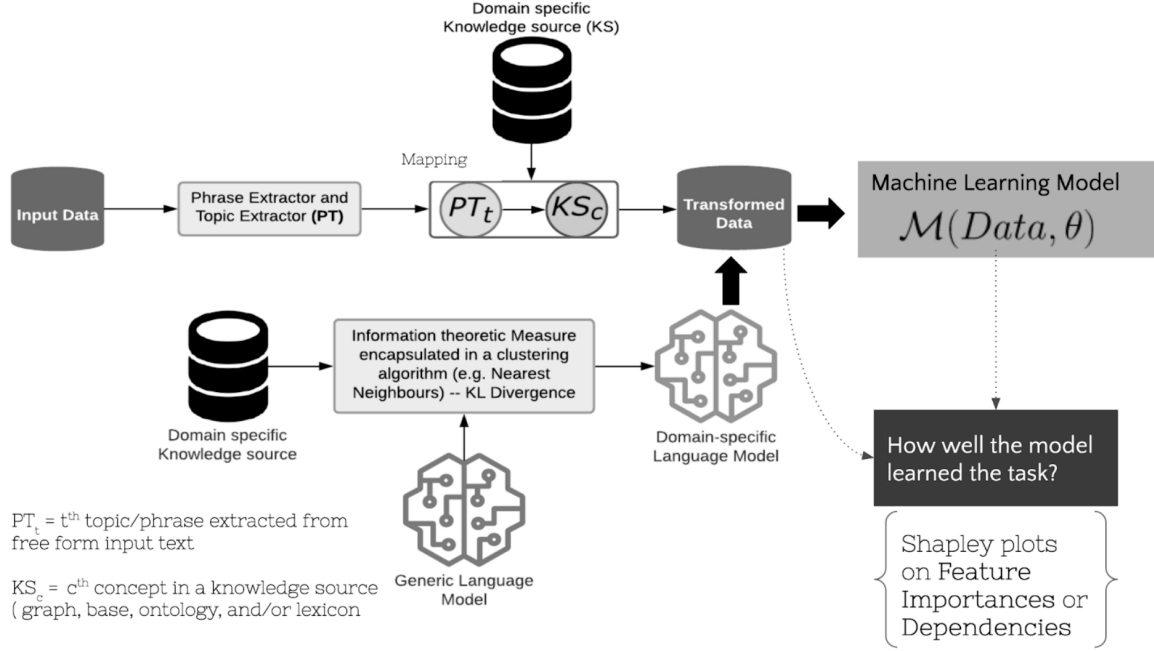


Figure 3.4 A generic architecture of **Shallow Infusion**.

for textual knowledge and graph embeddings for graphical knowledge (see Figure 3.4). An advantage of these methods is that they tokenize the input into smaller phrases and can therefore effectively handle misspellings, abbreviations, or etymologically similar text. However, they ignore relationship semantics between entities in external knowledge and are consequently greatly limited in their applicability in domains where contextualization through knowledge is required.

To see how shallow infusion can be applied in current state-of-the-art models, we note that recent advances in deep networks employ language models that use an attention mechanism to define the context of words given their neighborhood in the input dataset. The current state-of-the-art transformer models, such as BERT, broke records for several NLP tasks, learned to capture long-term dependencies and context by training on large amounts of text. Several other works have seen ground-breaking results with several Transformer-based successors of BERT (e.g., Roberta, XLNet, and Transformer-XL). However, context sensitivity and handling uncertainty and risk, has not been resolve, in spite of scaling the parameters of these models from millions to billions. This has a consequence, a large-scale

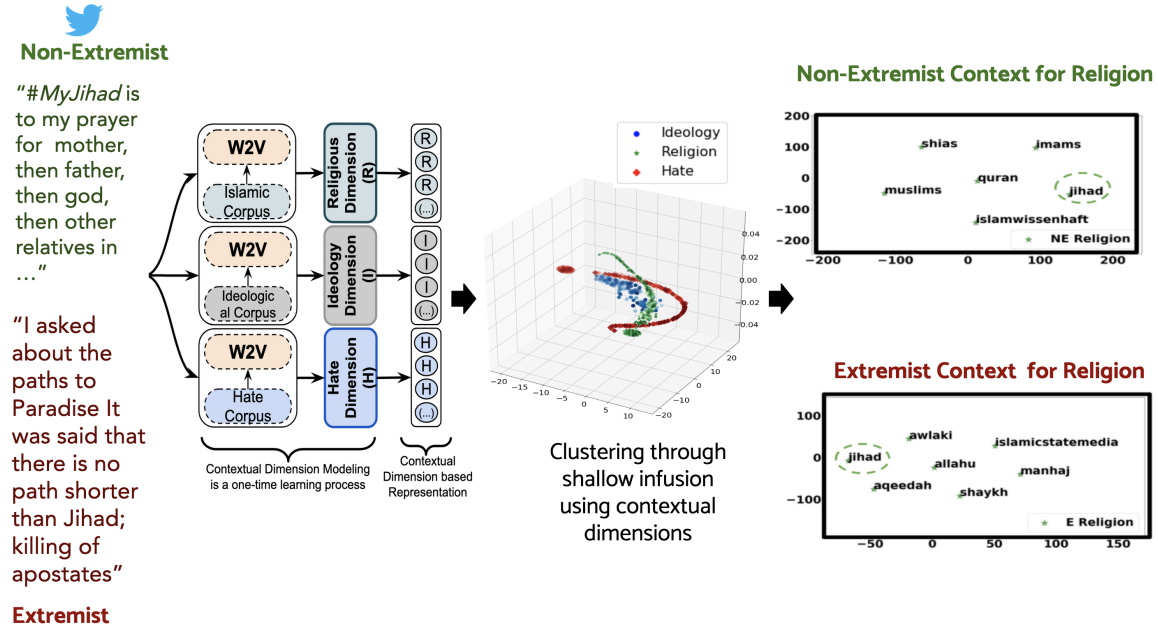


Figure 3.5 A shallow infusion process using contextual dimensions from the radicalization literature. The visualization is performed using T-SNE method [12]. Explainable view of the clustering is provided in Figure 2.3

model memorizes the patterns in the dataset on which it is trained and tested, and is difficult to adapt the model in a similar or related tasks. For instance, a model learned to identify and classify harassment on social media with simple scaling of parameters is prone to misclassification on a near-related problem of “radicalization in social media” [85]. Kursuncu et al. leveraged multiple domain-specific perspective models in enriching the representation of extremist communication on social media (see Figure 3.5). The approach provided the necessary knowledge required by a model to minimize false alarms. In the context of “harassment on social media,” a potential improvement in a machine learning model was made through the infusion of cyberbullying vocabulary knowledge [107].

3.2 METHOD UNDER **Shallow Infusion**

Following are some of the well-known methods of shallow infusion that are well studied and used by the NLP community. There is a long list of methods that are classified under **Shallow Infusion** that are mentioned in table 3.1.

Table 3.1 Other methods that are classified under shallow infusion. However, not all of them are supportive of system-level explainability (SysEx). Methods which are SysEx are also capable of user-level explainability with manual effort comprising of search and retrieval over related knowledge sources [1].

Methods	Approach	Type of Explainability
Term Frequency and Inverse Document Frequency (TF-IDF) [108]	Bag of Words	✗ (SysEx)
Retrofitting	Bag of Concepts [109]	✓ (SysEx)
	Verb Phrase/ Noun Phrase [110]	✓ (SysEx)
	Sentiments and Emotion Lexicons	✓ (SysEx)
	Topic Modeling [111]	✗ (SysEx)
Latent Dirichlet Allocation	Semantic Role Labeling [112]	✓ (SysEx)
Predict than Explain [113]	-	✓ (SysEx)
Explain than Predict [114]	-	✓ (SysEx)
Embeddings	Word2Vec/ GLoVe [51]	✓ (SysEx)
	FastText [115]	✗ (SysEx)
	ELMo [116]	✓ (SysEx)
	BERT/ RoBERTa	✓ (SysEx)
Transformers	GPT-2, GPT-3, XLNet, ProphetNet	✗ (SysEx)
	T5 and Longformers [117]	✓ (SysEx)
Reinforcement Learning	Neural Policy Gradient methods using GLUE-based rewards [118]	✓ (SysEx)
Multirelational Reinforcement Learning	Functional Policy Gradient Methods [77]	✓ (SysEx)
Reinforcement Learning with Deterministic Search	Combining search and value iteration or Policy Gradients [112]	✓ (SysEx)

Word Embeddings: This is the simplest form of shallow infusion. Here, the objective is to provide the model with “background” that the training data alone could not provide. The background information is available as large text corpora (for example GloVe is trained on 6B tokens) and a shallow neural network or a statistical model is trained in an unsupervised setting to capture the domain-specific meanings of words.

The popular examples include but are not restricted to Word2Vec (skip-gram and CBOW algorithm) and GloVe. The representation of words as n-dimensional vectors (e.g., n=300) makes them easily transferable and task-agnostic within a particular domain. As a result, numerous pre-trained word embeddings are available for many languages¹ and domains².

Enriched Word Embeddings: In this class of algorithms, the pre-trained word embeddings are enriched using additional information such as domain-specific lexicons/taxonomies and morphology of words. As a post-processing technique, **retrofitting** leverages semantic lexicons such as WordNet in modifying the embeddings [110]. For example, retrofitting enforces the embedding of the word *incorrect* to be in a similar vicinity to other related words such as *wrong*, *flawed* and *false* in the embedding space. **Counter-fitting**, an approach similar to retrofitting, introduces synonymy and antonymy constraints to the word-relatedness when refining word embeddings [119]. As a result, it prevents the word *inexpensive* to be closer to words such as *pricey* and *costly* even though they are related via an antonym relation. **Fast-Text** leverages information within the text to improve the learned embeddings [115]. It considers morphology of words – particularly, sub-word information – and represents a word as a bag of character n-grams in learning the embeddings. This allows misspelled words, rare words, and abbreviations to have a similar meaning to their original forms. Moreover, this further enables deriving embeddings for words that did not appear in the training data.

Deep Neural Language Models: The primary difference in this class of models is the use of deep neural architectures with language modeling objectives – i.e., learning to predict the next word conditioned on the given context by probabilistically modeling

¹<http://bit.do/multi-lang>

²<http://bit.do/bionlp>

words in a language. ELMo (also ULMFiT [120]) marks a significant step in this direction by capturing the *context* in which a word is used in a sentence [121]. By training a task-specific Bi-LSTM network to model the language from both forward and backward directions, ELMo represents a particular word as a combination of corresponding hidden layers. The current state-of-the-art neural language modeling is inspired by the advent of Transformers – a simple, solely attention-based mechanism that disregards the need too sue recurrent and convolutional neural networks. Transformer-based BERT, a model that broke records for several NLP tasks, learns to capture long term dependencies and context by training on large amounts of text. It further fine-tunes the knowledge gained, by specifically training on a supervised-learning task. Last year has seen ground-breaking works with several Transformer-based successors of BERT (e.g. RoBERTa [122], XLNet [123], and Transformer-XL [124]) coming into light navigating the modern NLP to new directions.

The combination of these **Shallow Infusion** methods along with strategies that brings out the benefits **Shallow Infusion** sees application in public health [106] [37], crisis management (e.g. natural disasters [24], pandemic [64]), autonomous driving [125] [126], epidemiology [1] [127] [88], sports [128], and others.

We want to focus on Social Media, which is a sore point of information in terms of actionable insights it can provide to stakeholders (e.g., emergency responders, healthcare providers) and the challenges involved in extracting insights. Such as semantic ambiguity and negation in the sentences. **Negation detection** is a crucial part as the presence of negated sentences can confound a classifier. For example, *I am not going to end my life because I failed a stupid test* is not suicidal, whereas *My daily struggles with depression have driven me to alcohol* reflects user’s mental health. The former sentence can give false positive, if we just extract “going to end my life” as a precursor to a suicide attempt. Gaur et al. employed a negation detection tool and probabilistic context-free grammar to supports negation extraction and negation resolution to improve classifier performance [10].

Among various social media platforms, we would be focusing on Reddit. Reddit is one of the largest social media platforms with >430 Million subscribers and 21 billion average screen visits per month across >130,000 subreddits. On a per month average, around 1.3 million subscribers anonymously post mental health-related content in 15 of the most active subreddits pertaining to mental health (MH) disorders (42,000 posts on r/SuicideWatch) [37]. The analysis of Reddit content is demanding due to a number of reasons, including interaction context, language variation, and the technical determination of clinical relevance. Correspondingly, the potential rewards of greater insight into mental illness are in general and suicidal thoughts and behavior specifically is great. Reddit platform enables free, unobtrusive, and honest sharing of mental health concerns because a patient is completely anonymous and so can open up without worrying about any social stigma or other consequences; thus, the content is less biased and of high quality compared to the content shared in survey questionnaires and interviews [129].

Through **Shallow Infusion** we seek answer to the following questions:

(a) Can *concept classes* and *entity normalization* procedures help AI algorithms to adapt to a task of assessing severity of suicide risk at an individual level? (b) Knowing that suicide is a terminal mental illness and patients drift in time across the spectrum of mental health disorders, what architectural choices need to be made to study suicide risk in *time-variant* and *time-invariant* manner?

3.3 **Shallow Infusion** FOR SUICIDE RISK SEVERITY DETECTION

Mental Health illness such as depression is a significant risk factor for suicidal ideation and behaviors, including suicide attempts. According to SAMHSA (Substance Abuse and Mental Health Services and Administration), 80% of the patients suffering from Borderline Personality Disorder have suicidal behavior, 5-10% whom commit suicide. According to Veen et al. the probability of admission to hospital increased over different levels of suicide risk; Suicide Ideation (12%), Suicide Behaviors (25%), and Suicidal Attempt (37%). It

is important to first classify the users along these suicide risk levels so that appropriate intervention strategies can be designed.

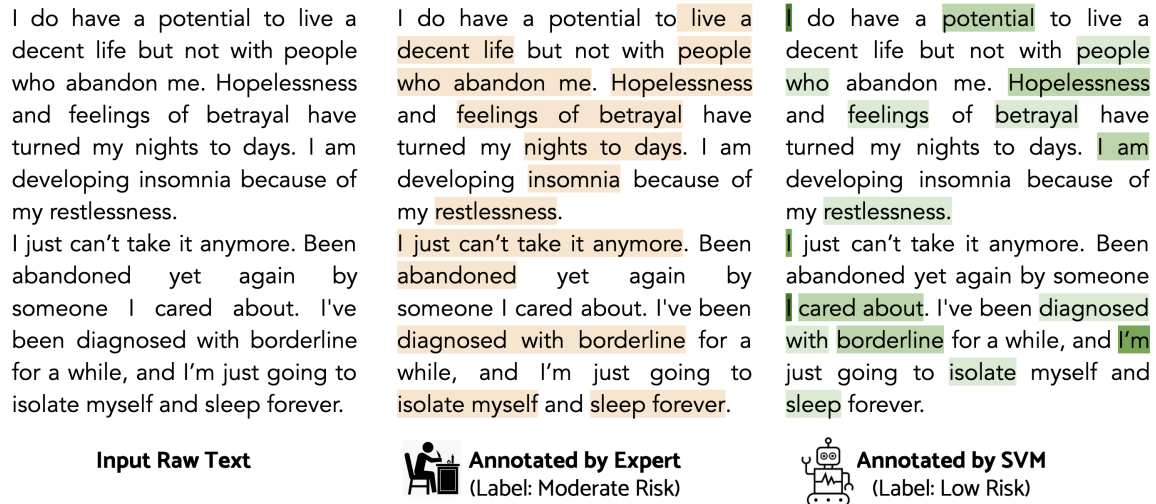


Figure 3.6 A view of input raw text being annotated by the expert and a model, respectively. It illustrates the gap between the “what a model understands as important features” compared with “how an annotator sees the text”.

Current AI models that predict suicide risk are not clinically grounded and explainable, as the labels used to label samples are not well-defined [130] (see Figure 3.3). Let us see this with an example, starting with figure 3.6. It illustrates how the annotators sees the posts and provide a label, and how an AI model sees the post through the lens of feature importance weights. This example is taken from the Reddit C-SSRS Suicide dataset, comprising of 500 posts labeled with following labels: *Supportive*, *Suicide Indication*, *Suicide Ideation*, *Suicide Behavior*, and *Suicide Attempt* [37]. Through visual inspection, it is evident that the phrases/token seem important to an expert is not given relatively close importance scores by the model. Scaling over the 500 posts, the model yielded a score of 53% recall³. As a next step, we replace the simple AI model, the support vector machine, with a large model, the convolutional network (CNN) (see Figure 3.7).

³In an order of severity levels: Suicide Indication → Suicide Ideation → Suicide Behavior → Suicide Attempt, if a model predicts a lower severity level than ground truth, it is counted in recall.

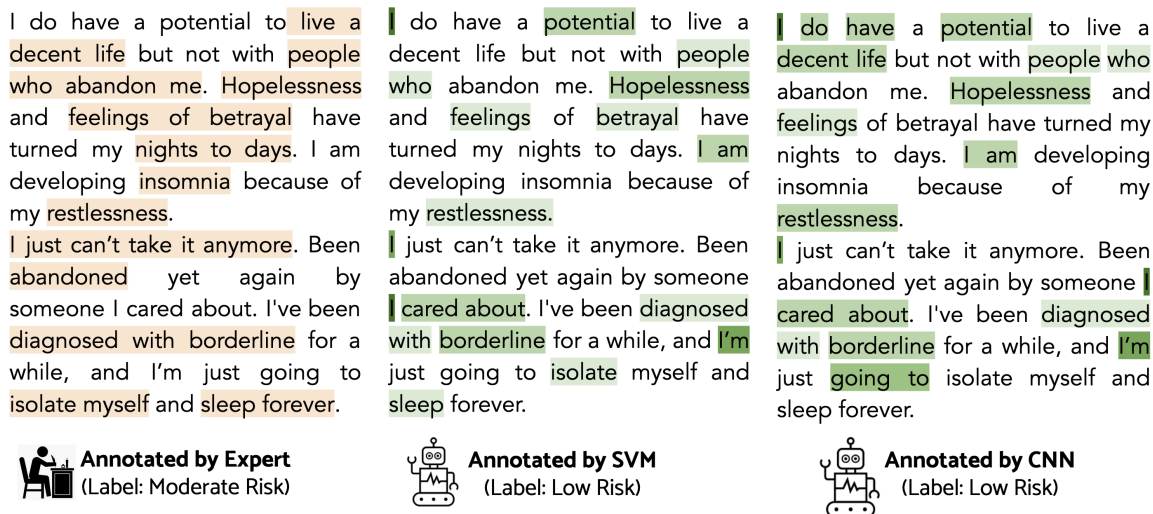


Figure 3.7 A snapshot illustrates that the discrepancy persists even with an increase in model complexity (from SVM to CNN). As a result, there is a misclassification. Since it is a case of suicide risk *severity* detection, a prediction of a low severity label can impact the quality of care a patient would receive.

Though CNN highlighted some more phrases/tokens, it did not contribute to increasing the severity level of the prediction. However, it improved the recall from 53% to 57% but failed to work well on such kind of data sample. An ingenious strategy is to make the model predict the severity level and a confidence score [131]. If the score is below a certain empirical threshold, the prediction is ignored, and the sample is sent to an expert for verification. It sounds like an intuitive strategy, and it worked well, yielding 85% recall but with only 50% sample coverage. It means the remaining 50% samples are sent to an expert for re-verification or re-annotation [132]. It is acceptable for low sample size datasets, but if we want to scale it across millions of samples to classify severity levels, experts would be overwhelmed.

This is where concept classes becomes crucial, as it minimizes the uncertainty and provides context capture. The inclusion of concept classes comes with few changes at the data and model levels. The data-level change is performed by extracting phrases (or n-grams) from the input text that have either string overlap or semantic similarity to the definitions or concepts that describe the severity level (see Table 3.2). We conduct the data-level change

Table 3.2 Suicide Risk Severity Lexicon. It can be downloaded from here

Suicide Risk Severity Class	Number of Concepts Per Class	Examples
Suicide Indication	1535	Pessimistic character, Suicide of relative, Family history of suicide
Suicide Ideation	472	Suicidal thoughts, Feeling suicidal, Potential suicide care
Suicide Behavior	146	Planning on cutting nerve, Threatening suicide, Loaded Gun, Drug-abuse
Suicide Attempt	124	Previous known suicide attempt, Suicidal deliberate poisoning, Goodbye Attempted suicide by self-administered drug, Suicide while incarcerated.

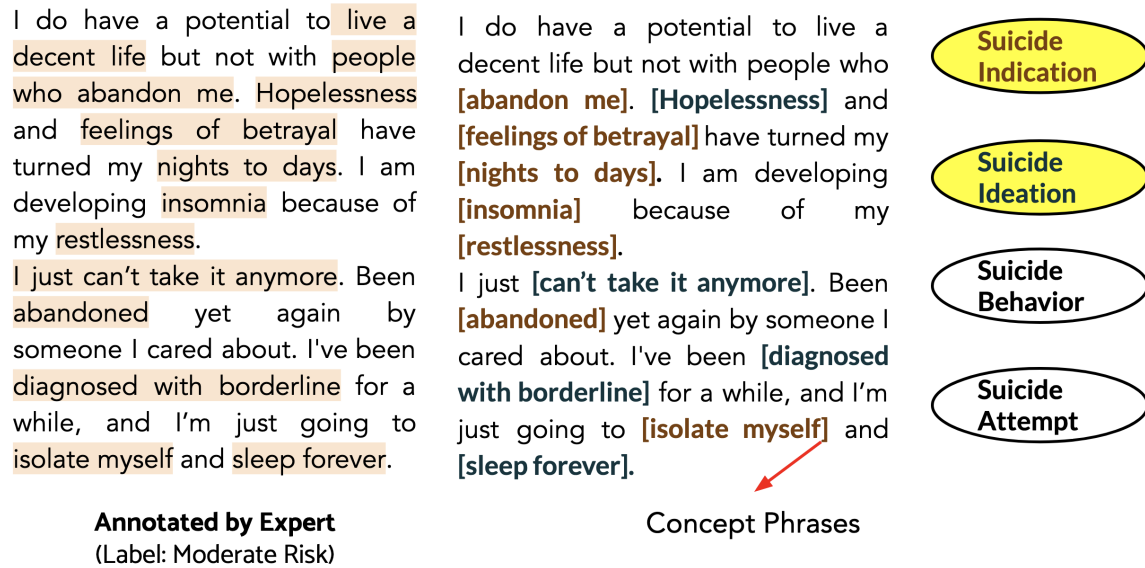


Figure 3.8 The text that contains some bracketed tokens is the transformed input text. The bracketed tokens are either similar to the concepts in the lexicon or definitions of the concept classes or present within them. Thus, we call them concept phrases. The elliptical shapes are an illustration of concept classes. Suicide Indication and Suicide Ideation are highlighted because the bracketed concept phrases are significantly similar to these classes. This transformed input text is input a model described in 3.10.

in two ways: (a) Dependency Parsing or Constituency Parsing resulting in noun-phrase or verb-phrase extraction, and finding its presence in lexicons and definitions. (b) Further, compute sentence embedding using Sentence BERT or phrase embedding using Concept-

Net and compute semantic similarity with the definition of severity levels or concepts in the lexicon. This results in Figure 3.8, where the bracketed texts are **concept phrases**. This term denotes phrases within the input text that have substantial similarities with definitions and lexicons.

By Definition, an “explainable data” is a resource created after processing the raw textual input using expert-curated knowledge sources with a purpose to understand an AI model’s behavior in classification or generation. An example illustration of the explainable data is shown figure 3.8, where an input text is pre-processed by identifying parts of the sentence that are similar to concepts in a related knowledge source (e.g. Lexicons, KGs). The bracketed tokens in the figure 3.8 are considered to be key-phrases and are termed as concept phrases after checking their presence or similarity with concepts in knowledge source. Among various ways to extract the key-phrases from the sentences, we considered constituency parsing to be a ubiquitous method across all NLP applications [133]. Figure 3.9 shows a parse tree of the first sentence in figure 3.8.

Figure 3.9 An example of constituency parse tree for the first sentence in figure 3.8. The image is created using Berkely Neural Constituency Parser, available online [here](#).

Parsing the constituency parse tree would yield noun phrases (NP) and verb phrases (VP) that are potential key-phrases reflecting on the topics of user's focus. "people who abandon me", "hopelessness", "feelings of betrayal" are some examples of NPs and VPs that are very similar to phrases identified by the annotators while annotating the post. After identification of the phrases, the next task is to check their similarity with the concepts in the knowledge source. Let us consider that we have two sources of knowledge: Lexicon (L) and Definitions (D) suitable to capture cues that describe suicide risk severity of an individual and the individual makes **P** posts, where an i^{th} post is represented as p_i . Then the method of identifying **concept phrases** using a lexicon (L) can be formulated as follows:

$$\begin{aligned}
& \{ \cos(NP_{p_i}, L) \}, NP_{p_i}, VP_{p_i} \in \text{constituency parse}(p_i) \\
& \cup \\
& \{ \cos(VP_{p_i}, L) \}, NP: \text{Noun Phrase}, VP: \text{Verb Phrase}, p_i \in \mathbf{P} \\
\text{Concept Phrases}_L(p_i) = & \cup \\
& \{ NP_{p_i} \cap \{w_0, w_1, w_2, \dots, w_n\}_{\in L} \} \\
& \cup \\
& \{ VP_{p_i} \cap \{w_0, w_1, w_2, \dots, w_n\}_{\in L} \} \\
\cos(NP_{p_i}, L) = & \text{for } np \in NP_{p_i} \text{ and } w \in L, \text{ if } \cos(\vec{np}, \vec{w}) > \delta \\
\cos(VP_{p_i}, L) = & \text{for } vp \in VP_{p_i} \text{ and } w \in L, \text{ if } \cos(\vec{vp}, \vec{w}) > \delta
\end{aligned}$$

Similarly, the method for identifying **concept phrases** using definitions (D) of concept classes can be formulated as following:

$$\begin{aligned}
& \{ \cos(NP_{p_i}, \vec{D}) \}, NP_{p_i}, VP_{p_i} \in \text{constituency parse}(p_i) \\
& \cup \\
& \{ \cos(VP_{p_i}, \vec{D}) \}, NP: \text{Noun Phrase}, VP: \text{Verb Phrase}, p_i \in \mathbf{P} \\
\text{Concept Phrases}_D(p_i) = & \cup \\
& \{ NP_{p_i} \cap \{w_0, w_1, w_2, \dots, w_n\}_{\in D} \} \\
& \cup \\
& \{ VP_{p_i} \cap \{w_0, w_1, w_2, \dots, w_n\}_{\in D} \} \\
& \cos(NP_{p_i}, \vec{D}) = \text{for } np \in NP_{p_i} \text{ if } \cos(\vec{np}, \vec{D}) > \delta \\
& \cos(VP_{p_i}, \vec{D}) = \text{for } vp \in VP_{p_i} \text{ if } \cos(\vec{vp}, \vec{D}) > \delta
\end{aligned}$$

An intersection of $\text{Concept Phrases}_L(p_i)$ and $\text{Concept Phrases}_D(p_i)$ is the total set of concept phrases in a post p_i . This process is to be followed across all the posts \mathbf{P} made by users, and it would result in a dataset with a set of identified concept phrases along with other tokens in the input texts. The utility of concept phrases is in preserving the context. Since embeddings are prone to lose semantics because of their distributional nature, concept phrases would retain semantics. Two ways in which we can create embeddings of concept phrases are: (a) Using pre-trained or fine-tuned sequential language models, they would create representations (e.g., BERT) by iterating over the concept phrases in either uni-directional (e.g., Recurrent Neural Networks, Long Short Term Memory) or bi-directional (e.g., Bi-LSTM, BERT) ways, and (b) Using word embedding models, that provides embeddings of individual words within a concept phrase and then we concatenate the embeddings and pass it through a dimensionality reduction method (e.g., Singular Value Decomposition, T-SNE) to match the dimensions of the AI model tasked for classification [73] [134] [135].

After creating the representations of the concept phrases, the representation of other tokens in the text are created using word embedding models. Final representation of the

input is through concatenation of token embeddings and concept phrase embeddings, and reducing it using dimensionality reduction. Denoising and simplifying vector space are the two main reasons for dimensionality reduction. Further, the size of the vector influences model structure, which means a lower size vector would require a simpler model and would have a low degree of freedom. This would also prevent the model from overfitting, a common phenomenon in ML/DL. Nguyen et al. provide some essential tips for selecting the dimensionality reduction method [136].

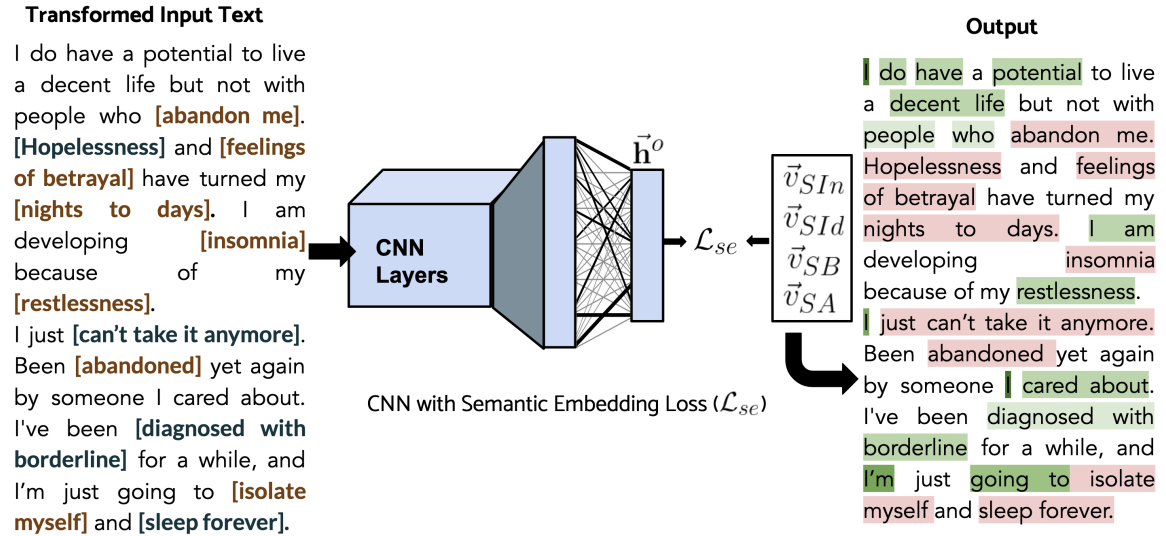


Figure 3.10 The transformed input text is the input to the CNN model that learns by computing semantic embedding loss (\mathcal{L}_{se}). This loss is defined because the concept classes have a representation form as vectors. \vec{v}_{SIn} represents the vectorized form of the definition and concepts that describe suicide indication. Likewise, \vec{v}_{SId} , \vec{v}_{SB} , and \vec{v}_{SA} represents the vectorized form of the suicide ideation, suicide behavior, and suicide attempt, respectively. \mathcal{L}_{se} compute the Euclidean distance between the representation of the input text and vectorized form of the concept classes. The output shows that by identifying concept phrases, the model learns their combined representation, resulting in an increase in their importance scores.

This forms the method to create numerical representation of the transformed input text in shown in figure 3.10. With this approach, an AI model would not tokenize the concept phrase rather consider them together, thus maintaining semantics. Now, we need to infuse knowledge into the AI model which would responsible of classifying the suicide risk severity of an individual. As you can see in Table 3.3, simply transforming the input yield sat-

Table 3.3 Shallow infusion improves recall when the model is tasked to predict suicide risk severity of a user. In such scenario Recall is the judge of model’s performance as high false negatives would result in wrong care plan for a patient with high-levels of suicide risk tendencies.

Model	Method	Recall
SVM with Linear Kernel	-	53%
CNN	-	57%
CNN [132]	Gambler Loss	62%
CNN [37]	Concept Phrases	74%
CNN [106]	Concept Phrases + Semantic Embedding Loss	84%

isfactory improvement over the baselines. However, significant boost was achieved when we performed shallow infusion of knowledge by introducing a new loss function, termed as, *semantic embedding loss*, which computes difference between the representation generated by the AI model at the outermost layer and different representations of concept classes created using embedding models. This can be formulated as follows:

$$\mathcal{L}_{se} = \min_j ||\vec{h}^o - \vec{v}_j||^2$$

where \vec{h}^o is the outermost representation of the AI model and \vec{v}_j is the j^{th} label among $\{\vec{v}_{SIIn}, \vec{v}_{SID}, \vec{v}_{SB}, \text{ and } \vec{v}_{SA}\}$. The results in the table 3.3 is recorded using ConceptNet (vocabulary= 417193, dimension= 300), a multi-lingual knowledge graph created from expert sources, crowd-sourcing, DBpedia, vocabulary derived from Word2Vec, and GLoVe. The recall score reported in the table 3.3 is computed in the following way:

$$FP = \frac{\sum_{i=1}^{N_T} I(r'_i > r_i^o)}{N_T}, FN = \frac{\sum_{i=1}^{N_T} I(r_i^o > r'_i)}{N_T}$$

where $\Delta(r_i^o, r'_i)$ is the difference between r_i^o and r'_i . r'_i and r_i^o are the predicted and actual response for i^{th} test sample.

Longitudinal and Cumulative Study of Suicide Risk: By introducing concept classes and its infusion into AI model, the approach opens-up avenues for wider applications of AI in suicide risk severity. Prediction of a suicide risk severity level is not always cumulative

Table 3.4 Example posts from a user ordered by timestamp (TS) and prediction from LSTM with semantic embedding loss. These examples illustrates the longitudinal efficiency brought into statistical LSTMs through shallow infusion.

Post 1 (TS 1):	“Homie, ... Im 27 yo, ... the <i>job is underpaying</i> - 700 euros per month. ... too afraid to search for a new job. ... fuck me, I guess? ... had these <i>thoughts of suicide</i> and these <i>fears to take charge of my life</i> from like the end of a high school. 10 years same <i>feelings of dread</i> , same <i>thoughts of killing myself</i> ”
Predicted Suicide Risk Severity:	Suicide Ideation
Post 2 (TS 2):	“One day sudden realization ... I gonna gather determination ... <i>roll over the bridge</i> . And my parents, or have a nice <i>heart attack!</i> <i>feel trapped</i> <i>nothing gonna change</i> . You will <i>end up</i> just like me. I will <i>roll over the bridge</i> ”
Predicted Suicide Risk Severity:	Suicide Behavior
Post 3 (TS 3):	“No wife, no house, no car, <i>no decent job</i> . Every single day ... <i>hating myself</i> at work Im going to <i>kill myself today</i> or tomorrow. Probably ... middle of next week, but the chances are ... <i>going to sleep forever</i> ”
Predicted Suicide Risk Severity:	Suicide Behavior
Post 4 (TS 4):	“I dont even go to the <i>exams</i> ... I might pass those exams... will <i>not graduate</i> playing some kind of a <i>Illness joke</i> ... <i>my poor family</i> .”
Predicted Suicide Risk Severity:	uninformative
User-level Predicted Suicide Risk Severity:	Suicide Ideation

of the posts made by a user, but also longitudinal. In the absence of concept classes, it is hard to capture whether the post made by a user is informative for predicting suicide risk severity or should be perform a cumulative prediction using the entire posts of the user or consider it length-wise and time-wise. Table 3.4 illustrates time-variant prediction of the AI model with semantic embedding loss, adapted to support temporal learning using long short term memory (LSTM) networks [106]. The italicized text are phrases which contributed to the representation of each post. These phrases had similarity to the concepts suicide risk severity lexicon [37]. Likewise, table 3.5 shows the predictions of CNN model with semantic embedding loss by cumulatively learning over the user’s post.

Through concept classes we were able to explore time-variant (TvarM) and time-invariant (TinvM) nature of suicide risk. Which happens to be a case in real-world, when a patient diagnosed with a suicide risk level, after months of treatment, commits suicide. A known reasons is associated with patient’s abrupt discontinuity from clinician meetings because patient is either switch clinicians or conceal truth regarding suicide risk-related develop-

Table 3.5 Example posts from a user(u_i) and prediction from TinvM. The italicized text are phrases which contributed to the representation of the post. These phrases had similarity to the concepts in medical knowledge bases

<p>User Post: “Homie, ... Im 27 yo, ... the job is underpaying - 700 euros per month... too afraid to search for a new job. ... fuck me, I guess? ... had these <i>thoughts of suicide</i> and these <i>fears to take charge of my life</i> from like the end of a high school. 10 years same feelings of dread, same <i>thoughts of killing myself</i>.” “One day ... sudden realization ... I gonna gather determination ... roll over the bridge. And my parents, or have a nice heart attack! <i>feel trapped</i>. ... nothing gonna change. You will end up just like me, <i>roll over the bridge</i>” “No wife, no house, no car, no decent job. Every single day ... hating myself at work Im going to <i>kill myself today</i> or tomorrow. Probably ... middle of next week, but the chances are ... <i>going to sleep forever</i>”. “I dont even go to the exams... I might pass those exams... will not graduate... playing some kind of a <i>Illness joke</i> ... my poor family.”</p> <p>User-level Predicted Suicide Risk Severity: Suicide Behavior</p>
--

ments between two different suicide risk levels. With the use of concept classes we were able to suicide risk, passively, in longitudinal and cumulative way.

Edge Cases - Supportive Users: Since we are using social media posts, one unforeseen challenge is the context overlap between users who are actually showing suicidal tendencies and the users who are sharing their past experiences. The later category of users is what we call “supportive user” and considering them as another concept class (e.g. no risk) can minimize the chances of false positives or false negatives. This can be seen by the reduction in perceived risk measure shown in Figure 3.13. In the following, we enumerate the key takeaways from the ROC plots created for each concept class, as shown in Figure 3.11. Qualitative inspection of the AI model with semantic embedding loss can be seen in Table 3.6.

1. The TinvM model identified 25% more suicide attempters compared to TvarM. Too few oscillations in suicide risk severity cause TinvM to be less vulnerable to false positives than TvarM. The solid lines in ROC curves in figure 3.11 shows a significant improvement in recall for TinvM (40% TPR (True Positive Rate) at 20% FPR (False

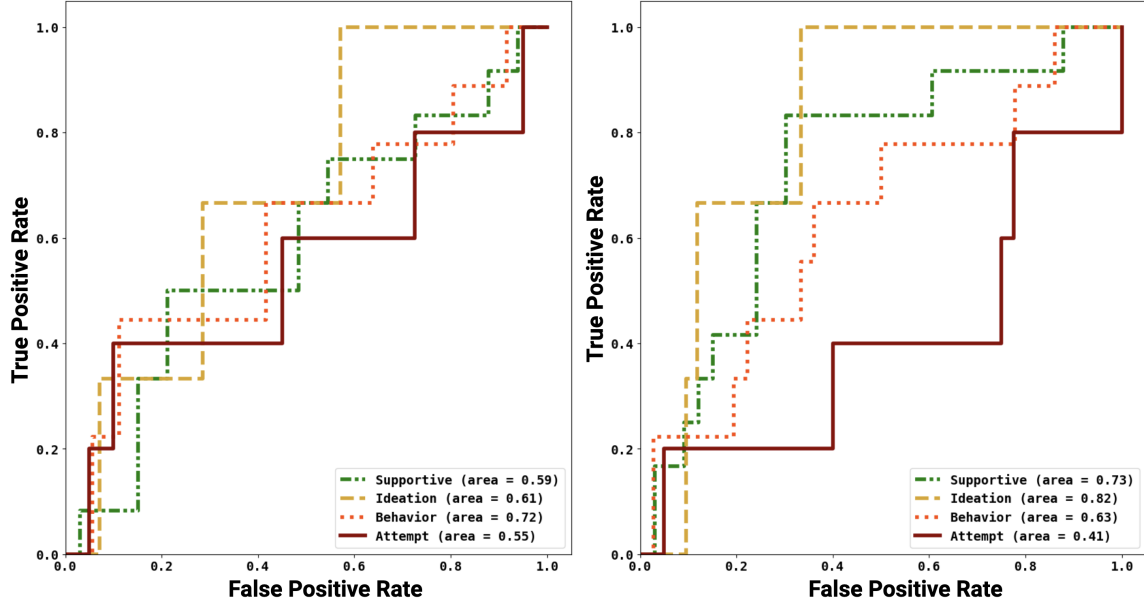


Figure 3.11 The ROC plots show the capability of either approach in detecting users with different levels of suicide risk severity based on their behavior over time on the SW subreddit. We notice that TvarM (right) effectively detects supportive and ideation users. TinvM (left) is capable of detecting behavior and attempts users. We also record that a hybrid of TinvM and TvarM is required for detecting users with suicidal behaviors.

- Positive Rate)) compared to TinvM. On the weak side, the TinvM showed a modest performance compared to a random and simple model due to difficulty separating supportive users from suicide attempters, which accounts for many false negatives.
2. One level less in suicide risk severity, the suicidal behavior users also did not show a significant change in suicide-related words (e.g. ‘loaded gun,’ ‘alcoholic parents,’ ‘slow poisoning,’ ‘scars of abuse’, etc.) causing TinvM model to identify 12.5% more users compared to TvarM. Further, TinvM predicted 20% of suicidal behavior users as supportive compared to 42% by TinvM, making it time-sensitive modeling susceptible to ignoring care for the user with severe mental illness.
 3. In contrast to users with suicide behaviors and attempt tendency, users with ideations show high oscillations in suicidal signals, making TvarM capable of correctly capturing 65% of the users, while 20% of the users were predicted with high severity levels. The false positives are due to overlap in content with behavior and attempt users be-

cause users with ideation explain behavior signs in the future tense. For example, in the following sentence, “*For not able to make anything right, getting abused, I would buy a gun and burn my brain,*” the user used a future tense to describe his ideations, developing a reasonable probability for false positives. A significant improvement of 26% in AUC for TvarM shows the low sensitivity and high specificity compared to TinvM.

4. Supportive users on Reddit account for the high false positives in the prediction of suicide assessment because of the substantial overlap in the content with users having ideation, behavior, and attempts. The time-variant methodology discreetly identifies semantic and linguistic markers which separate supportive users from users with a high risk of suicide. The use of past tense, words like "experience," "sharing," "explain," "been there," "help you," and subordinate conjunctions were consistent in temporal learning; however, their importance is overridden by suicide-related words in TinvM, leading to high false positives. From ROC curves in figure 3.11, TvarM is more specific than and less sensitive than TinvM with 20% improvement in AUC and $TPR = 1.0$ at $FPR=0.38$ compared to $TPR=1.0$ at $FPR=0.6$.

3.3.2 DESCRIPTION OF EXPLAINABLE DATA

For the purpose of annotation, we randomly picked 500 users from a set of 2181 potential suicidal users. In the annotated data, each user on an average has 31.5 posts within the time frame of 2005 to 2016. The Dataset is publicly available [here](#).

The annotated data comprises of 22% supportive users, 20% users with some suicidal indication but cannot be classified as suicidal, 34% users with suicidal ideation, 15% users with suicidal behaviors, and 9% users have made an attempt (success or fail) to commit suicide. Supportive users constitutes 1/5th of the total data size and prior studies have ignored them.

Table 3.6 Qualitative comparison of TinvM and TvarM models representative posts from users who are either supportive or showing signs of suicide ideations, behaviors or attempt. Pred.: Predictions, SW: r/SuicideWatch

TinvM Pred.	TvarM Pred.	SW Reddit Post or Comments
True Label: Support		
Ideation	Support	“Of many experiences of paranoia, anxiety, guilt, forcing me to jump into a death pithole,.... I realized how worthy I m of many things ... would be giving you my experience on this subreddit”
	Support	“I was a loner, facing increase strokes of anxiety and paranoia, that I went on driving myself into a pithole. I was missing one person who I cared the most I feel tired and careless towards anything... Guilt of not saving her”
True Label: Behavior		
Behavior	Behavior	“Please listen, I doubt myself and think committing suicide to escape my situation. Patience, I heard countless times but dying is still a bold decision for me.”
	Attempt	“This may be my last appearance. A thoughtful attempt to take my life is what I left with. I have ordered the materials required for my Suicide this evening. I also have a backup supplier in case my primary source sees through my lies and refuses sale.”
True Label: Ideation		
Behavior	Ideation	“Thank you. I actually am not on any medication. I was on Zyprexa and then Seroquel for quite a while but stopped taking the anti-psychotics about a year ago.”
	Ideation	“Anyway, Ive been thinking about seeing my shrink for a while. Maybe get back on the anti-depressants or something. Thank you though for the thoughtful post. It actually means a lot to me since I dont have many friends”
True Label: Attempt		
Attempt	Ideation	“My dad asked to step out of the house. I feared the ugly look and how disgusting I am looking. I tried therapy, talked to strangers. Everday is a torture for me. I like crafts but feel lack of energy in my self”
	Ideation	“Dark overwhelming sadness and hyperactive behavior is what describes me. I am trying to live my time to see if something changes for me”

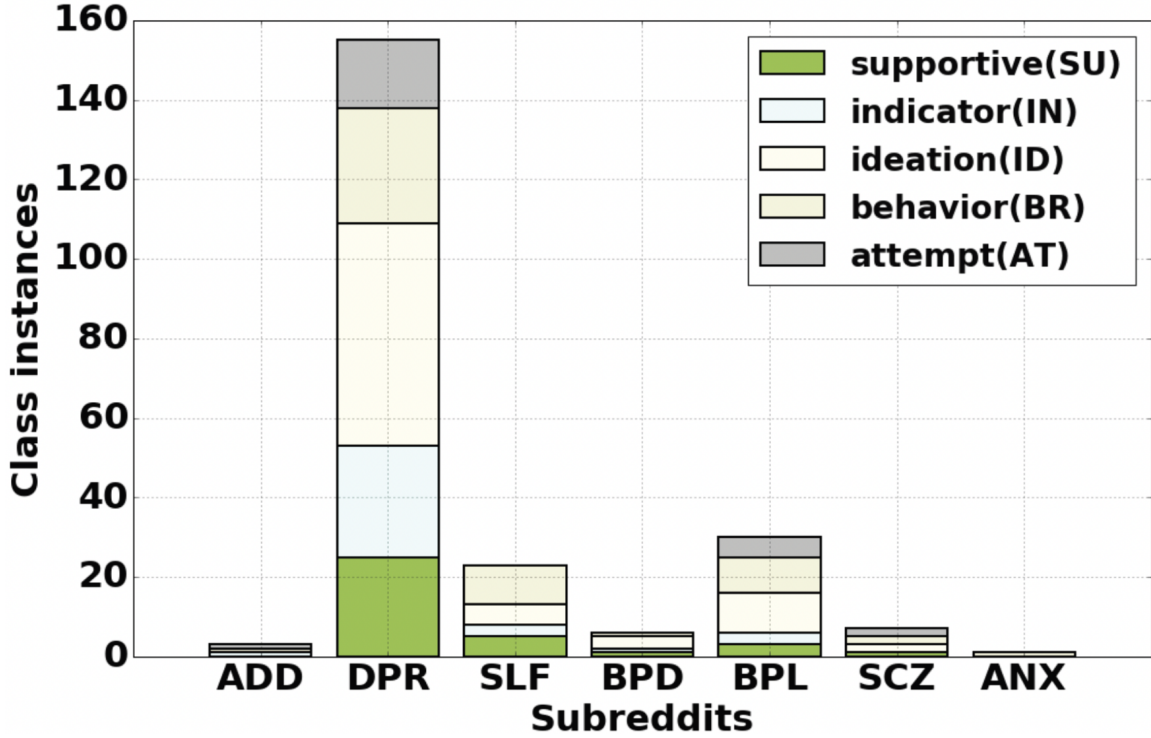


Figure 3.12 Distribution of 500 annotated users in different mental health subreddits. ADD: Addiction, DPR: Depression, SLF: Self Harm, BPD: Borderline Personality Disorder, BPL: Bipolar Disorder, SCZ: Schizophrenia, and ANX: Anxiety

Table 3.7 shows posts from redditors and their associated suicide risk severity level. To identify which mental health subreddits (except SW) contributed most to suicidality, we mapped potential suicidal Redditors to their subreddits (see Figure 3.12).

EVALUATION OF ANNOTATION

Four practicing clinical psychiatrists were involved in the annotation process. Each expert received 500 users dataset comprising of 15755 posts. We perform two annotation analysis defined for ordinal labels: **(1)** A pair-wise annotator agreement using Krippendorff metric (α) to identify the annotator with highest agreement with others, **(2)** An incremental group wise annotator agreement to find the robustness of the earlier annotator [137]. For group wise agreement, we denote a set of annotators as G with cardinality ($|G|$) range from 2 to 4. α is calculated as $1 - (\frac{D_o(A_j, S)}{D_e})$, where $D_o(A_j, S)$ is observed disagreement and D_e is

Table 3.7 Paraphrased posts from candidate suicidal redditors and associated suicide risk severity level.SU: Supportive users or no-risk users.

Always time for you to write your happy ending doesnt need to be spelled out with alcohol and Xanax.... keep an open mind	SU
Ive never really had a regular sleep schedule....no energy to hold a conversation....no focus on study....barely eat and sleep....fluffy puppy dog face	IN
Sometimes I literally cant bear to move....my depression....since I was 14....suffering rest of my life....only Death is reserved for me.	ID
Driving a sharp thing over my nerve. Extreme depression and loneliness.... worthless excuse for a life....used everything from wiring to knife blades	BR
I am going to off myself today...loaded gun to my head..determined....huge disappointment....screwed family life....breaks my heart everyday.	AT

expected disagreement. The pairwise annotator agreement is a subset of group-wise and we formally define it as:

$$D_o(A_j, S) = \frac{1}{N \cdot |S|} \sum_{i=1}^N \sum_{m \in S} |A_j^i - S_m^i|^2, S \subset G \setminus \{A_j\} \quad (3.1)$$

$$D_e = \frac{2}{N \cdot |G|(|G| - 1)} \sum_{i=1}^N \sum_{m, q \in G, m \neq q} |G_m^i - G_q^i|^2 \quad (3.2)$$

where A_j is the annotator having highest agreement in pairwise α . S is the subset of a group of annotators G that excludes A_j . G_m^i and G_q^i represents the two annotators m and q within the group G^i . i is the index over all the users in the dataset. Results of pairwise and group wise annotators agreement is in Table 3.8. We observe a substantial agreement between the annotators⁴.

Extension of the dataset to study longitudinal and transverse Suicide Risk: To assess which of the suicide risk levels are time-variant and which are time-invariant, we utilize the aforementioned dataset of 500 Reddit users. The created dataset allows Time-invariant suicide risk assesement of an individual on Reddit, ignoring time-based ordering

⁴<http://homepages.inf.ed.ac.uk/jeanc/maptask-coding-html/node23.html>

Table 3.8 (left). Pairwise annotator agreement, (right). Group wise annotator agreement. A, B, C, and D are annotators. Agreement scores are for Time-invariant modeling of suicide risk severity dataset.

	B	C	D
A	0.79	0.73	0.68
B	-	0.68	0.61
C	-	-	0.65

	B	B&C	B&C&D
A	0.79	0.70	0.69

of posts. For Time-Variant suicide risk assessment, the posts needed to be ordered with respect to time and be independently annotated. Following the annotation process highlighted in Gaur et al. [37] using modified C-SSRS labeling scheme, post-level annotation was performed by the same four psychiatrists with an inter-rater agreement of 0.88 (Table 3.9a) and a group-wise agreement of 0.76 (Table 3.9b). The annotated dataset of 448 users comprises 1170 supportive (throwaway account: 421, Non-throwaway account: 437) and uninformative(throwaway account: 115, Non-throwaway account: 197) posts. For throw-away accounts, the dataset had 37 supportive users (S), 63 users with suicide ideation (I), 23 users with suicide behavior (B), and 17 users had past experience with suicide attempt (A). User distribution within non-throwaway accounts is as follows: 85 S users, 115 I users, 76 B users, and 33 A users.

Table 3.9 Inter-rater reliability agreement using Krippendorff metric. A,B,C, and D are mental healthcare providers as annotators. The annotations provided by MHP “B” showed the highest pairwise agreement and were used to measure incremental groupwise agreement for the robustness in the annotation task. Agreement scores are for Time-variant modeling of suicide-risk severity dataset.

	B	C	D
A	0.82	0.79	0.80
B	-	0.85	0.88
C	-	-	0.83

	A	A&C	A&C&D
B	0.82	0.78	0.76

(a) Pairwise reliability agreement

(b) Groupwise reliability agreement

3.3.3 EXPLAINABILITY AS A METRIC: PERCEIVED RISK MEASURE (PRM)

It is defined to better characterize the difficulty in classifying a data item while developing a robust classifier in the face of *difficult to unambiguously* annotate datasets . It captures the intuition that if a data item is difficult for human annotators to classify unambiguously, it is unreasonable to expect a machine algorithm to do it well, or in other words, misclassifications will receive reduced penalty. On the other hand, if the human annotators are in strong agreement about a classification of a data item, then we would increase the penalty for any misclassification. This measure captures the biases in the data using disagreement among annotators. Based on this intuition, we define PRM as the ratio of disagreement between the predicted and actual outcomes summed over disagreements between the annotators multiplied by a reduction factor that reduces the penalty if the prediction matches any other annotator. We formally define it as;

$$PRM = \frac{1}{N_T} \sum_{i=1}^{N_T} \left(\frac{1 + \Delta(r'_i, r_i^o)}{1 + \sum_{m,q \in G^i, m \neq q} \Delta(G_m^i, G_q^i)} \cdot \frac{\sum_{m \in G^i} I(r'_i = G_m^i)}{|G^i|} \right) \quad (3.3)$$

Where the denominator is the disagreement between G_m^i and G_q^i annotators summed over all annotators in a group G^i . $\frac{\sum_{m \in G^i} I(r'_i = G_m^i)}{|G^i|}$ is the risk reducing factor calculated as the ratio of agreement of prediction with any of the annotators over the total number of annotators. In cases where r' disagrees with all the annotators in G , the risk reducing factor is set to 1.

Influence of Concept Classes on PRM: On analyzing models' behavior using PRM, figure 3.13 illustrates that concept phrases showed a reduction of 11.4% from SVM-Linear (the baseline) to SVM-Linear working on data with concept phrases. The CNN model provides an opportunity to learn through a new semantic embedding loss method, which reduces the uncertainty in classification. It is noticeable in figure 3.13 that CNN working on the transformed dataset with concept phrases benefits further if the model learns with semantic embedding loss wherein the predicted outcome is compared with concept classes

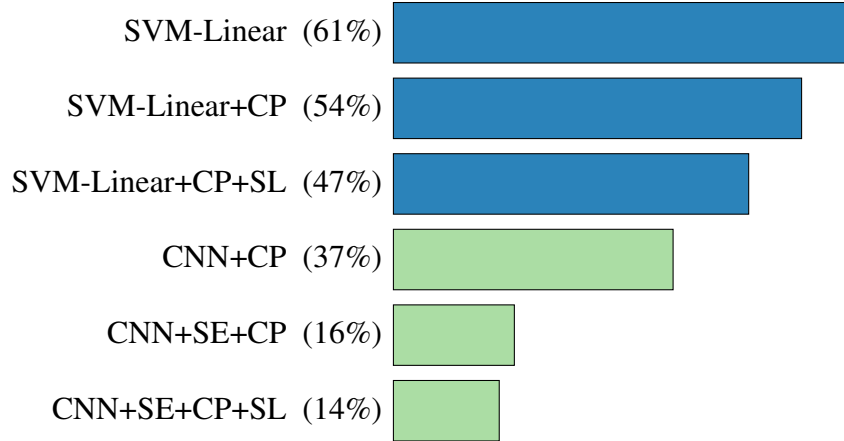


Figure 3.13 Results showing reduction in Perceived Risk Measure through an ablation of Concept Phrase (CP), Supportive Label (SL), and Semantic Embedding Loss (SE).

(↓ 57%). Suppose we extend the set of concept classes by modeling users who show supportive behavior online, and characterize it with keywords similar to the ones shown in table 3.2. In that case, the model better distinguishes between suicide risk classes and no risk class. This results in further minimizing risk (↓ 13% from SVM-Linear+CP to SVM-Linear+CP+SL; ↓ 12.5% from CNN+SE+CP to CNN+SE+CP+SL).

3.3.4 SUMMARY

We presented the notion of *concept classes* as one of the many methods of shallow knowledge infusion to abridge the gap between observational input and expected output. This chapter mainly uses external knowledge to make AI context-sensitive and user-level explainable in high consequence applications. Specifically, we show how the suicide severity lexicon can transform the observational data and outcome labels, so that model’s learning behavior can be gauged for uncertainty and context sensitivity. We introduce a perceived risk measure metric to quantify uncertainty in the presence of annotators’ agreements and disagreements among themselves and with the model’s outcome over a data sample. There are certain limitations of Shallow Infusion:

Model Interpretability: The approaches described in this chapter are large concerns with explainable data creation and semantic labels so that model’s learning can be grounded

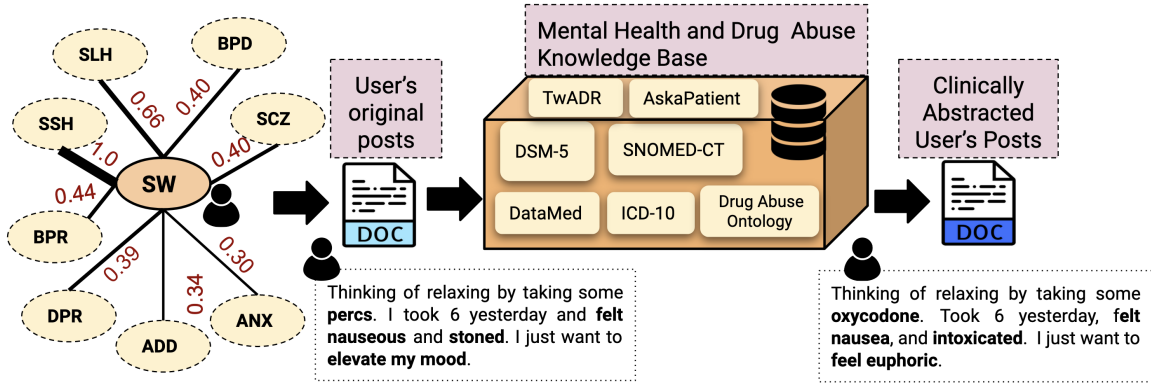


Figure 3.14 The transient posting of potential suicidal users in other subreddits, requires careful consideration to appropriately predict their suicidality. Hence, we analyze their content by harnessing their network and bringing their content if it overlaps with other users within r/SuicideWatch (SW). We found, Stop Self Harm (SSH) > Self Harm (SLH) > Bipolar (BPR) > Borderline Personality Disorder (BPD) > Schizophrenia (SCZ) > Depression (DPR) > Addiction (ADD) > Anxiety (ANX) to be most active subreddits for suicidal users. After aggregating their content, we perform MedNorm using Lexicons to generate clinically abstracted content for effective assessment.

in the domain. We haven't inspected the internal mechanics of the model and methods for knowledge infusion.

Domain Specific: Approaches under shallow knowledge infusion, which are essentially embedding-based, are highly domain specific or task-specific. Thus, their transferability is a challenge because of the rigid parametric knowledge learned by the model. Further, the concept classes required to make the model explainable and adaptive in a domain can hurt transferability across multiple domains. It is because not all domains have concept classes.

Modeling Uncertainty: We discussed perceived risk measure as a metric to assess uncertainty in predictions; however, we did not enforce "uncertainty handling" within the model's learning behavior. We also touched upon the gambler's loss function as a method to model uncertainty, but its statistical nature removes many samples [132]. Thus we need an approach at the intersection of semantic and statistical.

Cost and User-explainability trade-off: There is a trade-off between the cost involved in creating explainable data and the need for user-level explainability in the application. Recent datasets have spent thousands of dollars on annotation, but still model shows a vast gap between its prediction and human-level performance. So, to employ shallow infusion methods, the trade-off requires a nudge.

For clarity, we restricted the application of shallow infusion to mental health, particularly suicide risk classification. Figure 3.14, shows an architecture to scale the explainable data creation approach described in this chapter using a wide-variety of domain-specific knowledge sources. Shallow knowledge infusion is applicable in various other applications, such as sub-event detection in dynamic tweet streams [24], asking better follow-up questions in mental health⁵, explainable clustering to study patient’s discourse in social media and clinical notes [55], infusing cognitive theories [138], crisis informatics [139], and others. In the next chapter, we will focus on model interpretability and keep the model capable of preserving context and sensible towards preventing unsafe prediction or natural language generation.

⁵<https://tinyurl.com/IIIT-AIISC-PRIMATE>

CHAPTER 4

SEMI-DEEP INFUSION

Compared to **Shallow Infusion**, **Semi-Deep Infusion** concerns with opening the black-box of AI system using knowledge sources. This chapter will provide a detailed grounding of model interpretability highlighting the state of the art methods that promises interpretable AI, the limitations of these methods, and how knowledge infusion in AI can help make model interpretable without sacrificing uncertainty handling, context sensitivity, and user-level explainability.

Significance

Semi-Deep infusion retains the representational richness of knowledge representation and allows use of a variety of knowledge in the infusion process. It develops strategies to augment knowledge representation with latent representations to make statistical model interpretable. It also introduce methods wherein the model learns to balance between knowledge and data.

We will discuss two specific application areas: **(a)** Classification of mental health conditions of users on Reddit using Diagnostic Statistical Manual for Mental Health Disorders. This would expand on the *concept classes* discussed in Chapter 3. There is prior research on the extraction of mental health-related information, including symptoms, diagnosis, and treatments from social media; however, our approach can additionally provide actionable information to clinicians about the mental health of a patient in diagnostic terms for web-based intervention. **(b)** Conceptual flow-based question generation for conversational information seeking (CIS) using knowledge graphs as source of meta-information. CIS is

a relatively new research area within conversational AI that attempts to seek information from end-users in order to understand and satisfy users' needs. If realized, such a system has far-reaching benefits in the real world; for example, a CIS system can assist clinicians in pre-screening or triaging patients in healthcare. A key open sub-problem in CIS that remains unaddressed in the literature is generating Information Seeking Questions (ISQs) based on a short initial query from the enduser having an ill-defined context. To address this open problem, we propose a novel approach for generating ISQs from just a short user query, given a large text corpus relevant to the user query.

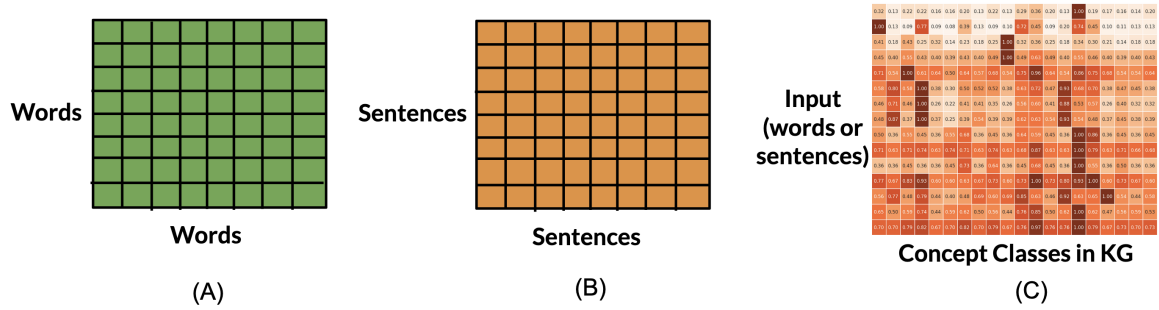


Figure 4.1 (A) & (B): An illustration of self-attention matrices computed in current attention-based transformer models and autoencoders. (C) The cross-attention matrix is what we desire and seek to achieve using autoencoders. We mainly use autoencoders as they are proven to be good representation generators and modulators. Credit: Image adapted from a Presentation

4.1 BENEFITS OF Semi-Deep Infusion

The methods concerns with innovation in loss functions and optimization functions for knowledge infusion.

Optimization Function: DL model learns by computing correlation between words or sentences which is analogous to self-attention matrices [140]. This chapter is interested in leveraging self-attention to learn a cross-correlation matrix between input words or sentences and concept classes in a knowledge source (KS). The matrix in figure 4.1(c) is the target matrix we want the model to learn. The intuition behind this

is to achieve interpretability in the model. One can use the matrix, visualize it using T-SNE [12] or can study the mapping scores to confirm whether the model was able to align words/sentences in the input with correct concept classes. From Chapter 3, concept classes give an understandable representation of the domain (e.g., the mental health domain’s concept classes are disorders mentioned in DSM-5), and mapping of independent words/sentences in the input allows us to judge model’s interpretation of the input. We will discuss a Semantic Encoding and Decoding-based optimization scheme that will enable the model to generate contextual feature vectors irrespective of the domain, as long as KS supports the scheme. Additionally, this process enables zero-shot learning using KS [10].

Constraints-based Loss Function: Such loss functions became important from the task of summarization in natural language processing. The constraints are placed to pick a sentence of desired characteristics for summary generation. Integer linear programming, inductive logic programming, planning, and others are areas where constraints-based loss functions have proved to be useful. In this chapter, we will describe a novel use-case of constraints-based loss function, which is to enforce a logical order and maintain semantic relations in a conversational agent tasked to full fill information needs of a user by asking information seeking questions in a conceptual flow. We will see how such a loss function in conjunction with KGs can improve the generation quality of a conversational agent and be safe when use in sensitive areas, like mental healthcare.

Model Interpretability: Figure 4.2 illustrate the complete pipeline of semi-deep infusion in achieving interpretability. The top-left part of the figure describe the working of a traditional ML/DL model who ends up giving a wrong prediction because it fails to capture contextual cues responsible for correct prediction. If we attempt to reason over the model, we won’t be able to make meaningful inferencing because of the sta-

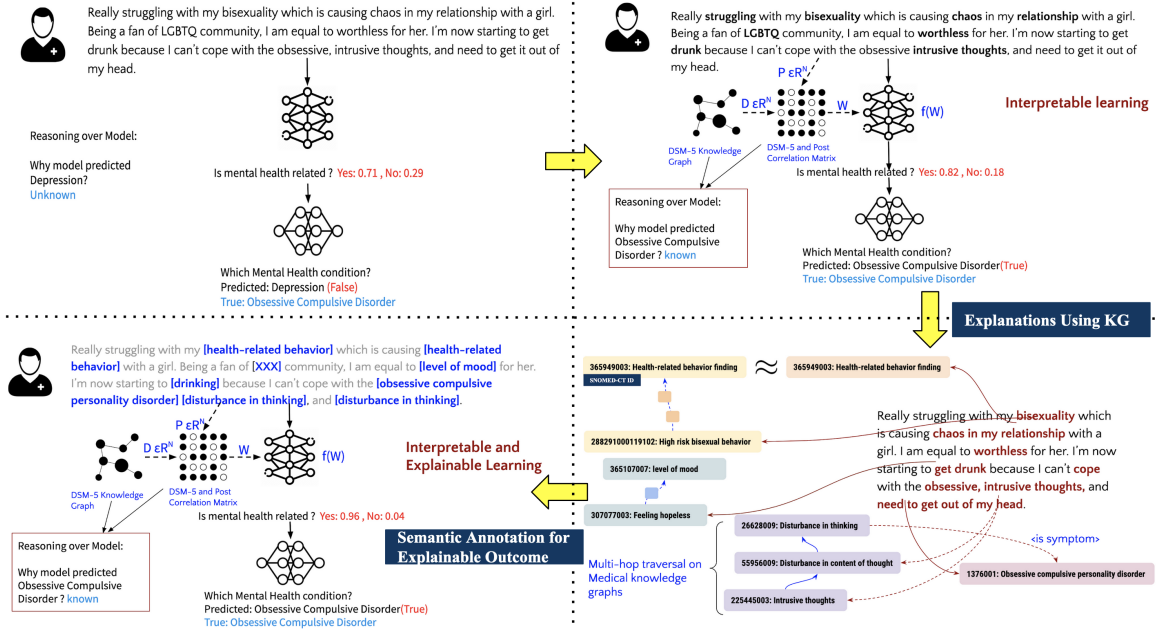


Figure 4.2 An overall pipeline illustrating the benefit of Semi-Deep Infusion in making ML/DL explainable and interpretable.

tistical feature vectors that provide inaccurate representation of real-world. However, if we enforce transformation of the input text using concepts (or concept classes) in KG within the model’s functional part (optimization function, loss function, or activation function), then we achieve model interpretability. It is because the resultant feature vector would be dominated by the phrases that mapped to a set of concepts (or concept classes) in KGs with high scores. It can be visualized as a heat-map shown in figure 4.1(c). This process yield two benefits: (a) reduction in mis-classification because the feature space is less varied and contextual, and (b) the reasoning over the model is possible.

User-level Explainability: A semi-deep knowledge infused model provide user-level explainability querying the KG using the concept (or concept class) and the word or phrase in the input having a maximum correlation with the concept (or concept class). For example, it can be seen in Figure 4.2 (bottom-right), which illustrates the multi-hop traversal in a KG using words or phrases identified as important by a semi-deep knowledge infused model. An additional benefit from multi-hop or single-hop traver-

sal is retrieving information that might overlap or inform the target label the model is supposed to predict. As a result, through user-level explainability, one can associate the traversed part of KG with the target label to measure the correctness of the model.

Context Sensitivity: A semi-deep knowledge infused model is context-sensitive. It semantically annotates the input context while learning using a cross-correlation matrix between words/sentences in input and concept (or concept class) in KG. It can be defined as either identifying *concept phrases* or substituting them with abstract concepts/categories. For instance, in bottom-left section of the figure 4.2, the words “bisexuality” and “relationship” are substituted with *Health-related behavior*. Such implicit transformation¹ of input text makes the model context-sensitive.

4.2 Semi-Deep Infusion

We define the second category of knowledge infusion, i.e., **Semi-Deep Infusion** as a paradigm that gauges the learning of a deep net and resolves the impedance mismatch by adding structural (e.g., dependency relations between words in a sentence) or symbolic (attention probability or constraints satisfaction) knowledge. Such an approach has been effective in a task-specific problem where the model is unable to learn complex representative features from the text. We categorize different perspectives of **Semi-Deep Infusion** of knowledge in the deep neural networks outlined for various natural language processing/understanding tasks (e.g. event detection, user classification, relationship extraction, reading comprehension, etc.) (see Table 4.1).

Teacher Forcing: In a deep learning framework comprising of an autoencoder, the capability of a decoder is enhanced through teacher forcing. In this procedure, the target labels (non-binary rather structured sentences) are fed word by word while training the decoder part of the autoencoder. The encoder provides the vectorized

¹it was explicit in shallow infusion

Table 4.1 Existing methods and approach that are classified based on whether they provide user-level explainable and knowledge-based interpretability. DLMs: Deep Language Models.

Method	Approach	Explainability	Knowledge-based Interpretability
Fine-tuning [74]	Any DLMs	User-Ex (✗), Sys-Ex (✓)	✗
Teacher Forcing [141]	Any DLMs	User-Ex (✗), Sys-Ex (✓)	✗
Professor Forcing [142]	Any DLMs	User-Ex (✗), Sys-Ex (✓)	✗
LSTMs	KG-LSTMs [143]	User-Ex (✗), Sys-Ex (✓)	✓
	KB-LSTMs [144]	User-Ex (✗), Sys-Ex (✓)	✓
GANs	KG-GANs [145]	User-Ex (✗), Sys-Ex (✗)	✓
	Self-Attention [146] [147]	User-Ex (✗), Sys-Ex (✓)	✓
	KG-Guided Attention [148]	User-Ex (✗), Sys-Ex (✓)	✓
Attention	CAGE [149]	User-Ex (✗), Sys-Ex (✓)	✗
	ERNIE v1.0 [150]	User-Ex (✗), Sys-Ex (✓)	✗
	ERNIE v2.0 [151]	User-Ex (✗), Sys-Ex (✓)	✗
	ERNIE v3.0 [152]	User-Ex (✗), Sys-Ex (✓)	✗
	Human Parity [153]	User-Ex (✓), Sys-Ex (✓)	✗
	K-BERT [83]	User-Ex (✗), Sys-Ex (✓)	✗
	K-Adapter [154]	User-Ex (✗), Sys-Ex (✓)	✗
	SenseBERT [155]	User-Ex (✗), Sys-Ex (✓)	✗
	KI-BERT [9]	User-Ex (✗), Sys-Ex (✓)	✗
	Integrated Gradients [156]	-	User-Ex (✗), Sys-Ex (✓)
Integrated Hessians [157]	-	User-Ex (✗), Sys-Ex (✓)	✗
	-	User-Ex (✗), Sys-Ex (✓)	✗
Autoencoders	Semantic Encoding and Decoding [10]	User-Ex (✓), Sys-Ex (✓)	✓
Reinforcement Learning	Deep Reinforcement Learning Methods with GLUE-based Rewards [118]	User-Ex (✗), Sys-Ex (✓)	✗
Multi-relational Reinforcement Learning	Relational Functional Policy Gradient Methods [77]	User-Ex (✓), Sys-Ex (✓)	✓
	Combining Search with Value Iteration, Policy Gradient, and Monte-Carlo Tree Search [112]	User-Ex (✗), Sys-Ex (✓)	✓
Reinforcement			
Search and Learning	ISEEQ [31]	User-Ex (✓), Sys-Ex (✓)	✓
	AlphaGo [158]	User-Ex (✗), Sys-Ex (✓)	✗
	PKiL [17]	User-Ex (✓), Sys-Ex (✓)	✓

representation of the input on which the decoder tries to learn. The procedure was first discussed by Williams et al. and has shown improvement in machine translation, entity extraction, and negation detection tasks [141] [142]. Understanding the procedure of teacher force, we identified two critical issues: (1) the representation provided by the encoder is not gauged in the teacher forcing method, and (2) the model memorizes the input patterns and is challenging to perform transfer learning

with the trained model. A teacher forced model can learn the correct representation of the input through following methods:

- **Redundancy:** In this learning process, the model is monitored for information loss through backpropagation and is replenished through replicating the input to the layers. Methods like skip connections or highway connections follow such a method [159].
- **Curriculum Learning:** A variation of forced learning is to introduce outputs generated from prior time steps during training to encourage the model to correct its own mistakes [160].

In the teacher forcing paradigm, during inference, the conditioning context may diverge during training when ground truth labels are given as input. As the encoder acts as a generator and the decoder behaves like a discriminator, their independent functioning affects model performance. Further, incorporating knowledge is on the decoder side, independent of the encoder. Hence, it is challenging to quantify the loss of information incurred on the encoder side. Our proposed approach on Deep Infusion regulates (1) where in a model, the latent weights are wrongly enforced and (2) How to adjust the weights leveraging external human-curated graphical knowledge sources.

Neural Attention Models (NAMs): Attention models highlight important features for pattern recognition/classification based on a hierarchical architecture of the content. The manipulation of attentional focus effectively solves real-world problems involving massive data [161]. On the other hand, some applications demonstrate the limitation of attentional manipulation in a set of problems such as sentiment (mis)classification and suicide risk [37], where feature presence is inherently ambiguous, just as in the radicalization problem. For example, in the suicide risk prediction task, references to the suicide-related terminology appear in the social media posts of both

victims and supportive listeners, and the existing NAMs fail to capture semantic relations between terms to help differentiate the suicidal from a supportive user. To overcome such limitations in a sentiment classification task [162], have augmented sentiment scores in the feature set for enhancing the learned representation and modified the loss function to respond to the values of the sentiment score during learning. However, Sheth et al. have pointed out the importance of using domain-specific knowledge, especially in cases where the problem is complex [13]. In an empirical study, Bian et al. showed the effectiveness of combining richer semantics from domain knowledge with morphological and syntactic knowledge in the text by modeling knowledge assistance as an auxiliary task that regularizes learning of the main objective in a deep neural network [163].

Professor Forcing and Learnable Knowledge Constraints: Professor forcing forms an architecture where the encoder (generator) competes with the decoder (discriminator) in improving the outcome, thus forming an Adversarial Network. Further, the improvement in the learning occurs by acting as a posterior regularizer and allowing the possibility of including rich structured domain knowledge. However, if knowledge constraints need to be infused in professor forcing, they need to be done apriori and not iteratively while learning. A recent study from Hu et al. focuses on infusing the knowledge as constraints in such an adversarial network by optimizing the Kullback-Leibler (KL) divergence [164]. However, the knowledge gathered for infusion is part of the dataset and does not exploit a human-curated Knowledge Graph. The study does relate to our objective by monitoring KL divergence. However, it does not provide an appropriate method for adding the relevant knowledge quantified from the KL score. However, in our Deep Infusion paradigm [?], we aim to define the quantification and inclusion of relevant knowledge to deep models to minimize the learning time and false alarm rate.

Graph Neural Network: Graph Neural Network is a type of neural network which directly operates on the graph structure [165]. A typical application of GNN is node classification. Essentially, every node in the graph is associated with a label, and we want to predict the nodes' label without ground-truth. In this process, the model generates an importance score for each node, and the connection weights form the weights of the relationship between the nodes. Marino et al. utilize knowledge graphs for multi-label classification of images using a KG [166]. In this and a similar study by Wang et al., the GNN framework can be seen as leveraging the structural property of the KG and quantifying itself using the input data [167]. However, the framework is restricted to the labels in the input dataset and their inter-relationships. Further, the GNN does not exploit the structural property and taxonomic relationships of the KG in identifying the relevant knowledge that can be applied to the learning of the neural network. Further, the hidden nodes in GNN are not the abstractions corresponding to a stratified knowledge in a KG; thus, the relationships between the labels are not well contextualized.

Neural Language Models: NLMs are a category of neural networks capable of learning sequential dependencies in a sentence, and preserve such information while learning a representation. In particular, LSTM (Long Short Term Memory) networks have emerged from the failure of RNNs (Recurrent Neural Networks) in remembering long-term information [168]. Concerning the loss of contextual information while learning, Cho et al, proposed a context feed forward LSTM architecture in which context is learned by the previous layer merged with forgetting and modulation gates of the next layer [169]. However, if erroneous contextual information is learned in previous layers, it is difficult to correct [170], which is a problem magnified by noisy data and content sparsity (e.g. Twitter, Reddit, Blogs). As the inclusion of structured knowledge (e.g., Knowledge Graphs) in deep learning, improves information retrieval [171], prior research has shown the significance of knowledge in the pur-

suit of improving NLMs, such as in commonsense reasoning [172]. The transformer NLMs such as BERT (including its variants BioBert and SciBERT), are still data dependent [173]. BERT has been utilized in hybrid frameworks such as in the creation of sense embeddings using BabelNet and NASARI [174]. Liu et al. proposed K-BERT, that enriches the representations by injecting the triples from KGs into the sentence [83]. As this incorporation of knowledge for BERT takes place in the form of attention, we consider the K-BERT as semi-deep infusion [175]. Similarly, ERNIE incorporated external knowledge to capture lexical, syntactic, and semantic information, enriching BERT [150].

Tree LSTMs: LSTMs are sequential models, whereas the sentences in the input corpus follow a grammatical tree structure (dependency or constituency). Hence, it is important to learn the contextual representation of the input following the same tree structure. Tree LSTMs replaces the nodes in the graph with LSTMs cells and vector representation of the words/phrases is given as input [176]. This model takes into account the structural (syntactic) property of the input, but the domain knowledge is ignored. A recent study from Yang et al. utilizes external knowledge bases (e.g. WordNet, NELL) to improve the performance of BiLSTMs by minimizing task-specific feature engineering [148]. Particularly, the study focused on improving entity and event extraction. Knowledge-based LSTM proposed in the study comprises an attention mechanism that acts as a sentinel to guide the model in deciding whether to use external knowledge and adaptively decide the level of abstractness in the information. Though the proposed architecture uses an external knowledge base as a separate component for each LSTM cell, it is uncertain how much of the external knowledge needs to be incorporated and to what level of abstraction the traversing of the knowledge base needs to be done to fulfill the information loss in the learning process.

Knowledge-based Neural Networks: Yi et al. introduced a knowledge-based, recurrent attention neural network (KB-RANN) improve model generalization by modifying the attention mechanism using domain knowledge. However, their domain knowledge is statistically derivable from the input data itself and is analogous to merely learning an interpolation function over the existing data. Dugas et al. proposed a modification in the neural network by adopting Lipschitz functions for its activation function [177]. Hu et al. proposed a combination of deep neural networks with logic rules by employing knowledge distillation procedure of transferring the learned tacit knowledge from larger neural network, to the weights of the smaller neural network in data-limited settings [164] [178].

These studies for incorporating knowledge in a deep learning framework have not explored declarative knowledge structures in the form of KGs (e.g., DBpedia, BabelNet, UMLS, Wikidata). However, Casteleiro et al. recently showed how the Cardiovascular Disease Ontology (CDO) provided context and reduced ambiguity, improving performance on a synonym detection task [179]. Shen et al. employed embeddings of entities in a KG, derived through Bi-LSTMs, to enhance the efficacy of NAMs [180]. Sarker et al. presented a conceptual framework for explaining artificial neural networks' classification behavior using background knowledge on the semantic web [181]. Makni et al. explained a deep learning approach to learn RDFS (Resource Description Framework Schema) rules from both synthetic and real-world semantic web data. They also claim their approach improves the noise-tolerance capabilities of RDFS reasoning [182]. All of the frameworks in the above subsections utilized external knowledge before or after the representation has been generated by NAMs, rather than within the deep neural network as in our approach [175]. We propose a learning framework that infuses domain knowledge within the latent layers of neural networks for modeling.

Figure 4.3

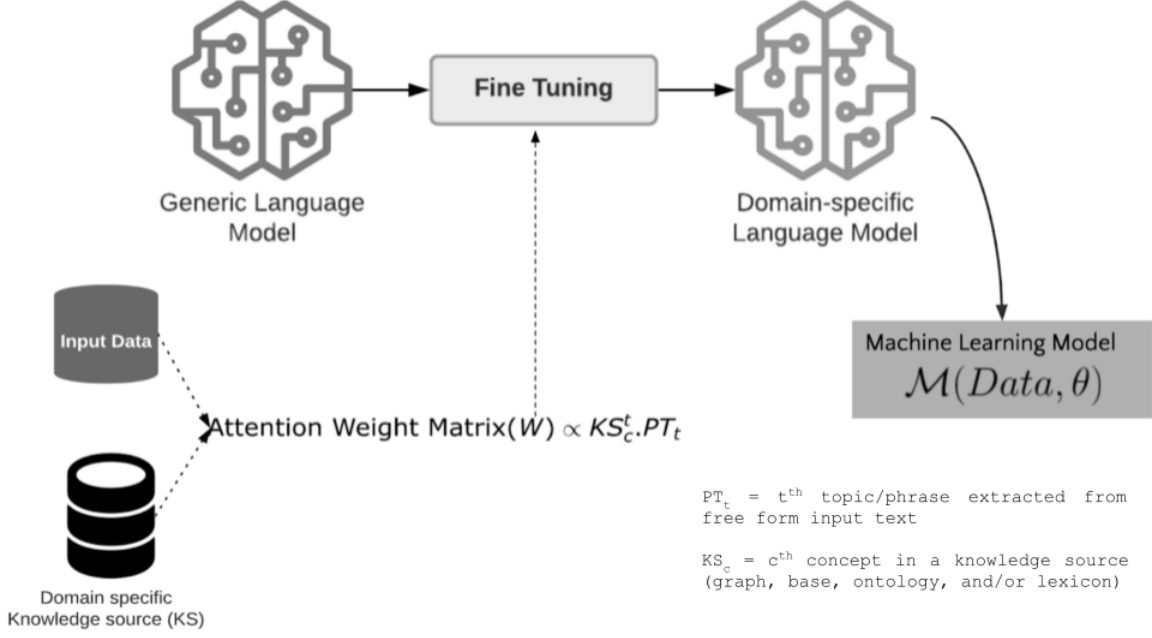


Figure 4.3 A general architecture of **Semi-Deep Infusion**. KS_c^t : c^{th} concept in a knowledge source (KS) that is similar to a t^{th} topic or phrase extracted from free form input text. **Semi-Deep Infusion** concerns with making AI model that learns a weight matrix which intersects with input observational data and expert knowledge.

4.3 SEMANTIC ENCODING AND DECODING OPTIMIZATION (SEDO)

In this section, we explain our semantic weighting algorithm, called SEDO, and its role in the DSM-5 multi-class classification, as illustrated in the Figure 4.4.

SEDO is an approach for obtaining a *discriminative weight* matrix between the DSM-5 lexicon and Reddit word embedding space after optimization utilizing the Sylvester equation [183]. Although the Sylvester equation has been used in computer vision within the context of ZSL [184], its utilization in creating a *discriminative weight* matrix between unstructured(e.g. Reddit) and structured data (DSM-5 Lexicon) has not been investigated. SEDO requires: (1) embedding space for each category in the DSM-5 lexicon, and (2) embedding space of each word in Word2Vec vocabulary created from Reddit data. Creating a link between embedding spaces of DSM-5 categories and Reddit data requires an energy function that will semantically maximize the number of matches while reducing

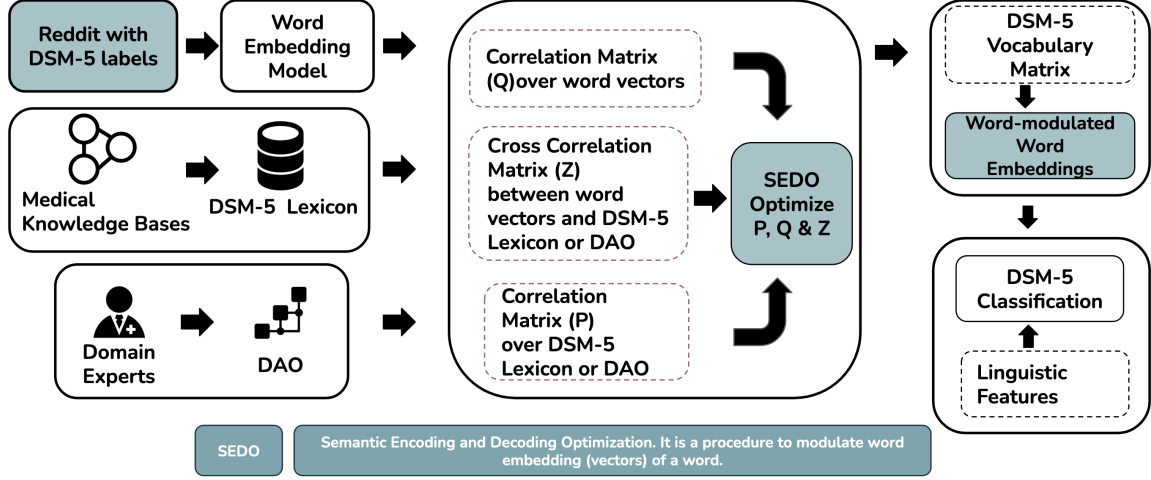


Figure 4.4 Proposed approach to DSM-5 classification using SEDO based word-vector modulation together with Horizontal Linguistic Features (HLF), Vertical Linguistic Features (VLF) and Fine-grained features (FGF). HLF includes, *number of definite articles, number of words per Reddit post, first person pronouns, number of pronouns, and subordinate conjunction*. VLF includes, *number of POS tags, similarity between Reddit posts made by a user, intra-subreddit similarity, and inter-subreddit similarity*. FGF includes, *sentiment scores, emotion scores, and readability scores*. These **Linguistic Features** are specific to mental health for which SEDO was used. Details of these features are presented here [10].

the number of mismatches. As we utilize the methodology of min-max separability [185] that provides precise differentiation between categories, we model our problem as the minimization of the semantic mismatch. SEDO formulate the function $E(\mathbf{R}, \mathbf{D})$ as minimizing the Frobenius norm² of difference between Reddit and DSM-5 embedding spaces (Equation 4.1).

$$E(\mathbf{R}, \mathbf{D}) = \min_W \{ \|\mathbf{R} - \mathbf{W}^T \mathbf{D}\|_F^2 + \delta \|\mathbf{W} \mathbf{R} - \mathbf{D}\|_F^2 \} \quad (4.1)$$

where \mathbf{R} represents the Reddit word embedding space, \mathbf{D} the DSM-5 embedding space, and \mathbf{W} the weight matrix to be minimized.

As we are mapping the Reddit (unstructured) embedding space to the DSM-5 (structured) embedding space, we call this process as decoding, and from DSM-5 to Reddit data as encoding. In Equation (3), the part before the “+” represents the encoding of DSM-5 cat-

²<http://mathworld.wolfram.com/FrobeniusNorm.html>

egories to Reddit data embedding space, while the part after “+” represents the decoding of Reddit data to DSM-5 categories. Furthermore, Equation (3) is a convex function; hence, we can expect a global optimal solution. Differentiating the Equation (3) with respect to "W" for minimization, involves following properties: $\text{Tr}(\mathbf{W}^T \mathbf{D}) = \text{Tr}(\mathbf{D}^T \mathbf{W})$ (cyclic property of trace)³ and $\text{Tr}(\mathbf{R}) = \text{Tr}(\mathbf{R}^T)$. A positive, symmetric and quasiseparable⁴ matrix show such properties. Hence, Equation 4.1 is transformed to

$$E(R, D) = \min_w \{ ||R^T - D^T W||_F^2 + \delta ||WR - D||_F^2 \} \quad (4.2)$$

$$\frac{d(E(R, D))}{d(W)} = -2(D)(R^T - D^T W) + 2\delta(WR - D)(R^T) \quad (4.3)$$

Setting LHS of the Equation 4.3 to zero $\frac{d(E(R, D))}{d(W)} = \mathbf{0}$ will result in an equation that is solvable using Sylvester equation. δ is a parameter for regularization during the optimization phase.

$$-DR^T + DD^T W + \delta WRR^T - \delta DR^T = 0 \quad (4.4)$$

$$(DD^T)W + W(\delta RR^T) = (1 + \delta)DR^T; 0 < \delta < 1 \quad (4.5)$$

Equation 4.5 represents the Sylvester equation form: $\mathbf{P}\mathbf{X} + \mathbf{X}\mathbf{Q} = \mathbf{Z}$ where \mathbf{P} is \mathbf{DD}^T and \mathbf{Q} is \mathbf{RR}^T , which represents self-correlation between DSM-5 and Reddit embedding spaces respectively, and \mathbf{Z} is \mathbf{DR}^T represent cross-correlation between DSM-5 and Reddit embeddings. The δ controls the knowledge infusion. A decrease in δ increases the infusion of knowledge in DSM-5 (D) to balance the left hand side of Equation 4.5 with right hand side. Figure 4.5 demonstrate the effect of δ .

DSM-5 Embedding Space: Each category in the DSM-5 Lexicon is represented by a set of concepts. These concepts can be U, B, or T. We created embedding of each category of

³<http://www2.math.ou.edu/~dmccullough/teaching/slides/maa2010.pdf>

⁴<https://goo.gl/mcgvcZ>

Really struggling with my bisexuality which is causing chaos in my relationship with a girl. Being a fan of LGBTQ community, I am equal to worthless for her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts, and need to get out of my head.
 Don't want to live anymore. Sexually assault, ignorant family members and my never ending loneliness brights up my path to death.
 I do have a potential to live a decent life but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights to days. I am developing insomnia because of my restlessness. I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

$\delta = 1.0$ (No Knowledge)

Really struggling with my bisexuality which is causing chaos in my relationship with a girl. Being a fan of LGBTQ community, I am equal to worthless for her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts, and need to get out of my head.
 Don't want to live anymore. Sexually assault, ignorant family members and my never ending loneliness brights up my path to death.
 I do have a potential to live a decent life but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights to days. I am developing insomnia because of my restlessness. I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

$\delta = 0.84$ (16% knowledge)

Really struggling with my bisexuality which is causing chaos in my relationship with a girl. Being a fan of LGBTQ community, I am equal to worthless for her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts, and need to get out of my head.
 Don't want to live anymore. Sexually assault, ignorant family members and my never ending loneliness brights up my path to death.
 I do have a potential to live a decent life but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights to days. I am developing insomnia because of my restlessness. I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

$\delta = 0.66$ (34% knowledge)

Really struggling with my bisexuality which is causing chaos in my relationship with a girl. Being a fan of LGBTQ community, I am equal to worthless for her. I'm now starting to get drunk because I can't cope with the obsessive, intrusive thoughts, and need to get out of my head.
 Don't want to live anymore. Sexually assault, ignorant family members and my never ending loneliness brights up my path to death.
 I do have a potential to live a decent life but not with people who abandon me. Hopelessness and feelings of betrayal have turned my nights to days. I am developing insomnia because of my restlessness. I just can't take it anymore. Been abandoned yet again by someone I cared about. I've been diagnosed with borderline for a while, and I'm just going to isolate myself and sleep forever.

$\delta = 0.71$ (29% knowledge)

Figure 4.5 δ controls the amount of knowledge infusion in SEDO for acceptable classification mental health disorder given a user's profile in the form of posts. Upon 34% knowledge infusion the model's recommendations matched five MHPs provided labels 84% of the times [10].

DSM-5 using trained Word2Vec model on Reddit corpus. We performed summation over concept vectors to 300 dimensions embedding for each DSM-5 Lexicon. Hence, DSM-5 embedding space is of dimensions 20 X 300. Self-correlation of DSM-5 embeddings (\mathbf{DD}^T) is performed using Pearson Correlation and creates a matrix of dimensions 20 X 20. Similarly, self-correlation of Reddit word-embedding space (\mathbf{RR}^T) creates a matrix of dimension 12808 X 12808. Cross-correlation between \mathbf{RR}^T and \mathbf{DD}^T creates a matrix of dimensions 20 X 12808.

Evaluation: The assessment of SEDO establishes it as a method that can utilize social media behaviors to estimate psychiatric diagnostic categories in a user. For the sake of simplicity, we replaced Random Forest's weighting function using SEDO, thus allowing semi-deep infusion. Likewise, with CNN, we employed CNN autoencoder with an optimization function defined using SEDO. Figure 4.6 demonstrates significant reduction in false alarms from SEDO. With knowledge infusion through SEDO, not only the feature set

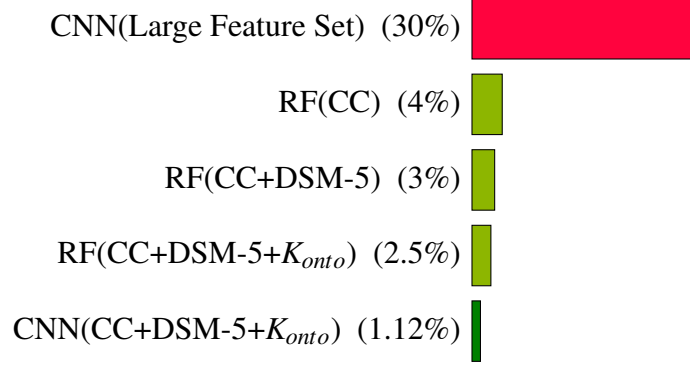


Figure 4.6 Results showing reduction in False Alarms by replacing statistical features with knowledge and its subsequent ablations of various form of knowledge. CC: Concept Classes, DSM-5: Diagnostic Statistical Manual for Mental Health Disorders, a knowledge source for mental healthcare practitioners, K_{onto} : Drug Abuse Ontology, a domain-specific ontology for substance use and addictive disorders, RF: Random Forest, CNN: Convolutional Neural Network. **Model(Features or Knowledge)**: It represents that either statistical features or concepts from knowledge sources are given as input to the model.

was reduced, but it also made the ML/DL model capable of working with different forms of knowledge.

4.4 ISEEQ FOR CONVERSATIONAL INFORMATION SEEKING

Traditional dialog agents in conversational information seeking have repeatedly focused on entities in the user query [186] [187]. Consequently, the generated questions are redundant and lack diversity, losing user engagement.

Further, the multi-turn conversations to support user engagement often results in irrelevant question generation by the agent. For instance, in Figure 4.7, a traditional dialog agent generated a question, “Do you want to know about economics,” which seems relevant to the user query; however, the user did not find it suitable. This is because economics is a vast subject, and the user is only interested in crucial economics concepts related to the gross domestic product (gdp), inflation, and employment. Hence, capturing the context and adapting the question generation to context is essential. Current conversational agents lack curiosity in question generation, which is critical for cohesive response [61] [62]. Curiosity in a conversational agent is defined by the agent’s capability to diversify the user query with

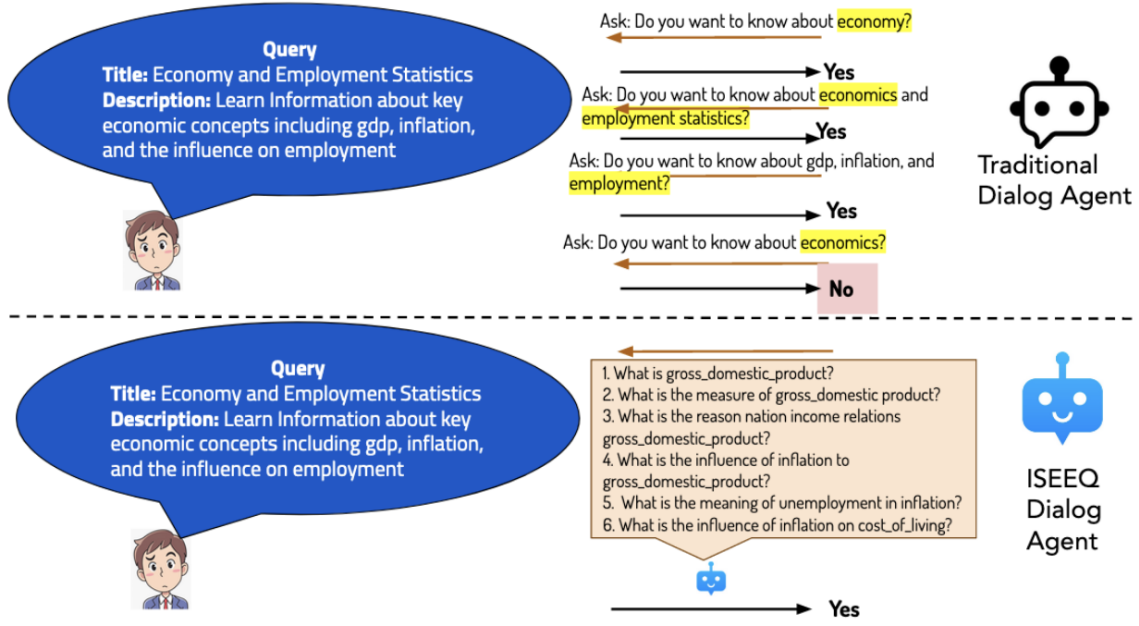


Figure 4.7 ISEEQ's one-shot procedural question generation

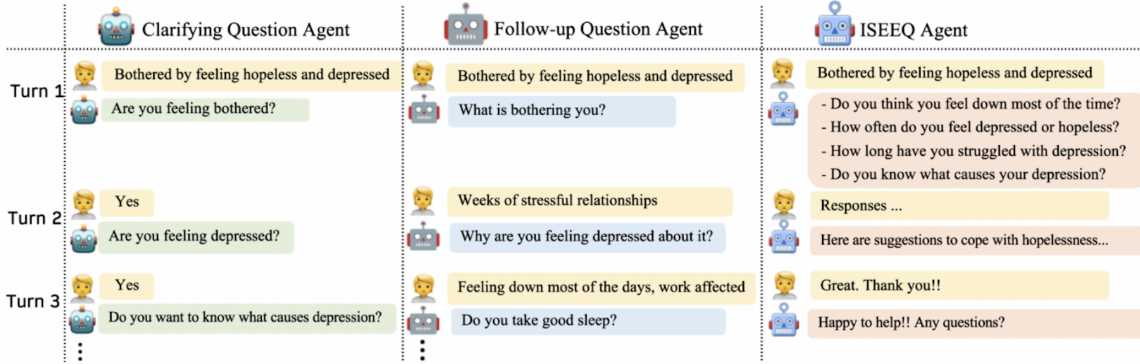


Figure 4.8 ISEEQ's generation of information seeking questions reduces the number of turns involved in providing the response needed by the end-user. Thus improving user engagement.

triples (explain it, if required) that are semantically related to entities in the query. Further, the agent retrieves meta-information using the diversified query for question generation. These properties sums up ISEEQ, an Information SEEking Question generation agent that generates a series of information seeking questions to gather the context of the user's query. *Another feature in conversational agent brought in ISEEQ is to force a conceptual flow while generating questions, defined by semantic relations and logical coherence between the generated questions* [31].

The problem of generating information seeking questions (ISQs) given an initial user’s information seeking-type (IS) query, in which ISEEQ specializes, has not been addressed in the literature so far. Apart from the general context of economics, illustrated in figure 4.7, consider the user IS query in mental health: “Bothered by feeling down or depressed. Need advice”. ISEEQ generated ISQs are: “How often do you feel depressed or hopeless?”, “How long have you struggled with depression?”, and others, which can be used either by the CIS or the healthcare provider to generate an appropriate response to the user’s needs. Another examples is shown in figure 4.8. ISQs differ from other question types (e.g., Clarifying questions, Follow-up questions [186–188]) by having a structure, covering objective details, and expanding on the breadth of the topic. For such a flow to exist between questions, ISQs require maximizing semantic relations and logical coherence. Semantic relations is synonymous to semantic similarity and can be computed using a variety of metrics, such as Cosine Similarity, BERTScore [189], Word Mover Distance [91], Concept Mover Distance [190], and others. Logical coherence can be considered synonymous to natural language inference or textual entailment, where the next question should entail previous in order to maintains consistency in the flow of context. Further, [191] describes clarifying questions are simple questions of facts, good to clarify the dilemma, and *confined to the entities in the query*. In contrast, ISQs go a step further with expanding the query context by *exploring relationships between entities in the query and linked entities in a KG*. Thus retrieving a diverse set of passages (or meta-information) that would provide a proper solution to a user query.

Components in ISEEQ: ISEEQ as a tool can automatically generate curiosity-driven and conceptual flow-based ISQs from a short user query. There are two major components in ISEEQ:

Dynamic Knowledge-aware Passage Retrieval: ISEEQ infuses IS queries with semantic information from knowledge graphs to improve unsupervised passage retrieval. Passages serve as meta-information for generating ISQs.

Reinforcement Learning for ISQs: To improve compositional diversity and legibility in QG, we allow ISEEQ to self-guide the generations through reinforcement learning in a generative-adversarial setting that results in ISEEQ-RL. I introduce entailment constraints borrowed from natural language inference (NLI) guidelines to expand ISEEQ-RL to ISEEQ-ERL to have smooth topical coherent transitions in the questions, achieving conceptual flow. ISEEQ-RL is a variant with reward on semantic relations, whereas in ISEEQ-ERL the reward is on both, semantic relations and logical coherence.

This structure of ISEEQ is defined to make following three contributions in conversational AI, which this chapter would provide answers for:

RQ1 Knowledge Infusion: Can expert-curated knowledge sources like knowledge graphs/bases related to the user query help in context retrieval and question generation?

RQ2 Conceptual Flow: Can ISEEQ generate ISQs having semantic relations and logical coherence?

Transferability or Zero Shot Test: Can ISEEQ generate ISQs in a cross domain setting and generate ISQs for new domains without requiring crowdsourced data collection?

Another utility coming from such a design for ISEEQ, is its forthcoming role of data creation agent to support annotation effort in CIS. Figure 4.9 illustrates the positioning of ISEEQ as dataset creation tool for training CIS agents. Looking at the past research on data creation for enhancing conversational AI, a key player is the pipeline of (a) crawling raw data, (b) setting up annotation guidelines, (c) sorting out crowdworkers and training them

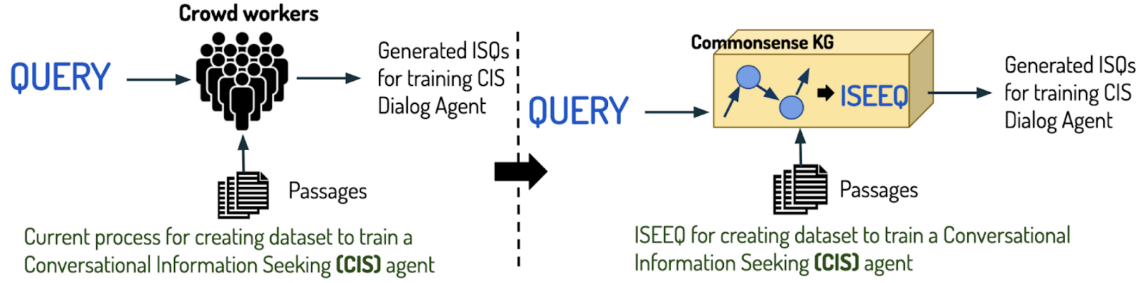


Figure 4.9 Minimize Annotation Effort in Conversational Information seeking

, and (d) handling crowdworkers agreement. What is more painful is the task of crowdworkers, which is characterized as follows: Given a user query, a crowdworker would (a) search the web with curiosity, (b) create good quality questions relevant to the user query, (c) shape the response to the created questions, and (d) maintain the flow of information using semantic relations and logical agreement between questions.

Currently, ISEEQ is capable of addressing following three types of IS queries illustrated through examples:

Title and Description: Online platforms such as Reddit showing this type of information-seeking behavior. For example: *Title*: “I am feeling down and depressed”. *Description*: “I am going through a rough patch in my life. With divorce proceedings and poor growth at work, I am feeling low and hopeless. What do you advise?”.

Topic and Aspects: Humans seek information on Google search, Twitter, WebMD, or MedicineNet by stating topics and aspects. For example: *Topic*: “Anxiety” and *Aspects*: “Panic Attacks, Trauma, Relationship, Self-detox”.

Description: This is relatively shorter in content compared to Title and Description, and Topic and Aspects. “Need Advice! I am bothered by feeling down or depressed”.

These three types of IS queries can be obtained from following datasets that are in use for preparing curiosity-driven conversational agents.

1. QADiscourse (QAD)

- Source for Passages: Wikipedia and WikiNews
- Training Samples: 125 User Queries with 25 ISQs per Query ($125 * 25 = 3,125$ Query-Question Pair)
- Testing Samples: 33 User Queries with 25 ISQs per Query
- ConceptNet KG hit percentage: 38.5%

2. Question Answer Meaning Representations (QAMR)

- Source for Passages: WikiNews, Wikipedia, and Newswire
- Training Samples: 395 User Queries with 63 ISQs per Query
- Testing Samples: 39 User Queries with 68 ISQs per Query
- ConceptNet KG hit percentage: 35.5%

3. Facebook Curiosity (FBC)

- Source for Passages: Geographic Wikipedia
- Training Samples: 8489 User Queries with 6 ISQs per Query
- Testing Samples: 2729 User Queries with 8 ISQs per Query
- ConceptNet KG hit percentage: 50%

4. Conversational Assistance Track Dataset (CAsT-19)

(Dataset only to test ISEEQ, train and test merged)

- Source for Passages: Microsoft MARCO⁵
- Training Samples: 30 User Queries with 9 ISQs per Query
- Testing Samples: 50 User Queries with 10 ISQs per Query
- ConceptNet KG hit percentage: 57%

⁵<https://microsoft.github.io/msmarco/>

The datasets exhibit following properties: (1) existence of semantic relations between questions, (2) logical coherence between questions, and (3) diverse context, that is, queries cover wider domains, such as health, sports, history, geography. Fundamentally, these datasets support the assessment of RQ1, RQ2, and RQ3.

QAD [98] dataset tests the ability of ISEEQ to generate questions that have logical coherence. The sources of queries are Wikinews and Wikipedia that consist of 8.7 Million passages. QAMR [192] dataset tests the ability of ISEEQ to generate questions with semantic relations between them. The source for creating IS queries is Wikinews, which consist of 3.4 Million passages. Both QAD and QAMR consist of only Description-type IS queries. FBC [193] is another dataset that challenges ISEEQ to have both semantic relations and logical coherence. This is because queries are described in the form of Topics and Aspects. The source for IS queries is Wikipedia having 3.3 Million geographical passages. Even though the questions in the dataset have logical coherence, they are relatively less diverse than QAMR and QAD. CAsT-19 [99] is the most challenging one for ISEEQ because of size, diversity in context, large number of passages, and IS queries are not annotated with passages. In CAsT-19, IS queries are provided with Topic and Description.

Adapting Datasets: Each dataset, except CAsT-19, has a query, a set of ISQs, and a relevant passage. For fairness in evaluation, we exclude the passages in the datasets; instead, we retrieve them from the sources using knowledge-aware passage retrieval and ranking components within ISEEQ. We also perform coreference resolution over ISQs using NeuralCoref to increase entity mentions [194]. For example, a question in CAsT-19 “What are the educational requirements required to become one?” is reformulated to “What are the educational requirements required to become a physician’s assistant?”.

4.5 ISEEQ ARCHITECTURE AND EVALUATION

Problem Definition: Given a short query ($q = w_1, w_2, w_3, \dots, w_n$) on any topic (e.g., mental health, sports, politics and policy, location, etc.) automatically generate ISQs in a con-

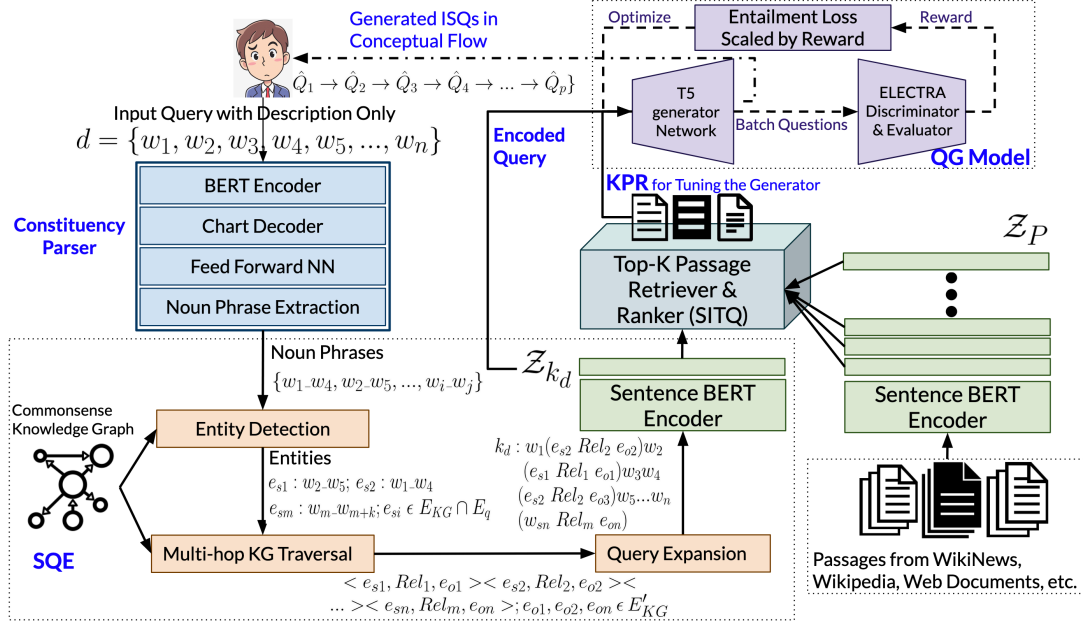


Figure 4.10 Overview of our approach. ISEEQ combines a BERT-based constituency parser, Semantic Query Expander (SQE), and Knowledge-aware Passage Retriever (KPR) to provide relevant context to a QG model for ISQ generations. The QG Model illustrates a structure of ISEEQ variants: ISEEQ-RL and ISEEQ-ERL. We train ISEEQ in generative-adversarial reinforcement learning setting that maximizes semantic relations and coherence while generating ISQs. (Patented with Samsung Research America)

ceptual flow ($ISQ : Q_1, Q_2, Q_3, \dots, Q_p$) to understand specificity in information needs of the user.

Our approach to address this problem, ISEEQ, is outlined in Figure 4.10. We describe in detail the main components of ISEEQ: semantic query expander (SQE), knowledge-aware passage retriever (KPR) and generative-adversarial Reinforcement Learning-based question generator (ISEEQ-RL) with Entailment constraints (ISEEQ-ERL). Inputs to ISEEQ are IS queries described in natural language. For instance, an IS query can be described with **Titles and Descriptions (T & D)**, **Descriptions only (D only)**, **Topics and Aspects (Tp & Asp)**, and others.

SQE: We expand the possibly short user input queries with the help of ConceptNet Commonsense Knowledge Graph (CNetKG) [45]. We first extract the entity set E_d in a user query description d using CNetKG. For this, we use the pre-trained self-attentive encoder-decoder-based constituency parser with BERT as the encoder for consistency in ISEEQ.

The parser is conditioned to extract noun phrases that capture candidate entities defining an IS query [195]. If the phrases have mentions in the CNetKG they are termed as entities⁶. Then a multi-hop triple (subject-entity, relation, object-entity) extraction over CNetKG is performed using depth first search on entity set \mathbf{E}_d . Triples of the form $\langle e_d, Rel_i, e_x \rangle$ and $\langle e_y, Rel_j, e_d \rangle$ are extracted where $e_d \in \mathbf{E}_d$. We keep only those triples where e_d ($\in \mathbf{E}_d$) appears as the subject-entity. We use this heuristic (1) to minimize noise and (2) gather more direct information about entities in \mathbf{E}_d . Finally, we contextualize d by injecting extracted triples to get k_d , a knowledge augmented query.

Take for example \mathbf{D} **only** IS query $d \in \mathbf{D}$, “Want to consider career options from becoming a physician’s assistant vs a nurse”. The extracted entity set \mathbf{E}_d for d is {career, career_options, physician, physician_assistant, nurse}. Then, the extracted triples for this entity set are $\langle \text{career_options}, \text{isrelatedto}, \text{career_choice} \rangle$, $\langle \text{career_options}, \text{isrelatedto}, \text{profession} \rangle$, $\langle \text{physician_assistant}, \text{is_a}, \text{PA} \rangle$, $\langle \text{physician}, \text{is_a}, \text{medical doctor} \rangle$, [...], $\langle \text{nurse}, \text{is_a}, \text{psychiatric_nurse} \rangle$, $\langle \text{nurse}, \text{is_a}, \text{licensed_practical_nurse} \rangle$, $\langle \text{nurse}, \text{is_a}, \text{nurse_practitioner} \rangle$, [...]. The knowledge augmented k_d is

Want to consider career options career options is related to career choice, profession from becoming a physician’s assistant physician assistant is a PA medical doctor, [...] vs a nurse nurse is a psychiatric nurse, licensed practical nurse, [...].

Next, we pass this into KPR. The set $\{k_d\}, \forall d \in \mathbf{D}$ is denoted by $\mathbf{K}_\mathbf{D}$ used by QG model in ISEEQ.

KPR: Given the knowledge augmented query k_d , KPR retrieve passages from a set \mathbf{P} and rank to get top-K passages $\mathbf{P}_{\text{top-K}}$. For this purpose, we make following specific improvements in the Dense Passage Retriever (DPR) described in [61]: (1) Sentence-BERT encoder for the passages $p \in \mathbf{P}$ and k_d . We create dense encodings of $p \in \mathbf{P}$ using Sentence-BERT, which is represented as \mathcal{Z}_p [196]. Likewise, encoding of k_d is represented as \mathcal{Z}_{k_d} .

⁶From here onwards we only use the term Entities, presuming check through exact match is performed using CNetKG

(2) Incorporate SITQ (Simple locality sensitive hashing (Simple-LSH) and Iterative Quantization) algorithm to pick top-K passages ($\mathbf{P}_{\text{top-K}}$) by using a normalized entity score (NES). SITQ is a fast approximate search algorithm over MIPS to retrieve and rank passages. It can be formalized as $\text{Score}(\mathbf{P}_{\text{top-K}}|k_d)$ where,

$$\text{Score}(\mathbf{P}_{\text{top-K}}|k_d) \propto \{\text{WMD}(\mathcal{Z}_{k_d}^T \mathcal{Z}_p)\}_{p \in \mathbf{P}}$$

$$\mathcal{Z}_{k_d} = \text{S-BERT}(k_d); \mathcal{Z}_p = \text{S-BERT}(p);$$

SITQ converts dense encodings into low-rank vectors and calculates the semantic similarity between the input query and passage using word mover distance (WMD) [91]. $\mathbf{P}_{\text{top-K}}$ from SITQ is re-ranked by NES, calculated⁷ for each $p \in \mathbf{P}_{\text{top-K}}$ as $\frac{\sum_{e_j \in k_d} \{\mathbb{I}(e_j=w)\}_{w \in p}}{|k_d|}$ and arrange in descending order. $\mathbf{P}_{\text{top-K}}$ consists of K passages with NES >80%. Execution of KPR is iterative and stopped when each query in the train set has at least one passage for generating ISQs.

We tested retrieving efficiency of KPR using encoding of e_d denoted by \mathcal{Z}_{e_d} and using the encoding of k_d denoted by \mathcal{Z}_{k_d} as inputs to KPR. Measurements were recorded using Hit Rate (HR) @ 10 and 20 retrieved passages. Mean Average Precision (MAP) is calculated with respect to ground truth questions in QAMR. There are two components in MAP: (a) *Relevance* of the retrieved passage in generating questions that have >70% cosine similarity with ground truth; (b) Normalize *Relevance* by the number of ground truth questions per input query. To get MAP, we multiply (a) and (b) and take mean over all the input queries. We computed MAP by setting $K = 20$ retrieved passages due to the good confidence from hit rate (a hyperparameter). KPR outperformed the comparable baselines on the QAMR Wikinews dataset and Table 4.2 shows that SQE improves the retrieval process⁸. A set of $\mathbf{P}_{\text{top-K}}$ for $\mathbf{K_D}$ is denoted by $\{\mathbf{P}_{\text{top-K}}\}_{k_d}, k_d \in \mathbf{K_D}$.

⁷an entity occurring multiple times in p is counted once

⁸KPR(\mathcal{Z}_{e_d}) & KPR(\mathcal{Z}_{k_d}) is executed for each CAsT-19 query.

Table 4.2 Evaluating retrievers. ECE: Electra Cross Encoder, (*): variant of (Clark et al. 2019), DPR: Dense Passage Retrieval.

Retrievers	HR@10	HR@20	MAP
TF-IDF + ECE [197]	0.31	0.45	0.16
BM25 + ECE*	0.38	0.49	0.23
DPR [198]	0.44	0.61	0.31
KPR(\mathcal{Z}_{e_d})	0.47	0.66	0.35
KPR(\mathcal{Z}_{k_d})	0.49	0.70	0.38

QG Model: ISEEQ leverages $\mathbf{K_D}$ and $\{\mathbf{P}_{\text{top-K}}\}_{k_d}$ to learn QG in generative-adversarial setting guided by a reward function. ISEEQ-RL contains T5-base as generator and Electra-base as discriminator to learn to generate IS-type questions. ISEEQ use the reward function to learn to selectively preserve terms from the IS query versus introducing diversity. Also, reward function prevent ISEEQ from generating ISQs that are loose in context or redundant.

Reward Function: Let q_i^n be the i^{th} question in the ground truth questions Q having n tokens and let \hat{q}_i^m be the i^{th} question in the list of generated questions, \hat{Q} having m tokens. We create BERT encodings for each of the n and m words in the question vectors. The reward (R_i) in ISEEQ-RL and ISEEQ-ERL is defined as:

$$\alpha \left[\frac{LCS(\hat{q}_i^m, q_i^n)}{|\hat{q}_i^m|} \right] + (1 - \alpha) \left[\sum_{\hat{w}_{ij} \in \hat{q}_i^m} \max_{w_{ik} \in q_i^n} \text{WMD}(\hat{w}_{ij}^T w_{ik}) \right] \quad (4.6)$$

where $\alpha[*]$ is a normalized longest common subsequence (LCS) score that capture word order and make ISEEQ-RL learn to copy in some very complex IS-type queries. $(1 - \alpha)[*]$ uses WMD to account for semantic similarity and compositional diversity. For a $q_i^n =$ ‘‘What is the average starting salary in the UK?’’, $(1 - \alpha)[*]$ generates $\hat{q}_i^m =$ ‘‘What is the average earnings of nurse in UK?’’

Loss Function in ISEEQ-RL: We revise cross entropy (CE) loss for training ISEEQ by scaling with the reward function because each $k_d \in \mathbf{K_D}$ are not only short but they also vary by context. Corresponding to each k_d , there are b ground truth questions $q_{1:b}$ and thus we normalize the revised CE loss by a factor of b . Formally, we define our CE loss in

$$\text{ISEEQ-RL}, \mathcal{L}(\hat{q}_{1:b}|q_{1:b}, \theta) = \frac{-\sum_{i=1}^b R_i \cdot \mathbb{I}(q_i^n = \hat{q}_i^m) \cdot \log \Pr(\hat{q}_i^m | \theta)}{b} \quad (4.7)$$

where $\mathbb{I}(q_i^n = \hat{q}_i^m)$ is an indicator function counting word indices in q_i^n that match word indices in \hat{q}_i^m . The CE loss over $\mathbf{K_D}$ in a discourse dataset is $\mathcal{L}(\hat{Q}|Q, \Theta)_t$, recorded after t^{th} epoch. Formally $\mathcal{L}(\hat{Q}|Q, \Theta)_t =$

$$\gamma \mathcal{L}(\hat{Q}|Q, \Theta)_{t-1} + (1 - \gamma) \mathcal{L}(\hat{q}_{1:b}|q_{1:b}, \theta) \quad (4.8)$$

Theoretically, ISEEQ-RL addresses RQ1, but weakly mandates conceptual flow while generating ISQs. Thus, it does not address RQ2.

Loss Function in ISEEQ-ERL: For instance, given $d2(\in \mathbf{D})$: “Bothered by feeling down or depressed”, ISEEQ-RL generations are: (\hat{q}_1) : What is the reason for the depression, hopelessness? and (\hat{q}_2) What is the frequency of you feeling down and depressed? Whereas, ISEEQ-ERL would re-order placing (\hat{q}_2) before (\hat{q}_1) for conceptual flow. To develop ISEEQ-ERL, we redefine the loss function in ISEEQ-RL by introducing principles of entailment as in NLI [199] [200]⁹. Consider $\hat{q}_{i|next}^m$ to be the next generated question after \hat{q}_i^m . We condition equation 4.7 on $y_{max} = \arg \max_Y \text{RoBERTa}(\hat{q}_i^m, \hat{q}_{i|next}^m)$, where $Y \in \{\text{neutral, contradiction, entailment}\}$ and $\Pr(y_{max}) = \max_Y \text{RoBERTa}(\hat{q}_i^m, \hat{q}_{i|next}^m)$. Formally, $\mathcal{L}(\hat{q}_{1:b}|q_{1:b}, \theta)$ in ISEEQ-ERL is:

Algorithm 1: Entailment Contrained Loss in ISEEQ-ERL

```

1 if  $y_{max} == \text{entailment}$  then
2   |  $\text{CE} - \Pr(y_{max})$  ;
3 else
4   |  $\text{RCE} = -\frac{\sum_{i=1}^b R_i (1 - \mathbb{I}(q_i = \hat{q}_i)) \Pr(\hat{q}_i | \theta)}{b}$  ;
5   |  $\text{RCE} - (1 - \Pr(y_{max}))$  ;
```

⁹We use RoBERTa pre-trained on Stanford NLI dataset to measure semantic relations and coherence between a pair of generated questions

Table 4.3 An ablation study showing improvement in the quality of ISQs after encodings of retrieved passages ($\mathbf{P}_{1:K}$) are concatenated with knowledge-augmented query (k_d) after SQE. The concatenation is performed for each $p \in \mathbf{P}_{1:K}$.

Model	Encoding	QAD	QAMR	FBC
		R-L/BRT/BScore/SR/LC(%)		
ISEEQ-RL	$P_{1:K}$	0.62/ 0.73/ 0.39/ 0.21/ 21.3	0.47/ 0.71/ 0.39/ 0.63/ 28	0.74/ 0.83/ 0.60/ 0.73/ 71.6
	$P_{1:K}$ + SQE	0.65/ 0.74/ 0.45/ 0.25/ 22	0.53/ 0.78/ 0.71/ 0.65/ 34.7	0.74/ 0.87/ 0.65/ 0.73/ 71.8
ISEEQ-ERL	$P_{1:K}$	0.62/ 0.76/ 0.44/ 0.26/ 24.6	0.54/ 0.80/ 0.73/ 0.68/ 36.3	0.71/ 0.84/ 0.61/ 0.77/ 78.2
	$P_{1:K}$ + SQE	0.67/ 0.79/ 0.50/ 0.27/ 25.7	0.57/ 0.83/ 0.77/ 0.68/ 37.0	0.79/ 0.89/ 0.66/ 0.78/ 79.4

Reverse Cross Entropy(RCE) complements CE (Equation 4.7) by checking $\hat{q}_{i|next}^m$ is semantically related and coherent to \hat{q}_i^m . Tuning of the loss after an epoch follows Equation 4.8.

ISEEQ Evaluation: ISEEQ-RL or ISEEQ-ERL generator uses top-p (nucleus) sampling¹⁰ with sum probability of generations equal 0.92, a hyperparameter that sufficiently removes the possibility of redundant QG [201]. We evaluate ISEEQ generations using Rouge-L (R-L), BERTScore (BScore) [189], and BLEURT (BRT) [80] that measure preservation of syntactic context, semantics, and legibility of generated question to human understanding, respectively. For conceptual flow in question generation, we define “semantic relations” (SR) and “logical coherence” (LC) metrics. To calculate SR or LC, we pair $\hat{Q}_{1:p}$ generated questions with Q . SR in the generations is computed across all pairs using RoBERTa pre-trained on semantic similarity tasks¹¹. LC between Q and $\hat{Q}_{1:p}$ is computed from counting the labels predicted as “entailment” by RoBERTa pre-trained on SNLI dataset¹².

¹⁰Top-p or Top-K sampling either works in ISEEQ

¹¹<https://huggingface.co/textattack/roberta-base-STS-B>

¹²<https://paperswithcode.com/lib/allennlp/roberta-snli>

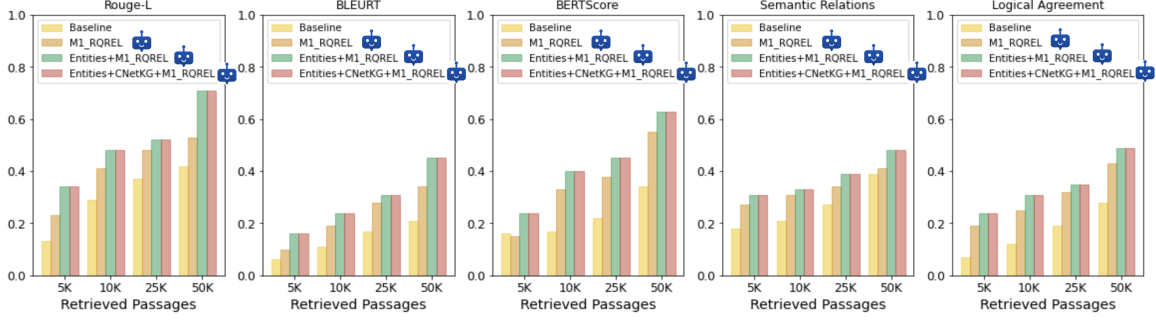


Figure 4.11 Dark Blue chatbot is a pictograph of ISEEQ-RL. With KG and minimal set of passages, ISEEQ-RL generated 27% more questions that are semantically similar to ground truth compared. Without entailment constraints, questions generated by KG triples never made to top-K. Hence performance of ISEEQ-RL with or without KG triples is same.

Baselines: As there exists no system to automatically generate ISQs, we considered transformer language models fine-tuned (TLMs-FT) on open domain datasets used for reading comprehension, and complex non-factoid answer retrieval as baselines. Specifically, T5 model fine-tuned (T5-FT) on WikipassageQA [202], SQUAD [203], and CANARD [204], and ProphetNet [205] fine-tuned on SQUADv2.0 are comparable baselines.

We substantiate our claims in RQ1, RQ2, and RQ3 by highlighting: (1) Multiple passage-based QG yields better results over single gold passage QG used in TLMs-FT (e.g. T5 fine tuned on SQUAD dataset; results in Table 4.4); (2) Knowledge-infusion through SQE significantly advance the process of QG (Table 4.3); (3) Pressing on conceptual flow in ISEEQ-ERL improve SR and LC in generations. Evidence from 12 human evaluations support our quantitative findings (Table 4.7); (4) We investigate the potential of ISEEQ-ERL in minimizing crowd workers for IS dataset creation through cross-domain experiments (Table 4.6).

Performance of ISEEQ-RL and ISEEQ-ERL: Datasets used in this research were designed for a CIS system to obtain the capability of multiple contextual passage retrieval and diverse ISQ generation. The process of creating such datasets requires crowd workers to take the role of a CIS system responsible for creating questions and evaluators to see whether questions match the information needs of IS queries. Implicitly, the process embed crowd workers’ curiosity-driven search to read multiple passages for generating ISQs.

Table 4.4 Scores on test set of datasets. In comparison to T5-FT CANARD, a competitive baseline, ISEEQ-ERL generated better questions across three datasets (30%↑ in QADiscourse, 7%↑ in QAMR, and 5%↑ in FB Curiosity). For fine-tuning we used SQUADv2.0.

Methods	SQE	QAD					QAMR					FBC				
		R-L	BRT	BScore	SR	LC(%)	R-L	BRT	BScore	SR	LC(%)	R-L	BRT	BScore	SR	LC(%)
T5-FT WikiPassageQA	-	0.37	0.43	0.16	0.17	10.0	0.19	0.51	0.38	0.36	17.0	0.65	0.78	0.54	0.51	47.3
	+Entities	0.39	0.45	0.16	0.17	10.0	0.20	0.53	0.38	0.36	17.5	0.65	0.78	0.54	0.52	47.4
	+Triples	0.41	0.46	0.16	0.18	11.0	0.20	0.53	0.39	0.37	17.8	0.65	0.78	0.55	0.52	47.3
T5-FT SQUAD	-	0.44	0.54	0.20	0.19	13.0	0.40	0.66	0.46	0.58	21.0	0.70	0.83	0.62	0.67	65.1
	+Entities	0.45	0.56	0.22	0.19	13.5	0.40	0.68	0.47	0.59	22.7	0.71	0.84	0.63	0.69	65.8
	+Triples	0.45	0.58	0.22	0.20	13.8	0.43	0.69	0.47	0.59	22.6	0.70	0.84	0.64	0.69	65.8
T5-FT CANARD	-	0.47	0.54	0.23	0.19	17.1	0.41	0.64	0.53	0.58	22.6	0.73	0.84	0.63	0.67	66.2
	+Entities	0.48	0.55	0.25	0.20	17.5	0.44	0.67	0.62	0.61	23.5	0.74	0.84	0.65	0.69	66.5
	+Triples	0.51	0.57	0.26	0.21	18.3	0.49	0.68	0.66	0.61	24.3	0.74	0.85	0.65	0.70	68.2
ProphetNet-FT SQUAD	-	0.31	0.44	0.14	0.17	12.2	0.35	0.59	0.38	0.36	21.5	0.63	0.78	0.53	0.67	63.2
	+Entities	0.31	0.44	0.14	0.17	12.7	0.37	0.60	0.41	0.37	22.1	0.65	0.78	0.54	0.67	63.3
	+Triples	0.34	0.45	0.15	0.18	13.0	0.37	0.61	0.43	0.37	22.3	0.65	0.79	0.56	0.69	64.0
ISEEQ-RL	-	0.57	0.72	0.40	0.22	20.0	0.50	0.75	0.67	0.64	29.4	0.71	0.84	0.62	0.69	68.2
	+Entities	0.64	0.72	0.41	0.23	22.0	0.52	0.77	0.68	0.64	33.1	0.72	0.85	0.63	0.71	69.8
	+Triples	0.65	0.74	0.45	0.25	22.0	0.53	0.78	0.71	0.65	34.7	0.74	0.87	0.63	0.73	71.8
ISEEQ-ERL	-	0.60	0.76	0.44	0.26	24.5	0.55	0.81	0.72	0.68	36.1	0.74	0.85	0.64	0.76	78.2
	+Entities	0.65	0.78	0.47	0.27	25.2	0.55	0.82	0.74	0.68	36.3	0.77	0.88	0.66	0.76	78.3
	+Triples	0.67	0.79	0.50	0.27	25.7	0.57	0.83	0.77	0.68	37.0	0.79	0.89	0.66	0.78	79.4

Baselines on employed datasets use single passage QG, with much of the efforts focusing on improving QG. Whereas ISEEQ generation enjoys the success from the connection of SQE, KPR, and novel QG model over baselines in CIS (see Table 4.4). With SQE, ISEEQ achieved 2-6% across all datasets. The knowledge-infusion in ISEEQ through SQE has shown to be powerful for baselines as well. Table 4.4 records 3-10%, 3-10%, and 1-3% performance gains of the baselines on QAD, QAMR, and FBC across five evaluation metrics, respectively. SQE allows baselines to semantically widen their search over the gold passages in datasets to generate diverse questions that match better with ground truth. Differently, ISEEQ-RL generations benefit from dynamic meta-information retrieval from multiple passages yielding hike of 20-35%, 6-13%, 3-10% on QAD, QAMR, and FBC, respectively, across five evaluation metrics. Especially, QG in CAsT-19 and FBC datasets advance because of KPR in ISEEQ-RL and ISEEQ-ERL (see Figures 4.11 and 4.12).

Most of the CAsT-19 and FBC queries required multiple passages to construct legible questions. For instance, an IS query : “Enquiry about History, Economy, and Sports in Hyderabad” ISEEQ retrieved following three passages: “History_Hyderabad”, “Economy_Hyderabad”, and “Sports_Hyderabad” which were missing in the set of passages in

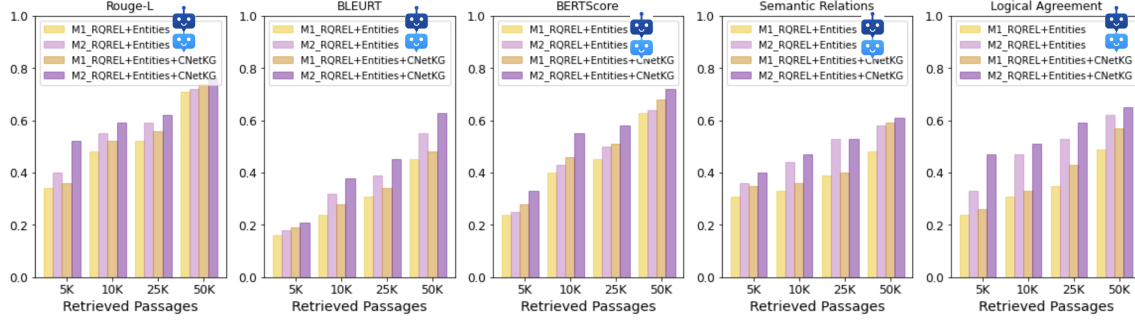


Figure 4.12 Light blue chatbot is a pictograph of ISEEQ-ERL. Forcing the entailment constraints in ISEEQ-ERL yielded high scores on Semantic Relation and Logical Agreement – Conceptual Flow. This performance is seen uniformly from minimal set of 5K passages to 50K passages. Though in initial epochs, “entity only” generated question takes precedence over questions generated using triples from KG. In higher epochs, questions generated from passages retrieved from ConceptNet information took precedence. Longer training cycles were there for ISEEQ-ERL over ISEEQ-RL (2 hours long).

Table 4.5 Performance of KPR on MS-MARCO passages while retrieving atleast one passage per IS query in CAsT-19. 269 is the size of CAST-19 train set. KPR covered the train set but left 16% of the IS queries in test set. Ret.Pass. : Retrieved Passages.

Ret.Pass.	DPR		KPR(\mathcal{L}_{e_d})		KPR(\mathcal{L}_{k_d})	
	Train	Test	Train	Test	Train	Test
5K	71	123	99	278	157	275
10K	96	133	154	301	194	316
25K	139	133	235	329	236	363
50K	173	144	269	358	269	402

FBC. Thus, TLM-FT baselines find it hard to construct legible ISQs using a single passage. Furthermore, ISEEQ-ERL advance the quality of ISQs over ISEEQ-RL by 7-14% and 6-10% in QAD and FBC (refer Table 4.3). This is because QAD and FBC questions require the QG model to emphasize conceptual flow.

Further, we examine the combined **performance of KPR** and ISEEQ-ERL on CAsT-19 dataset. KPR retrieve $\sim 50K$ passages sufficient to generate questions for 269 IS queries¹³. Table 4.5 depicts $KPR(\mathcal{L}_{e_d})$ retrieval performance match $KPR(\mathcal{L}_{k_d})$, with later supported 72% of queries in training set compare to 57% by $KPR(\mathcal{L}_{e_d})$. Also, it outperforms DPR,

¹³one query can have multiple passages

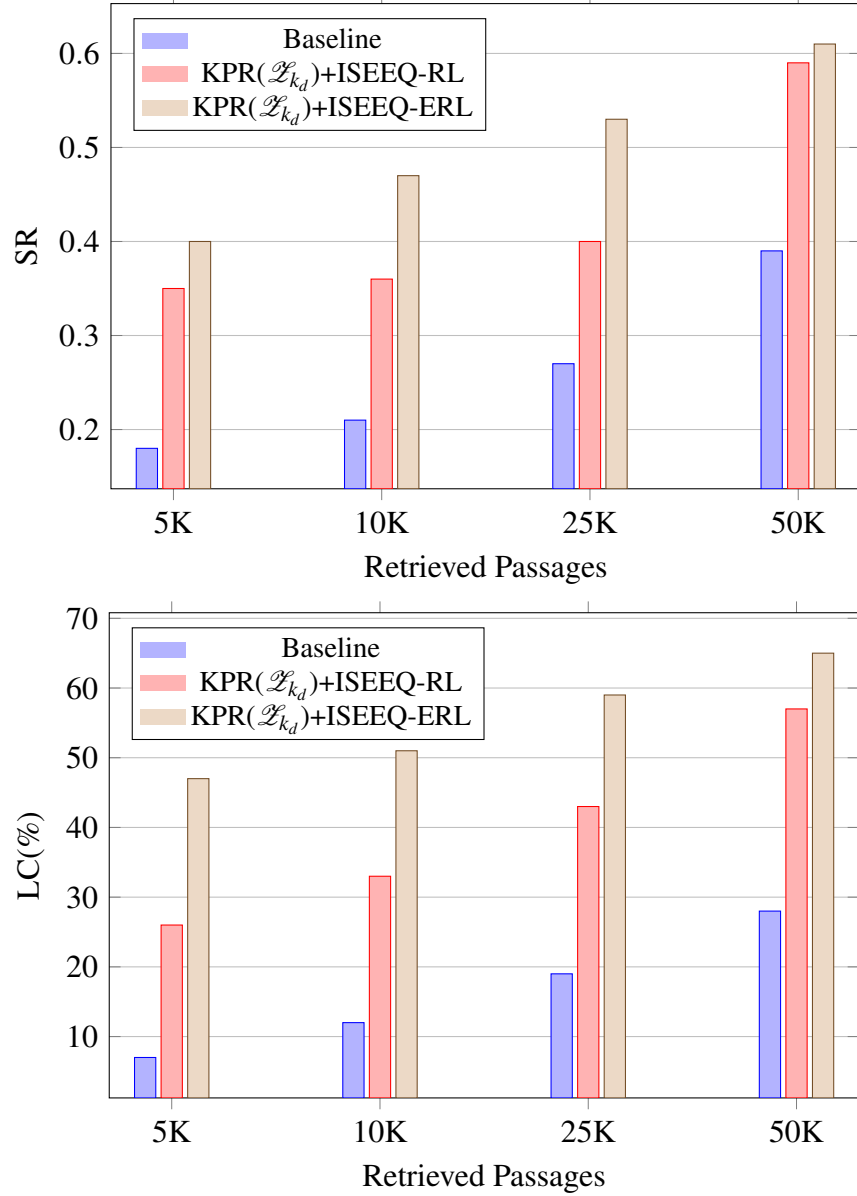


Figure 4.13 Improvement in performance of ISEEQ-ERL over ISEEQ-RL and Baseline: T5-FT CANARD concerning SR and LC in generated ISQs. This experiment was performed on CAsT-19 with *unannotated* passages.

which supported 30% queries in train set (see Table 4.5). In test time, KPR(\mathcal{Z}_{k_d}) supported 84% queries that were used to generate questions by ISEEQ-ERL and evaluated with ground truth for SR and LC (see Figure 4.13). Apart from monotonic rise in SR and LC scores shown by ISEEQ, ISEEQ-ERL generations achieved better coherence than

counterparts with 5K passages (Figure 4.12). We attribute the addition of entailment check and RCE for conceptual flow-based QG improvements.

Transferability Test for RQ3: We examine the performance of ISEEQ-ERL in an environment where the train and test dataset belong to a different domain. For instance, QAMR is composed of IS queries from Wikinews, whereas FBC is composed of IS queries from geography category in Wikipedia.

Table 4.6 Transferability test scores using ISEEQ-ERL to answer RQ3. gray cell: ISEEQ-ERL trained and tested on same dataset. dark gray cell: shows acceptable cross-domain {Train-Test} pairs, where train size is smaller than test size.

Test →	QAD			QAMR			FBC			CAsT19		
Train ↓	R-L/BRT/BScore/SR/LC(%)											
QAD	0.67/ 0.27/	0.79/ 25.7	0.50/	0.56/ 0.64/	0.79/ 33.1	0.75/	0.62/ 0.71/	0.70/ 73.5	0.55/	0.76/ 0.60/	0.48/ 64.2	0.64/
QAMR	0.73/ 0.28/	0.89/ 27.7	0.62/	0.57/ 0.68/	0.83/ 37.0	0.77/	0.74/ 0.75/	0.89/ 77.8	0.67/	0.67/ 0.57/	0.41/ 58.6	0.57/
FBC	0.70/ 0.31/	0.73/ 33.0	0.56/	0.61/ 0.67/	0.85/ 35.8	0.72/	0.79/ 0.78/	0.89/ 79.4	0.66/	0.75/ 0.67/	0.37/ 66.5	0.76/
CAsT-19	0.58/ 0.23/	0.69/ 25.2	0.51/	0.52/ 0.61/	0.73/ 33.4	0.70/	0.63/ 0.73/	0.77/ 76.5	0.57/	0.74/ 0.61/	0.48/ 65.0	0.68/

From experiments in Table 4.6, we make two deductions: (1) ISEEQ-ERL provided acceptable performance in generating ISQs for {Train-Test} pairs, where train size is smaller than test size: {QAD-QAMR} and {QAMR-FBC}¹⁴. (2) ISEEQ-ERL trained on a *narrow domain dataset* (FBC) generated far better ISQs for IS queries in generic domain. The transferability test show ISEEQ-ERL’s ability to create new datasets for training and development of CIS systems.

Human Evaluation: We carried out 12 blind evaluations of 30 information-seeking queries covering mental health (7), politics and policy (6), geography (5), general health (3), legal news (2), and others (4). Each evaluator rate ISQs from the ground-truth dataset (S1), ISEEQ-ERL (S2), and T5-FT CANARD (S3) using Likert score where 1 is the lowest

¹⁴S-BERT Pairwise similarity between retrieved passages in QAD, QAMR, FBC, and CAsT-19 is <9%, which is minimal.

Table 4.7 Assessment of human evaluation. G1: ISQs are diverse in context and non-redundant. G2: ISQs are logically coherent and share semantic relations. >: difference is statistically significant. SD: Standard Deviation. S1, S2, and S3 are ground truth, ISEEQ-ERL, and T5-FT CANARD, respectively.

	Response: Mean (SD)			F(2, 957) (p-value)	LSD post-hoc (p < 0.05)
	S1	S2	S3		
G1	3.756 (1.14)	3.759 (1.06)	3.518 (1.08)	5.05 (6.5e-3)	S1>S3, S2>S3
G2	3.803 (1.10)	3.843 (1.02)	3.503 (1.06)	9.71 (6.63e-5)	S1>S3, S2>S3

and 5 is the highest. A total of 570 ISQs (On average 7 by S1, 7 by S2, and 4 by S3) were evaluated on two guidelines, described in Table 4.7. We measured their statistical significance by first performing one-way ANOVA and then using Least Significant Difference (LSD) post-hoc analysis (as performed in [206]). Across the 30 queries on both guidelines, both S1 and S2 are better (statistically significant) than S3 whereas, even though S2 mean is better than S1, there is no statistical significance between the two systems on the scores (we may say they are comparable).

4.6 SUMMARY

Semi-Deep Infusion introduces a class of methods that allows knowledge infusion through model parameters and optimization methods. In the recommender system tasked to predict psychiatric disorders based on human input, we saw how one form of knowledge could help statistical ML/DL models curtail the chance of false alarms. Further, the explainability aspect is assessed through the attention-based highlighting of tokens and its overlap with highlighting performed by mental health professionals. The interpretability aspect is achieved by the hyperparameter δ that helps the model in deciding: “when to take support of knowledge” and “when to move forward with data alone”. There are two open questions that Semi-Deep Infusion cannot address:

- What if the human input has an implicit hierarchical and abstract relationship and one-shot parametric knowledge infusion at the input is insufficient in capturing it and leveraging it for explainable decision making?
- What if, to capture implicit relationships, we require heterogeneous knowledge infusion from multiple knowledge sources. If so, each knowledge source would represent a different semantic parametric space. Simple aggregation, concatenation, or similar arithmetic would not work as they introduce noise [9]. How can we have heterogeneous knowledge infusion in a stratified manner, and How can we control the amount of heterogeneous knowledge infusion?

In conversational artificial intelligence, we saw how knowledge could help maintain semantic and conceptual flow in open domain question generation. Its application in closed-domain, like mental healthcare, has been discussed in these two studies by Roy et al. [17] and Gupta et al. [207]. The ISEEQ agent is interpretable through entailment-driven question generation, reflecting human information-seeking behavior. It is explainable by checking retrieved passages and their use in question generation. Two open questions reflect on the limitation in Semi-Deep Infusion:

- Questions asked in the real world are not only chit-chat or information-seeking [208]. They can also be diagnostic, like in the case of mental health interview [23], where there is an inherent process in the asked questions. Asking a context-controlled and right follow-up question can yield actionable decisions. The question is: How to develop a method that infuses the inherent process in question-answering or dialog into the agent’s decision making pipeline?
- The dialog agent discussed exploiting knowledge’s structural properties in passage retrieval. What if the agent’s decision-making architecture has a graphical structure, like Graph neural network, but with nodes and edges representing concepts and relationships in KG?

These open questions defines the path for **process knowledge infusion** and **deep knowledge infusion**.

CHAPTER 5

PROCESS KNOWLEDGE INFUSION

Benchmarking datasets that assess the natural language understanding capabilities of large language models fall short in accelerating models to achieve user-level explainability, safety, uncertainty, and risk handling¹ [27]. These challenges are associated with the limitations of AI in restricting its learning tasks to classification and generation, which are single shots. In comparison, real-world applications demand an orchestrated response going through a multi-step process of learning the high-level needs of the user, then drilling down to specific needs, and subsequently yielding a structured response having a conceptual flow. For example, triaging patients in mental health requires clinical process knowledge manifested in a clinical questionnaire. Figure 5.1 illustrates a scenario where the agent maps user input to a sequence of yes or no questions to compile suicide risk severity. The agent can keep track of user-provided cues and ask appropriate follow-up questions through these ordered sets of questions. Upon receiving the required information to derive appropriate severity labels, the agent's outcome can be explained to mental healthcare professionals (MHPs) for appropriate intervention. Similar but more complex applications include using ADOS (Autism Diagnostic Observation Schedule) to evaluate children with autism or using MoCA (Montreal Cognitive Assessment) score to measure the cognitive decline in post-stroke Aphasia patients [209]².

To train conversational agents for such functionality requires specialized datasets grounded in the knowledge that enables AI systems to exploit the duality of data and knowledge

¹<https://tinyurl.com/KiL-MentalHealth-NLU>

²<https://tinyurl.com/ABC-Cohort-Repo>

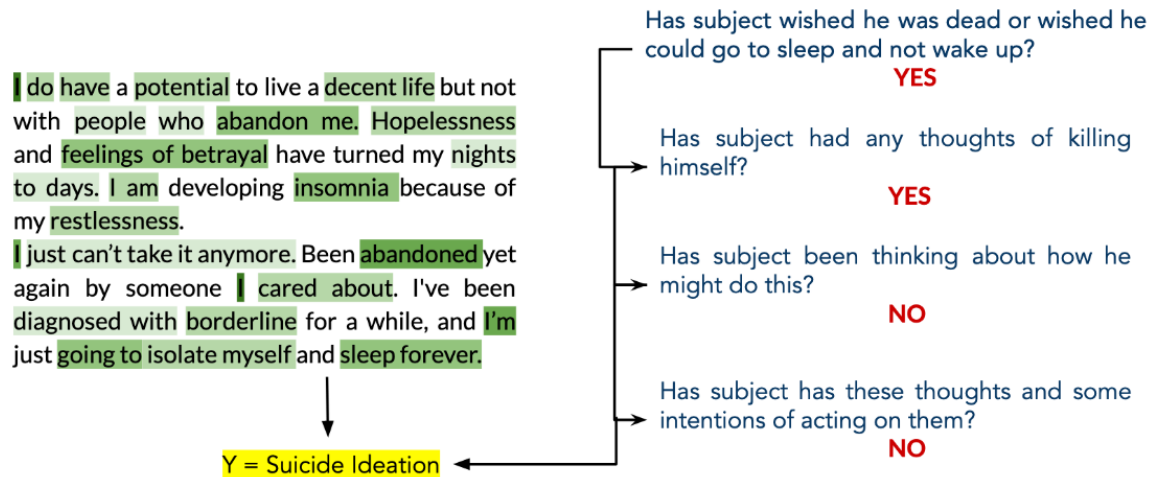


Figure 5.1 An illustration of a classification task that benefits from process knowledge. Here, an AI model using a process knowledge structure would consume the user's input, extract conceptual cues that can answer questions in process knowledge, and provide a classification label. The figure illustrates this process in assessing suicide risk severity using a partial sequence of questions from the C-SSRS. The highlighted text on the left is concept phrases that contribute to the yes/no in the C-SSRS questions.

for human-like decision-making [175]³. Further, to develop agents that learn from such process knowledge-integrated datasets we require interpretable and explainable learning mechanisms⁴. These learning mechanisms have been characterized under the umbrella of KiL.

Significance

Deep Language Models suffer from factual incorrectness, irrelevant sentence generation, and failure to maintain conceptual flow. Some of the generated sentences lose the semantic relations between current and previous generations. Such generations may have severe consequences in critical applications, like Mental Health. Incorporating the process knowledge help address these limitations. This chapter discusses **Process Knowledge Infused Learning** in the context of Language Models in Mental Health.

³<https://tinyurl.com/duality-data-knowledge>

⁴<https://tinyurl.com/petrinet-workflow>

5.1 PROCESS KNOWLEDGE AND ITS INFUSION INTO STATISTICAL AI

Process knowledge incorporates flow of work and task for specific goals. PK is incorporation of this flow in the knowledge representation. An example of process model or PK in a clinical domain, would be the use of clinical guidelines or clinical protocols for diagnosis or treatment. An instance of this example are the clinical practice guidelines (CPGs) by The American Academy of Family Physicians (AAFP) to support systematic clinical assessment and decision making.

On the other hand, U.S. Departments of Agriculture (USDA) and Health and Human Services (HHS) develops Dietary Guidelines for Americans ⁵ that serves as an recommendation for meeting nutrient needs, promote health, and prevent disease. An AI system adapted to process knowledge can handle uncertainty in prediction, and the predicted outcomes are safe and user-level explainable. Further, an AI system can consider process knowledge as meta-information to capture the sequential context necessary for carrying out a structured conversation. Also, it allows the developer of the AI system to probe the internal decision-making of AI systems using application-specific guidelines or specifications that inform the synchrony between the end-users thought process and the model's functioning.

This unique form of knowledge differs from other forms of knowledge in the following manner: (a) knowledge graph: it is structured but not ordered. Knowledge graphs can support context capture but cannot enforce conceptual flow⁶. (b) Semantic lexicons: this is a flattened form of knowledge graph that makes deep language models context-sensitive and add constraints but cannot enforce conceptual flow [210]⁷. (c) Ontologies are curated schematic forms of knowledge graphs with classes, instances, and constraints. Thus, ontologies can provide stricter control over context and constraints. If defined, an ontol-

⁵<https://tinyurl.com/american-dietary>

⁶<https://tinyurl.com/KI-summarization>

⁷<https://tinyurl.com/lex-to-flex>

ogy can enforce order in question generation using deep language models [211]⁸. Process knowledge is represented differently for different applications. For instance, to assess the severity of suicide risk, the process knowledge used is C-SSRS, which is similar to a flow chart. On the other hand, the GAD-7-based process knowledge is used to assess anxiety severity which has a flattened structure (see Figure 5.2). These characteristic properties of **PK** and its infusion into statistical AI would yield a new class of neuro-symbolic algorithms that would drive the question:

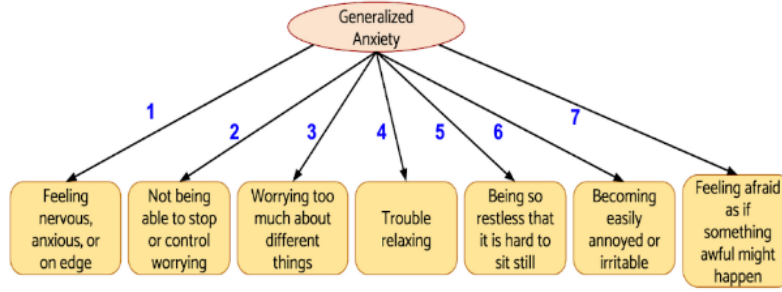
What if we could use the annotator’s labels and the process or guidelines used to label them and explicitly control the learning of a model to recover the guideline or process (instead of implicitly)?

Such an algorithm would, by design, be explainable and emulate the human model of similarity between data points. For the task of classification, a process knowledge-infused AI system would solicit the use of interpretable machine learning algorithms (e.g., Decision Trees, Random Forest) that can enforce structure in decision making over traditional deep language model-based classification⁹. In NLG, the biggest concern with deep generative language models is that they hallucinate when either asking questions or providing responses in a conversational setting. Along with the issue of hallucination, there have been extensive study about the inappropriate and unsafe risk behaviors of language models¹⁰. Efforts to pair these language models with passage retrievers and rankers have been proposed to control incoherent, irrelevant, and factually incorrect responses and questions; however, the order, like the one defined in process knowledge, is far from being realized [212]. Such process knowledge-based NLG is even more crucial in the field of healthcare NLP, where each response from the agent can have severe consequences. These concerns are further discussed with the help of a use case: **Mental Health**.

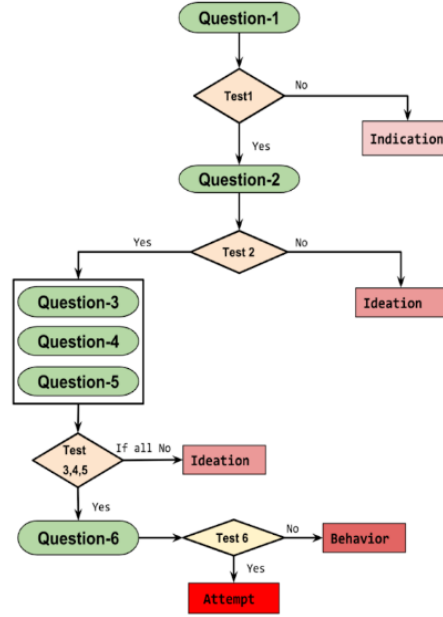
⁸<https://tinyurl.com/MCQ-generation-ontology>

⁹<https://tinyurl.com/PK-iL-suicide>

¹⁰<https://tinyurl.com/LaMDA-dialog>



(a)



(b)

Figure 5.2 Illustration of process knowledge for different purposes. (Left) A process knowledge to assess anxiety disorder in an individual using GAD-7 questionnaire. (Right) A process knowledge to assess severity of suicide risk in an individual

5.2 BENEFITS OF PROCESS KNOWLEDGE INFUSION

The purpose of enforcing a process structure in DL is to:

Minimize Hallucination: Conversational Agents (CAs) are susceptible to irrelevant and sometimes harmful questions when generating FQs or responses to a patient suffering from depression [213]. The primary reason for irrelevant and harmful questions is that CAs cannot incorporate contextual information in generating appropriate follow-up questions (FQs) (see Figure 5.3). Further, the sensitivity of the conversation and

Several years of persistent drowsiness continue to wreak havoc in my life

REQUESTING ADVICE



- o [...] "I started experiencing **general anxiety** for no apparent reason. Also, felt stressed a lot for no real reason either." "I woke up one day after a good night's sleep and felt a little tired. Over the months and years, it has very gradually worsened to the point where I'm at now where I can hardly function due to the **overwhelming fatigue/brain fog**. It is **24/7**."
- o I **feel drunk/drugged**, especially as the day progresses. Come 8 PM, I feel so spaced out, the **fatigue is unbearable**. [...] The **gradual cognitive decline is shocking** (Concentration, alertness, memory, focus, processing/recalling/stumbling over words and sentences, confusion).
- o My motor skills have also declined badly as well. I bump into things, my **reflexes are awful**, I miss things when I try and grab them, etc. [...] **Eyes are dry/watery** off and on. Numbness all over body. Tough to explain, but **extremities/body parts feel so utterly exhausted** and numb-like to the touch.[...] **Senses feel dulled** (Hearing, taste, smell, touch) all feel weakened and suppressed, especially as the day goes on and I feel more tired.
- o **Alcohol hits me much harder now**. As a result, I feel intoxicated more easily/quickly. [...] I am **emotionally flat lined**. [...] I **feel like I need lots of sleep to "recover"**, but all sleep really does is reset me a little bit and is very unrefreshing. [...] Everything feels like it's suppressed by this fog that has gradually gotten worse over time.
- o Realistically, the last kind of test I can do at this point is a **sleep study**, which is in a few weeks.[...]



Questions asked by Reddit Users



- Does the fatigue wax and wane?
- Could something in your mind be depleting your life force, or do you believe it is a physical/medical problem with your body's systems?
- Have you had your ferritin and iron levels checked?
- Have you had your inflammatory markers checked?
- Any autoimmune diseases in your family?
- Any chronic sinus issues/ear issues?



Questions generated by DLMs

- How long ago did the worsening chronic fatigue continue to cripple my life?
- What kind of activity continues to cripple my life?
- What was the estimated dose of suicide?
- How could body parts feel far less hot in other parts of my life?
- What has caused the gradual cognitive decline?
- What was the reason that increasing depression has a negative effect on me?
- A sleep study consists of what?
- What kind of test does someone try to do at this point?
- What has degenerated my muscles?

Figure 5.3 Reddit is a rich source for bringing crowd perspective in training DLMs over conversational data. On the **left** is a sample post from r/depression_help which sees inquisitive interaction from other Reddit users. At the **top-right** are the FQs asked by the Reddit users in the comments. These FQs are aimed at understanding the severity of the mental health situation of the user and are hence, diagnostically relevant. At the **bottom-right** are the questions generated by DLMs. It can be seen that these are not suitable FQs.

a controlled generation process are essential characteristics of patient-clinician interactions, which are difficult to embed in DLM-based CAs. Therefore, question generation (QG) in mental health is challenging. In this chapter, we will discuss **PRIMATE**, a **PR**ocess knowledge **I**ntegrated **M**ental **h**ealth **d**ata **S**et, that would train deep language models (DLMs) to capture information from user input that can answer questions in clinical questionnaires (see Figure 5.1 for examples in PRIMATE). Unanswered questions are the ones that would be used by agents to generate information seeking questions for user (or patient) [207] [214].

Maximize Uncertainty Handling: A major concern in DLMs is about safety. In the figure 5.4, we illustrate the utility of “sanity checks” that are put into use through process knowledge. The model without **PK** is susceptible to risky generations, which might have severe consequences to patient’s mental health. The concept of “sanity

Table 5.1 This is an example of how a process knowledge-integrated dataset is constructed in collaboration with mental healthcare providers. The leftmost column presents example questions mental healthcare providers (MHPs) asked. The MHPs provided Tag and Rank shown in the rightmost columns representing process knowledge. The middle column provides a series of questions gathered using Google SERP API (<https://tinyurl.com/G-SERP-api>) and Bing Search API (<https://tinyurl.com/bing-search-api>) logically ordered by MHPs.

GAD-7 Question (x)	Paraphrases (Y)	Process Knowledge (P) (Tag, Rank)
Feeling nervous, anxious, or on edge	Do you feel nervous anxious or on edge	(Yes/No,1)
	How likely are you to feel this way	(Degree/frequency,2)
	Any ideas on what may be causing this	(Causes,3)
	Have you tried any remedies to feel less nervous	(Remedies,4)
	Are you also feeling any other symptoms such as jitters or dread	(OSI, 5)
Not being able to stop or control worrying	Do you feel not able to stop or control worrying	(Yes/No,1)
	How likely are you to feel this way	(Degree/frequency,2)
	Any thoughts on what may be causing this	(Causes,3)
	Have you tried any remedies to stop worrying	(Remedies,4)
	Are you also feeling any other symptoms	(OSI, 5)

check” is pretty simple. It introduce constraint-based knowledge¹¹ into DLMs using one of the following methods:

- **Textual Entailment Constraints (TEC)** is a directional relationship between sentences in a response or questions. If the two sentences share semantic relations and logically agree, they are entailed. If the two sentences are synonymous based on the entities they contain, they are neutral. If the second sentence refutes the information in the first sentence, they are contradictory.

¹¹a part of **KiL**

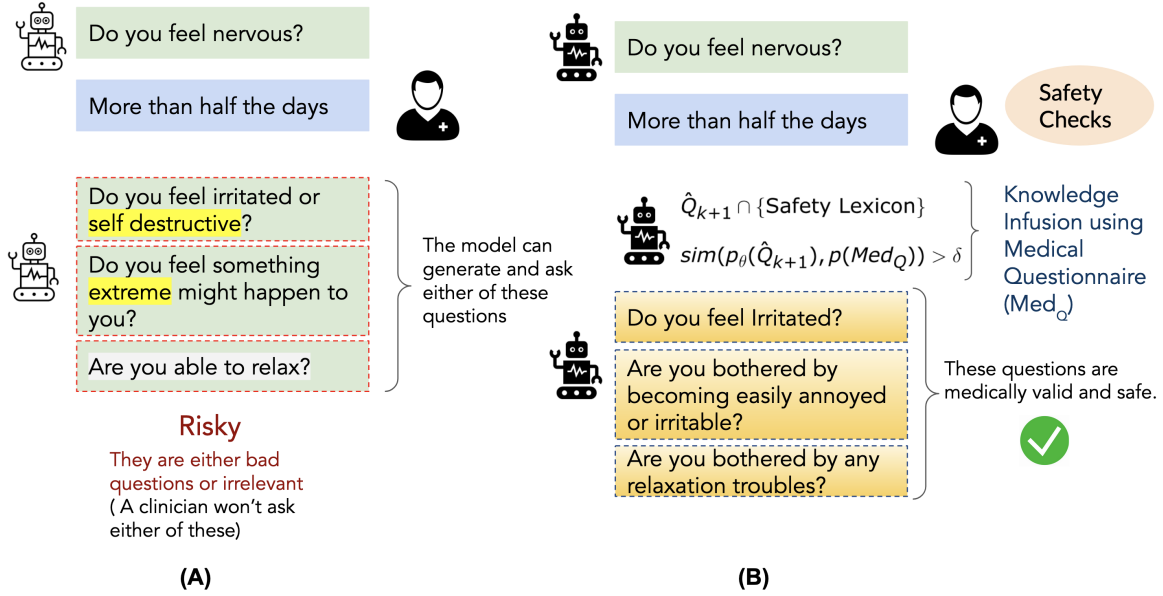


Figure 5.4 Inclusion in sanity checks in the neural response generation or question generation model. The sanity checks entails a check on the quality of generation by computing either jaccard index between the tokens in the generation and medical lexicons or compute cosine similarity between the generated sentence and medical questionnaire (Med_Q). \hat{Q}_{k+1} : Generated Question. $p_{\theta}(\hat{Q}_{k+1})$: Encoding of the generated question and $p(Med_Q)$: Encoding of a question in the list of questions in medical questionnaire. δ : a threshold, above which the generated question is accepted.

Such constraints are manifestations of process knowledge in clinical practice. In machine-understandable form, we can model them as Rules containing Tags and Rank (see Figure 5.1).

- **Rules (Tag and Rank):** These rules can help structure the question generation process, which is random and unsafe in current state-of-the-art NLG models¹². For instance, if the conditional probability function within an AI model, defined as $P(\hat{Q}_{k+1}|\hat{Q}_k)$ is augmented with a Tag containing the following labels: { Yes/No, Degree/Frequency, Causes, Treatment/Remedies } then the model can learn to follow a definite process:

- If \hat{Q}_k is Yes then \hat{Q}_{k+1} is about Degree/Frequency
- If \hat{Q}_k is Degree/Frequency then \hat{Q}_{k+1} is about Causes

¹²<https://tinyurl.com/adaptive-education>

- If \hat{Q}_k is Causes then \hat{Q}_{k+1} is about Treatment/Remedies
- If \hat{Q}_k is Treatment/Remedies then \hat{Q}_{k+1} ask about Information on Other Side Effects

Here, \hat{Q}_{k+1} is the next generated question given \hat{Q}_k , a previous generated and accepted question.

Apart from these, the ability to capture context and provide user-level explanations can be enhanced using **PK**. To understand it better, we created an alternative version of “Reddit C-SSRS Suicidality” dataset, discussed in Chapter 3.

Table 5.2 Attention visualization based explanations in C-SSRS 1.0

Prediction: Suicide Ideation Ground Truth: Suicide Indication Model: Language Model
<p>‘A book is usually what I do when Im getting down, but it doesnt work when I start getting panicky. Ill try the carbs, the caffeine doesnt work because Ive gotten it in a movie theater and had a soda with me...’, ‘A few reasons. I feel backed into a corner mostly. And Im Tired of being Tired of everything. If that makes sense.’, ‘Thank you! I understand its a sad thing. But I also want people to realize that there can be humor in anything and its the best way to deal with this. Its how I would do it. ’, ‘I really dont want to ask for help. Id rather not let anyone know Im having these kind of issues.’</p>

Table 5.3 A process knowledge-guided improved explanations in C-SSRS 2.0

Prediction: Suicide Ideation Ground Truth: Suicide Indication Model: PK-iL(W2V) with Gaussian Kernel
<p>‘A book is usually what I do when Im getting down, but it doesnt work when I start getting panicky. Ill try the carbs, the caffeine doesnt work because Ive gotten it in a movie theater and had a soda with me...’, ‘A few reasons. I feel backed into a corner mostly. And Im Tired of being Tired of everything. If that makes sense.’, ‘Thank you! I understand its a sad thing. But I also want people to realize that there can be humor in anything and its the best way to deal with this. [...]</p>
Explanation: 1. Wish to be dead (<i>no</i>) → Suicide Indication

Table 5.3 is an example of an annotated post in C-SSRS 2.0 and its equivalent in C-SSRS 1.0. The added explanations seems to benefit any LMs in suicidality detection task by enforcing explainability as precursor to classification.

5.3 C-SSRS 2.0

For the sake of convenience, let us take a quick recap of the C-SSRS 1.0, which is discussed by Gaur et al. [37]. Gaur et al. created the “Reddit C-SSRS Suicide Dataset” comprising 500 users and approx 16,000 posts identified as informative for suicidality detection. The dataset was created after multiple rounds of semantic filtering of 94,000 users and 3.4 million posts made from 2005 to 2017 (a.k.a C-SSRS 1.0). C-SSRS 1.0 has been a gold standard (annotator agreement of 0.79) in suicide risk severity detection. It employed clinical knowledge in SNOMED-CT, DSM-5 [10], TwADR lexicon [215], AskaPatient lexicon [215], and i2b2 suicide notes [216] to identify users who have produced an overt signal indicating a positive instance of suicide risk. The label distribution comprises 20% users with suicide indication, 34% users with suicide ideation, 36% users with suicide behavior or attempt. An additional label called “supportive users” constitutes 10% of the dataset. These users share experiences about suicidality.

Creation of C-SSRS 2.0 : To create C-SSRS 2.0, we focused on 450 users with active labels of suicide risk severity. These are *suicide indication*, *suicide ideation*, *suicide behavior or attempt*. We ignored supportive users as they were meant to prevent false alarms.

The task of providing a label to users in the dataset was less time-consuming in terms of maintaining the quality of annotation compared to an annotating task involving marking yes/no against 6 questions for each user while reading their long-winded text. Further, each severity level in the C-SSRS has a main question and sub-sentences (questions in CSSRS : [link](#)). To make the annotation task manageable for four practicing MHPs, we asked them to glean through posts and select 28 users while keeping uniform distribution across suicide risk severity levels. For each user, an MHP provides a sequence of yes/no labels

to each information-seeking sub-sentence in C-SSRS as a proxy of the clinical process, which informs the user’s final suicide risk severity label. There were conflicts between MHPs regarding the explanations they provided for the users’ content and the labels already mentioned in C-SSRS 1.0. After multiple rounds of correction and meetings with the authors of C-SSRS 1.0, we resolved the conflicts, while maintaining high standards in the annotation process [217]. On the 28 users dataset, we attained satisfactory annotator agreement scores of 0.83 on the Fleiss Kappa scale. This part of C-SSRS 2.0 contains 471 posts with an average of 6 sentences per post.

We leverage the annotation of 28 users to annotate the remaining 422 users by computing cosine similarity between the neural representation of the user’s text and questions in the C-SSRS (see Figure 5.5). We used state-of-the-art LMs for generating neural representations of questions and users’ posts. This approach is inspired by Coppersmith et al., who used language models to estimate the similarity between annotated sentences and sentences to be annotated for labeling [218].

The use of LM in the labeling process allowed us to create highlights over parts of the text, which is a part of attention visualization [219]. Next, we provide the outcome to MHPs for evaluation, which involves (1) checking the meaningfulness of yes/no labeling against the highlighted parts of the text for each user and (2) checking its correctness with the final suicide risk severity label provided to the user. In this process, we received an agreement score of 0.74 among the same 4 MHPs who now took the role of evaluators. A certain level of disagreements were hard to resolve, making C-SSRS 2.0 a silver-standard dataset. A snapshot of the dataset is presented [here](#)¹³.

Method for C-SSRS 2.0: In the context of mental health, it can be envisioned as an ordered sequence of questions that an AI algorithm should follow in order to be explainable to MHPs and support acceptable classification of severity of illness in the patient. In this

¹³Though the dataset contains social media content, we did practice anonymity using the guidelines described by Benton et al. [220]. Since we make C-SSRS 2.0 public for research use, we use a Data Use Agreement for responsible dissemination of the dataset [221]

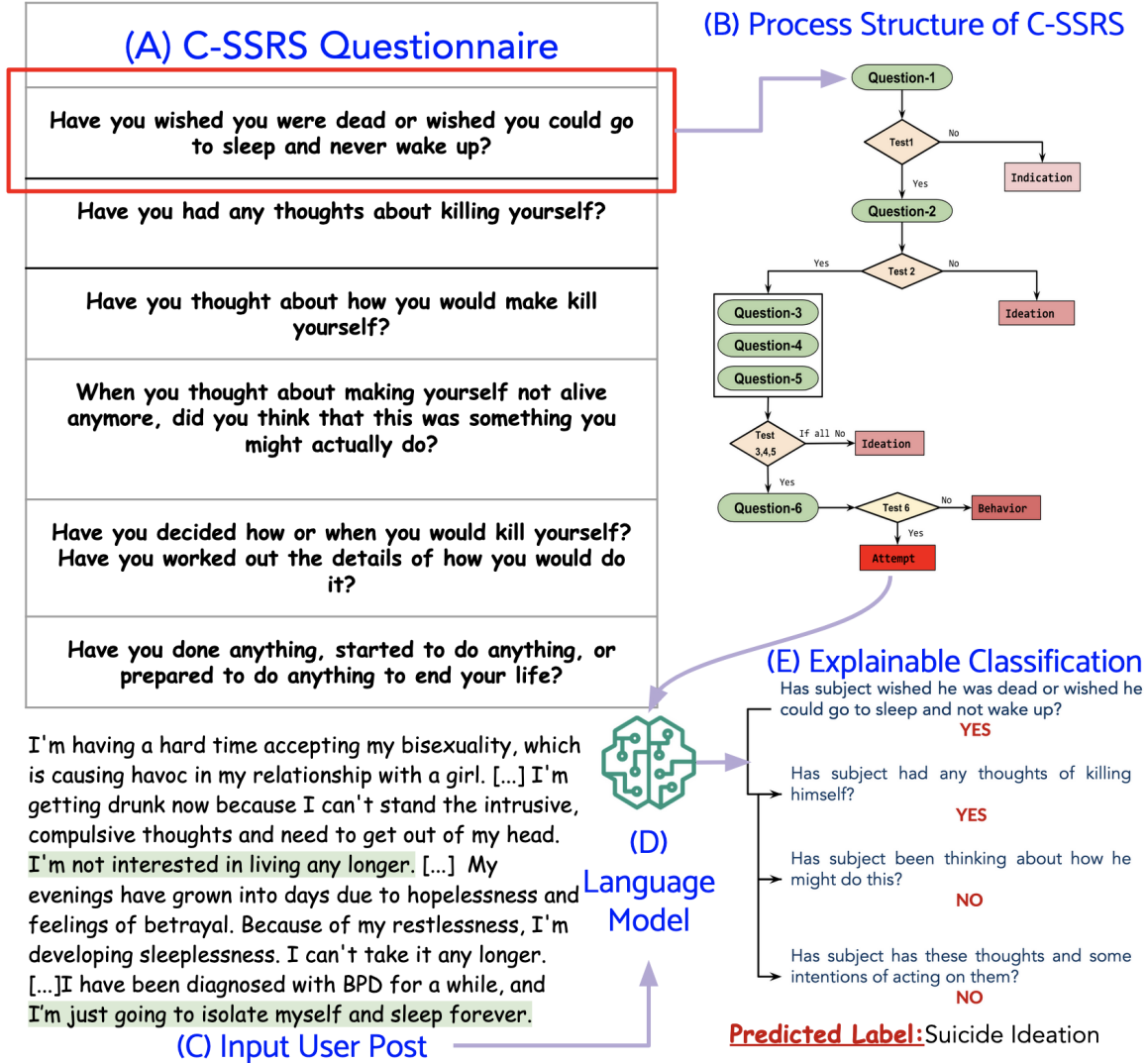


Figure 5.5 An overview diagram of our proposed PK-iL approach using C-SSRS 2.0 Dataset. The process structure of C-SSRS is stacked over a fine-tuned or end-to-end pre-trained LM to provide yes/no responses to questions in C-SSRS. Highlighted parts of the posts contribute to yes in C-SSRS.

section, we define a general algorithm for C-SSRS 2.0, but it is applicable for any other dataset having structural **PK**.

The **PK** we use for suicide risk severity detection is C-SSRS, and it can be visualized as a decision tree according to figure 5.5(B). It makes intuitive sense to utilize a decision tree model to estimate the suicide risk severity level for a user. We formulate the problem for the algorithm as follows: Given a user's post $X = x$, $I_{yes}(q_i)$ and $I_{no}(q_i)$ represents if the post follows a *yes* path or a *no* path to the question q_i in C-SSRS, then the probability of a

label y for a user post can be written as a polynomial of the form.

$$y = \sum_{l \in \text{Leaves}} p_l \prod_{i=1}^{N_q} (I_{\text{yes}}(q_i))(1 - I_{\text{yes}}(q_i)) \quad (5.1)$$

where N_q is the number of questions in the decision tree, which is the number of questions in C-SSRS-like PK. *Leaves* is a set of all leaves that lead to the label y . p_l is computed as the ratio of the number of annotators that chose that path for the example to the total number of annotators - this in some sense captures the inter-annotator agreement in C-SSRS 2.0 [222]. To interpret it, consider a particular post from a user. If among four annotators, two annotators labeled the **PK** as the path $1.2 \rightarrow 2.2 \rightarrow 4$. It is equivalent to $1 \rightarrow 2 \rightarrow 4$ for classifying the user to one of the C-SSRS levels. Then the probability of $y = \text{Behavior or Attempt}$ for that post is 0.50^{14} .

Assertion. *For any model $\mathcal{M}(y)$ that approximates the probability of y for a post according to Eq. 5.1, let the inter-annotator agreement for the post labeled as y be $\mathcal{A}(y)$. Then best approximation for the post, $\mathcal{M}^*(y) \leq \mathcal{A}(y)$*

We claim the above as an assertion instead of a theorem as it is trivial to see that $\prod_{i=1}^{N_q} (I_{\text{yes}}(q_i))(1 - I_{\text{yes}}(q_i)) \leq 1$ always holds, and therefore any approximation is upper bounded by the inter-annotator agreement. Improving upon the inter-annotator agreement means infusing additional external knowledge not present in the ground truth. and hence, gives us the opportunity to label unseen data at par with the human annotators, while explicitly capturing their annotation process in the learned model.

In order to compute $I_{\text{yes}}(q_i)$ and $I_{\text{no}}(q_i)$, where $I_{\text{no}}(q_i) = 1 - I_{\text{yes}}(q_i)$, we compute the similarity between the question q_i in C-SSRS with parts of the sentence highlighted by a LM¹⁵. At a user-level, the similarity can be understood as the inner product between the neural representation of the question and parts of the posts to determine $I_{\text{yes}}(q_i)$ and

¹⁴Note here that the sub-sentences in C-SSRS are not stored in the tree leaves

¹⁵There are several options in NLP literature to construct representations of text. Such as TF-IDF, Hashing Vectorizer, etc. Since LMs show remarkable efficiency, we used them to automate dataset annotation.

$I_{no}(q_i)$. Consider a q_i in C-SSRS; “*Have you thought about being dead or what it is like to be dead*”, it is being answered as *yes* by the response “*Rarely is a day where **I don’t suffer from thoughts of self-harm***” because of the bold-faced phrases. We use a concatenation representation padded with zeros according to the longest text fragment to construct the neural representations of posts semantically. The inner product-based similarity can be formulated as: $\mathcal{S}(x_{sub}^R, q_i^R) \equiv \mathcal{K}\left(\frac{x_{sub}^R}{|x_{sub}^R|}, \frac{q_i^R}{|q_i^R|}\right) \geq \pm\theta_i$, where x^R and q_i^R denotes neural representation of text (x) and question (q_i). \mathcal{K} denote a similarity function. The normalization of the representations by size makes an inner product a valid similarity measure in the range -1 to $+1$. Now, we formally develop the algorithm for the C-SSRS 2.0; Process Knowledge-infused Learning(**PK-iL**).

We define a function that predicts the probability of post label being $Y = y$ according to the **PK** as follows:

$$P(Y = y | X = x) = \sum_{l \in Leaves} p_l \prod_{i=1}^{N_q} \mathbb{1}_{x_{sub} \in x} \left(\mathcal{S}(x_{sub}^R, q_i^R) \geq \pm\theta_i \right) \quad (5.2)$$

where $x_{sub} \in x$ is a fragment of the post x (e.g., word, phrase or sentence). \pm signifies if we are checking if the question q_i is answered as *yes* or *no* by fragment x_{sub} in post x with confidence θ_i . Using $\bigvee_{k=1}^K z_k = (\sum_{k=1}^K z_k \geq 0.5)$, we have:

$$P(Y = y | X = x) = \sum_{l \in Leaves} p_l \prod_{i=1}^{N_q} \sum_{x_{sub} \in x} \left(\mathcal{S}(x_{sub}^R, q_i^R) \geq \pm\theta_i \right) \geq 0.5 \quad (5.3)$$

We can optimize Eq. 5.2 using Bernoulli Loss (\mathcal{L}) for an input post ($X = x$) and label ($Y = x$) as follows:

$$\begin{aligned} \mathcal{L}(\{\theta_i\}_{i=1}^{N_q}) &= P(Y = y | X = x) \log(P(Y = y | X = x)) + \\ &\quad (1 - P(Y = y | X = x)) \log(1 - P(Y = y | X = x)) \end{aligned} \quad (5.4)$$

Since $\mathcal{L}(\{\theta_i\}_{i=1}^{N_q})$ is strongly convex, we use Newton’s optimization method to learn the parameters of the model.

We propose an algorithm for **PK-iL** which is general enough to allow test with different LMs suitable to the task and **PK** suitable to any domain. However, in our experimental

Algorithm 2: Process Knowledge-infused Learning (**PK-iL**)

```
1 Compute  $p_l \forall$  leaves  $l$  from the ground truth;
2 Choose Kernel  $\mathcal{K}$ , fragment size, and CE model for representation
3 Initialize  $\theta_i, \forall i \leftarrow 1$  to  $N_q$ 
4                                      $\triangleright$  Begin Newton's method
5 for  $k \leftarrow 1$  to  $K$  do
6   for  $\theta_i$ , where  $i \leftarrow 1$  to  $N_q$  do
7     Compute  $\theta'_i = \nabla_{\theta_i} \mathcal{L}(\theta_i)$ 
8     Compute  $\theta''_i = \nabla \theta'_i = \nabla_{\theta_i} (\nabla_{\theta_i} \mathcal{L}(\theta_i))$ 
9     Set  $\theta_i = \theta_i - \frac{\theta'_i}{\theta''_i + 1}$ 
10                                      $\triangleright$  add 1 to avoid divide by zero error
11 return  $\theta_i, \forall i \leftarrow 1$  to  $N_q$ 
```

results we will evaluate **PK-iL** both quantitatively and qualitatively evaluation using the C-SSRS 2.0 dataset. Prediction of final suicide risk severity level at a user-level is carried out by choosing the summand in Equation 5.3 that has the highest value once normalized by dividing by the sum of the summands, in order for it to be a probability.

Quantitative Analysis: PK-iL for LMs for C-SSRS 2.0: From figure 5.6, it is interesting that the W2V, trained using the Continuous Bag of Words method, is the best performing model in the Baseline, Cosine Similarity, and the Gaussian Kernel case. Upon inspection of the embeddings, we hypothesize that W2V (which has been trained from scratch on the suicide-related post corpus), captures contextual dependencies between suicidality tokens and phrases much better than LMs. LMs need to be fine-tuned on vast amounts of data to adapt against non-suicidality term-related contexts that they have trained on using massive corpora. Our analysis notes that for domain-specific tasks such as mental health-related prediction, it is perhaps better to train contextual dependencies between words and phrases from scratch, as pre-trained models are already heavily biased towards the contextual dependencies in their training corpora [223]. Across all LMs, we see that **PK-iL** improves the accuracy of the vanilla models by up to almost 15% points for LongFormer. Although to confirm our statement, we have to rule out the effects of collecting more data, adding/deleting features, etc., using neural representations and limited data alone, explic-

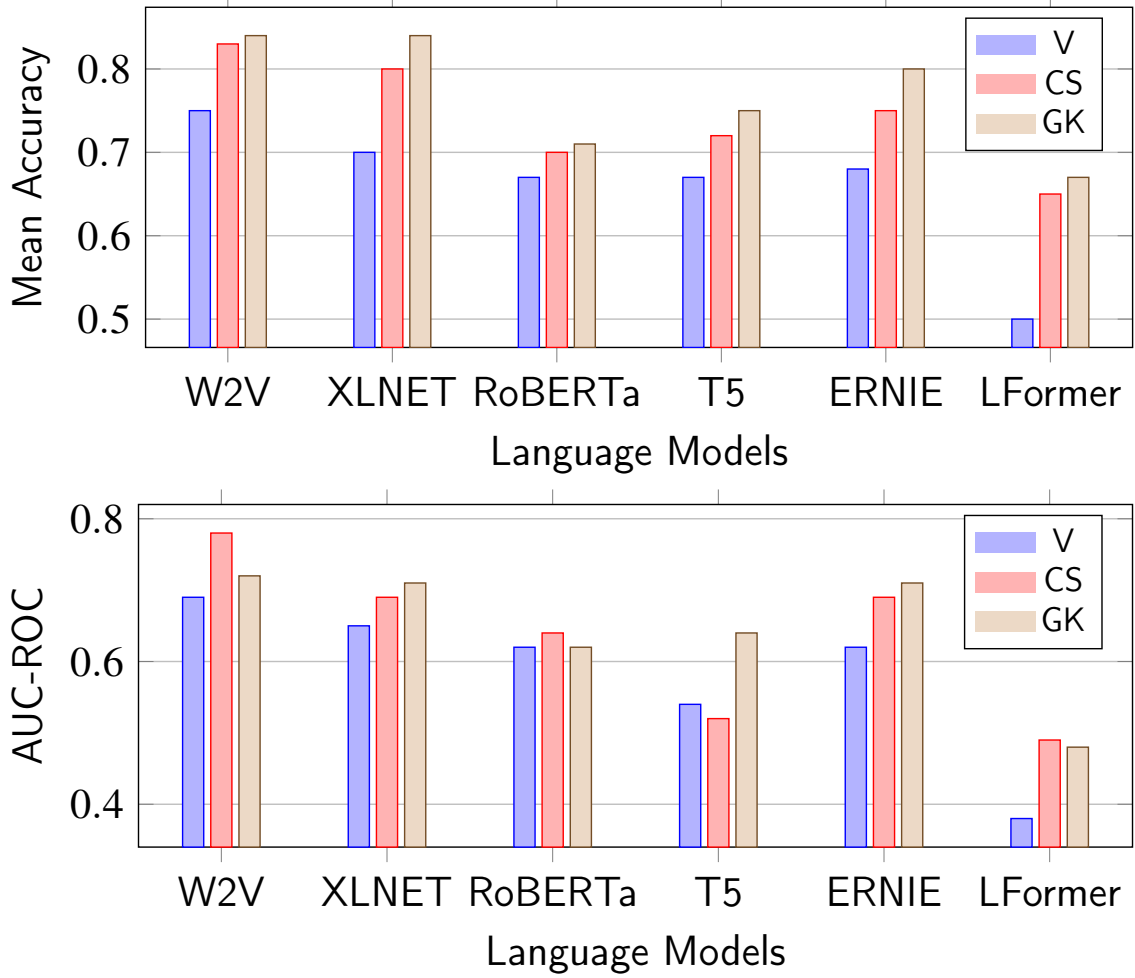


Figure 5.6 Mean Accuracy and AUC-ROC scores, rounded up, for all LMs used in PK-iL algorithm over C-SSRS 2.0 dataset. There are two variants of PK-iL that was evaluated: (a) PK-iL with Cosine Similarity (CS) and (b) PK-iL with Gaussian Kernel (GK) for Kernel choice for each language model of representation. V: The LMs in their vanilla state.

itly controlling the learned model with process knowledge shows significant performance gains.

Qualitative Analysis: In Table 5.2, we see W2V associating phrases and words that characterize a low mood with suicidal ideation. In real life, such words may raise triggers in the minds of a clinician and may benefit their analysis. However, the human annotator seems to have labeled this as an *indication* based on the $x_{sub}^R = \text{“there can be humor in everything”}$ of the post. Recall that **PK-iL** deals with whole fragments of text. The highest threshold

Table 5.4 Explanations provided from W2V on C-SSRS 2.0.

Prediction: Suicide Behavior or Attempt Ground Truth: Suicide Behavior or Attempt Model: Model that uses guidelines, inputs and labels
<p>‘I wish I could give a shit about what would make it to the front page. I have been there and got nothing. Same as my life. I do have a gun.’, ‘I thought [...]. I am not on a ledge or something, but I do have my gun in my lap.’, ‘No. I made sure she got an education and she knows how to get a job. I also have recently bought her clothes to make her more attractive. She has told me she only loves me because I buy her things. ’</p>
Explanation: 1. Wish to be dead (yes) → 2. Non-Specific Active Suicidal Thoughts (yes) → Active Suicidal Ideation with Some Intent to Act, without Specific Plan (yes) → Behavior or Attempt

among the similarity functions in Equation 3 corresponded to the fragment highlighted and path **1**, which is question 1 in C-SSRS ‘*Wish to be dead*’, and hence the model picks indication with probability equal to inter-annotator agreement of $y = \text{suicide indication}$ at that leaf (see Table 5.4 with Table 5.2). To perform an expert evaluation of the explanations provided by W2V and the next best XLNet, we compute Spearman correlation with ground-truth C-SSRS 2.0 for 50 hold-out test users. Vanilla XLNet and **PK-iL**(XLNet) reported 47% and 51% correlations with expert annotated explanations, respectively. Whereas vanilla W2V and **PK-iL**(W2V) reported 66% and 70% correlation, respectively. Overall, subject to annotator agreements, such an explanation is more informative to the clinician about the models’ prediction.

The use of clinical **PK** in dataset creation to support user-level explainability is a new topic and provides a new direction in mental healthcare research using social media content. Prior algorithmic research on either classification of mental health conditions or predicting the severity of a mental health condition has found it hard to explain the inner functioning of their methods and has hence, led to poor adaptation among MHPs [224] [225] [226]. The dataset and it’s associated algorithm that we report here present the first effort toward

making algorithms explainable for MHPs. Next, we will show you another explainable dataset in mental healthcare, designed with a specific objectives: (a) Allow CAs gather what user knows about their conditions, and (b) Ask safe and medically appropriate follow-up question to prevent what is shown in Figure 5.4. The purpose of the next section to prevent CAs fro generating incoherent and irrelevant questions as shown in Figure 5.3.

5.4 PRIMATE

The approach to data collection for PRIMATE involves scraping posts and comments from r/depression_help, a subreddit on Reddit, which is meant to provide advice and support to help individuals suffering from depression. The posts on this subreddit contain flair tags such as *SEEKING HELP*, *SEEKING ADVICE*, and *REQUESTING SUPPORT*. We filter down the data curated from this subreddit based on the flair tag attribute to retain only *advice*, *help* or *support* seeking posts and their comments. After filtering, our dataset had approximately 21,000 posts. Each post contains a title, description, and comments. On average, each post has 5 comments. Next, we chunked the main text of each post into smaller groups of sentences (chunks) of less than 512 tokens while making sure no sentence is segmented in between. The motivation for chunking is to ensure no context is lost from the post due to the limitation of T5 to process 512 tokens as input (DLMs in general suffer from such representation limits). We also appended the post title to each chunk to ensure that main idea of each post was captured in it's chunks. This curated dataset tests T5's capability to generate FQs similar to any of the questions in the extended PHQ-9 questionnaire.

Extending PHQ-9 to support FQ generation: PHQ-9 questions are subject to different interpretations depending on patient-MHP interaction. Additionally, nine questions are limited in scope for use in tasks like fine-tuning and similarity-based performance evaluations. Therefore, to increase the strength of PHQ-9, we collaborated with MHPs to create

sub-questions for each question in PHQ-9. First, we used Google SERP API¹⁶ and Microsoft Bing Search API¹⁷ to retrieve “People-Also-Ask” questions. For each question, we retrieved 40 questions by manually searching and assessing their relevance to PHQ-9 questions. Next, we provided the set of 360 questions to three MHPs for assessment. MHPs evaluated the questions on two grounds: (a) Whether they would ask such a question to a patient? (relevance) (b) If yes, when should such a question be asked? (rank). Based on their ratings, we created a final set of 134 sub-questions for the nine questions in PHQ-9¹⁸ resulting in a total of 143 questions.

Models for FQ Generation: We used an off-the-bench T5-base QG model that was fine-tuned on the SQuAD 2.0 question generation dataset [227] [**Model 1**]. Next, we fine-tuned Model 1 on r/depression_help posts and comments. To align with our task of making T5 generate relevant FQs, we filtered out comments which were non-interrogative. We kept only the interrogative statements asked by Reddit users in the comments [**Model 2**]. Not all interrogative comments by Reddit users are *diagnostically relevant* FQs (Eg: “Can you use MS Excel?”, “Were you interactions on FaceTime?”). To remove such questions, we further filtered the dataset by calculating the maximum BLEURT score between the question (present in the comments) and the questions in extended PHQ-9. We applied a threshold of 0.60 to this score¹⁹. This removed harmful and diagnostically irrelevant questions while preserving contextual, semantically relevant, and legible questions [**Model 3**]. See Fig 5.3 for examples of diagnostically relevant questions.

Analysis of Models for Question Generation: Out of the 21k questions, performance of Models 1, 2, and 3 were examined on those 2003 posts that had at least one interrogative comment. Each of the three models was made to generate FQs in sets of 5, 10, and 15

¹⁶<https://serpapi.com/>

¹⁷<https://www.microsoft.com/en-us/bing/apis/bing-web-search-api>

¹⁸Questions in extended PHQ-9 : link

¹⁹empirically judged

Table 5.5 Examples of questions generated by T5 when tasked to generate FQs when the user query for the **post** in Figure 5.3 was provided as input. **Model 1**, which is a pre-trained T5 [2], often generates questions which are irrelevant, unsafe, incoherent, and redundant. **Model 2**, which is T5 fine-tuned on r/depression_help seems to be relatively coherent and inquisitive compared to **Model 1**. However, both models generate questions about the topic that user has discussed in their query. As a result, we see that pre-trained and fine-tuned DLMs fail to generate FQs. By enforcing FQ generation using using a dataset curated using extended PHQ-9, generated questions have been mostly inquisitive. This is shown by **Model 3**. Still, a lot of generations are around the problem the user mentioned.

Model 1 QGs	Model 2 QGs	Model 3 QGs
<ul style="list-style-type: none"> • How long ago did the worsening chronic fatigue continue to cripple my life? [<i>*User already mentions in the title of the post he has been facing the fatigue for years</i>] • What was the estimated dose of suicide? [<i>*Unsafe incoherent question</i>] • How could body parts feel far less hot in other parts of my life? [<i>*Incoherent question</i>] • What has caused the gradual cognitive decline? [<i>*It's the physician who needs to answer this question after the conversation</i>] • What effect is that I'm less optimistic, and it's embarrassing ? [<i>*Incoherent attempt to generate a question from the sentence in the post which mentions "I am still pretty optimistic , but I 'm just too out of it , it 's embarrassing"</i>] • What was the reason that increasing depression has a negative effect on me? [<i>*Irrelevant and incoherent</i>] • A sleep study consists of what? [<i>*Irrelevant to the diagnostic procedure</i>] • What did Caffeine feel in my body? [<i>*Wrong pronoun used and irrelevant to the diagnosis</i>] 	<ul style="list-style-type: none"> • Have you ever tried many depressants or mental health things? [<i>*Redundant question as user has already mentioned in the post text that he has tried antidepressants</i>] • Are you seeing a therapist? [<i>*Redundant question as user had already mentioned in the post text that he has seen a therapist in the past</i>] • Do you even ever try to focus on negative things and/or pain, relief, and patience? [<i>*Incoherent question</i>] • And what if you forgot to fucking realize that, then it will be fucking so much that you'll be fucking fucking can we ask what they want. [<i>*Unsafe objectionable question</i>] • Do you mean heavy bipolar disorder? [<i>*Irrelevant to the context of the post</i>] • Are you aware about the qualities of antidepressants? [<i>*Inquisitive in nature but irrelevant from a diagnostic perspective</i>] • What does antidepressant work at? [<i>*Irrelevant from a diagnostic perspective</i>] • Do you have a close friend you can go to who you can talk with, that way you can get out of the house? [<i>*FQ</i>] 	<p>Are you on any antidepressants? [<i>*User mentions in the post he has tried them before but no mention has been made if he is still on them</i>]</p> <ul style="list-style-type: none"> • Do you have any positive thoughts? [<i>*FQ</i>] • Are you in any danger of hurting yourself? [<i>*Slightly unsafe</i>] • Is it that you aren't happy with your feelings? [<i>*Irrelevant question</i>] • Have you tried some exercise? [<i>*Redundant question as user already mentions he has tried it</i>] • Do you wake up frequently? [<i>*FQ</i>] • How is your sleep quality? [<i>*FQ</i>] • When you wake up, what do you do? [<i>*FQ</i>] • Is there anything that helps you calm the symptoms for now? [<i>*FQ</i>] • What are your hobbies? [<i>*Generic FQ</i>] • What are your interests? [<i>*Generic FQ</i>]

Table 5.6 In this example, the generated questions from both Model 2 and Model 3 seem to be relevant FQs, but they are not assessing the severity of the mental health condition, despite Model 3 being fine-tuned on a dataset filtered by PHQ-9 questions. In comparison to the qualitative outcome in Table 5.5, this showcases the inability of T5 to support mental health triage.

Post text	Model 1 QGs	Model 2 QGs	Model 3 QGs
I feel depressed . I have felt depressed for a while. I have done a lot of things to try and help myself <i>through therapy</i> ,[...] , staying busy , going to university . all I have found is that each time I complete any of these things that I still feel like I am empty . When I do not do those things for more than 10 mins by myself that my thoughts just get dark and I be little everything I have ever done . [...] Also I get depressed because I have dreams where I ’m constantly bombarded by women who are not my wife who want to have relationships with me and it ’s a nightly battle to not give in to it when I ’m dreaming . I know it sounds silly but I fear that if I give into the dreams that I might do it in real life . <i>I do n’t want to tell my wife</i> because she does n’t understand other weird dreams that I have had in the past .	<ul style="list-style-type: none"> • What if I do not do it for more than 10 mins by myself? [<i>*Closed book question already answered in the post and irrelevant to the diagnostic procedure</i>] • What might happen to any women that want to have relationships with me? [<i>*Irrelevant to the diagnostic process</i>] • What do the women at work do? [<i>*Irrelevant to the diagnostic process</i>)] • What kind of dreams are I was depressed? [<i>*Closed book question generated directly from a statement in the post</i>] 	<p>Would you be happy with them? [<i>*Irrelevant to the diagnostic process</i>]</p> <ul style="list-style-type: none"> • What’s your job? [<i>*Irrelevant</i>] • Are you able to reach out to your doctor or psychiatrist? [<i>*User has already mentioned in the original post that he has talked to his therapist</i>] • what are your dreams? [<i>*Closed book question already answered in the post</i>] • What do you enjoy about your dreams? [<i>*FQ</i>] • Do you want to talk to people about what dreams are ? [<i>*FQ</i>] • Do you feel as satisfied as yourself when you are dreams? [<i>*FQ</i>] 	<p>Have you talked to a therapist? [<i>*Already answered by the user in the post itself</i>]</p> <ul style="list-style-type: none"> • Is there anything that helps you calm your feeling for now? [<i>*FQ</i>] • “Have you ever gotten any help?” [<i>*Already answered in the query</i>] • Do you feel isolated? [<i>*FQ</i>] • What are your hobbies? [<i>*FQ</i>] • What are your interests? [<i>*FQ</i>] • How long have you been waiting for your wife to talk about these dreams? [<i>*FQ</i>] • Have you told your wife you’re depressed or not? [<i>*Inquisitive in nature but already answered by the user in original post</i>]

through nucleus sampling [201]. For a generated question, BLEURT score was computed with each question in Extended PHQ-9 and the maximum among those scores was taken as the score for the generated question. A clear distinction between models 1, 2, and 3 is the nature of the questions asked. Model 1 generated closed book questions, whereas Model 2 and 3 seem to show some inquisitive nature and seem more focused on the mental health

domain, which can be attributed to the after effect of finetuning on Reddit (see Table 5.5 and 5.6). We captured the performance of the models quantitatively using 'hit rate' as a metric. For a generated question (\hat{q}), we denote :

$$score(\hat{q}) = \max(bleurt_score(\hat{q}, q_1), bleurt_score(\hat{q}, q_2), \dots, bleurt_score(\hat{q}, q_{143}))$$

where $q_1, q_2, \dots, q_{143} \in \text{Extended-PHQ-9}$. Across all 2003 posts, we had $C = 2575$ chunks²⁰. Let total number of questions generated by a model be $|\hat{\mathbf{Q}}|$ and $|\hat{Q}|$ denote the number of question generated by the model for a given chunk. For experimentation, we set $|\hat{Q}|$ to have values $\{5, 10, 15\}$. Thus, $|\hat{\mathbf{Q}}| = |\hat{Q}| * C$. Then the **Hit Rate** for a model was computed as:

$$\text{Hit Rate}(\text{model}, |\hat{Q}|) = \frac{\sum_{\hat{q} \in \hat{\mathbf{Q}}} \mathbf{I}(score(\hat{q}) > \delta)}{|\hat{\mathbf{Q}}|}$$

where δ is the threshold on the similarity between generated question in a chunk and sub-questions in PHQ-9 and $I[\varphi]$ is the indicator function taking values 0 or 1 for a predicate φ .

Inference on Table 5.7: (1) Regardless of fine-tuning and filtering based on PHQ-9 questions, inherently, T5 does not capture the meaning and usage of the words in the mental health context. Moreover, T5 fails to generate legible and relevant FQs as safe as PHQ-9 questions. Therefore, we scrutinize the generated FQs by mapping them to most similar questions in extended PHQ-9. Examples of irrelevant generations by T5 that it thought were relevant are: (a) “Wtf?” (generated FQ) was found most similar to “Do you have hope?” (PHQ-9) (b) “What did Boyfriend suffocate me with during his break up a week after I got a diagnosis?” (generated FQ) was found most similar to “What do you think makes you a failure” (PHQ-9). The previous generated question is redundant as the answer to it was already present in the original post. **(2)** Many generated questions contain extreme language due to the informal nature of the Reddit platform, which is very sensitive issue, especially in the mental health domain. Examples are: “Did you f****ing realize

²⁰Chunking was done as DLM accepts a maximum input length of 512 tokens.

Table 5.7 Experimental results comparing different models in generating questions that match the sub-questions in PHQ-9. \hat{Q} is the set of generated questions in each chunk. The performance is recorded over all the generated questions (\hat{Q}). δ was used as the threshold on the similarity between generated question and PHQ-9 sub-questions while calculating hit rate. BLEURT records semantic similarity, whereas Rouge-L records the longest common subsequence exact match between generated question and PHQ-9 sub-questions. The highest performance on semantic and string similarity is bolded. Acceptable performance in Model 3 achieved using PHQ-9 motivated us to prepare **PRIMATE**.

$ \hat{Q} (\downarrow)$	Hit Rate on BLEURT			Hit Rate on Rouge-L		
$\delta(\rightarrow)$	0.4	0.5	0.7	0.4	0.5	0.7
Model 1: Pre-trained T5						
5	0.5417	0.1233	0.0020	0.1241	0.0386	0.0005
10	0.5400	0.1203	0.0010	0.1290	0.0400	0.0010
15	0.5368	0.1250	0.0013	0.1266	0.0384	0.0009
Model 2: Fine-Tuned T5 on r/depression_help						
5	0.6657	0.2804	0.0097	0.3445	0.1560	0.0100
10	0.6691	0.2792	0.0104	0.3481	0.1590	0.0098
15	0.6726	0.2787	0.0104	0.3476	0.1588	0.0094
Model 3: T5 Fine-tuned on r/depression_help filtered by PHQ-9						
5	0.9489	0.7088	0.1261	0.7457	0.4937	0.0903
10	0.9542	0.7126	0.1272	0.7460	0.5002	0.0947
15	0.9514	0.7098	0.1274	0.7484	0.4945	0.0916

that f***ing people are f***ing too?” (generated FQ) was found to be the most similar to “What do you think makes you a failure?”. Thus, T5 and its variants need to capture “what the user knows and has already mentioned in his post” by checking which PHQ-9 questions are already answerable using the user’s post before generating the next probable FQs in order to avoid redundancy.

Creation of PRIMATE: We present **PRIMATE**, a dataset consisting of Reddit posts containing user situations describing their health conditions and whether the questions in PHQ-9 are answerable using the content in the posts. Each question is attributed with a binary “yes” or “no” label stating whether the user’s description already contains the answer to that question (see Table 5.8). **PRIMATE** was created from a month long annotation-evaluation

Table 5.8 Distribution of 2003 posts in **PRIMATE** according to whether the text in the post answers a particular PHQ-9 question. Through this imbalance, **PRIMATE** presents its importance in training DLM(s) to identify potential FQs in PHQ-9 that would guide a generative DLM(s) to conduct a discourse with a patient with a vision to assist MHPs in triage. Q1-Q9 are described in Figure 5.7

PHQ-9 Questions	Number of Posts	
	With Answer (Yes)	W/o Answer (No)
Q1	1679	324
Q2	1664	339
Q3	686	1317
Q4	949	1054
Q5	530	1473
Q6	195	1808
Q7	741	1262
Q8	196	1807
Q9	374	1629

A User's Post	Process Knowledge Annotation using PHQ-9
<i>Should I use the psychological help service that my university provides for free ?.</i>	Q1: Feeling bad about yourself or that you are a failure or have let yourself or your family down, YES
Lately I have been [feeling really low (Q2, Q3)].	Q2: Feeling down depressed or hopeless, YES
[I can't make myself leave the bed (Q3, Q9)],	Q3: Feeling tired or having little energy, YES
[I start crying out of the blue and everything is just so heavy (Q1, Q4)]. I think I have [always suffered from some kind of depression (Q2)] but I have never been to therapy because [I could not afford it (Q1)] on my own and [my family did not ever suspect anything (Q1)]. Now I live on my own in another city .	Q4: Little interest or pleasure in doing things, YES
[...] my university provides psychological help for students for free . Do you think I should give it a go ? [.....] I have nothing to lose because it's free . Did you ever try anything like that ?	Q5: Moving or speaking so slowly that other people could have noticed Or the Opposite being so fidgety or restless that you have been moving around a lot more than usual, NO
	Q6: Poor appetite or overeating, NO
	Q7: Thoughts that you would be better off dead or of hurting yourself in some way, NO
	Q8: Trouble concentrating on things such as reading the newspaper or watching television, NO
	Q9: Trouble falling or staying asleep or sleeping too much, YES

Figure 5.7 A post in **PRIMATE** which is annotated with PHQ-9. The questions marked “YES” are answerable by DLMs using the mental health specific cues from user text. The questions marked “NO” are the questions a DLM should consider asking as FQs. Sentences within [] were taken as signals that the “YES” marked questions had already been answered in the post .

cycle between MHPs and crowd workers. A total of five crowd workers performed this task, achieving an initial annotator agreement of 67% using Fleiss kappa. Subsequently, the MHPs assessed the quality of annotations and provided their suggestion for improvement,

Table 5.9 The MCC score for all 9 questions across different thresholds is in the range 0 to +1 (low to high positive relationships). The MCC for some configurations runs into a divide by zero error, and we replace this value with 0.0. **Unable**: model is unable to learn cues to determine answerability in a post. **Maybe**: model is uncertain whether a particular PHQ-9 question is answerable or not. **Certain**: answerability can be determined by the model with high reliability. Class-Type: Classification Type when $\delta = 0.9$

$\delta (\rightarrow)$	0.5	0.7	0.9	Class-
PHQ-9(\downarrow)	MCC	MCC	MCC	Type
Q1	0.0	0.17	0.17	Unable
Q2	0.43	0.45	0.52	Certain
Q3	0.41	0.46	0.33	Maybe
Q4	0.14	0.19	0.13	Unable
Q5	0.63	0.65	0.66	Certain
Q6	0.47	0.43	0.27	Unable
Q7	0.66	0.68	0.7	Certain
Q8	0.1	0.0	0.0	Unable
Q9	0.62	0.56	0.39	Maybe

leading to an acceptable agreement score of 85%. A sample annotated post in **PRIMATE** is shown in Figure 5.7.

Method for PRIMATE: With PRIMATE, we propose the question generation task as a binary classification problem followed by question generation. We train BERT²¹ (a transformer-based DLM) as a classifier on the **PRIMATE** dataset. We report the Matthews Correlation Coefficient (MCC) scores in table 5.9. MCC is a reliable metric to assess a model’s classification over an imbalanced dataset, particularly useful when we are interested in all four categories of confusion matrix: true positives (answerable questions (AQ)), true negatives (FQ candidates), and false alarms (false negatives and positives). As **PRIMATE** shows a disproportional distribution of AQs (yes) and FQs (no), MCC is an appropriate metric [229]. We base our analysis on the consistency of BERT classifier on varying threshold (δ) in table 5.9. A score between 0.0 to 0.30 (Type **Unable**) on MCC means the model is only able to find a negligible to weak positive relationship between input and output. In our context, a score in this range for a particular PHQ-9 question means

²¹BERT end-to-end training perform well compared to baselines Electra [197], and MedBERT [228]

that model is unable to effectively learn the cues needed to judge the answerability of that question in user posts. A score between 0.30 and 0.40 (Type **Maybe**) means that the model is able to learn a moderately positive relationship, interpreted as ambiguity in the model to judge whether a particular PHQ-9 question is answerable from user posts. MCC scores between 0.40 to 0.70 (Type **Certain**) for a question in PHQ-9 means that the model can effectively judge whether that question is answerable in user posts . Any score above 0.70 makes the model’s judgements even more reliable.

Future Direction in PK for Mental Health: Our experiments show that PRIMATE can train DLMs to judge *whether a user’s description of their mental health condition already contains an answer to a particular question in PHQ-9*, which would eventually guide coherent FQ generations. We leave our approach for FQ generation as future work on process knowledge²² [17]. Further, we are yet to scale our understanding to other mental health disorders, such as anxiety using GAD-7 and Suicidality using C-SSRS [230]. Further, we are yet to investigate whether PRIMATE, along with the knowledge in SCID can make DLMs transferable across multiple mental health disorders, especially the ones comorbid with depression. Also, there is a need for a clinically explainable safety metric for our task.

5.5 SUMMARY

This chapter demonstrated the importance of data and process knowledge to adapt DLMs for generating FQs that would assist MHPs in triaging depression. Our experiments show that without process knowledge, DLMs hallucinate by generating unsafe, incoherent, and irrelevant questions that are not helpful for MHPs in pre-screening or triaging. The challenge lies in the inability of the DLMs to judge from the set of generated questions, which is a potential effective FQ to ask based on the user information. The improved question generation performance of DLMs fine-tuned on conversational data filtered by process knowledge encouraged us to prepare **PRIMATE** and **C-SSRS 2.0**. Both **PRIMATE** and

²²under review; contact Manas Gaur

C-SSRS 2.0 can train DLMs to judge ‘whether a user’s description of their mental health condition already contains an answer to a particular question in PHQ-9 or C-SSRS, which would eventually guide coherent FQ generations.

Methodologically, we demonstrated two forms of knowledge infusion; (a) Shallow Infusion in PRIMATE and (b) Semi-Deep Infusion in C-SSRS 2.0. The shallow infusion in PRIMATE aims to prevent hallucination and minimize uncertainty. The semi-deep infusion also prevents hallucination and minimizes uncertainty, along with it, provides user-level explanations. However, we are yet to scale our understanding of other mental health disorders. Further, we are yet to investigate whether PRIMATE and C-SSRS 2.0, along with the knowledge in SCID, can make DLMs transferable across multiple mental health disorders, especially those comorbid with depression and suicide.

Certain challenges remain, despite Shallow and Semi-Deep Infusion of process knowledge. First, by enforcing knowledge-based filtering at the input (Shallow Infusion), we restrict the exploration capabilities of DLMs (reduction in the degree of freedom). What if we allow selective knowledge-guided filtering and infusion at every layer of DLMs to estimate which questions a DLM can answer and which it is not. Second, enforcing the structure in PK into a DLM using a decision tree does not account for abstraction and specificity, the two critical components in mental health interviews (or any general interview). What if we construct a knowledge-infused DLM, which upon getting a user’s input, asks questions in the order defined clinical questionnaires. To work towards these challenges, we will lay the groundwork for Deep Infusion as the future of Neuro-symbolic AI.

CHAPTER 6

DEEP KNOWLEDGE INFUSION

We define the third category of knowledge infusion, i.e., *Deep Infusion of Knowledge*, as a paradigm that couples the latent representation learned by deep neural networks with KGs exploiting the semantic relationships between entities. This chapter will provide theoretical background to achieve *Deep Infusion*, as illustrated in Figures 6.1 and 6.2. We aim to:

- Quantify the information loss.
- Identify the relevant knowledge at an appropriate level of abstraction.
- Appropriately combine identified concepts in KGs with a latent representation of data.

Significance

Deep infusion supports the stronger weaving of different forms of knowledge at different levels of abstraction, that typically map to different layers in a deep neural network architecture.

Recent research that are inline with the general theme of KiL leverage what is called as parametric knowledge [61]. For instance, Dai et al. define a term a called “Knowledge neuron” within deep neural network which is described as a hidden node that represents the knowledge infused at the input [231]. Further, we argue that there is a subtle difference between the terms *knowledge infusion* and *knowledge injection* [232]. Methods under the latter theme are similar to *Shallow Infusion* and *Semi-Deep Infusion* categories. Whereas, *knowledge infusion* is about the use of stratified representation of knowledge

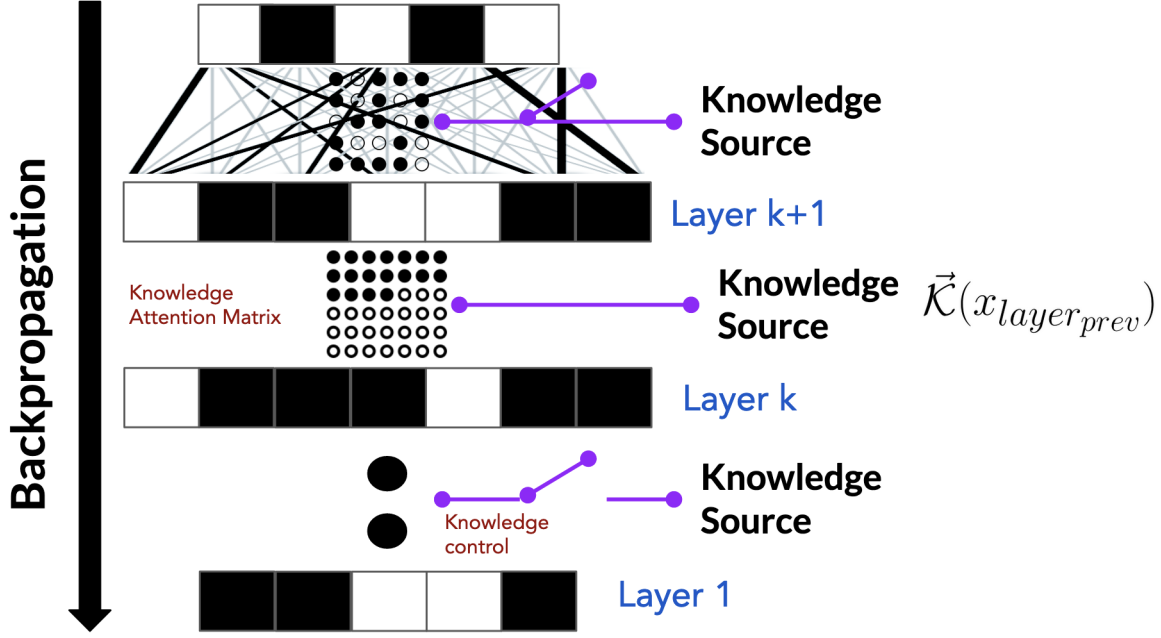


Figure 6.1 An illustration of deep knowledge infusion. The procedure provides an improvement over existing DL architectures by including (a) layer-wise knowledge augmentation(\mathcal{K}) and (b) monitoring correct infusion through knowledge attention matrix. The later component controls the information flow between the previous layer ($x_{layer_{prev}}$) and the next layer ($x_{layer_{next}}$).

representing different levels of abstraction that would be merged at various layers of deep neural network and not just at the input. As we understand the levels of abstraction represented by different layers in a deep neural network, we can look to transfer knowledge that aligns with the corresponding later in the layer-wise learning process in DL. We argue that Deep Infusion within the latent layers of neural networks will boost the performance of neural networks as an integral component of AI models deployed in applications. With a Deep Infusion of such structured knowledge, it will reveal patterns missed by shallow and semi-deep infusions because of sparse feature occurrence, ambiguity, and noise. At the same time, Deep Infusion would retain the explainability, interpretability, and uncertainty handling aspect of ML/DL. In this chapter, we will lay down theoretical positioning of the Deep Infusion and discuss novel technological components that are required for knowledge infusion in current and most-widely used neural language models.

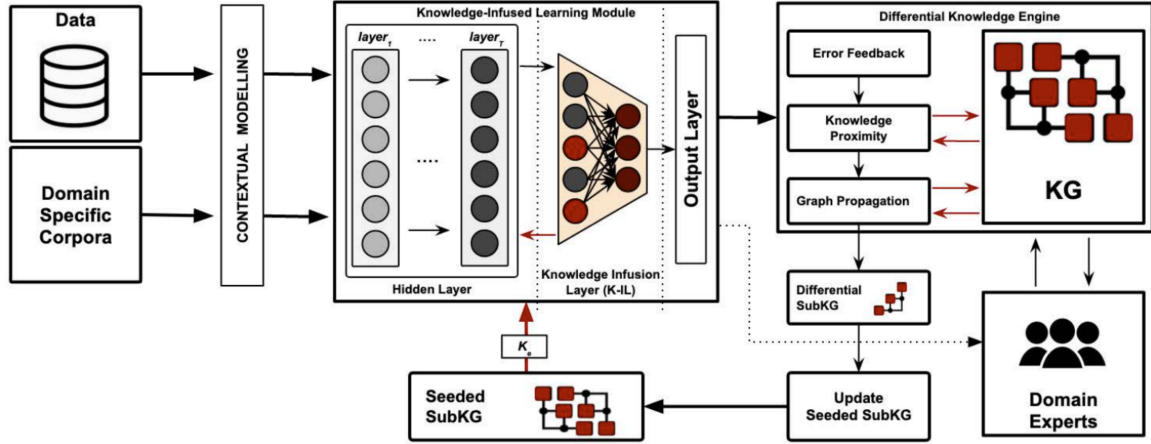


Figure 6.2 Overall Architecture: Contextual representations of data are generated, and domain knowledge amplifies the significance of specific important concepts that are missed in the learning model. Classification error determines the need for updating a Seeded SubKG with more relevant knowledge, resulting in a Seeded SubKG that is more refined and informative to our model.

6.1 DEEP INFUSION MODULE

Each layer in a neural network architecture produces a latent representation of the input vector. As neural network consists of an input layer, hidden layers and output layer, external information has been incorporated before the input layer and after the output layer. Infusion after the input, within the hidden layer or before the output layer have not been investigated. we infuse knowledge within the neural network while the latent representation is transmitted between layers including hidden layers. The infusion of knowledge during the representation learning phase raises the following central research questions, (i) *Knowledge-Aware Loss Function (K-LF)*: How do we decide whether to infuse knowledge or not at a particular stage in learning between layers, and how to measure the incorporation of knowledge? (ii) *Knowledge Modulation Function (K-MF)*: How to merge latent representations with knowledge representations, and How to propagate the knowledge through the learned representation?

Configurations of neural networks can be designed in various ways depending on the problem. As our aim is to infuse knowledge within the neural network, such opera-

tion can take place (i) before the output layer (e.g., SoftMax), (ii) between hidden layers (e.g., reinforcing the gates of an NLM layer, modulating the hidden states of NLM layers, Knowledge-driven NLM dropout and recurrent dropout between layers). To illustrate (i), we describe our initial approach to neural language models that fuses knowledge before the output layer.

Following the figure 6.2, in the subsequent subsections, we explain: (a) Creation of Knowledge representations (e.g., Knowledge embeddings, K_e), (b) Knowledge Infusion Layer is responsible for the two proposed functions. In these subsections, we provide an initial approach that, we believe, will shed the light towards a reliable and robust solutions with more research and rigorous experimentation.

K_e : Knowledge Embedding Creation We generate representation of knowledge in the Seeded SubKG as embedding vectors. We create an embedding of each concept and their relations in the Seeded SubKG using the perspective models (R, I, V), and merge these embeddings through the proximity of their concepts and relations in the graph. Unlike traditional approaches that compute the representation of each concept in the KGs by simply taking average of embedding vectors of concepts, we leverage the existing structural information of the graph. This procedure is formally defined:

$$K_e = \sum_{ij} [C_i, C_j] \otimes D_{ij} \quad (6.1)$$

where K_e is the representation of the concepts enriched by the relationships in the Seeded-KG, (C_i, C_j) is the relevant pair of concepts in the Seeded-KG, D_{ij} is the distance measure (e.g., Least Common Subsumer [233]) between the two concepts C_i and C_j . We will further examine novel methods building upon our initial approach above as well as existing tools that include TRANS-E [234], TRANS-H [235], and HOLE [236] for the creation of embeddings from KGs.

Knowledge Infusion Layer: In a many-to-one NLM [237] network with \mathbf{T} hidden layers, the \mathbf{T}^{th} layer contains the learned representation before the output layer. The output layer

Algorithm 3: Routine for Infusion of Knowledge in NLMs

```
1 Data :  $NLM_{type}, \#Epochs, \#Iterations, K_e$ 
2 Output:  $\vec{M}_T$ 
3 for  $ne=1$  to  $\#Epochs$  do
4   Compute  $\vec{h}_T, \vec{h}_{T-1} \leftarrow \text{TrainingNLM}(NLM_{type}, \#Iterations)$ 
5   while  $(\mathbf{D}_{KL}(\vec{h}_{T-1} || \vec{K}_e) - \mathbf{D}_{KL}(\vec{h}_T || \vec{K}_e)) > \varepsilon$  do
6      $\triangleright \varepsilon$ : acceptance threshold
7     Compute  $h_T \leftarrow \sigma(W_{hk} * (\vec{h}_T \oplus \vec{K}_e) + b_{hk})$   $\triangleright \sigma$ : sigmoid activation
8     Compute  $W^{hk} \leftarrow W^{hk} - \eta_k \nabla(\text{K-LF})$   $\triangleright \eta$ : learning rate
9     Compute  $\vec{M}_T \leftarrow \vec{h}_T \odot W^{hk}$ 
10 return:  $\vec{M}_T$ 
```

(e.g., SoftMax) of the NLM model will estimate the error to be back-propagated. As discussed above, knowledge infusion can take place between hidden layers or just before the output layer. We will explore techniques for both scenarios. In this subsection, we explain the Knowledge Infusion Layer (*Ki-layer*) which takes place just before the output layer.

Algorithm 3 takes the type of NLM, number of epochs, iterations and the seeded knowledge graph embedding K_e as input, and returns a knowledge infused representation of the hidden state \mathbf{M}_T . In line 4, the infusion of knowledge takes place after each epoch without obstructing the learning of the vanilla NLM model and is explained in lines 5-10. Within the knowledge infusion process (lines 7-9), we optimize the loss function in equation 2 with convergence condition defined as the reduction in the difference between the \mathbf{D}_{KL} of h_T and h_{T-1} in the presence of K_e . Considering the vanilla structure of a NLM, \mathbf{M}_T is utilized by the fully connected layer for classification.

To illustrate an initial approach in figure 6.3, we use LSTMs as NLMs in our neural network. *Ki-layer* functions add an additional layer before the output layer of our proposed neural network architecture. This layer takes the latent vector (h_{T-1}) of the penultimate layer, the latent vector of the last hidden layer (h_T) and the knowledge embedding (K_e), as input. In this layer, we define two particular functions that will be critical for merging the latent vectors from the hidden layers and the knowledge embedding vector from the KG.

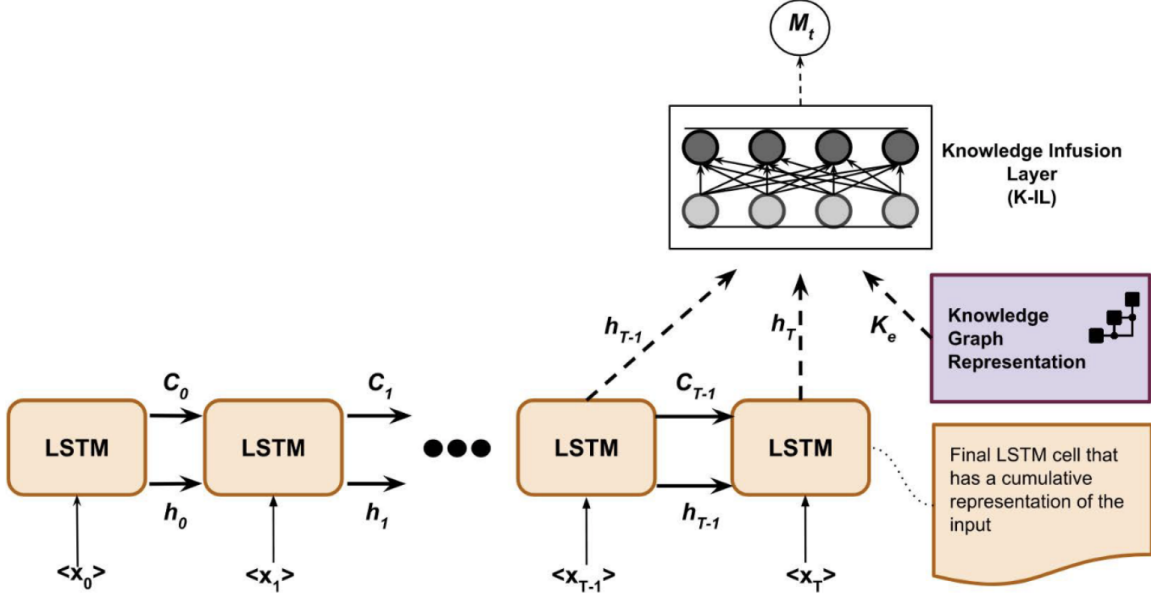


Figure 6.3 Inner Mechanism of the Knowledge Infusion Layer in an LSTM Network

Note that the dimensions of these vectors are the same because they are created from the same models (e.g., contextual models), which makes the merge operation of those vectors possible and valid.

Knowledge-Aware Loss Function (K-LF): In neural networks, hidden layers may de-emphasize important patterns due to the sparsity of certain features during learning, which causes information loss. In some cases, such patterns may not even appear in the data. However, such relations or patterns may be defined in KGs with even more relevant knowledge. We call this information gap between the learned representation of the data and knowledge representation as differential knowledge. Information loss in a learning process is relative to the distribution that suffered the loss. Hence, we propose a measure to determine the differential knowledge and guide the degree of knowledge infusion in learning. As our initial approach to this measure, we developed a two state regularized loss function by utilizing Kullback Leibler (KL) divergence. Our choice of KL divergence measure is largely influenced by the Markov assumptions made in language modeling and have been highlighted in [238]. The K-LF measure estimates the divergence between the hidden rep-

representations ($\mathbf{h}_{T-1}; \mathbf{h}_T$) and knowledge representation (K_e), to determine the differential knowledge to be infused.

Formally we define it as:

$$\mathbf{K-LF} = \min \mathbf{D}_{KL}(\vec{h}_T || \vec{K}_e); \text{ s.t. } \mathbf{D}_{KL}(\vec{h}_T || \vec{K}_e) < \mathbf{D}_{KL}(\vec{h}_{T-1} || \vec{K}_e) \quad (6.2)$$

where \mathbf{h}_{T-1} is an input for convergence constraint.

We minimize the *relative entropy* for information loss to maximize the information gain from the knowledge representation (e.g., K_e). We will compute differential knowledge ($\nabla \mathbf{K-LF}$) through such optimization approach; thus, the computed differential knowledge will also determine the degree of knowledge to be infused in the *Ki-layer*. $\nabla \mathbf{K-LF}$ will be computed in the form of embedding vectors, and the dimensions from K_e will be preserved.

Knowledge Modulation Function (K-MF): We need to merge the differential knowledge representation with the learned representation. However, such operation cannot be done arbitrarily., We explain an initial approach for the K-MF to modulate the learned weight matrix of the neural network with the hidden vector through an appropriate operation (e.g., Hadamard pointwise multiplication). This operation at the \mathbf{T}^{th} layer can be formulated as:

Equation for $W^{hk} = W^{hk} - \eta_k * \nabla \mathbf{K-LF}$, where W^{hk} is the learned weight matrix infusing knowledge, η_k is learning momentum [239], $\nabla \mathbf{K-LF}$ is differential knowledge. The weight matrix (W^{hk}) is computed through the learning epochs utilizing the differential knowledge embedding ($\nabla \mathbf{K-LF}$). Then we merge W^{hk} with the hidden vector \mathbf{h}_T through the K-MF. Considering that we use Hadamard pointwise multiplication as our initial approach, we formally define the output \mathbf{M}_T of K-MF as: This operation at the \mathbf{T}^{th} layer can be formulated as:

$$\vec{M}_T = \vec{h}_T \odot W^{hk} \quad (6.3)$$

where \mathbf{M}_T is Knowledge-Modulated representation, \mathbf{h}_T is the hidden vector and W^{hk} is the learned weight matrix infusing knowledge. Further investigations of techniques for K-MF, will be one of the main research topics in the agenda of this proposed research.

6.2 DIFFERENTIAL KNOWLEDGE ENGINE

In deep neural networks, each epoch generates an error that is back-propagated until the model reaches a saddle point in the local minima, and the error is reduced in each epoch. The error indicates the difference between probabilities of actual and predicted labels, and such difference can be used to enrich the Seeded SubKG in our proposed knowledge-infused deep learning framework.

In this section, we discuss the sub-knowledge graph operations that are based on the difference between the learned representation of our knowledge-infused model (\mathbf{M}_T), and the representation of the relevant sub-knowledge graph from the R-KG, which we call as differential sub-knowledge graph. We define *Knowledge Proximity function* to generate the *Differential Sub-knowledge Graph*, and *Update Seeded SubKG* to insert the differential sub-knowledge graph into the Seeded SubKG.

Knowledge Proximity: Upon the arrival of the learned representation from the knowledge-infused learning model, we query the KG for retrieving related information to the respective data point. In this particular step, it is important to find the optimal proximity between the concept and its related concepts. For example, from the “South Carolina” concept, we may traverse the surrounding concepts with a varying number of hops (empirically decided). Finding the optimal number of hops towards each direction from the concept in question is still an open research question. As we find optimal proximity of a particular concept in the KG, we propagate KG based on the proximity starting from the concept in question.

Differential SubKG: Once we obtain the SubKG from the graph propagation, we create differential SubKG that will reflect the difference in knowledge from the Seeded SubKG. For this procedure, we plan to carry out research formulating the problem using variational autoencoders to extract such SubKG as we call *differential subKG* (\mathbf{D}_{kg}) and, we believe it will provide missing information in the Seeded-KG.

Update function: The differential subKG generated as a result of minimizing knowledge proximation is considered as input factual graph to the update procedure. As a result, the procedure dynamically evolves the Seeded SubKG with missing information from differential SubKG. We plan to utilize *Lyapunov stability theorem* [240] and *Zero Shot learning* to update the Seeded-KG using D_{kg} . D_{kg} and Seeded-KG represent two knowledge structures requiring a process of transfer the knowledge from one structure to another [241]. We define it as the process of generating semantic mapping weights that encodes and decodes the two semantic spaces. We plan to utilize the Lyapunov stability constraint and Sylvester optimization approach: Given two semantic spaces belonging to a domain D , we tend to attain an equilibrium position defined as:

$$||S_{kg} - W * D_{kg}||_F = \alpha * ||W * S_{kg} - D_{kg}||_F \quad (6.4)$$

$||\cdot||_F$ represents Frobenius norm and α is a proportionality constant belong to \mathbb{R} . Equation 6.4 reflects lyapunov stability theorem and to achieve such a stable state we define our optimization function as follows:

$$\mathcal{L} \equiv \min(||S_{kg} - WD_{kg}||_F - \alpha * ||WS_{kg} - D_{kg}||_F), \alpha > 0, W \in \mathbb{R} \times \mathbb{R} \quad (6.5)$$

Equation 6.5 is solvable using Sylvester optimization and its derivation is defined in a recent study [10].

Let us investigate how Deep Infusion can happen in deep neural language models, that are gaining popularity in various application areas like computational social science, conversational artificial intelligence, multi-agent systems, and others.

6.3 DEEP INFUSION IN NEURAL LANGUAGE MODELS

The neural language models (NLMs) are designed to gather parametric knowledge after pre-training over a large-scale natural language corpus. This parametric memory is utilized in downstream applications in the following forms: (a) Fine-tuning of NLMs on a domain-specific tasks [242], (b) Augmenting the NLMs with external knowledge at the input layer and tuning it end-to-end [9], (c) Leveraging a pre-trained NLM and passing the generated representation through the knowledge-aware generative model for contextualized representation learning [150], and (d) Probing (Edge and Structured) the NLMs at each layer for checking the accuracy of parametric memory [243] [244]. These state-of-the-art methods are consistent with our definition of Shallow Infusion and Semi-Deep Infusion but can be improved towards Deep Infusion. We provide a positive direction for deep infusion as answers to the following questions:

When does a NLM require Non-parametric Knowledge? An intermediate representation between two hidden layers, denoted by h_{out}^{l-1} and h_{in}^l is often studied as the model attention in current transformer models. Essentially, these representations can be enquired for model’s learning behavior. A distributional drift (a.k.a variational inference) between *gold representation* (a.k.a knowledge representation) and h_{output}^{l-1} can be considered as a signal for non-parametric knowledge infusion. Of the various methods to measure variational inference¹, KL divergence is the most widely used.

How to infuse non-parametric knowledge seamlessly? Let us consider the most widely used multi-lingual KG, ConceptNet [45] as the source for *knowledge representation*.

¹<https://ermongroup.github.io/cs228-notes/inference/variational/>

Since it would be tedious and error-prone to measure the variational inference between every node in ConceptNet and h_{out}^{l-1} , we construct a SubKG of ConceptNet (S^{kg}) by computing exact and cosine similarity between input and concepts in S^{kg} . Now, we use h_{out}^{l-1} to traverse each node in S^{kg} by computing a distance score measured using KL divergence. We formally define it as:

$$KL(h_{out}^{l-1}, S^{kg}) = \left\{ h_{out}^{l-1} \log \frac{h_{out}^{l-1}}{S_i^{kg}} \right\}_i ; \text{where } i \in \text{Nodes in } S^{kg}$$

$KL(h_{out}^{l-1}, S^{kg})$ yields a set of nodes with its KL divergence scores. The nodes with scores above a threshold (δ , often defined empirically) are recorded as visited nodes, and their representations are used in infusion. The infusion of knowledge happens following the equation 6.3, which can be formalized as follows:

$$\tilde{h}_{out}^{l-1} = h_{out}^{l-1} \odot S_0^{kg} \odot S_1^{kg} \odot S_2^{kg} \dots \odot S_{j-1}^{kg};$$

Where $j \in \{S^{kg}\}_i$ is the set of nodes with acceptable KL scores. After the infusion of external knowledge, the model needs to be regularized, which is done by updating the backpropagation update of weights and dropout strategies. We leverage the dual form of deep neural network for updating the weights of neurons. The dual form focuses on attention, thus informing us about the importance of each neural connection between two hidden layers. The dropout is made deterministic by thresholding over the attention matrix created between \tilde{h}_{out}^{l-1} and h_{in}^l , as described by Faldu et al. [9]. To appreciate the importance of the dual form of the neural network, I would like to direct the readers to the paper by Irie et al. [245].

Due to deep knowledge infusion, the model's predicted outcome would differ from gold truth by some margins. However, this would show the model's thoughtful prediction (or classification), where the end-user would notice the likelihood of prediction of other labels or generations (if it is a language generation model) that seems similar to ground truth. This would happen because, in deep infusion, the model

would be trained end-to-end with marginalized loss², defined as:

$$P(X) = \sum_y P(X, Y = y) = \sum_y P(X|Y = y) * P(Y = y);$$

Where X is input & Y is a class label or a natural language generation from the deep neural network. This loss would enable the model to preserve the input semantics ($P(X)$) by generating its probabilities from the model's prediction or generation.

How to leverage external knowledge's inherent abstraction in enhancing it? The reason for having S^{kg} is to allow the hierarchical concepts in KG to be infused into the upper layers of the deep neural network. Maintaining a set of visited nodes, starting from the lowermost layers, supports traversing higher-order concepts in KG when representations from this and above are generated through non-linear activation. This structuring of knowledge infusion is based on the assumption that (a) non-linear activation allows the neural network to exploit all possible syntactic combinations of input tokens, which might yield a representation of concepts in KG (not present in input), and (b) these combinations represent a closed world that can be studied with input and semantically related concepts in S^{kg} [247].

Such a training methodology (a) introduces explainability intrinsically into the model's behavior. (b) The trace over the S_{kg} created during the model's training provides a clue on the model's interpretations of the input. (c) The deterministic nature of dropout, governed by knowledge-infused attention matrices, enables uncertainty handling. (d) And the context capture is always the centric component in Knowledge Infusion, which in Deep Infusion is achieved by computing variational inference between the latent, hidden representation and knowledge nodes in KG.

²It can be seen as Beam Search Optimization [246]

6.4 SUMMARY

Combining deep learning and knowledge graphs in a hybrid neural-symbolic learning framework will further enhance performance and accelerate the convergence of the learning processes. Specifically, the impact of this improvement in susceptible domains such as health and social science will be significant concerning their implications for real-world deployment. Furthermore, adopting tools that automate tasks that require knowledge and intelligence, and are traditionally done by humans, will improve with the help of this framework that marries deep learning and knowledge graph techniques. Specifically, we envision that the infusion of knowledge as described in this framework will capture information for the corresponding domain in finer granularity of abstraction. We believe that this approach will provide reliable solutions to the problems faced in deep learning. Hence, in real-world applications, resolving these issues with knowledge graphs and deep learning in a hybrid neuro-symbolic framework will significantly contribute to fulfilling AI's promise [248].

BIBLIOGRAPHY

- [1] Ugur Kursuncu, Manas Gaur, Usha Lokala, Anurag Illendula, Krishnaprasad Thirunarayan, Raminta Daniulaityte, Amit Sheth, and I Budak Arpinar. "what's ur type?" contextualized classification of user types in marijuana-related communications using compositional multiview embedding. In *IEEE/WIC/ACM International Conference on Web Intelligence(WI'18)*, 2018.
- [2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.
- [3] Manas Gaur, Saeedeh Shekarpour, Amelie Gyrard, and Amit Sheth. empathi: An ontology for emergency managing and planning about hazard crisis. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 396–403. IEEE.
- [4] Bin Chen, Xiao Dong, Dazhi Jiao, Huijun Wang, Qian Zhu, Ying Ding, and David J Wild. Chem2bio2rdf: a semantic framework for linking and data mining chemogenic and systems chemical biology data. *BMC bioinformatics*, 11(1):1–13, 2010.
- [5] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035, 2019.
- [6] Manas Gaur, Ankit Desai, Keyur Faldu, and Amit Sheth. Explainable ai using knowledge graphs. In *ACM CoDS-COMAD Conference*, 2020.
- [7] Shan Jiang, William Groves, Sam Anzaroot, and Alejandro Jaimes. Crisis sub-events on social media: A case study of wildfires. In *International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, July*, volume 1, 2019.
- [8] Maulik R Kamdar, Tymor Hamamsy, Shea Shelton, Ayin Vala, Tome Eftimov, James Zou, and Suzanne Tamang. A knowledge graph-based approach for exploring the us opioid epidemic. *arXiv preprint arXiv:1905.11513*, 2019.

- [9] Keyur Faldu, Amit Sheth, Prashant Kikani, and Hemang Akbari. Ki-bert: Infusing knowledge context for better language and domain understanding. *arXiv preprint arXiv:2104.08145*, 2021.
- [10] Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. "let me tell you about your mental health!" contextualized classification of reddit posts to dsm-5 for web-based intervention. 2018.
- [11] K Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10, 2008.
- [12] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [13] Amit Sheth, Sujana Perera, Sanjaya Wijeratne, and Krishnaprasad Thirunarayan. Knowledge will propel machine understanding of content: Extrapolating from current examples. *arXiv preprint arXiv:1707.05308*, 2017.
- [14] Amit Sheth and Krishnaprasad Thirunarayan. The inescapable duality of data and knowledge. *arXiv e-prints*, pages arXiv–2103, 2021.
- [15] Ugur Kursuncu, Manas Gaur, and Amit Sheth. Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning. 2020.
- [16] Manas Gaur, Ugur Kursuncu, Amit Sheth, Ruwan Wickramarachchi, and Shweta Yadav. Knowledge-infused deep learning. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 309–310, 2020.
- [17] Kaushik Roy, Manas Gaur, Qi Zhang, and Amit Sheth. Process knowledge-infused learning for suicidality assessment on social media. *arXiv preprint arXiv:2204.12560*, 2022.
- [18] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. *Advances in neural information processing systems*, 28, 2015.
- [19] Saeedeh Shekarpour, Edgar Marx, Sören Auer, and Amit Sheth. Rquery: rewriting natural language queries on knowledge graphs to alleviate the vocabulary mismatch problem. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

- [20] Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv preprint arXiv:2106.04554*, 2021.
- [21] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lambda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [22] Marti A Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992.
- [23] Gaur Manas, Vamsi Aribandi, Ugur Kursuncu, Amanuel Alambo, Valerie L Shalin, Krishnaprasad Thirunarayan, Jonathan Beich, Meera Narasimhan, Amit Sheth, et al. Knowledge-infused abstractive summarization of clinical diagnostic interviews: Framework development study. *JMIR Mental Health*, 8(5):e20865, 2021.
- [24] Chidubem Arachie, Manas Gaur, Sam Anzaroot, William Groves, Ke Zhang, and Alejandro Jaimes. Unsupervised detection of sub-events in large scale disasters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 354–361, 2020.
- [25] Qiwei Han, Inigo Martinez de Rituerto de Troya, Mengxin Ji, Manas Gaur, and Leid Zejnilovic. A collaborative filtering recommender system in primary care: Towards a trusting patient-doctor relationship. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 377–379. IEEE, 2018.
- [26] Qiwei Han, Mengxin Ji, Inigo Martinez de Rituerto de Troya, Manas Gaur, and Leid Zejnilovic. A hybrid recommender system for patient-doctor matchmaking in primary care. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 481–490. IEEE, 2018.
- [27] Amit Sheth, Manas Gaur, Kaushik Roy, and Keyur Faldu. Knowledge-intensive language understanding for explainable ai. *arXiv preprint arXiv:2108.01174*, 2021.
- [28] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*, 2020.
- [29] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu,

- Miruna Clinciu, Dipanjan Das, Kaustubh D Dhole, et al. The gem benchmark: Natural language generation, its evaluation and metrics. *arXiv preprint arXiv:2102.01672*, 2021.
- [30] Sujan Perera, Pablo N Mendes, Adarsh Alex, Amit P Sheth, and Krishnaprasad Thirunarayan. Implicit entity linking in tweets. In *International Semantic Web Conference*, pages 118–132. Springer, 2016.
 - [31] Manas Gaur, Kalpa Gunaratna, Vijay Srinivasan, and Hongxia Jin. Iseeq: Information seeking question generation using dynamic meta-information retrieval and knowledge graphs. *arXiv preprint arXiv:2112.07622*, 2021.
 - [32] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, 2018.
 - [33] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.
 - [34] Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33:16722–16735, 2020.
 - [35] Leslie Valiant. *Probably Approximately Correct: Nature’s Algorithms for Learning and Prospering in a Complex World*. Basic Books, Inc., USA, 2013.
 - [36] Ugur Kursuncu, Manas Gaur, and Amit Sheth. Knowledge infused learning (kil): Towards deep incorporation of knowledge in deep learning. *arXiv preprint arXiv:1912.00512*, 2019.
 - [37] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, pages 514–525, 2019.
 - [38] Xuehao Liu, Sarah Jane Delany, and Susan McKeever. Wider vision: Enriching convolutional neural networks via alignment to external knowledge bases. *arXiv preprint arXiv:2102.11132*, 2021.

- [39] Yoonna Jang, Jungwoo Lim, Yuna Hur, Dongsuk Oh, Suhyune Son, Yeonsoo Lee, Donghoon Shin, Seungryong Kim, and Heuiseok Lim. Call for customized conversation: Customized conversation grounding persona and knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10803–10812, 2022.
- [40] Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017.
- [41] Xuanli He, Islam Nassar, Jamie Ryan Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Generative models advance self-training and knowledge distillation. 2021.
- [42] Dim P Papadopoulos, Ethan Weber, and Antonio Torralba. Scaling up instance annotation via label propagation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15364–15373, 2021.
- [43] Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. Unsupervised enrichment of persona-grounded dialog with background stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 585–592, 2021.
- [44] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.
- [45] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.
- [46] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [47] David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Richard Andrew Taylor. Benchmark and best practices for biomedical knowledge graph embeddings. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 167. NIH Public Access, 2020.

- [48] Vinh Nguyen, Hong Yung Yip, Goonmeet Bajaj, Thilini Wijesiriwardene, Vishesh Javangula, Srinivasan Parthasarathy, Amit Sheth, and Olivier Bodenreider. Context-enriched learning models for aligning biomedical vocabularies at scale in the umls metathesaurus. In *Proceedings of the ACM Web Conference 2022, WWW '22*, page 1037–1046, New York, NY, USA, 2022. Association for Computing Machinery.
- [49] Hong Yung Yip, Vinh Nguyen, and Olivier Bodenreider. Construction of umls metathesaurus with knowledge-infused deep learning. 2019.
- [50] Amit Sheth, Swati Padhee, and Amelie Gyrard. Knowledge graphs and knowledge networks: the story in brief. *IEEE Internet Computing*, 23(4):67–75, 2019.
- [51] T Mikolov, I Sutskever, K Chen, GS Corrado, and J Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.
- [52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.
- [53] Shreyansh P Bhatt, Hemant Purohit, Andrew Hampton, Valerie Shalin, Amit Sheth, and John Flach. Assisting coordination during crisis: a domain ontology based approach to infer resource needs from tweets. In *Proceedings of the 2014 ACM conference on Web science*, pages 297–298. ACM, 2014.
- [54] Ugur Kursuncu, Manas Gaur, Carlos Castillo, Amanuel Alambo, Krishnaprasad Thirunarayan, Valerie Shalin, Dilshod Achilov, I. Budak Arpinar, and Amit Sheth. Modeling islamist extremist communications on social media using contextual dimensions: Religion, ideology, and hate. 3(CSCW), 2019.
- [55] Rohith K Thiruvalluru, Manas Gaur, Krishnaprasad Thirunarayan, Amit Sheth, and Jyotishman Pathak. Comparing suicide risk insights derived from clinical and social media data. In *AMIA Annual Symposium Proceedings*, volume 2021, page 364. American Medical Informatics Association, 2021.
- [56] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.
- [57] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

- [58] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020.
- [59] Samuel R. Bowman. When combating hype, proceed with caution. *CoRR*, abs/2110.08300, 2021.
- [60] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380. Association for Computational Linguistics, 2018.
- [61] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.
- [62] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.
- [63] Jinjin Zhao, Shreyansh Bhatt, Candace Thille, Dawn Zimmaro, and Neelesh Gattani. Interpretable personalized knowledge tracing and next learning activity recommendation. In *Proceedings of the Seventh ACM Conference on Learning@ Scale*, pages 325–328, 2020.
- [64] Manas Gaur, Kaushik Roy, Aditya Sharma, Biplav Srivastava, and Amit Sheth. “who can help me?”: Knowledge infused matching of support seekers and support providers during covid-19 on reddit. In *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 265–269. IEEE, 2021.
- [65] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.
- [66] Muhammad Rehman Zafar and Naimul Mefraz Khan. Dlime: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. *arXiv preprint arXiv:1906.10263*, 2019.

- [67] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
- [68] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.
- [69] Yinchong Yang, Volker Tresp, Marius Wunderle, and Peter A Fasching. Explaining therapy predictions with layer-wise relevance propagation in neural networks. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 152–162. IEEE, 2018.
- [70] Wojciech Samek, Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, and Klaus-Robert Müller. Interpreting the predictions of complex ml models by layer-wise relevance propagation. *arXiv preprint arXiv:1611.08191*, 2016.
- [71] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [72] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.
- [73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [74] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*, 2020.
- [75] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.
- [76] Maria Riveiro and Serge Thill. “that’s (not) the output i expected!” on the role of end user expectations in creating explanations of ai systems. *Artificial Intelligence*, 298:103507, 2021.

- [77] Kaushik Roy, Qi Zhang, Manas Gaur, and Amit Sheth. Knowledge infused policy gradients with upper confidence bound for relational bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer, 2021.
- [78] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, 2004.
- [79] Preksha Nema and Mitesh M Khapra. Towards a better metric for evaluating question generation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3950–3959, 2018.
- [80] Thibault Sellam, Dipanjan Das, and Ankur Parikh. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, 2020.
- [81] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34, 2021.
- [82] Milan Gritta, Ruoyu Hu, and Ignacio Iacobacci. Crossaligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language understanding. *arXiv preprint arXiv:2203.09982*, 2022.
- [83] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908, 2020.
- [84] Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya, and Amit Sheth. Multi-task learning framework for mining crowd intelligence towards clinical treatment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 271–277, 2018.
- [85] Ugur Kursuncu, Manas Gaur, Carlos Castillo, Amanuel Alambo, Krishnaprasad Thirunarayan, Valerie Shalin, Dilshod Achilov, I Budak Arpinar, and Amit Sheth. Modeling islamist extremist communications on social media using contextual di-

mensions: religion, ideology, and hate. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–22, 2019.

- [86] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*, 2018.
- [87] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
- [88] Usha Lokala, Raminta Daniulaityte, Francois Lamy, Manas Gaur, Krishnaprasad Thirunarayan, Ugur Kursuncu, and Amit P Sheth. Dao: An ontology for substance use epidemiology on social media and dark web. *JMIR Public Health and Surveillance*, 2020.
- [89] Bill Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*, 2005.
- [90] Joseph Reagle and Manas Gaur. Spinning words as disguise: Shady services for ethical research? *First Monday*, 2022.
- [91] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015.
- [92] Daniel Cera, Mona Diabb, Eneko Agirrec, Inigo Lopez-Gazpioc, Lucia Speciad, and Basque Country Donostia. Semeval-2017 task 1: Semantic textual similarity multilingual and cross-lingual focused evaluation.
- [93] S Wijeratne, L Balasuriya, A Sheth, and D Doran. Emojinet: An open service and api for emoji sense discovery. *arXiv preprint arXiv:1707.04652*, 2017.
- [94] Shankar Iyer, Nikhil Dandekar, Kornél Csernai, et al. First quora dataset release: Question pairs. *data. quora. com*, 2017.
- [95] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

- [96] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, 2018.
- [97] Pedro Rodriguez, Paul Crook, Seungwhan Moon, and Zhiguang Wang. Information seeking in the spirit of learning: a dataset for conversational curiosity. In *Empirical Methods in Natural Language Processing*, 2020.
- [98] Valentina Pyatkin, Ayal Klein, Reut Tsarfaty, and Ido Dagan. Qadiscourse-discourse relations as qa pairs: Representation, crowdsourcing and baselines. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2804–2819, 2020.
- [99] Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1985–1988, 2020.
- [100] Ido Dagan, Bill Dolan, Bernardo Magnini, and Dan Roth. Recognizing textual entailment: Rational, evaluation and approaches–erratum. *Natural Language Engineering*, 16(1):105–105, 2010.
- [101] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.
- [102] James W Pennebaker. Linguistic inquiry and word count: Liwc 2001.
- [103] Margaret M Bradley and Peter J Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Technical report C-1, the center for research in psychophysiology . . . , 1999.
- [104] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.
- [105] Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198, 2017.

- [106] Manas Gaur, Vamsi Aribandi, Amanuel Alambo, Ugur Kursuncu, Krishnaprasad Thirunarayan, Jonathan Beich, Jyotishman Pathak, and Amit Sheth. Characterization of time-variant and time-invariant assessment of suicidality on reddit using c-ssrs. *PloS one*, 16(5):e0250448, 2021.
- [107] Thilini Wijesiriwardene, Hale Inan, Ugur Kursuncu, Manas Gaur, Valerie L Shalin, Krishnaprasad Thirunarayan, Amit Sheth, and I Budak Arpinar. Alone: A dataset for toxic behavior among adolescents on twitter. In *International Conference on Social Informatics*, pages 427–439. Springer, 2020.
- [108] Akiko Aizawa. An information-theoretic perspective of tf-idf measures. *Information Processing & Management*, 39(1):45–65, 2003.
- [109] Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.
- [110] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard H Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. In *HLT-NAACL*, 2015.
- [111] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
- [112] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what’s next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, 2017.
- [113] Sawan Kumar and Partha Talukdar. Nile: Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, 2020.
- [114] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, 2019.
- [115] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

- [116] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.
- [117] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*, 2020.
- [118] Mostafa Dehghani, Yi Tay, Alexey A Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.
- [119] Nikola Mrksic, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J Young. Counter-fitting word vectors to linguistic constraints. In *HLT-NAACL*, 2016.
- [120] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL (1)*, 2018.
- [121] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.
- [122] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [123] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.
- [124] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019.
- [125] Sreyasi Nag Chowdhury, Ruwan Wickramarachchi, Mohamed H Gad-Elrab, Daria Stepanova, and Cory Henson. Towards leveraging commonsense knowledge for autonomous driving. 2021.

- [126] Ruwan Wickramarachchi, Cory Henson, and Amit Sheth. Knowledge-infused learning for entity prediction in driving scenes. *Frontiers in big Data*, 4, 2021.
- [127] Ramnath Kumar, Shweta Yadav, Raminta Daniulaityte, Francois Lamy, Krishnaprasad Thirunarayan, Usha Lokala, and Amit Sheth. edarkfind: Unsupervised multi-view learning for sybil account detection. In *Proceedings of The Web Conference 2020*, pages 1955–1965, 2020.
- [128] Shreyansh Bhatt, Manas Gaur, Beth Bullemer, Valerie Shalin, Amit Sheth, and Brandon Minnery. Enhancing crowd wisdom using explainable diversity inferred from social media. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 293–300. IEEE, 2018.
- [129] Matthew R Jamnik and David J Lane. The use of reddit as an inexpensive source for high-quality data. *Practical Assessment, Research, and Evaluation*, 22(1):5, 2017.
- [130] Manas Gaur, Keyur Faldu, and Amit Sheth. Semantics of the black-box: Can knowledge graphs help make deep learning systems more interpretable and explainable? *IEEE Internet Computing*, 25(1):51–59, 2021.
- [131] Ziyin Liu, Zhikang Wang, Paul Pu Liang, Russ R Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. *Advances in Neural Information Processing Systems*, 32, 2019.
- [132] Ramit Sawhney, Atula Neerkaje, and Manas Gaur. A risk-averse mechanism for suicidality assessment on social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 628–635, 2022.
- [133] Kazi Saidul Hasan and Vincent Ng. Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1262–1273, 2014.
- [134] Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*, 2018.
- [135] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380, 2022.

- [136] Lan Huong Nguyen and Susan Holmes. Ten quick tips for effective dimensionality reduction. *PLoS computational biology*, 15(6):e1006907, 2019.
- [137] Guillermo Soberón, Lora Aroyo, Chris Welty, Oana Inel, Hui Lin, and Manfred Overmeien. Measuring crowd truth: Disagreement metrics combined with worker behavior filters. In *CrowdSem 2013 Workshop*, volume 2, 2013.
- [138] Hemant Purohit, Valerie L Shalin, and Amit P Sheth. Knowledge graphs to empower humanity-inspired ai systems. *IEEE Internet Computing*, 24(4):48–54, 2020.
- [139] Yasas Senarath, Jennifer Chan, Hemant Purohit, and Ozlem Uzuner. Evaluating the relevance of umls knowledge base for public health informatics during disasters. In *ISCRAM 2021 Conference Proceedings–18th International Conference on Information Systems for Crisis Response and Management*, 2021.
- [140] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pages 9355–9366. PMLR, 2021.
- [141] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.
- [142] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.
- [143] KM Annervaz, Somnath Basu Roy Chowdhury, and Ambedkar Dukkipati. Learning beyond datasets: Knowledge graph augmented neural networks for natural language processing. In *NAACL-HLT*, 2018.
- [144] Liwei Cai and William Yang Wang. Kbgan: Adversarial learning for knowledge graph embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1470–1480, 2018.
- [145] Che-Han Chang, Chun-Hsien Yu, Szu-Ying Chen, and Edward Y Chang. Kg-gan: Knowledge-guided generative adversarial networks. *arXiv preprint arXiv:1905.12261*, 2019.
- [146] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American*

- [147] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.
- [148] Bishan Yang and Tom Mitchell. Leveraging knowledge bases in lstms for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1436–1446, 2017.
- [149] Joey Bose, Ricardo Pio Monti, and Aditya Grover. Cage: Probing causal relationships in deep generative models. 2021.
- [150] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*, 2019.
- [151] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975, 2020.
- [152] Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.
- [153] Yichong Xu, Chenguang Zhu, Shuohang Wang, Siqu Sun, Hao Cheng, Xiaodong Liu, Jianfeng Gao, Pengcheng He, Michael Zeng, and Xuedong Huang. Human parity on commonsenseqa: Augmenting self-attention with external attention.
- [154] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuan-Jing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. K-adapter: Infusing knowledge into pre-trained models with adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, 2021.
- [155] Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. Sensebert: Driving some sense into bert. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, 2020.

- [156] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.
- [157] Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22:104–1, 2021.
- [158] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.
- [159] Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Timothy Nguyen, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Architecture matters in continual learning. *arXiv preprint arXiv:2202.00275*, 2022.
- [160] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.
- [161] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 843–852. IEEE, 2017.
- [162] Khuong Vo, Dang Pham, Mao Nguyen, Trung Mai, and Tho Quan. Combination of domain knowledge and deep learning for sentiment analysis. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pages 162–173. Springer, 2017.
- [163] Jiang Bian, Bin Gao, and Tie-Yan Liu. Knowledge-powered deep learning for word embedding. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 132–148. Springer, 2014.
- [164] Zhiting Hu, Zichao Yang, Russ R Salakhutdinov, LIANHUI Qin, Xiaodan Liang, Haoye Dong, and Eric P Xing. Deep generative models with learnable knowledge constraints. *Advances in Neural Information Processing Systems*, 31, 2018.
- [165] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

- [166] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2673–2681, 2017.
- [167] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5329–5336, 2019.
- [168] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
- [169] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014.
- [170] Nicolas Y Masse, Gregory D Grant, and David J Freedman. Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467–E10475, 2018.
- [171] Amit Sheth and Pavan Kapanipathi. Semantic filtering for social data. *IEEE Internet Computing*, 20(4):74–78, 2016.
- [172] Hugo Liu and Push Singh. Commonsense reasoning in and over natural language. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 293–306. Springer, 2004.
- [173] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [174] Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8758–8765, 2020.
- [175] Amit Sheth, Manas Gaur, Ugur Kursuncu, and Ruwan Wickramarachchi. Shades of knowledge-infused learning for enhancing deep learning. *IEEE Internet Computing*, 23(6):54–63, 2019.

- [176] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, 2015.
- [177] Charles Dugas, Yoshua Bengio, François Bélisle, Claude Nadeau, and René Garcia. Incorporating functional knowledge in neural networks. *Journal of Machine Learning Research*, 10(Jun):1239–1262, 2009.
- [178] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [179] Mercedes Arguello Casteleiro, George Demetriou, Warren Read, Maria Jesus Fernandez Prieto, Nava Maroto, Diego Maseda Fernandez, Goran Nenadic, Julie Klein, John Keane, and Robert Stevens. Deep learning meets ontologies: experiments to anchor the cardiovascular disease ontology in the biomedical literature. *Journal of biomedical semantics*, 9(1):1–24, 2018.
- [180] Ying Shen, Yang Deng, Min Yang, Yaliang Li, Nan Du, Wei Fan, and Kai Lei. Knowledge-aware attentive neural network for ranking question answer pairs. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 901–904. ACM, 2018.
- [181] Md Kamruzzaman Sarker, Ning Xie, Derek Doran, Michael Raymer, and Pascal Hitzler. Explaining trained neural networks with semantic web technologies: First steps. *arXiv preprint arXiv:1710.04324*, 2017.
- [182] Bassem Makni and James Hendler. Deep learning for noise-tolerant rdfs reasoning. *Semantic Web*, 10(5):823–862, 2019.
- [183] Amrudin Agovic and Arindam Banerjee. Gaussian process topic models. *arXiv preprint arXiv:1203.3462*, 2012.
- [184] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.
- [185] Adil M Bagirov and Julien Ugon. Supervised data classification via max-min separability. In *Continuous Optimization*. 2005.

- [186] Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655*, 2018.
- [187] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. Generating clarifying questions for information retrieval. In *Proceedings of The Web Conference 2020*, pages 418–428, 2020.
- [188] Suppanut Pothirattanachaikul, Takehiro Yamamoto, Yusuke Yamamoto, and Masatoshi Yoshikawa. Analyzing the effects of "people also ask" on search behaviors and beliefs. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 101–110, 2020.
- [189] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.
- [190] Dustin S Stoltz and Marshall A Taylor. Concept mover’s distance: measuring concept engagement via word embeddings in texts. *Journal of Computational Social Science*, 2(2):293–313, 2019.
- [191] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 167–175, 2021.
- [192] Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. Crowdsourcing question-answer meaning representations. *CoRR*, abs/1711.05885, 2017.
- [193] Pedro Rodriguez, Paul A Crook, Seungwhan Moon, and Zhiguang Wang. Information seeking in the spirit of learning: A dataset for conversational curiosity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8153–8172, 2020.
- [194] Kevin Clark and Christopher D Manning. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, 2016.
- [195] Nikita Kitaev and Dan Klein. Constituency parsing with a self-attentive encoder. In *ACL (1)*, 2018.

- [196] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.
- [197] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.
- [198] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.
- [199] Ishan Tarunesh, Somak Aditya, and Monojit Choudhury. Trusting roberta over bert: Insights from checklisting the natural language inference task. *arXiv preprint arXiv:2107.07229*, 2021.
- [200] Yifan Gao, Chien-Sheng Wu, Jingjing Li, Shafiq Joty, Steven CH Hoi, Caiming Xiong, Irwin King, and Michael Lyu. Discern: Discourse-aware entailment reasoning network for conversational machine reading. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2439–2449, 2020.
- [201] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- [202] Daniel Cohen, Liu Yang, and W Bruce Croft. Wikipassageqa: A benchmark collection for research on non-factoid answer passage retrieval. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1165–1168, 2018.
- [203] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
- [204] Ahmed Elgohary Ghoneim and Denis Peskov. Canard: A dataset for question-in-context rewriting. 2019.
- [205] Weizhen Qi, Yeyun Gong, Yu Yan, Jian Jiao, Bo Shao, Ruofei Zhang, Houqiang Li, Nan Duan, and Ming Zhou. Prophetnet-ads: A looking ahead strategy for generative retrieval models in sponsored search engine. In *CCF International Conference on*

Natural Language Processing and Chinese Computing, pages 305–317. Springer, 2020.

- [206] Kalpa Gunaratna, Amir Hossein Yazdavar, Krishnaprasad Thirunarayan, Amit Sheth, and Gong Cheng. Relatedness-based multi-entity summarization. In *IJCAI: proceedings of the conference*, volume 2017, page 1060. NIH Public Access, 2017.
- [207] Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Pon-nurangam Kumaraguru, and Amit Sheth. Learning to automate follow-up question generation using process knowledge for depression triage on reddit posts. 05 2022.
- [208] Chao-Yi Lu and Sin-En Lu. A survey of approaches to automatic question generation: from 2019 to early 2021. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 151–162, 2021.
- [209] Roger D Newman-Norlund, Sarah E Newman-Norlund, Sara Sayers, Samaneh Nemat, Nicholas Riccardi, Chris Rorden, and Julius Fridriksson. The aging brain cohort (abc) repository: The university of south carolina’s multimodal lifespan database for studying the relationship between the brain, cognition, genetics and behavior in healthy aging. *Neuroimage: Reports*, 1(1):100008, 2021.
- [210] Gary Libben. From lexicon to flexicon: The principles of morphological transcendence and lexical superstates in the characterization of words in the mind. *Frontiers in Artificial Intelligence*, 4, 2021.
- [211] Katherine Stasaski and Marti A Hearst. Multiple choice question generation utilizing an ontology. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 303–312, 2017.
- [212] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo. Re2g: Retrieve, rerank, generate.
- [213] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine*, 176(5):619–625, 2016.
- [214] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*, 2022.

- [215] Nut Limsopatham and Nigel Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1014–1023, 2016.
- [216] Wenbo Wang, Lu Chen, Ming Tan, Shaojun Wang, and Amit P Sheth. Discovering fine-grained sentiment in suicide notes. *Biomedical informatics insights*, 5:BII–S8963, 2012.
- [217] Danielle L Mowery, Craig Bryan, and Mike Conway. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 89–98, 2015.
- [218] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.
- [219] Jesse Vig. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*, 2019.
- [220] Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 94–102, 2017.
- [221] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.
- [222] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*, 2020.
- [223] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.
- [224] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *arXiv preprint arXiv:2203.15144*, 2022.

- [225] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preotiu-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.
- [226] Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625, 2020.
- [227] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for squad. *CoRR*, abs/1806.03822, 2018.
- [228] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- [229] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.
- [230] Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 147–156, 2020.
- [231] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, 2022.
- [232] Jian Yang, Gang Xiao, Yulong Shen, Wei Jiang, Xinyu Hu, Ying Zhang, and Jinghui Peng. A survey of knowledge enhanced pre-trained models. *arXiv preprint arXiv:2110.00269*, 2021.
- [233] Franz Baader, Baris Sertkaya, and Anni-Yasmin Turhan. Computing the least common subsumer wrt a background terminology. *Journal of Applied Logic*, 5(3):392–420, 2007.

- [234] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.
- [235] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, volume 14, pages 1112–1119, 2014.
- [236] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al. Holographic embeddings of knowledge graphs. In *AAAI*, volume 2, pages 3–2, 2016.
- [237] Prashanth Gurunath Shivakumar, Haoqi Li, Kevin Knight, and Panayiotis Georgiou. Learning from past mistakes: Improving automatic speech recognition output via noisy-clean phrase context modeling. *arXiv preprint arXiv:1802.02607*, 2018.
- [238] Chris Longworth. *Kernel methods for text-independent speaker verification*. PhD thesis, University of Cambridge, 2010.
- [239] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- [240] Mingxia Liu, Daoqiang Zhang, and Songcan Chen. Attribute relation learning for zero-shot classification. *Neurocomputing*, 139:34–46, 2014.
- [241] Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. Knowledge transfer for out-of-knowledge-base entities: a graph neural network approach. *arXiv preprint arXiv:1706.05674*, 2017.
- [242] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. Transfertransfo: A transfer learning approach for neural network based conversational agents. 2019.
- [243] Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*, 2018.
- [244] David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. Probing for constituency structure in neural language models. *arXiv preprint arXiv:2204.06201*, 2022.

- [245] Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. *arXiv preprint arXiv:2202.05798*, 2022.
- [246] Sam Wiseman and Alexander M Rush. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1296–1306, 2016.
- [247] David J Acunzo, Daniel M Low, and Scott L Fairhall. Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. *NeuroImage*, 251:119005, 2022.
- [248] Keyur Faldu, Amit Sheth, Prashant Kikani, Manas Gaur, and Aditi Avasthi. Towards tractable mathematical reasoning: Challenges, strategies, and opportunities for solving math word problems. *arXiv preprint arXiv:2111.05364*, 2021.

Review

Microplastics in the Ecosystem: An Overview on Detection, Removal, Toxicity Assessment, and Control Release

Bhamini Pandey ¹, Jigyasa Pathak ¹, Poonam Singh ^{1,*}, Ravinder Kumar ², Amit Kumar ^{3,*}, Sandeep Kaushik ⁴ and Tarun Kumar Thakur ⁴

¹ Department of Applied Chemistry, Delhi Technological University, Delhi 110042, India

² Department of Chemistry, Gurukula Kangri (Deemed to be University), Haridwar 249404, Uttarakhand, India

³ School of Hydrology and Water Resources, Nanjing University of Information Science and Technology, Nanjing 210044, China

⁴ Department of Environmental Science, Indira Gandhi National Tribal University Amarkantak, Madhya Pradesh 484887, India

* Correspondence: poonam@dtu.ac.in (P.S.); amitkdah@nuist.edu.cn (A.K.)

Abstract: In recent decades, the accumulation and fragmentation of plastics on the surface of the planet have caused several long-term climatic and health risks. Plastic materials, specifically microplastics (MPs; sizes < 5 mm), have gained significant interest in the global scientific fraternity due to their bioaccumulation, non-biodegradability, and ecotoxicological effects on living organisms. This study explains how microplastics are generated, transported, and disposed of in the environment based on their sources and physicochemical properties. Additionally, the study also examines the impact of COVID-19 on global plastic waste production. The physical and chemical techniques such as SEM-EDX, PLM, FTIR, Raman, TG-DSC, and GC-MS that are employed for the quantification and identification of MPs are discussed. This paper provides insight into conventional and advanced methods applied for microplastic removal from aquatic systems. The finding of this review helps to gain a deeper understanding of research on the toxicity of microplastics on humans, aquatic organisms, and soil ecosystems. Further, the efforts and measures that have been enforced globally to combat MP waste have been highlighted and need to be explored to reduce its potential risk in the future.

Keywords: microplastics; environmental pollution; covid-19; detection techniques; toxicity assessment

Citation: Pandey, B.; Pathak, J.; Singh, P.; Kumar, R.; Kumar, A.; Kaushik, S.; Thakur, T.K. Microplastics in the Ecosystem: An Overview on Detection, Removal, Toxicity Assessment, and Control Release. *Water* **2023**, *15*, 51. <https://doi.org/10.3390/w15010051>

Academic Editor: Grzegorz Nałęcz-Jawecki

Received: 30 November 2022

Revised: 19 December 2022

Accepted: 20 December 2022

Published: 23 December 2022



Copyright: © 2022 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/>).

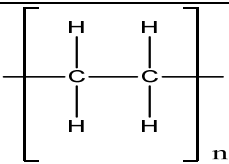
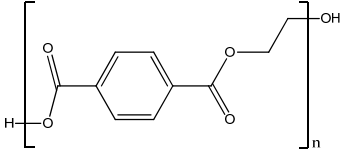
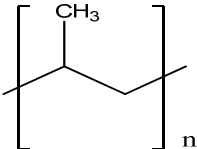
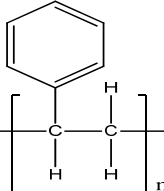
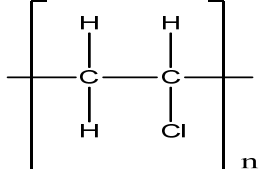
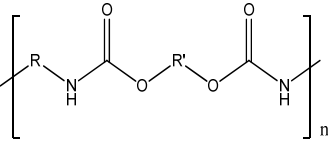
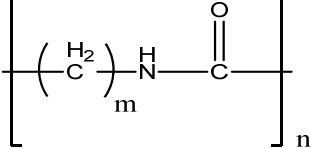
1. Introduction

Today's world relies heavily on plastic on a global scale, infiltrating almost every aspect of human lives. Plastics are organic polymers that exhibit exceptional properties such as durability, flexibility, lightness, and mechanical and thermal stability which contribute to their widespread applications in construction, food and packaging industries, pharmaceuticals, and many more sectors [1]. Despite annual expansion in the plastic industry, the demand for plastic does not seem to be decreasing. The amount of plastic generation is estimated to reach approximately 33 billion tons by the year 2050 [2]. The environmental impact of plastic has been a considerable concern for government entities, the scientific community, and the general public, regardless of its long-term industrial benefit [3]. The production and distribution of plastics possessing high degradation resistance are increasing at a rapid pace, which has serious environmental and ecological consequences. Geyer et al. [4] reported the contamination of the marine environment by 4–12 million metric tons of land-generated plastic waste by 2010.

Environmental pollution caused by plastic debris has become increasingly apparent

in the past few decades. Although the size of plastic debris can range from microscopic particles to pieces measuring several meters in length, the focus of public concern is currently on synthetic microplastics having a diameter of less than 5 mm [5]. The term microplastics (MPs) was first given by Thompson et al. [6]. Table 1 presents a number of frequently used commodity plastics along with their structure, applications, and associated hazards. For instance, polypropylene is a commonly used commodity plastic that may naturally degrade in approximately 30 years, potentially causing an unknown harmful impact on the biosphere [7]. In addition, these MPs also serve as carriers of various hazardous pollutants in biomedical and cosmetic products as well as some organic contaminants such as polychlorinated biphenyls (PCBs), dichlorodiphenyltrichloroethane (DDT), polycyclic aromatic hydrocarbons (PAHs), etc. [8]. The increasing contribution of microplastics to the environment has resulted in MP pollution becoming a global issue.

Table 1. Applications and hazards associated with commodity microplastics.

Polymer	Structure	Applications	Toxic Effects	References
Polyethylene (PE)		Packaging	Detrimental to environment	[9]
Polyethylene terephthalate (PET)		Packaging	Disruption of endocrine system	[10]
Polypropylene (PP)		Automotives and furniture	Carcinogenic and cytotoxic	[11]
Polystyrene (PS)		Food packaging	Inhibition of growth and mortality	[12]
Polyvinyl chloride (PVC)		Constructions and buildings	Damage to immune system and causes infertility	[13]
Polyurethane (PU)		Constructions and buildings	Cause neurological impairment	[14]
Polyamide (PA)		Textiles and automobiles	Liver damage	[15]

1.1. Properties of MPs

A vast array of products made from plastic are used in day-to-day life, including packaging, containers, coatings, bags, etc. Since microplastics are chemically stable, they can last for thousands of years or longer in the environment [16]. Around 90% of the total plastics produced in the world are polymeric materials, including PET, PS, PVC, and PE [17]. The physicochemical properties of these polymers determine how these micro-sized particles interact under different environmental conditions. The interaction between MPs and biota depends upon the size of the plastics. Three different sizes of micro-sized PS beads were examined to determine their impact on the survival and development of marine copepods *T. japonica* [18]. Their findings revealed that the MPs of PS may have detrimental effects on marine copepods, including decreased survival and retarded development [18]. Another important property determining the interaction between MPs and biological systems is particle shape. Au et al. [19] estimated the impacts of the shape of polypropylene MPs (beads and fibers) on the development, reproduction, and egestion of amphipod *Hyaella azteca*. Compared to beads, MP fibers exhibited more toxicity owing to their prolonged residence time in the gut, causing food to be egested more slowly and the growth of amphipods to be significantly slower [19]. The irregularities in the shape of MPs result in their more rapid attachment to the external and internal surfaces of the terrestrial or aquatic biota.

Several chemical characteristics determine the chemical nature of microplastics, including functional groups, surface polarities, stability, and crystallinity. The chemical properties of MPs are associated with their affinity for chemicals, as opposed to their physical properties that directly affect ingestion, egestion, or cause physical injury to marine and terrestrial biota. Microplastics tend to accumulate other pollutants from the ecosystem primarily due to their polarities and functional groups. The adsorption of 18 perfluoroalkyl (PFAS) compounds on three different MPs (PS, PE), and carboxylate polystyrene (PS-COOH) was studied by Llorca et al. [20], where it was observed that PS and PS-COOH exhibited higher affinity for PFASs than PE. In addition, it was also concluded that the interactions between PFASs and microplastics increased the toxicity of hydrophobic contaminants in terrestrial and aquatic ecosystems. The crystallinity of a polymer determines the physicochemical properties of MPs, i.e., permeability, density, etc., which successively govern their hydration and swelling properties. Chen et al. [21] illustrated that the degree of crystallinity of MPs alters with degradation time. They studied the biodegradation of polycaprolactone (PCL) MPs and observed that MPs became more crystalline upon degradation, demonstrating a preference for degradation in the amorphous region. This may lead to the formation of crystallites exhibiting different toxicity in comparison to the parent microplastic materials [21].

1.2. Primary and Secondary MPs

There are a variety of routes through which microplastics can enter the environment, and they can be classified as primary or secondary MPs, based on their origin and usage. Primary MPs are materials purposely fabricated for a particular application; however, secondary MPs are formed as a byproduct of the fragmentation and breakdown of larger microplastics via hydrolysis, UV rays, mechanical friction, etc. [22]. The common sources of both primary and secondary MPs are given in Figure 1.

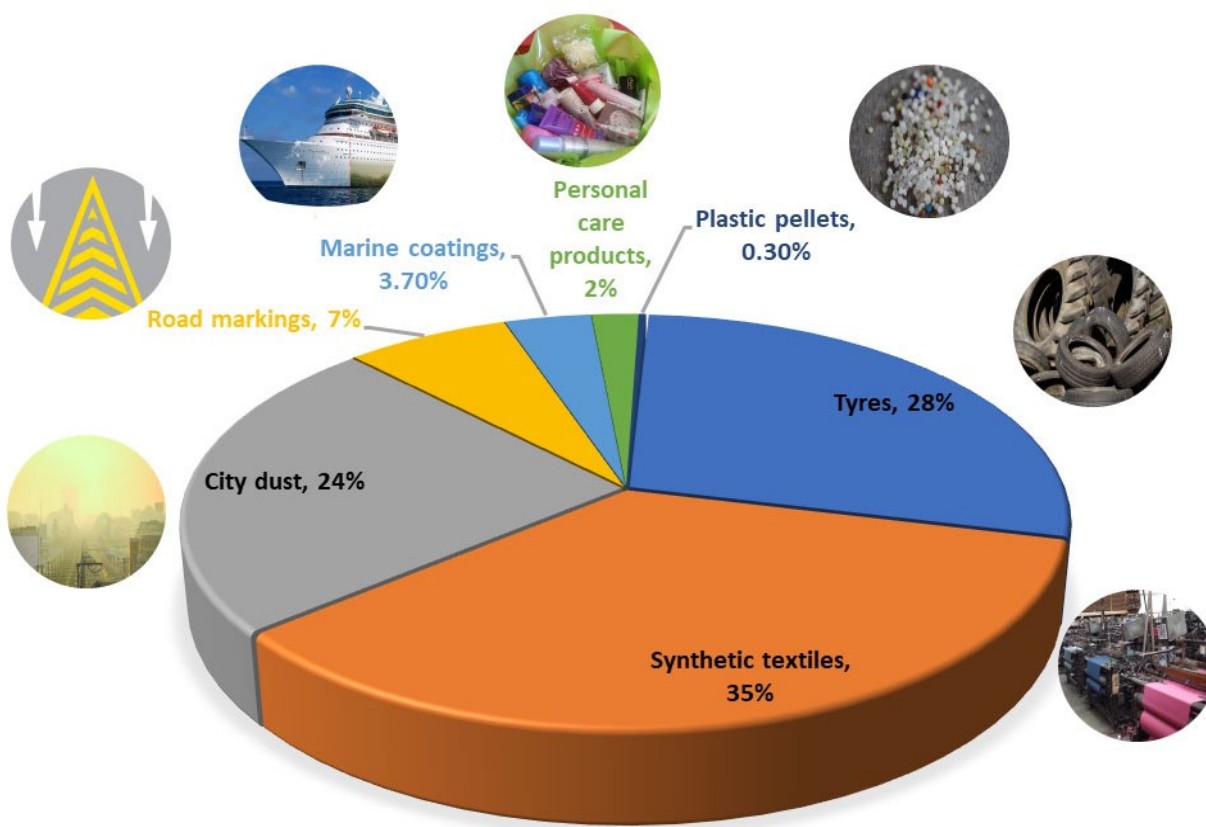


Figure 1. Sources of primary and secondary MPs [23].

Generally, primary microplastic materials come from various sources, including plastic pellets, vectors for drugs, and cosmetics products. It is still unclear where and how primary microplastics are produced, especially concerning the amounts of each type of microplastic that are released. Plastic pellets and flakes, used in making plastic products, are among the primary sources of microplastics [24]. These MPs can be released into the environment through an accidental loss during transportation or contamination during processing if they are not handled properly [25]. Certain segments of personal care products, including hand cleaners, sanitizers, facial cleansers, sunscreen, and toothpaste, use microplastic particles as exfoliants. The market has gradually been replaced by products containing microbeads instead of natural materials such as pumice, apricots, walnut peel, etc. [26,27]. According to a survey conducted by Cosmetics Europe (2012) in Switzerland, European Union, and Norway, polyethylene accounted for 93% of MPs employed in skin care products [28]. The usage of MPs in medical applications, such as tooth polish for dentistry, and pharmaceutical carriers, is also widespread. These MPs from cosmetic and medical products are released into the natural environment after usage, leading to aggravation of MPs pollution.

Fragmentation of plastic materials triggers the release of secondary MPs in the ecosystem, which occurs when plastic is degraded into smaller pieces as a result of various processes. Plastic waste enters the nearby water bodies because of littering and improper waste handling. Various natural weathering processes such as UV irradiation, pH changes, biological activities, exposure to particular chemicals, etc., result in the fragmentation or degradation of plastic waste into secondary MPs (Figure 2) [16,29,30].

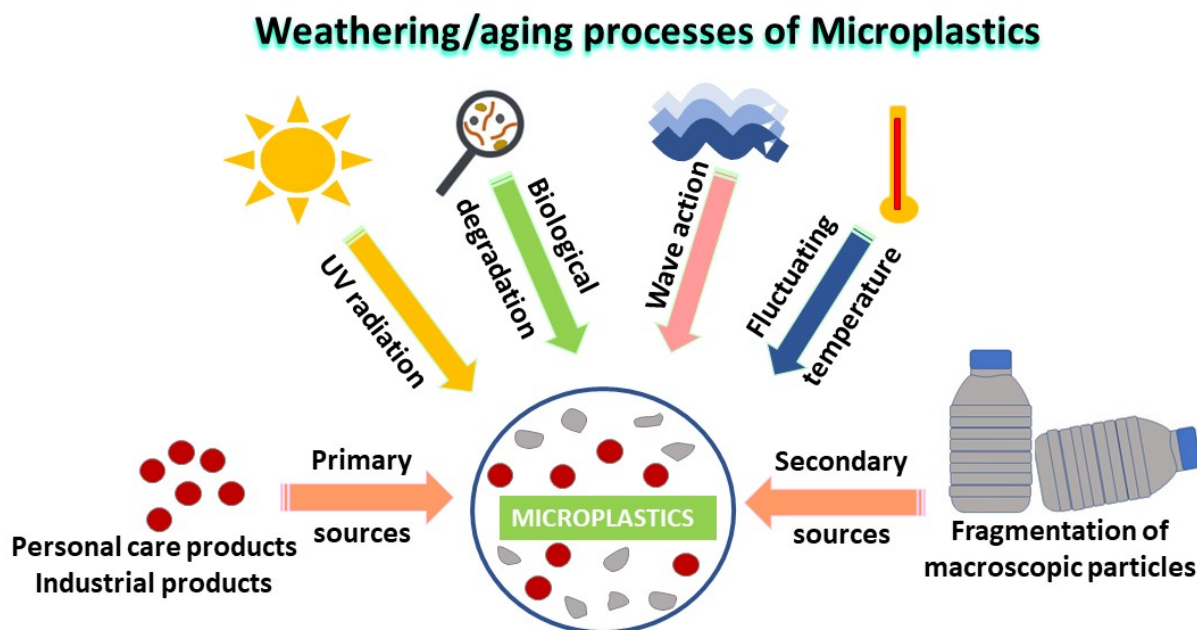


Figure 2. Weathering or ageing processes of microplastics.

Fishing gears, fish cages, and fishnets are also sources of secondary MPs, although these items are not intended to release microplastic particles into water resources, they do so when they deteriorate over time. Among fishing waste, nylon nets and fibrous ropes are the most commonly lost wastes during fishing activities [24]. It has been identified that washing synthetic clothes is also considered a significant source of MP contamination in the environment. According to a study by Napper and Thompson [31], in garment industry, each cloth product releases 1900 MP particles into the wastewater during the washing process [25]. Automobile tire abrasion has also been considered another source of MPs in the environment. As vehicles are driven on roads, the elastomers on their tires wear out and abrasion of tires occurs, which results in fine dust pollution as well as MP pollution. Approximately 0.81 Kg of abrasion from tires gets into the environment annually [26].

1.3. Impacts of COVID-19 on Release of MPs in the Environment

Pandemic-threatening contagious diseases have emerged and spread across history regularly. Humankind has already suffered from various pandemics and epidemics, including plague, cholera, famine, and Middle East respiratory syndrome coronavirus (MERS-CoV) [32,33]. A global pandemic has been sweeping the world since December 2019 caused by the novel coronavirus (SARS-CoV-2) suspected of causing a severe respiratory illness, termed COVID-19. The WHO (World Health Organization) declared COVID-19 a global pandemic in March 2020, and since then, preventive measures have been adopted to control its spread. The excessive usage of single-use plastic (SUP) materials such as face masks, disposable utensils, personal protective equipment (PPE), food packaging plastics, etc., during COVID-19 led to MP discharge into the environment [34]. Therefore, a sudden increase in plastic pollution can be observed through a significant amount of generated biowaste and medical waste during the COVID-19 pandemic. In the aftermath of the COVID-19 outbreak, the global demand for personal protective equipment increased significantly, with 65 and 129 billion pairs of gloves and masks consumed each month, respectively [35]. According to Benson et al., global plastic waste has increased by 1.6 million tonnes since the beginning of the pandemic. Every day approximately 3.4 billion single-use face masks and shields are discarded [36]. In addition, according to Peng et al. (2021), as of 23 August 2021, the plastic waste generated during the

pandemic by 193 countries reached over eight million tons and washed into the ocean globally more than 25 thousand tons which is approximately 1.5% of total plastic discharged into the aquatic environment [37]. The common sources, environmental processing of generated waste, and the fate of MPs generated during COVID-19 are given in Figure 3. Morgana et al. [38] experimented to confirm the release of MPs into water from face masks made up of polypropylene (PP). They carried out the fragmentation and deterioration of three-layered surgical masks in water via a rotating blender in order to mimic the circular waves and motions of water in oceans. The experimental outcomes showed that an enormous quantity of microplastics can be discharged from a singular mask under weathering conditions. Additionally, they also reported that as the exposure time and shear intensity increased, the release of MPs from disposable masks increased too [38].

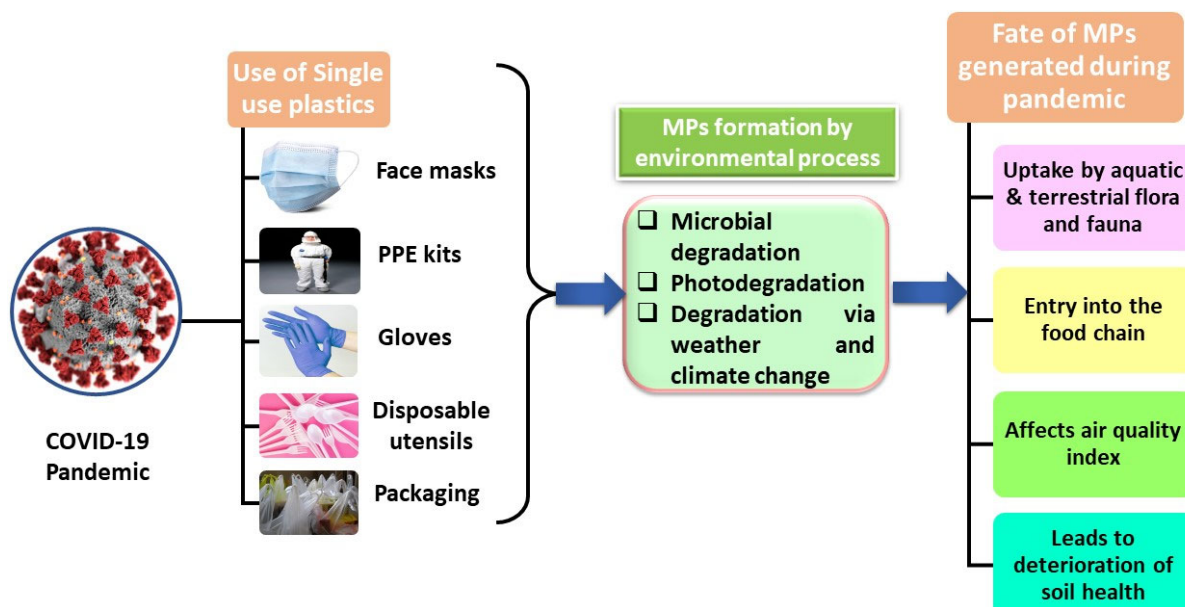


Figure 3. Sources, environmental processing, and fate of MPs generated during COVID-19.

Consequently, overloading the existing facilities might lead to paralysis of the waste disposal and recycling industry as a result of this sudden increase in waste. The mismanagement in the disposal of plastic waste might cause MPs to accumulate in terrestrial and marine ecosystems [39]. Hence, pandemics such as COVID-19 pose a serious threat to humankind, and in order to combat their outbreaks the use of PPE kits, face masks, and other polymer products cannot be avoided, leading to the discharge of excessive plastic waste into the ecosystem. Aquatic organisms and terrestrial plants easily accumulate the released MPs from the water and soil, allowing them to be readily consumed by humans and ultimately enter the food chain [40]. Therefore, in order to reduce MP pollution caused by improper disposal of face masks and PPE kits, environmental awareness about proper waste disposal should be implemented as a part of long-term preventive measures.

This study explains how microplastics are generated, transported, disposed and quantify in the environment based on their sources and physicochemical properties. Moreover, the quantification techniques and methodologies for microplastic removal from aquatic systems have been briefly discussed. Additionally, the study also examines the impact of COVID-19 on global plastic waste production. This review aims to gain a better understanding of research on the toxic effects of microplastics on humans, aquatic life forms, and soil ecosystems.

2. Detection and Identification of MPs

Due to the exorbitant usage of commodity plastics worldwide, MPs with a wide array of attributes are produced and the detection and analysis of MPs is a prerequisite condition for their effective removal from aquatic systems. A wide variety of physical and chemical techniques, presented in Figure 4, are employed for the quantification of MPs since reliability on a single identification method poses a risk of skipping or missing out on some categories of MPs. Physical detection is frequently used as a preliminary step for easy, low cost and rapid detection of MPs based on their appearance, color, and size. Physical detection does not effectively remove small-sized MPs but is useful for the identification of colored and larger MPs (>500 μm). Therefore, chemical approaches are used to identify the composition and structure of MPs. These include destructive and non-destructive techniques such as SEM-EDX, PLM, FTIR, Raman spectroscopy, and GC-MS.

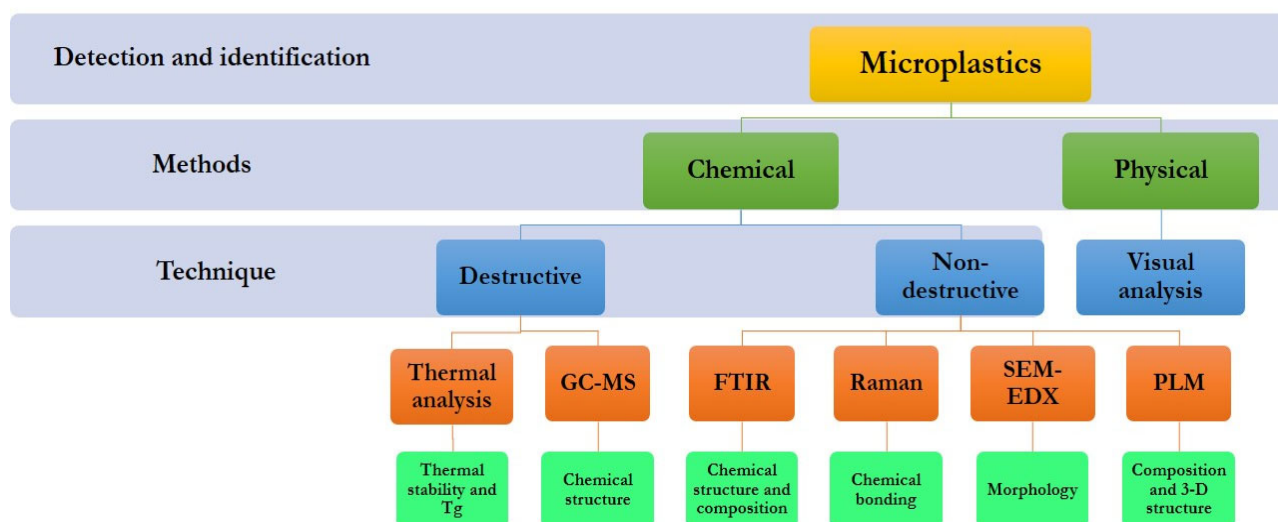


Figure 4. Methods for detection and identification of microplastics in aquatic systems.

2.1. Identification of Morphology

Scanning Electron Microscopy (SEM) is often used for the morphological analysis of MPs since it captures high-resolution images of MP surface, thereby providing information about the surface texture and deformities that helps in distinguishing MPs from other materials present in the wastewater. This technique is often used in conjunction with the Energy Dispersive X-Ray spectroscopy (EDX) technique for the analysis of constituent elements of MPs. There are certain limitations associated with SEM-EDX, i.e., high cost, low efficiency, and inability to detect colored MPs that hinder its applicability for MP detection. To improve the MP detection ability of SEM-EDX, MPs are often stained with fluorescent dyes such as Safranin T, Nile Red, and fluorescein isophosphate at high temperatures to reduce error probability.

The development of advanced microscopic techniques such as Polarized Light Microscopy (PLM) has proved to be quite efficient and useful in the determination of the type of MP. The PLM technique takes advantage of the anisotropic property of polymers and involves the passage of unpolarised light through MP particles that are placed between cross-polarizers. The polarized light emitted from the polarisers imparts information about the crystallinity of MPs and hence aids in the identification of MP polymer type. It is a reliable technique but cannot be used with thick and opaque samples.

2.2. Identification of Chemical Structure and Composition

To further improve the identification of MPs and to determine the chemical composition of MPs, certain destructive and non-destructive techniques are employed. FTIR is used for the detection of IR-active MPs by irradiation of the samples with infrared radiation and noting the changes in the dipole moments of the structural chemical bonds present in the sample. A comparison of the obtained sample spectrum with reference spectrums provides information about the composition of the MPs. Although FTIR is a suitable method for the identification of agglomerates and smaller particles, its functionality for the detection of larger particles is hindered due to limitations associated with sample size, difficult sample preparation, and the labor and time-intensive nature of this technique. To improve the detection efficiency, the FPA-FTIR (Focal Plane Array- Fourier Transform Infrared spectroscopy) technique is employed since it provides a larger spectrum for the MP particles.

The drawbacks of FTIR can be overcome by using Raman spectroscopy based on the principle of inelastic light scattering by polarized molecules. It provides images of MP particles with finer spatial resolutions of 1 μm , better than that of FTIR and the results remain unaffected by the thickness and shape of the MPs. It can be used for the identification of non-polar functional moieties and the detection efficiency can be improved by the addition of fluorescent tools as it is a highly sensitive technique. Contamination of samples with dyes, inorganic, organic, and microbial materials strongly impacts the results of Raman spectroscopy. The usage of evolved techniques such as surface-enhanced Raman spectroscopy and Raman tweezers can further improve the detection accuracy for MPs.

2.3. Identification of Thermal Properties and Chemical Bonding

Destructive identification techniques such as TGA (Thermo gravimetric Analysis), DSC (Differential Scanning Calorimetry), and GC-MS (Gas Chromatography-Mass Spectrometry) can be used as alternatives to spectroscopic methods for MP identification. TGA and DSC are used for the determination of the polymers on basis of their thermal stability and the glass transition temperature, which varies for each polymer type. The TGA and DSC plots obtained on the thermal treatment of samples are compared with the reference plots to identify the MP type and its characteristics. GC-MS is another popular and reliable technique used for polymer identification in bulk mixture samples. This technique can even be used for nanosized plastics with ease and the detection accuracy can be significantly increased by treatment of samples at elevated temperatures. This is conducted in the TD-GC-MS and pyro-GC-MS techniques. They involve the high-temperature degradation of bulk samples, followed by their segregation via gas chromatography and the subsequent analysis using mass spectrometry. These techniques offer high precision, and sensitivity and can provide qualitative and quantitative results. However, the reproducibility of results highly depends on the sample purity, sample preparation, and thermal treatment conditions. Hence, even with the myriad of methods available for the quantification of MPs, each comes with its drawbacks. Thus, there is still scope for the optimisation of detection and identification methods of MP particles. Further, the feasibility of the usage of chemical methods for MP detection still needs in-depth investigation since the interaction and accumulation of MPs on other materials may strongly influence the detection and identification capability of the above-mentioned methods.

3. Removal Methods

The rising level of microplastics in our surroundings poses an imminent threat to human and ecosystem health. Hence, there is an urgent need to devise methodologies for the detection and removal of MPs in order to prevent their bioaccumulation. A variety of

physicochemical and biological methods have been devised for MP removal and these methods can be classified into three categories (Figure 5):

1. Filtration and segregation
2. Surface adhesion and growth
3. Deterioration

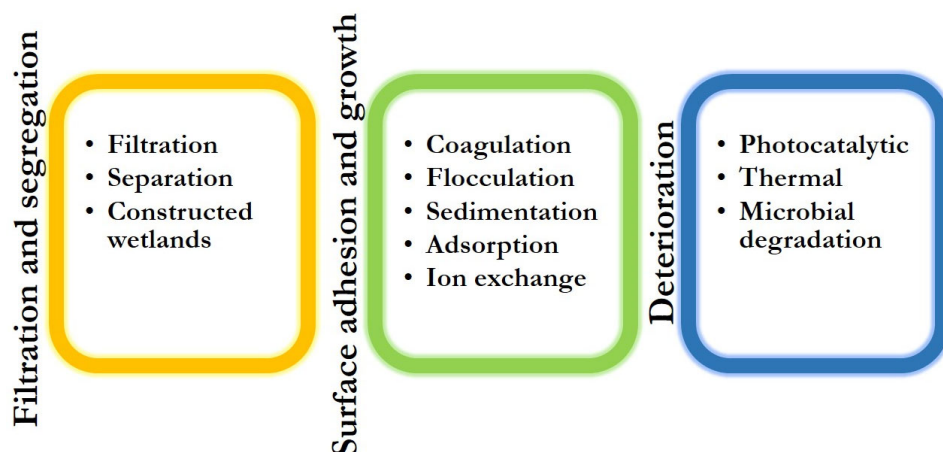


Figure 5. Classification of methods used for the removal of microplastics from aqueous media [41–46].

- **Filtration and segregation methods:** These methods involve the separation of MPs from contaminated water by physical barriers such as membranes and filter mechanisms. These physical barriers only allow the passage of liquids, thereby separating microplastics from aqueous media. However, these methods are often found to be ineffective in the removal of microplastics from sludge waste with higher viscosities. In addition, filtration methods require intensive manpower and require the movement of enormous quantities of water for the separation of micro- and nano-sized microplastics present in minimal concentrations. By using these methods, we only obtain information about the quantification of separated microplastics and do not gather any information about microplastic pollutant type and structure. To obtain detailed information about the type and structure of MPs, we need to adopt other characterisation techniques [47].
- **Surface adhesion and growth methods:** This method involves the capture and attachment of MPs onto the surface of the added materials (e.g., coagulants, disinfectants, oxidants, surfactants, etc.), causing them to form macrostructures such as aggregates, facilitating their easy removal. This methodology utilizes techniques such as coagulation, flocculation and sedimentation (CFS), adsorption, and ion exchange. Unlike the filtration and segregation methods, these methods are efficient, easy to handle and monitor, and are even helpful in the removal of other pollutants. However, due to a lack of information, they are still only performed at the pilot scale instead of large-scale operations. However, these methods possess certain limitations, i.e., they are often time-intensive and ineffective for the uptake of smooth, small-sized microplastics due to a lack of sufficient surface area to either adhere to the surface of the added materials or form flocs [48].
- **Deterioration methods:** Another method used for the separation of microplastics is the deterioration method which makes use of the action of external factors such as radiation, heat, and microorganisms to bring about changes in the physiological structure of MPs and break them down into simpler molecules such as CO₂, H₂O, H₂S, methane, etc. Photocatalytic, thermal and microbial degradation fall under this category. Degradation methods are one of the most efficient methods for combating

MP waste but these methods are not much explored and still need further in-depth studies for understanding the detailed mechanisms involved in degradation to fully exploit their potential. The breakdown capacities efficiencies can also be enhanced which can ultimately lead to a reduced degradation time span [49]. Table 2 presents the advantages and disadvantages of the above-mentioned removal methods.

Table 2. Advantages and disadvantages of methods employed for microplastic removal from aqueous media.

Removal Method	Advantages	Disadvantages
Filtration [41]	<ul style="list-style-type: none"> • High removal efficiency • Stable effluent quality • Easy to handle 	<ul style="list-style-type: none"> • Membrane fouling • Possibility of secondary MP formation • Frequent cleaning required • Little information about the mechanisms involved
Constructed wetlands [42]	<ul style="list-style-type: none"> • Less maintenance • Low operating cost 	<ul style="list-style-type: none"> • Influence of external factors not fully understood
Coagulation-Flocculation-Sedimentation [43]	<ul style="list-style-type: none"> • Simple and easy to operate • Ability to capture and remove small-sized microplastics 	<ul style="list-style-type: none"> • High requirement for chemicals • Majorly studied only in laboratories • Not widely studied at the commercial level
Adsorption and ion exchange [44]	<ul style="list-style-type: none"> • Recyclability of adsorbents and ion exchangers • Can remove MPs less than 100 µm 	<ul style="list-style-type: none"> • Large time spans for adsorption and ion exchange required • Handling adsorbents may be difficult
Photocatalytic degradation [45]	<ul style="list-style-type: none"> • No requirement for chemicals • Environment-friendly • Applicable for multiple MPs types such as PE, PS, PET, etc. 	<ul style="list-style-type: none"> • Low efficiency • May produce harmful by-products • No selectivity
Microbial degradation [46]	<ul style="list-style-type: none"> • Low cost • Simple and flexible usage 	<ul style="list-style-type: none"> • Less information available • High degradation time • May produce secondary MPs

Since the above-mentioned methods are often unable to efficiently and completely remove microplastics when used singularly, they are used in conjunction with each other, thereby forming the primary, secondary, and tertiary stages of wastewater purification (Figure 6). Some of these removal methods are described below in the following sub-sections.

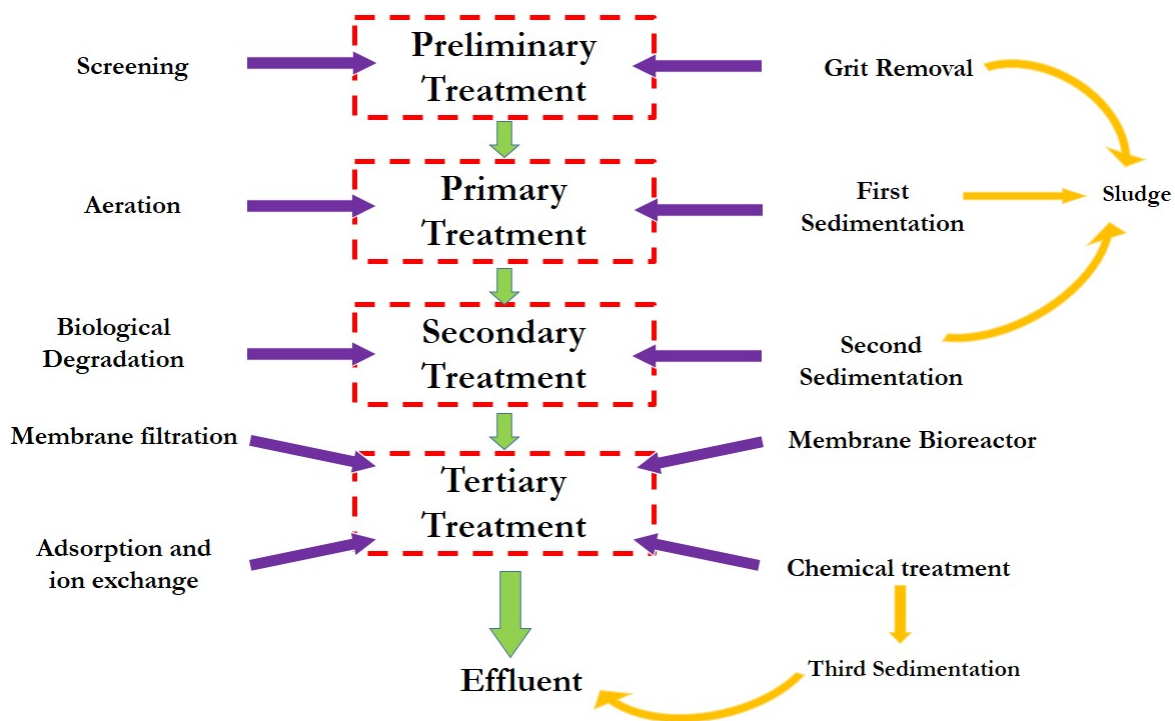


Figure 6. Schematic representation of methods and stages in wastewater treatment.

3.1. Membrane Filtration

Membrane filtration is an advanced technique that has been recently developed especially for MP removal and involves the movement of polluted water across a membrane with pore sizes varying according to variation in the shape and size of MP particles. The use of membranes is being adopted increasingly due to the low energy requirements, facile and flexible operation, easy scalability, and stability of the method. Membrane filtration is an umbrella term and is used for multiple methods such as ultrafiltration, microfiltration, nanofiltration, and, reverse osmosis [50–53]. Membrane bioreactors (MBR) and dynamic membrane (DM) systems have also been established for efficient MP removal. Membranes with varied pore sizes and external conditions such as pressure, and pumping shear stress help in the effective operation of these systems. The permeability and selectivity of membranes, their durability, the size and concentration of MPs, and influent flux are key factors that influence MP removal efficiency by a large degree. Membrane processes have been known to exhibit removal efficiencies of up to 99.9% when used in combination with other techniques. For instance, Lares et al. (2018) devised an advanced MBR system incorporated with WWTP to analyze MP removal. They compared the MP removal efficiency of the MBR device with a conventional activated sludge method and concluded that MBR showed higher removal efficiency towards MP removal than the conventional activated sludge method [35]. Tadsuwan et al. [54] investigated the outcome of coupling ultrafiltration with a pre-existing water treatment plant and reported an increase in removal efficiency from 86.14% in a traditional water treatment plant to 96.97% combined with an ultrafiltration setup. Li et al. [55] formulated a dynamic membrane (DM) on a 90 μm mesh via synthetic wastewater filtration and investigated the impact of control parameters on the functioning of the DM. They concluded that the DM was formed rapidly and was quite effective in MP removal, and it was promoted by the increased motion of solids and concentration of influent particles [55].

The membranes used in membrane filtration often suffer from fouling phenomenon caused by the interaction and accumulation of MPs in the membrane pores and the

growth of microorganisms on them, leading to their clogging, and thereby reducing their removal capability (Figure 7). Therefore, a need often arises to pre-treat the polluted water with disinfecting agents or coagulants to prevent this phenomenon; although pre-treatment of water often reduces the probability of membrane fouling. A study by Xing et al. describes a low-dosage UV/Chlorine pre-oxidation strategy to prevent membrane fouling. They reported the pre-oxidation method was for reducing membrane fouling by 49% [56]. It is also suspected that cleaning and backwashing of membranes may aggravate the problem of MP release. Membrane filtration is quite successful in the removal of fragment and pellet MPs due to synergistic interactions between the MP particle and membrane material and pores. However, they are less useful for fiber MPs since fiber-type MPs may move longitudinally through the pores of the membrane. Ziajahromi et al. [57] prepared an MP sampling device and studied the microplastics present in effluents released from three different WWTPs that used treatment methods such as CFS method, biological treatment, disinfection/de-chlorination processes, and ultrafiltration, followed by reverse osmosis (RO) process. They detected the presence of microplastic fibers in the effluents even after reverse osmosis and attributed it to the existence of membrane defects [57]. Thus, an in-depth study of the MF methods is required to enhance their efficiency for MP removal at the pilot scale.

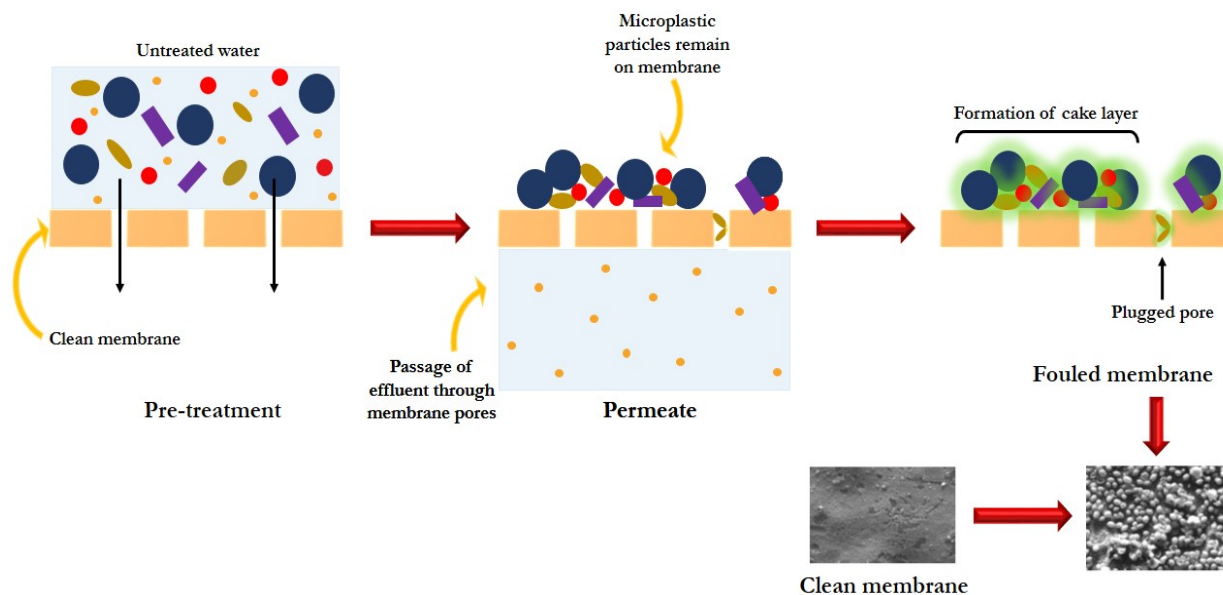


Figure 7. Schematic representation of microplastic removal by membrane filtration and membrane fouling [58].

3.2. Adsorption

The adsorption method takes precedence over many other methods for pollutant removal from aqueous media due to its facile nature, high efficiency, economical usage and other advantages. Adsorption is a surface phenomenon and involves the uptake of pollutants on the adsorbent surface by means of weak Van Der Waal interactions. A variety of materials such as carbonaceous materials, zeolites, polymers, and inorganic clays have been utilized for the removal of MPs from water sources. The high number of adsorption sites, nature, and strength of the adsorbent are deciding factors in the adsorptive and regenerative capabilities of adsorbent materials. Tang et al. synthesized magnetic carbon nanotubes and reported the adsorption of PE, PET, and PA microplastic particles. The adsorbent exhibited 100% removal efficiency for the MP particles and showed maximum adsorption capacity for PE, then PET and least for PA, and exhibited <80% removal efficiency even after four adsorption cycles [59]. Recently biosorption has

emerged as a viable and effective method for MP uptake (Figure 8). The use of biomass, bacteria, fungi, algae, seaweed and other industrial and agricultural biowaste is being highly favoured as it does not lead to the discharge of secondary MPs into water bodies. The adsorption process with such biomaterials generally proceeds via physical adsorption, ion exchange, chelation, microprecipitation and complexation mechanisms in the extracellular technique. The presence of hydroxyl, amine, carboxyl, and phosphonate groups in the cell walls of microbes and plant bodies aids in microplastic adsorption. In a recent study by Sundbæk et al. [36], the marine algae *Fucus vesiculosus* was used for the adsorption of MPs. The constituent carboxylic groups of alginic acid present in algae cell walls are responsible for the binding of MPs to the adsorbent surface. A detailed report by Siipola et al. showed the usage of steam-activated pine and spruce bark-based biochar for the purification of urban wastewater and runoff. They focused on determining the effect of features of biochar adsorbent, such as chemical composition and particle size, and concluded that these bioadsorbents were suitable even for the removal of very small-sized microplastic particles. They performed the mechanism for retention of MP particles on bioadsorbent surfaces and still more research needed to be conducted to gain a deep understanding of the adsorption mechanism [60]. Sun et al. fabricated chitin and graphene oxide-based compressive sponges that can adsorb MP particles at pH 6–8 with a high removal efficiency of 89.7% and offers recyclability for up to three cycles. The sponge was also found to be biocompatible as it did not inhibit the growth of green algae on its surface and could be broken down by microorganisms present in the soil, confirming its biodegradability [61]. Although adsorption is an easy and effective method, drawbacks such as time and labor intensiveness limit its advantageous usage. Thus, it may be used in combination with other advanced techniques for better MP removal efficiencies.

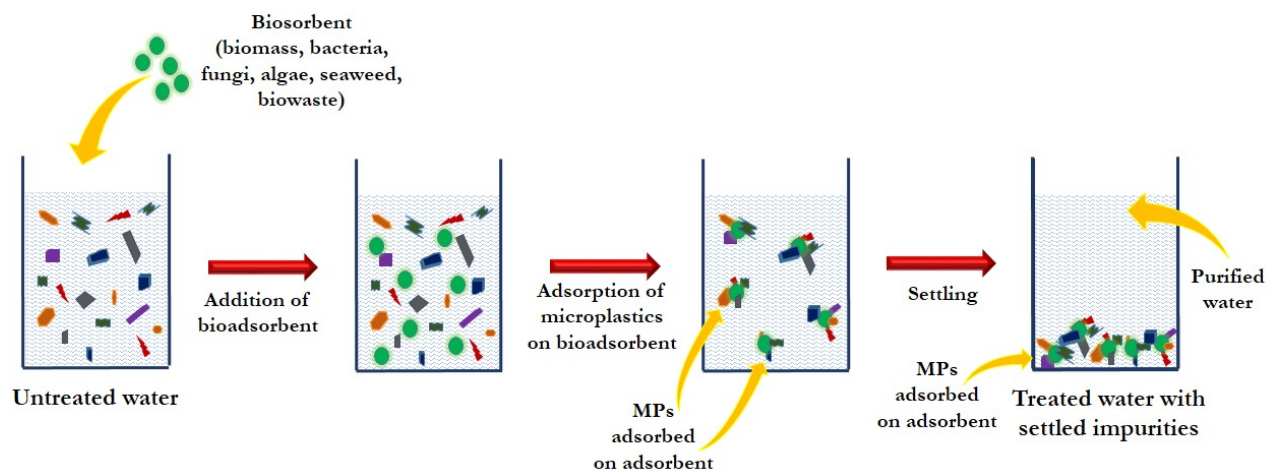


Figure 8. Schematic diagram illustrating the bio-adsorption of microplastics.

3.3. Coagulation, Flocculation and Sedimentation (CFS)

The combination of coagulation, flocculation and sedimentation is the most widely used method for MP uptake from water sources. The CFS mechanism involves the heterogeneous separation of solids and liquids and is monitored by the density of solids and liquids (Figure 9). Coagulation involves the destabilisation of suspended particles in a colloid by the addition of a coagulant material such as metal salts. It is a rapid method often followed by flocculation. The flocculation technique involves slow mixing for long time intervals, leading to the aggregation of previously destabilized particles to form large aggregates (flocs) that can then be easily removed by sedimentation. Flocs formed during stirring are influenced by the characteristics of aqueous media such as ionic strength, pH, divalent cations, natural organic matter, and particulate/colloidal matter.

Zhang et al. [62] synthesized a magnesium hydroxide- Fe_3O_4 -based magnetic coagulant and applied it for microplastic removal. They reported a removal efficiency of 87.2% and also explored the influence of aging time on floc formation. They also concluded that removal efficiency above 85% can be maintained in water samples within the pH range of 5–8 and that charge neutralisation is an important mechanism involved in microplastic removal [62]. The sedimentation technique is based on the gravitational settling of suspended aggregates and is impacted by microplastic particle density. It is especially helpful in the elimination of irregularly shaped MP fragments since angular and irregularly shaped particles can be easily captured to form bigger aggregates that can settle down due to increased density. These methods are often used together to enhance MP removal capacities and the removal efficiency of the CFS method is monitored by the physiochemical and morphological properties of MPs, i.e., shape, size, and surface properties. These methods are often employed as primary or secondary treatment methods and are often used in conjunction with other advanced techniques to maximize MP removal efficiency. The CFS method is more successful in removing fibrous MPs than spherical or fragmented MPs due to the availability of a larger surface area, facilitating more interaction with flocculating agents. Peydayesh et al. (2021) investigated the uptake of carboxylated PS microspheres from various water samples using a lysozyme amyloid fibril natural bioflocculant by CFS technique with 98.2% removal efficiency [37]. Pivokonský et al. tested raw and purified water from two different drinking water treatment plants (DWTPs) and identified that the CFS method is quite effective in eliminating microplastic particles from the water samples [63]. Lapointe et al. examined the performance of the CFS method for the sequestration of pristine and weathered PE, PS and PEST microplastic particles and noted that the removal efficiency was found to be maximum at 97% for weathered PEST particles. They also explored the use of settled water turbidity as a possible indicator for the removal of MPs [43]. The CFS method suffers from drawbacks such as high chemical consumption, large power requirements, and frequent electrode replacement, limiting their cost-effective usage for water purification.

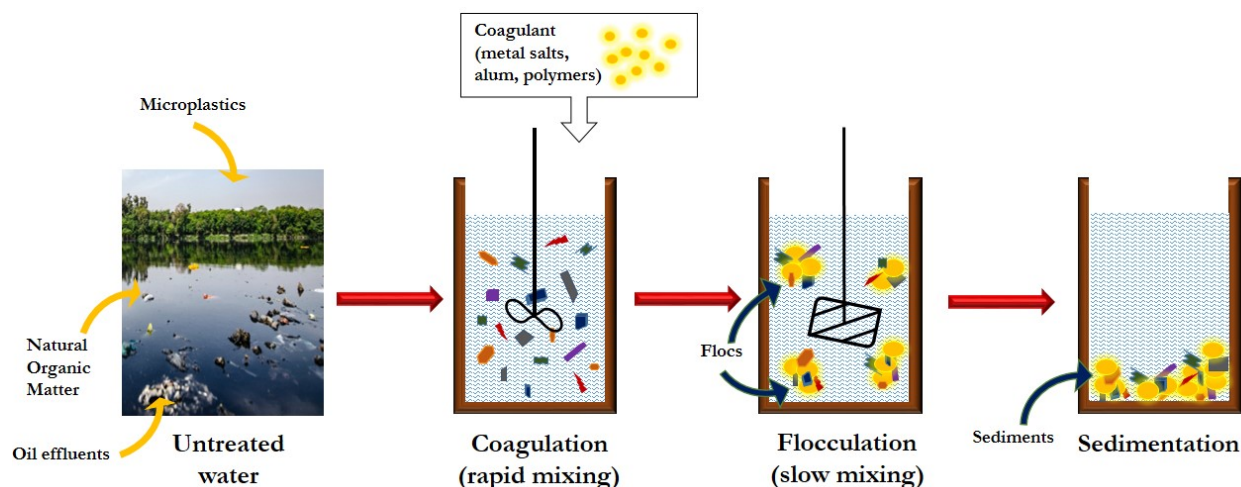


Figure 9. Microplastic removal by coagulation, flocculation and sedimentation processes.

3.4. Biological Degradation

The biodegradation of MPs is an environmentally benign method for the eradication of MPs from aqueous systems. A variety of microbes such as fungi, diatoms, bacteria, biofilms, etc., are reported to induce the degradation of PE, PP, and PS microplastics. These microbes grow and form colonies on the MP surface and use them as carbon sources, thereby leading to their degradation, producing secondary MPs, CO_2 , methane, H_2O , and biomass (Figure 10) [64]. The biodeterioration process may be aerobic or an-

aerobic, depending on the presence of oxygen, and is influenced by external factors such as temperature, humidity, UV and solar radiation. The efficacy of biodegradation mainly depends on the type of polymer, its characteristics and morphology, and the molecular weight. Auta et al. [38] studied the impact of isolated *Bacillus gottheili* bacteria on the deterioration of PE, PET, PP, and PS over a span of 40 days [38]. They observed changes in the surface texture as well as the formation of grooves and cracks and concluded that bacteria affect the surface and bulk properties of MPs. The marine fungus *Zalerionmaritimum* was used by Paco et al. [39] for the degradation of PE pellets and they observed molecular changes in the MPs along with a reduction in PE pellet mass and size, confirming their degradation. Biodegradation of PVC microplastics using the larvae of *Tenebrio molitor* was investigated by Peng et al. and they observed partial biodegradation of polymer along with the formation of smaller chlorinated organic compounds and reduction in M_n by 32.8%. They concluded that the ingestion rate was slow and the mineralisation of PVC microplastic powder was only partial [65]. Huang et al. [66] surveyed the distribution of microplastics on biofilms consisting of filamentous algae in the Middle Route of the South-to-North Water Diversion Project (SNWDP) in China, a regulated canal and reported that MPs were concentrated on the biofilms, with small PET fibers being the major category MPs present in the biofilms. They concluded that these biofilms could be used as a sink for microplastics. Although, the usage of microorganisms for MP degradation is a flexible and tunable technique, the time duration required for these methods is substantial. Additionally, there are challenges in scaling up these methods, and they are suspected to release secondary MPs along with the non-reusability of the microbes. The degradation mechanisms involved require further in-depth investigation to fully utilize the potential of microbes for MP degradation.

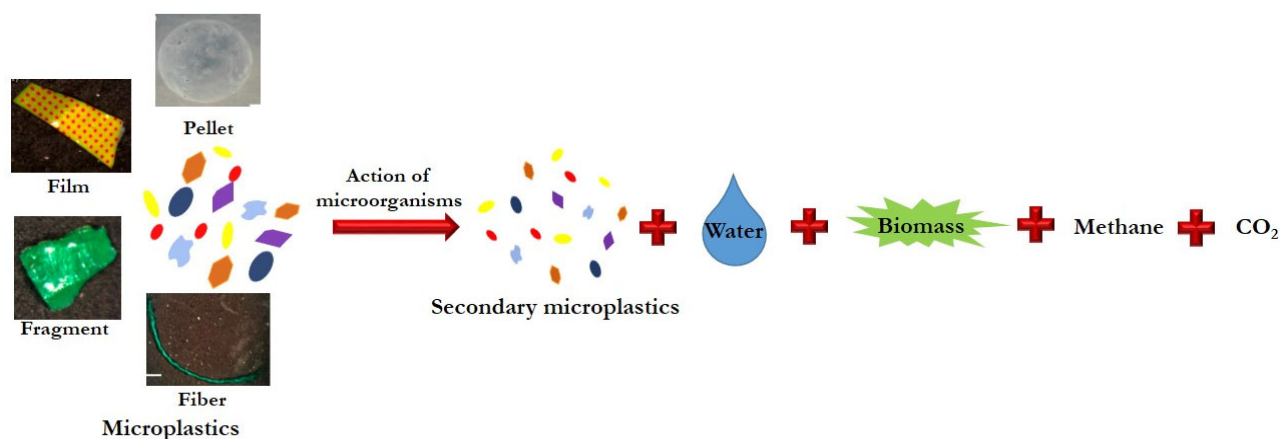


Figure 10. Microbial degradation of microplastics [67].

3.5. MP Shape, Size, and Polymer Type and Their Impact on Efficiency of Removal Methods

The shape, size and polymer type of MPs have a vital impact on the effective extraction since particles with different shapes, sizes and compositions exhibit varying properties and toxicities. Shape and size are helpful in the analysis of accumulation patterns of MPs in water bodies since low-density MP particles, such as fibers and fragments, tend to stay afloat on the water surface while MP particles such as pellets that possess relatively higher densities sink to the bottom of water bodies. Due to their ubiquitous presence, there arises the need to understand and identify the transportation and dispersal mechanisms across soil and water systems.

3.5.1. Impact of Shape of MP

MPs occur in many shapes in aquatic systems but the majority of MPs exist as fragments, pellets, spheres, films, and fibers. Among these forms, fragments and fibers are the most prevalent in water bodies [40,41]. The biggest contributor of fiber MPs to water bodies is the garment industry, leaching fiber-type MPs into aquatic systems via effluents released from their washing processes. Film-type MPs usually originate from the weathering of packaging materials and plastic bags while the usage of MP pellets for abrasion applications in the cosmetics industry contributes to their release. Fragment MPs are usually secondary MPs and often arise from the degradation of bigger plastic objects. It has been observed that fiber MPs are typically eliminated in the primary treatment step by techniques such as coagulation, flocculation and sedimentation (CFS). While during the secondary treatment stage, maximum removal of fragment-type MPs occurs since their lamellar structure facilitates their agglomeration and subsequent removal. Substantial removal of MP pellets is seen in the tertiary treatment stage, involving the use of advanced filtration and oxidation techniques. The tertiary treatment processes are most favoured for the removal of MPs with very small sizes and have distinct features. However, it was noted that the MP removal efficacy of tertiary treatment stages is lesser than those of primary and secondary treatment stages.

3.5.2. Impact of Size of MP

The size of MP particles also influences the determination of removal efficiency. During the primary treatment processes, large MP particles having sizes larger than 0.5 mm are easily removed by methods such as flocculation, sedimentation and grease removal. Large-sized fiber and film-type MPs are easily removed by flocculation and grease removal methods due to their low densities while small MP pellets sink to the bottom of containers due to action of gravity on these high-density particles. Additionally, due to weathering phenomena, primary microplastics often break down into secondary MPs that are ingested by aquatic organisms and lead to bioaccumulation and toxicity. Thus, the removal of secondary MPs holds utmost importance but they cannot be extracted by conventional methods, thereby requiring the use of sophisticated techniques and instrumentation for their efficient removal.

3.5.3. Impact of Polymer Type of MP

The ubiquitous utilisation of PE, PP, PET, PU, PA, PAAm, PVC, PES, PS, PEVA and other such polymers can be ascribed to their versatile properties and stability. Among these, PE and PS are the most highly favoured materials due to their excellent impact and chemical resistance, low production costs, and easy workability, thus finding application in multiple industries. Due to their vast usage, the proportion of PE and PS MPs present in aquatic systems is much more than other commodity MPs. PE and PS MPs are positively charged, and hence they can be effectively removed from wastewater by using secondary treatment processes. Therefore, the impact of the morphological attributes and types of polymers on the efficient removal of MPs needs to be studied in detail. It will also help in gaining a deeper understanding of the removal mechanisms by conventional and advanced MP removal technologies.

4. Accumulation of MPs in the Ecosystem and Their Toxicity Assessment

The environment can be affected by MPs in a variety of physical, chemical, and biological ways. Various physical injuries may occur to animals when they become entangled in microplastics in the environment, including drowning, suffocation, strangulation, and starvation [68]. The chemical impact of MP on the environment is attributed to the adsorbed chemicals onto plastic surfaces. MPs are composed of highly hydrophobic materials, making them a potentially toxic chemical reservoir. Furthermore, the presence of excessive levels of MPs in ecosystems can also impair the normal physiological func-

tioning of living organisms [27]. Throughout the food web, MPs may pose an environmental risk because of their bioavailability. Aquatic and terrestrial environments contain a high concentration of MPs, which would be present in food products consumed by humans.

4.1. Impacts of Microplastics on Human Health

A recent report released on Microplastics in Drinking-water (2019) by World Health Organisation highlighted the ubiquitous presence of MPs in the ecosystem and raised concerns about their adverse effects on human health [69]. Growing concern over microplastics' potential health impacts has been raised since microplastics can enter the human food supply via the ingestion of terrestrial foods and seafood. The existence of MPs in foods consumed by humans has also been highlighted by many groups (Table 3).

Table 3. Presence of MPs in food items and drinks consumed by humans.

Consumable Products	Polymer Types	Size	MPs Concentration	References
<i>Seafood</i>				
Bivalve (oyster, mussel, Manila clam, and scallop)	PE, PP, PS, PES, PEVA, PET, PUR	0.1–0.2 mm	0.97 (0–2.8) particles/individual 0.15 (0–1.8) particles/g	[70]
Canned Sardines	PE, PET, PVC, PP	190–3800 µm	6 MPs per item	[71]
Fish	PET, PP, PUR, PES	<500 µm	2.2 ± 0.89 MPs/individual	[72]
<i>Acanthopagrus australis</i> (Yellowfin bream)	PET, RY, PES	-	Mean 0.6 MPs/fish	[73]
Pelagic and demersal fish	Cellulose, PA, RY	0.13–14.3 mm	1.90 particles/individual	[74]
<i>Engraulis japonicus</i> (Japanese anchovy)	PE, PP, PS	150–1000 µm	Mean 2.3 MPs/individual	[75]
<i>Fenneropenaeus indicus</i> (Indian white shrimp)	PA, PES, PE, PP	0.157–2.785 mm	0.39 ± 0.6 items/shrimp 0.04 ± 0.07 items/g	[76]
<i>Mytilus edulis</i> (Mussels)	CPH, PET, PES PE,	0.033–4.7 mm	0.9–4.6 particles/individual 1.5–7.6 particles/g	[77]
<i>Meat</i>				
Poultry, cows, and pigs	PP, PE, PET	<5 mm	Poultry manure: 667 ± 990 particles/kg Cow manure: 74 ± 129 particles/kg Pig manure: 902 ± 1290 particles/kg	[78]
Chicken gizzards	PS	0.1–5 mm	10.2 ± 13.8 particles/g	[79]
<i>Salts</i>				
Salt	CPH, PE, PET	<200 µm	Lake salt: 43–364 particles/kg Rock salt: 7–204 particles/kg Sea salts: 550–681 particles/kg	[80]
Sea/lake/rock salt	PE, PET, PP	<500 µm	Lake salt: 28–462 particles/kg Rock salt: 0–148 parti-	[81]

			cles/kg Sea salts: 0–1674 parti- cles/kg	
<i>Drinks</i>				
Tea	PA, PET	25 μ m	~11.6 microplastics/cup of the beverage	[82]
Drinking water	PET, PE, PA, PP	0.005–0.1 mm	1 \pm 8 particles/L (beverage cartons) 118 \pm 88 particles/L (returnable plastic bottles)	[83]
Milk	Polysulfone	0.1–5 mm	6500 particles/m ³	[84]
Drinking water	PES, PVC, PE, PA, EP	0.05–0.105 mm	0–7000 particles/L	[85]
Beer	-	0.1–5 mm	0–14.3 particles/L	[86]
<i>Sugar and honey</i>				
Honey	PP, PE, PAAm	0.013–0.25 mm	54 particles/L (industrial honey) 67 particles/L (craft honey)	[87]
Honey	-	0.01–9 mm	166 \pm 147 particles/kg (fibers) 9 \pm 9 particles/kg (fragments)	[88]
Sugar			217 \pm 123 particles kg ⁻¹ (fibres) 32 \pm 7 particles kg ⁻¹ (fragments)	

Note(s): PE—Polyethylene, PUR—polyurethane, PP—Polypropylene, PA—Polyamide, PET—Polyethylene terephthalate, PAAm—Polyacrylamide, PES—Polyester, PVC—Polyvinyl chloride, PS—Polystyrene, PEVA—Poly (ethylene-co-vinyl acetate), RY—Rayon, EP—Epoxy resin, CPH—Cellophane.

In addition, MPs are also absorbed into the body when they are inhaled and come in contact with skin [89,90]. Microplastics are ingested mainly through food products such as table salt, mussels, sugar, commercial fish, and even water, which are contaminated with microplastics. MPs may enter the digestive tract, triggering inflammatory responses, increasing intestinal permeability, and altering the metabolism [91]. The toxic effects of MPs exposure in humans are presented in Figure 11.

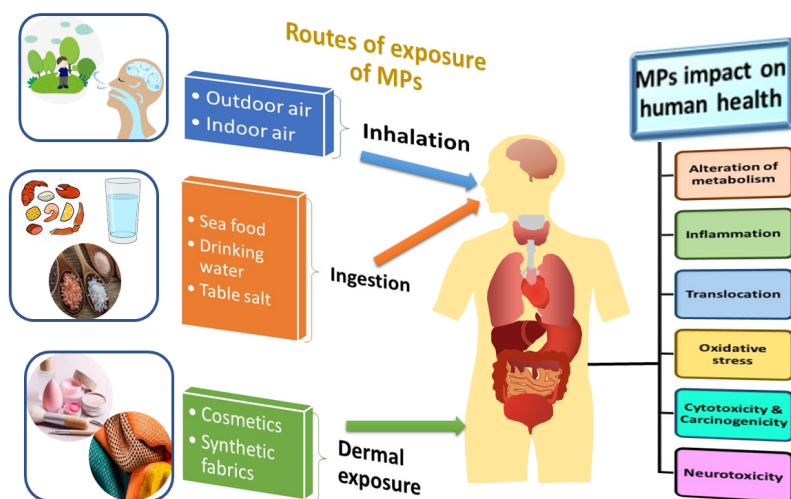


Figure 11. Schematic illustration of exposure of MPs on human health.

Various sources release MPs into the atmosphere, including textiles, abrasion of car tires, buildings, etc., and microplastics resuspension from surfaces. Prata (2018) reported that the amount of MPs inhaled by individuals per day ranges from 26 to 130 MPs. Consequently, this could be hazardous to human health because of their polymeric structure which makes their removal from the respiratory system difficult and they release toxic organic pollutants and plasticizers from their surface [92]. The risk of inhalation of MPs from wearing different masks during the COVID-19 pandemic was investigated by Li et al. [58]. They reported that wearing N-95 masks posed a lower risk of inhalation of fiber-like MPs as compared to the activated carbon masks [93].

Currently, almost negligible studies have been conducted to evaluate the associated risks of dermal exposure to MPs in humans. Although the extensive usage of microbeads in personal care products and synthetic fabrics along with the presence of microplastics in indoor dust particles leads to considerable human exposure to MPs via dermal contact. Microplastic beads having a size of less than 1 mm have been extensively utilized in facial scrubs, toothpaste, and dentures [94]. Human exposure to MPs via dermal contact has not been comprehensively studied, a few studies have only assessed the per capita consumption of MPs. A study by Napper et al. [26] showed that usage of facial scrubs by the UK population alone is discharging 40.5–215 mg of polyethylene microbeads person⁻¹ day⁻¹. Although human skin is susceptible to penetration by particles less than 100 nm in size, microplastics may penetrate through hair follicles, open wounds, or sweat glands to cause skin damage. It is imperative that in-depth research be carried out on human dermal exposure to MPs via cosmetics, settled dust particles, and fabric fibers so that the significance and health risks associated with these exposure routes can be determined [95].

The risk of ingesting microplastics has not yet been quantified completely due to the relatively limited amount of research. Although several groups have performed in vitro studies to assess the toxicological effects of microplastics on human health, there is still a lack of availability of data on in vivo studies. In this section, some of the studies which examined the toxicity of MPs on human cell models have been discussed. The cell viability and cytotoxicity of microplastics in terms of oxidative stress were investigated by Schirizzi et al. [96] on cerebral (T98G) and epithelial (HeLa) human cells. The results demonstrated that in both cases, i.e., with exposure to polyethylene (3–16 µm) and polystyrene (10 µm), cell viability was not affected. Wu et al. [97] studied the cytotoxicity of 0.1 and 5 µm polystyrene MPs on human Caco-2 cell lines. The results indicated that both MPs exhibited weak cytotoxicity and displayed negligible changes in membrane integrity, whereas both sizes lead to disruptions in mitochondrial potential with larger MP sizes producing greater disruption than smaller MP sizes. Stock et al. [98] also studied the cytotoxic effects of polystyrene MPs (sizes ranging from 1, 4, and 10 µm) on the human Caco-2 cells and monocyte-like THP-1 cells. The results indicated that MPs of 1 µm size affected the cell viability of Caco-2 cells. Hesler et al. [99] performed in vitro analysis of the toxicological effects of modified polystyrene (0.5 µm) at the human intestinal and placental barrier. The MPs exhibited no genotoxicity and weak embryotoxicity. In vitro analysis performed by Xu et al. [100] confirmed the cytotoxicity induced by PVC particles (2 µm) in human pulmonary cells. A particle's size and density determine the deposition of MPs into the human respiratory system, smaller and less dense particles penetrate the lungs more deeply. The toxicology of these particles needs to be further researched on a multidisciplinary and international scale in order to understand their long-term impact on humans.

4.2. Impacts of MPs on Aquatic Environments

As microplastics (MPs) accumulate in the environment, aquatic life is becoming more vulnerable. Secondary microplastics resulting from the fragmentation of automobile tires, packaging materials, paints, synthetic fabrics, etc., are the primary source of contamination of water resources [101]. Poor waste management is also responsible for

introducing microplastics into freshwater through runoff from surface and agricultural areas (Figure 12). It has been identified that wastewater and sewage treatment plant effluent discharges are a chief source of introducing MPs into the freshwater. As MPs float on the water surface, disperse throughout the column, and accumulate in seawater sediments, they can be consumed by diverse aquatic organisms occupying a variety of habitats [1]. The impact of MPs on aquatic environments can be classified as physical and chemical. Aquatic species can be physically impacted by the MPs by entanglement or by ingestion, with the former being the more common. Allsopp et al. [102] revealed that entanglement results in the suffocation, drowning, or starvation of aquatic animals. It has been reported that ingested plastic fragments have caused physical injuries in animals, including ulcerations and rupture of the digestive tract. Aquatic organisms are incapable of differentiating between MPs and natural prey items resulting in accidental uptake of MPs, leading to an influx of microplastics into the aquatic food web [103].

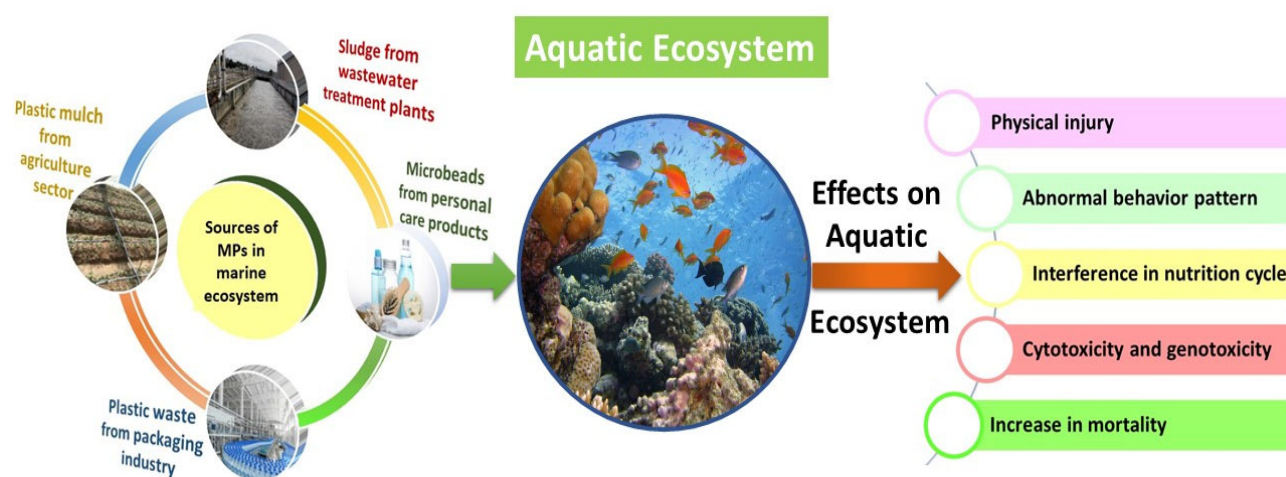


Figure 12. An overview of sources and toxic effects of MPs on aquatic environment.

A number of adverse effects may be induced by MPs on aquatic species, such as behavioral changes, slow metabolism, and disruption in growth and reproduction. When aquatic organisms are severely overloaded with MPs, they display lethargic swimming and feeding behavior. Yin et al. [104] showed that the accumulation of MPs in the digestive tract may result in abnormal behavior in fish. A study by Chen et al. [105] revealed that exposure to PS-MPs at 1 mg/L concentration suppressed the catalytic activity of acetylcholinesterase (AChE) on zebrafish larvae. The inhibition of AChE activity subsequently results in the over-stimulation of receptors and may result in paralysis and death as a result of a significant build-up of AChE in synaptic clefts. Consequently, prolonged exposure to MPs could influence the nutritional status of fish, thus affecting their health and growth.

According to Sussarellu et al. [106], polystyrene (PS) MPs adversely affect oyster reproduction and feeding by altering their food intake and energy distribution. The quality of oocytes, motility of sperm, and egg production were all reduced in oysters exposed to micro-sized polystyrene. The eggs and sperm are released in the sea for external fertilisation by oysters, but because of the intake of MPs, the sperm's speed and count were significantly reduced [106]. Two-month exposures of adult oysters (*Pinctada margaritifera*) to 6 and 10 µm polystyrene microbeads were conducted by Gardon et al. [107] to examine the effect of MP on their physiology. *P. margaritifera*'s assimilation efficiency, energy balance, and reproduction were significantly affected by PS-MPs. Cole et al. [108] examined how MP consumption affects the fertility, feeding habits, and functioning of the copepod *Calanus helgolandicus*, which because of its size, lipid content, and opulence, is a vital prey species for many fish larvae. As a consequence of exposure to 20

μm PS microbeads, the carbon biomass of copepods was reduced by 40%, resulting in an energy reduction and increased consumption of lipids, which affected their growth [108]. Banaee et al. [109] studied the effect of polyethylene MPs on various biochemical parameters of blood on *Emys orbicularis* (pond turtle). The MPs exposure adversely affected all the parameters studied which indicated liver and kidney dysfunction [109]. As a result of exposure to PS microplastics, *Danio rerio*'s (zebrafish) metabolic pathways were changed and its lipid and energy metabolism was altered [110]. In addition, Lei et al. studied the toxic effects of five varieties of MPs on Zebrafish and reported that *D. rerio* was observed to develop intestinal damage after exposure to microplastic particles. Moreover, they also demonstrated that MPs' lethality is not dependent on their chemical composition, but on their size.

In a study by Kaposi et al. [111] PE microspheres (10–45 μm) were exposed to *Tripneustes gratilla* (sea urchin) larvae for 5 days. They concluded that although a significant decrease was observed in larval body width, the current MP levels in the ocean present only a limited threat to marine invertebrates. Weber et al. [112] indicated that despite high levels of MP contamination, *Gammarus pulex* (amphipod) did not show significant effects on development, survival, feeding behavior, or metabolism (glycogen, lipid storage). Zhang et al. [62] assessed the toxic effects of variably sized PVC MP on *Skeletonema costatum* (marine algae). The results suggested that small-sized PVC MPs adversely affected photosynthesis and inhibited the growth of microalgae. PVC MPs have been exposed to marine *Perna Viridis* (Asian green mussels) by Rist et al. [113]. It was observed that there was an increase in mortality after MP exposure, with decreased filtration and respiration rates, as well as reduced motility [113]. A study by Rochman et al. [114] demonstrated the impact of short-term exposure to PE fragments in marine *Oryzias latipes* (Japanese medaka fish) and observed that this led to the bioaccumulation of chemicals and early tumour development. When *Arenicola marina* L. (lugworms) was exposed to a high concentration of PVC MPs, the immune function of lugworms was impaired and a high mortality rate was observed in a study by Browne et al. [115]. Table 4 summarizes the toxic effects of some of the MPs on different aquatic species.

Table 4. Effects of MPs on aquatic organisms.

Aquatic Organisms	Polymer Types	Size	Effects	References
Zebrafish Larvae	PS	45 μm	Suppressed catalytic performance of AchE	[105]
Oyster	PS	2 and 6 μm	Reduced sperm count and speed	[106]
<i>Pinctada margaritifera</i> (Oyster)	PS	6 and 10 μm	Reduced assimilation efficiency and reproduction	[107]
<i>Calanus helgolandicus</i> (Copepods)	PS	20 μm	Reduction in carbon biomass	[108]
<i>Emys orbicularis</i> (Pond turtle)	PE	-	Adverse impact on the liver and kidney functioning	[109]
<i>Danio rerio</i> (Zebrafish)	PS	5 and 20 μm	Inhibited liver functions and metabolism of fish	[110]
<i>Danio rerio</i> (Zebrafish)	PA, PE, PVC, PP	70 μm	Damage to intestine	[116]
<i>Tripneustes gratilla</i> (Sea urchin)	PE	10–45 μm	Decreased larval width and survival affected by 50%	[111]
<i>Gammarus pulex</i> (Amphipoda)	PET	10–150 μm	Metabolic rate, behavior, and growth were not affected	[112]
<i>Skeletonema costatum</i> (Microalgae)	PVC	1 μm and mm	Inhibition in growth and affected photosynthesis	[117]
<i>Perna viridis</i>	PVC	1–50 μm	Negative impacts on physiological	[113]

			functions of mussels	
(Asian green mussel)				
<i>Scrobicularia plana</i> (Bivalve mollusc)	PS	20 µm	MPs inhibited antioxidant activity, damaged DNA, and caused neurotoxicity and oxidative stress.	[118]
<i>Euphausia superba</i> (Antarctic Kill)	PE	27–32 µm	Loss in weight	[119]
<i>Oryzias latipes</i> (Japanese medaka fish)	LDPE	-	Resulted in formation of tumours, liver damage, and accumulation of toxic chemicals	[114]
<i>Oryzias latipes</i> (Japanese medaka fish)	PE	<1 mm	Adverse effects on reproduction and growth	[120]
<i>Arenicola marina</i> L. (Lugworms)	PVC, PS	<10 µm	Mortality and dysfunction of immune system	[115]
<i>Ostrea edulis</i> (Flat Oysters)	HDPE and PLA	Varying sizes	Increase in respiration rate	[121]
<i>Crangon crangon</i> L. (Brown shrimp)	-	200–1000 µm	No adverse impact on the shrimp's nutritional condition	[122]
<i>Ciona intestinalis</i> (Sea squirt)	PS	1 µm	Negative effects on growth and food intake	[123]
<i>Crepidula onyx</i> (Mollusca)	PS	2 µm	Growth inhibition	[124]

Note(s): PP—Polypropylene, PA—Polyamide, LDPE—Low-density polyethylene, PET—Polyethylene terephthalate, PE—Polyethylene, PVC—Polyvinyl chloride, PS—Polystyrene, PLA—Polylactic acid, HDPE—High-density polyethylene.

Green [121] studied the impact of high-density PE and PLA on *Ostrea edulis* (Flat Oysters). It was observed that the rate of respiration increased with an increase in microplastic concentration. In a study by Devriese et al. [122], it was reported that microplastic ingested by *Crangon crangon* L. (Shrimps) showed no significant negative impact on nutritional conditions. In conclusion, in-depth research is required to better comprehend the impacts of microplastics on marine biota, and further research is necessary to fill knowledge gaps.

4.3. Impacts of MPs on Soil

Scientists have paid minimal attention to MPs pollution in soil environments as compared to marine ecosystems, despite a more significant accumulation of MPs in terrestrial soils. Microplastic contamination is especially prevalent in agricultural and urban soils as a consequence of human activities. The soil ecosystem is exposed to a wide range of MPs due to the over-exploitation and haphazard management of plastic wastes [125]. Agricultural practices, including plowing and harvesting, influence the horizontal distribution of MPs in soil, whereas vertical distribution is governed by soil macropores and the cracking of soil [126]. The extent of transportation, deposition, and retention of MPs are influenced by numerous aspects such as (1) human activities including littering and inefficient waste handling, (2) physicochemical properties of plastics including size, density, etc., and (3) atmospheric conditions (temperature, rainfall, speed of wind). The movement of microplastics in the soil can be facilitated by a variety of plant processes such as uprooting and root development as well as various organisms (vertebrates, earthworms, etc.) inputs [127]. For instance, earthworms can swallow and excrete microplastics, the movement of anecic earthworms is capable of vertically transporting microplastics from shallow to deep soils, and geophagous earthworms are responsible for horizontally spreading them across wide areas (as shown in Figure 13) [128]. In addition, soil microarthropods have been found to consume earthworm casts containing concentrated microplastics [129]. As microplastics migrate, soil properties, including microbial

diversity as well as soil structure and function are altered which could have a negative effect on plants and animals, as well as threaten the quality and safety of food.

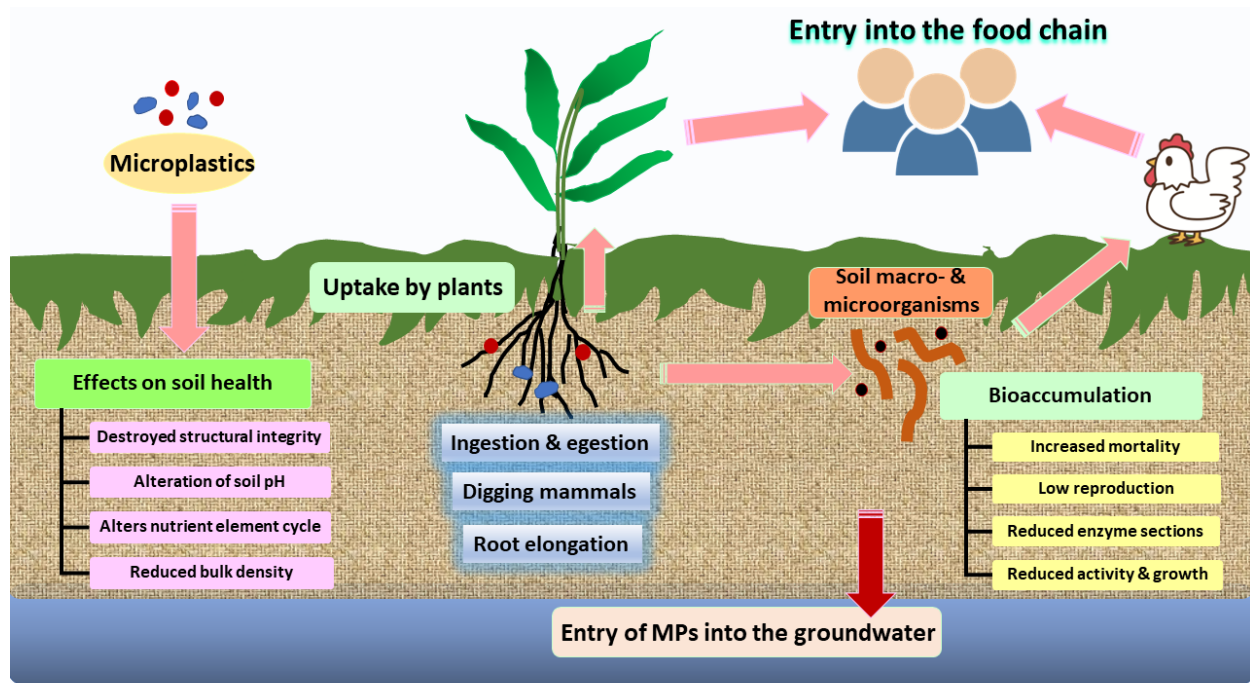


Figure 13. The toxic effects of microplastics on soil biota.

To study the interactions between MPs with soil biota several studies have been performed by various groups and their impacts on the health of soil biota have been examined. In contrast to aquatic habitats, the ecotoxicological impacts of MPs on soil fauna have been the subject of very little research, and most of the studies were performed in laboratories. Coa et al. (2017) studied the impact of PS (58 μm) MPs on the health of *Eisenia Foetida* (earthworms) in dry soil. They found that MPs at concentrations greater than 1% (*w/w*) decreased earthworm growth and also increased earthworm mortality [130]. Similar findings were presented by Huerta Lwanga et al. [131] where polyethylene MPs (0.2–1.2% concentration) affected the growth and mortality of *Lumbricus Terrestris* (earthworms).

Lahive et al. [132] demonstrated that varying sizes of microplastics affect the rate of reproduction of *Enchytraeus crypticus* (soil worm) differently. It was observed that smaller-sized particles (i.e., 20 μm) affected the survival and reproduction to a greater extent as compared to the larger particles (i.e., 160 μm) which could be attributed to the ingestion of a larger number of smaller-sized MPs by soil worms. A study by Rillig et al. [133] revealed that exposure of *Lumbricus terrestris* (anecic earthworm) to the PE microplastics resulted in the transport of MPs deeper into the soil. There are potential consequences of this movement including, other soil biotas that may be exposed to MPs and microplastics may remain underground for extended periods of time. A study by Lei et al. [134] stated that the exposure of PS microplastics (size 1 μm) to *Caenorhabditis Elegans* (Roundworm) for 3 days lead to reduced survival rate and growth. It was observed that nematode survival, development, and cholinergic and GABAergic neurons were most affected by the polystyrene particles. Zhu et al. [135] also demonstrated that PVC microplastic exposure to the collembolan gut leads to a 28.8 and 16.8% inhibition of reproduction and growth of the soil organisms, respectively. Song et al. demonstrated that *Achatina Fulica* (snails) experienced different reductions in food intake and excretion after exposure to PET microfibers for 28 days, and microfibers caused significant villous damage

to the snails' digestive tract [136]. Kim and An [137] studied the effect of PE and PS MPs on *Lobella sokamensis* (soil springtail). It was reported that MPs may accumulate in the cavities created by springtail thereby inhibiting their mobility. In a study by Ju et al. [138] a decrease in the survival and reproduction rate of *Folsomia candida* (soil springtail) was observed on exposure to different concentrations of polyethylene MPs.

Yi et al. [139] demonstrated that the impacts of MPs on the soil ecosystem are also affected by the shape of the MPs. It was observed that PP fibers were more effective at inhibiting urease and alkaline phosphatase enzyme activities than the PP microsphere [139]. Wan et al. [140] depicted that MPs alter the water evaporation of soil which may lead to the drying of soil. A significant amount of soil water evaporation occurred as a consequence of MPs' presence in the soil as they created channels for water to move through. In addition, increasing the concentrations and reducing the size of MP contributed to more pronounced effects. As a consequence, the microplastic uptake damages the key functions of soil animals which are critical to biodiversity and soil health. Details of some of these studies of MP pollution on soil and soil biota are given in Table 5.

Table 5. Impact of MP pollution on properties of soil and soil biota.

Soil Biota and Properties	Polymer Types	Size	MPs Effects	References
<i>Eisenia Foetida</i> (Earthworm)	PS	58 µm	Inhibition in growth and increased mortality	[130]
<i>Lumbricus Terrestris</i> (earthworm)	PE	≥ 50 µm	Growth inhibition and mortality	[131]
<i>Lumbricus terrestris</i> (Earthworm)	PE	40.7 ± 3.8 µm	Cellular stress	[141]
<i>Eisenia fetida</i> (Earthworm)	PE	250–1000 µm	Gut damage	[142]
<i>Enchytraeus crypticus</i> (Soil worm)	PA	20 and 160 µm	Rate of reproduction was affected	[132]
<i>Lumbricus terrestris</i> (Anecic earthworm)	PE	Varying sizes	Earthworms transported MPs deeper into the soil	[133]
<i>Caenorhabditis Elegans</i> (Roundworm)	PS	1–5 µm	MPs caused reduction in body growth and low survival rate	[134]
<i>Caenorhabditis elegans</i> (Nematode)	PS	1 µm	Oxidative stress and intestinal damage	[143]
<i>Folsomia candida</i> (Collembolans)	PVC	80–250 mm	Inhibition of reproduction and growth	[135]
<i>Achatina Fulica</i> (snail)	PET	76.3 µm	Reduction in food intake and damage to digestive tract	[136]
<i>Lobella sokamensis</i> (Soil springtail)	PE and PS	0.47~1155 µm	Movement inhibition	[137]
<i>Folsomia candida</i> (Soil springtail)	PE	281 µm	Decreased survival and reproduction rate	[138]
Soil enzyme (urease and phosphatase)	Membranous PE, PP microsphere and fibrous PP	-	Inhibition of enzymatic activity	[139]
Soil property	PE	2, 5 and 10 mm	Increased water evaporation of soil leading to soil drying	[140]
<i>Triticum aestivum</i> (wheat plant)	PE	-	Inhibited the vegetative and reproductive growth	[144]

<i>Lepidium sativum</i> (cress seed)	-	< 5 mm	Delayed germination rate and growth of its root	[145]
<i>Vicia faba</i> (Broad bean)	PS	5 µm	Oxidative damage, Inhibition of plant growth, and induced genotoxicity and ecotoxicity	[146]
<i>Allium fistulosum</i> (Spring onions)	PEHD, PA, PES, PET, PP, and PS	Varying sizes	Affected plant performance	[147]
<i>Lactuca sativa</i> L. var. <i>ramose</i> Hort (Lettuce)	PS	23 µm	Lettuce's growth rate, photosynthesis, and chlorophyll content were significantly reduced by MPs	[148]
<i>Lycopersicon esculentum</i> Mill (Tomato)	PET, PP, PE	0.4–2.6 mm	MP sludge stimulated tomato plant growth but delayed the production and yield	[149]

Note(s): PS—Polystyrene, PET—Polyethylene terephthalate, PE—Polyethylene, PP—Polypropylene, PEHD—Polyethylene high density, PA—Polyamide, PVC—Polyvinyl chloride, PES—Polyester.

In terms of microplastics' effects on terrestrial plants, there is still a lack of research and knowledge. Qi et al. [144] demonstrated that polyethylene MP films (1% *w/w*) negatively inhibited the reproductive and vegetative growth of the *Triticum aestivum* (wheat plant) in dry soil. After 8 and 24 h of MPs exposure, Bosker et al. [145] witnessed that *Lepidium sativum* (cress seed) capsules accumulated MPs and resulted in a delayed germination rate and growth of its root, respectively. Jiang et al. [146] stated that the accumulation of a large number of polystyrene MPs in the root tips of the *Vicia faba* plant could significantly result in oxidative damage, inhibit plant growth, and induced genotoxicity and ecotoxicity. Likewise, de Souza Machado et al. [147] explored the performance of *Allium fistulosum* (spring onions) when exposed to different MPs (0.2% *w/w*). The results indicated that MP exposure could alter plant biomass, elemental composition, and root traits, while the actual effects were different depending on the particle type. Gao et al. [148] demonstrated that PE microplastics demonstrated negative impacts on the growth, photosynthesis, and chlorophyll content of lettuce. Hernández-Arenas et al. [149] studied the effect of sludge containing PP, PET, and PE MPs on *Lycopersicon esculentum* Mill (Tomato). The results revealed that the growth of tomato plants in soils containing MPs was accelerated, while fruit production was delayed. As plants are an important part of the terrestrial ecosystem and MPs are prevalent, future research is needed to examine several kinds of MP particles, different soil conditions, and a wider range of plant species to investigate the potential consequences of MP pollution [150].

5. Protocols and Existing Infrastructure in Place for Controlling MP Release

Keeping in view the surmounting issue of microplastic accumulation in nature, international organisations and governments are endorsing concepts such as the circular economy and the six 'R's—Reduce (raw material usage), Redesign (designing reusable and recyclable products), Remove (avoid usage of single-use plastics), Reuse (refurbishment of old products), Recycle (repeated usage of products) and Recover (regeneration and resynthesis) for sustainable growth and development [150]. Since polymers and plastics have become an indispensable part of the global economy, microplastic waste generation and release into water bodies need to be closely monitored and there is a need to undertake effective measures to minimize their detrimental impacts on the ecosystem. There is an immediate need to take action based on available evidence of MP waste while

also taking precautionary approaches towards MP extraction to remove tangible future threats. Governments, organisations, industries, and the public need to work together in order to overcome these issues. Organisations such as UNEP, IMO, ICO and FAO are currently working to combat the problem of MP accumulation in aquatic bodies. An example of such action is the Canadian government which has banned the use of MPs in cosmetic products since these pellets sink to the bottom of oceans and rivers and accumulate [63]. Even after the ban, the pre-existing MPs pose a serious and imminent threat to the ecosystem since their degradation is touted to take many years.

Additionally, financial tools such as fines, taxes, fees, subsidies, deposit-refund schemes, and incentives have also proven to be effective in promoting the recycling of products which leads to a reduction in dumping and subsequent accumulation of MP pollutants in aquatic ecosystems. For instance, Ecuador requires extensive use of PET bottles for supplying clean and potable drinking water. Therefore, a bottle deposit scheme of US\$ 0.02 per bottle was introduced in the country in 2011 to motivate people to deposit bottles, thereby making the collection of plastic bottles easier. It was noted that PET bottle recycling increased to 80% in 2012 from 30% in 2011, with 1.13 million bottles out of 1.40 million bottles being recycled. Measures such as the imposition of port fees (Port of Rotterdam, Netherlands), product bans (Canada), tourist fees (Galapagos Archipelago, Ecuador), and littering fines (California, USA) that involve diligence by local and government action have also proven to be successful in tackling plastic accumulation and promotion of recycling across the world. The Indian government has enforced a country-wide ban on the use, import, manufacture, stocking, distribution, and selling of single-use plastics. Fines have also been imposed in case of failing to comply with the provisions of the ban. These steps have been taken to reduce India's contribution to global plastic waste stockpiles that have currently reached epidemic proportions.

The reduction and removal of MPs in marine ecosystems is a focal point of goal no. 14—Life Under Water (focus on marine ecosystem health) of the Sustainable Development Goals put forth by the United Nations. Particularly, it aims to improve aquatic ecosystem health by reducing sources of marine pollution, especially microplastic release by 2025. In the year 2019, member countries of the G20 summit released a statement termed the *Osaka Blue Ocean Vision* wherein the Ministry of Foreign Affairs of Japan launched the MARINE initiative that aims to reduce MP pollution of oceans and assist developing countries in plastic waste management by using the life cycle approach and innovative solutions [151]. The US Marine Debris Program developed the NOAA marine debris program that aims at the removal of MP from marine bodies without the requirement of any sophisticated instruments [152]. Collection and analysis systems such as the Manta net, Albatross device, and PLEX (PLastic EXplorer) instrument have also been involved for water remediation purposes [153–155].

In recent years, attempts have been made to find materials that can serve as alternatives to commodity plastics. The use of compostable and biodegradable plastics and paper is being promoted for use as packaging material while natural products such as walnut shell powder and mineral powder have replaced micro-sized plastic pellets in cosmetic products. Polylactic acid (PLA), and starch, sugarcane, and mushroom-based biomaterials are being promoted for use in various applications as substitutes for PE and PS due to their benign nature and biodegradability. Although these substitutes are biocompatible and degrade more easily than microplastics, their use is not always economical and environment-friendly since they raise production costs and require more exploitation of natural resources. Hence, there is still scope for improvement in the adoption of these alternative materials. Efforts are being made at multiple levels to overcome the problem of MP pollution but despite all these efforts, the influx of MPs into water bodies, soil, and human, plant, and animal physiology is rising at an astonishing rate. Hence, there is a need for strict rules and stringent action from the policymakers and the public to safeguard the future.

6. Conclusions

Global plastic production and usage are expected to increase in the coming years, which may lead to increased microplastic pollution in the environment. In the current research, it has been determined that microplastics have diverse sources, and the way in which they occur, transport, and evolve depends on a wide range of natural conditions as well as their physicochemical characteristics (e.g., size, crystallinity, shape, density, etc.). The quantification and identification of microplastics are essential prerequisites to ensure their efficacious removal. The combined use of visual analysis and spectroscopic techniques is especially helpful in the detection of microplastics in aquatic systems but they possess certain drawbacks. Thus, keeping in view the extensive amount of MP waste generated globally, efforts need to be made to develop better techniques for MP detection and identification to minimize misidentification. To extract MP waste from wastewater, many treatment methods have been developed to facilitate rapid and efficient removal. Conventional and advanced methods such as CFS, membrane filtration, adsorption, and biological degradation are often used in conjunction and have the potential to exhibit removal efficiencies as high as >99%. Although, issues such as membrane fouling, adsorption site blockage, particle selectivity, and lack of reusability plague these treatment methods, and the effect of size, shape, and polymer type on the removal efficacy is not very clear. These limitations call for further research into these methods to improve the MP removal capability of these technologies. Most of the studies have examined the toxicology of microplastics in the marine environment, but their impact on soil biota and human health has not been fully explored. Moreover, further research should be conducted on how microplastics affect human cells in vivo. Since microplastic waste accumulation has reached epidemic proportions at the global scale, certain protocols and infrastructure has been enforced to reduce further generation MPs and attempts are being made to combat MP waste accumulation at the international, national, and local levels.

Author Contributions: The presented work was carried out by the contribution of all the authors. Conceptualisation: B.P. and J.P.; Writing—original draft: B.P. and J.P.; Visualisation: P.S., R.K. and A.K.; Review and editing: P.S., R.K., T.K.T., S.K. and A.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Delhi Technological University, DTU (India) for providing the necessary infrastructure to carry out the research. Author BP would also like to thank CSIR, India for providing SRF fellowship.

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

References

1. Xu, S.; Ma, J.; Ji, R.; Pan, K.; Miao, A.J. Microplastics in Aquatic Environments: Occurrence, Accumulation, and Biological Effects. *Sci. Total Environ.* **2020**, *703*, 134699. <https://doi.org/10.1016/j.scitotenv.2019.134699>.
2. Rochman, C.M.; Browne, M.A.; Halpern, B.S.; Hentschel, B.T.; Hoh, E.; Karapanagioti, H.K.; Rios-Mendoza, L.M.; Takada, H.; Teh, S.; Thompson, R.C. Classify Plastic Waste as Hazardous. *Nature* **2013**, *494*, 169–171. <https://doi.org/10.1038/494169a>.
3. Faunce, T.; Kolodziejczyk, B. Nanowaste: Need for Disposal and Recycling Standards. *G20 Insights* **2017**, *148*, 202–213.
4. Geyer, R.; Jambeck, J.R.; Law, K.L. Production, Use, and Fate of All Plastics Ever Made. *Sci. Adv.* **2017**, *3*, e1700782.
5. Gall, S.C.; Thompson, R.C. The Impact of Debris on Marine Life. *Mar. Pollut. Bull.* **2015**, *92*, 170–179. <https://doi.org/10.1016/j.marpolbul.2014.12.041>.
6. Thompson, R.C.; Olsen, Y.; Mitchell, R.P.; Davis, A.; Rowland, S.J.; John, A.W.G.; McGonigle, D.; Russell, A.E. Lost at Sea: Where Is All the Plastic? *Science* **2004**, *304*, 838. <https://doi.org/10.1126/science.1094559>.
7. Alsabri, A.; Tahir, F.; Al-Ghamdi, S.G. Environmental Impacts of Polypropylene (PP) Production and Prospects of Its Recycling in the GCC Region. *Mater. Today Proc.* **2022**, *56*, 2245–2251. <https://doi.org/10.1016/j.matpr.2021.11.574>.

8. Anderson, J.C.; Park, B.J.; Palace, V.P. Microplastics in Aquatic Environments: Implications for Canadian Ecosystems. *Environ. Pollut.* **2016**, *218*, 269–280. <https://doi.org/10.1016/j.envpol.2016.06.074>.
9. Mangaraj, S.; Goswami, T.K.; Mahajan, P.V. Applications of Plastic Films for Modified Atmosphere Packaging of Fruits and Vegetables: A Review. *Food Eng. Rev.* **2009**, *1*, 133–158. <https://doi.org/10.1007/s12393-009-9007-3>.
10. Bach, C.; Dauchy, X.; Severin, I.; Munoz, J.F.; Etienne, S.; Chagnon, M.C. Effect of Temperature on the Release of Intentionally and Non-Intentionally Added Substances from Polyethylene Terephthalate (PET) Bottles into Water: Chemical Analysis and Potential Toxicity. *Food Chem.* **2013**, *139*, 672–680. <https://doi.org/10.1016/j.foodchem.2013.01.046>.
11. Hwang, J.; Choi, D.; Han, S.; Choi, J.; Hong, J. An Assessment of the Toxicity of Polypropylene Microplastics in Human Derived Cells. *Sci. Total Environ.* **2019**, *684*, 657–669. <https://doi.org/10.1016/j.scitotenv.2019.05.071>.
12. Tawfik, M.S.; Huyghebaert, A. Polystyrene Cups and Containers: Styrene Migration. *Food Addit. Contam.* **1998**, *15*, 592–599. <https://doi.org/10.1080/02652039809374686>.
13. Nakamura, S.; Nakajima, K.; Yoshizawa, Y.; Matsubae-Yokoyama, K.; Nagasaka, T. Analyzing Polyvinyl Chloride in Japan with The Waste Input-Output Material Flow Analysis Model. *J. Ind. Ecol.* **2009**, *13*, 706–717. <https://doi.org/10.1111/j.1530-9290.2009.00153.x>.
14. Das, A.; Mahanwar, P. A Brief Discussion on Advances in Polyurethane Applications. *Adv. Ind. Eng. Polym. Res.* **2020**, *3*, 93–101. <https://doi.org/10.1016/j.aiepr.2020.07.002>.
15. Hearle, J.W.S. Textile Fibers: A Comparative Overview. *Encycl. Mater. Sci. Technol.* **2001**, *2001*, 9100–9116. <https://doi.org/10.1016/b0-08-043152-6/01643-0>.
16. Barnes, D.K.A.; Galgani, F.; Thompson, R.C.; Barlaz, M. Accumulation and Fragmentation of Plastic Debris in Global Environments. *Philos. Trans. R. Soc. B Biol. Sci.* **2009**, *364*, 1985–1998. <https://doi.org/10.1098/rstb.2008.0205>.
17. Andrady, A.L. The Plastic in Microplastics: A Review. *Mar. Pollut. Bull.* **2017**, *119*, 12–22. <https://doi.org/10.1016/j.marpolbul.2017.01.082>.
18. Lee, K.W.; Shim, W.J.; Kwon, O.Y.; Kang, J.H. Size-Dependent Effects of Micro Polystyrene Particles in the Marine Copepod *Tigriopus Japonicus*. *Environ. Sci. Technol.* **2013**, *47*, 11278–11283. <https://doi.org/10.1021/es401932b>.
19. Au, S.Y.; Bruce, T.F.; Bridges, W.C.; Klaine, S.J. Responses of *Hyaella Azteca* to Acute and Chronic Microplastic Exposures. *Environ. Toxicol. Chem.* **2015**, *34*, 2564–2572. <https://doi.org/10.1002/etc.3093>.
20. Llorca, M.; Schirizzi, G.; Martínez, M.; Barceló, D.; Farré, M. Adsorption of Perfluoroalkyl Substances on Microplastics under Environmental Conditions. *Environ. Pollut.* **2018**, *235*, 680–691. <https://doi.org/10.1016/j.envpol.2017.12.075>.
21. Chen, D.R.; Bei, J.Z.; Wang, S.G. Polycaprolactone Microparticles and Their Biodegradation. *Polym. Degrad. Stab.* **2000**, *67*, 455–459. [https://doi.org/10.1016/S0141-3910\(99\)00145-7](https://doi.org/10.1016/S0141-3910(99)00145-7).
22. Stolte, A.; Forster, S.; Gerdts, G.; Schubert, H. Microplastic Concentrations in Beach Sediments along the German Baltic Coast. *Mar. Pollut. Bull.* **2015**, *99*, 216–229. <https://doi.org/10.1016/j.marpolbul.2015.07.022>.
23. Priot, D.; Boucher, J. *Primary Microplastics in the Oceans: A Global Evaluation of Sources*; IUCN: Gland, Switzerland, 2017; ISBN 0231137079.
24. An, L.; Liu, Q.; Deng, Y.; Wu, W.; Gao, Y.; Ling, W. Sources of Microplastic in the Environment. *Handb. Environ. Chem.* **2020**, *95*, 143–159. https://doi.org/10.1007/698_2020_449.
25. Karlsson, T.M.; Arneborg, L.; Broström, G.; Almroth, B.C.; Gipperth, L.; Hassellöv, M. The Unaccountability Case of Plastic Pellet Pollution. *Mar. Pollut. Bull.* **2018**, *129*, 52–60. <https://doi.org/10.1016/j.marpolbul.2018.01.041>.
26. Napper, I.E.; Bakir, A.; Rowland, S.J.; Thompson, R.C. Characterisation, Quantity and Sorptive Properties of Microplastics Extracted from Cosmetics. *Mar. Pollut. Bull.* **2015**, *99*, 178–185. <https://doi.org/10.1016/j.marpolbul.2015.07.029>.
27. Atugoda, T.; Vithanage, M.; Wijesekara, H.; Bolan, N.; Sarmah, A.K.; Bank, M.S.; You, S.; Ok, Y.S. Interactions between Microplastics, Pharmaceuticals and Personal Care Products: Implications for Vector Transport. *Environ. Int.* **2021**, *149*, 106367. <https://doi.org/10.1016/j.envint.2020.106367>.
28. Gouin, T.; Avalos, J.; Brunning, I.; Brzuska, K.; De Graaf, J.; Kaumanns, J.; Koning, T.; Meyberg, M.; Rettinger, K.; Schlatter, H.; et al. Use of Micro-Plastic Beads in Cosmetic Products in Europe and Their Estimated Emissions to the North Sea Environment. *SOFW J.* **2015**, *141*, 40–46.
29. Nagaraj, V.; Skillman, L.; Li, D.; Ho, G. Review—Bacteria and Their Extracellular Polymeric Substances Causing Biofouling on Seawater Reverse Osmosis Desalination Membranes. *J. Environ. Manage.* **2018**, *223*, 586–599. <https://doi.org/10.1016/j.jenvman.2018.05.088>.
30. Montarsolo, A.; Mossotti, R.; Patrucco, A.; Caringella, R.; Zoccola, M.; Pozzo, P.D.; Tonin, C. Study on the Microplastics Release from Fishing Nets. *Eur. Phys. J. Plus* **2018**, *133*, 494. <https://doi.org/10.1140/epjp/i2018-12415-1>.
31. Napper, I.E.; Thompson, R.C. Release of Synthetic Microplastic Plastic Fibres from Domestic Washing Machines: Effects of Fabric Type and Washing Conditions. *Mar. Pollut. Bull.* **2016**, *112*, 39–45. <https://doi.org/10.1016/j.marpolbul.2016.09.025>.
32. Jan Kole, P.; Löhr, A.J.; Van Belleghem, F.G.A.J.; Ragas, A.M.J. Wear and Tear of Tyres: A Stealthy Source of Microplastics in the Environment. *Int. J. Environ. Res. Public Health* **2017**, *14*, 1265. <https://doi.org/10.3390/ijerph14101265>.
33. Piret, J.; Boivin, G. Pandemics Throughout History. *Front. Microbiol.* **2021**, *11*, 631736. <https://doi.org/10.3389/fmicb.2020.631736>.

34. Shams, M.; Alam, I.; Mahbub, M.S. Plastic Pollution During COVID-19: Plastic Waste Directives and Its Long-Term Impact on The Environment. *Environ. Adv.* **2021**, *5*, 100119. <https://doi.org/10.1016/j.envadv.2021.100119>.
35. Prata, J.C.; Silva, A.L.P.; Walker, T.R.; Duarte, A.C.; Rocha-Santos, T. COVID-19 Pandemic Repercussions on the Use and Management of Plastics. *Environ. Sci. Technol.* **2020**, *54*, 7760–7765. <https://doi.org/10.1021/acs.est.0c02178>.
36. Benson, N.U.; Bassey, D.E.; Palanisami, T. COVID Pollution: Impact of COVID-19 Pandemic on Global Plastic Waste Footprint. *Heliyon* **2021**, *7*, e06343. <https://doi.org/10.1016/j.heliyon.2021.e06343>.
37. Peng, Y.; Wu, P.; Schartup, A.T.; Zhang, Y. Plastic Waste Release Caused by COVID-19 and Its Fate in the Global Ocean. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2111530118. <https://doi.org/10.1073/pnas.2111530118>.
38. Morgana, S.; Casentini, B.; Amalfitano, S. Uncovering the Release of Micro/Nanoplastics from Disposable Face Masks at Times of COVID-19. *J. Hazard. Mater.* **2021**, *419*, 126507. <https://doi.org/10.1016/j.jhazmat.2021.126507>.
39. Okuku, E.; Kiteresi, L.; Owato, G.; Otieno, K.; Mwalugha, C.; Mbuche, M.; Gwada, B.; Nelson, A.; Chepkemboi, P.; Achieng, Q.; et al. The Impacts of COVID-19 Pandemic on Marine Litter Pollution along the Kenyan Coast: A Synthesis after 100 Days Following the First Reported Case in Kenya. *Mar. Pollut. Bull.* **2021**, *162*, 111840. <https://doi.org/10.1016/j.marpolbul.2020.111840>.
40. Haque, F.; Fan, C. Prospect of Microplastic Pollution Control under the “New Normal” Concept beyond COVID-19 Pandemic. *J. Clean. Prod.* **2022**, *367*, 133027. <https://doi.org/10.1016/j.jclepro.2022.133027>.
41. Poerio, T.; Piacentini, E.; Mazzei, R. Membrane Processes for Microplastic Removal. *Molecules* **2019**, *24*, 4148. <https://doi.org/10.3390/molecules24224148>.
42. Wang, Q.; Hernández-Crespo, C.; Du, B.; Van Hulle, S.W.H.; Rousseau, D.P.L. Fate and Removal of Microplastics in Unplanted Lab-Scale Vertical Flow Constructed Wetlands. *Sci. Total Environ.* **2021**, *778*, 146152. <https://doi.org/10.1016/j.scitotenv.2021.146152>.
43. Lapointe, M.; Farner, J.M.; Hernandez, L.M.; Tufenkji, N. Understanding and Improving Microplastic Removal during Water Treatment: Impact of Coagulation and Flocculation. *Environ. Sci. Technol.* **2020**, *54*, 8719–8727. <https://doi.org/10.1021/acs.est.0c00712>.
44. Joo, S.H.; Liang, Y.; Kim, M.; Byun, J.; Choi, H. Microplastics with Adsorbed Contaminants: Mechanisms and Treatment. *Environ. Challenges* **2021**, *3*, 100042. <https://doi.org/10.1016/j.envc.2021.100042>.
45. Lee, Q.Y.; Li, H. Photocatalytic Degradation of Plastic Waste: A Mini Review. *Micromachines* **2021**, *12*, 907. <https://doi.org/10.3390/mi12080907>.
46. Mohanan, N.; Montazer, Z.; Sharma, P.K.; Levin, D.B. Microbial and Enzymatic Degradation of Synthetic Plastics. *Front. Microbiol.* **2020**, *11*, 580709. <https://doi.org/10.3389/fmicb.2020.580709>.
47. Dey, T.K.; Uddin, M.E.; Jamal, M. Detection and Removal of Microplastics in Wastewater: Evolution and Impact. *Environ. Sci. Pollut. Res.* **2021**, *28*, 16925–16947. <https://doi.org/10.1007/s11356-021-12943-5>.
48. Kim, K.T.; Park, S. Enhancing Microplastics Removal from Wastewater Using Electro-Coagulation and Granule-Activated Carbon with Thermal Regeneration. *Processes* **2021**, *9*, 617. <https://doi.org/10.3390/pr9040617>.
49. Arpia, A.A.; Chen, W.H.; Ubando, A.T.; Naqvi, S.R.; Culaba, A.B. Microplastic Degradation as a Sustainable Concurrent Approach for Producing Biofuel and Obliterating Hazardous Environmental Effects: A State-of-the-Art Review. *J. Hazard. Mater.* **2021**, *418*, 126381. <https://doi.org/10.1016/j.jhazmat.2021.126381>.
50. Ma, B.; Xue, W.; Hu, C.; Liu, H.; Qu, J.; Li, L. Characteristics of Microplastic Removal via Coagulation and Ultrafiltration during Drinking Water Treatment. *Chem. Eng. J.* **2019**, *359*, 159–167. <https://doi.org/10.1016/j.cej.2018.11.155>.
51. Barbier, J.S.; Dris, R.; Lecarpentier, C.; Raymond, V.; Delabre, K.; Thibert, S.; Tassin, B.; Gasperi, J. Microplastic Occurrence after Conventional and Nanofiltration Processes at Drinking Water Treatment Plants: Preliminary Results. *Front. Water* **2022**, *2022*, 4. <https://doi.org/10.3389/frwa.2022.886703>.
52. Yaranal, N.A.; Subbiah, S.; Mohanty, K. Identification, Extraction of Microplastics from Edible Salts and Its Removal from Contaminated Seawater. *Environ. Technol. Innov.* **2021**, *21*, 101253. <https://doi.org/10.1016/j.eti.2020.101253>.
53. Hidayatullah, H.; Lee, T.G. A Study on Characteristics of Microplastic in Wastewater of South Korea: Identification, Quantification, and Fate of Microplastics during Treatment Process. *Mar. Pollut. Bull.* **2019**, *146*, 696–702. <https://doi.org/10.1016/j.marpolbul.2019.06.071>.
54. Tadsuwan, K.; Babel, S. Microplastic Abundance and Removal via an Ultrafiltration System Coupled to a Conventional Municipal Wastewater Treatment Plant in Thailand. *J. Environ. Chem. Eng.* **2022**, *10*, 107142. <https://doi.org/10.1016/j.jece.2022.107142>.
55. Li, L.; Xu, G.; Yu, H.; Xing, J. Dynamic Membrane for Micro-Particle Removal in Wastewater Treatment: Performance and Influencing Factors. *Sci. Total Environ.* **2018**, *627*, 332–340. <https://doi.org/10.1016/j.scitotenv.2018.01.239>.
56. Xing, J.; Wang, H.; Cheng, X.; Tang, X.; Luo, X.; Wang, J.; Wang, T.; Li, G.; Liang, H. Application of Low-Dosage UV/Chlorine Pre-Oxidation for Mitigating Ultrafiltration (UF) Membrane Fouling in Natural Surface Water Treatment. *Chem. Eng. J.* **2018**, *344*, 62–70. <https://doi.org/10.1016/j.cej.2018.03.052>.
57. Ziajahromi, S.; Neale, P.A.; Rintoul, L.; Leusch, F.D.L. Wastewater Treatment Plants as a Pathway for Microplastics: Development of a New Approach to Sample Wastewater-Based Microplastics. *Water Res.* **2017**, *112*, 93–99. <https://doi.org/10.1016/j.watres.2017.01.042>.

58. Ben Hassan, I.; Ennouri, M.; Lafforgue, C.; Schmitz, P.; Ayadi, A. Experimental Study of Membrane Fouling during Crossflow Microfiltration of Yeast and Bacteria Suspensions: Towards an Analysis at the Microscopic Level. *Membranes* **2013**, *3*, 44–68. <https://doi.org/10.3390/membranes3020044>.
59. Tang, Y.; Zhang, S.; Su, Y.; Wu, D.; Zhao, Y.; Xie, B. Removal of Microplastics from Aqueous Solutions by Magnetic Carbon Nanotubes. *Chem. Eng. J.* **2021**, *406*, 126804. <https://doi.org/10.1016/j.cej.2020.126804>.
60. Siipola, V.; Pflugmacher, S.; Romar, H.; Wendling, L.; Koukkari, P. Low-Cost Biochar Adsorbents for Water Purification Including Microplastics Removal. *Appl. Sci.* **2020**, *10*, 788. <https://doi.org/10.3390/app10030788>.
61. Sun, C.; Wang, Z.; Chen, L.; Li, F. Fabrication of Robust and Compressive Chitin and Graphene Oxide Sponges for Removal of Microplastics with Different Functional Groups. *Chem. Eng. J.* **2020**, *393*, 124796. <https://doi.org/10.1016/j.cej.2020.124796>.
62. Zhang, Y.; Zhao, J.; Liu, Z.; Tian, S.; Lu, J.; Mu, R.; Yuan, H. Coagulation Removal of Microplastics from Wastewater by Magnetic Magnesium Hydroxide and PAM. *J. Water Process Eng.* **2021**, *43*, 102250. <https://doi.org/10.1016/j.jwpe.2021.102250>.
63. Kumar, A.; Upadhyay, P.; Prajapati, S.K. Impact of microplastics on riverine greenhouse gas emissions: A view point. *Environ. Sci. Pollut. Res.* **2022**, *2022*, 1–4. <https://doi.org/10.1007/s11356-022-23929-2>.
64. Danso, D.; Chow, J.; Streita, W.R. Plastics: Environmental and Biotechnological Perspectives on Microbial Degradation. *Appl. Environ. Microbiol.* **2019**, *85*, AEM-01095. <https://doi.org/10.1128/AEM.01095-19>.
65. Peng, B.Y.; Chen, Z.; Chen, J.; Yu, H.; Zhou, X.; Criddle, C.S.; Wu, W.M.; Zhang, Y. Biodegradation of Polyvinyl Chloride (PVC) in *Tenebrio Molitor* (Coleoptera: Tenebrionidae) Larvae. *Environ. Int.* **2020**, *145*, 106106. <https://doi.org/10.1016/j.envint.2020.106106>.
66. Huang, S.; Peng, C.; Wang, Z.; Xiong, X.; Bi, Y.; Liu, Y.; Li, D. Spatiotemporal Distribution of Microplastics in Surface Water, Biofilms, and Sediments in the World's Largest Drinking Water Diversion Project. *Sci. Total Environ.* **2021**, *789*, 148001. <https://doi.org/10.1016/j.scitotenv.2021.148001>.
67. Jimenez-Cárdenas, V.; Luna-Acosta, A.; Gómez-Méndez, L.D. Differential Presence of Microplastics and Mesoplastics in Coral Reef and Mangrove Fishes in Isla Grande, Colombia. *Microplastics* **2022**, *1*, 477–493. <https://doi.org/10.3390/microplastics1030034>.
68. Prokić, M.D.; Radovanović, T.B.; Gavrić, J.P.; Faggio, C. Ecotoxicological Effects of Microplastics: Examination of Biomarkers, Current State and Future Perspectives. *TrAC Trends Anal. Chem.* **2019**, *111*, 37–46. <https://doi.org/10.1016/j.trac.2018.12.001>.
69. *Microplastics in Drinking-Water*; World Health Organization: Geneva, Switzerland, 2019;
70. Cho, Y.; Shim, W.J.; Jang, M.; Han, G.M.; Hong, S.H. Abundance and Characteristics of Microplastics in Market Bivalves from South Korea. *Environ. Pollut.* **2019**, *245*, 1107–1116. <https://doi.org/10.1016/j.envpol.2018.11.091>.
71. Karami, A.; Golieskardi, A.; Choo, C.K.; Larat, V.; Karbalaee, S.; Salamatinia, B. Microplastic and Mesoplastic Contamination in Canned Sardines and Sprats. *Sci. Total Environ.* **2018**, *612*, 1380–1386. <https://doi.org/10.1016/j.scitotenv.2017.09.005>.
72. Ghosh, G.C.; Akter, S.M.; Islam, R.M.; Habib, A.; Chakraborty, T.K.; Zaman, S.; Kabir, A.H.M.E.; Shipin, O.V.; Wahid, M.A. Microplastics Contamination in Commercial Marine Fish from the Bay of Bengal. *Reg. Stud. Mar. Sci.* **2021**, *44*, 101728. <https://doi.org/10.1016/j.rsma.2021.101728>.
73. Halstead, J.E.; Smith, J.A.; Carter, E.A.; Lay, P.A.; Johnston, E.L. Assessment Tools for Microplastics and Natural Fibres Ingested by Fish in an Urbanised Estuary. *Environ. Pollut.* **2018**, *234*, 552–561. <https://doi.org/10.1016/j.envpol.2017.11.085>.
74. Lusher, A.L.; McHugh, M.; Thompson, R.C. Occurrence of Microplastics in the Gastrointestinal Tract of Pelagic and Demersal Fish from the English Channel. *Mar. Pollut. Bull.* **2013**, *67*, 94–99. <https://doi.org/10.1016/j.marpolbul.2012.11.028>.
75. Tanaka, K.; Takada, H. Microplastic Fragments and Microbeads in Digestive Tracts of Planktivorous Fish from Urban Coastal Waters. *Sci. Rep.* **2016**, *6*, 1–8. <https://doi.org/10.1038/srep34351>.
76. Daniel, D.B.; Ashraf, P.M.; Thomas, S.N. Abundance, Characteristics and Seasonal Variation of Microplastics in Indian White Shrimps (*Fenneropenaeus Indicus*) from Coastal Waters off Cochin, Kerala, India. *Sci. Total Environ.* **2020**, *737*, 139839. <https://doi.org/10.1016/j.scitotenv.2020.139839>.
77. Li, J.; Qu, X.; Su, L.; Zhang, W.; Yang, D.; Kolandhasamy, P.; Li, D.; Shi, H. Microplastics in Mussels along the Coastal Waters of China. *Environ. Pollut.* **2016**, *214*, 177–184. <https://doi.org/10.1016/j.envpol.2016.04.012>.
78. Wu, R.T.; Cai, Y.F.; Chen, Y.X.; Yang, Y.W.; Xing, S.C.; Liao, X. Di Occurrence of Microplastic in Livestock and Poultry Manure in South China. *Environ. Pollut.* **2021**, *277*, 116790. <https://doi.org/10.1016/j.envpol.2021.116790>.
79. Huerta Lwanga, E.; Mendoza Vega, J.; Ku Quej, V.; de los Chi, J.A.; del Cid Sanchez, L.; Chi, C.; Escalona Segura, G.; Gertsen, H.; Salánki, T.; van der Ploeg, M.; et al. Field Evidence for Transfer of Plastic Debris along a Terrestrial Food Chain. *Sci. Rep.* **2017**, *7*, 1–7. <https://doi.org/10.1038/s41598-017-14588-2>.
80. Yang, D.; Shi, H.; Li, L.; Li, J.; Jabeen, K.; Kolandhasamy, P. Microplastic Pollution in Table Salts from China. *Environ. Sci. Technol.* **2015**, *49*, 13622–13627. <https://doi.org/10.1021/acs.est.5b03163>.
81. Kim, J.S.; Lee, H.J.; Kim, S.K.; Kim, H.J. Global Pattern of Microplastics (MPs) in Commercial Food-Grade Salts: Sea Salt as an Indicator of Seawater MP Pollution. *Environ. Sci. Technol.* **2018**, *52*, 12819–12828. <https://doi.org/10.1021/acs.est.8b04180>.
82. Hernandez, L.M.; Xu, E.G.; Larsson, H.C.E.; Tahara, R.; Maisuria, V.B.; Tufenkji, N. Plastic Teabags Release Billions of Micro-particles and Nanoparticles into Tea. *Environ. Sci. Technol.* **2019**, *53*, 12300–12310. <https://doi.org/10.1021/acs.est.9b02540>.

83. Schymanski, D.; Goldbeck, C.; Humpf, H.U.; Fürst, P. Analysis of Microplastics in Water by Micro-Raman Spectroscopy: Release of Plastic Particles from Different Packaging into Mineral Water. *Water Res.* **2018**, *129*, 154–162. <https://doi.org/10.1016/j.watres.2017.11.011>.
84. Kutralam-Muniasamy, G.; Pérez-Guevara, F.; Elizalde-Martínez, I.; Shruti, V.C. Branded Milks—Are They Immune from Microplastics Contamination? *Sci. Total Environ.* **2020**, *714*, 136823. <https://doi.org/10.1016/j.scitotenv.2020.136823>.
85. Mintenig, S.M.; Löder, M.G.J.; Primpke, S.; Gerdts, G. Low Numbers of Microplastics Detected in Drinking Water from Ground Water Sources. *Sci. Total Environ.* **2019**, *648*, 631–635. <https://doi.org/10.1016/j.scitotenv.2018.08.178>.
86. Kosuth, M.; Mason, S.A.; Wattenberg, E.V. Anthropogenic Contamination of Tap Water, Beer, and Sea Salt. *PLoS One* **2018**, *13*, 1–18. <https://doi.org/10.1371/journal.pone.0194970>.
87. Diaz-Basantes, M.F.; Conesa, J.A.; Fullana, A. Microplastics in Honey, Beer, Milk and Refreshments in Ecuador as Emerging Contaminants. *Sustainability* **2020**, *12*, 5514. <https://doi.org/10.3390/SU12145514>.
88. Liebezeit, G.; Liebezeit, E. Non-Pollen Particulates in Honey and Sugar. *Food Addit. Contam. Part A* **2013**, *30*, 2136–2140. <https://doi.org/10.1080/19440049.2013.843025>.
89. Amato-Lourenço, L.F.; Carvalho-Oliveira, R.; Júnior, G.R.; dos Santos Galvão, L.; Ando, R.A.; Mauad, T. Presence of Airborne Microplastics in Human Lung Tissue. *J. Hazard. Mater.* **2021**, *416*, 126124. <https://doi.org/10.1016/j.jhazmat.2021.126124>.
90. Prata, J.C.; da Costa, J.P.; Lopes, I.; Duarte, A.C.; Rocha-Santos, T. Environmental Exposure to Microplastics: An Overview on Possible Human Health Effects. *Sci. Total Environ.* **2020**, *702*, 134455. <https://doi.org/10.1016/j.scitotenv.2019.134455>.
91. Pironti, C.; Ricciardi, M.; Motta, O.; Miele, Y.; Proto, A.; Montano, L. Microplastics in the Environment: Intake through the Food Web, Human Exposure and Toxicological Effects. *Toxics* **2021**, *9*, 1–29. <https://doi.org/10.3390/toxics9090224>.
92. Prata, J.C. Airborne Microplastics: Consequences to Human Health? *Environ. Pollut.* **2018**, *234*, 115–126. <https://doi.org/10.1016/j.envpol.2017.11.043>.
93. Li, L.; Zhao, X.; Li, Z.; Song, K. COVID-19: Performance Study of Microplastic Inhalation Risk Posed by Wearing Masks. *J. Hazard. Mater.* **2021**, *411*, 1–9. <https://doi.org/10.1016/j.jhazmat.2020.124955>.
94. Salvioni, L.; Morelli, L.; Ochoa, E.; Labra, M.; Fiandra, L.; Palugan, L.; Prosperi, D.; Colombo, M. The Emerging Role of Nanotechnology in Skincare. *Adv. Colloid Interface Sci.* **2021**, *293*, 102437. <https://doi.org/10.1016/j.cis.2021.102437>.
95. Revel, M.; Châtel, A.; Mouneyrac, C. Micro(Nano)Plastics: A Threat to Human Health? *Curr. Opin. Environ. Sci. Heal.* **2018**, *1*, 17–23. <https://doi.org/10.1016/j.coesh.2017.10.003>.
96. Schirinzi, G.F.; Pérez-Pomeda, I.; Sanchís, J.; Rossini, C.; Farré, M.; Barceló, D. Cytotoxic Effects of Commonly Used Nanomaterials and Microplastics on Cerebral and Epithelial Human Cells. *Environ. Res.* **2017**, *159*, 579–587. <https://doi.org/10.1016/j.envres.2017.08.043>.
97. Wu, B.; Wu, X.; Liu, S.; Wang, Z.; Chen, L. Size-Dependent Effects of Polystyrene Microplastics on Cytotoxicity and Efflux Pump Inhibition in Human Caco-2 cells. *Chemosphere* **2019**, *221*, 333–341. <https://doi.org/10.1016/j.chemosphere.2019.01.056>.
98. Stock, V.; Böhmert, L.; Lisicki, E.; Block, R.; Cara-Carmona, J.; Pack, L.K.; Selb, R.; Lichtenstein, D.; Voss, L.; Henderson, C.J.; et al. Uptake and Effects of Orally Ingested Polystyrene Microplastic Particles in Vitro and in Vivo. *Arch. Toxicol.* **2019**, *93*, 1817–1833. <https://doi.org/10.1007/s00204-019-02478-7>.
99. Hesler, M.; Aengenheister, L.; Ellinger, B.; Drexel, R.; Straskraba, S.; Jost, C.; Wagner, S.; Meier, F.; von Briesen, H.; Büchel, C.; et al. Multi-Endpoint Toxicological Assessment of Polystyrene Nano- and Microparticles in Different Biological Models in Vitro. *Toxicol. Vitro* **2019**, *61*, 104610. <https://doi.org/10.1016/j.tiv.2019.104610>.
100. Xu, H.; Hoet, P.H.M.; Nemery, B. In Vitro Toxicity Assessment of Polyvinyl Chloride Particles and Comparison of Six Cellular Systems. *J. Toxicol. Environ. Heal. Part A* **2002**, *65*, 1141–1159. <https://doi.org/10.1080/152873902760125372>.
101. Guzzetti, E.; Sureda, A.; Tejada, S.; Faggio, C. Microplastic in Marine Organism: Environmental and Toxicological Effects. *Environ. Toxicol. Pharmacol.* **2018**, *64*, 164–171. <https://doi.org/10.1016/j.etap.2018.10.009>.
102. Allsopp, M.; Walters, A.; Santillo, D.; Johnston, P. *Plastic Debris in the World's Oceans, Greenspace*; UN Environment Programme, UNEP: Nairobi, Kenya, 2006.
103. Carbery, M.; O'Connor, W.; Palanisami, T. Trophic Transfer of Microplastics and Mixed Contaminants in the Marine Food Web and Implications for Human Health. *Environ. Int.* **2018**, *115*, 400–409. <https://doi.org/10.1016/j.envint.2018.03.007>.
104. Yin, L.; Chen, B.; Xia, B.; Shi, X.; Qu, K. Polystyrene Microplastics Alter the Behavior, Energy Reserve and Nutritional Composition of Marine Jacopever (*Sebastes Schlegelii*). *J. Hazard. Mater.* **2018**, *360*, 97–105. <https://doi.org/10.1016/j.jhazmat.2018.07.110>.
105. Chen, Q.; Gundlach, M.; Yang, S.; Jiang, J.; Velki, M.; Yin, D.; Hollert, H. Quantitative Investigation of the Mechanisms of Microplastics and Nanoplastics toward Zebrafish Larvae Locomotor Activity. *Sci. Total Environ.* **2017**, *584–585*, 1022–1031. <https://doi.org/10.1016/j.scitotenv.2017.01.156>.
106. Sussarellu, R.; Suquet, M.; Thomas, Y.; Lambert, C.; Fabioux, C.; Pernet, M.E.J.; Le Goïc, N.; Quillien, V.; Mingant, C.; Epelboin, Y.; et al. Oyster Reproduction Is Affected by Exposure to Polystyrene Microplastics. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 2430–2435. <https://doi.org/10.1073/pnas.1519019113>.
107. Gardon, T.; Reisser, C.; Soye, C.; Quillien, V.; Le Moullac, G. Microplastics Affect Energy Balance and Gametogenesis in the Pearl Oyster *Pinctada Margaritifera*. *Environ. Sci. Technol.* **2018**, *52*, 5277–5286. <https://doi.org/10.1021/acs.est.8b00168>.

108. Cole, M.; Lindeque, P.; Fileman, E.; Halsband, C.; Galloway, T.S. The Impact of Polystyrene Microplastics on Feeding, Function and Fecundity in the Marine Copepod *Calanus Helgolandicus*. *Environ. Sci. Technol.* **2015**, *49*, 1130–1137. <https://doi.org/10.1021/es504525u>.
109. Banaee, M.; Gholamhosseini, A.; Sureda, A.; Soltanian, S.; Fereidouni, M.S.; Ibrahim, A.T.A. Effects of Microplastic Exposure on the Blood Biochemical Parameters in the Pond Turtle (*Emys Orbicularis*). *Environ. Sci. Pollut. Res.* **2021**, *28*, 9221–9234. <https://doi.org/10.1007/s11356-020-11419-2>.
110. Lu, Y.; Zhang, Y.; Deng, Y.; Jiang, W.; Zhao, Y.; Geng, J.; Ding, L.; Ren, H. Uptake and Accumulation of Polystyrene Microplastics in Zebrafish (*Danio Rerio*) and Toxic Effects in Liver. *Environ. Sci. Technol.* **2016**, *50*, 4054–4060. <https://doi.org/10.1021/acs.est.6b00183>.
111. Kaposi, K.L.; Mos, B.; Kelaher, B.P.; Dworjanyan, S.A. Ingestion of Microplastic Has Limited Impact on a Marine Larva. *Environ. Sci. Technol.* **2014**, *48*, 1638–1645. <https://doi.org/10.1021/es404295e>.
112. Weber, A.; Scherer, C.; Brennholt, N.; Reifferscheid, G.; Wagner, M. PET Microplastics Do Not Negatively Affect the Survival, Development, Metabolism and Feeding Activity of the Freshwater Invertebrate *Gammarus Pulex*. *Environ. Pollut.* **2018**, *234*, 181–189. <https://doi.org/10.1016/j.envpol.2017.11.014>.
113. Rist, S.E.; Assidqi, K.; Zamani, N.P.; Appel, D.; Perschke, M.; Huhn, M.; Lenz, M. Suspended Micro-Sized PVC Particles Impair the Performance and Decrease Survival in the Asian Green Mussel *Perna Viridis*. *Mar. Pollut. Bull.* **2016**, *111*, 213–220. <https://doi.org/10.1016/j.marpolbul.2016.07.006>.
114. Rochman, C.M.; Hoh, E.; Kurobe, T.; Teh, S.J. Ingested Plastic Transfers Hazardous Chemicals to Fish and Induces Hepatic Stress. *Sci. Rep.* **2013**, *3*, 1–7. <https://doi.org/10.1038/srep03263>.
115. Browne, M.A.; Niven, S.J.; Galloway, T.S.; Rowland, S.J.; Thompson, R.C. Microplastic Moves Pollutants and Additives to Worms, Reducing Functions Linked to Health and Biodiversity. *Curr. Biol.* **2013**, *23*, 2388–2392. <https://doi.org/10.1016/j.cub.2013.10.012>.
116. Lei, L.; Wu, S.; Lu, S.; Liu, M.; Song, Y.; Fu, Z.; Shi, H.; Raley-Susman, K.M.; He, D. Microplastic Particles Cause Intestinal Damage and Other Adverse Effects in Zebrafish *Danio Rerio* and Nematode *Caenorhabditis Elegans*. *Sci. Total Environ.* **2018**, *619–620*, 1–8. <https://doi.org/10.1016/j.scitotenv.2017.11.103>.
117. Zhang, C.; Chen, X.; Wang, J.; Tan, L. Toxic Effects of Microplastic on Marine Microalgae *Skeletonema Costatum*: Interactions between Microplastic and Algae. *Environ. Pollut.* **2016**, *220*, 1282–1288. <https://doi.org/10.1016/j.envpol.2016.11.005>.
118. Ribeiro, F.; Garcia, A.R.; Pereira, B.P.; Fonseca, M.; Mestre, N.C.; Fonseca, T.G.; Ilharco, L.M.; Bebianno, M.J. Microplastics Effects in *Scrobicularia Plana*. *Mar. Pollut. Bull.* **2017**, *122*, 379–391. <https://doi.org/10.1016/j.marpolbul.2017.06.078>.
119. Dawson, A.; Huston, W.; Kawaguchi, S.; King, C.; Cropp, R.; Wild, S.; Eisenmann, P.; Townsend, K.; Bengtson Nash, S.M. Uptake and Depuration Kinetics Influence Microplastic Bioaccumulation and Toxicity in Antarctic Krill (*Euphausia Superba*). *Environ. Sci. Technol.* **2018**, *52*, 3195–3201. <https://doi.org/10.1021/acs.est.7b05759>.
120. Chisada, S.; Yoshida, M.; Karita, K. Ingestion of Polyethylene Microbeads Affects the Growth and Reproduction of Medaka, *Oryzias Latipes*. *Environ. Pollut.* **2019**, *254*, 113094. <https://doi.org/10.1016/j.envpol.2019.113094>.
121. Green, D.S. Effects of Microplastics on European Flat Oysters, *Ostrea Edulis* and Their Associated Benthic Communities. *Environ. Pollut.* **2016**, *216*, 95–103. <https://doi.org/10.1016/j.envpol.2016.05.043>.
122. Devriese, L.I.; van der Meulen, M.D.; Maes, T.; Bekaert, K.; Paul-Pont, I.; Frère, L.; Robbens, J.; Vethaak, A.D. Microplastic Contamination in Brown Shrimp (*Crangon Crangon*, Linnaeus 1758) from Coastal Waters of the Southern North Sea and Channel Area. *Mar. Pollut. Bull.* **2015**, *98*, 179–187. <https://doi.org/10.1016/j.marpolbul.2015.06.051>.
123. Messinetti, S.; Mercurio, S.; Scari, G.; Pennati, A.; Pennati, R. Ingested Microscopic Plastics Translocate from the Gut Cavity of Juveniles of the Ascidian *Ciona Intestinalis*. *Eur. Zool. J.* **2019**, *86*, 189–195. <https://doi.org/10.1080/24750263.2019.1616837>.
124. Lo, H.K.A.; Chan, K.Y.K. Negative Effects of Microplastic Exposure on Growth and Development of *Crepidula Onyx*. *Environ. Pollut.* **2018**, *233*, 588–595. <https://doi.org/10.1016/j.envpol.2017.10.095>.
125. Nizzetto, L.; Futter, M.; Langaas, S. Are Agricultural Soils Dumps for Microplastics of Urban Origin? *Environ. Sci. Technol.* **2016**, *50*, 10777–10779.
126. Möller, J.N.; Löder, M.G.J.; Laforsch, C. Finding Microplastics in Soils: A Review of Analytical Methods. *Environ. Sci. Technol.* **2020**, *54*, 2078–2090. <https://doi.org/10.1021/acs.est.9b04618>.
127. Gabet, E.J.; Reichman, O.J.; Seabloom, E.W. The Effects of Bioturbation on Soil Processes and Sediment Transport. *Annu. Rev. Earth Planet. Sci.* **2003**, *31*, 249–273. <https://doi.org/10.1146/annurev.earth.31.100901.141314>.
128. Hurley, R.R.; Nizzetto, L. Fate and Occurrence of Micro(Nano)Plastics in Soils: Knowledge Gaps and Possible Risks. *Curr. Opin. Environ. Sci. Heal.* **2018**, *1*, 6–11. <https://doi.org/10.1016/j.coesh.2017.10.006>.
129. Gutiérrez-López, M.; Salmon, S.; Trigo, D. Movement Response of Collembola to the Excreta of Two Earthworm Species: Importance of Ammonium Content and Nitrogen Forms. *Soil Biol. Biochem.* **2011**, *43*, 55–62. <https://doi.org/10.1016/j.soilbio.2010.09.010>.
130. Cao, D.; Wang, X.; Luo, X.; Liu, G.; Zheng, H. Effects of Polystyrene Microplastics on the Fitness of Earthworms in an Agricultural Soil. *IOP Conf. Ser. Earth Environ. Sci.* **2017**, *61*, 012148. <https://doi.org/10.1088/1755-1315/61/1/012148>.

131. Huerta Lwanga, E.; Gertsen, H.; Gooren, H.; Peters, P.; Salánki, T.; van der Ploeg, M.; Besseling, E.; Koelmans, A.A.; Geissen, V. Incorporation of Microplastics from Litter into Burrows of *Lumbricus Terrestris*. *Environ. Pollut.* **2017**, *220*, 523–531. <https://doi.org/10.1016/j.envpol.2016.09.096>.
132. Lahive, E.; Walton, A.; Horton, A.A.; Spurgeon, D.J.; Svendsen, C. Microplastic Particles Reduce Reproduction in the Terrestrial Worm *Enchytraeus Crypticus* in a Soil Exposure. *Environ. Pollut.* **2019**, *255*, 113174. <https://doi.org/10.1016/j.envpol.2019.113174>.
133. Rillig, M.C.; Ziersch, L.; Hempel, S. Microplastic Transport in Soil by Earthworms. *Sci. Rep.* **2017**, *7*, 1–6. <https://doi.org/10.1038/s41598-017-01594-7>.
134. Lei, L.; Liu, M.; Song, Y.; Lu, S.; Hu, J.; Cao, C.; Xie, B.; Shi, H.; He, D. Polystyrene (Nano)Microplastics Cause Size-Dependent Neurotoxicity, Oxidative Damage and Other Adverse Effects in *Caenorhabditis Elegans*. *Environ. Sci. Nano* **2018**, *5*, 2009–2020. <https://doi.org/10.1039/c8en00412a>.
135. Zhu, D.; Chen, Q.L.; An, X.L.; Yang, X.R.; Christie, P.; Ke, X.; Wu, L.H.; Zhu, Y.G. Exposure of Soil Collembolans to Microplastics Perturbs Their Gut Microbiota and Alters Their Isotopic Composition. *Soil Biol. Biochem.* **2018**, *116*, 302–310. <https://doi.org/10.1016/j.soilbio.2017.10.027>.
136. Song, Y.; Cao, C.; Qiu, R.; Hu, J.; Liu, M.; Lu, S.; Shi, H.; Raley-Susman, K.M.; He, D. Uptake and Adverse Effects of Polyethylene Terephthalate Microplastics Fibers on Terrestrial Snails (*Achatina Fulica*) after Soil Exposure. *Environ. Pollut.* **2019**, *250*, 447–455. <https://doi.org/10.1016/j.envpol.2019.04.066>.
137. Kim, S.W.; An, Y.J. Soil Microplastics Inhibit the Movement of Springtail Species. *Environ. Int.* **2019**, *126*, 699–706. <https://doi.org/10.1016/j.envint.2019.02.067>.
138. Ju, H.; Zhu, D.; Qiao, M. Effects of Polyethylene Microplastics on the Gut Microbial Community, Reproduction and Avoidance Behaviors of the Soil Springtail, *Folsomia Candida*. *Environ. Pollut.* **2019**, *247*, 890–897. <https://doi.org/10.1016/j.envpol.2019.01.097>.
139. Yi, M.; Zhou, S.; Zhang, L.; Ding, S. The Effects of Three Different Microplastics on Enzyme Activities and Microbial Communities in Soil. *Water Environ. Res.* **2020**, *93*, 24–32. <https://doi.org/10.1002/wer.1327>.
140. Wan, Y.; Wu, C.; Xue, Q.; Hui, X. Effects of Plastic Contamination on Water Evaporation and Desiccation Cracking in Soil. *Sci. Total Environ.* **2019**, *654*, 576–582. <https://doi.org/10.1016/j.scitotenv.2018.11.123>.
141. Prendergast-Miller, M.T.; Katsiamides, A.; Abbass, M.; Sturzenbaum, S.R.; Thorpe, K.L.; Hodson, M.E. Polyester-Derived Microfibre Impacts on the Soil-Dwelling Earthworm *Lumbricus Terrestris*. *Environ. Pollut.* **2019**, *251*, 453–459. <https://doi.org/10.1016/j.envpol.2019.05.037>.
142. Rodríguez-Seijo, A.; da Costa, J.P.; Rocha-Santos, T.; Duarte, A.C.; Pereira, R. Oxidative Stress, Energy Metabolism and Molecular Responses of Earthworms (*Eisenia Fetida*) Exposed to Low-Density Polyethylene Microplastics. *Environ. Sci. Pollut. Res.* **2018**, *25*, 33599–33610. <https://doi.org/10.1007/s11356-018-3317-z>.
143. Yu, Y.; Chen, H.; Hua, X.; Dang, Y.; Han, Y.; Yu, Z.; Chen, X.; Ding, P.; Li, H. Polystyrene Microplastics (PS-MPs) Toxicity Induced Oxidative Stress and Intestinal Injury in Nematode *Caenorhabditis Elegans*. *Sci. Total Environ.* **2020**, *726*, 138679. <https://doi.org/10.1016/j.scitotenv.2020.138679>.
144. Qi, Y.; Yang, X.; Pelaez, A.M.; Huerta Lwanga, E.; Beriot, N.; Gertsen, H.; Garbeva, P.; Geissen, V. Macro- and Micro-Plastics in Soil-Plant System: Effects of Plastic Mulch Film Residues on Wheat (*Triticum Aestivum*) Growth. *Sci. Total Environ.* **2018**, *645*, 1048–1056. <https://doi.org/10.1016/j.scitotenv.2018.07.229>.
145. Bosker, T.; Bouwman, L.J.; Brun, N.R.; Behrens, P.; Vijver, M.G. Microplastics Accumulate on Pores in Seed Capsule and Delay Germination and Root Growth of the Terrestrial Vascular Plant *Lepidium Sativum*. *Chemosphere* **2019**, *226*, 774–781. <https://doi.org/10.1016/j.chemosphere.2019.03.163>.
146. Jiang, X.; Chen, H.; Liao, Y.; Ye, Z.; Li, M.; Klobučar, G. Ecotoxicity and Genotoxicity of Polystyrene Microplastics on Higher Plant *Vicia Faba*. *Environ. Pollut.* **2019**, *250*, 831–838. <https://doi.org/10.1016/j.envpol.2019.04.055>.
147. De Souza Machado, A.A.; Lau, C.W.; Kloas, W.; Bergmann, J.; Bachelier, J.B.; Faltin, E.; Becker, R.; Görlich, A.S.; Rillig, M.C. Microplastics Can Change Soil Properties and Affect Plant Performance. *Environ. Sci. Technol.* **2019**, *53*, 6044–6052. <https://doi.org/10.1021/acs.est.9b01339>.
148. Gao, M.; Liu, Y.; Song, Z. Effects of Polyethylene Microplastic on the Phytotoxicity of Di-n-Butyl Phthalate in Lettuce (*Lactuca Sativa* L. Var. *Ramosa* Hort). *Chemosphere* **2019**, *237*, 124482. <https://doi.org/10.1016/j.chemosphere.2019.124482>.
149. Hernández-Arenas, R.; Beltrán-Sanahuja, A.; Navarro-Quirant, P.; Sanz-Lazaro, C. The Effect of Sewage Sludge Containing Microplastics on Growth and Fruit Development of Tomato Plants. *Environ. Pollut.* **2021**, *268*, 115779. <https://doi.org/10.1016/j.envpol.2020.115779>.
150. Kumar, A.; Mishra, S.; Pandey, R.; Yu, Z.G.; Kumar, M. Microplastics in terrestrial ecosystems: Un-ignorable impacts on soil characterises, nutrient storage and its cycling. *TrAC Trends Anal. Chem.* **2023**, *158*, 116869. <https://doi.org/10.1016/j.trac.2022.116869>.
151. Ministry of Foreign Affairs of Japan Japan's "MARINE Initiative" toward Realization of the Osaka Blue Ocean Vision; Ministry of Foreign Affairs of Japan: Tokyo, Japan, 2019; p. 2050.
152. Marine Debris Program, N. *Laboratory Methods for the Analysis of Microplastics in the Marine Environment: Recommendations for Quantifying Synthetic Particles in Waters and Sediments*; NOAA Marine Debris Division: Silver Spring, MD, USA, 2015.

153. Wu, J.F.; Yao, H.X.; Yuan, X.; Lin, B.Q. Dissolved organic carbon response to hydrological drought characteristics: Based on long-term measurements of headwater streams. *Water Res.* **2022**, *215*, 115252. <https://doi.org/10.1016/j.watres.2022.118252>.
154. Pasquier, G.; Doyen, P.; Kazour, M.; Dehaut, A.; Diop, M.; Duflos, G.; Amara, R. Manta Net: The Golden Method for Sampling Surface Water Microplastics in Aquatic Environments. *Front. Environ. Sci.* **2022**, *10*, 1–12. <https://doi.org/10.3389/fenvs.2022.811112>.
155. Zobkov, M.B.; Esiukova, E.E.; Zyubin, A.Y.; Samusev, I.G. Microplastic Content Variation in Water Column: The Observations Employing a Novel Sampling Tool in Stratified Baltic Sea. *Mar. Pollut. Bull.* **2019**, *138*, 193–205. <https://doi.org/10.1016/j.marpolbul.2018.11.047>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Multimedia information hiding method for AMBTC compressed images using LSB substitution technique

Rajeev Kumar¹ · Aruna Malik²

Received: 11 April 2021 / Revised: 13 July 2022 / Accepted: 2 November 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

Communication bandwidth plays a significant role in real-time communication. For this, absolute moment block truncation coding (AMBTC) has been popular. An AMBTC based high capacity multimedia data hiding method for covert communication is proposed in this paper. The proposed method applies AMBTC to the host image to get the compressed image in the form of AMBTC trios. The trios are then classified into three categories based on the difference between their quantization levels. The proposed method then adaptively conceals the secret message bits in the bitmaps of the trios based on their type. Additionally, the proposed method conceals two bits of secret data in every combination of quantization levels irrespective of their block types using the LSB substitution technique. The experimental results show that the proposed AMBTC based method is perform better to the other related data hiding based methods in terms of embedding capacity while providing comparable PSNR.

Keywords Data hiding · AMBTC · LSB substitution · Stego-image

Abbreviations

AMBTC	Absolute Moment Block Truncation Coding
BTC	Block Truncation Coding
DCT	Discrete Cosine Transform
DWT	Discrete Wavelet Transform
JNC	Joint Neighbour Coding
LSB	Least Significant Bit
PSNR	Peak Signal-to-Noise Ratio

✉ Aruna Malik
arunacsrke@gmail.com

Rajeev Kumar
rajeevkumar@dtu.ac.in

¹ Department of CSE, Delhi Technological University New Delhi, New Delhi, India

² Department of CSE, National Institute of Technology Jalandhar, Punjab, India

1 Introduction

Data hiding is an art and science of concealing important data in a cover file [11, 12] which can be an image, video, text audio, or any other multimedia file. There are two main objectives of any multimedia communication system: 1) authenticity 2) confidential communication, for embedding the data in cover media. The later objective is achieved by steganography techniques and the first one by watermarking techniques. The watermarking technique imperceptibility or perceptibility modifies the cover media so that the authenticity of the media can be determined at the receiving end. Watermarking has its applications not only in copyright protection but also in the field of content authentication, broadcast monitoring, content description, proof of ownership, and transaction tracking. In contrast, steganography is mainly used for confidential or private multimedia communication. The steganography techniques modify the cover medium in such a way that the modification is not detectable. In this way, steganography can hide even the presence of the secret information or communication [1, 2]. The applications of steganography that can be information alteration protection, media related database systems, storing secret data and confidential communication, and digital content distribution through the confidential control system.

To judge the quality of data hiding techniques, there exist three main parameters, 1) visual quality, 2) data hiding capacity, and 3) robustness. For the steganography techniques, the first two are the prime parameters [25]. Visual quality is related to the difference between the original image with respect to the image after data hiding whereas the data hiding capacity is the capacity of hiding maximum data in the cover image pixels. There are three domains namely, spatial domain, compression domain, and frequency domain, in which the data can be concealed [17]. In the case of the spatial domain, the intensity values of pixels are altered to innocuously hide the secret information. [22]. In the frequency domain, the cover medium is first transformed into a frequency domain using either *discrete cosine transform* (DCT) or discrete wavelet transform (DWT), and then secret payload bits are fused into coefficient values of the transformed image. Further, these coefficients can also retrace the frequency form of the image. The compressed domain-based data hiding involves the conversion of the cover medium into compressed codes followed by embedding of data into these codes.

To tackle with the problems related to limited communication bandwidth [8, 14, 18], compression domain methods are usually preferred. Various compression domain based techniques have been proposed to address this problem; however, it is recommended to adopt lossy compression techniques [8, 14, 18], when the cover medium is video or image. BTC is one of the robust and efficient methods because of its simplicity and high compression ratio. Further refinements of BTC are implemented in the form of AMBTC [19] to improve its performance. In AMBTC, preservation of the first absolute moment is done along with the mean, in contrast to standard deviation. The AMBTC method is computationally simpler than BTC. In the proposed work, a compression domain based data hiding approach using AMBTC has been followed. The proposed data hiding technique is an extension of Ou et al. [26] scheme and uses LSB substitution to additionally accommodate the secret data in quantization levels along with the bit-plane. This technique also preserves the quality of the image even after additionally hiding the data. To support this statement, experimental results are provided in section 6, which justify that along with embedding a large amount of secret payload as

compared to some other related techniques, the proposed method also maintains the PSNR value which reflects marked image quality. The proposed method will be applicable in the scenarios such as medical imaging, military communication etc., where real-time covert communication is highly desirable.

The rest of the paper is structured as follows: motivation and contribution of the proposed work are discussed in Section 2. Section 3 is dedicated to a brief introduction to AMBTC. Related works are discussed in section 4. The proposed method is discussed in section 5. Section 6 contains the experimental results and comparative analysis. At last, section 7 is provided with the concluding remarks.

2 Motivation and contribution

It has been noticed that the categorization of AMBTC trios in three categories namely, smooth, moderately, and highly complex to hide the secret data has shown noteworthy enhancement in the performance. The majority of data hiding methods for AMBTC compressed images have not been able to use the complex blocks to embed the secret information bits due to their sensitivity to distortion. This motivated us to propose a new AMBTC based hiding scheme using LSB substitution which optimally uses the compressed image blocks for data hiding. The contribution of this paper is summarized as follows:

- The proposed information hiding method classifies the compressed image blocks (also known as AMBTC trios) into three categories based on the block-wise texture of the host image.
- Next, the proposed LSB substitution method optimally conceals the bits of secret payload in an image block (AMBTC trio) based on its category so that balance between embedding capacity and PSNR value can be maintained.
- Further, the proposed information hiding method conceals secret payload into quantization levels using the LSB substitution technique.
- Experimental results provide evidence of the proposed AMBTC based data hiding method's superiority on the existing and related data hiding methods as far as embedding capacity is concerned. Furthermore, our method provides comparable quality marked images.

In the next section, AMBTC is briefly reviewed.

3 Absolute moment block truncation coding (AMBTC)

AMBTC is a type of lossy image compression technique specially designed for grayscale images. It is a block-based technique that is efficient and simple as compared to other image compression standards. The main idea of AMBTC revolves around the quantization of the pixels into two levels in a block-wise manner [9]. The encoding and decoding processes of the AMBTC technique are discussed below:

At the encoder side, firstly, the input image is non-overlappingly partitioned into non $N \times N$ sized blocks where N is preferably set to 4. Then a mean value (Avg_i) is computed by using Eq. (1) for each image block.

$$\text{Avg}_i = \sum_{j=1}^{N \times N} \alpha_j / N \times N \quad (1)$$

where α_j is the intensity value of j^{th} pixel for i^{th} each block and N is a representative of the image block size. A bitmap for each block is generated using the following two rules:

- If $\alpha_j < \text{Avg}_i$ then the j^{th} pixel of the i^{th} block is denoted by '0'.
- Else, the pixel is denoted by '1'.

Next, the following Eqs. (2) and (3) are used to compute the low L_i and high H_i mean values, respectively.

$$L_i = \sum_{\alpha_j < \text{Avg}_i}^{N \times N} \alpha_j / q \quad (2)$$

$$H_i = \sum_{\alpha_j \geq \text{Avg}_i}^{N \times N} \alpha_j / (N \times N - q) \quad (3)$$

Here, q represents the count of pixels that are $\geq \text{Avg}$. Thus, each block of the image is represented by L_i , H_i and the respective bitmap. To re-generate the blocks of the image at the decoder side, the '0' & '1' in the bitmap are substituted by L_i and H_i , respectively. The next section is devoted to the discussion of some prominent methods of data hiding for AMBTC and BTC based compressed images.

4 Related works

This section briefly reviews the evolution of BTC and AMBTC based data hiding methods [3–7, 9, 13, 15, 16, 20, 21, 23, 24, 26–30]. To the best of our knowledge, the first BTC based data hiding method was introduced by Chuang et al. [6], which is applicable to grey-scale cover images. Chuang et al. [6] method applies the BTC method to compress the host image to get low and high mean values and one bit-plane per block. Then, the BTC encoded blocks are categorized into smooth or complex categories using a user-defined threshold. Ultimately, selective bit-planes are used to carry the secret data. However, this results into a lower compression ratio than the other variants of BTC. Chang et al. [3] discussed a BTC based reversible data hiding (RDH) method for colored images. The method first compresses each channel of colored host image using BTC and then uses a genetic algorithm to select a nearly optimal bit-plane out of three planes so that the compression rate can be improved. However, the method can embed only three bits (on average) in each BTC-encoded block. Chen et al. [5] asserted a data hiding method that utilizes the difference of quantization values for data hiding. Though this method provides better PSNR, offers limited embedding capacity.

In 2011, Li et al. [20] proposed a novel RDH method for BTC compressed images. This method uses a histogram shifting approach along with a bit-plane flipping strategy with an aim to provide a very less distorted stego-image. However, the method has a limitation as far as embedding capacity is concerned. To address the problem of limited capacity, Sun et al. [28] introduced a RDH method using joint neighbour coding (JNC). JNC is used to hide bits of secret payload in quantization tables to enhance the embedding capacity. Although it offers data embedding four times greater than the number of blocks, it has some overhead. Zhang et al. [30] come up with BTC based RDH scheme which first losslessly compresses the BTC encoded images by exploiting the secret data characteristics. After applying the BTC method to the image, BTC compressed codes are obtained. These codes consist of low and high mean

tables and a bit plane sequence that are further utilized for carrying the secret data. However, this method requires the overhead of information such as the key for sequence generation, size of cover image, etc., so that reversibility can be maintained.

In 2013, Lin et al. [23] discussed a new AMBTC based RDH scheme that first investigates the redundancy of the compressed blocks and then classifies compressed blocks into the embeddable or non-embeddable category. Next, the secret data is embedded based on some pre-defined rules in embeddable blocks. However, in the process of data hiding the compression ratio is compromised. Pan et al. [27] discuss a reference matrix based RDH scheme for AMBTC compressed images in which high & low mean values of each AMBTC trio carry secret data. Ou et al. [26] introduce a new data hiding method with an idea to minimize distortion. In this method, the AMBTC compressed blocks/trios are tagged as smooth or complex based on the difference between quantization levels. In the case of smooth trios, the bits of secret payload are embedded in the bit-planes and the quantization levels are re-computed based on the new bit-plane. In the case of complex trios, the bits of secret payload are concealed by doing an exchange of the position of the quantization levels and correspondingly flipping the bits of the bit-plane. However, the complex block's bit-plane is not utilized for embedding the secret information due to its sensitivity.

Huang et al. [9] discuss hybridized data hiding method based on AMBTC. This method is an enhancement of Ou et al. scheme [26] for embedding some additional secret data bits in both complex and smooth blocks. Recently Kumar et al. [16] discuss a new AMBTC compression method based on image interpolation for improving the compression ratio. The method makes the user of Weber's law for categorizing the image blocks into smooth and complex blocks and then only stores half of the bits of the smooth block's bitmap to improve the compression ratio. At the receiving end, the complete smooth block is predicted using defined equations for interpolation. Thus, the method is able to maintain the image quality of the compressed image while further reducing the required number of bits to represent the image. Hong [7] discusses a new data hiding method that is based on the concept of pixel pair matching. The method further reduces the stego-image distortion by efficiently processing the smooth blocks. Additionally, the adaptive pixel pair matching technique helps in concealing the bits of secret payload into the quantization levels. Li et al. [21] discuss an AMBTC based bi-stretch RDH algorithm that hides the bits of the secret payload by exploiting the properties of AMBTC coefficients.

Chen and Chi [4] introduced a new BTC based data hiding and blind decoding scheme. This scheme categorizes the AMBTC trios in three categories (instead of two) based on their roughness properties. Next, the bits of secret payload are embedded based on the block type so that both optimal quality and capacity can be achieved. The work of [4, 26] is extended by Kumar et al. [13] by introducing a new AMBTC based information hiding method. Kumar et al. first categorizes the image into three types of blocks based on their correlation and then embed the secret message using hamming distance and pixel value differencing in case the image block is not a smooth one. Thus, the method is able to embed a few bits of secret data into complex blocks as well. The concept of hamming distance/code is further utilized in [24, 29] to improve the performance. In this paper, a new high capacity AMBTC based data hiding method is proposed by adopting the idea of two thresholds from [4, 13]. The proposed method also uses the LSB substitution technique for additional embedding in quantization levels. In 2019, Kumar and Jung [10] did a survey of data hiding techniques based on block truncation coding. In the survey, the authors presented a taxonomy based on the working methods. The survey concludes that the performance of BTC based data hiding techniques is highly limited

to the performance of block truncation coding. They also suggested to make use of adaptive techniques to fully exploit the BTC qualities.

5 Proposed method

The detailed algorithm in the form of steps of the proposed method is discussed. The proposed method is an extension of Ou et al. scheme [26]. The proposed method first employs AMBTC compression to compress the original host image and gets the AMBTC trios. Next, the trios are categorized into either *Smooth*, *Less_Complex*, or *Highly_Complex* trios based on two user-provided thresholds. It is to be noted that the values of both the threshold must be an even number. In the case of the *smooth* trio, the proposed method replaces the bits of the bit-plane by bits of secret payload. In the case of *Less_Complex* trios, the four innermost bits of the bit-plane are simply substituted by the four secret bits payload. Subsequently, the low and high mean values of the block are re-computed using Eqs. (4)–(5) to reduce the alteration. However, as far as *Highly_Complex* trios are concerned, no embedding is done into the bit-plane to avoid large distortion. Furthermore, the quantization levels of each trios are used to conceal two bits of the payload by simply replacing their first LSB by secret data bits. The detailed embedding algorithm of the our method is provided in sub-section 5.1.

5.1 Embedding algorithm

Input- I : Host image of $M \times M$ pixels, $thr1$: Smooth_threshold, $thr2$: Complex_threshold, S : secret data bitstream.

Output- Marked trios/codes.

BEGIN

Step 1: Read host image I in raster scan order to non-overlappingly and divide into blocks of 4×4 (pixels)

Step 2: Compute two quantization levels L_i and H_i hereafter, named say a_i and b_i , respectively, using Eqs. (2) & (3) and a bit-plane B_i for each host image block IB_i using rules defined in Section 3.

Step 3: Calculate absolute difference e.g., D_i between a_i and b_i .

Step 4: If $D_i \leq thr1$, means IB_i is a smooth trio. Substitute all the bits of B_i with 16 bits of secret data (S) to get new bit-plane B'_i .

Step 5: Re-Compute quantization levels (a_i and b_i) as a'_i and b'_i w.r.t. B'_i using Eqs. 4 and 5, respectively, as defined follows.

$$a'_i = \frac{1}{16-q} \sum_{x_i \in G_0} x_i \quad (4)$$

$$b'_i = \frac{1}{q} \sum_{x_i \in G_1} x_i \quad (5)$$

where q represents count of 1's in B'_i .

Step 6: If $|a'_i - b'_i| \leq thr1$. It means the smoothness of the block is not changed so add $\{a'_i, b'_i, B'_i\}$ into I_S otherwise $|a'_i - b'_i| > thr$, which means that the smoothness property of the block has been violated. So, to preserve its smoothness, add $\{a_i, b_i, B'_i\}$ into I_S .

Step 7: If $(|a'_i - b'_i| == thr1)$ or $(|a_i - b_i| == thr1)$ of the obtained trio from **Step 6** then $b'_i = b'_i - 1$ or $b_i = b_i - 1$ (whichever is applicable).

Step 8: Replace first LSBs of both low and high mean values with bits of secret data using LSB substitution method. Go to Step 3 and start the process for the next block.

Step 9: If $thr2 \leq D_i > thr1$ means IB_i is a *Less_Complex* block. In case of *Less_Complex* block, the innermost sub-blocks of the bit-plane (which has four pixels) is replaced by 4-bits of secret data (S) to get new bit-plane B'_i .

Step 10: Re-Compute quantization levels (a_i and b_i) as a'_i and b'_i w.r.t B'_i using Eqs. 4 and 5.

Step 11: If $thr2 \leq |a'_i - b'_i| > thr1$ means that the *Less_Complex* property of the block is preserved, add $\{a'_i, b'_i, B'_i\}$ into I_S . If $thr2 > |a'_i - b'_i| \leq thr1$ it means that the *Less_Complex* property has been violated. So, to preserve its *Less_Complex* property, add old a_i and b_i into I_S to get $\{a_i, b_i, B'_i\}$.

Step 12: If $(|a'_i - b'_i| == thr1 + 1)$ or $(|a_i - b_i| == thr1 + 1)$ of obtained trio from Step 11 then check if $(b'_i \% 2 == 1)$ or $(b_i \% 2 == 1)$ (whichever is applicable) then $b'_i = b'_i + 1$ or $b_i = b_i + 1$.

Step 13: If $(|a'_i - b'_i| == thr2)$ or $(|a_i - b_i| == thr2)$ of the obtained trio from Step 11 then $b'_i = b'_i - 1$ or $b_i = b_i - 1$ (whichever is applicable).

Step 14: Replace the first LSBs of both the quantization levels by the bits of secret payload using the LSB method. Repeat the above process from Step 3 for the next block.

Step 15: If $D_i > thr2$, then go to next step.

Step 16: If $D_i == thr2 + 1$ and $(b_i \% 2 == 1)$ then $b_i = b_i + 1$.

Step 17: Replace the first LSBs of both the quantization levels by the bits of secret payload using the LSB method. Repeat the above process from Step 3 for the next block.

END

The process is repeated until the whole image is visited, thus, a stego-image in the form of trios is obtained. In the next sub-section, the extraction process is described.

5.2 Extraction process

The extraction process defines the steps to extract the bits of the secret payload from the marked AMBTC codes. A detailed algorithm for the extraction process is given below:

Input- Marked codes, $thr1$: Smooth_threshold, $thr2$: Complex_threshold.

Output- S_D : Bit stream of secret message/payload

Step 1: Compute absolute difference D_i between a'_i and b'_i .

Step 2: If $D_i \leq thr1$, extract all the 16 bits of the bit-plane and add them into S_D .

Step 3: If $thr2 \leq D_i > thr1$, then extract 4 bits of the innermost sub-block of the bit-plane and add them to S_D .

Step 4: If $D_i > thr2$, then Go to Step 5.

Step 5: Extract the first LSB of a'_i and b'_i and append them to S_D . Then go to step 6.

Step 6: Go back to step 1 until all the trios are visited.

Thus, all the bits of the secret payload (S_D) are extracted from the marked AMBTC codes.

5.3 An example of embedding and extraction procedure

Here, the embedding procedure and extraction procedure are discussed using an example. First of all, the embedding procedure is explained with the help of a suitable diagram which is given in Fig. 1. The data bit stream of the secret data that is considered a sequence of 0 and 1 as $(10100101001011101010101001)_2$. There are two threshold values required as defined the proposed method which are considered as $thr1$ (*Smooth_threshold*) and $thr2$

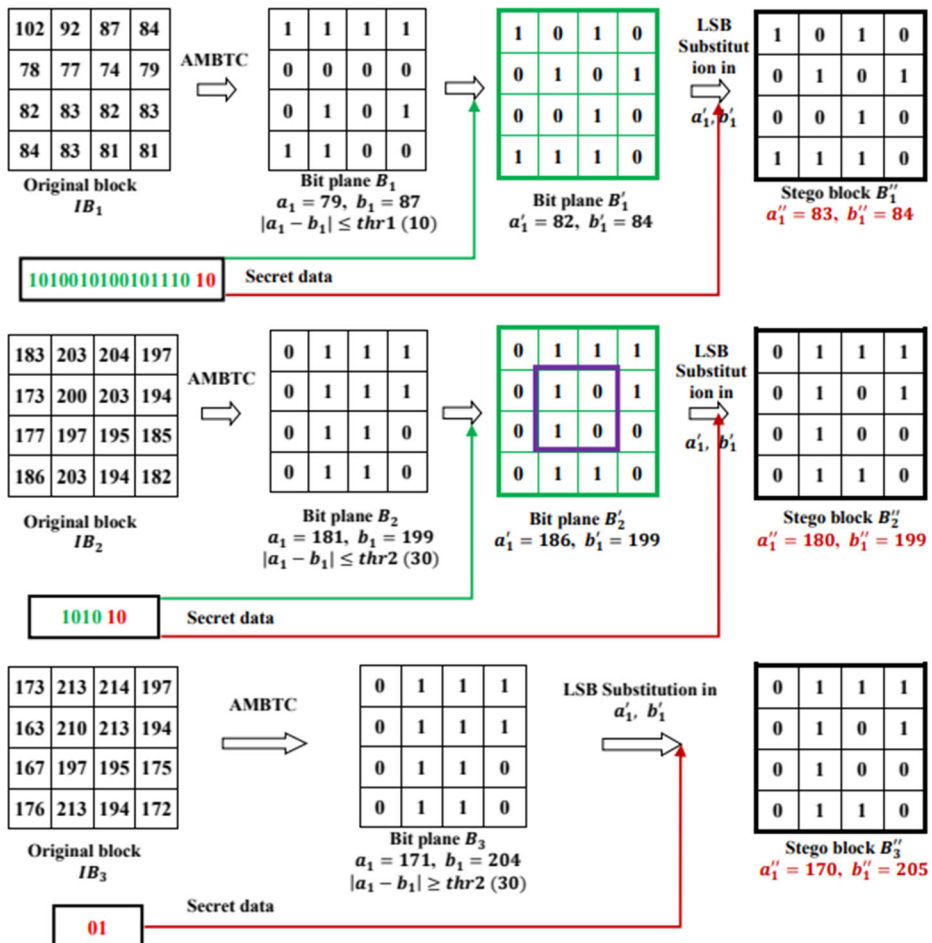


Fig. 1 Proposed embedding procedure example

(*Complex_threshold*) with values 10 and 30 only for the example, respectively. A 4×4 pixels block size is considered for all three different blocks of the image. In the first block called IB_1 16 values are considered as 102, 92, 87, 84, 78, 77, 74, 79, 82, 83, 82, 83, 84, 83, 81, and 81. In the second block called IB_2 , another 16 values of the image pixels are considered as 183, 203, 204, 197, 173, 200, 203, 194, 177, 197, 195, 186, 203, 194, 182, and 185, whereas in the third block called IB_3 , another 16 values of the image pixels are considered as 173, 213, 214, 197, 163, 210, 213, 194, 167, 197, 195, 175, 176, 213, 194, and 172. All the three blocks IB_1 , IB_2 , and IB_3 are compressed with the help of the AMBTC technique and then we get three trios as $\{79, 87, B_1\}$, $\{181, 199, B_2\}$ and $\{171, 204, B_3\}$, respectively. The blocks IB_1 , IB_2 , and IB_3 have 83, 192, and 191 mean values, respectively. We have only two values in the bit plan either 0 or 1. Find absolute difference (D_1) for the first block i.e., $D_1 = |a_1 - b_1| = |79 - 87| = 8 < thr1 = 8 < 10$. D_1 value is less than the threshold value thus, this block is considered the smooth block for data hiding. So, IB_1 is considered for hiding the secret data, and bits of B_1 is replaced with the 16 number of secret bits. According to the embedding process of the ABMTC, a'_1 and b'_1 two new quantization levels are computed using Eqs. (4) and (5). Now, we check the modification of the old and new a'_1 and b'_1 such as $|a'_1 - b'_1|$ i.e., $|82 - 84| = 2$ that is below the $thr1$ which indicates the preservation of smoothness property for the block. Now, The LSB substitution method is computed to the quantization levels a'_1 and b'_1 so that the 2 next secret data bits can be considered for embedding. Then, updated values for quantization levels are $a'_1 = 83$, $b'_1 = 84$. Therefore, secret data embedded in the smooth trio is i.e., $S_1 = (1010010100101110\ 10)_2$ which is subsequently detached as $S = S - S_1$, and the remaining secret data stream is applied for the next block.

For IB_2 , D_2 is computed as $D_2 = |a_1 - b_1| = |181 - 199| = 18$, which is in the range of $((thr1 + 1) \text{ to } thr2)$ i.e., (11 to 30). Thus, IB_2 is called a *Less_Complex* block. IB_2 is considered for hiding the secret data where four innermost bits of B_1 are substituted with the secret data four bits as shown in Fig. 1 through the violet color rectangle. Now, we check the modification of the old and new a'_1 and b'_1 such as $|a'_1 - b'_1|$ i.e., $|186 - 199| = 13$ that is in the range of 11 to 30 which indicates the preservation of *Less_Complex* property for the block. Now, The LSB substitution method is computed to the quantization levels a'_1 and b'_1 so that the 2 next secret data bits can be considered for embedding. Then, updated values for quantization levels are $a'_1 = 180$, $b'_1 = 199$. Therefore, secret data embedded in the *Less_Complex* trio is i.e., $S_2 = (101010)_2$ which is subsequently detached as $S = S - S_2$, and the remaining secret data stream is applied for the next block. For IB_3 , D_3 is computed as $D_3 = |a_1 - b_1| = |171 - 204| = 33$, i.e., larger than the threshold value i.e., 30. Thus, IB_3 is called a *Highly_Complex* block. The secret data bits (01) are then embedded and thus, the updated values for quantization levels are $a'_1 = 170$, $b'_1 = 205$. The embedded secret data in the highly complex trio is i.e., $S_3 = (01)_2$ which is subsequently detached as $S = S - S_3$, and the remaining secret data stream is applied for the next block. The same process is followed for all the blocks of the cover image. The extraction procedure of our method is the just reverse as we discussed in the embedding process.

6 Results and discussions

This section performs the comparative analyse of results for our data hiding method against some of the existing and relevant methods like Chuang et al. [6], Ou et al. [26], Huang et al. [9], Hong [7] and Kumar et al. [15]. Figure 2 shows the twelve number of grayscale images of pixels 512×512 that are considered for its critical analysis. The experiments are carried out in MATLAB and the secret data is generated as a random sequence of bits. Two parameters considered for the purpose of comparative study and performance measure are peak signal to noise ratio (PSNR) and embedding capacity. The embedding capacity that is measured in “number of bits”, denotes the amount of secret payload that is accommodated in the image. On the other hand, the PSNR is the metric to study the visual quality of the marked or stego image. The PSNR is calculated using Eq. (6) as follows.

$$\text{PSNR} = 10\log_{10} \left[\frac{255*255}{\text{MSE}} \right] \quad (6)$$

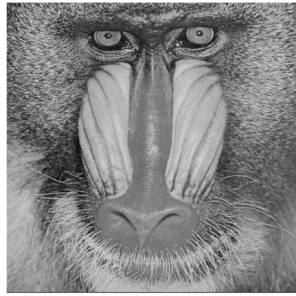
where the mean squared error (MSE) signifies the mean squared error among the grayscale cover image and the marked image or stego image obtained after embedding the secret payload.

As discussed, the proposed information hiding method marks the image blocks into three categories. This in turn, directly influences the image quality and hiding capacity of the marked or stego image. In simple words, as the value of thresholds increases, the image quality of the marked image decreases while the embedding capacity increases. In regard to this, the proposed method experimental results are taken on different thresholds for the test original images as presented in Fig. 2a-l, which are provided in Table 1. As capacity and quality requirements may vary from application to application, the value of the thresholds can be tuned accordingly. The stego images are given in Fig. 3.

An evaluation of performance parameters of our method with other related and popular AMBTC based methods like Kumar et al. [15], Chuang et al. [6] and Ou et al. [26] has been done. The working of our scheme is closely related to Ou et al. [26] scheme. The average embedding capacity and PSNR offered by our scheme, Chuang et al. [6], Ou et al. [26] & Kumar et al. [15], are 82,517, 150,761, 140,051, & 102,630 bits and 29.42, 30.90, 30.34, & 31.31 dB, respectively at the threshold values $\text{thr1} = 10$ and $\text{thr2} = 30$. At the threshold values $\text{thr1} = 20$ and $\text{thr2} = 50$, the average embedding capacity and PSNR offered by our scheme, Chuang et al. [6], Ou et al. [26] & Kumar et al. [15] are 233,629, 190,224, 190,157, & 129,741 bits and 28.74, 30.25, 29.26, & 30.82 dB, respectively. The average hiding capacity and PSNR offered by our scheme, Chuang et al. [6], Ou et al. [26] & Kumar et al. [15] are 254,249, 211,284, 211,070 & 144,361 bits, and 28.39, 29.51, 28.27 & 30.47 dB, respectively at the threshold values $\text{thr1} = 30$ and $\text{thr2} = 60$. At the threshold values $\text{thr1} = 40$ and $\text{thr2} = 70$, the average embedding capacity and PSNR offered by our scheme, Chuang et al. [6], Ou et al. [26] & Kumar et al. [15] are 266,430, 225,205, 225,600 & 158,542 bits and 27.71, 28.72, 27.32, & 29.88 dB, respectively. The average embedding capacity and PSNR offered by our scheme, Chuang et al. [6], Ou et al. [26] and Kumar et al. [15] are 274,864, 237,457, 236,940 & 413,207 bits and 26.91, 27.93, 26.33 & 28.66 dB, respectively at the threshold values $\text{thr1} = 50$ and $\text{thr2} = 80$. From Table 1 and the average results, it is evident from Table 1 that our method outperforms [6, 15, 26] in terms of embedding capacity. However, it does not perform



(a) Lena



(b) Baboon



(c) Plane



(d) Peppers



(e) Boats



(f) Barbara



(g) House



(h) Houses



(i) Zelda



(j) Clown



(k) Kiel



(l) Lighthouse

Fig. 2 Twelve cover images of size 512×512 pixels (a). Lena (b). Baboon (c). Plane (d). Peppers (e). Boats (f). Barbara (g). House (h). Houses (i). Zelda (j). Clown (k). Kiel (l). Lighthouse

as better as Ou et al. [26] in terms of the image quality with a little difference in the PSNR values.

Table 1 Comparison of PSNR and capacity of the proposed and existing Ou et al. [26] Chuang et al. [6], Kumar et al. [15] methods using various threshold values

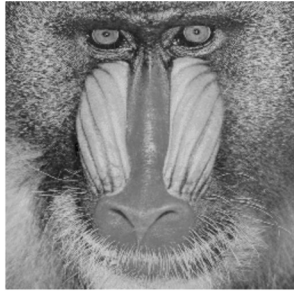
Methods	Metrics	Lena	Baboon	Plane	Peppers	Boats	Barb	House	Houses	Zelda	Clown	Kiel	Lighthouse
Threshold values: thr1 = 10 and thr2 = 30													
Proposed method	PSNR	31.05	25.98	30.31	31.82	30.04	28.40	31.71	28.32	32.82	28.12	26.24	28.27
	Capacity	215,856	91,588	215,300	214,136	102,360	196,272	173,452	249,276	250,756	185,320	158,928	136,968
Ou et al. [26]	PSNR	32.67	26.92	31.91	33.36	31.32	29.22	33.07	29.88	34.25	31.52	27.86	28.89
	Capacity	172,579	50,374	178,099	183,529	148,609	121,759	181,451	136,774	171,315	136,774	162,355	165,523
Chuang et al. [6]	PSNR	32.03	26.85	31.54	32.37	31.04	29.01	33.11	28.78	32.98	31.03	26.92	28.53
	Capacity	166,608	36,256	172,496	178,288	141,040	112,400	175,436	128,416	167,849	128,416	130,134	143,276
Kumar et al. [15]	PSNR	33.11	28.69	31.98	33.20	31.02	29.84	32.89	33.52	34.01	29.45	27.56	30.45
	Capacity	102,102	100,002	100,701	101,010	102,001	101,001	101,278	104,045	104,453	104,578	104,545	105,845
Threshold values: thr1 = 20 and thr2 = 50													
Proposed method	PSNR	30.72	25.03	29.97	31.12	29.79	27.64	31.01	27.36	32.12	27.23	25.56	27.34
	Capacity	256,572	155,404	246,284	260,000	218,420	247,140	219,372	271,440	280,628	237,516	202,592	208,184
Ou et al. [26]	PSNR	31.78	26.58	31.34	32.42	30.62	28.23	32.85	29.04	33.89	30.61	27.66	28.01
	Capacity	216,529	106,534	209,794	224,989	191,914	158,794	210,946	194,354	221,107	194,944	173,921	178,865
Chuang et al. [6]	PSNR	30.43	26.11	30.39	30.78	29.63	28.19	32.45	28.07	31.22	29.30	26.66	27.89
	Capacity	213,488	96,160	206,304	222,512	187,232	151,904	209,864	190,254	217,009	190,464	190,123	206,578
Kumar et al. [15]	PSNR	32.97	28.56	31.64	32.97	30.87	29.08	32.19	32.56	33.31	28.56	27.18	30.02
	Capacity	132,818	105,818	131,685	146,874	148,061	121,869	127,198	116,209	124,325	146,774	138,209	117,061
Threshold values: thr1 = 30 and thr2 = 60													
Proposed method	PSNR	30.12	24.86	29.48	30.59	29.11	26.97	30.62	26.72	31.78	26.86	26.89	26.79
	Capacity	271,124	190,784	258,868	271,804	253,704	269,012	245,848	280,920	289,948	258,504	225,716	234,760
Ou et al. [26]	PSNR	30.91	26.02	30.75	31.64	29.69	26.99	31.92	28.75	32.01	29.70	27.20	27.50
	Capacity	234,004	141,919	223,039	238,969	217,264	187,879	230,769	219,354	238,764	219,169	186,139	198,139
Chuang et al. [6]	PSNR	29.06	25.04	29.34	29.58	28.05	26.83	31.04	27.45	30.83	27.85	27.10	27.15
	Capacity	232,128	133,904	220,432	237,424	214,272	182,928	227,896	216,504	231,040	216,304	201,243	218,767
Kumar et al. [15]	PSNR	32.68	28.39	31.25	32.62	30.54	28.41	31.99	32.12	32.97	28.19	26.81	29.77
	Capacity	157,370	109,198	154,269	158,678	173,345	143,741	133,674	125,689	133,645	167,762	141,333	133,637
Threshold values: thr1 = 40 and thr2 = 70													
Proposed method	PSNR	29.55	24.55	28.97	29.85	27.71	26.74	30.06	26.02	30.87	26.01	25.97	26.29
	Capacity	280,796	219,452	268,052	278,400	266,868	281,504	262,884	285,932	292,840	272,680	241,616	246,140
Ou et al. [26]	PSNR	30.02	25.17	29.97	30.88	28.83	27.22	30.87	27.65	31.79	28.68	26.65	27.01
	Capacity	245,569	173,974	233,329	247,219	231,544	207,904	246,789	236,348	246,543	236,534	194,567	202,144
Chuang et al. [6]	PSNR	27.78	23.58	28.06	28.50	26.76	25.47	30.16	26.78	29.46	28.41	26.21	26.76

Table 1 (continued)

Methods	Metrics	Lena	Baboon	Plane	Peppers	Boats	Barb	House	Houses	Zelda	Clown	Kiel	Lighthouse
Kumar et al. [15]	Capacity	244,464	168,096	231,408	246,224	229,504	204,288	240,765	234,859	239,429	234,832	206,545	226,787
	PSNR	32.16	28.18	31.14	32.23	30.26	28.18	31.84	30.22	32.06	27.84	26.29	28.17
Threshold values: thr1 = 50 and thr2 = 80													
Proposed method	PSNR	28.85	23.50	28.07	29.04	26.90	25.97	29.51	25.11	29.62	25.03	25.38	26.01
Ou et al. [26]	Capacity	286,100	244,616	274,148	282,612	275,000	288,232	273,744	290,224	294,288	280,384	253,556	255,468
	PSNR	29.30	24.11	29.20	30.29	28.00	26.31	30.04	27.05	30.24	27.96	26.11	26.56
Chuang et al. [6]	Capacity	252,154	204,319	240,709	251,854	241,834	223,534	251,022	245,847	253,215	245,854	215,687	223,456
	PSNR	26.80	22.00	26.93	27.66	25.58	24.17	29.98	26.21	29.07	25.42	25.89	26.34
Kumar et al. [15]	Capacity	251,488	200,464	239,280	251,168	240,480	220,960	247,802	244,687	248,654	244,768	218,967	234,567
	PSNR	31.01	26.06	30.01	31.32	29.11	28.05	30.21	29.05	30.02	26.14	25.14	27.87
	Capacity	229,827	135,414	222,881	235,277	203,652	160,203	1,604,450	1,512,578	160,676	172,475	160,456	160,601



(a) Lena



(b) Baboon



(c) Plane



(d) Peppers



(e) Boats



(f) Barb



(g) House



(h) Houses



(i) Zelda



(j) Clown



(k) Kiel



(l) Lighthouse

Fig. 3 Stego-images for the proposed method (a). Lena (b). Baboon (c). Plane (d). Peppers (e). Boats (f). Barb (g). House (h). Houses (i). Zelda (j). Clown (k). Kiel (l). Lighthouse

For the standard grayscale images such as Lena, Peppers, Boats, and Baboon, we have demonstrated the results for capacity versus PSNR are shown in Fig. 4. The figure makes it

clear that the proposed method has greater data hiding capacity while preserving the quality of the marked image to an extent with respect to other relevant methods such as Chuang et al. [6], Ou et al. [26], Huang et al. [9], and Hong [7] because of efficient utilization of quantization levels for hiding the secret information. It can also be inferred from Fig. 4a–d that the proposed scheme is capable of maintaining the quality of the marked image while increasing the embedding capacity in contrast to other related methods i.e., as Chuang et al. [6], Ou et al. [26], Huang et al. [9], and Hong [7]. Figure 4a shows the capacity in bits with respect to the PSNR in dB for the existing Chuang et al. [6], Ou et al. [26], Huang et al. [9], Hong [7] methods, and the proposed method. The experimental results that the proposed method gives better embedding capacity along with the image visual quality. Figure 4b shows the capacity in bits with respect to the PSNR in dB for the existing Chuang et al. [6], Ou et al. [26], Huang et al. [9], Hong [7] methods, and our method. The experimental results that our method gives better embedding capacity along with the image visual quality. Figure 4c shows the capacity in bits with respect to the PSNR in dB for the existing Chuang et al. [6], Ou et al. [26], Huang et al. [9], Hong [7] methods and our method. The experimental results that our methods better embedding capacity along with the image visual quality. Figure 4d shows the capacity in bits with respect to the PSNR in dB for the existing Chuang et al. [6], Ou et al. [26], Huang et al. [9], Hong [7] methods, and the proposed method. The experimental results that our method gives better embedding capacity along with the image visual quality. The reason behind the

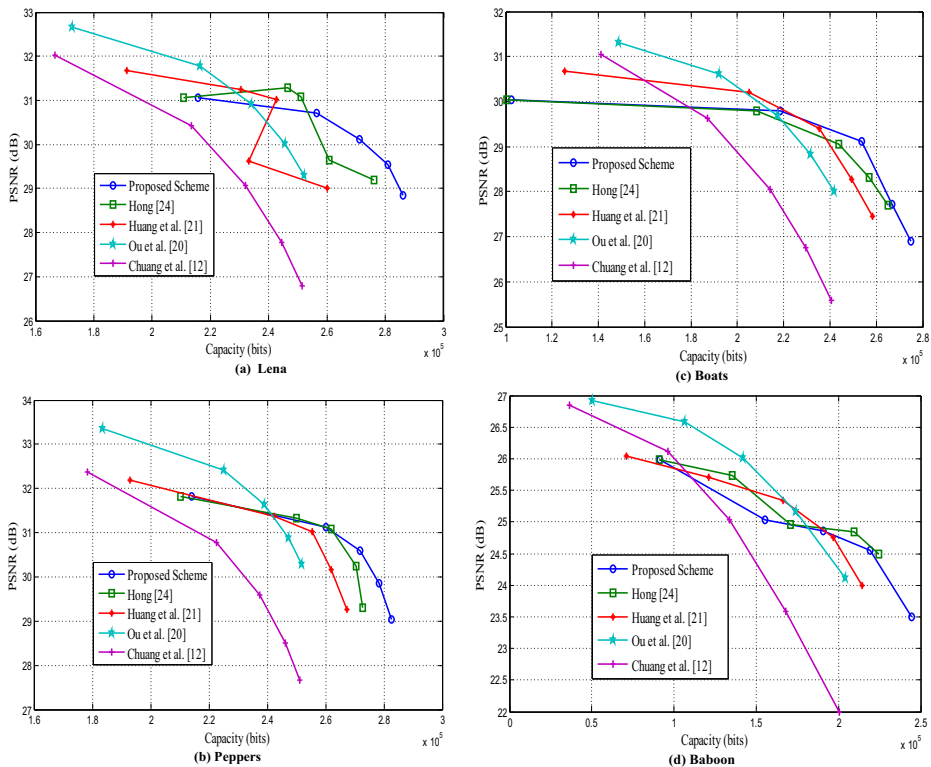


Fig. 4 Analysis of the proposed and exiting methods for test images (a) Lena, (b) Peppers, (c) Boats, and (d) Baboon

performance of our method is the optimal use of bit-plane and quantization levels for embedding the secret information.

On the basis of experimental results and above discussion, it is stated that the proposed scheme outshines other related methods as far as embedding capacity is concerned. For further analysis, the results at different thresholds for our scheme are viz-a-viz with its parent method [26] and provided in Table 1.

For further validation of the proposed method's performance, we also undertake a study on hundreds of images for two threshold values namely $thr1 = 20$ & $thr2 = 50$ and $thr1 = 40$ & $thr2 = 70$. However, to avoid repetition, the results have been shown only for forty-five images which can be seen in Fig. 5. The corresponding results of embedding capacity and PSNR with different threshold values are recorded in Table 2 for these images. The highest and lowest values of the embedding capacity are 288,716 & 294,320 bits and 54,692 & 65,108



Fig. 5 Cover images

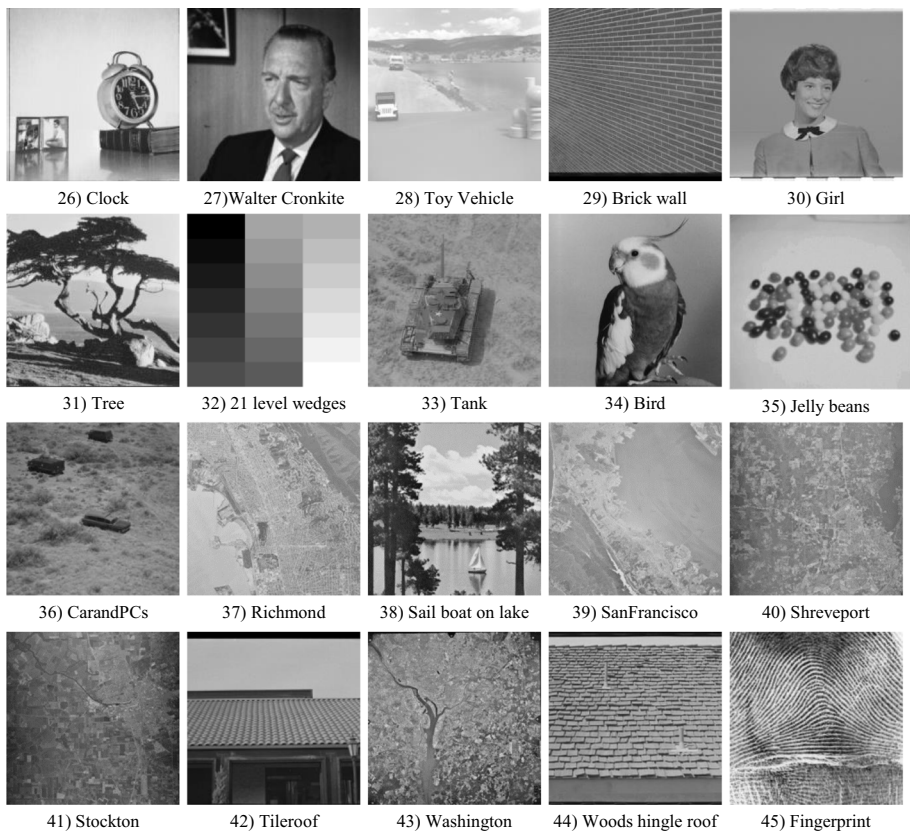


Fig. 5 (continued)

bits for $thr1 = 20$ & $thr2 = 50$ and $thr1 = 40$ & $thr2 = 70$, respectively. The highest and lowest values of PSNR are 42.97 & 40.87 dB and 20.72 & 20.17 dB for $thr1 = 20$ & $thr2 = 50$ and $thr1 = 40$ & $thr2 = 70$, respectively. The average embedding capacity and PSNR are 234,708 & 26,064 bits and 31.95 & 34.81 dB at the threshold values $thr1 = 20$ & $thr2 = 50$ and $thr1 = 40$ & $thr2 = 70$, respectively. From Table 2, it can be seen that our method has a high capacity while preserving the visual quality at the same time. This is possible because of additionally hiding two bits into every pair of quantization level, and four bits in the *Less_Complex* blocks.

7 Conclusion

This paper has proposed a new AMBTC based information hiding method using the LSB replacement technique. The proposed method has used two thresholds (more specifically, two even valued thresholds) for marking the AMBTC trios into three different categories based on their texture. Next, the bits of the secret information payload are adaptively concealed inside the bit-plane based on the marking of the trio category. Additionally, the LSB substitution-

Table 2 Evaluation of image quality & embedding capacity of the proposed method at different values of two thresholds

So. No.	Images	PSNR		Capacity	
		Threshold values		Threshold values	
		thr1 =20 and thr2 =50	thr1 =40 and thr2 =70	thr1 =20 and thr2 =50	thr1 =40 and thr2 =70
1	Moon surface	35.55	34.81	288,716	294,136
2	Airplane1	37.39	35.91	272,080	277,728
3	Splash	34.08	33.04	274,928	281,016
4	Truck	33.61	32.76	282,232	293,332
5	Tiffany	32.25	30.77	265,080	284,196
6	House1	28.56	26.96	202,220	245,344
7	Couple	27.86	26.55	222,028	272,744
8	Elaine	28.46	27.70	247,552	287,800
9	Man	27.30	25.97	211,924	260,948
10	Pixel ruler	25.26	23.28	115,420	153,424
11	Airplane U-2	37.79	36.49	284,136	290,100
12	APC	36.79	35.94	288,136	292,960
13	Airport	31.17	29.60	254,960	286,764
14	Stream & bridge	29.77	28.30	244,820	283,792
15	Aerial	29.02	27.44	230,916	275,464
16	Truck & APCs	31.73	30.68	271,680	291,916
17	Brodatz Grass	25.24	23.43	125,120	236,040
18	San Diego	29.86	28.44	247,824	288,144
19	Oakland	31.01	29.95	267,500	293,888
20	Pentagon	34.79	33.94	268,165	272,998
21	Woodland Hills	30.20	28.81	253,256	289,992
22	Earth from space	31.42	29.89	256,360	286,120
23	Hexagonal hole	20.72	20.17	54,692	93,688
24	Chemical plant	31.28	29.76	260,112	289,996
25	Resolution chart	28.75	27.98	76,808	87,040
26	Clock	33.48	31.67	249,920	267,176
27	Walter Cronkite	35.66	34.16	270,732	279,280
28	Toy Vehicle	37.88	36.73	253,140	258,324
29	Brick wall	30.36	28.46	230,680	285,496
30	Girl	36.65	35.35	242,312	248,300
31	Tree	30.63	28.85	235,800	267,732
32	21 level wedge	42.97	40.87	65,472	67,108
33	Tank	34.99	34.18	287,072	294,320
34	Bird	32.67	32.08	235,420	247,486
35	Jelly beans	36.60	34.88	200,092	209,568
36	CarandPCs	30.77	30.06	274,024	293,152
37	Richmond	32.33	31.51	275,760	291,632
38	Sail boat on lake	30.31	28.64	240,132	273,980
39	SanFrancisco	35.31	34.43	269,692	277,492
40	Shreveport	33.05	32.37	282,772	293,960
41	Stockton	34.02	33.36	285,732	294,060
42	Tilerof	31.88	29.99	241,336	272,504
43	Washington	29.33	27.80	238,452	285,892
44	Woods hingle roof	27.52	25.55	180,876	244,176
45	Fingerprint	31.63	29.85	235,800	267,732
Average		31.95	34.81	234,708	260,643

based method is utilized to conceal bits of the secret payload in the quantization levels of every trio. Thus, the proposed information hiding method achieves a decent increase in the data

embedding capacity without compromising the PSNR. Further, the proposed LSB replacement method maintains the same level of compression ratio as the original AMBTC method has. Experimentally, it is also proved that the proposed information hiding method outperforms the existing related methods for capacity while providing comparable quality of marked images. In future work, embedding in the encrypted and compressed domain to provide privacy preserving covert communication for low bandwidth applications can be explored.

Data availability There is no external data used in this work. Other information will be available on request to Dr. Aruna Malik (arunacsrke@gmail.com).

Declarations

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

References

1. Abdulla AA, Jassim SA, Sellahehwa H (2013) Efficient high-capacity steganography technique. *Proceedings Volume 8755, Mobile Multimedia/Image Processing, Security, and Applications 2013*. (875508). <https://doi.org/10.1117/12.2018994>
2. Abdulla AA, Jassim SA, Sellahehwa H (2014) Secure steganography technique based on Bitplane indexes. In: *Proc. of 2013 IEEE International Symposium Multimedia*, <https://doi.org/10.1109/ISM.2013.55>
3. Chang C, Lin C, Fan Y (2008) Lossless data hiding for color images based on block truncation coding. *Pattern Recogn* 41:2347–2357
4. Chen YY, Chi KY (2017) Cloud image watermarking: high quality data hiding and blind decoding scheme based on block truncation coding. *Multimedia Systems* 25:551–563. <https://doi.org/10.1007/s00530-017-0560-y>
5. Chen J, Hong W, Chen T, Shiu C (2010) Steganography for BTC compressed images using no distortion technique. *Imaging Sci J* 58(4):177–185
6. Chuang J, Chang C (2006) Using a simple and fast image compression algorithm to hide secret information. *Int J Comput Appl* 28(4):329–333
7. Hong W (2018) Efficient data hiding based on block truncation coding using pixel pair matching technique. *Symmetry* 10(2):1–18
8. Hu YC, Chang CC (1999) Quadtree-segmented image coding schemes using vector quantization and block truncation coding. *Opt Eng* 39(2):464–471
9. Huang YH, Chang CC, Chen YH (2016) Hybrid secret hiding schemes based on absolute moment block truncation coding. *Multimed Tools Appl* 76:6159–6174. <https://doi.org/10.1007/s11042-015-3208-y>
10. Kumar R (2019) Jung, KH. A systematic survey on block truncation coding based data hiding techniques. *Multimed Tools Appl* 78:32239–32259. <https://doi.org/10.1007/s11042-019-07997-0>
11. Kumar R, Chand S, Singh S (2018) A reversible data hiding scheme using pixel location. *Int Arab J Inf Technol* 15(4):763–768
12. Kumar R, Chand S, Singh S (2018) An improved histogram-shifting-imitated reversible data hiding based on HVS characteristics. *Multimedia Tools Appl* 77(11):13445–13457
13. Kumar R, Kim D, Jung K (2019) Enhanced AMBTC based data hiding method using hamming distance and pixel value differencing. *J Info Sec Appl* 47:94–103. <https://doi.org/10.1016/j.jisa.2019.04.007>
14. Kumar R, Kumar N, Jung KH (2020) Color image steganography scheme using gray invariant in AMBTC compression domain. *Multimedia Syst Signal Process* 31:1145–1162. <https://doi.org/10.1007/s11045-020-00701-8>
15. Kumar N, Kumar R, Malik A (2021) Low bandwidth data hiding for multimedia systems based on bit redundancy. *Multimedia Tools Appl* 81:35027–35045. <https://doi.org/10.1007/s11042-021-10832-0>
16. Kumar R, Kumar N, Jung KH (2022) Enhanced interpolation-based AMBTC image compression using Weber's law. *Multimedia Tools Appl* 81:20817–20828. <https://doi.org/10.1007/s11042-022-12634-4>
17. Langelaar G, Setyawan I, Lagendijk R (2000) Watermarking digital image and video data: a state-of-the-overview. *IEEE Signal Process Mag* 17(5):20–46

18. Lee J, Chiou Y, Guo J (2013) A high capacity lossless data hiding scheme for JPEG images. *J Syst Softw* 86:1965–1975
19. Lema M, Mitchell O (1984) Absolute moment block truncation coding and its application to color images. *IEEE Trans Commun* 32:1148–1157
20. Li C, Lu Z, Su Y (2011) Reversible data hiding for BTC-compressed images based on bit plane flipping and histogram shifting of mean tables. *Inf Technol J* 10(7):1421–1426
21. Li F, Bharanitharan K, Chang CC, Mao Q (2015) Bi-stretch reversible data hiding algorithm for absolute moment block truncation coding compressed images. *Multimed Tools Appl* 75:16153–16171. <https://doi.org/10.1007/s11042-015-2924-7>
22. Lin I, Lin Y, Wang C (2009) Hiding data in spatial domain images with distortion tolerance. *Comput Standards Interfaces* 31(2):458–464
23. Lin C, Liu X, Tai W, Yuan S (2013) A novel reversible data hiding scheme based on AMBTC compression technique. *Multimed Tools Appl* 74(11):3823–3842
24. Lin C-C, Lin J, Chang C-C (2021) “Reversible data hiding for AMBTC compressed images based on matrix and hamming coding”, *electronics*, 10(3). <https://doi.org/10.3390/electronics10030281>
25. Malik A, Singh S, Kumar R (2018) Recovery based high capacity reversible data hiding scheme using even-odd embedding. *Multimedia Tools Appl* 77(12):15803–15827
26. Ou D, Sun W (2014) High payload image steganography with minimum distortion based on absolute moment block truncation coding. *Multimed Tools Appl* 74(21):9117–9139
27. Pan J, Li W, Lin C (2014) Novel reversible data hiding scheme for AMBTC-compressed images by reference matrix. *Multidiscipl Soc Networks Res* 473:427–436
28. Sun W, Lu Z, Wen Y (2013) High-performance reversible data hiding for block truncation coding compressed images. *Signal Image Video Process* 7(2):297–306
29. T-T Xia, J. Lin, C-C Chang and T-C Lu, “A novel adjustable reversible data hiding method for AMBTC-compressed codes using hamming distance”, *Int J Embed Syst*, Vol. 14, No. 4, pp. 313–323, 2021.
30. Zhang Y, Guo S, Lu Z, Luo H (2013) Reversible data hiding for BTC-compressed images based on lossless coding of mean tables. *IEICE Trans Commun* 96(2):624–631

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Nano-inspired smart medicines targeting brain cancer: diagnosis and treatment

Raksha Anand¹ · Lakhan Kumar¹ · Lalit Mohan¹ · Navneeta Bharadvaja¹

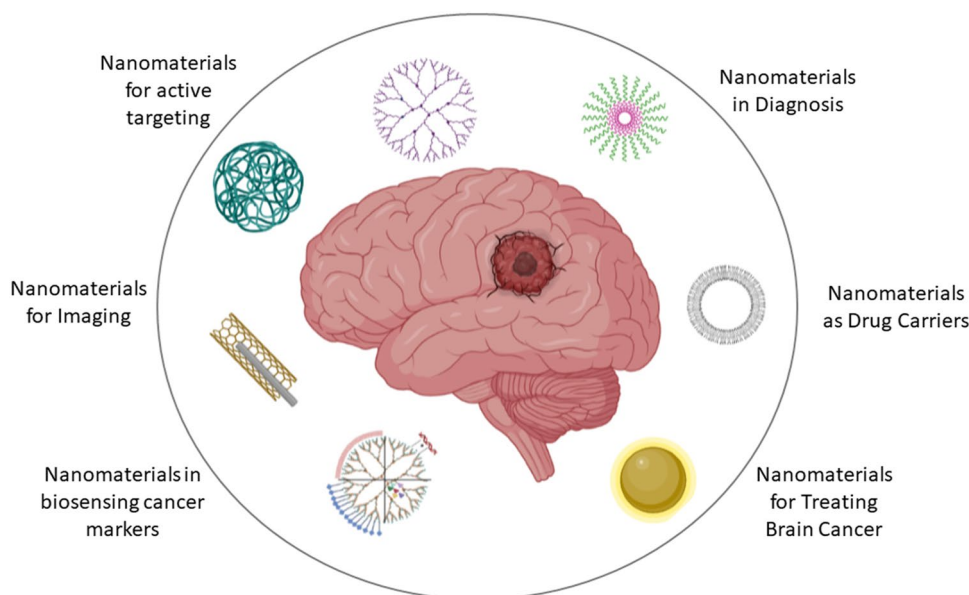
Received: 6 July 2022 / Accepted: 1 November 2022

© The Author(s), under exclusive licence to Society for Biological Inorganic Chemistry (SBIC) 2022

Abstract

Cancer, despite being the bull's eye for the research community, accounts for a large number of morbidity and mortality. Cancer of the brain is considered the most intractable, with the least diagnosis rates, hence treatment and survival. Despite the extensive development of therapeutic molecules, their targeting to the diseased site is a challenge. Specially tailored nanoparticles can efficiently deliver drugs and genes to the brain to treat tumours and diseases. These nanotechnology-based strategies target the blood–brain barrier, the local space, or a specific cell type. These nanoparticles are preferred over other forms of targeted drug delivery due to the chances for controlled delivery of therapeutic cargo to the intended receptor. Targeted cancer therapy involves using specific receptor-blocking compounds that block the spreading or growth of cancerous cells. This review presents an account of the recent applications of nano-based cancer theragnostic, which deal in conjunct functionalities of nanoparticles for effective diagnosis and treatment of cancer. It commences with an introduction to tumours of the brain and their grades, followed by hurdles in its conventional diagnosis and treatment. The characteristic mechanism of nanoparticles for efficiently tracing brain tumour grade and delivery of therapeutic genes or drugs has been summarised. Nanocarriers like liposomes have been widely used and commercialized for human brain cancer treatment. However, nano-inspired structures await their translational recognition. The green synthesis of nanomaterials and their advantages have been discussed. The article highlights the challenges in the nano-modulation of brain cancer and its future outlook.

Graphical Abstract



Keywords Brain cancer · Nanomedicine · Cancer theranostics · Nano-modulation · Green nanomaterials

Extended author information available on the last page of the article

Introduction

Out of all the ailments contributing to the worldwide morbidity and mortality of the population, Cancer leads the index. Cancer of the brain and the Central Nervous System (CNS) is the third most common type after breast and colorectal cancer. However, brain cancer still accounts for only 0.2% of the globally registered cases of cancer. Tumours from elsewhere in the body share characteristics among themselves; the brain's cancer is unique, attributing to the organ it perches. The countries with higher Human Development Index (HDI) have been studied to have higher mortality rates [1].

The damaged brain tissues, called brain lesions, can result from encephalitis, injuries or strokes, arterio-venous-malformations, and potential tumours. These tumours of the brain could be cancerous, i.e., with malignancy, or simply benign. Depending upon the area of their origin, brain tumours can be categorized as primary or secondary brain tumours. For the primary brain tumours, the site of origin would be the Central Nervous System (CNS), like a glioma [2]. In contrast, the secondary tumours arise in other body parts, ultimately metastasizing to the brain or the nearby tissues. Colon cancer, melanoma, and breast cancers are some of the common cancers that spread to the brain [3]. Brain tumours are named as per their location or the type of cells they are made of. For example, medulloblastoma is the tumour of the cerebellum, whereas Glioma is the tumour arising from the glial cells [4].

The otherwise selectively permeable Blood–Brain Barrier (BBB), upon signalling alterations induced due to brain cancer, transmutes. The leakiness, which is the overall increase in the permeability of the tumour, is accessed by the rise in the concentration of biological factors in the region of the tumour as compared to the adjacently flowing blood [5].

Brain tumours or the tumours of CNS do not occur in stages. However, they have been classified into four different grades depending on their malignancy. The properties

of each grade are listed below. For example, the most frequently occurring malignant brain tumour, the Glioma, has four defined grades, as depicted in Fig. 1. Grades I and II are considered low grades, whereas grades III and IV classify as high-grade tumours as per World Health Organisation (WHO) [6].

Grade I: Non-infiltrative type Pilocytic astrocytoma.

Grade II: Slightly infiltrative type Diffuse astrocytoma.

Grade III: Infiltrative type Anaplastic astrocytoma, and,

Grade IV: Readily infiltrative type Glioblastoma multiforme.

The variation in the incidence rate of these tumours varies with gender, age, and community. The mortality rate due to brain and CNS tumours is highest in areas with the highest incidence. Some of such incidence-based observations have been summarized below:

- Considering the overall incidence of Brain/CNS tumours, its higher in females, specifically for non-malignant type. Whereas the malignant type of tumour has a higher incidence in males [7].
- In younger children aged between 0 and 4 years, a higher incidence of solid, malignant brain/CNS tumours is seen. Whereas, with an increase in age, people over 40 are at a higher risk of developing such cancers. For non-malignant tumours, with an increase in age, the incidence has been found to increase. Glioblastomas are the most common type of non-malignant tumour seen in adults. Hence, brain cancers are considered the most common type of lethal solid-paediatric tumours [8].
- Depending on the race, the incidence of these tumours varies. The black community is more prone to benign types, such as non-malignant meningioma. Whereas malignant tumours like Gliomas have a high incidence

Fig. 1 Grades of brain cancer

Grade I	Grade II	Grade III	Grade IV
<ul style="list-style-type: none"> • Slow growing, benign type • Do not spread, and can be cured with surgery • Limited to children 	<ul style="list-style-type: none"> • Slight abnormal morphology • Chances of recurrence and spread in nearby healthy areas • Frequent in younger aged 	<ul style="list-style-type: none"> • Malignant cells with abnormal morphology • Quick growth of cells seen • Low prophecy • BBB might be damaged 	<ul style="list-style-type: none"> • Malignant, aggressively growing cancer cells • Significant necrosis and angiogenesis spotted

in the white community. However, brain cancer is most commonly seen in the States [7].

Nanomaterials promote biological interactions owing to their size and structures. Nanotechnology also provides opportunities to alter drug pharmacokinetic and biodistribution properties, along with limited/controlled release at the target site. Chemically and biologically modified nanoparticles are also effective in delivering siRNA to the tumour region of the brain [9]. These selected nanoparticles must be cyto-compatible, permeable through BBB, intensively selective, specific, and biocompatible [6]. These carriers would be considered effective only if they would protect the cargo from degradation on their way to the target and could penetrate the target efficiently, releasing the drug in a controlled and non-toxic manner. The payload must offer the least toxicity or aggregation in the non-targeted sites [10]. The nano-composites, combinations of organic and inorganic nano-components, are effective “Cancer-Nano-Theranostics” [11, 12].

The current review summarizes the latest developments in diagnosing and treating brain cancers/ tumours using nanomaterials. Also, it focuses on the specific targeting of tumour tissues using various surface-targeting molecules such as antibodies and aptamers. This review also describes various nanostructures that can be used as a carrier for transporting drugs across the blood–brain barrier to treat brain cancers. Lastly, the article provides insight into the recent developments in the field of green-technology-based nanomaterials for targeting brain tumours and the advantages of shifting towards green technology approaches.

Drug transport across the blood–brain barrier (BBB)

The Blood–Brain Barrier (BBB), a component of the neurovascular unit (NVU), shields the inner CNS from direct contact with the blood, tightly regulating the transport between the brain tissues and the bloodstream [13]. It is responsible for maintaining the homeostasis and microenvironment of the brain and the surroundings [14]. BBB consists of tightly connected brain capillary endothelial and microvascular cells. Another multi-functional class, called mural cells, is responsible for controlled capillary contractions and hence angiogenesis and healing of wounds [12]. The BBB comprises endothelial cells, pericytes assorted basal membrane, transporters, astrocytes, and immune cells [4]. The increased levels of growth hormone fenestrate the endothelial layer. Tight junctions are hampered. This allows the previously blocked substances to cross the barrier. This adapted permeability of BBB can be exploited for treating brain cancer [15]. Receptors and transporters on the BBB allow selected

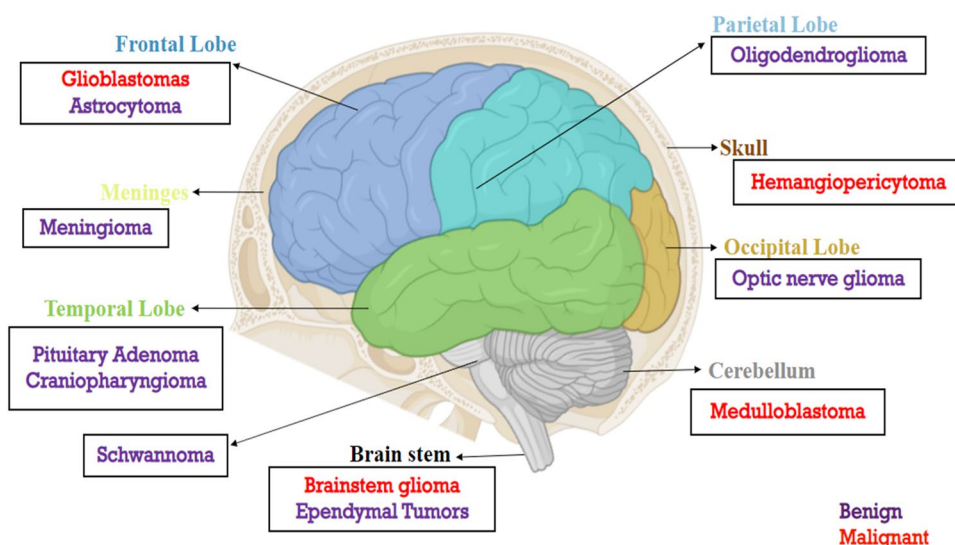
antibodies and molecules to pass through. These peptides and antibodies have been tagged on the surface of the nanoparticles for the BBB receptors and transporters to recognize and allow the penetration of designed nanoparticles across the barrier. This adsorption, recognition, and penetration mechanism is called “receptor/transporter-mediated transcytosis” [16]. These nanoparticles now have to reach their designated site of action.

Surface-modified nanoparticles have been found to penetrate the BBB expertly. TAT, the cell-penetrating peptide or albumin-coated nanoparticles, having a high positive surface charge, attaches electrostatically to the BBB and penetrates through it [17]. Gold nanoparticles coated with glucose in the range of 4 nm were found to reach astrocytes, effectively crossing the BBB [13]. Introducing some cell-surface receptors like fibroblast growth factor-inducible 14 (Fn14) targets nanodrugs across the brain barriers. Similar nanodrugs targeting glioblastoma multiforme with Decreased Adhesivity and Receptor Targeting (DART-NPs) have been developed. These imparted high-target binding and reduced non-specific interactions with off-target proteins [18]. Some of the highly prevalent types of brain cancer, along with their localization have been illustrated in Fig. 2.

Nano-therapeutic approach for brain cancer

The field of nanotechnology dedicated to medical resolutions contributes to Nanomedicine [13]. The NP-based carrier has therapeutic cargo, a functional group (probe or reporter), and a target vector [5]. It has found its application in the efficient diagnosis and treatment of tumours of the brain. Nano-therapeutic particles, upon systemic administration into the system, can cross the Blood–Brain Barrier (BBB) as well as the Blood–Brain-Tumour Barrier (BBTB) and hence can deliver the complex molecular cargo to the target site [19–21]. In the case of metastatic brain tumours, the BBB disrupts and leads to the formation of the Blood-Tumour barrier. These “Smart NPs” have been designed for targeted delivery aiming at anti-angiogenesis, gene therapy, etc. Nanocarriers can overpower cancer resistance [13]. In addition to the previously described characteristics, the nanomaterials offer high mobility and quantum effect due to electron movement. There are several ways in which nanomedicine has waived off the limitations that traditional therapy offers. Due to the Enhanced Permeability and Retention (EPR) effect of tumours, they can accumulate higher amounts of nanoparticles than normal tissues. This effect is accompanied by leakiness of vascular and substandard lymphatic drainage. This leads to the accumulation of nanomaterial in the tumour [22]. However, in the case of brain tumours, a higher accumulation of nanoparticles due to the EPR effect

Fig. 2 A representation of some common brain tumours as per their locations in the human brain



is sometimes confused with general increasing tumour size. Such candidates which are non-toxic, easily manufacturable, bio-compatible and degradable, stable in blood, non-immunogenic and non-allergic, and easily modifiable as the target but stay stable upon administration are preferred as nanomedicine [23]. Other nanotechnology-based strategies include targeting the tumour's intra- or extracellular space [24]. Some of the nano-inspired therapies for brain cancer have been illustrated in Fig. 3.

The nano-carrier encapsulated drugs have the capacity to bypass the drug efflux pumps on the membranes of tumour cells, which otherwise limit the intracellular concentration of conventional drugs. This nano-assisted-drug delivery thus enhances the concentration and activity of anti-cancerous compounds in the tumour site [25]. The nanomedicines can cross the BBB, making the cargo delivery targeted to the brain possible. Since the nanomedicine formulations can be in the form of gels, there is a provision for the controlled

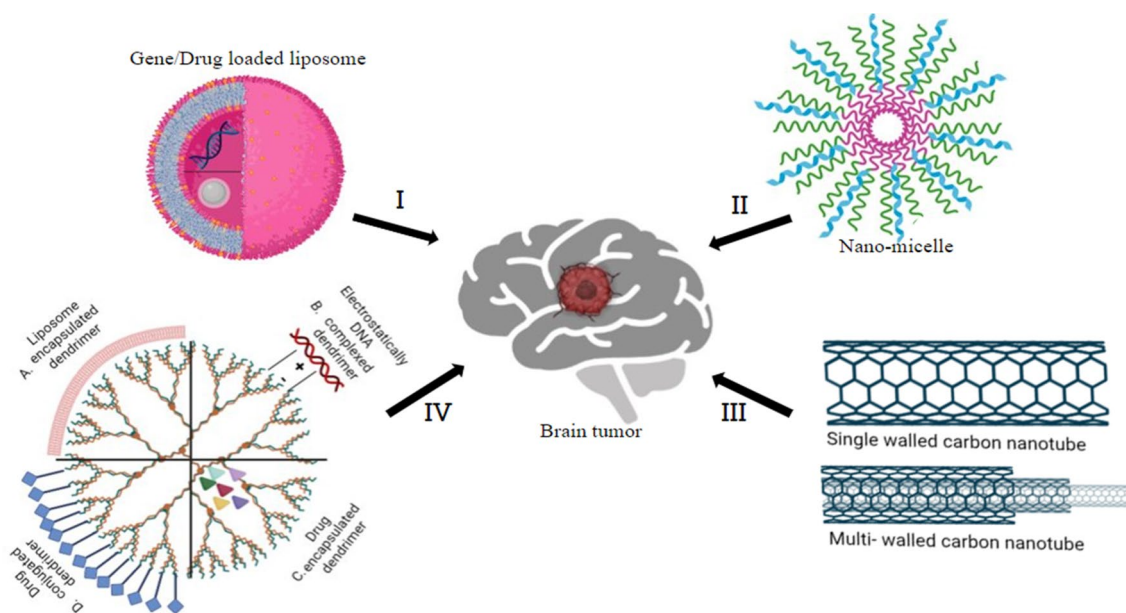


Fig. 3 Some common nanotherapeutic approaches for treating brain cancer. (I) Liposomes which can be loaded with therapeutic genes/drugs and targeted to the tumour site, (II) ligand-tagged polymeric micelle, (III) Single or multi-walled carbon nanotubes (SW-CNTs and MW-CNTs), which can be loaded with therapeutic cargo, and effectively delivered through the Blood Brain Barrier, (IV) Some of the

different types of dendrimers used in brain tumour treatment: **A** dendrimer encapsulated within liposome for targeted delivery; **B** DNA/nucleic acid electrostatically complexed over dendrimers; **C** simple dendrimers as drug carriers; and **D** drug/ligand conjugated on the surface of dendrimer against specific receptors)

release of a laden drug followed by its implantation, for example, in the intra-cerebral region. This promises an efficacious clinical response. Additionally, nanomedicine leads to a reduced dosage size and toxicity. These nanoparticles have been used as cytostatic delivery vehicles for anti-cancerous drugs [3]. Beyond surface tailoring, targeting nanoparticles to a specific cell type or organelle in the right amount is required for the least possible side effects. Hence, nanoparticles can be targeted to regulate functions like neuroprotection, immunomodulation, microenvironment remodelling, and improving oxidative stress, etc. [24]. The overall mechanism for nanomaterial-targeted brain cancer treatment can be explained as—Nanomaterial circulation in the blood, BBB receptor/transporter recognition, intracellular transport targeting the diseased cells and its internalization followed by the release of the therapeutic payload intracellularly at the target site [26].

Several factors affect the nanomaterial uptake, accumulation, and biodistribution in tumour cells. They are as follows: the size of nanomedicine, the composition and targeting, its shape and branching, its flexibility/rigidity, its core, and architecture [9]. There are two defined strategies for targeting nanoparticles in tumour cells, active and passive targeted delivery. Unlike the active targeting process, the passively targeted nanoparticles have no surface ligand for specific targeting. These take advantage of size, effectively lesser than 100 nm, by reaching the intra-tumoural space through the tumours' compromised hyper-vascular and lymphatic systems [12]. Hence, active targeting ensures a significant increase in the amount of therapeutic cargo being delivered. For example, glioblastoma cells overexpressing CD44 receptors have been actively targeted [27].

Nanostructures for diagnosis and biosensing of brain cancer

The brain tumour tracers earlier depended upon the disruption of the BBB, which could be easily detected by contrast-enhanced MRI or CT scans. Thus, it could be considered an efficient detection strategy for brain tumours with blemished BBB, for example, brain metastases and gliomas [28].

The pathophysiology exhibited by brain tumours is too complex to be imaged through conventional techniques. Active targeting of tumour cells through nanoparticles provide scope for sensitive and high-resolution imaging of brain tumour and their associated microenvironment. Biocompatible nanoparticle probes have been designed with the ability to cross the Blood–Brain Barrier.

Before the emergence of nanotechnology, brain cancer, and tumours diagnosis were done with the help of techniques such as positron emission tomography (PET), photoacoustic imaging (PA), computed tomography (CT), optical imaging, fluorescence imaging (FL). Gadolinium (Gd) chelates

administration enables the imaging and determination of the outline of the tumour during surgeries using magnetic resonance imaging (MRI) [29]. The only drawback of this method is the short half-life of Gd, due to which its frequent administration is required to maintain its level in the blood. Intraoperative ultrasound is one of the non-optical methods to procure integrated images of brain tissues but is insufficient for detecting superficial or minute tumours of the brain. Several neurophotonic techniques involving thermal imaging, Raman spectroscopy, fluorescence spectroscopy, and optical coherence tomography are currently being used for elucidating brain tumours and cancers [30]. These techniques lacked sensitivity, accuracy, and specificity. With the advancements in the field of nanotechnology, biosensing and bioimaging have gained heights in terms of research and advancements. Nanotechnology has enabled us to obtain clinical facts with high accuracy and precision without the use of invasive techniques [31]. The development of nano-chip/nanoarrays has enabled easy diagnosis and imaging in the case of brain tumours as they have combined optics, magnetism, and electric properties. With advancements in the field of nanotechnology, nanomaterials are proving to be highly potent candidates for imaging and diagnosis of tumours and cancers of the brain [30].

Carbon dots (CDs)

Carbon dots (CDs) have tunable optical characteristics, making them excellent agents for bioimaging. Their size is in the range of 10 nm, which enables facilitated penetration into the neural barrier. Carbon dots (CDs) are a category of carbon-based nanomaterials with high crystallinity. CDs can be categorized into two types: (i) carbon quantum dots (CQDs) which possess excitation-independent luminescence and (ii) CDs with excitation-dependent luminescence. The very first use of CDs for cellular imaging was done during 2011 [32, 33]. CDs are found to be one of the best candidates for tumour theranostic because of photostability, enhanced water solubility, adjustable fluorescence emission and excitation, and biocompatibility [34, 35].

Since the first use of CDs for imaging, CDs have emerged as the best imaging tool for the early detection of brain tumours. Du et al., reported the use of Gd³⁺-polymer-loaded CDs for MR-fluorescence-based imaging of gliomas [36]. In a study conducted by Zhao et al., red-light emitting CDs (R-CDs) were employed for the imaging of deep-situated brain glioblastomas (GBM) through a liposome-mediated delivery system in mice model [37].

Zheng et al., reported the synthesis of CD-Asp, a novel CD synthesized via pyrolysis of D-glucose and L-aspartic acid. These CD-Asp demonstrated high contrast fluorescent images in vivo after tail vein injection into mice model. Enhanced fluorescent signals were detected in the site of

glioma as compared to normal portions indicating that these novel CD-Asp can cross BBB easily and can target C6 glioma cells without the use of another targeting molecules [32]. Fan et al., reported the synthesis of a novel “pH-responsive fluorescent graphene quantum dots (pRF-GQDs)”. These pRF-GQDs have been shown to pose minimal toxicity and have been demonstrated to display a sharp fluorescent transition at pH 6.8 (pH that corresponds to the acidic extracellular microenvironment of solid tumours) between green and blue [38]. A study conducted by Li et al., demonstrated the use of CDs functionalized by multiple paired α -carboxyl and amino groups to facilitate the binding of these CDs to large neutral amino acid transporter 1 expressed in most tumours. These CDs were demonstrated to selectively accumulate in human tumour xenografts in orthotopic mouse model of human gliomas [39].

Wang et al., demonstrated the green synthesis of fluorescent nitrogen-doped CDs using hydrothermal heating of milk. The resultant N-CDs were used for imaging human brain glioma cancer cell line, U87 cells [40]. In a study Ruan et al., have reported the synthesis of CDs using heat treatment of glycine. The resultant CDs were demonstrated to have high serum stability and low cytotoxicity for facilitated in-vivo imaging. These CDs were shown to be taken up by C6 glioma cells in vitro in a time dependent and concentration dependent manner which enabled enhanced imaging [41].

Magnetic NPs

Magnetic NPs (MNPs) is one of the best probes for imaging and diagnosis of cancer due to their exclusive size and properties. Some examples of MNPs are “superparamagnetic iron oxide NPs” (SPIONs) and “ultra-superparamagnetic iron oxide NPs” (USPIONs) [42]. These MNPs are highly stable, susceptible to the magnetic field, and possess desired physical properties for imaging brain cancer and tumour. For imaging of glioma, a study has reported the use of “folic acid (FA)-conjugated bovine serum albumin (BSA)-coated SPIONs.” In the case of glioma U251 cells, FA-BSA-SPIONs have demonstrated high cellular uptake and biocompatibility [43].

Metallic NPs

Gold nanoparticles (AuNPs) have previously been investigated to be used for various rheumatological disorders. A number of enzymes and ligands were attached to the surface of AuNPs for cancer diagnosis and immune analysis. AuNPs possess certain optical properties of plasmon resonance, enabling the imaging of biological disorders [44]. AuNPs can easily be detected through CT, MR, and FUS imaging techniques. AuNPs, when administered

intravenously, are proven to enhance the fluorescence intensity in the case of FL imaging compared to when administered directly in tumour tissue. FL imaging provides better visualization of glioma through intravenous infusion of AuNPs [45]. Presently, various bio-compatible templates have been under study to be used along with AuNPs for the imaging of glioma. A study was performed on a rat glioma model with synthesized AuNPs functionalized with chlorotoxin peptide and labelled using ^{131}I . The functionalized AuNPs exhibited cytocompatibility and the property of X-ray attenuation which could cross the model's BBB [46].

Quantum dots

Quantum dots (QDs) are an emerging pulp for diagnosing and imaging cancer cells due to their flexibility and tunable properties, enabling its use along with fluorescence, with significant Stokes shifts and narrow emission bands for enhanced imaging [47]. QDs can also be used for mapping brain abnormalities during surgery. A recent study used PEG-coated QDs as nanoprobe for glioma imaging. These QDs were integrated with asparagine-glycine-arginine peptides (NGR) targeted towards CD13 glycoprotein in tumour cells. It was observed that NGR-PEG-QDs could attach and target CD13 present on glioma tissues in in-vivo settings [48]. A study conducted by Wang et al., reported the use of a new near-infrared window (NIF-II) fluorescent agent, graphene quantum dots doped with nitrogen and boron (N-B-GQDs). These N-B-GQDs have been demonstrated to be highly stable in serum and to be highly photostable. N-B-GQDs have been used for imaging glioma xenograft in mouse model [49].

Polymeric nano-vehicles

Due to their biodegradability and biocompatibility, polymers are in demand to enhance MR imaging in gliomas. In a recent study, red fluorescent carbonized polymer dots were used for real-time imaging during surgery [29]. The nano-polymer used had soaring internalization capacity in the case of glioma cells and exhibited low toxicity for the neighbouring cells, long excitation wavelength, and high photostability. Semiconducting polymers have also been prepared using a fluorination strategy to produce fluorescence to diagnose tumours in the brain accurately [50]. Compared to non-fluorinated counterparts, bright near-infrared-designed polymers yield enhanced images up to three-fold. Nanostructure coordination polymers are another polymeric nano-vehicle with low toxicity to healthy cells, high stability, and high contrast ability when linked to paramagnetic iron moiety [51].

Multimodal imaging using nanomaterials

Multimodal imaging is currently being investigated for better and more precise imaging of brain gliomas. Nanomaterials enable club diagnostics and therapeutics together for better performance [52]. In recent times, multimodal imaging has been a promising technique based on the active targeting of targeted brain tumour tissues by employing aptamers or ligands to make imaging more sensitive. Furthermore, MR imaging can also be integrated with these diverse multimodal imaging techniques [53].

Extracellular vesicles and exosomes

Gliomas require susceptible, non-invasive, and precise methods of prognosis and diagnosis. Extracellular vesicles carry forward molecules from the parent cells to the designated places. When coated with nanoparticles, these vesicles cross BBB through the transcytosis process [54]. Extracellular vesicles present in gliomas are known to express EGFR proteins whose presence can be used to accurately diagnose malignant tumours [55]. Thus, extracellular vesicle-based brain tumour diagnosis can be a reliable and an accurate diagnosis option. On similar grounds, exosomes are also nano-sized extracellular lipid bilayer vesicles. Exosomes are known to express both coding and non-coding RNAs along with specific lipids, which can be exploited to be used as a diagnostic tool that can cross BBB [56].

The brain tumours have been accessed using nanoparticles to visualize lesions and off-site characterization. For example, contrast agents under the brand names Dotarem[®]; Gadovist[®]; and Gadavist[®]- Enhanced MRI have been practiced [57].

Nanostructures for brain cancer treatment

The typical tumour treatment dealt with the surgical resection of the abnormal cell mass. However, in the case of the most common type of brain cancer, glioma, a malignant type, there is no confirmation of complete removal of cancerous cells from the body. Hence, effective treatment processes had to be identified. The median survival duration of the patient could be increased up to a year by involving radio- and chemo-therapies. Technological advancements allow for the real-time profiling of brain tumours, and robust screening of the tumour's genomic, transcriptomic, proteomic, metabolomic, and epigenomic condition [4]. The currently available and accepted standard-of-care treatments for brain cancer include irradiation, chemotherapy, and surgery. However, the diagnosis and treatment of brain cancer have witnessed a paradigm shift owing to nanomedicine formulations. There has been considerable development in the field of nanotechnology, which enables the development of

certain nanostructures that could aid in treating brain cancers. Some of the most commonly used nanostructures for treatment have been discussed below. Table 1 summarizes some approved/studied nanodrugs and nano-gene-therapy for human brain cancer treatment.

Gold nanoparticles (AuNPs)

AuNPs possess various attributes involving biocompatibility, quickly modifiable shapes and sizes, ease of being conjugated with different functional moieties making them one of the best candidates for theragnostic applications [30]. AuNPs are known to exhibit cytotoxic effects caused due to oxidative stress. AuNPs possess mono disparity, large surface area, ease of fabrication, binding ability to various biomolecules, and diagnostic properties [58]. Various therapeutic molecules or drugs can be loaded onto AuNPs with the help of covalent bonding or electrostatic force of attraction. The versatility in the size of AuNPs facilitates their free movement through the circulatory system. Thus, blood circulation can be used as a tool to target AuNPs directly toward the tumour cells [59]. “Carboxymethyl xanthan gum-coated AuNPs” (CMXG-AuNPs) were developed through microwave irradiation. These NPs were eco-friendly, non-toxic, and cost-effective, as xanthan gum was used for the biosynthesis of AuNPs. Doxorubicin (DOX) was loaded onto these CMXG-AuNPs with the help of electrostatic forces. It was found that the anti-tumour efficacy of DOX, when encapsulated with CMXG-AuNPs, was about 4.6 times higher than free DOX. AuNPs conjugated with curcumin to enable pH-dependent release of curcumin, which in turn was friendly with the normal healthy cells while restricting the growth, proliferation and migration of glioma cells [60–63].

Silver nanoparticles (AgNPs)

AgNPs exhibit various properties in biological systems based on differences in shape, size, concentration of particles, route of administration, and specific target. Research on AgNPs has intensified recently due to their unique physico-chemical and biological characteristics and well-established anti-cancerous and anti-bacterial properties [64]. The anti-cancerous and anti-bacterial properties of AgNPs can be attributed to the release of Ag⁺ ions in the cells because of the destabilization of AgNPs, leading to the massive production of reactive oxygen species (ROS). ROS causes oxidative stress and damages various cellular components, including DNA, lipids, and proteins, ultimately leading to cell death [65]. Salazar-Garcia et al. demonstrated the deteriorating effects of AgNPs on C6 rat glioma cell lines after treating them with variable amounts of AgNPs for 24 h. Compared to the control, the viability of C6 glioma cells decreased by about 21% after AgNPs treatment [66]. Multifunctional

Table 1 Some of the approved/studied nanodrugs and nano-gene-therapy for human brain cancer treatment

Nano-drug (Name and type)	Cancer type	Function/Mechanism	Toxicity levels	References
DepoCyt Injectable liposome	Lymphomatous meningitis Meningeal leukaemia Leptomeningeal metastasis Neoplastic meningitis	Cytarabine, its active component, gets administered directly into the cerebrospinal fluid (CSF)	Arachnoiditis – Inflamed mid-layers of the membrane around the brain Momentary hike in protein and leukocytes levels of CSF Some cases of neurotoxicity	[75, 76]
Encapsulated Doxorubicin Immuno-conjugated liposome Available in the market as Myocet [®] ; Doxil [®] ; Caelyx [®]	Glioblastoma multiforme Other lesions and brain tumours	Anti-EGFR/PEG/Interleukin conjugated liposomes were introduced intravenously Liposomes should transport drugs toward the tumour and away from potential toxicity zones	Termed as having toxicity in a favourable range compared to other conventional treatments Better cardiac and myelo- safety	[77, 78]
SGT-53 p53 encoding cDNA with plasmid encapsulated in cationic liposome; anti-TfR conjugated	Recurrent glioblastoma Glioblastoma multiforme	Restoration of wtp53 apoptotic function; hence down-regulation of O6-methylguanine-DNA-methyl transferase (MGMT) Can be used in combination with standard chemotherapeutic agents	Currently being accessed for toxicity Undergoing clinical trials	[79–81]
AGuIX [®] NPs Activation and Guiding of Irradiation by X-ray; polysiloxane based nanoparticles	Brain metastases Glioblastoma	Tumour volume reduction Can cross the BBB These ultra-nanoparticles get accumulated into tumours by the EPR effect, and retention provides therapy	Currently being accessed for toxicity in humans Undergoing clinical trials	[82]

nano-platforms consisting of AgNPs and alisertib as drugs were prepared and conjugated with chlorotoxin, a targeting peptide against the MMP-2 receptor expressed in brain tumour cells. The effect and biodistribution of nanoparticles were examined using ^{90m}Tc , and a significant tumour reduction was found due to the developed AgNPs-alisertib-chlorotoxin conjugate [67].

Zinc-oxide nanoparticles (ZnO NPs)

Besides AuNPs and AgNPs, ZnO NPs have also been attributed to exhibit anti-cancerous activities and are inexpensive as compared to the other two [68]. ZnO NPs exhibit high catalytic activity, adsorption capabilities, biocompatibility, and enhanced electron-transfer kinetics [69]. Various ZnO nanostructures, including nano-nails, nanobelts, nanowires, nanobridges, nanoribbons, and nanotubes, have been synthesized, and these nanostructures have been reported to cross the BBB or reach the brain through neural transportation system after oral ingestion [70]. The presence of apolipoprotein E on the surface of ZnO NPs helps in crossing BBB by the NPs [71]. ZnO NPs cause energy deprivation, induce oxidative stress in the microglia cells and eventually lead to cellular damage [72]. Wahab et al. demonstrated cell growth inhibition and apoptosis by ZnO NPs in HeLa and U87 cell lines, which was dose-dependent. The cell death was also attributed to an increased micronuclei formation in U87 cells. Amongst different ZnO fabricated nanomaterials, nanosheets and NPs have been found most effective as an anti-tumour agent against HeLa and U87 cells [73]. ZnO NPs have also been reported to cause neurotoxicity in mature rat brain cells, followed by oral administration of 40 and 100 mg/kg ZnO NPs within seven days of ingestion [74].

Active target

Active targeting refers to slight modification on the surface of nanoparticles to decrease their uptake in the normal tissues and increase uptake and accumulation in tumour tissues. Active targeting of tumour cells involves targeting surface membrane molecules or proteins that are usually upregulated in the case of cancer cells [83]. The commonly employed targeting molecules include aptamers, antibodies and their fragments, oligopeptides or small molecules, transferrin, and folic acid. NPs coupled with these molecules can easily localize into targeted tumour cells, expressing complementary receptors or antigens on their cell surface and ultimately enhancing drug delivery [84]. Certain ligand-receptor interactions may also lead to receptor-mediated endocytosis, which further aids in the delivery of NPs to the target. Through active targeting, cancer therapies can be made more precise and efficient. Nanoparticles are subjected to surface modifications for enhanced uptake and

accumulation of the cargo in tumour cells. For this purpose, the surface membrane proteins upregulated in cancerous cells are targeted. There are common molecules, like antibody fragments, aptamers, and so on, which have been approved by the FDA to be used as targeting molecules in the treatment of cancer; however, none in the name of therapy for brain tumours [3].

Monoclonal antibodies

IgG is the most widely used monoclonal antibody for targeting. The antigen binding site amounts to only a fraction of the entire size composition of antibodies. F(ab')₂ fragments of antibody retain both antigen binding sites linked together with the help of disulfide linkage and can be used for targeting cancerous cells without increasing much of the overall size of therapeutic agents [9]. In certain tumours, such as breast cancer, there is an up-regulation of specific growth factors, such as HER2/neu, and can be effectively targeted through anti-HER2/neu antibodies [85]. Liposomes conjugated with antibodies against glial fibrillary acidic proteins have been studied for their ability to cross the BBB. Antibodies have also been developed to target transferrin receptors. These receptors are extensively expressed on the BBB endothelia, which mediates transcytosis [86].

Aptamers

Folded single-stranded oligonucleotides with a length of about 25–100 amino acids having the ability to bind to their molecular targets are known as aptamers [87]. Nanoparticles conjugated with aptamers exhibit an increased in vitro cytotoxicity compared to non-conjugated nanoparticles. In a recent study, “EpCAM-fluoropyrimidine RNA aptamer-modified doxorubicin-loaded PLGA-b-PEG nanoparticles” with the ability to bind to epithelial cells employing extracellular domain have been reported in the case of a lung cancer model and on similar lines aptamers can be designed for brain cancer targeting [9].

Small molecules

Various small molecules involving growth factors, receptor ligands, peptides, and carbohydrates can be used for targeting. Explicit examples of these include transferrin, folic acid (FA), and arginylglycylaspartic acid (RGD) peptides [88]. FA is necessary for cell survival and proliferation due to its role in DNA synthesis, repair, and methylation. Humans' folate receptors (FR) have a high affinity for FA. The expression of FR is minimal or negligible in the normal cells but is highly upregulated in the cases of brain, lung, ovarian, breast, and colorectal cancers [89, 90]. The covalent conjugation of molecules such as liposomes or other small

molecules to FA does not affect its binding ability with FR. It thus can deliver drugs to the target cells through endocytosis [91]. The iron-transporting glycoprotein transferrin (Tf) is known to provide iron to cells employing receptor-mediated endocytosis. The transferrin receptors (TfR) are found to be expressed in lower levels in the case of normal cells and are overexpressed in certain classes of tumours [92]. The binding affinity of Tf to TfR in the case of tumour cells is ten to hundred times more effective when compared to normal cells [93]. Thus, the drug carrier can effectively be labelled through Tf for its effective transport inside the targeted tumour cells. Advantages of nano-therapy include targeted delivery of therapeutic cargo, reduced side effects of anti-cancer compounds, enhanced efficacy of treatment, and better survival rates of patients.

Nanostructures as drug carriers

The capacity of nanocarriers to deliver the cargo to desired space is a challenge. Much research has been conducted to understand the interrelation between characterized nanomaterial and its drug-loading concept. The mechanism of conjugation and drug release at the target tumour is another broad area of research. Various nanostructures have been recognized as efficient drug delivery types, some of which have been briefly discussed below.

Liposomes

Liposomes are lipid-based, colloidal nanocarriers with brain-targeting and selective drug delivery efficacy [94]. These nano-phospho-lipid-carriers can enclose aqueous component enclosed in the phospholipid bilayer. This uniqueness in the structure of nano-liposomes (NLs) allows them to deliver both hydrophilic and hydrophobic cargo [95]. Studies have reported the BCF (Bi-molecular-Corona-Fingerprints) coated cationic-nano-liposomes' ability to bind with overexpressed receptors on the BBB. Such liposomal nanocarriers have shown anti-tumour activity on glioblastomas [46]. Some of the most widely accepted nano-liposomes for treating brain cancer include Poly-Ethylene-Glycolated Liposomal Doxorubicin and Liposomal daunorubicin [31]. Other kinds include solid lipid NPs (SLN) and nanostructured lipid carriers (LNCs) [94].

Nano-micelles

Ligand-mediated delivery of drugs can be an efficient way of treating cancer. For example, an intractable brain cancer type, glioblastoma, has overexpression of integrins $\alpha\beta3/5$ at the angiogenic sites, which has high affinity with a ligand combination, RGD. Here comes the idea of introducing polymeric micelles associated with cyclic, targeted ligands [96]. Such

nano-micelles accumulate into tumours through transcytosis, a type of active internalization.

Dendrimers

Dendrimers are well-defined, three-dimensional polymeric, sphere-shaped nanocarriers that could be multivalently hyper-branched [55]. Various generations of designed dendrimers have been produced that offer sustained drug-, gene-release, and anti-angiogenic abilities [10]. These 1–10 nm dia ranging nanoparticles have several hydrophobic pockets that can be exploited for systemically delivering the bioactive payload. Dendrimers are pretty flexible with modifications and offer mono-dispersion abilities [97]. Their structures can be differentiated into three distinct domains: core, branches, and; functional groups at the terminals. The branches and the bridging among them create a radial geometry called “generation.” The higher the generations, the more spherical progression would be of the dendrimers, along with an increase in their drug-loading capacity. Hence, a higher-generation dendrimer can be used for loading a larger quantity of DNA or drug [98]. Some widely used dendrimers for brain cancer treatment are Poly(amidoamine) (PAMAM), Poly(propylene imine) (PPI), and Poly-L-lysine(PLL).

Carbon nanotubes (CNTs)

These cylindrical, nano-diameter tubes made of carbon allotropes have excellent penetrating efficiency attributed to their structure. This approach for drug delivery not only opens ways through the BBB but also allows for the extended circulation of drugs and moiety functionalization [82]. Nucleic acids, proteins, and drugs have been delivered into cells using functionalized Carbon Nano Tubes (f-CNTs) [30].

Carbon dots

The synthesis of CDs is simple, making them an outstanding candidate for drug delivery to the brain as well [34]. The ultra-small size of CDs allows them to cross the BBB easily, to deliver neurological therapeutics across BBB to treat brain cancers and other neurological disorders [99]. A recent study on CDs demonstrated photothermal therapeutic effects by converting NIR light to heat which kills cancerous cells as well as suppresses oncogenic growth in the model [49].

Green technology-based nanomaterials for brain tumour targeting

The nanomaterials must be surface modified to regulate their interactions with biological systems. Surface functionality imparts stability and localization to these nanoparticles by

altering their physicochemical features. This might involve additional steps in nanomedicine preparation, along with some chemicals (for reducing and capping, catalysis, and so on) that might add toxicity to these nanomedicines. For example, chemically synthesized magnetic ferrous oxide and superparamagnetic nanoparticles (IONs and SP-IONs) have gained much attention for brain cancer treatment. They must be surface coated with hydrophilic polymers to reduce cytotoxicity. This, in turn, reduces the nanoparticle's cargo (drug/gene) carrying capacity along with the altered hydrophilicity of such nanoparticles. These are less efficiently internalized by the cell [100]. However, magnetosomes, bi-lipid membraned, iron-rich nanoparticles having magnetic properties, biologically synthesized by magneto-tactic bacteria can cover all the above-mentioned limitations [101]. Hence, there is a scope for the utilization of such bio-factories for the production of green nanomedicine.

The involvement of living organisms or derived biomolecules for the sustainable and non-toxic production of nanoparticles is termed “green synthesis” of nanomaterials. The choice of capping and reducing agents decides the size and morphology of the green nanoparticles formed by altering the reaction dynamics. Microorganisms like algae, fungi, bacteria, yeast, and plant extracts have been used to synthesize these “nano-medical particles.” The green synthesis of nanomaterials through organisms or extracts from natural sources can occur either extrinsically or intrinsically. For example, bacterial strains can accumulate ions from their surroundings, reducing them to their elemental state by enzymes and metabolites. For the reduction to happen intracellularly, the ions are transported inside the organism's cell, whereas the surface accumulated ions are reduced extracellularly [102]. Green synthesis approaches reduce the toxicity caused due to the functionalization of chemically synthesized nanomaterials. It also tends to reduce the environmental and health impact of chemically or physically synthesized nanomaterials.

Challenges while treating brain tumours

Due to the exclusive intrinsic conditions of neurons and their microenvironment, brain tumour detection and its treatment are complex tasks. The unique epigenetic and developmental features of the brain and its tumour render it resistant to the available treatments. Since there exists a severe disparity among individuals with a brain tumour, there is a need for precision therapy over conventional treatment. In addition, the pronounced intra-tumoural-heterogeneity among tumour cells adds to the existing challenges towards effective treatment [103]. The degree of anaplasia poses another challenge while grading tumours through stereo-tactical surgical biopsies [104]. The Blood–Brain Barrier (BBB) restricts drug

delivery at the target site due to its selective permeability [4]. The drugs used for cancer treatment offer limited cures due to their low biodistribution and even lower bioavailability to the tumour. There have been reported cases of either quick flush-out of drugs from the body or their accumulation, leading to organ toxicity [105]. As per Global Cancer Registry, approximately 30,000 out of 18 million registered cases were of brain cancer [1]. This rarity of occurrence of such cancers has subjected it to a gap in the required scientific and funding attention [103]. The otherwise damaged BBB, as seen in Grade III and Grade IV brain cancer, could also result from therapy leading to the formation of reactive tissues.

Post-delivery, all the nanoparticles are subjected to opsonization and ultimately removed by the reticuloendothelial cells (REC), irrespective of the shape, size, and composition [8]. Hence, timely and adequate removal of such nanomedicine is required.

Conclusion and future prospects

The potential nanoparticles must be considered for their pharmacokinetic and toxicity studies. These young therapeutics should be accessed for their long-term effects. There is much scope for researching the tumour microenvironment stimulated drug release through nanomedicines. “Cancer nano-theranostic” aims at critically imaging tumours and delivering the right amount of drug at the right site. Nanotherapy would be considered selective and efficient upon systemic portage of therapeutic cargo to the body's primary and metastatic cancer cells.

The conventional diagnostic approaches for brain cancer are inefficient in detecting the grade and degree of anaplasia. The heterogeneity among cells of brain tumours is another challenge. Further, chemotherapy has led to developing of multi-drug resistant (MDR) phenotypes. Nowadays, smart nanomedicine has taken over the conventional treatment of brain cancer and has improved survival rates. However, the field of medical oncology demands more data on the mechanism, efficiency, safety, and toxicity of such nanomedicines for them to be accepted as the primary mode of treatment of cancers. All the nanomedicines with recognized potential in diagnosing and treating brain cancer either involve liposomes or have been limited to rodent models. The medical community demands rigorous translation of the research output in the field.

Author contributions RA: Concept and writing-Original draft and revisions; LK: Concept, writing-original draft and revisions; LM: Concept and writing-Original draft and revisions; and NB- Supervision and concept, and editing.

Data availability Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose. There is no any conflict of interests.

References

- Khazaei Z, Goodarzi E, Borhaninejad V et al (2020) The association between incidence and mortality of brain cancer and human development index (HDI): an ecological study. *BMC Public Health* 20:1696. <https://doi.org/10.1186/s12889-020-09838-4>
- Kaur J, Gulati M, Kapoor B et al (2022) Advances in designing of polymeric micelles for biomedical application in brain related diseases. *Chem Biol Interact* 361:109960. <https://doi.org/10.1016/J.CBI.2022.109960>
- Mesfin FB, Al-Dhahir MA (2022) Gliomas. *StatPearls*
- Park JH, de Lomana ALG, Marzese DM et al (2021) A systems approach to brain tumor treatment. *Cancers (Basel)*. <https://doi.org/10.3390/CANCERS13133152>
- Belykh E, Shaffer KV, Lin C et al (2020) Blood-brain barrier, blood-brain tumor barrier, and fluorescence-guided neurosurgical oncology: delivering optical labels to brain tumors. *Front Oncol* 10:739. <https://doi.org/10.3389/FONC.2020.00739/BIBTEX>
- Kim HS, Lee DY (2022) Nanomedicine in clinical photodynamic therapy for the treatment of brain tumors. *Biomedicines* 10:96. <https://doi.org/10.3390/biomedicines10010096>
- Meyers JD, Doane T, Burda C, Basilion JP (2013) Nanoparticles for imaging and treating brain cancer. *Nanomedicine (Lond)* 8:123. <https://doi.org/10.2217/NNM.12.185>
- Mostafavi E, Medina-Cruz D, Vernet-Crua A, et al (2021) Green nanomedicine: the path to the next generation of nanomaterials for diagnosing brain tumors and therapeutics? 18:715–736. <https://doi.org/10.1080/17425247.2021.1865306>
- Cerna T, Stiborova M, Adam V et al (2016) Nanocarrier drugs in the treatment of brain tumors. *J Cancer Metast Treat* 2:407–416. <https://doi.org/10.20517/2394-4722.2015.95>
- Dwivedi N, Shah J, Mishra V et al (2016) Dendrimer-mediated approaches for the treatment of brain tumor. *J Biomater Sci Polym Ed* 27:557–580. <https://doi.org/10.1080/09205063.2015.1133155>
- Chen Q, Tan K, Lin Q, et al (2022) Nanotechnology: a better diagnosis and treatment strategy for brain tumour? 25:33–47. <https://doi.org/10.22186/25.3.1.2>
- d'Angelo M, Castelli V, Benedetti E et al (2019) Theranostic nanomedicine for malignant gliomas. *Front BioengBiotechnol* 7:325. <https://doi.org/10.3389/FBIOE.2019.00325/BIBTEX>
- Tzeng SY, Green JJ (2013) Therapeutic nanomedicine for brain cancer. *Ther Deliv* 4:687–704. <https://doi.org/10.4155/tde.13.38>
- Kadry H, Noorani B, Cucullo L (2020) A blood–brain barrier overview on structure, function, impairment, and biomarkers of integrity. *FluidsBarriers CNS* 17:1. <https://doi.org/10.1186/S12987-020-00230-3> (17:1–24)
- Rhea EM, Banks WA (2019) Role of the blood-brain barrier in central nervous system insulin resistance. *Front Neurosci* 13:521. <https://doi.org/10.3389/FNINS.2019.00521/BIBTEX>
- Fisusi FA, Schätzlein AG, Uchegbu IF (2018) Nanomedicines in the treatment of brain tumors. 13:579–583. <https://doi.org/10.2217/NNM-2017-0378>
- Haque S, Norbert CC, Patra CR (2021) Nanomedicine: future therapy for brain cancers. *Nano Drug Deliv Strateg Treat Cancers*. <https://doi.org/10.1016/B978-0-12-819793-6.00003-5>
- Wadajkar AS et al (2021) Surface-Modified Nanodrug Carriers for Brain Cancer Treatment. In: Agrahari V, Kim A, Agrahari V (eds) *Nanotherapy for Brain Tumor Drug Delivery*. *Neuromethods*, vol 163. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-1052-7_5
- Gao H (2016) Perspectives on dual targeting delivery systems for brain tumors. *J Neuroimmune Pharmacol* 12:1. <https://doi.org/10.1007/S11481-016-9687-4> (12:6–16)
- Gao H (2016) Progress and perspectives on targeting nanoparticles for brain drug delivery. *Acta Pharm Sin B* 6:268–286. <https://doi.org/10.1016/J.APSB.2016.05.013>
- Jena LN, Bennie LA, McErlean EM et al (2021) Exploiting the anticancer effects of a nitrogen bisphosphonate nanomedicine for glioblastoma multiforme. *J Nanobiotechnology* 19:1–18. <https://doi.org/10.1186/S12951-021-00856-X/TABLES/2>
- Kemp JA, Kwon YJ (2021) Cancer nanotechnology: current status and perspectives. *Nano Convergence* 8:1. <https://doi.org/10.1186/S40580-021-00282-7> (8:1–38)
- Jena L, McErlean E, McCarthy H (2020) Delivery across the blood-brain barrier: nanomedicine for glioblastoma multiforme. *Drug Deliv Transl Res* 10:304–318. <https://doi.org/10.1007/S13346-019-00679-2/FIGURES/2>
- Hanif S, Muhammad P, Chesworth R et al (2020) Nanomedicine-based immunotherapy for central nervous system disorders. *Acta Pharmacol Sin* 41:7. <https://doi.org/10.1038/s41401-020-0429-z> (41:936–953)
- Tang W, Fan W, Lau J et al (2019) Emerging blood–brain-barrier-crossing nanotechnology for brain cancer theranostics. *Chem Soc Rev* 48:2967–3014. <https://doi.org/10.1039/C8CS00805A>
- Ruan S, Zhou Y, Jiang X, Gao H (2021) Rethinking CRITID procedure of brain targeting drug delivery: circulation, blood brain barrier recognition, intracellular transport, diseased cell targeting, internalization, and drug release. *Adv Sci* 8:2004025. <https://doi.org/10.1002/ADVS.202004025>
- Houston ZH, Bunt J, Chen KS et al (2020) Understanding the uptake of nanomedicines at different stages of brain cancer using a modular nanocarrier platform and precision bispecific antibodies. *ACS Cent Sci* 6:727–738. https://doi.org/10.1021/ACSCENTSCI.9B01299/ASSET/IMAGES/LARGE/OC9B01299_0004.JPEG
- van't Root M, Lowik C, Mezzanotte L (2017) Targeting nanomedicine to brain tumors: latest progress and achievements. *Curr Pharm Des* 23:1953–1962. <https://doi.org/10.2174/1381612822666161227153359>
- Liu Y, Liu J, Zhang J et al (2018) Noninvasive brain tumor imaging using red emissive carbonized polymer dots across the blood-brain barrier. *ACS Omega* 3:7888–7896. https://doi.org/10.1021/ACSOMEGA.8B01169/ASSET/IMAGES/LARGE/AO-2018-011692_0003.JPEG
- Mukhtar M, Bilal M, Rahdar A et al (2020) Nanomaterials for diagnosis and treatment of brain cancer: recent updates. *Chemosensors* 8:1–31. <https://doi.org/10.3390/chemosensors8040117>
- Wen CJ, Sung CT, Aljuffali IA et al (2013) Nanocomposite liposomes containing quantum dots and anticancer drugs for bioimaging and therapeutic delivery: a comparison of cationic, PEGylated and deformable liposomes. *Nanotechnology*. <https://doi.org/10.1088/0957-4484/24/32/325101>
- Zhang W, Sigdel G, Mintz KJ et al (2021) Carbon dots: A future blood–brain barrier penetrating nanomedicine and drug nanocarrier. *Int J Nanomed* 16:5003–5016. <https://doi.org/10.2147/IJN.S318732>

33. Wu H, Su W, Xu H et al (2021) Applications of carbon dots on tumour theranostics. *View* 2:20200061. <https://doi.org/10.1002/viw.20200061>
34. Ashrafizadeh M, Mohammadinejad R, Kailasa SK et al (2020) Carbon dots as versatile nanoarchitectures for the treatment of neurological disorders and their theranostic applications: a review. *Adv Colloid Interface Sci* 278:102123. <https://doi.org/10.1016/j.cis.2020.102123>
35. Calabrese G, de Luca G, Nocito G et al (2021) Carbon dots: AN innovative tool for drug delivery in brain tumors. *Int J Mol Sci*. <https://doi.org/10.3390/ijms222111783>
36. Du Y, Qian M, Li C, et al (2018) Facile marriage of Gd3+ to polymer-coated carbon nanodots with enhanced biocompatibility for targeted MR/fluorescence imaging of glioma. *Int J Pharm* 552:84–90. <https://doi.org/10.1016/j.ijpharm.2018.09.010>
37. Zhao Y, Xie Y, Liu Y et al (2022) Comprehensive exploration of long-wave emission carbon dots for brain tumor visualization. *J Mater Chem B*. <https://doi.org/10.1039/d2tb00322h>
38. Fan Z, Zhou S, Garcia C et al (2017) pH-Responsive fluorescent graphene quantum dots for fluorescence-guided cancer surgery and diagnosis. *Nanoscale* 9:4928–4933. <https://doi.org/10.1039/C7NR00888K>
39. Li S, Su W, Wu H et al (2020) Targeted tumour theranostics in mice via carbon quantum dots structurally mimicking large amino acids. *Nat Biomed Eng* 4:704–716. <https://doi.org/10.1038/s41551-020-0540-y>
40. Wang L, Zhou HS (2014) Green synthesis of luminescent nitrogen-doped carbon dots from milk and its imaging application. *Anal Chem* 86:8902–8905. https://doi.org/10.1021/AC502646X/SUPPL_FILE/AC502646X_SI_001.PDF
41. Ruan S, Qian J, Shen S et al (2014) A simple one-step method to prepare fluorescent carbon dots and their potential application in non-invasive glioma imaging. *Nanoscale* 6:10040–10047. <https://doi.org/10.1039/C4NR02657H>
42. Sonali VMK, Singh RP et al (2018) Nanotheranostics: Emerging strategies for early diagnosis and therapy of brain cancer. *Nanotheranostics* 2:70–86. <https://doi.org/10.7150/NTNO.21638>
43. Wang X, Tu M, Tian B et al (2016) Synthesis of tumor-targeted folate conjugated fluorescent magnetic albumin nanoparticles for enhanced intracellular dual-modal imaging into human brain tumor cells. *Anal Biochem* 512:8–17. <https://doi.org/10.1016/j.ab.2016.08.010>
44. Bagheri S, Yasemi M, Safaie-Qamsari E, et al (2018) Using gold nanoparticles in diagnosis and treatment of melanoma cancer. 46:462–471. <https://doi.org/10.1080/21691401.2018.1430585>
45. Smilowitz HM, Meyers A, Rahman K et al (2018) Intravenously-injected gold nanoparticles (AuNPs) access intracerebral F98 rat gliomas better than AuNPs infused directly into the tumor site by convection enhanced delivery. *Int J Nanomed* 13:3937–3948. <https://doi.org/10.2147/IJN.S154555>
46. Zhao L, Li Y, Zhu J et al (2019) Chlorotoxin peptide-functionalized polyethylenimine-entrapped gold nanoparticles for glioma SPECT/CT imaging and radionuclide therapy. *J Nanobiotechnol* 17:1–13. <https://doi.org/10.1186/S12951-019-0462-6/FIGURES/7>
47. McHugh KJ, Jing L, Behrens AM et al (2018) Biocompatible semiconductor quantum dots as cancer imaging agents. *Adv Mater* 30:1706356. <https://doi.org/10.1002/ADMA.201706356>
48. Huang N, Cheng S, Zhang X et al (2017) Efficacy of NGR peptide-modified PEGylated quantum dots for crossing the blood–brain barrier and targeted fluorescence imaging of glioma and tumor vasculature. *Nanomedicine* 13:83–93. <https://doi.org/10.1016/J.NANO.2016.08.029>
49. Wang H, Mu Q, Wang K et al (2019) Nitrogen and boron dual-doped graphene quantum dots for near-infrared second window imaging and photothermal therapy. *Appl Mater Today* 14:108–117. <https://doi.org/10.1016/J.APMT.2018.11.011>
50. Liu Y, Liu J, Chen D et al (2020) Fluorination enhances NIR-II fluorescence of polymer dots for quantitative brain tumor imaging. *Angewandte Chemie Int Ed* 59:21049–21057. <https://doi.org/10.1002/ANIE.202007886>
51. Suárez-García S, Arias-Ramos N, Frias C et al (2018) Dual T1/T2 nanoscale coordination polymers as novel contrast agents for MRI: a preclinical study for brain tumor. *ACS Appl Mater Interfaces* 10:38819–38832. https://doi.org/10.1021/ACSAMI.8B15594/SUPPL_FILE/AM8B15594_SI_001.PDF
52. Huang X, Deng G, Liao L et al (2017) CuCo2S4 nanocrystals: a new platform for multimodal imaging guided photothermal therapy. *Nanoscale* 9:2626–2632. <https://doi.org/10.1039/C6NR09028A>
53. Ho YN, Shu LJ, Yang YL (2017) Imaging mass spectrometry for metabolites: technical progress, multimodal imaging, and biological interactions. *Wiley Interdiscip Rev Syst Biol Med* 9:e1387. <https://doi.org/10.1002/WSBM.1387>
54. Hallal S, Ebrahimkhani S, Shivalingam B et al (2019) The emerging clinical potential of circulating extracellular vesicles for non-invasive glioma diagnosis and disease monitoring. *Brain Tumor Pathol* 36:2. <https://doi.org/10.1007/S10014-019-00335-0> (36:29–39)
55. Wang H, Jiang D, Li W et al (2019) Evaluation of serum extracellular vesicles as noninvasive diagnostic markers of glioma. *Theranostics* 9:5347–5358. <https://doi.org/10.7150/THNO.33114>
56. Rufino-Ramos D, Albuquerque PR, Carmona V et al (2017) Extracellular vesicles: Novel promising delivery systems for therapy of brain diseases. *J Control Release* 262:247–258. <https://doi.org/10.1016/J.JCONREL.2017.07.001>
57. Vaneckova M, Herman M, Smith M et al (2015) Gadobenate dimeglumine (MultiHance) or gadoterate meglumine (Dotarem) for brain tumour imaging? An intra-individual comparison. *Cancer Imaging* 15:P15. <https://doi.org/10.1186/1470-7330-15-S1-P15>
58. Fan Z, Fu PP, Yu H, Ray PC (2014) Theranostic nanomedicine for cancer detection and treatment. *J Food Drug Anal* 22:3–17. <https://doi.org/10.1016/J.JFDA.2014.01.001>
59. Norden AD, Drappatz J, Wen PY (2008) Novel anti-angiogenic therapies for malignant gliomas. *Lancet Neurol* 7:1152–1160. [https://doi.org/10.1016/S1474-4422\(08\)70260-6](https://doi.org/10.1016/S1474-4422(08)70260-6)
60. Alle M, reddyKim GBTH et al (2020) Doxorubicin-carboxymethyl xanthan gum capped gold nanoparticles: microwave synthesis, characterization, and anti-cancer activity. *Carbohydr Polym* 229:115511. <https://doi.org/10.1016/J.CARBPOL.2019.115511>
61. Ruan S, Xiao W, Hu C et al (2017) Ligand-mediated and enzyme-directed precise targeting and retention for the enhanced treatment of glioblastoma. *ACS Appl Mater Interfaces* 9:20348–20360. https://doi.org/10.1021/ACSAMI.7B02303/SUPPL_FILE/AM7B02303_SI_001.PDF
62. Ruan S, He Q, Gao H (2015) Matrix metalloproteinase triggered size-shrinkable gelatin-gold fabricated nanoparticles for tumor microenvironment sensitive penetration and diagnosis of glioma. *Nanoscale* 7:9487–9496. <https://doi.org/10.1039/C5NR01408E>
63. Ruan S, Yuan M, Zhang L et al (2015) Tumor microenvironment sensitive doxorubicin delivery and release to glioma using angiopep-2 decorated gold nanoparticles. *Biomaterials* 37:425–435. <https://doi.org/10.1016/J.BIOMATERIALS.2014.10.007>
64. le Ouay B, Stellacci F (2015) Antibacterial activity of silver nanoparticles: A surface science insight. *Nano Today* 10:339–354. <https://doi.org/10.1016/J.NANTOD.2015.04.002>
65. Dayem AA, Hossain MK, Lee S, bin et al (2017) The role of reactive oxygen species (ros) in the biological activities of metallic nanoparticles. *Int J Mol Sci* 18:120. <https://doi.org/10.3390/IJMS18010120>

66. Salazar-García S, García-Rodrigo JF, Martínez-Castañón GA et al (2020) Silver nanoparticles (AgNPs) and zinc chloride (ZnCl₂) exposure order determines the toxicity in C6 rat glioma cells. *J Nanopart Res* 22:1–13. <https://doi.org/10.1007/S11051-020-04984-7/FIGURES/8>
67. Locatelli E, Naddaka M, Uboldi C et al (2014) Targeted delivery of silver nanoparticles and alisertib: In vitro and in vivo synergistic effect against glioblastoma. *Nanomedicine* 9:839–849. https://doi.org/10.2217/NNM.14.1/SUPPL_FILE/SUPPL_MATERIAL.DOCX
68. Jin T, Sun D, Su JY et al (2009) Antimicrobial efficacy of zinc oxide quantum dots against *Listeria monocytogenes*, *Salmonella enteritidis*, and *Escherichia coli* O157:H7. *J Food Sci* 74:M46–M52. <https://doi.org/10.1111/J.1750-3841.2008.01013.X>
69. Jin BJ, Bae SH, Lee SY, Im S (2000) Effects of native defects on optical and electrical properties of ZnO prepared by pulsed laser deposition. *Mater Sci Eng, B* 71:301–305. [https://doi.org/10.1016/S0921-5107\(99\)00395-5](https://doi.org/10.1016/S0921-5107(99)00395-5)
70. Ostrovsky S, Kazimirsky G, Gedanken A, Brodie C (2009) Selective cytotoxic effect of ZnO nanoparticles on glioma cells. *Nano Res* 2:11. <https://doi.org/10.1007/S12274-009-9089-5> (2:882–890)
71. Shim KH, Hulme J, Maeng EH et al (2014) Analysis of zinc oxide nanoparticles binding proteins in rat blood and brain homogenate. *Int J Nanomed* 9(Suppl 2):217–224. <https://doi.org/10.2147/IJN.S58204>
72. Sharma AK, Singh V, Gera R et al (2017) Zinc oxide nanoparticle induces microglial death by NADPH-oxidase-independent reactive oxygen species as well as energy depletion. *Mol Neurobiol* 54:6273–6286. <https://doi.org/10.1007/S12035-016-0133-7/FIGURES/13>
73. Wahab R, Kaushik NK, Verma AK et al (2011) Fabrication and growth mechanism of ZnO nanostructures and their cytotoxic effect on human brain tumor U87, cervical cancer HeLa, and normal HEK cells. *J Biol Inorg Chem* 16:431–442. <https://doi.org/10.1007/S00775-010-0740-0/FIGURES/7>
74. Attia H, Nounou H, Shalaby M (2018) Zinc oxide nanoparticles induced oxidative DNA damage, inflammation and apoptosis in rat's brain after oral exposure. *Toxics* 6:29. <https://doi.org/10.3390/TOXICS6020029>
75. Chamberlain MC (2012) Neurotoxicity of intra-CSF liposomal cytarabine (DepoCyt) administered for the treatment of leptomeningeal metastases: a retrospective case series. *J Neurooncol* 109:143–148. <https://doi.org/10.1007/s11060-012-0880-x>
76. Domínguez AR, Hidalgo DO, Garrido RV, Sánchez ET (2005) Liposomal cytarabine (DepoCyt®) for the treatment of neoplastic meningitis. *Clin Transl Oncol* 7:232–238. <https://doi.org/10.1007/BF02710168>
77. Gaillard PJ, Appeldoorn CCM, Dorland R et al (2014) Pharmacokinetics, brain delivery, and efficacy in brain tumor-bearing mice of glutathione pegylated liposomal doxorubicin (2B3-101). *PLoS ONE* 9:e82331. <https://doi.org/10.1371/journal.pone.0082331>
78. Rafiyath SM, Rasul M, Lee B et al (2012) Comparison of safety and toxicity of liposomal doxorubicin vs. conventional anthracyclines: a meta-analysis. *Exp Hematol Oncol* 1:10. <https://doi.org/10.1186/2162-3619-1-10>
79. Neil S, Nemunaitis J, Nunan R et al (2012) 51. Results of a phase I trial of SGT-53: a systemically administered, tumor-targeting immunoliposome nanocomplex incorporating a plasmid encoding wtp53. *Mol Ther* 20:S21. [https://doi.org/10.1016/S1525-0016\(16\)S5855-5](https://doi.org/10.1016/S1525-0016(16)S5855-5)
80. Kim S-S, Harford JB, Moghe M et al (2018) Combination with SGT-53 overcomes tumor resistance to a checkpoint inhibitor. *Oncoimmunology* 7:e1484982. <https://doi.org/10.1080/2162402X.2018.1484982>
81. Banerjee K, Núñez FJ, Haase S et al (2021) Current approaches for glioma gene therapy and virotherapy. *Front Mol Neurosci*. <https://doi.org/10.3389/fnmol.2021.621831>
82. Bort G, Lux F, Dufort S et al (2020) EPR-mediated tumor targeting using ultrasmall-hybrid nanoparticles: from animal to human with theranostic AGuIX nanoparticles. *Theranostics* 10:1319–1331. <https://doi.org/10.7150/thno.37543>
83. Dawidczyk CM, Russell LM, Searson PC (2014) Nanomedicines for cancer therapy: State-of-the-art and limitations to pre-clinical studies that hinder future developments. *Front Chem* 2:69. <https://doi.org/10.3389/FCHEM.2014.00069/ABSTRACT>
84. Ediriwickrema A, Saltzman WM (2015) Nanotherapy for cancer: targeting and multifunctionality in the future of cancer therapies. *ACS Biomater Sci Eng* 1:64–78. <https://doi.org/10.1021/ab500084g>
85. Peer D, Karp JM, Hong S et al (2007) Nanocarriers as an emerging platform for cancer therapy. *Nat Nanotechnol* 2:12. <https://doi.org/10.1038/nnano.2007.387> (2:751–760)
86. Bray N (2015) Biologics: Transferrin' bispecific antibodies across the blood-brain barrier. *Nat Rev Drug Discov* 14:14. <https://doi.org/10.1038/NRD4522>
87. Alibolandi M, Ramezani M, Abnous K et al (2015) In vitro and in vivo evaluation of therapy targeting epithelial-cell adhesion-molecule aptamers for non-small cell lung cancer. *J Control Release* 209:88–100. <https://doi.org/10.1016/J.JCONREL.2015.04.026>
88. Yang Z, Tang W, Luo X et al (2015) Dual-ligand modified polymer-lipid hybrid nanoparticles for docetaxel targeting delivery to Her2/neu overexpressed human breast cancer cells. *J Biomed Nanotechnol* 11:1401–1417. <https://doi.org/10.1166/jbnn.2015.2086>
89. Shan L, Liu M, Wu C et al (2015) Multi-small molecule conjugations as new targeted delivery carriers for tumor therapy. *Int J Nanomed* 10:5571–5591. <https://doi.org/10.2147/IJN.S85402>
90. Wu Q, Zheng H, Gu J et al (2022) Detection of folate receptor-positive circulating tumor cells as a biomarker for diagnosis, prognostication, and therapeutic monitoring in breast cancer. *J Clin Lab Anal* 36:e24180. <https://doi.org/10.1002/JCLA.24180>
91. Pan X, Lee RJ (2005) Tumour-selective drug delivery via folate receptor-targeted liposomes. 1:7–17. <https://doi.org/10.1517/17425247.1.1.7>
92. Ruan S, Qin L, Xiao W et al (2018) Acid-responsive transferrin dissociation and GLUT mediated exocytosis for increased blood-brain barrier transcytosis and programmed glioma targeting delivery. *Adv Funct Mater* 28:1802227. <https://doi.org/10.1002/ADFM.201802227>
93. Guo L, Zhang H, Wang F et al (2015) Targeted multidrug-resistance reversal in tumor based on PEG-PLL-PLGA polymer nano drug delivery system. *Int J Nanomed* 10:4535–4547. <https://doi.org/10.2147/IJN.S85587>
94. Nsairat H, Khater D, Odeh F et al (2021) Lipid nanostructures for targeting brain cancer. *Heliyon* 7:e07994. <https://doi.org/10.1016/j.heliyon.2021.e07994>
95. Sapkota R, Dash AK (2021) Liposomes and transferosomes: a breakthrough in topical and transdermal delivery. *Ther Deliv* 12:145–158. <https://doi.org/10.4155/tde-2020-0122>
96. Miura Y, Takenaka T, Toh K et al (2013) Cyclic RGD-linked polymeric micelles for targeted delivery of platinum anticancer drugs to glioblastoma through the blood-brain tumor barrier. *ACS nano* 7(10):8583–92
97. Ruan S, Xie R, Qin L et al (2019) Aggregable nanoparticles-enabled chemotherapy and autophagy inhibition combined with anti-PD-1 antibody for improved glioma treatment. *Nano Lett* 19:8318–8332. https://doi.org/10.1021/ACS.NANOLETT.9B03968/SUPPL_FILE/NL9B03968_SI_001.PDF

98. Kesharwani P, Iyer AK (2015) Recent advances in dendrimer-based nanovectors for tumor-targeted drug and gene delivery. *Drug Discov Today* 20:536–547
99. Zhou Y, Peng Z, Seven ES, Leblanc RM (2018) Crossing the blood-brain barrier with nanoparticles. *J Control Release* 270:290–303. <https://doi.org/10.1016/J.JCONREL.2017.12.015>
100. Weerathunge P, Pooja D, Singh M et al (2019) Transferrin-conjugated quasi-cubic SPIONs for cellular receptor profiling and detection of brain cancer. *Sens Actuators B Chem* 297:126737. <https://doi.org/10.1016/J.SNB.2019.126737>
101. Alphandéry E, Idhah A, Adam C et al (2017) Chains of magnetosomes with controlled endotoxin release and partial tumor occupation induce full destruction of intracranial U87-Luc glioma in mice under the application of an alternating magnetic field. *J Control Release* 262:259–272. <https://doi.org/10.1016/J.JCONREL.2017.07.020>
102. Patra S, Mukherjee S, Barui AK et al (2015) Green synthesis, characterization of gold and silver nanoparticles and their potential application for cancer therapeutics. *Mater Sci Eng, C* 53:298–309. <https://doi.org/10.1016/J.MSEC.2015.04.048>
103. Aldape K, Brindle KM, Chesler L et al (2019) Challenges to curing primary brain tumours. *Nat Rev Clin Oncol* 16:8. <https://doi.org/10.1038/s41571-019-0177-5> (16:509–520)
104. de Campos Vieira Abib S, Chui CH, Cox S et al (2022) International Society of Paediatric Surgical Oncology (IPSO) Surgical Practice Guidelines. *Ecancermedicalscience*. <https://doi.org/10.3332/ECANCER.2022.1356>
105. Senapati S, Kumar Mahanta A, Kumar S, Maiti P (2018) Controlled drug delivery vehicles for cancer treatment and their performance. *Signal Transduct Target Ther*. <https://doi.org/10.1038/s41392-017-0004-3>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Raksha Anand¹ · Lakhan Kumar¹ · Lalit Mohan¹ · Navneeta Bharadvaja¹

✉ Navneeta Bharadvaja
navneetab@dce.ac.in

Raksha Anand
r27anand@gmail.com

Lakhan Kumar
adarsh.lakhan@gmail.com

Lalit Mohan
lalitmohan2405@gmail.com

¹ Plant Biotechnology Laboratory, Department of Biotechnology, Delhi Technological University, New Delhi, Delhi, India



Natural polyphenols: a promising bioactive compounds for skin care and cosmetics

Navneeta Bharadvaja¹ · Shruti Gautam¹ · Harshita Singh¹

Received: 12 June 2022 / Accepted: 23 November 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

The physiological and morphological aspects of skin suffer from frequent change. Numerous internal and external factors have direct impact on inducing various skin problems like inflammation, aging, cancer, oxidative stress, hyperpigmentation etc. The use of plant polyphenols as a photo-ecting agent is gaining popularity nowadays. Polyphenols are known to enhance endogenous antioxidant system of skin thereby preventing various skin diseases. The biological activity of plant polyphenols is dependent on their physicochemical properties for overcoming the epidermal barriers to reach the specific receptor. Several evidences have reported the vital role polyphenols in mitigating adverse skin problems and reverting back the healthy skin condition. The interest in plant derived skin care products is emerging due to the changing notion of people to shift their focus towards use of plant-based products. The present review draws an attention to uncover the protective role of polyphenols in prevention of various skin problems. Several in vitro and in vivo studies have been summarized that claims the efficacious nature of plant extract having dermatological significance.

Keywords Skin · Polyphenols · Anti-aging · Photoprotectant · Anti-melanogenesis · Antioxidant

Abbreviations

MMP	Matrix metalloproteinase
UV rays	Ultraviolet rays
ROS	Reactive oxygen species
EGCG	Epigallo catechin-3-gallate
NO	Nitric oxide
H ₂ O ₂	Hydrogen peroxide
COX-2	Cyclooxygenase-2
GTP	Green tea polyphenol
EC	Epicatechin
ECG	(-) Epicatechin-3-gallate
IL	Interleukin
TNF	Tumor necrosis factor
MAPK	Mitogen-activated protein kinase
iNOS	Inducible nitric oxide synthase
TGF	Transforming growth factor
AP-1	Activator protein 1
HaCaT	Human keratinocyte cell culture
TGM	Transglutaminase

FLG	Filaggrin gene
HAS	Hyaluronic acid synthase
MPO	Myeloperoxidase
SOD	Superoxide dismutase
CAT	Catalase
TRP	Tyrosinase related protein
DNA	Deoxyribonucleic acid

Introduction

Skin is one of the complex organs covering the complete part of the body. It performs a crucial function to protect the body from extreme temperature conditions and external factors by acting as a barrier to harmful microbes and harmful rays of sun [1]. However, the skin is continuously prone to physiological and morphological falloffs [2]. The major cause of skin damage has been attributed to the harmful UV rays and oxidative stress leading to cutaneous damage, skin cancer, premature aging, hyper pigmentation etc. The endogenous antioxidant system of skin comprises of many enzymatic and non-enzymatic antioxidants. Enzymatic antioxidants like glutathione reductase (for maintaining cellular control of reactive oxygen species), glutathione peroxidase (for protecting cell against oxidative damage) [3], superoxide

✉ Navneeta Bharadvaja
navneeta@dtu.ac.in

¹ Plant Biotechnology Laboratory, Department of Biotechnology, Delhi Technological University, Delhi 110042, India

dismutase and catalase (for degrading hydrogen peroxide) etc. protect cell from oxidative damages. The non-enzymatic antioxidants like glutathione, tocopherol, and ubiquinol etc. functions to scavenge free radicals. Due to the excess free radical formation, there is initiation of various degenerative processes in skin that diminishes the effectiveness of endogenous antioxidant system, thereby causing several skin problems. For example, exposure of skin cells to excessive UVA radiations leads to reduction in the activity of catalase and superoxide dismutase enzyme [1]. The increase in oxidative stress leads to degradation of proteins by upregulation of enzymes like matrix metalloproteinase (MMPs) which can facilitate aging and cancer development and other skin related issues. For enhancement of endogenic antioxidant system activity for restorance of the normal physiological condition of skin, the antioxidants are either supplied in the form of diet or applied topically as cosmetics. Nowadays natural products are gaining attention in cosmetology for combating various skin-related pathologies like sunburn, hyper pigmentation, premature skin aging, skin allergies, tumors etc. Plant based skin care products are gaining attention due to their safe nature compared to the synthetically derived cosmetics. The research for skin care products should necessarily not only focus to improve skin appearance but also to prevent various skin disorders. Polyphenols are most widely used bioactive substances due to their anti-carcinogenesis, anti-hyperpigmented, anti-oxidative, anti-aging, antibacterial and anti-inflammatory properties.

Polyphenols consists of large family of naturally occurring secondary metabolites generated by plants in response to any kind of biotic and abiotic stress. They are widely available in plants throughout both the hemisphere and are conferring different range of colors to fruits and flowers of plants. Polyphenols ranges in terms of their structure and properties and are widely present in different vegetative and

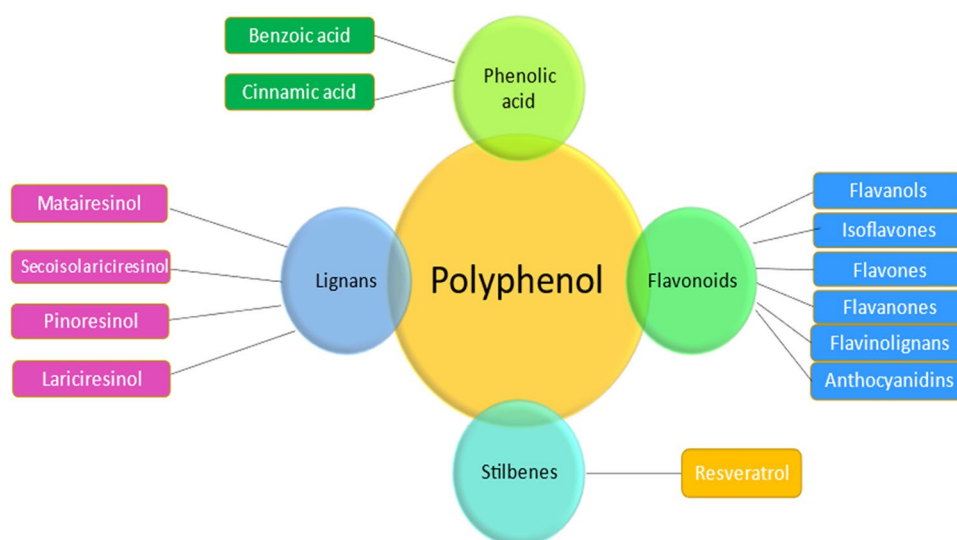
generative organs of plants like leaves, root, flower, fruits, bark, seeds etc. and imparts them some kind of bitterness and astringency [4]. The monomer unit in polyphenols is phenol ring and all these phenolic entities are derived via a common intermediate, either from phenylalanine, or shikimic acid. Primarily these polyphenols are present as conjugated forms with one or more sugar residue linked to hydroxyl group and in some case there is also direct linkage of sugar with aromatic carbon [5]. Till now more than 8000 polyphenols have been identified, out of which 4000 belongs to flavonoids. Polyphenols are classified into many classes (mentioned in Fig. 1) based on number of phenol ring present and the binding properties of the ring structure.

The presence of hydroxyl and phenyl group in phenol ring provides them immunomodulatory, anti-inflammatory and antioxidant properties. Their further division into subclasses is based on the interaction of their respective phenyl ring with the oxygen, carbon and organic acid molecule. The basic classification of polyphenols distinguishes flavonoids and non-flavonoids compound. Flavonoids, the most diverse and largest group of polyphenols in secondary metabolites is found mainly as *o*-glycosides in vacuolar juices present in plant cell. The non-specific activity of flavonoids is absorption of UV radiations, neutralization of ROS, anti-inflammatory, antimicrobial activities, metal chelation etc. The ROS over here, has severe implications of skin damage. Some of the ROS converts into a stable H_2O_2 which in turn gets converted to two molecules of water and a single molecule of oxygen [5] as depicted in equation below.



It has been seen that the flavonoids extracted from grape tea leaves and its seeds, oligomers of pycogenol are known for their protective effect on skin against radical stress. The

Fig. 1 Classification of polyphenol



metal chelating activity of flavonoids is major factor influencing anti-aging property of flavonoids. Flavonoids like quercetin, kaempferol and myricetin are effective at inhibiting histamine release. The flavonoids can be considered as cosmeceutical compound due to its ability to improve skin-blood microcirculation in order to enhance the factor that limits tissue growth and renewal. The major classes of polyphenols are flavonoids, stilbenes, lignans, anthocyanins, phenolic acids, phenolic alcohols, terpenes. The presence of polyphenols can be found in different plant parts like flavanols in tea, flavanones in citrus fruits, apples, quercetin (a type of flavanols) in onions and tea. Hydroxycinnamic acid which is a type of phenolic acid is abundant in coffee and other fruits and vegetables. Resveratrol is a type of stilbene found in the peel of dark colored grapes (Afaq and Katiyar). Procyanidins are oligomeric catechin found abundantly in grapes, grape seeds, apples, red wine, cocoa, cranberry etc. [6].

Anthocyanidins and anthocyanin are water soluble pigments prominent in large number of vegetables, flowers and fruits mainly grapes, grapes extract and berries [7]. The different molecular structure of these polyphenols imparts them various biological properties. With the growing widespread use of polyphenols, the focus has shifted towards expanding and exploring their health benefits. There have been numerous studies reported to confirm that the use of various plant polyphenols contributing towards attaining a healthy skin by their various molecular actions which include regulating DNA repair mechanism, regulating the anti-inflammatory activities, modulating various signaling pathways contributing towards initiation of photo carcinogenesis and

immunosuppression and preventing from DNA damage. The review article presents the role of polyphenols in skin care, research studies on various plants used in different formulations for correcting skin problems and nanocarrier formulations for skin.

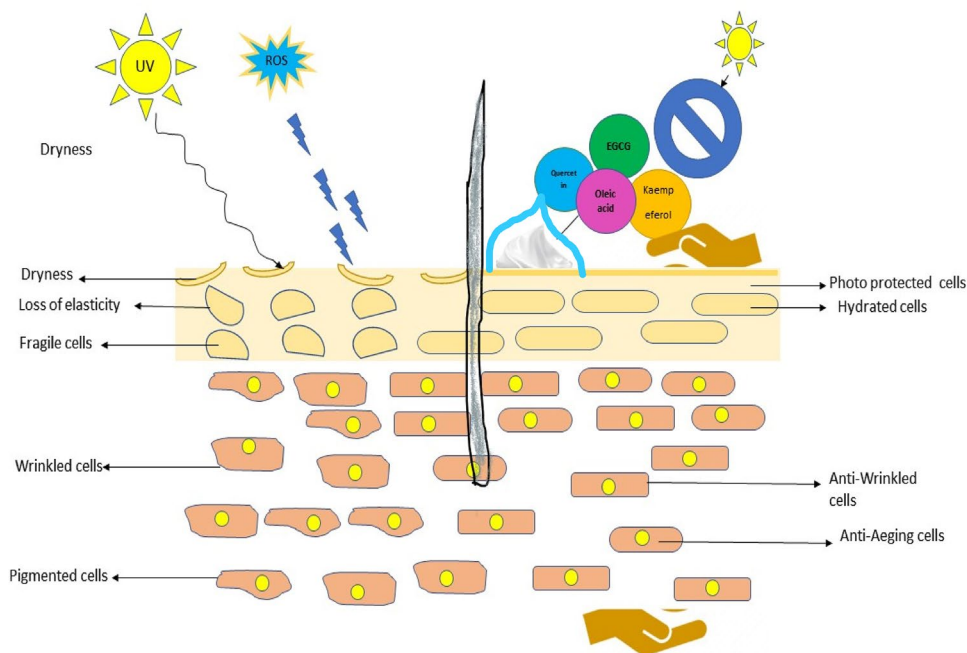
Protective role of polyphenols in skin care

Photo-protective effects

Polyphenols are immensely exploited as promising photo-protective agent for preventing skin photodamage and skin melanoma induced due to harmful radiations. Polyphenols are used as a natural agent in cosmetics to suppress and reverse the damage caused due to UV exposure is presented in Fig. 2. The presence of chromophores structure in phenolic compounds imparts them ability to absorb the UV radiation [8]. Hence, most of the polyphenols have UV absorbing properties. They can effectively block the UV radiations to penetrate into the stratum corneum when used topically. The polyphenols have great UV absorbing properties and can effectively absorb entire wavelength of UVB spectrum.

The antioxidant, anti-inflammatory and ROS scavenging ability of polyphenol acts as natural sunscreen and also provides effective photo-protective effects. Many studies have exhibited the effectiveness of plant polyphenols in ameliorating the damages caused due to UV exposure. The polyphenols like green tea polyphenols, proanthocyanidins, silymarin have proved effective against UV induced damages like inflammation, DNA damage, oxidative stress, generation of

Fig. 2 Photoprotective effects of polyphenols



free radical and immunosuppression. Different researches have proved that polyphenols inhibits the signaling molecules associated with inflammation caused due to UVB exposure which ultimately reflects the photoprotective and anti-photo carcinogenic effect of polyphenols [9].

Anti-oxidant effects

The photo of harmful UV rays leads to chronic damage to skin's endogenous antioxidant defense system leading to oxidative stress and cutaneous damage resulting in various skin disorders, premature skin aging, leading to cancer development and immunosuppression. The topical use of natural plant phenols has shown significant reduction in ROS level in skin. The use of EGCG has shown promising effects in ameliorating the skin damage caused due to excessive ROS generation in skin. Polyphenols have found to be effective in enhancing endogenous antioxidant defense system. Polyphenols have shown promising effects in suppression of lipid peroxidation, decrease the level of UV induced NO and H₂O₂ thereby acting as a potent natural antioxidant. Earlier, it was believed that the polyphenols fight against free radical directly by scavenging free radicals but now many studies have proved that polyphenols directly interact with the enzyme or receptor involved in signal transduction process and modulates the redox status of cell. The antioxidant potential of Polyphenols reported improvement in the cell survival rate and induced apoptosis and tumor prevention. Metal ions are the chief source of generation of free radicals that plays a key role in causing oxidative damages. The effectiveness of polyphenols such as catechol and gallol in chelating metal ions is well documented. Such as the tea polyphenol are known to chelate copper ions and prevent peroxidation of low-density lipoproteins. These also have significant effect in protecting from heavy metal toxicity. These polyphenols affects signal transduction pathways by

modulating the endocrine system which ultimately leads to alter the hormones and affects the physiological process due to binding of metal ions and enzyme cofactors [9].

Anti-inflammatory properties

The polyphenol has ability to inhibit phospholipase and cyclooxygenase activity thereby downregulating the inflammation activity. The exposure to harmful UVB radiation initiates the UV induced inflammatory responses which are prominently visible as redness of skin and erythema. The inflammatory responses are characterized by various factors like generation of free radical, enhancement of inflammatory cytokines, cutaneous edema, dermal blood vessel dilation, along with enhancement in the protein and enzyme expression of COX-2. Further, the inflammation due to UV irradiation triggers the development of tumor [9]. There have been evidences that proof the anti-inflammatory properties of plant polyphenols. Green tea polyphenols are considered to be the most versatile polyphenols for anti-inflammatory properties. Many studies have suggested that GTPs have shown positive effects in reducing the myeloperoxidase activity and UVB induced infiltration of inflammatory leukocytes. These polyphenols are also effective at reducing the level of proinflammatory cytokines when tested in UV irradiated skin. The polyphenol resveratrol has shown significant inhibitory effect in UVB induced skin thickness which is a marker symptom in edema due to skin inflammation [9]. There have been various research studies mentioned in the review that proves the anti-inflammatory nature of various plant polyphenols rich extract used in treating skin problems. Pathways associated with different polyphenols are mentioned in Table 1.

Anti-aging effects

Skin is continuously being subjected to aging process both by extrinsic and intrinsic factors. Intrinsic aging is

Table 1 Pathways associated with polyphenols for skin care

Skin care role	Pathways/protein involved	References
Photo-protective	MAPK, ps66Shc	[10, 11]
Antioxidant	Nrf2-ARE, MAPK	[12]
Anti-inflammatory	Nf-kB inhibition	[13]
Anti-aging effects	TOR pathway	[14]
Skin whitening effects	Melanin synthesis pathway	[15]
Anti-acne	Inhibition of NF-kb activation, reduction in inflammation	[16]
Collagenesis	TGFβ/Smad pathway	[17]
Free radical scavenging activity	Decrease in fluoride induced superoxide radical, reduction in lipid peroxidation	[18]
Antimicrobial	Inhibition of NF-kB activation, liberation of IL-8, reduction in inflammation	[19]
Wound healing	Reduction in p53, iNOS and IL-6	[20]
Cutaneous leishmaniasis	Downregulation of kinase	[21]

recognized by skin atrophy along with loss of skin elasticity thereby slowing the overall metabolic activities whereas extrinsic aging is influenced by extrinsic harmful UV radiations leading to premature aging characterized with thickened epidermis, collagen degradation and increased melanogenesis. Various research studies have proved that polyphenols act as potent anti-aging agent. Recently, a report proposed the effectiveness of phenolic extracts as an anti-ageing material for the cosmetics world [8]. There have been evidences which prove that topical polyphenols being used as cosmetics have potency to reverse the histological changes in the skin caused due to exposure to harmful UV rays and chronological aging process. They do so by improving vasculature, repairing damages caused to keratinocyte ultra-structure, depositing new papillary dermal collagen, normalizing hyper keratinization and improving enzymes responsible for moisturizing effects. Topical application of these polyphenols have shown promising effects being a natural sunscreen in providing additional photo protection and protecting skin from premature aging [22].

Skin whitening effects

The main role of melanin pigment is to prevent skin from damage but due to its accumulation in different regions of skin there is development of pigmented patches which are visibly considered as an aesthetic problem. There may also be excessive generation of melanin pigments and reactive species due to overexposure skin of UV irradiation that cause various different types of skin injuries including melasma, inflammation, freckles and aging. In this regard, polyphenols especially quercetin glycosides are found to have anti-melanogenesis property. They function by inhibition of melanin synthesis pathway [23]. Crude phenol extract, from a medicinal plant, *Malpighia emarginata* DC exhibited lightening of skin when exposed to UVB irradiation. Similar effects were seen in another report by Smeriglio et al. wherein phenolic extract from *Alna Cordata* [24]. Another polyphenol Arbutin, commonly found in dried leaves of plant like bearberry is a naturally occurring β -D-glucopyranoside that is found to have inhibiting action on melanosomal tyrosinase enzyme. The polyphenols functions via inhibition of proliferation of melanocytes and blocks the synthesis of melanin via inhibiting tyrosinase enzyme in melanocytes. The chemical structure of plant polyphenols has been illustrated in Fig. 3. Various polyphenols-based cosmetics have been mentioned in Table 1 of online resources named ESM 1 (supplementary file) and their roles in Fig. 1 of online resource named ESM 1 (supplementary file).

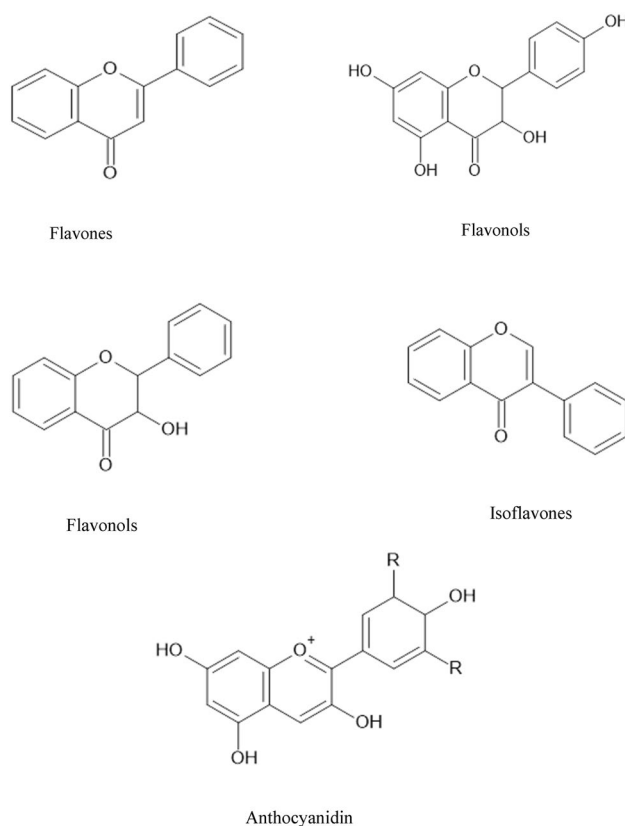


Fig. 3 Chemical structure of polyphenols

Polyphenol rich plants for skin problems

There has been a considerable growth for using natural phytochemicals especially polyphenols for various skin cosmetic formulations. The growing trend of natural plant-based products has increased competition among companies to find diverse plant based natural formulations effective for various skin problems which has ultimately increased the horizon of research. A summary of research studies focusing various plant-based extracts used for different skin problems has been presented in the Table 2.

Polyphenol rich nanocarrier drugs for skin

The dietary benefits of various edible plant are mainly contributed by its phytochemical composition including various polyphenols. But several studies have shown that when these polyphenols rich extract are applied topically the barriers of stratum corneum impedes the transdermal absorption which leads to minimized penetration of these active ingredients into the cells. This has led to the discovery of novel drug delivery system (NDDS) which are mainly involved in the enhanced and site specific delivery

Table 2 Plants with several dermatological importance

Plant species	Main polyphenols present	Properties	References
<i>Acacia nilotica</i>	Epigallocatechin-3-gallate, flavonoids, phenolic compounds	Anti-tumor, anti-inflammatory properties, anti-aging, radical scavenging	[25]
<i>Achillea</i> spp.	1,8-Cimneole	Anti-tyrosinase	[26]
<i>Achillea sivasica</i>			
<i>Agastache rugosa kuntze</i>	Phenylpropanoids, terpenoids, rosmarinic acid, tilianin, acacetin, agastachoside, methyl flavones, agastinol agastenol, apigenin, quercetin	Anti-aging, antioxidant, anti-elastase, anti-hyaluronidase	[27]
<i>Benincasa hispida</i>	Triterpenoids, flavonoids, carotene, glycosides, B sitosterin, saccharides, uronic acids	Anti-aging	[28]
<i>Camellia sinensis</i>	(-)-Epicatechin-3-gallate (ECG), (-)-epicatechin (EC), EGCG	Anti-aging, anti-carcinogenic anti-inflammatory	[29]
<i>Centella asiatica</i>	Ursane- and oleanane-type pentacyclic triterpenid, madecassoside, madecassic acid, asiaticoside, asiatic acid	Skin aging, anti-hyaluronidase, activity, anti-elastase, inhibits H ₂ O ₂ , induced antioxidant, anti-inflammatory property	[30]
<i>Coffea arabica</i>	Chlorogenic acid, quinic acid, ferulic acid, hydroxycinnamic acid, cafestol, kahweol	Wound healing, antioxidant, UV protectant, anti-photo carcinogenesis, anti-inflammatory	[31]
<i>Coriandrum sativum</i> L.	Quercetin, kaempferol, acacetin, <i>P</i> -coumaric acid, vanillic acid, cis and trans form of ferulic acid	Upregulates oxidative defense system. Protect from UVb induced skin damage	[32]
<i>Crataegus pinnatifida</i> Bge.	Oligomeric procyanidins and their glycosides, chlorogenic acid, rutin, quercetin, isoquercetin, epicatechin, gallic acid, 4-amino benzoic acid	Antioxidant, anti-photoaging collagenase inhibitory activity, anti-photo-aging, anti-inflammatory	[33, 34]
<i>Curcuma longa</i>	Curcumin	Anti-aging agent antioxidant anti-psoriasis agent	[35, 36]
<i>Cyclopia</i> spp.	Terpenoids, phenolic compounds		
	Xanthones, flavonones, heperidin, mangiferin	Antioxidant and anti-inflammatory, reduced signs of skin peeling, sunburn, reduced erythema, edema, skin hardening, modulated epidermal hyperplasia	[37]
<i>Embllica officinalis</i>	Flavonoids, phenolic acids like ellagic and gallic acid, tannins like puniglucoin, pedunculagin and emblicanin	Broad-spectrum antioxidant, antitumor anti-wrinkle, anti-tyrosinase, anti-inflammatory	[38]
<i>Eugenia dysenterica</i>	Quercetin and gallic acid	Lower down collagenase activity, reduce the ageing effects	[39]
<i>Foeniculum vulgare</i>	Linolenic, oleic and linoleic acid	Anti-photoaging	[40]
<i>Fragaria vesca</i> L.	Flavonoids, catechins, phenolic acids, ellagitannins, proanthocyanidins, ellagic acid	Anti-melanogenic, antioxidant, photoprotectant	[41]
<i>Hippophae rhamnoides</i>	Casuarinin, B-carotene and tocopherol quercetin, kaempferol, isorhamnetin, catechin, flavonoids, quercetin, oleic acid and linoleic acids, procyanidins, quercetin, kaempferol, myricetin, isorhamnetin, kaempferol-3-rutinoside	Anti-melanogenic properties, anti-aging effects reduced skin melanin and helped melasma patients, anti-inflammatory, UV-induced skin anti-aging regulating skin hydration, antioxidant, photoprotectant	[42, 43]
<i>Hydrangea serrata</i> (Thunb.) Ser	Hydrangenol	Photoprotectant, antioxidant, moisturizing properties, antiaging	[44, 45]
<i>Hypericum perforatum</i>	Quercetin	Inhibit the damage caused due to UVB	[46]
<i>Ixora parviflora</i>	Chrysin 5- <i>O</i> -B-D-xylopyranoside, chlorogenic acid	Anti-photoaging properties, antioxidant photo-protective effects	[47]
<i>Michelia alba</i>	(-) <i>N</i> -formylanonaine sesquiterpenes, terpenes, aporphines, benzenoids, oxoaporphines, steroids, lignans	Tyrosinase inhibiting activity, antioxidant properties. Protect the fibroblast from cellular oxidative damages due to UVB	[48, 49]
<i>Momordica charantia</i>	Charatin, flavonoids, normordin	Cytoprotective, antioxidative, tissue remodeling, hydrating, moisturizing and anti-pigmentation properties	[50]

Table 2 (continued)

Plant species	Main polyphenols present	Properties	References
<i>Myristica fragrans</i> Houtt	Macelignan	Antimelanogenic, inhibition of UVB induced inflammation and photo-aging of skin	[51]
<i>Panax ginseng</i>	Chlorogenic acid, syringic acid, kaempferol, quercetin, resveratrol, naringenin, gentisic acid, rutin, catechin, <i>N</i> - and <i>P</i> -coumaric acid, vanillic acid	Anti-skin aging, photo-protective effect, anti-melanogenic, anti-tyrosinase activity	[52, 53]
	Flavonoids		
<i>Patrinia villosa</i>	Kaempferol	Anti-melanogenic	[54]
<i>Penthorum Chinese pursh</i>	Quercetin	Antioxidant, anti-aging agent, UVB protectant	[55]
<i>Polypodium leucotomos</i>	Ferulic, caffeic acids, cinnamic acid chlorogenic acids	Antioxidant, anti-photoaging, anti-photocarcinogenesis	[56]
<i>Populus nigra</i>	Caffeic acids, isofurlic, cinnamic, salicin	Anti-oxidant activity along with anti-aging	[57]
<i>Punica granatum</i>	Ellagitannins, ellagic acid, punicalagins, anthocyanins	Anti-photoaging, anti-metastatic, anti-proliferative, antioxidant, photo-protection	[58]
<i>Rhus coriaria</i> L.	Tannins, myricetin derivatives, quercetin and gallic acid flavonoids, phenolic acid and gallotannins	Antioxidant, anti-carcinogenic, antifibrogenic, genoprotective effect	[59, 60]
<i>Spatholobus suberectus</i>	Formononetin, butin	Anti-tyrosinase activity, inhibiting UVB induced ROS production	[61]
<i>Theobroma cacao</i> L.	Monomeric (-) epicatechin and (+) catechin, proanthocyanidin, flavonols	Antioxidant, anti-inflammatory, immunomodulatory, photo protection	[62]

of therapeutic biomolecules [63]. These drug carriers facilitate the delivery of big molecules across epidermal barriers in skin. The advancement in nanotechnology has made possible the delivery of complete therapeutic molecule at its targeted site without any degradation. These nanocarriers are designed to enhance polyphenols transport across the epithelium, enhance their distribution and its pharmacokinetics, improve stability, reduced irritation of skin, controlled release of bioactive and thereby enhancing its efficacy at cellular level [64].

The therapeutic activity of polyphenols imparted to the skin depends upon its concentration that reaches the skin. These nanocarriers ensure to transport large size polyphenols across epidermis and enable their biological functions. Many researches are being carried on the next generation cosmetics i.e., nanocarriers for their enhanced activity. The photo protection imparted by polyphenols is related to their amount which reaches the skin cell. Many polyphenols are found to have excellent clinical use but due to their poor bioavailability has limited their use. EGCG being potential candidate for sunscreens but having poor absorption capacity nano-transfersomal formulations of EGCG has been developed to enhance permeation and efficient delivery into stratum corneum [65]. These nano-transferosomes gave excellent protection from UV radiations in addition to their anti-aging and antioxidant activities by their higher permeation and deposition capacity into cells which increased the cell viability, lowered intracellular free radicals and reduced expression of matrix metalloproteinases in HeLa cells [66].

Nanocarriers is one of the emerging drug delivery strategies. Nanoparticles are classified into different categories namely (a) liposomes-based nanoparticles wherein liposomes are in form of circular enclosed vesicles in colloidal size range where outer phospholipids layer contains water soluble drug to be delivered. (b) Nano capsule comprising inner active ingredient and out polymembrane surface. (c) Solid lipid nanoparticles in which lipid colloidal carries dispersed in aqueous solution. (d) Aggregation-based nanocrystal where numerous atoms form cluster and are, widely used for the transportation of poor soluble drugs. Lipid based noncarriers products are widely used in cosmetics sector. second generation lipid carries possess the ability from leakage and contain high drug capacity [67]. Beside this lipid-based nanoparticle for topical use include nano emulsions, nanostructure lipid carriers and solid liquid nanoparticles. In a study, solid lipid nanoparticle containing resveratrol is developed whose uptake and transportation is studied in keratinocytes. This nanoparticle prevents the active ingredient i.e., resveratrol from undergoing photo degradation and lipoperoxidation. Nano cosmetics containing flavonoids inside are either present in single or a mix of different ingredients such as resveratrol

and curcumin [68]. Various nanoparticle along with their examples and bioactivity is mentioned in Table 3.

Catechin being a good photo protectant has limited bioavailability due to its less half time and high rate of biotransformation. So, nano based emulsion and gel of catechin was developed to enhance its permeability across skin to maximize its photo protection and sunscreen property in a sustained manner [81]. The nanoemulsion made from ethyl acetate fraction of pomegranate rich in gallic acid, ellagic acid and punicalagin enhanced the delivery of polyphenols to skin when compared to free solution of pomegranate. The research showed enhancement in antioxidant property against free radicals due to UV irradiation upon topical application of nanoemulsion. This also enhanced the absorption capacity of UV radiations upon its application and hence can be incorporated in sunscreen products [82]. It has also been seen that curcumin loaded nanocarriers enhanced their delivery and penetration capability and can be used in cream formulations to treat skin conditions like aging, acne, UV induced damages, skin cancer etc. [83]. The nanoberries developed by conjugating blueberry extract and liposomes were found effective in successful penetration of antioxidants upon topical application and protected skin from photodamage when tested in zebrafish.

Future prospects

Plants being easily available, rich in natural phytochemicals mainly polyphenols are cheap source for gaining traditional and excellent therapeutic health benefits [84]. Plant based cosmetics has achieved a new attraction in the eyes of customers for their safe results in skin care products. However, the complete exact knowledge of their biological activity and phytochemical composition of plants extract is predominant but not adequate. The important issues that must be administered are plant identification, harvesting including their post-harvest treatment. The main principal step is the identification of key ingredient from the raw extract and to know about the molecular target of the phytochemical affecting the activity for providing desired results. The knowledge of genetic profile of that particular plant and the effect of environment must be taken care of as this can significantly affect the phytochemical profile of the extracted chemical affecting the well-being of customer. Some species of medicinal plants being extensively used in cosmetics are being on the verge of getting endangered. It is therefore significant to devise an alternative biotechnological way to produce phytochemicals without altering their chemical composition and promoting the culturing of endangered medicinal plants via cell culturing methods. The growth of green chemistry being

Table 3 Nanoparticle from polyphenols and their bioactivity

Nanocarrier	Polyphenol	Method of preparation	Bioactivity details	References
Lipid nanoparticles	Resveratrol	Phase inversion temperature	Increased skin hydration	[69]
Solid-lipid nanoparticles	Quercetin	Ultra-sonification	Deeper penetration Delayed in UV damage	[70]
Solid-lipid nanoparticles	EGCG	High pressure homogenization	Antioxidant properties	[71]
Nano transferosomes	EGCG + hyaluronic acid	High pressure homogenization	UV protection, skin ageing and antioxidant properties	[72]
Solid-lipid nanoparticles	Curcumin-resveratrol	High shear homogenization	Anti-melanogenic	[73]
Nanostructure lipid nanoparticles	Curcumin	Homogenization	Psoriasis treatment	[74]
Niosomes	Curcumin	Reverse evaporation	Restoration of protein level, reduction in epidermal hyperplasia, anti-ageing, anti-wrinkle	[75]
Nanoencapsulation	Catechin	Solvent evaporation sonification	Antioxidant, antiaging	[76]
Liposphere	Curcumin + tacrolimus	Homogenization	Psoriasis treatment Deeper penetration	[77]
Liquid crystal nanoparticles	Curcumin	Homogenization	Anti-bacterial efficacy, deeper penetration	[78]
Liposomes	Hibiscus sabdariffa	Ethanol injection	Permeation in skin, lowered skin irritation, good stability	[79]
Emulsification based polymeric nanoparticles	Naringenin + PGLA (poly D, L lactide-c—glycolide solution	Homogenization	Increase skin retention Protection from free radical	[80]

customer friendly has led to increase the horizon for using plant extract in cosmetics due to their inherent nature of being safe than synthetic chemicals used in cosmetics. But their side effects must also be addressed and needs to be properly investigated. Targeted bioactive chemicals should be used in in-vitro and in-vivo pharmaceutical research to know the efficacy and side effects of product in long term use. Despite all the hurdles discussed polyphenols still continue to be a potent and serious candidate in the therapeutic sector to enhance skin properties. Polyphenols provide a vast field of exploration and seems to be open for research studies for various skin-protecting benefits.

Conclusion

The skin-protecting effects of polyphenols has been demonstrated by various research studies as mentioned in the review. It is clear from the review that bioactive ingredients from plants impose beneficial effects on skin whether applied topically or taken orally and protects the skin from various environmental factors and improving its function. This is important to offer an alternative to existing synthetic topical medications for maintaining youthful healthy skin. The polyphenols are a great source of antioxidants and these properties of polyphenols have been attributed due to different chemical structure of polyphenols. The ability of polyphenols to interfere in different metabolic pathways to attenuate the skin problems make them promising attractive candidate for use in cosmetic products. The different studies conclude that polyphenols have significant photo-protective, anti-inflammatory, antimicrobial, anti-melanogenic, anti-aging, antioxidant, anti-carcinogenic properties. The expanding popularity of plants-based product for cosmetic use needs extensive research on the polyphenols to get proper information on the interaction of polyphenols and molecular markers involved in various skin problems. The proper research is still lagging to extrapolate the results obtained in vitro studies in model organism and the use of effective chemical for human skin. Moreover, it is important for the patients and researchers to realize that compliance is very necessary while using natural cosmetics as they are slow in action than conventional synthetically derived cosmetics. The current scenario seems to be slow and inadequate with respect to research. More research is needed to investigate and evolve plant-based therapeutics as an alternative to chemically derived synthetic medications.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11033-022-08156-9>.

Funding The authors did not receive financial support from any organization for the submitted work.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Research involving human and/or animal participants This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Gallo RL (2017) Human skin is the largest epithelial surface for interaction with microbes. *J Invest Dermatol* 137(6):1213–1214
- Zouboulis CC, Ganceviciene R, Liakou AI, Theodoridis A, Elewa R, Makrantonaki E (2019) Aesthetic aspects of skin aging, prevention, and local treatment. *Clin Dermatol* 37(4):365–372. <https://doi.org/10.1016/j.clindermatol.2019.04.002>
- Zedan H, Abdel-Motaleb AA, Kassem NMA, Hafeez HAA, Hussein MRA (2015) Low glutathione peroxidase activity levels in patients with vitiligo. *J Cutan Med Surg* 19(2):144–148
- Ratz-Lyko A, Arct J, Majewski S, Pytkowska K (2015) Influence of polyphenols on the physiological processes in the skin. *Phytother Res* 29(4):509–517. <https://doi.org/10.1002/ptr.5289>
- Eskandari M, Rembiesa J, Startaitė L, Holfors A, Valančiūtė A, Faridbod F, Ganjali MR, Engblom J, Ruzgas T (2019) Polyphenol-hydrogen peroxide reactions in skin: in vitro model relevant to study ROS reactions at inflammation. *Anal Chim Acta* 1075:91–97
- Afaq F, Katiyar SK (2011) Polyphenols: skin photoprotection and inhibition of photocarcinogenesis. *Mini Rev Med Chem* 11(14):1200–1215. <https://doi.org/10.2174/13895575111091200>
- Moreira LC, de Ávila RI, Veloso D et al (2017) In vitro safety and efficacy evaluations of a complex botanical mixture of *Eugenia dysenterica* DC. (Myrtaceae): prospects for developing a new dermocosmetic product. *Toxicol In Vitro* 45:397–408
- de Lima Cherubim DJ, Buzanello Martins CV, Oliveira Fariña L, da Silva de Lucca RA (2020) Polyphenols as natural antioxidants in cosmetics applications. *J Cosmet Dermatol* 19(1):33–37
- Mandal SM, Chakraborty D, Dey S (2010) Phenolic acids act as signaling molecules in plant-microbe symbioses. *Plant Signal Behav* 5(4):359–368. <https://doi.org/10.4161/psb.5.4.10871>
- Chiang HS, Wu WB, Fang JY, Chen BH, Kao TH, Chen YT, Huang CC, Hung CF (2007) UVB-protective effects of isoflavone extracts from soybean cake in human keratinocytes. *Int J Mol Sci* 8:651–661
- Wang XF, Huang YF, Wang L, Xu LQ, Yu XT, Liu YH, Li CL, Zhan JY, Su ZR, Chen JN, Zeng HF (2016) Overec activity of pogostone against UV-induced skin premature aging in mice. *Exp Gerontol* 77:76–86
- Yan Z, Zhong Y, Duan Y, Chen Q, Li F (2020) Antioxidant mechanism of tea polyphenols and its impact on health benefits. *Anim Nutr* 6(2):115–123. <https://doi.org/10.1016/j.aninu.2020.01.001>
- Kundu JK, Surh YJ (2007) Epigallocatechin gallate inhibits phorbol ester-induced activation of NF-KB and CREB in mouse skin role of P38 MAPK. *Ann N Y Acad Sci* 1095:504–512. <https://doi.org/10.1196/annals.1397.054>
- Liu L, Guo P, Wang P, Zheng S, Qu Z, Liu N (2021) The review of anti-aging mechanism of polyphenols on *Caenorhabditis elegans*.

- Front Bioeng Biotechnol 1:635768. <https://doi.org/10.3389/fbioe.2021.635768>
15. Hanamura T, Uchida E, Aoki H (2008) Skin-lightening effect of a polyphenol extract from Acerola (*Malpighia emarginata* DC.) fruit on UV-induced pigmentation. Biosci Biotechnol Biochem 72(12):3211–3218. <https://doi.org/10.1271/bbb.80421>
 16. Saric S, Notay M, Sivamani RK (2016) Green tea and other tea polyphenols: effects on sebum production and acne vulgaris. Antioxidants (Basel) 6(1):2. <https://doi.org/10.3390/antiox6010002>
 17. Semkova M, Hsuan J (2021) TGF β -1 induced cross-linking of the extracellular matrix of primary human dermal fibroblasts. Int J Mol Sci 22:984
 18. Xia EQ, Deng GF, Guo YJ, Li HB (2010) Biological activities of polyphenols from grapes. Int J Mol Sci 11(2):622–46. <https://doi.org/10.3390/ijms11020622>
 19. Truong VL, Jeong WS (2021) Cellular defensive mechanisms of tea polyphenols: structure-activity relationship. Int J Mol Sci 22(17):9109. <https://doi.org/10.3390/ijms22179109>
 20. Shi HP, Most D, Efron DT, Tantry U, Fischel MH, Barbul A (2001) The role of iNOS in wound healing. Surgery 130(2):225–259. <https://doi.org/10.1067/msy.2001.115837> (Erratum in: **Surgery** 2001 Nov;130(5):808)
 21. Feily A, Yaghoobi R, Reza M (2009) The potential utility of green tea extract as a novel treatment for cutaneous leishmaniasis. J Altern Complement Med 5:815–816
 22. Mena F, Mena A, Tréton J (2013) Polyphenols against skin aging. In: Polyphenols in human health and disease. Elsevier Inc., Amsterdam, pp 819–830. <https://doi.org/10.1016/B978-0-12-398456-2.00063-3>
 23. Hanamura T, Uchida E, Aoki H (2008) Skin-lightening effect of a polyphenol extract from Acerola (*Malpighia emarginata* DC.) fruit on UV-induced pigmentation. Biosci Biotechnol Biochem 72(12):3211–3218
 24. Smeriglio A, D'Angelo V, Denaro M, Trombetta D, Raimondo F, Germanò M (2019) Polyphenol characterization, antioxidant and skin whitening properties of *Alnus cordata* (Loisel.) Duby stem bark. Chem Biodivers. <https://doi.org/10.1002/cbdv.201900314>
 25. Kalaivani T, Mathew L (2010) Free radical scavenging activity from leaves of *Acacia nilotica* (L.) Willd. Ex Delile, an Indian medicinal tree. Food Chem Toxicol 48:298–305. <https://doi.org/10.1016/j.fct.2009.10.013>
 26. Haliloglu Y, Ozek T, Tekin M, Goger F, Can Baser KH, Ozek G, Can Baser H (2017) Phytochemicals, antioxidant, and antityrosinase activities of *Achillea sivasica* Çelik and Akpulat. Int J Food Prop. <https://doi.org/10.1080/10942912.2017.1308954>
 27. Zielińska A, Matkowski (2014) Phytochemistry and bioactivity of aromatic and medicinal plants from the genus Agastache (Lamiaceae). Phytochem Rev 13:391–416. <https://doi.org/10.1007/s11101-014-9349-1>
 28. Sabale V, Kunjwani H, Sabale P (2011) Formulation and in vitro evaluation of the topical antiageing preparation of the fruit of *Benincasa hispida*. J Ayurveda Integr Med 2:124–128
 29. OyetakinWhite P, Tribout H, Baron E (2012) Protective mechanisms of green tea polyphenols in skin. Oxid Med Cell Longev 2012:560682. <https://doi.org/10.1155/2012/560682>
 30. Kim YJ, Cha HJ, Nam KH, Yoon Y, Lee H, An S (2011) *Centella asiatica* extracts modulate hydrogen peroxide-induced senescence in human dermal fibroblasts. Exp Dermatol 20:998–1003. <https://doi.org/10.1111/j.1600-0625.2011.01388.x>
 31. Affonso RCL, Voytena APL, Fanan S, Pitz H, Coelho DS, Horstmann AL, Pereira A, Uarrota VG, Hillmann MC, Varela LAC, Ribeiro-Do-Valle RM, Maraschin M, Phytochemical, Composition (2016) Antioxidant activity, and the effect of the aqueous extract of coffee (*Coffea arabica* L.) bean residual press cake on the skin wound healing. Oxid Med Cell Longev. <https://doi.org/10.1155/2016/1923754>
 32. Park G, Kim HG, Kim YO, Park SH, Kim SY, Oh MS (2012) *Coriandrum sativum* L. protects human keratinocytes from oxidative stress by regulating oxidative defense systems. Skin Pharmacol Physiol 25:93–99. <https://doi.org/10.1159/000335257>
 33. Jurikova T, Sochor J, Rop O, Mlcek J, Balla S, Szekeres L, Adam V, Kizek R (2012) Polyphenolic profile and biological activity of chinese hawthorn (*Crataegus pinnatifida* Bunge) fruits. Molecules 17(12):14490–14509. <https://doi.org/10.3390/molecules171214490>
 34. Moon HI, Kim T, Cho HS, Kim EK (2010) Identification of potential and selective collagenase, gelatinase inhibitors from *Crataegus pinnatifida*. Bioorg Med Chem Lett. <https://doi.org/10.1016/j.bmcl.2009.12.059>
 35. Panahi Y, Fazlollahzadeh O, Atkin SL, Majeed M, Butler AE, Johnston TP, Sahebkar A (2019) Evidence of curcumin and curcumin analogue effects in skin diseases: a narrative review. J Cell Physiol 234(2):1165–1178. <https://doi.org/10.1002/jcp.27096>
 36. Perrone D, Ardito F, Giannatempo G, Dioguardi M, Troiano G, Lo Russo L, DE Lillo A, Laino L, Lo Muzio L (2015) Biological and therapeutic activities, and anticancer properties of curcumin. Exp Ther Med 10(5):1615–1623. <https://doi.org/10.3892/etm.2015.2749>
 37. Petrova A, Davids LM, Rautenbach F, Marnewick JL (2011) Photoprotection by honeybush extracts, hesperidin and mangiferin against UVB-induced skin damage in SKH-1 mice. J Photochem Photobiol B 103(2):126–139. <https://doi.org/10.1016/j.jphotobiol.2011.02.020>
 38. Chaikul P, Kanlayavattanakul M, Somkumnerd J, Lourith N (2021) *Phyllanthus emblica* L. (amla) branch: a safe and effective ingredient against skin aging. J Tradit Complement Med 11(5):390–399. <https://doi.org/10.1016/j.jtcme.2021.02.004>
 39. Moreira LC, de Ávila RI, Veloso DFMC, Pedrosa TN, Lima ES, do Couto RO, Lima EM, Batista AC, de Paula JR, Valadares MC, (2017) In vitro safety and efficacy evaluations of a complex botanical mixture of *Eugenia dysenterica* DC. (Myrtaceae): Prospects for developing a new dermocosmetic product. Toxicol In Vitro 45(Pt 3):397–408. <https://doi.org/10.1016/j.tiv.2017.04.002>
 40. He W, Huang B (2011) A review of chemistry and bioactivities of a medicinal spice: *Foeniculum vulgare*. J Med Plants Res 5:3595–3600
 41. Gasparrini M, Forbes-Hernandez TY, Afrin S, Reboredo-Rodriguez P, Cianciosi D, Mezzetti B, Quiles JL, Bompadre S, Battino M, Giampieri F (2017) Strawberry-based cosmetic formulations protect human dermal fibroblasts against UVA-induced damage. Nutrients 9(6):605. <https://doi.org/10.3390/nu9060605>
 42. Kwon DJ, Bae YS, Ju SM, Goh AR, Choi SY, Park J (2011) Casuarinin suppresses TNF- α -induced ICAM-1 expression via blockade of NF- κ B activation in HaCaT cells. Biochem Biophys Res Commun 409(4):780–785. <https://doi.org/10.1016/j.bbrc.2011.05.088>
 43. Hwang IS, Kim JE, Choi SI, Lee HR, Lee YJ, Jang MJ, Son HJ, Lee HS, Oh CH, Kim BH, Lee SH, Hwang DY (2012) UV radiation-induced skin aging in hairless mice is effectively prevented by oral intake of sea buckthorn (*Hippophae rhamnoides* L.) fruit blend for 6 weeks through MMP suppression and increase of SOD activity. Int J Mol Med 30(2):392–400. <https://doi.org/10.3892/ijmm.2012.1011>
 44. Myung DB, Han HS, Shin JS, Park JY, Hwang HJ, Kim HJ, Ahn HS, Lee SH, Lee KT (2019) Hydrangenol isolated from the leaves of *Hydrangea serrata* attenuates wrinkle formation and repairs skin moisture in UVB-irradiated hairless mice. Nutrients 11(10):2354. <https://doi.org/10.3390/nu11102354>
 45. Myung DB, Lee JH, Han HS, Lee KY, Ahn HS, Shin YK, Song E, Kim BH, Lee KH, Lee SH, Lee KT, Myung DB, Lee JH, Han HS, Lee KY, Ahn HS, Shin YK, Song E, Kim BH, Lee KH, Lee SH, Lee KT (2020) Oral intake of *Hydrangea serrata* (Thunb.) Ser.

- leaves extract improves wrinkles, hydration, elasticity, texture, and roughness in human skin: a randomized, double-blind, placebo-controlled study. *Nutrients* 12(6):1588. <https://doi.org/10.3390/nu12061588>
46. Zhu X, Zeng X, Zhang X, Cao W, Wang Y, Chen H, Wang T, Tsai HI, Zhang R, Chang D, He S, Mei L, Shi X (2016) The effects of quercetin-loaded PLGA-TPGS nanoparticles on ultraviolet B-induced skin damages in vivo. *Nanomedicine* 12(3):623–632. <https://doi.org/10.1016/j.nano.2015.10.016>
 47. Wen KC, Chiu HH, Fan PC, Chen CW, Wu SM, Chang JH, Chiang HM (2011) Antioxidant activity of *Ixora parviflora* in a cell/cell-free system and in UV-exposed human fibroblasts. *Molecules* 16(7):5735–5752. <https://doi.org/10.3390/molecules16075735>
 48. Chiang HM, Chen HC, Lin TJ, Shih IC, Wen KC (2012) *Michelia alba* extract attenuates UVB-induced expression of matrix metalloproteinases via MAP kinase pathway in human dermal fibroblasts. *Food Chem Toxicol* 50(12):4260–4269. <https://doi.org/10.1016/j.fct.2012.08.018>
 49. Wang HM, Chen CY, Chen CY, Ho ML, Chou YT, Chang HC, Lee CH, Wang CZ, Chu IM (2010) (-)-N-formylanonaine from *Michelia alba* as a human tyrosinase inhibitor and antioxidant. *Bioorg Med Chem* 18(14):5241–5247
 50. Park SH, Yi YS, Kim MY, Cho JY (2019) Antioxidative and anti-melanogenesis effect of *Momordica charantia* methanol extract. *Evid Based Complement Altern Med*. <https://doi.org/10.1155/2019/5091534>
 51. Cho Y, Kim KH, Shim JS, Hwang JK (2008) Inhibitory effects of macelignan isolated from *Myristica fragrans* Houtt. on melanin biosynthesis. *Biol Pharm Bull* 31(5):986–989. <https://doi.org/10.1248/bpb.31.986>
 52. Kang TH, Park HM, Kim YB, Kim H, Kim N, Do JH, Kang C, Cho Y, Kim SY (2009) Effects of red ginseng extract on UVB irradiation-induced skin aging in hairless mice. *J Ethnopharmacol* 123(3):446–451. <https://doi.org/10.1016/j.jep.2009.03.022>
 53. Lee HJ, Kim JS, Song MS, Seo HS, Moon C, Kim JC, Jo SK, Jang JS, Kim SH (2009) Photoprotective effect of red ginseng against ultraviolet radiation-induced chronic skin damage in the hairless mouse. *Phytother Res* 23(3):399–403. <https://doi.org/10.1002/ptr.2640>
 54. Jeong D, Park SH, Kim MH, Lee S, Cho YK, Kim YA, Park BJ, Lee J, Kang H, Cho JY (2020) Anti-melanogenic effects of ethanol extracts of the leaves and roots of *Patrinia villosa* (Thunb.) Juss through their inhibition of CREB and induction of ERK and autophagy. *Molecules* 25(22):5375. <https://doi.org/10.3390/molecules25225375>
 55. Jeong D, Lee J, Park SH, Kim YA, Park BJ, Oh J, Sung GH, Aravinthan A, Kim JH, Kang H, Cho JY (2019) Antiphotoreaging and antimelanogenic effects of *Penthorum chinense* pursh ethanol extract due to antioxidant- and autophagy-inducing properties. *Oxid Med Cell Longev* 2019:9679731. <https://doi.org/10.1155/2019/9679731>
 56. Parrado C, Mascaraque M, Gilaberte Y, Juarranz A, Gonzalez S (2016) Fernblock (*Polypodium leucotomos* extract): molecular mechanisms and pleiotropic effects in light-related skin conditions, photoaging and skin cancers, a review. *Int J Mol Sci* 17(7):1026. <https://doi.org/10.3390/ijms17071026>
 57. Spagnol CM, Di Filippo LD, Isaac VLB, Correa MA, Salgado HRN (2017) Caffeic acid in dermatological formulations: in vitro release profile and skin absorption. *Comb Chem High Throughput Screen* 20(8):675–681. <https://doi.org/10.2174/1386207320666170602090448>
 58. Turrini E, Ferruzzi L, Fimognari C (2015) Potential effects of pomegranate polyphenols in cancer prevention and therapy. *Oxid Med Cell Longev* 2015:938475. <https://doi.org/10.1155/2015/938475>
 59. Gabr SA, Alghadir AH (2019) Evaluation of the biological effects of lyophilized hydrophilic extract of *Rhus coriaria* on myeloperoxidase (MPO) activity, wound healing, and microbial infections of skin wound tissues. *Evid Based Complement Alternat Med* 2019:5861537. <https://doi.org/10.1155/2019/5861537>
 60. Nozza E, Melzi G, Marabini L, Marinovich M, Piazza S, Khalilpour S, Dell'Agli M, Sangiovanni E (2020) *Rhus coriaria* L. fruit extract prevents UV-A-induced genotoxicity and oxidative injury in human microvascular endothelial cells. *Antioxidants* (Basel) 9(4):292. <https://doi.org/10.3390/antiox9040292>
 61. Lee MH, Lin YP, Hsu FL, Zhan GR, Yen KY (2006) Bioactive constituents of *Spatholobus suberectus* in regulating tyrosinase-related proteins and mRNA in HEMn cells. *Phytochemistry* 67(12):1262–1270. <https://doi.org/10.1016/j.phytochem.2006.05.008>
 62. Scapagnini G, Davinelli S, Di Renzo L, De Lorenzo A, Olarte HH, Micali G, Cicero AF, Gonzalez S (2014) Cocoa bioactive compounds: significance and potential for the maintenance of skin health. *Nutrients* 6(8):3202–3213. <https://doi.org/10.3390/nu6083202>
 63. Ajazuddin, Saraf S (2010) Applications of novel drug delivery system for herbal formulations. *Fitoterapia* 81:680–689. <https://doi.org/10.1016/j.fitote.2010.05.001>
 64. Działo M, Mierziak J, Korzun U, Preisner M, Szopa J, Kulma A (2016) The potential of plant phenolics in prevention and therapy of skin disorders. *Int J Mol Sci* 17(2):160. <https://doi.org/10.3390/ijms17020160>
 65. Avadhani KS, Manikkath J, Tiwari M, Chandrasekhar M, Godavarthi A, Vidya SM, Hariharapura RC, Kalthur G, Udupa N, Mutalik S (2017) Skin delivery of epigallocatechin-3-gallate (EGCG) and hyaluronic acid loaded nano-transfersomes for antioxidant and anti-aging effects in UV radiation induced skin damage. *Drug Deliv* 24(1):61–74
 66. Działo M, Mierziak J, Korzun U, Preisner M, Szopa J, Kulma A (2016) The potential of Plant phenolics in prevention and therapy of skin disorders. *Int J Mol Sci* 17(2):160. <https://doi.org/10.3390/ijms17020160>
 67. García-Pinel B, Porras-Alcalá C, Ortega-Rodríguez A, Sarabia F, Prados J, Melguizo C, López-Romero JM (2019) Lipid-based nanoparticles: application and recent advances in cancer treatment. *Nanomaterials* (Basel) 9(4):638. <https://doi.org/10.3390/nano9040638>
 68. Sheng X, Zhu Y, Zhou J, Yan L, Du G, Liu Z, Chen H (2021) Antioxidant effects of caffeic acid lead to protection of *Drosophila* intestinal stem cell aging. *Front Cell Dev Biol* 9:735483. <https://doi.org/10.3389/fcell.2021.735483>
 69. Montenegro L, Parenti C, Turnaturi R, Pasquinucci L (2017) Resveratrol-loaded lipid nanocarriers: correlation between in vitro occlusion factor and in vivo skin hydrating effect. *Pharmaceutics* 9:58. <https://doi.org/10.3390/pharmaceutics9040058>
 70. Bose S, Du Y, Takhistov P, Michniak-Kohn B (2013) Formulation optimization and topical delivery of quercetin from solid lipid based nanosystems. *Int J Pharm* 441:56–66
 71. Shtay R, Keppler JK, Schrader K, Schwarz K (2019) Encapsulation of (-)-epigallocatechin-3-gallate (EGCG) in solid lipid nanoparticles for food applications. *J Food Eng* 244:91–100. <https://doi.org/10.1016/j.jfoodeng.2018.09.008>
 72. Avadhani KS, Manikkath J, Tiwari M, Chandrasekhar M, Godavarthi A, Vidya SM, Hariharapura RC, Kalthur G, Udupa N, Mutalik S (2017) Skin delivery of epigallocatechin-3-gallate (EGCG) and hyaluronic acid loaded nano-transfersomes for antioxidant and anti-aging effects in UV radiation induced skin damage. *Drug Deliv* 24(1):61–74
 73. Gumireddy A, Christman R, Kumari D et al (2019) Preparation, characterization, and in vitro evaluation of curcumin- and

- resveratrol-loaded solid lipid nanoparticles. AAPS PharmSciTech 20:145. <https://doi.org/10.1208/s12249-019-1349-4>
74. Esposito E, Sticozzi C, Ravani L, Drechsler M, Muresan XM, Cervellati F et al (2015) Effect of new curcumin-containing nanostructured lipid dispersions on human keratinocytes proliferative responses. Exp Dermatol 24(6):449–454
 75. Gupta NK, Dixit VK (2011) Development and evaluation of vesicular system for curcumin delivery. Arch Dermatol Res 303(2):89–101. <https://doi.org/10.1007/s00403-010-1096-6>
 76. Aljuffali IA, Lin CH, Yang SC, Alalaiwe A, Fang JY (2022) Nanoencapsulation of tea catechins for enhancing skin absorption and therapeutic efficacy. AAPS PharmSciTech 23(6):187. <https://doi.org/10.1208/s12249-022-02344-3>
 77. Jain A, Doppalapudi S, Domb AJ, Khan W (2016) Tacrolimus and curcumin co-loaded liposphere gel: synergistic combination towards management of psoriasis. J Control Release 243:132–145
 78. Archana A, Sri KV, Madhuri M, Kumar CA (2015) Curcumin loaded nano cubosomal hydrogel: preparation, in vitro characterization and antibacterial activity. Chem Sci Trans 4:75–80
 79. Pinsuwan S, Amnuait T, Ungphaiboon S, Itharat A (2010) Liposome-containing *Hibiscus sabdariffa* calyx extract formulations with increased antioxidant activity, improved dermal penetration and reduced dermal toxicity. J Med Assoc Thai 93(Suppl 7:S216-26):S216–26
 80. Joshi H, Hegde AR, Shetty PK, Gollavilli H, Managuli RS, Kalthur G, Mutalik S (2018) Sunscreen creams containing naringenin nanoparticles: formulation development and in vitro and in vivo evaluations. Photodermatol Photoimmunol Photomed 34(1):69–81. <https://doi.org/10.1111/phpp.12335>
 81. Scapagnini G, Davinelli S, Di Renzo L, De Lorenzo A, Olarte HH, Micali G, Cicero AF, Gonzalez S (2014) Cocoa bioactive compounds: significance and potential for the maintenance of skin health. Nutrients 6(8):3202–3213. <https://doi.org/10.3390/nu6083202>
 82. Ajazuddin S (2010) Applications of novel drug delivery system for herbal formulations. Fitoterapia 81:680–689. <https://doi.org/10.1016/j.fitote.2010.05.001>
 83. Rafiee Z, Nejatian M, Daeihamed, Jafari SM (2019) Application of curcumin-loaded nanocarriers for food, drug and cosmetic purposes. Trends Food Sci Technol 88:445–458. <https://doi.org/10.1016/j.tifs.2019.04.017>
 84. Singh H, Bharadvaja N (2021) Treasuring the computational approach in medicinal plant research. Prog Biophys Mol Biol 164:19–32. <https://doi.org/10.1016/j.pbiomolbio.2021.05.004>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Neural Underpinnings of Decoupled Ethical Behavior in Adolescents as an Interaction of Peer and Personal Values

Manvi Jain¹, Karsheet Negi², Pooja S. Sahni¹, Jyoti Kumar¹

¹Department of Design, Indian Institute of Technology, Delhi, 110116, India

²Department of Design, Delhi Technological University, Delhi, 110042, India

Abstract

In the present study, we are trying to understand how peer unethical behavior stimulates the decoupling of emotions in adolescents. We have simulated an interactive game-based environment in order to stimulate participants to make decisions that are found to be correlated with their virtual partner's decisions. The responses given by participants were also recorded as neural signals using an EEG to study neurophysiological correlates of different decision-making behavioral patterns. There was an active correlation between personality values and decision-making. Preliminary analysis was focused on studying the differences in lower brain frequencies (0.1-4Hz) when the participants developed 'frustration', in contrast to when they experienced 'gratitude'. The study presents three case studies in which delta frequencies increased in cases when frustration was experienced and decreased when gratitude was experienced. The study focused on understanding the neural underpinnings of corresponding modified behavior in adolescents. The findings highlight an increase in delta frequencies when apparent 'frustration' was developed in adolescents due to their peers' unethical behavior. The delta frequencies lowered when participants were tested for ethical behavior. The results concluded that based on personality value types, adolescents tend to develop frustration toward perceived unethical behavior and carry it over to other unrelated peers. This study is highly explorative in nature, with preliminary analysis using only three case studies, having a small sample size. However, the novelty of this study brings about new dimensions to social cognition and personality studies.

Keywords: *Moral behavior, Adolescent, Social cognition, Peer influence, Schwartz Model, Personality, Emotions, Frustration, Gratitude*

Introduction

Ethical Behavior in Adolescents: Some theories state that morality becomes a part of an individual's self-concept during adolescence (Colby and Damon 1992; Hardy and Carlo 2005; Moshman n.d.) forming one's moral identity. Strong moral identity engages individuals in moral actions as it drives them toward a sense of obligation to behave in ways that align with their moral values (Hardy and Carlo 2005; Blasi 1983). Moral judgment in adolescents contributes majorly to morally relevant actions that may include responsibility towards society, prosocial acts, moral actions toward peers, etc. (Hart, Atkins, and Donnelly, n.d.; Hertz and Krettenauer 2016).

Role of Personal values in Ethical Behavior: From a social-cognitive perspective, moral identity is a cognitive representation of moral values, goals, traits and behavioral scripts (Pohling et al. 2016; Hannah, Avolio, and May 2011). Therefore, personal values and personality can be identified as specific aspects of moral behavior. Values are

used to characterize cultural groups, societies, and individuals, to trace change over time, and to explain the motivational bases of attitudes and behavior. Recent theoretical and methodological developments (Schwartz 1992; Berry et al. 1997) have brought about a resurgence of research on values. Schwartz's theory of values identifies ten distinct types of values and describes the dynamic relations among them. This list of values contains some contrasting values (e.g., benevolence and power) whereas some are compatible with each other (e.g., conformity and security) (Schwartz 2012). The circular structure in Figure 1 represents pairs of major contrasting values namely Self-Transcendence and Self-Enhancement that include opposite values. Self-transcendence includes universalism and benevolence which positively correlate with empathy and ethical competence (Hofmann-Towfigh 2007). Self-enhancement, on the other hand, is found to be negatively correlated with moral judgment and discourse competence.

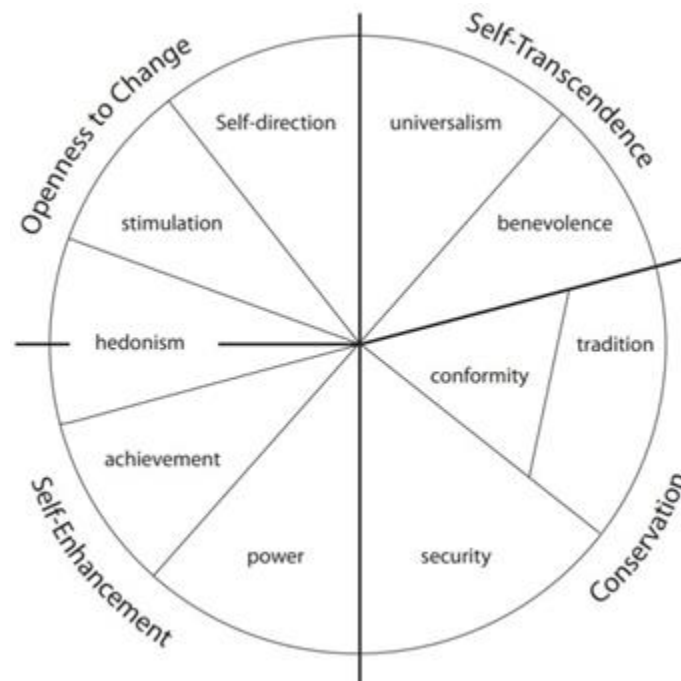


Figure 1: The circular structure of ten major values given in Schwartz's theory of values. It highlights two pairs of major contrasting groups of values including Self-Transcendence and Self-Enhancement that include opposite values.

Social factors in Behavior Change: The value systems are generally stable over time; however they may be prone to change depending upon one's circumstances (Rokeach 1973). According to (Padilla-Walker and Carlo 2014), goals to behave in a moral manner can be motivated by certain internalized (e.g., personal values) and external (e.g., punishment or reward) concerns. At middle and high school years in adolescence, the role of peers becomes critical in affecting prosocial behavior in individuals. It is highly likely that positive feedback from one's peers in the form of social approval and acceptance of their behavior may increase their empathetic, prosocial and moral behavior. In contrast, negative feedback for the behavior that is moral according to the individual, may result in reversing their original behavior. Social influence works at multiple levels, from gradual and permanent modifications in mood, language, gestures, etc., to more immediate influences on an individual's social attitudes and activities (Burnett et al. 2011).

Measures of Behavioral Change: In the past, several studies measured interaction of changes in adolescent behavior with age, group influence etc. A review suggests that behavioral economic paradigms which engage participants in structured competitive or cooperative interactions, reveal subtle differences in the degree of mental perspective-taking (Burnett et al. 2011). Such tasks are quantified as the amount of reward money/tokens exchanged with social partners (Berg, Dickhaut, and McCabe 1995; Binmore 2007). Neuroimaging studies show heightened activity in the reward system of the brain when such a game paradigm is introduced to adults. However, in similar circumstances, peer influence plays an important role in the case of adolescents. As previously mentioned, some studies suggest positive societal influence, on the other hand, some studies (Geier and Luna 2009) report peer influence on potentially harmful behaviors such as increase in risk taking behavior in simulated driving studies, especially in adolescents. The same can be understood from a neuropsychological perspective. The present study uses a novel behavioral game paradigm that is designed to study neural underpinnings of peer influence on adolescent brains. This is implemented by simulating conditions which aim to develop frustration towards peers. Neurophysiologically, frustration is correlated with dominance of lower brain frequencies (delta - 0.1-4Hz and theta - 4-8Hz). Hypothetically, in the present study, if an adolescent develops frustration, they tend to behave in a way that is morally unacceptable and in contrast, if they lack frustration biomarkers, it is assumed that they may behave in a morally correct manner.

In further sections, the methodology used in the study has been discussed where the game paradigm is elaborated. Followed by the methods, there is a results section that discusses the outcomes and finally the discussion section elaborates on inferences drawn on the basis of outcomes.

Methods

Participants: A cohort of 16 school girls aged 14-16 years ($M=14.6$ years, $SD=1.8$ years) participated in this study. All the participants were taken from the class 9 of the same school to control any differences due to education type, socio-economic status or other ecological factors. In the present study which is a part of a larger study with the present dataset, only 3 representative case studies have been discussed.

Game-based task: An interactive game environment was designed consisting of different levels of subitizing tasks (Clements 1999). The task requires the participant to intuitively identify the majority color of the objects given on the screen and press the corresponding button. The level of difficulty ranged from trial to trial (as shown in Figure 2). The task was presented in three blocks of 10 games each. The ten games consisted of 10 rounds each.

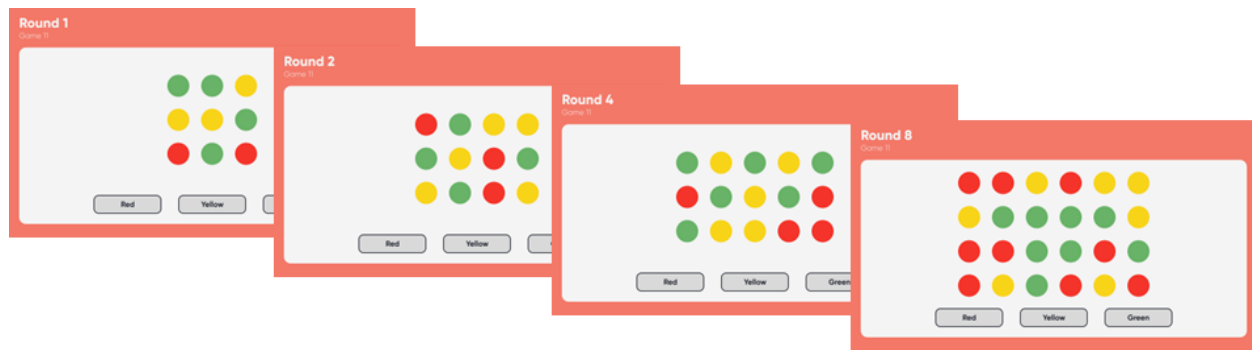


Figure 2: Level of difficulty in Subitizing task: The participant is required to intuitively identify the majority color of circles. The round 1 is the easiest level consisting of 3*3 object matrix, rounds 2,3 are medium level consist of 3*4 object matrix, rounds 4,5,6,7 are low difficulty level consist of 3*5 object matrix and rounds 8,9,10 are high difficulty level consist of 5*6 object matrix.

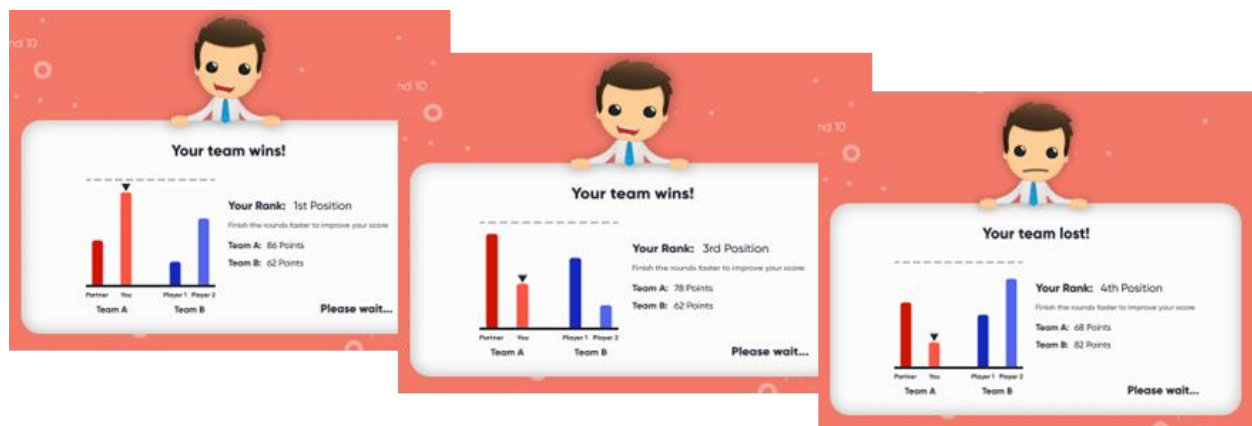


Figure 3: Result screens: The participant received different relative positions in the game.

At the beginning of each block, different virtual partners and opponent teams are introduced and staged as the participant is playing a multiplayer online game with them. However, the responses given by the partner are predefined and there is no virtual partner/team connected in real time. At the end of each game (after every 10 rounds), a result screen was shown to participants which represents the relative position of the participant as compared to their virtual partner and opponent team (Figure 3). The winning team receives 5 token rewards and a team member of the winning team randomly receives the reward to share it with the other member by pressing the corresponding button, e.g., press button 4 to share 4 reward tokens, and so on. The responses of the virtual partner are predetermined in an order-effect based manner in order to evoke emotional responses in participants' brains. For the same purpose, each block was designed differently, the first block was the conditioning block for the participants to instill moral judgment toward their partner by giving 80% reward to the team member who came in first position, the second block was the frustration block in which the virtual partner was depicted to be unethical in sharing rewards when the participant came in first position, and the final block was test of morality block which required the participant to share rewards when the partner came in different positions making the team win or lose correspondingly, to test moral behavior of the participant.

Data collection: *Psychological* - For psychological assessment of personal values of the participants, two scales were used, namely recently revised Portrait Values Questionnaire (PVQ-RR) (Schwartz 2021) and Interpersonal Reactivity Index (IRI) (Davis 1983). Other supporting data was also collected using a demographic scale consisting of questions about average grades in school, family education level, etc.; *Neurophysiological* - While playing the game, the participants were wearing a 64 channel EEG device that collects continuous neural signals. Some event-based markers were introduced in the EEG dataset at specific instances of reward sharing/receiving for further analysis. The sampling frequency of the EEG device is 256 Hz. Bandpass or IIR filters were kept between 0.1 to 45 Hz with a notch filter of 50 Hz. For ocular correction, channels Fp1 and Fp2 were used as EOG channels.

Results

The results of this study include a multivariate analysis of three cases that include participant/subject no. 6, 7 and 9. Psychological tests were also given to participants to report for personal values such as empathy, self-transcendence and self-enhancement. A scale called IRI was used to measure scores on components of empathy in participants namely cognitive empathy and affective empathy. The other scale used was PVQ-RR which measured two values in personality namely, self-transcendence which is composed of benevolence and universalism and in contrast, self-enhancement which is composed of hedonism, achievement and power (refer to figure 1). The scores of the three subjects for PVQ-RR psychological scale (Mean=15.33, S.D. =6.8) are as follows: Subject 6 scored 62 on ST and 39 on SE, Subject 7 scored 67 on ST and 54 on SE, Subject 9 scored 53 on ST and 43 on SE. Along with these scales, average grades in school were also measured which were found to be similar for all subjects (Mean= 84.10, SD= 0.8). For IRI, Subject 6 scored 29 on CE and 27 on AE, Subject 7 scored 37 on CE and 38 on AE, Subject 9 scored 36 on CE and 15 on AE. Figure 4 shows scores of the three subjects for PVQ-RR and IRI scales. Along with these scales, average grades in school were also measured for the three subjects which was found to be similar for all (Mean= 84.10, SD= 0.8).

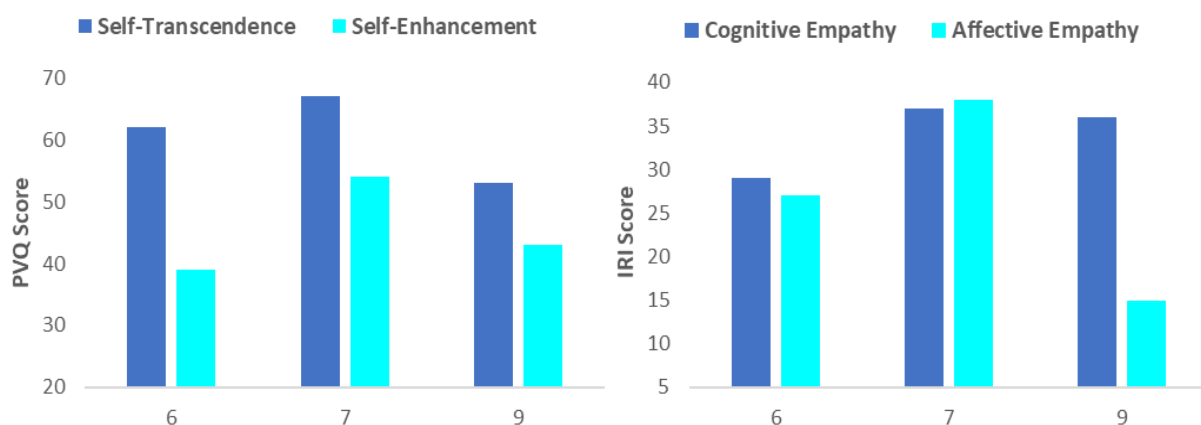


Figure 4. Scores for subject 6, 7 and 9 on PVQ-RR and IRI psychological scales.

The behavioral results are presented in Figure 4 which shows the average number of reward tokens shared by the participants with their apparent virtual partner in the two comparative blocks. The first block i.e., conditioning block

is compared with the third block i.e., test of morality block. The average for the two blocks only includes the rounds in which participants came in fourth position and the virtual partner was given first position.

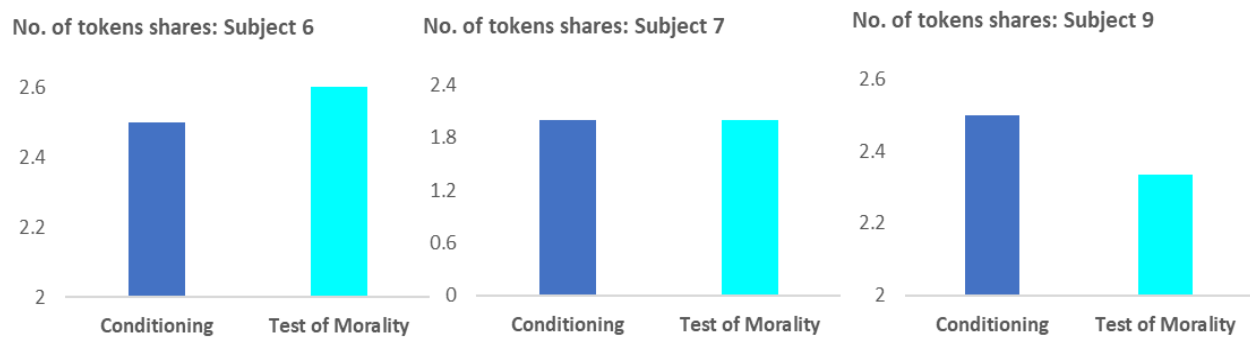


Figure 5: Behavioral results: For subject 6, 7 and 9, the average number of reward tokens (ranging from 1 -5), shared by the participant are plotted for the first (conditioning) block and third (test of morality) block.

Neurophysiological analysis was carried out in MATLAB-based software (Brainstorm) using spectral analysis techniques. The specific events of receiving or sharing reward tokens were extracted for an epoch of 1000 milliseconds to observe spectral changes across events. The events with significant differences across conditions and subjects both behaviorally and neurophysiologically were found to be the rounds in which participants came in fourth position and the virtual partner was given first position. Therefore, the EEG signals for the aforementioned events were analyzed. Relative spectral power for different frequency ranges was calculated by subtracting power for second and third blocks from that in the rounds of the first block. The relative power differences across conditions in all subjects were found to be significant for delta frequency range (0.1 to 4Hz) as shown in Figure 5.

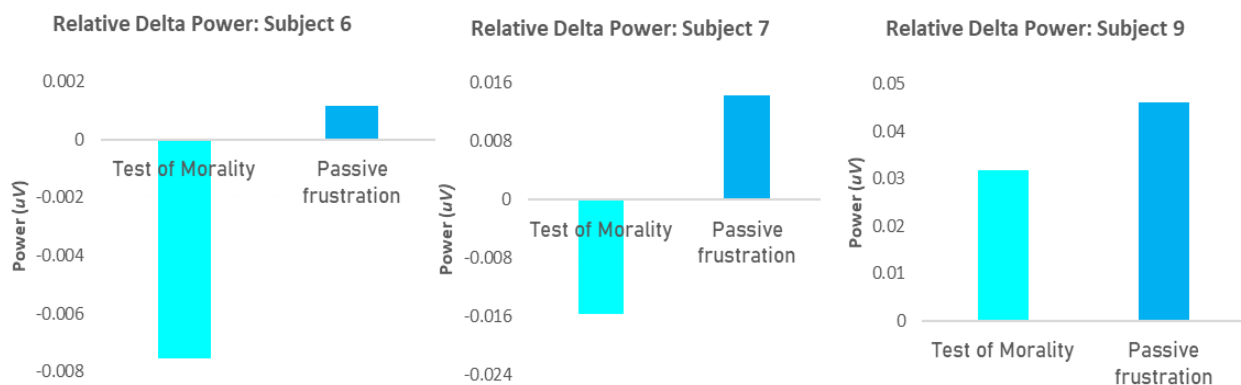


Figure 6: Neurophysiological results: Relative power for delta range (0.1 to 4 Hz) represented for second and third block in the subjects 6, 7 and 9.

Conclusion

Case study 1: Subject 6 scored higher on ST than SE with a difference of 23 which appears to be significant; scored equally on empathy representing positive personality traits. Relative delta power for the frustration block does not have significant value whereas for the third block, power has negative value. This depicts that the subject did not develop frustration towards the partner and showed gratitude, also evidenced in Figure 2 having more for reward sharing in the third block.

Case study 2: Subject 7 scored higher on ST than SE with a difference of 13 (mean=15.33) which is not significant; scored equally on empathy representing positive personality traits. Relative delta power for the frustration block has positive value whereas for the third block, power has negative value. This depicts that some level of frustration was developed toward the partner, however it was not carried over in the third block, also evidenced by neutral reward sharing.

Cast study 3 (Subject 9): Subject 9 scored higher on ST than SE with a difference of 10 (mean=15.33) which is not significant. They also scored differently on empathy components, having a higher score for cognitive (36) as compared to affective (15) empathy representing more self-enhancement traits in their personality. Relative delta power for passive frustration blocks has positive value for both blocks. This depicts some level of frustration developed toward the partner, and it was carried over to the next one, also shown by lesser reward sharing in the third block.

Conclusively, the results show that based on personality value types, adolescents tend to develop frustration toward perceived unethical behavior and carry it over to other unrelated peers. This study is highly explorative in nature, with preliminary analysis using only three case studies, having a small sample size. However, the novelty of this study brings about new dimensions to social cognition and personality studies.

References:

1. Berg, J., Dickhaut, J. & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1), 122–142.
2. Berry, J. W., Poortinga, Y. H., Pandey, J., Dasen, P. R. & Saraswathi, T. S. (1997). *Handbook of Cross-cultural Psychology: Theory and method*. John Berry.
3. Binmore, K. (2007). *Game Theory: A Very Short Introduction*. OUP Oxford.

4. Burnett, S., Sebastian, C., Cohen Kadosh, K. & Blakemore, S.-J. (2011). The social brain in adolescence: evidence from functional magnetic resonance imaging and behavioural studies. *Neuroscience and Biobehavioral Reviews*, 35(8), 1654–1664.
5. Clements, D. H. (1999). Subitizing: What Is It? Why Teach It? *Teaching Children Mathematics*, 5(7), 400–405.
6. Colby, A. & Damon, W. (1992). *Some Do Care: Contemporary Lives of Moral Commitment*. Free Press.
7. Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1), 113–126.
8. Hannah, S. T., Avolio, B. J. & May, D. R. (2011). Moral Maturation and Moral Conation: A Capacity Approach to Explaining Moral Thought and Action. *AMRO*, 36(4), 663–685.
9. Hardy, S. A. & Carlo, G. (2005). Identity as a Source of Moral Motivation. *Human Development*, 48(4), 232–256.
10. Moshman, D. (n.d.). *Adolescent rationality and development: Cognition, morality, and identity*. <https://doi.org/10.4324/9780203835111/adolescent-rationality-development-david-moshman>
11. Padilla-Walker, L. M. & Carlo, G. (2014). *Prosocial Development: A Multidimensional Approach*. Oxford University Press.
12. Pohling, R., Bzdok, D., Eigenstetter, M., Stumpf, S. & Strobel, A. (2016). What is Ethical Competence? The Role of Empathy, Personal Values, and the Five-Factor Model of Personality in Ethical Decision-Making. *Journal of Business Ethics: JBE*, 137(3), 449–474.
13. Schwartz, S. H. (1992). Universals in the Content and Structure of Values: Theoretical Advances and Empirical Tests in 20 Countries. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 25, pp. 1–65). Academic Press.
14. Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1116>
15. Schwartz, S. H. (2021). A Repository of Schwartz Value Scales with Instructions and an Introduction. *Online Readings in Psychology and Culture*, 2(2), 9.

Offload 802.11 scanning to low power device

Vishal Bhargava

Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
vishalbharg@gmail.com

N.S. Raghava

Department of Electronics & Communciation
Delhi Technological University
Delhi, India
nsraghava@dce.ac.in

Abstract—In this generation, Wi-Fi throughput is increasing day by day, with the speed of Wi-Fi data also increasing exponentially. From full HD, 2K to 4K, and 8K video watching standards are improving very fast for users. Users want to roam without any impact on user browsing. The challenge is to backward scan without interrupting the existing connection during roaming. This paper focuses on a low power external device specially designed for scanning purposes. An external Wi-Fi radio is introduced to make scanning faster and improve power saving during the connection process.

Keywords—802.11, Scanning, Connection, Authentication, Association, Wi-Fi

I. INTRODUCTION

Wi-Fi technology has seen tremendous growth in the last decade. Wi-Fi old generation (802.11a, 802.bg, 802.11n, 802.11ac) to Wi-Fi 6 (802.11ax). Now Wi-Fi 7 (802.11be) is also knocking on the door. Throughput & Security-wise, these 802.11 standards [1] are improving significantly. However, the Wi-Fi connection procedure is still the same as the first standard. Devices are increasing exponentially, so the environment noise is also increasing parallelly. Wi-Fi connection time still needs lots of improvement.

Wi-Fi connection is an IEEE 802.11 specification defined standard process (Figure 1), which is affected by some factors [2,3].

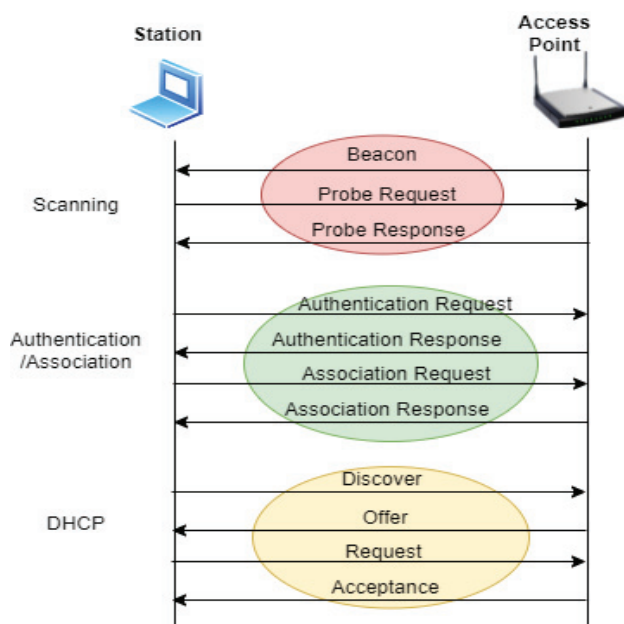


Fig. 1. 802.11 Connection Process

In the connection process first step is to identify and find the AP (access point) with whom STA (station device) wants to establish a Wi-Fi connection. This step is called scanning. The STA device scans networks on different channels to check the RF performance and AP (s) capabilities [4]. If STA finds its desired AP, it initiates a connection with a particular AP. To identify the correct access point, a station stays for some time on every scanned channel; that time is called dwell time [5].

Two types of scanning are performed via station [6] (Figure 2).

- Active Scan: Wi-Fi station device sends probe request and receives probe response in response.
- Passive Scan: Wi-Fi Station device waits for beacon from APs. Station receives a beacon during dwell time. On Dynamic Frequency Selection (DFS) channels, passive scanning is performed.



Fig. 2. Wi-Fi Scanning Methods

Before any connection scan is the crucial phase, the station always needs to be ready with the latest result. That's why STA performs background scan even current connection is working fine. The contribution of this paper is to propose an offloaded low power scanning method to create a fast connection.

The organization of this paper is as follows: Section II presents related work and research objectives. Our proposed system architecture and its conclusion are discussed in section III, and finally, conclusions and future work are drawn in section IV.

II. RELATED WORK & RESEARCH OBJECTIVES

Multi-dimensional ways used by researchers to solve Wi-Fi scan problems. In the IoT (Internet of Things) world, devices become very time savvy; even a minor save of time is a significant achievement. 802.11ba [7] has the same strategy to use low power devices, but it is only used for power saving purposes. Here only the Rx chain (WURx) is active, and it does not transmit anything.

Rishabh et al. [8] suggested a unique access point in a common area which can provide a scan list to the station. The station needs to send a vendor-specific probe request to AP, and in response, the station got the scan list. Station and access point have very different power capabilities, so the following approach is not suitable in a realistic environment.

Researchers used caching of previous connection information [9] where using guarded action system (GAS) information element scan and connection related information shared on the particular network element and all access points talk with that element before creating a new connection. If a piece of old connection information is found on the network element, the same is used via AP to create the latest connection.

The author in [10] proposes a fast authentication algorithm to make a quick connection, especially in roaming scenarios. The proposed method allows users to do advanced authentication before moving to other APs. Here the radius server plays a vital role in transferring information to other apps. In this research, somehow, the packets are travelling into the air, which can't give the same performance as the same device.

Cross-layer approach is also a working area where one layer of work can offload to another network layer. DHCP offload to the lower layer is a way to improve connection time from seconds to milliseconds. In [11], Pre-allocation of DHCP is performed to enhance connection time. With probe request, DHCP allocation is done via AP with the DHCP server. But it's a waste of resources as 99% APs are just used for scan purposes via stations. To overcome the above problem, In the paper [12], DHCP packets merge with Authentication and association packets, and data link layer packets offload to the MAC layer.

Sometimes, finding an AP and maintaining an existing connection seems a very tough task, especially in a dense environment [13,14] like shopping malls, railway stations, or a university where so many users work on the same frequency at the same time. Fast scanning [15] and selective active scanning [16] are discussed to improve Wi-Fi scanning. Unfortunately provided solution does not fit under the real environment, and practicality is not addressed in the paper. We will overcome this issue in this paper.

III. SYSTEM ARCHITECTURE

The proposed methodology towards this direction is the work on a new amendment to the Wi-Fi scanning process, which introduces a scan-oriented Radio. This radio is an additional interface with extremely low power consumption that is used to perform active and passive on a requirement basis. In contrast, the device's primary radio is to perform another task or switch off. This paper describes the IEEE 802.11 scanning mechanism and protocol via secondary radio, discusses its work model, investigates software and

hardware dependencies, evaluates different test cases and how much throughput improves via the proposed method.

The proposed System architecture is shown in figure 3. Scan devices have an additional Wi-Fi radio with an extremely low power consumption processor. Main Device has a high clock application processor which performs the main task, and scan request comes from the operating system or application to the device. Instead of being taken care of via the main device, it offloads to the secondary device.

Scan device has small memory, which saves scan results. Until that time, the main device is in a switched-off state. Results are shared in common memory, which can be accessed via the main device.

The device is operating on channel 36 (5180 MHz). So the secondary device will scan all channels from the channel list except the operating channel of the main device. In this case, channel 36 will not scan via the secondary device. As the main device already operates on channel 36, it already receives the beacons from other access points (running on the same channel), and it can fill the scan list for channel 36. This way, both devices will not interfere with each other.

In case any disconnection happens, the main device can directly read the scan list, and it can connect to the desired AP. Scan device can serve the purpose of background scanning. When scanning is not required, the scan device can be completely switched off.

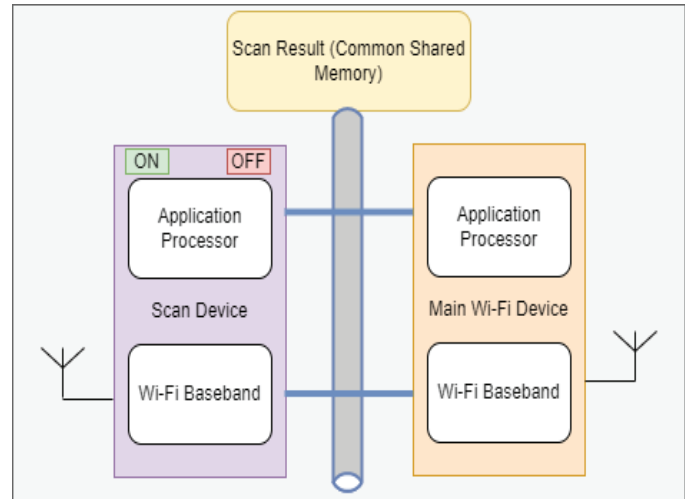


Fig. 3. Proposed system architecture

IV. SIMULATION

Existing simulator MATLAB, NS (network simulator) doesn't provide this kind of facility, and no hardware is available to demonstrate this. Own windows device driver-based simulator developed to perform the test. NDIS Miniport [17] driver handles IOCTL (input/output control) from the application layer. In the driver, two different elements are created. One works as a Main Wi-Fi device, and it talks to the application layer. Other elements work as scan devices, and the Main Wi-Fi device only interacts with this element (figure 4).

An application creation which performs an initial and background scan initiates a Wi-Fi connection and transfers some random data.

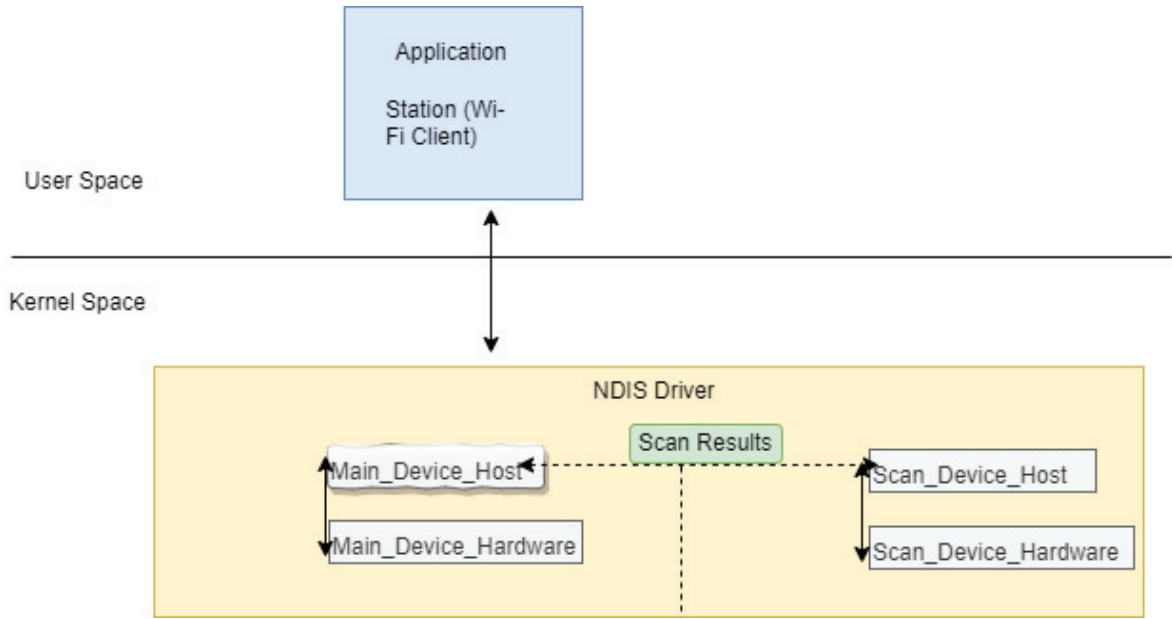


Fig. 4. Simulator Device Model

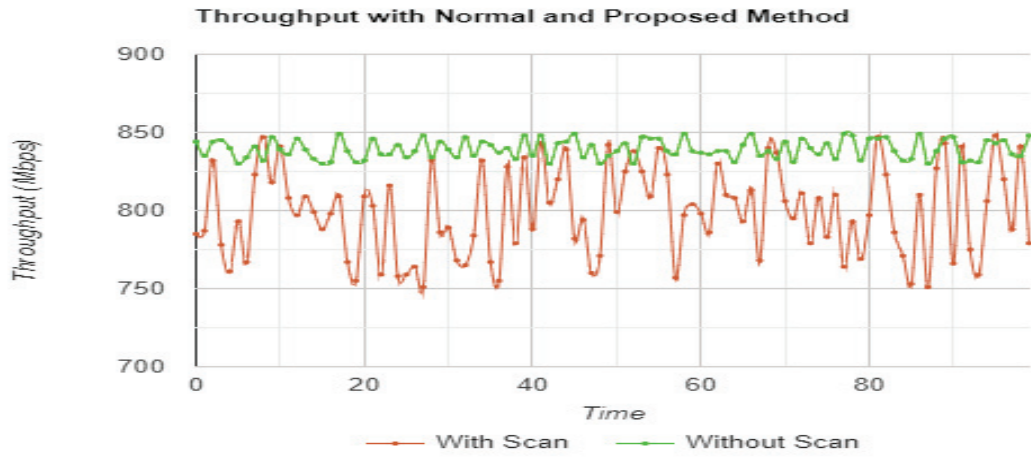


Fig. 5. Scan impact on throughput

Application sent “OID_802_11_BSSID_LIST_SCAN” for scan purposes and other OIDs for a different purposes. The flow chart is shown in figure 6. To perform simulation, application when the main device got the scan OID (Object identification), it first sends radio on request to scan device, and after that scan, request sent to scan device. The scan device performs the scanning, and the scan result sends to the main device, which communicates to the application and the main device sends a radio off command to the scan device. The radio on/Off command simulates the power on/off of the secondary device.

Here identify some user scenarios problems where the proposed method can deliver the best performance and optimize the network behaviour:

1. High Throughput Run: Some high throughput is running on the device, for example, a user is watching a high definition of video. If any scan request comes to the Wi-Fi device, it can degrade the user experience. Proposed behaviour saves from the above degradation.

We took Dlink DWA-X1850 [18] windows dongle and ran throughput with Netgear 802.11ax AP (Configuration shown in table 1) and a scan given every 100 ms via script.

TABLE I. AP CONFIGURATION

Frequency	Mode	Channel	SSID	Bandwidth
2.4 GHz	11n	6	TestNgr	20 MHz
5.0 GHz	11ac	36	TestNgr	80 MHz

It clearly shows how background scans impact throughput, and without scan (offload to another device), throughput can be much better, as shown in figure 5.

2. Roaming improvement: The operating system (OS) periodically monitors the RSSI (Received Signal Strength Indicator). If RSSI goes below the defined threshold, the OS asks the Wi-Fi device to roam on another AP. It generally happens in shopping malls and railway stations where many Access points are installed with the same SSID (Service Set Identifier)

and security. So, if RSSI is going down, the application can send a scan request, which can go to the scan device. It can help in roaming performance improvement.

3. Power-save: The main device took more power than compared to scan device. In the case of 1st device, after offload scan to a secondary device, 1st device can go on low power to save power.
4. Auto-Channel Selection: The proposed way is beneficial for access point auto channel without impacting throughput on the network. Scan device identifies an operating channel that minimizes interference from other devices without interrupting other tasks.

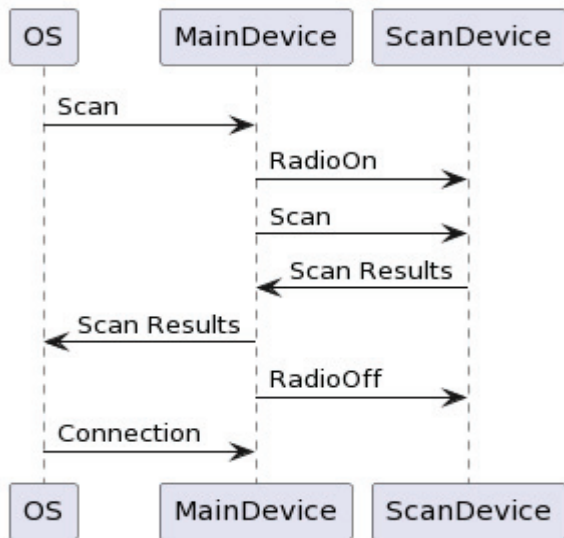


Fig. 5. Scan offload flow chart

V. CONCLUSION & FUTURE WORK

In this paper, we have discussed how scan offload to the secondary device can help to improve 802.11 scanning performance by considering some unique factors that remain unaddressed by the existing scan mechanism. The paper specifically identified test cases in which scan offload can help to improve user behaviour, reduce power consumption, and maintain the connection in a roaming environment. Mechanism and advantages discussed.

Although the suggested approach improves user experience and reduces connection time during roaming, it still has some points which must be taken care of in the future.

- The different device increases the cost, so cost reduction should be taken care of. Hardware can be cheaper with time.
- The machine learning approach can decide when to offload the scan to the different device.
- Both devices should have the same interpretation for access point RSSI.
- Scan devices can be used for other purposes.

REFERENCES

- [1] IEEE Std 802.11ba™ -20 2 1, IEEE Standard for Information Technology, Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 3: Wake -Up Radio Operation , March 2021.
- [2] Seneviratne, S., Seneviratne, A., Mohapatra, P. and Tournoux, P., "Characterizing WiFi connection and its impact on mobile users," Proceedings of the 8th ACM international workshop on Wireless network testbeds, experimental evaluation & characterization - WiNTECH '13, pp. 81-88, 2013.
- [3] C. Pei, Z. Wang, Y. Zhao, Z. Wang, Y. Meng, D. Pei, Y. Peng, W. Tang, X. Qu, "Why it Takes so Long to Connect to a Wi-Fi Access Point?," IEEE INFOCOM 2017 - IEEE Conference on Computer Communications, 2017, pp. 1-9, doi: 10.1109/INFOCOM.2017.8057164.
- [4] Singh, Dhananjay, "Channel Scanning and Access Point Selection Mechanisms for 802.11 Handoff: A Survey" (2020). Computer Science and Engineering Master's Theses. 15. Online available at https://scholarcommons.scu.edu/cgi/viewcontent.cgi?article=1014&context=cseng_mstr.
- [5] "IEEE Standard for Information Technology - Telecommunications and information exchange between systems - Local and Metropolitan Area Networks - Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications - Physical Layer Parameters and Specifications for 1000 Mb/s Operation over 4 pair of Category 5 Balanced Copper Cabling, Type 1000BASE-T," in IEEE Std 802.3ab-1999 , vol., no., pp.1-144, 26 July 1999, DOI: 10.1109/IEEESTD.1999.90568.
- [6] Majumder, Abhishek, Samir Nath, and Sudipta Roy. "A Scanning Technique Based on Selective Neighbour Channels in 802.11 Wi-Fi Networks." International Conference on Machine Intelligence and Signal Processing pp.83-96, Springer, Singapore, 2019.
- [7] Hyun-hee Park and Eui-Jik Kim, "Wake-up Radio-resilient Scanning Mechanism for Mobile Device in IEEE 802.11ba" Sensors and Materials, Vol. 30, No. 12, 2018, pp. 2961–2968
- [8] R. Gupta and V. Singh, "Reduce 802.11 Scanning Time Using Special Device to Provide Scan Results," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 1276-1278, DOI: 10.1109/ICRITO48877.2020.9197842.
- [9] Lei Wang, "Method and apparatus for accelerated link setup", US Patent 738,589, Dec 01 2015.
- [10] R. Syahputri and S. Sriyanto, "Fast and secure authentication in IEEE 802.11i wireless LAN," in Uncertainty Reasoning and Knowledge Engineering (URKE), 2012 2nd International Conference on, aug. 2012, pp. 158–161
- [11] A. Zúquete and C. Frade, "Pre-Allocation of DHCP Leases: A Cross-Layer Approach," 2011 4th IFIP International Conference on New Technologies, Mobility and Security, Paris, 2011, pp. 1-5, DOI: 10.1109/NTMS.2011.5720663.
- [12] Bhargava V., Raghava NS. (2021) Reduce 802.11 Connection Time Using Offloading and Merging of DHCP Layer to MAC Layer. In: Paiva S., Lopes S.I., Zitouni R., Gupta N., Lopes S.F., Yonezawa T. (eds) Science and Technologies for Smart Cities. SmartCity360° 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 372. Springer, Cham. https://doi.org/10.1007/978-3-030-76063-2_13
- [13] M. M. Surur and N. Surantha, "Performance Evaluation of Dense Wi-Fi Network Based on Capacity Requirement," 2019 International Conference on Information Management and Technology (ICIMTech), 2019, pp. 466-471, doi: 10.1109/ICIMTech.2019.8843775.
- [14] K. Sui et al., "Understanding the Impact of AP Density on WiFi Performance Through Real-World Deployment," 2016 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN), 2016, pp. 1-6, doi: 10.1109/LANMAN.2016.7548845.
- [15] S. Jin, M. Choi, L. Wang, and S. Choi, "Fast scanning schemes for IEEE 802.11 wlans in virtual ap environments," Computer Networks, vol. 55, pp. 2520–2533, 07 2011.
- [16] S. Waharte, K. Ritzenthaler, and R. Boutaba, "Selective active scanning for fast handoff in WLAN using sensor networks," in Mobile and Wireless Communication Networks (E. M. Belding-Royer, K. Al Agha, and G. Pujolle, eds.), pp. 59–70, 2005.

- [17] NDIS Miniport drivers available at <https://docs.microsoft.com/en-us/windows-hardware/drivers/network/ndis-miniport-drivers2>.
- [18] AX1800 Wi-Fi 6 USB adapter, online available at <https://www.dlink.com/en/products/dwa-x1850-ax1800-wi-fi-6-usb-adapter>.

Open Source Software Based Electronic Health Record Management System

Javtreshwar Singh GILL¹, Himanshu MITTAL², Kunal BANSAL³, Varsha SISAUDIA⁴

kunal.bansal35@gmail.com

^{1,2,3,4}Department of Information Technology, Delhi Technological University,
varsha@dtu.ac.in

Abstract. The countries of the Third World are beginning to use Electronic Health Records, however, the lack of a standard or open-source records in the health industry has resulted in a new set of problems. Countries like India lack a unified and standardized model for the same. Moreover, these electronic records are limited to modern and mostly private hospitals and clinics making it difficult for the people of rural areas to get access to this better standard of medical services. This paper explores the concept of an open-source Electronic Health Record system and discusses how it can change the Indian Medical industry.

Keywords. EHR, India, DigiMed, OCR, Open-Source

1. Introduction

Medical record keeping is undergoing a transition in the developing countries. All the Third World countries are currently fighting a war internally, against the rising medical issues due to increasing waste, global warming and other issues. The problem is that the major population effected with diseases are in the Third World countries. This would not be a problem if these nations had good medical practices and services. But most of these countries are overpopulated and the use of a pen and paper approach has made life very difficult for the residents.

1.1. Electronic Health Record (EHR)

Paper based shortcomings can be overcome by using a modern and digital approach, Electronic Health Records (EHR). The major problem of record and data transfer is easily overcome by EHRs. They can be easily transferred at very high speeds. It is only limited by the data upload rate at the sender's location and the data download rate at the receiver's location. This not only leads to a saving in time but also personnel, which is needed for physical transfer of reports and forms.

2. Literature Review

2.1. History

The work on Electronic Medical/Health Records goes a long time back to the beginning of internet in the mid to late 90s. The early papers discussed about the shortcomings of the then current EMRs which were very segregated and almost all were proprietary [1]. It became increasingly difficult for the parties involved as the number of EMRs grew. At one time, the United States had nearly 230 different vendors for EMRs used by hospitals around the country. [2].

2.2. Existing Electronic Medical Records (EMRs) in Developing Countries

It has been very difficult for the Third World Countries to develop Electronic Medical Records and even more difficult to use them in production. We describe a few of them here.

1. AMRS, Kenya: Mosoriot Medical Record System (MMRS) [4].
2. PIH-EMR, Peru: "Partners in Health" [5].

2.3. EHR Scenario in India

Electronic Health Records have begun to crop up around the country but are still limited to the private sector. They include Max Health [6], Apollo [7], Sankara Nethralaya [8], Fortis, etc. In the public health care domain, All India Institute of Medical Sciences (AIIMS) and the Postgraduate Institute of Medical Education and Research (PGIMER) are the few which make use of Electronic Health Records.

The main benefit is disease surveillance which helps protect mass outbreaks in highly populated countries like India. However, there were concerns about confidentiality and security of data [9] [10].

3. Architecture

The choice and implementation of the architecture to be used in an Electronic Medical/Health Record is very important and needs to be carefully analysed and tested before making a final decision. These parts are discussed in this section.

3.1. Data Model

The first step in building an Electronic Health Record System is the choice of the kind of database or data model that would be used because the most important part of any EHR is its capacity to hold as well as release data as per the requirement.

For our application, we have chosen a SQL database, specifically, PostgreSQL [14]. It is an open-source SQL database.

3.2. Network Architecture

The various kinds of network architectures that can be used are mentioned below:

1. Stand-alone Systems: A stand-alone system uses a single machine to handle both the backend database and the frontend interface.
2. Local Area Network (LAN) Systems: LAN systems are operated in small geographical area where a decent internet connection is present.
3. Wide Area Network (WAN) Systems: WAN system works like a LAN system but at a much larger geographical stage. A WAN network can work over huge areas, even across states and countries.

4. Advantages of Electronic Health Records (EHRs)

The previous section detailed about the various constituents that need to come together to create an Electronic Medical Record. This chapter will detail the advantages that an Electronic Medical Record holds over the traditional paper-based approach. These advantages are clear and help the argument, the need for adopting Electronic Medical Records all over the world, especially in developing countries.

4.1. Electronic Health Record Functionality

Electronic Medical Records are not just there to record a patient's health, but also carry in them various data sets required for clinical research called Common Clinical data sets (CCDS). CCDS is used to store clinical information at the level of the patients. This includes the statistics, diagnoses, management, lab results, drug history, vital signs, vaccination reports, radiology results as well as their allergies [13] [14].

4.2. EHR Data: Primary Usage

The data used by doctors and physicians is directly from the Electronic Health Records (EHR) as they need the correct entered information. They need to interact with the database using the basic CRUD operations. These operations refer to creating, reading, updating, and deleting of records in the database. To improve the experience of the doctors, we can enhance the database's capability to perform CRUD operations quickly [15].

4.3. EMR Data: Secondary Usage

The data obtained from EHRs is primarily used for curing the diseases of the patients. However, another use for the data is to altogether prevent the cause of diseases as well as finding better cures for them which are cheaper and much more widely available. Simply said, the data is used to perform research on various diseases to learn more about their occurrence, causes and use this data to help procure better treatments [16].

5. Modern Techniques to Enhance EHR

The recent rise in technologies like Machine Learning, Optical Character Recognition, Artificial Intelligence, etc, has resulted in a plethora of new opportunities for the enhancement and upgradation of the classical Electronic Health Records. These

techniques and technologies have made the trivial tasks of data entry much simpler. In this section, we explain what these technologies are and how they can result in far superior Electronic Health Records.

5.1. Optical Character Recognition (OCR)

Optical Character Recognition solves the problem of identifying characters which are fed into the system optically i.e., an image is provided. It works off-line i.e.; it works after the complete text has been input whereas on-line recognition works as the characters are drawn in real time. Hand written as well as printed characters can be read, but the efficiency is based on the quality of the input documents as well as the training model that has been used [17].

The major application of Optical Character Recognition is for the conversion of old handwritten records into their digital counterparts. This is one of the reasons that OCR has become mainstream in recent times.

5.2. Machine Learning (ML)

Machine learning is a subject that deals with the concept of using computers to simulate learning activities as done by humans. It also involves the study of self-improvement methods used by computers to obtain new knowledge and skills based on current knowledge.

Compared to humans, computers can learn using much more data and can also do it in comparatively shorter duration of time. As a result, all the advancement in the Machine Learning domain will have a direct effect on the human society [18].

6. Results

The previous section provides some guidelines on how to build an EMR/EHR. In this section, we will showcase an application built using those guidelines and how it has panned out. The application, DigiMed, has been built using open-source tools and frameworks that are widely and easily available to realize the main objective of the paper.

6.1. Technical Stack Used

We have tried to use the best open-source tools available to build this Electronic Health Records System. The reasoning behind it is simple, we need to use an open-source product to increase the inclusiveness of the EHR system so that it can be widely applied without the problems of inter-EHR compatibility and transfer of data. This technical stack helps to build a system which will be compatible and will be easy to convert because of its open-source nature.

6.2. Database

The database that is used for the application is a dynamic database. The database is changing as per the needs and requirements. The database being used is PostgreSQL, which is an open-source SQL database.

6.3. Authentication

The authentication screen in the application provides a user the option to either register or login if already registered.

6.4. Patient Home Screen

The patient home screen is available after login and provides the patient the multiple options. The patient can book an appointment from the list of doctors that are available, specifying the date and time for the same. After booking an appointment, the patient can check the status of the appointment, whether the doctor has accepted or rejected the appointment, from the 'My Appointments' screen. All the old appointments are present in the archive of appointments.

6.5. Doctor Home Screen

The doctor's home screen is accessible post successful login and provides the options mentioned ahead. The doctor can upload a form and get the written text via the use of Optical Character Recognition. The other screen is the 'My Appointment' screen where a doctor can either accept an appointment or reject. The changes are then reflected on both the doctor and the patient screens. Like patient, the doctor also has an archive of appointments.

7. Conclusion

The growth and development in the Information Technology industry has made it simpler to maintain and manage the data of patients in a digital format across the various levels of healthcare system. The advancement along with the support of Indian Government and launch of Digital India movement has helped in the progression of healthcare systems. This paper has tried to highlight the need of an open standard for Electronic Medical/Health Records and also provided a starting step for the same. All the technical stack used in the development of DigiMed is open source and freely available. The objective of this paper was to provide a starting step for building an inclusive and open-sourced EHR system and the same has been realized using the creation of DigiMed. The application provided in the results section is built for mobile phones but can easily be ported over to be used on the web.

7.1. Future Works

This paper deals with the very basics of building an Electronic Health Record and should be used as a first reference point. Expanding on the work, the additions that can be made are:

Data Protection: Although the basic data protection has been applied, encryption algorithms can be used to protect a patient's personal information in case of a database leak or breach. This protection can be used to help separate a patient's reports from their identity so that even if a report is leaked, it cannot be linked to the patient.

References

- [1] Kohane IS, Greenspun P, Fackler J, Cimino C, Szolovits P. Building national electronic medical record systems via the World Wide Web. *Journal of the American Medical Informatics Association*. 1996 May 1;3(3):191-207.
- [2] Berners-Lee T, Cailliau R, Luotonen A, Nielsen HF, Secret A. The world-wide web. *Communications of the ACM*. 1994 Aug 1;37(8):76-82.
- [3] Voelker R. Conquering HIV and stigma in Kenya. *Jama*. 2004 Jul 14;292(2):157-9.
- [4] Anokwa Y. Delivering Better HIV Care in Sub-Saharan Africa Using Phone-Based Clinical Summaries and Reminders.
- [5] Fraser HS, Jazayeri D, Mitnick CD, Mukherjee JS, Bayona J. Informatics tools to monitor progress and outcomes of patients with drug resistant tuberculosis in Peru. In *Proceedings of the AMIA Symposium 2002* (p. 270). American Medical Informatics Association.
- [6] Max healthcare. <https://www.maxhealthcare.in/hospitals-in-india>
- [7] Apollo Hospital. <https://www.apollohospitals.com/patient-care/clinical-quality-and-outcomes/it-excellence/>
- [8] Sankara Nethralaya. <https://www.sankaranethralaya.org/patient-care-eye-q.html>
- [9] Radhakrishna K, Goud BR, Kasthuri A, Waghmare A, Raj T. Electronic health records and information portability: a pilot study in a rural primary healthcare center in India. *Perspectives in health information Management*. 2014;11(Summer).
- [10] Dornan L, Pinyopornpanish K, Jiraporncharoen W, Hashmi A, Dejkriengkraikul N, Angkurawaranon C. Utilisation of electronic health records for public health in Asia: a review of success factors and potential challenges. *BioMed research international*. 2019 Jul 8;2019.
- [11] Kavitha R, Kannan E, Kotteswaran S. Implementation of cloud based Electronic Health Record (EHR) for Indian healthcare needs. *Indian Journal of Science and Technology*. 2016 Jan;9(3):1-5.
- [12] Chun JR, Hong HG. Factors affecting on personal health record. *Indian Journal of Science and Technology*. 2015 Apr;8(S8):173-9.
- [13] Pai MM, Ganiga R, Pai RM, Sinha RK. Standard electronic health record (EHR) framework for Indian healthcare system. *Health Services and Outcomes Research Methodology*. 2021 Sep;21(3):339-62.
- [14] PostgreSQL. <https://www.postgresql.org/>
- [15] Bates DW, Cohen M, Leape LL, Overhage JM, Shabot MM, Sheridan T. Reducing the frequency of errors in medicine using information technology. *Journal of the American Medical Informatics Association*. 2001 Jul 1;8(4):299-308.
- [16] Fraser H, Biondich P, Moodley D, Choi S, Mamlin B, Szolovits P. Implementing electronic medical record systems in developing countries. *Journal of Innovation in Health Informatics*. 2005;13(2):83-95.
- [17] Galvão J. Access to antiretroviral drugs in Brazil. *The Lancet*. 2002 Dec 7;360(9348):1862-5.
- [18] Kuo CC, Ting P, Teng WG, Chen PM, Chen MS, Chen JC. Multimedia over IP for thin clients: building a collaborative resource-sharing prototype. *Concurrent Engineering*. 2004 Sep;12(3):175-83.



Performance of adaptive radial basis functional neural network for inverter control

Alka Singh¹ · Amarendra Pandey¹

Received: 7 April 2022 / Accepted: 7 December 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022

Abstract

A single phase grid connected photovoltaic (PV) system is susceptible to a number of power quality (PQ) problems, including power factor, harmonic current, voltage fluctuations, and load unbalance. Compensation is needed to address these PQ concerns. In this study, a single phase shunt active power filter is presented to handle power quality issues using novel and straight forward radial basis function neural network (RBFNN) controller architecture and to ensure maximum power flow between PV and grid using a maximum power point tracker control technique. The design takes into account a single neuron in the hidden layer, and the network is trained on-line to be suitable for inverter control to reduce power quality (PQ) issues. The newly developed controller has a single input for the load current and is able to isolate the fundamental component of the current. Tracking is fast and achieved within one cycle. The trained model shows exceptional results for load compensation under various loading conditions. With the suggested RBFNN controller, both findings from simulation and from experiments have been shown to work.

Keywords Adaptive control · Power quality · Compensator · Harmonics · Neural network · Renewable energy

1 Introduction

The power demand is increasing day-by-day so maintaining a balance between the energy consumption and energy productivity is imperative. Integrating renewable energy source (RES) sources into utility grid is a growing concern globally [1]. The adoption of RES not only meets the energy shortfall but also reduces environmental pollution to a great extent. However, the uncertainty and intermittency of renewable energy sources need to be looked into carefully. The stability and safety of power grid especially in case of faults need detailed studies with RES energy sources.

In renewable energy generation system, photovoltaic (PV) based power generation has been thoroughly explored and widely utilized. The PV technology has seen exponential growth from 1992 till date [1, 2]. The PV is worldwide recognized as one of the most promising RES technology and large scale implementation projects supported by various countries demonstrate this trend [2]. Medium to small scale consumers

have received several incentives and government grants for adopting PV technology. Table 1 shows the recent and estimated capacity of PV capacity (in GW) worldwide along with the cumulative growth in percentage points [3].

Due to such large scale PV integration, the power systems are growing and getting complex. Moreover, the use of other renewable energy sources such as wind and increase in power electronic-based loads has affected the nature and performance of power systems [1]. PV panels show poor conversion efficiency while operating normally, and their output is non-linear and dependent on temperature and weather conditions. Thus, the maximum power point tracking (MPPT) approach is utilized to harvest maximum power with maximum efficiency from PV panels [4]. Several MPPT approaches have been proposed in the literature. Perturb and Observe (P and O) MPPT method is applied in this research work because of its simplicity, strong tracking ability, accuracy and ease of implementation.

Moreover, the grid operator faces greater challenges (like power quality (PQ), efficiency and reliability) to maintain grid stability and reliability because of high penetration of PV. Due to this various grid codes have been proposed to regulate seamless integration of PV system with distributed grid [5]. It is a major challenge to supply reliable and good

✉ Amarendra Pandey
amarendra8109@gmail.com

¹ Department of Electrical Engineering, Delhi Technological University, Delhi, India

Table 1 Recent and estimated capacity (GWp)

Year-end	2016	2017	2018	2019	2020	2021
Cumulative (GW)	306.5	403.3	512	633	– 770	– 950
Annual New(GW)	76.8	99	109	121	121–154	160–200
Cumulative growth (%)	32%	32%	27%	24%	24%	27%

quality power round the clock. Voltage fluctuations, voltage imbalance, voltage sag, harmonics and transients are all well-known PQ issues that impair power system performance even in the wake of contemporary technological advancements. Harmonic distortion is one of the most significant concerns among these PQ issues [6, 7]. Studies indicate that the degree of harmonic distortion in a system rises as the system's technology advances and consumer loads become predominantly non-linear. Such PQ problems can cause problems on the consumer side at the point of common coupling (PCC) for a number of reasons; sometimes sensitive instruments in hospitals and medical equipment may malfunction because of PQ issues [8]. The severity of PQ issues and resulting losses has motivated engineers to search for new, improved and effective solutions for conventional PQ problems such as reactive power compensation, load unbalancing, poor power-factor and voltage regulation [9, 10]. Cost effective solutions include passive filters; however fast and accurate control within 1–2 cycles and transition from lagging to leading vars is possible only using shunt compensators. Shunt compensators can enhance the system's power quality and minimize issues including harmonics, reactive power burden, low power factor, and load unbalancing [10].

A single phase insulated gate bipolar transistor (IGBT) based shunt compensator is an effective and reliable means of providing PQ improvement. Shunt compensators can be effectively controlled to achieve load compensation [6, 7]. Many control strategies have been established for the effective operation of the shunt active power filter (SAPF) [11–20] and some of them include Synchronous Reference Frame (SRF) theory discussed in [11] and Instantaneous Reactive Power theory (IRPT) proposed by Akagi [12–14], Symmetrical Component based Instantaneous Power Theory and Composite observer based theory are discussed in [15]. The need for transformation increases the system complexity of SRF and IRPT. Fourier transform based methods such as Fast Fourier Transform and discrete Fourier transform are discussed in [16]. Various control algorithms including artificial neural networks (ANN) based have been designed for control of shunt compensator. The use of conventional back propagation technique is discussed in [17]. The use of functional link artificial neural networks, Legendre polynomials, Chebyshev functional ANN, Spline polynomials have been discussed for PQ mitigation in recent papers [8, 18–20].

ANNs belong to the family of artificial adaptive systems that are inspired by the way the human brain works. Inspired by this, ANNs are trained to construct a relationship between input variables and output variables using observational data. Nodes are the basic building blocks of an ANN; they are also known as processing elements (PE) and the connections. Every node has a function that converts its own global input into output shown in Fig. 2a [21]. RBFNN is a potential ANN technique with several positive attributes such as easy design, good generalization, strong tolerance to input noise, online learning ability, very good tracking speed, fast convergence and stability. The RBFNN essentially comprises input, output layer and a single hidden layer. At each node of the hidden layer, a set of radial basis functions may be present. From the hidden to the output layer, adjustable weights are used to obtain the desired output. The RBFNN finds a numerous applications in the field of approximation, clustering, forecasting, estimation, direct and inverse system modeling and adaptive control [22]. Some of the applications of RBFNN controller are in the control of unified power flow controller [23, 24], permanent magnet synchronous machine [25], thyristor controlled series compensator [26] and microgrids [27]. Moreover, RBFNN has been designed for non-electrical aspects such as rainfall forecasting [28] and unmanned aircraft [29]. However, the RBFNN structure in all these papers involves a large number of neurons in the hidden layer and elaborate training.

It is proposed to design RBFNN as a non-linear adaptive filter for mitigation of PQ problems in this paper. The proposed RBFNN structure is self-adaptive and unique. It is designed to control the SAPF and improve power quality. The SAPF may be further integrated to PV panels in one of two ways: single stage or double stage. A PV array coupled to a single phase power distribution system has been described and designed in this work. The PV source is connected to the system through a single phase H-bridge converter and is designed to function as an active shunt filter. Thus, studies on SAPF connected to utility grid and RES integrated SAPF are discussed in detail. The objectives of this research paper include:

- (1) Designing a self-adaptive RBFNN having a simplified structure.
- (2) Design hidden layer having only one neuron and train RBFNN using Least Means Square (LMS) technique.

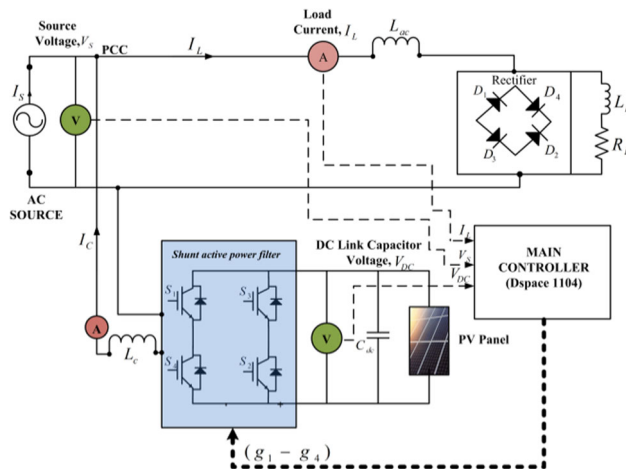


Fig. 1 System configuration

- (3) Study system performance under several test cases viz. non-linear load with high harmonic content.
- (4) Establish Lyapunov Stability criteria for the proposed technique
- (5) Evaluate the system performance when 2.7 kW PV array is integrated to the investigated SAPF system

2 System configurations

Figure 1 shows the schematic diagram of a single phase H bridge shunt compensator connected to a system feeding non-linear loads. The shunt compensator comprises four switches (S_1, S_2, S_3, S_4) which are effectively controlled using the RBFNN controller. The control scheme is designed to correctly track the fundamental component of load current under all load variations. It takes load current as inputs and generates four gating pulses (G_1, G_2, G_3 , and G_4) appropriately. The RBFNN technique is effectively designed to work as an adaptive filter and provide PQ improvement. Further, the shunt compensator controlled as SAPF is also integrated to PV arrays. Simulation and hardware results of the developed system are illustrated in the paper.

Figure 2a shows the configuration of generalized RBFNN model with multiple inputs, multiple neurons in the hidden layer and a single output. This is the conventional method involving multiple weights of associated neurons and is a complex approach. The proposed adaptive RBFNN controller in Fig. 2b shows the proposed configuration with one input (load current), single neuron in hidden layer and output (y_1) corresponding to the extracted weight. This network is trained online using the Least means Square (LMS) technique. Only two parameters viz. weight (w_{11}) and centre (c_{11})

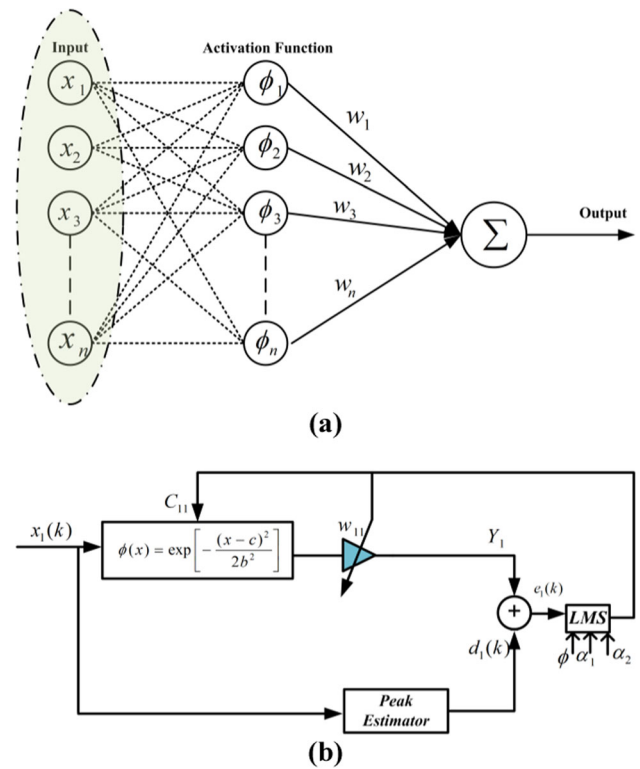


Fig. 2 Block diagram of **a** General RBFNN Model **b** Proposed RBFNN single neuron-based model

are required to be trained to obtain the fundamental component of load current (Y_1). The trained network can then be utilized for PQ improvement. The details of the proposed controller including the peak estimated current (d_1), error (e_1) and constants (α_1, α_2) are discussed in the next Section.

3 Mathematical formulation

This section discusses the mathematical details and stability of the developed RBFNN controller.

3.1 Design aspects

In Fig. 2a, the activation function of the j th node in the hidden layer is

$$\Phi_j = \exp\left(-\frac{z_j^2}{2b_j^2}\right) \quad (1)$$

where $z_j = ||(X - c_j)||$ and $X = [x_1, x_2, x_3 \dots x_n]$ denote the input vector for the NN, $c_j = [c_{j1}, c_{j2}, c_{j3} \dots c_{jn}]$ denotes the centre vector for the j th node, b_j is basis width of the j th node. The output of the generalized RBFNN model is

$$y_n(k) = w_1\Phi_1 + w_2\Phi_2 + \dots w_n\Phi_n \quad (2)$$

where w_1, w_2, \dots, w_n are the weights to be updated.

The conventional RBFNN controller in Fig. 2a computes single output once all the weights have been tuned. In contrast, the proposed configuration in Fig. 2b considers only a single neuron in the hidden layer which is updated using generalized LMS technique as shown in Eq. (3)

$$w_{kj}(k+1) = w_{kj}(k) - \alpha_1 \frac{\partial E}{\partial w_{kj}} \quad (3)$$

where α_1 is the convergence factor and < 1 , the single weight w_{11} is updated as

$$w_{11}(k+1) = w_{11}(k) - \alpha_1 \frac{\partial E}{\partial w_{11}} \quad (4)$$

where $E = 0.5 * (d_1 - y_1)^2$ denotes the error and is computed using the desired output (d_1) and actual output (y_1). Substituting

$$\frac{\partial E}{\partial w_{11}} = -(d_1 - y_1)\Phi_1 \quad (5)$$

The weight equation is now updated as

$$w_{11}(k+1) = w_{11}(k) + \alpha_1(d_1 - y_1)\Phi_1 \quad (6)$$

Similarly, the centre c_{ji} can be updated as

$$c_{11}(k+1) = c_{11}(k) - \alpha_2 \frac{\partial E}{\partial c_{11}} \quad (7)$$

where α_2 is another convergence factor < 1 . Using the chain rule for derivatives,

$$\frac{\partial E}{\partial c_{11}} = \frac{\partial E}{\partial y_1} \frac{\partial y_1}{\partial \Phi_1} \frac{\partial \Phi_1}{\partial z_1} \frac{\partial z_1}{\partial c_{11}} \quad (8)$$

These partial derivatives are computed as

$$\begin{aligned} \frac{\partial E}{\partial y_1} &= -(d_1 - y_1), \quad \frac{\partial y_1}{\partial \Phi_1} = w_{11}, \\ \frac{\partial \Phi_1}{\partial z_1} &= -\frac{\Phi_1 z_1}{2b_1^2} \frac{\partial z_1}{\partial c_{11}} \text{ and } z_1 \frac{\partial z_1}{\partial c_{11}} = -(x_1 - c_1) \end{aligned} \quad (9)$$

The updation for center, c_{11} is obtained as

$$c_{11}(k+1) = c_{11}(k) + \alpha_2(d_1 - y_1)w_{11}(k) \frac{\Phi_1}{2b_1^2} (x_1 - c_1) \quad (10)$$

For simplification, b_1^2 is taken as 1.0. The diagram showing the online updation of weight and center for system is shown in Fig. 2b. This is extended for the development of single-phase controller in Fig. 3. The weights computed from

proposed RBFNN (w) using a gain factor of $k = 0.33$ and added to the output of the PI controller (w_{loss}). A standard hysteresis current controlled loop is used for current control and gating pulses are generated.

3.2 Determination of the reference compensating current

In practice, unexpected transients impact the system performance owing to rapid changes in load. Monitoring and regulating the DC link voltage (V_{DC}) are crucial for efficient SAPF compensation service. To reduce the error in DC link voltage (V_{DC}), a typical conventional proportional-integral (PI) feedback controller is required and may be easily constructed. The error can be expressed as follows:

$$e_{\text{DC}} = V_{\text{DCref}} - V_{\text{DC}} \quad (11)$$

For switching losses requirement, the reference value of DC link (V_{DCref}) is subtracted from real time DC link voltage (V_{DC}), and thus switching loss will be

$$w_{\text{loss}} = k_p e_{\text{DC}} + k_i \int e_{\text{DC}} dt \quad (12)$$

where k_p and k_i are the PI controller's proportional and integral gains. The overall active power demand of the load is computed by estimating the basic active power component of load current and the loss component determined from dc link voltages (V_{DC}).

$$w_{\text{eff}} = w_{\text{loss}} + w_p \quad (13)$$

Furthermore, the required fundamental reference current (i_s^*) is calculated by multiplying the predicted fundamental active component of the load (w_{eff}) by the unit in-phase component (u_p).

$$i_s^* = w_{\text{eff}} \times u_p \quad (14)$$

where u_p is unit vector template generated by using second order generalized integrator (SOGI) filter as shown in Fig. 3.

$$V_t = \sqrt{v_{\text{sp}}^2 + v_{\text{sq}}^2}; \quad u_p = \frac{v_{\text{sp}}}{v_t} \quad (15)$$

where V_t is maximum amplitude of voltage signals and v_{sp} is in-phase and v_{sq} is quadrature components.

3.3 Stability analysis

Lyapunov's stability is one of the most essential pillars of control theory and it remains popular among researchers due to its simplicity, universality and usefulness [30]. As per Lyapunov method, proposed system's energy reduces along with

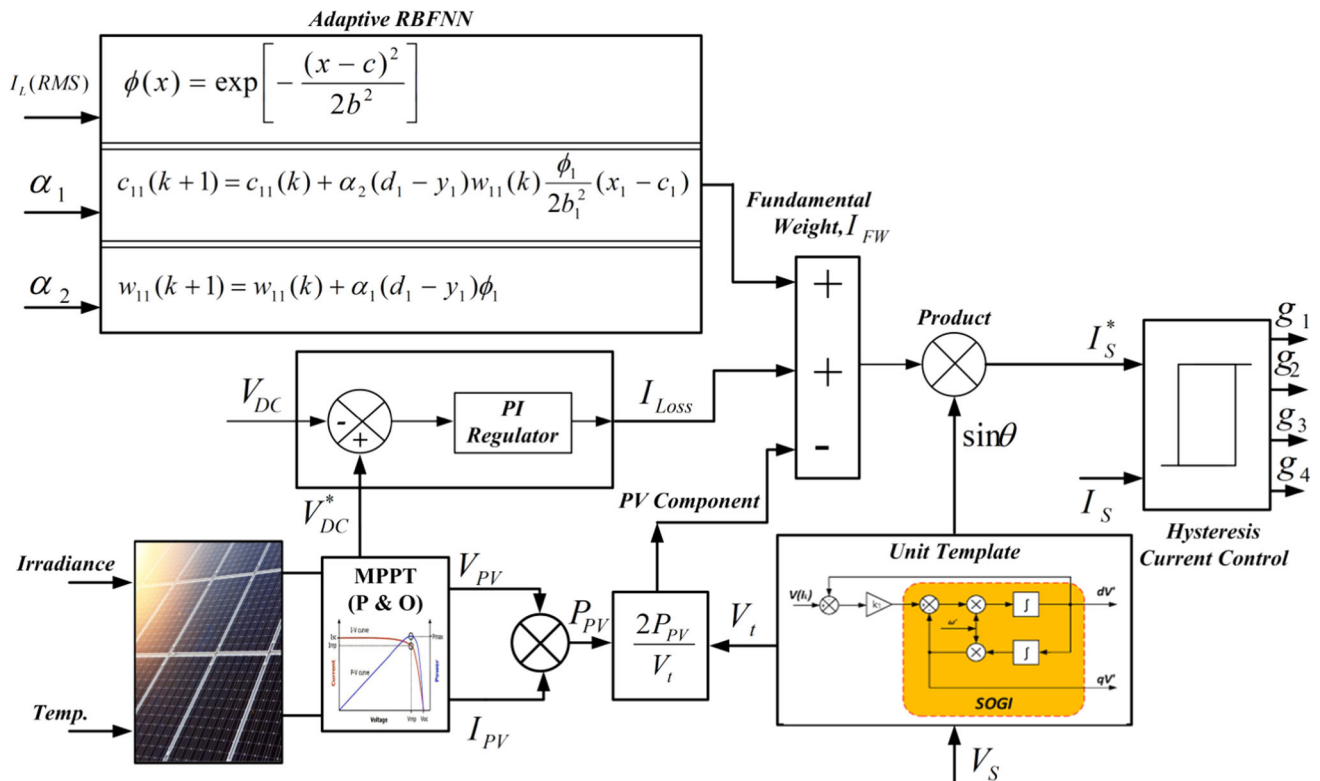


Fig. 3 Adaptive RBFNN controller design for single-phase system

trajectories of the proposed system. The Lyapunov's stability theorem states that a nonlinear system is globally asymptotically stable if the Lyapunov function $F(k)$ has following properties [31]

$$\begin{cases} F(0) \\ F(k) > 0 \text{ for all } k \neq 0. \\ \dot{F}(k) < 0 \text{ for all } k \neq 0 \end{cases}$$

For the given system, the Lyapunov stability theory can be established considering a Lyapunov function is

$$F(k) = e(k)^2 \quad (16)$$

which is a positive definite function for all i

$$F(k) = \Delta F(k) = e(k)^2 - e(k-1)^2 \quad (17)$$

where

$$e(k) = d(k) - y(k) \text{ And } y(k) = w^T(k)\Phi_1(k)$$

Moreover, from Eq. (6), the weight updation is

$$w(k) = w(k-1) + \alpha_1 e(k)\Phi_1(k)$$

hence $\Delta F(k)$ is simplified as

$$\Delta F(k) = (d(k) - y(k))^2 - e(k-1)^2 \quad (18)$$

Table 2 Simulation and experimental parameters for considered system

Parameters	Simulation	Experimental
Grid Voltage	$V_S = 110 \text{ V}$ (without and with PV)	$V_S = 40 \text{ V}$ (without & with PV)
Grid frequency	$\omega_o = 2\pi \times 50 \text{ rad/s}$	$\omega_o = 2\pi \times 50 \text{ rad/s}$
Sampling time	$T_s = 50 \mu\text{s}$	$T_s = 50 \mu\text{s}$
Interfacing inductor	$L_s = 3.5 \text{ mH}$	$L_s = 3.5 \text{ mH}$
Irradiance	$I_{rr} = 1000 \text{ W/m}^2$	$I_{rr} = 1000 \text{ W/m}^2$
PV rating	2.7 kW	500 W
DC-link capacitor	$C_{DC} = 4700 \mu\text{F}$	$C_{DC} = 5600 \mu\text{F}$
Feeder parameters	$L_S = 0.5 \text{ mH}$, $R = 0.02 \Omega$	$L_S = 0.5 \text{ mH}$, $R = 0.02 \Omega$
Controller gains	$k_p = 1.5$ and $k_i = 0.33$	Adjusted and tuned
Non-linear load	Single phase diode rectifier with $R = 80 \Omega$ and $L = 100 \text{ mH}$	Single phase diode rectifier with $R = 90 \Omega$ and $L = 100 \text{ mH}$
Convergence Parameters	$\alpha_1 = 10^{-2}$ and $\alpha_2 = 50 \times 10^{-3}$	$\alpha_1 = 10^{-2}$ and $\alpha_2 = 50 \times 10^{-3}$

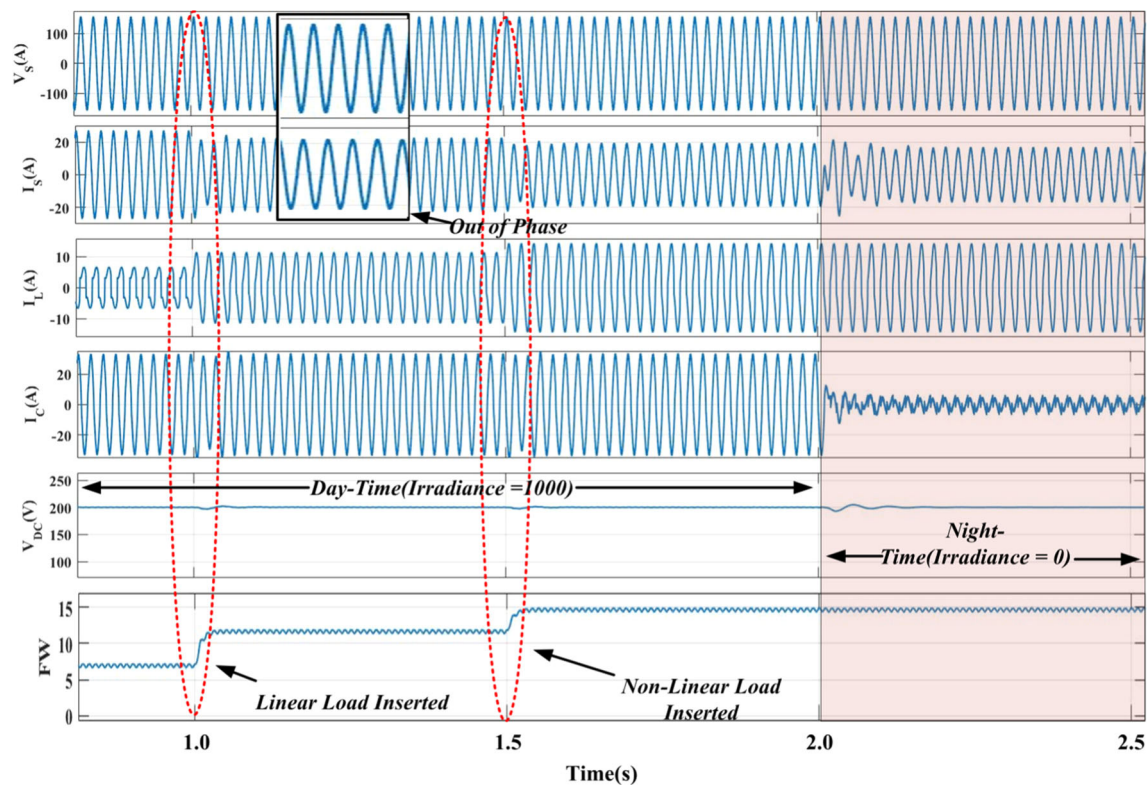


Fig. 4 Performance of controller showing **a** Source voltage, V_s **b** Source current, I_s **c** Load current, I_L **d** Converter current, I_C **e** DC link voltage, V_{DC} **f** Fundamental weight, FW

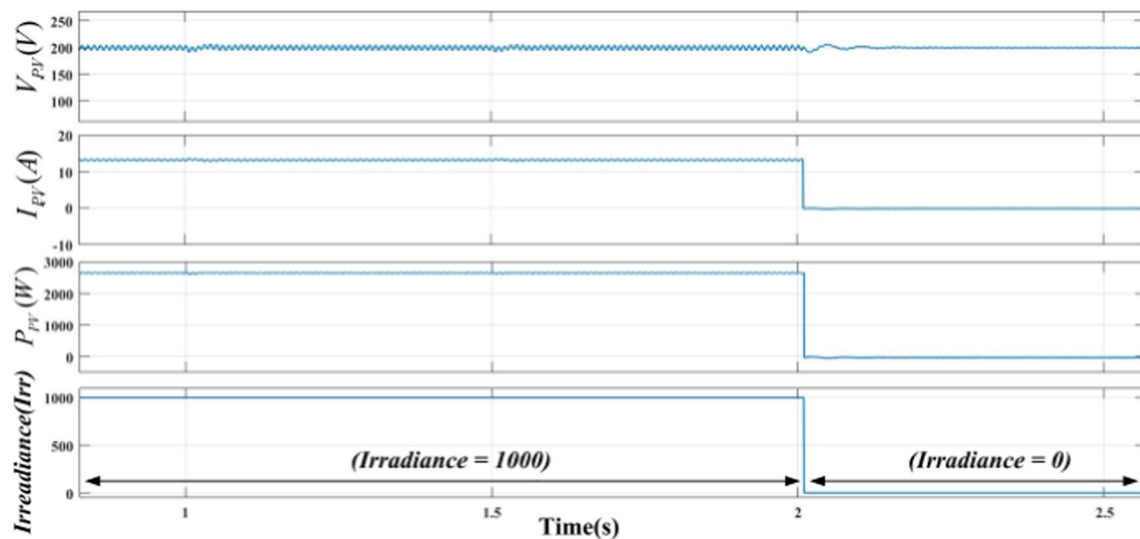


Fig. 5 Waveform of **a** PV voltage, V_{PV} **b** PV current, I_{PV} **c** PV power, P_{PV} **d** Irradiance, I_{rr}

$$= \left(d(k) - w^T(k) \Phi_1(k) \right)^2 - e(k-1)^2 \quad (19) \quad = e^2(k) \left[1 - \alpha_1 \Phi_1^2(k) \right]^2 - e(k-1)^2 \quad (22)$$

$$= \left(d(k) - w(k-1) \Phi_1(k) - \alpha_1 e(k) \Phi_1^2(k) \right)^2 - e(k-1)^2 \quad (20) \quad = e^2(k) \left[1 - 2\alpha_1 \Phi_1^2(k) + \alpha_1^2 \Phi_1^4(k) \right] - e(k-1)^2 \quad (23)$$

$$= \left(e(k) - \alpha_1 e(k) \Phi_1^2(k) \right)^2 - e(k-1)^2 \quad (21) \quad \text{Now, parameter } \alpha_1 \text{ is quite small, hence } \frac{2}{1} \text{ is negligible, hence}$$

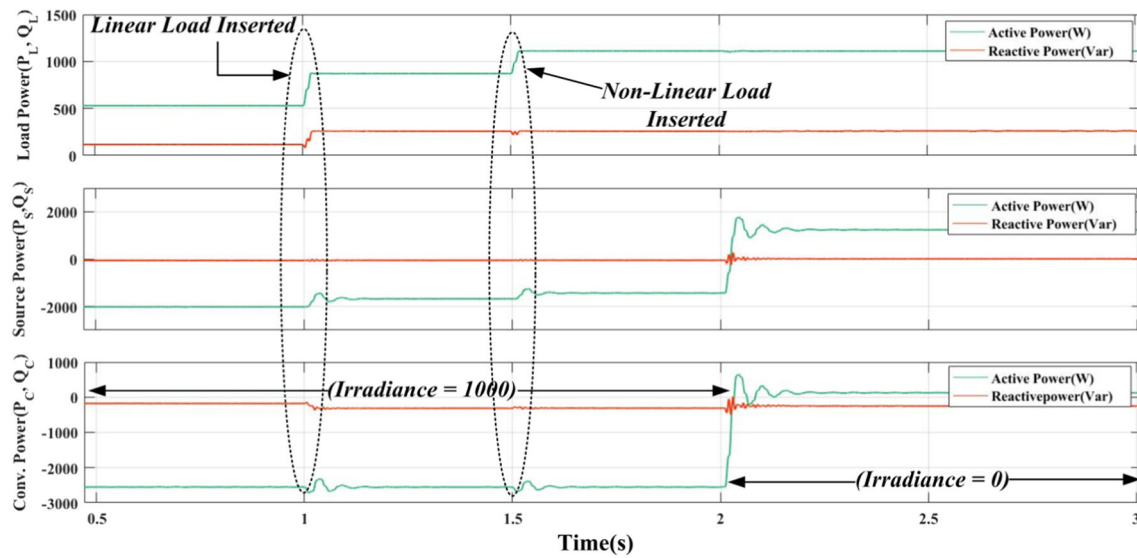


Fig. 6 Waveform of **a** Load Power (P_L , Q_L) **b** Source Power (P_S , Q_S) **c** Converter Power (P_C , Q_C)

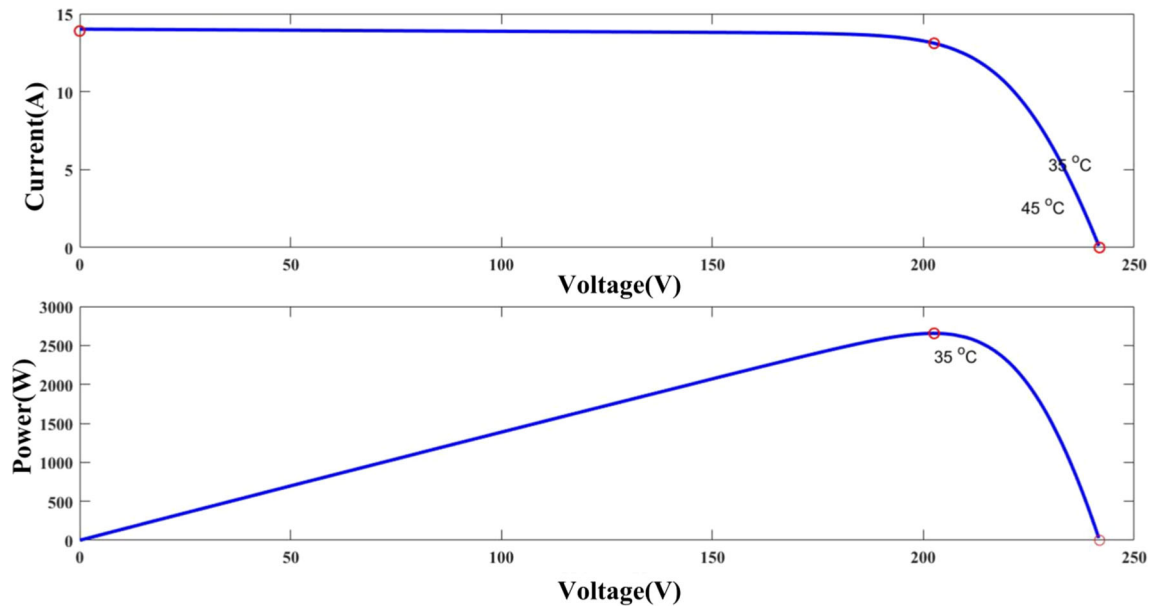


Fig. 7 PV and IV curves for PV module for different irradiance conditions and fixed temperature (35 °C)

$$\Delta F(k) \sim -2\alpha_1 \Phi_1^2(k) e^2(k) \quad (24)$$

which is negative for all k except equilibrium point \hat{k} , as per Lyapunov theorem, it is concluded that the proposed system is asymptotically stable.

4 Result and discussions

This Section presents the results with the developed adaptive RBFNN technique for a single-phase grid-connected

and PV integrated power distribution system feeding non-linear loads and linear loads. The system is simulated and its operation under various conditions is investigated using Simulink/MATLAB. The developed RBFNN network as shown in Fig. 3 is first trained to correctly extract the peak value of the fundamental component of load current. Table 2 highlights the design value of parameters of the proposed system for simulation as well as experimental studies.

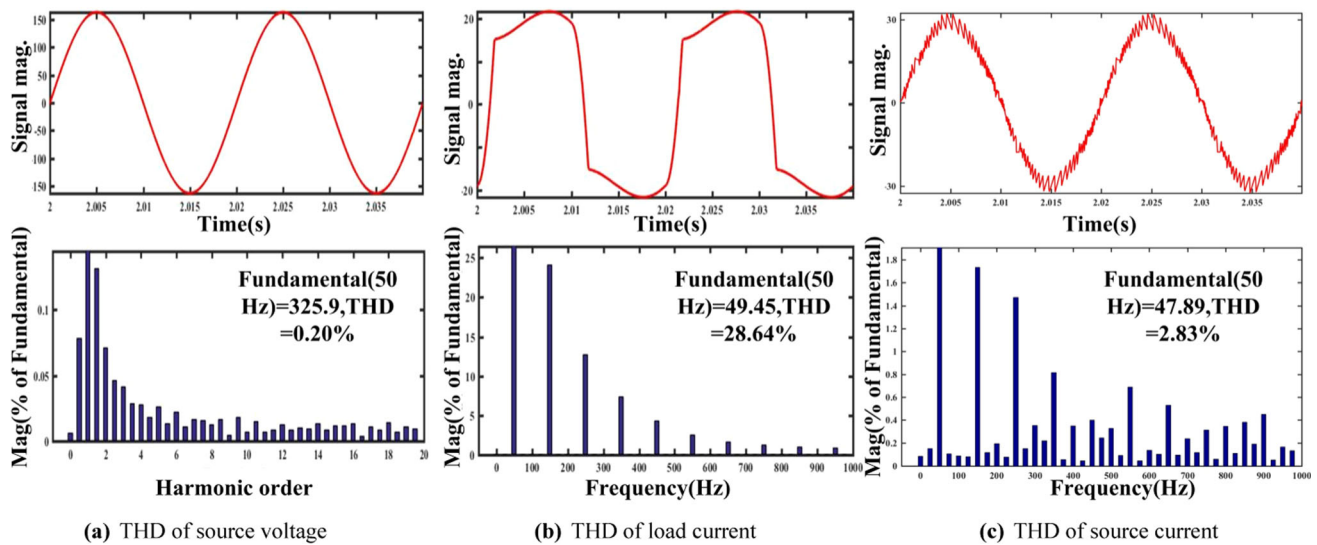


Fig. 8 **a** THD of source voltage, **b** THD of load current, **c** THD of source current

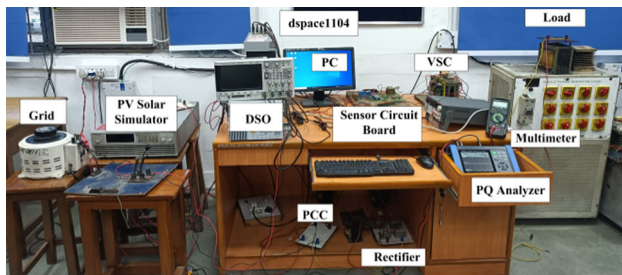


Fig. 9 Experimental setup

4.1 Simulation results and discussion

Figure 4 depicts the MATLAB/Simulink-based outcomes under various operating situations. Figure 4 shows the plots for supply voltage (V_s), supply currents (I_s), load currents (I_l), compensator currents (I_c), DC link voltage (V_{dc}) and fundamental weight (FW). A single-phase supply of 110 V, 50 Hz is system ac input and both linear and non-linear loads are considered. Initially, the connected load comprises linear and non-linear components. Both day time and night time cases are explored. During daytime PV array with irradiance of 1000 W/m² is simulated upto $t = 2$ s and beyond $t = 2$ s PV array irradiance is reduced to 0 W/m² (considered as night time). Thus, the results in Fig. 4 highlight the effect of PV integration till $t = 2$ s and its subsequent removal thereafter. The night time operation corresponds to SAPF operation without RES integration. The tracked weight is shown in Fig. 4 and it is observed that the actual weight tracking is fast even when the linear load inserted at $t = 1$ s and a non-linear load increased at $t = 1.5$ s. Convergence is reached within two cycles irrespective of the load changes. Both the parameters.

Figure 5 shows output parameters of PV panels PV output voltage in volt (V_{PV}), PV output current in ampere (I_{PV}), PV output power in watt (P_{PV}) and irradiance in W/m² (I_{rr}). Before $t = 2$ s I_{rr} is 1000 W/m² and at $t = 2$ s irradiance is reduced to 0 W/m² to simulate the night time conditions. Accordingly output power of PV changes from 2.7 kW to 0 kW as shown in Fig. 5.

Figure 6 shows the load active and reactive power (P_L , Q_L), supply active and reactive power (P_S , Q_S) and converter active and reactive power (P_C , Q_C). For this case, the load demand before $t = 1$ s is around 600 W and the power supplied by PV array is 2700 W and, therefore, the remaining 2000 W PV power is supplied back to the grid. At $t = 1$ s additional linear load is inserted in the system so the load demand increases to 850 W and during this time, the surplus PV power of 1750 W is supplied back to the grid. Further at $t = 2$ s, non-linear load is increased which enhances the load demand up to 1100 W. Now, a net surplus of 1500 W PV power is injected back to the grid. The DC link voltage exhibits a little perturbation during load changes but stabilizes due to PI controller action.

The simulation results of the system in Fig. 4 after $t = 2$ s pertain to zero irradiance value which can be considered as night time operation. During this period, the generated PV output power is zero and the entire active power demand of the load is supplied by the grid as shown in Fig. 6. At $t = 2$ s, there is a sudden removal of PV array at the inverter DC link. Under this large transient also, it is observed that the DC link voltage is still regulated. The PI controller on the DC link voltage regulates it to a reference value of 200 V. All the reactive power demand of the load is supplied by the VSC which is working as a SAPF or shunt compensator. Power balance is achieved between the inverter, grid and the load. Figure 7

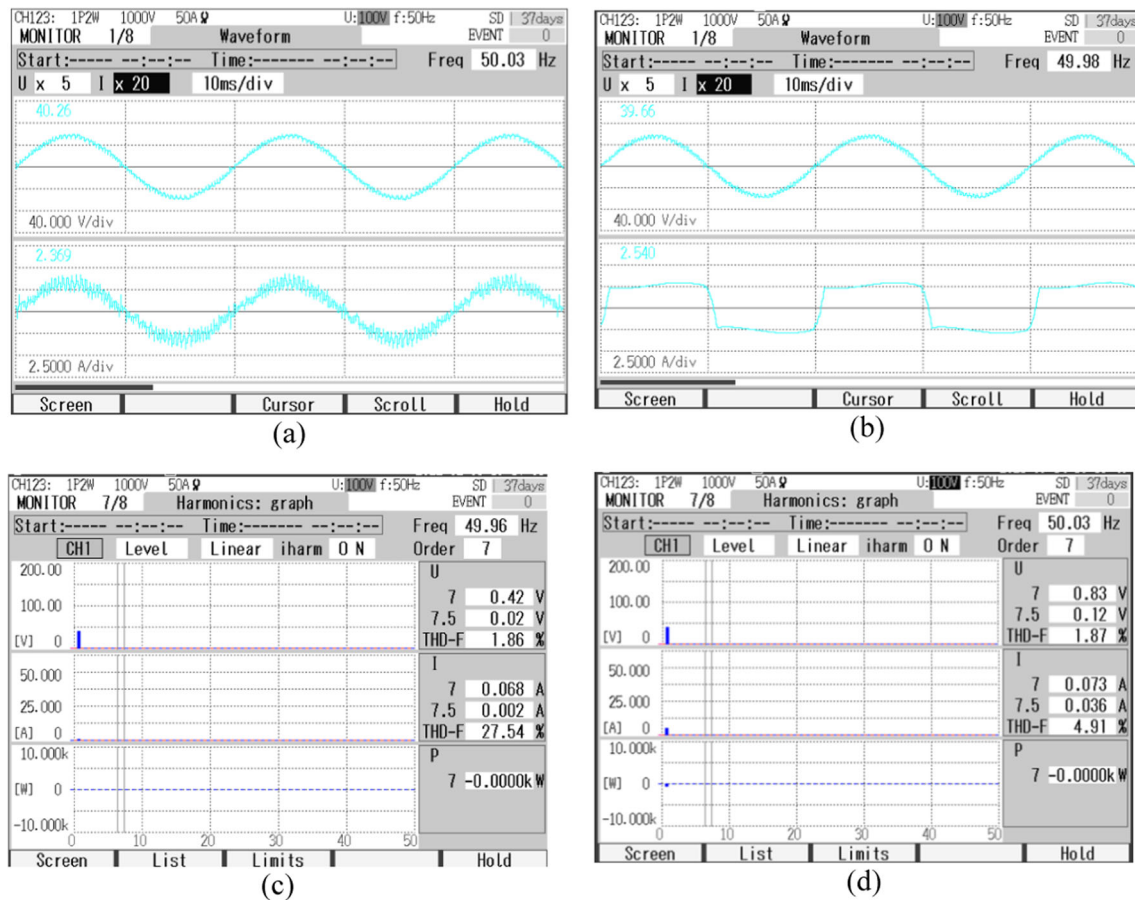


Fig. 10 Waveforms showing **a** Source voltage (V_s) and grid current (I_s) **b** Source voltage (V_s) and load current (I_L) **c** THD of V_s (1.87%) and I_L (27.54%) **d** THD of V_s (1.87%) and I_s (4.91%)

shows the PV curve of array at given temperature and irradiance during the operation of SAPF PV; and it is observed that the array is delivering maximum possible power to the system extracted using the P&O MPPT technique.

Figure 4 shows PQ improvement results and focuses on pf correction and harmonic mitigation. Severe load variation is introduced at $t = 1$ s and 2 s. The compensator injects the necessary currents and the grid currents are sinusoidal. Thus, the developed adaptive RBFNN controller achieves load compensation. Figure 8 shows the total harmonic distortion (THD) in source voltage, source current and load current to be 0.18%, 2.92% and 28.71% respectively during closed-loop operation of SAPF. The proposed controller is able maintain THD level of source current as per IEEE 519 standard.

4.2 Experimental results and discussion

An experimental setup of a single phase system has been developed to validate the simulation results of the proposed system. Figure 9 depicts the prototype hardware setup built in the lab. The performance of adaptive RBFNN is investigated

in steady-state and dynamic changes under various loading scenarios for both the cases viz. SAPF without and with PV integrated system.

4.2.1 Experimental results without PV integration

The steady-state performance of the system is shown in Fig. 10. Figure 10a shows waveform of source voltage (V_s) and source current (I_s). Figure 10b shows the waveform of source voltage (V_s) and load current (I_L). The THD in load current is 27.54% as shown in Fig. 10c. The THD in source voltage is 1.87% and THD in source current 4.91% shown in Fig. 10d which as per IEEE 519 standard. The proposed single phase SAPF performs harmonic reduction satisfactorily for supply currents. In addition, the supply voltage and current have an in-phase relationship showing almost unity power factor operation.

The dynamic state performance of proposed system is taken during load disturbances with help of digital storage oscilloscope. The dynamic performance of the system is shown in Fig. 11a–d in which source voltage (V_s), source current (I_s), load current (I_L), DC link voltage (V_{DC}), reference

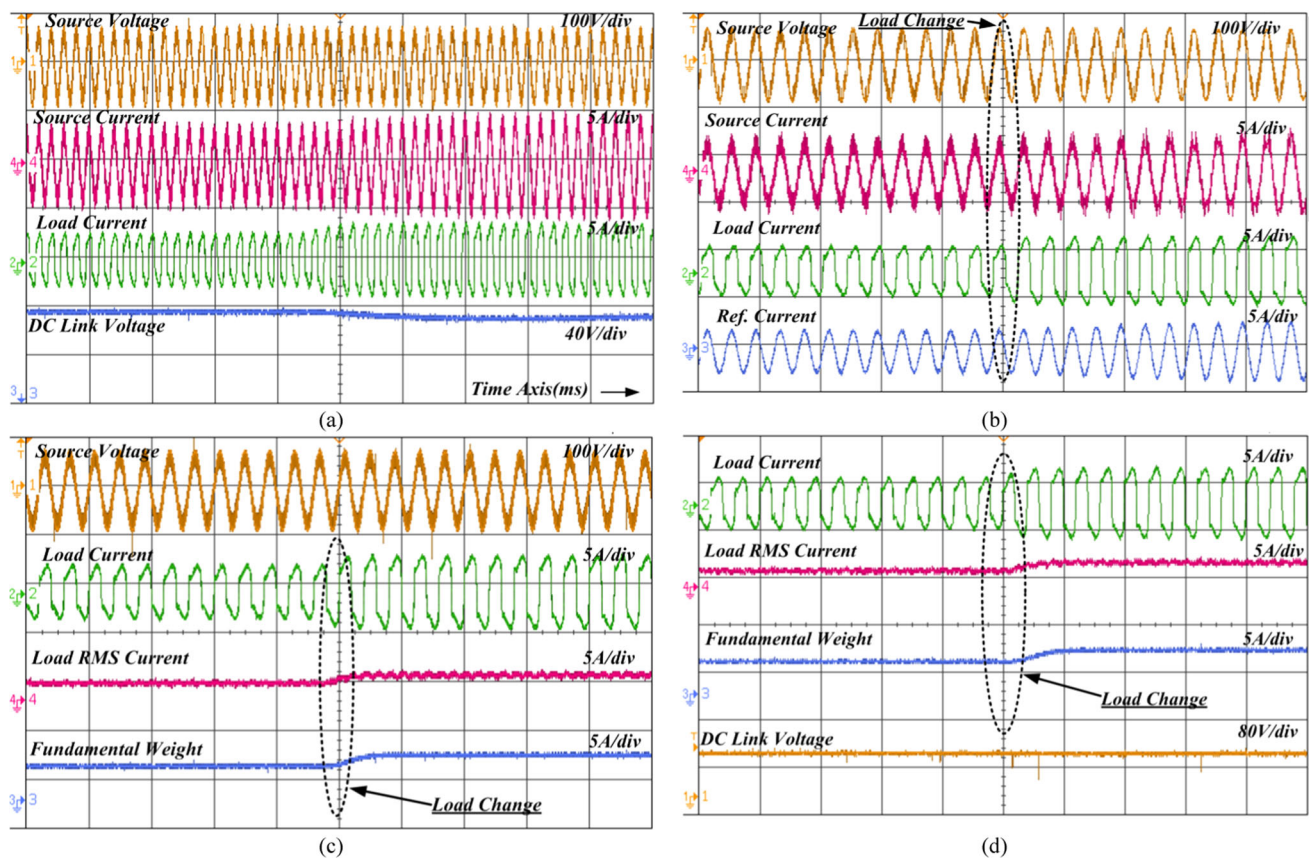


Fig. 11 Dynamic response during load increase using adaptive RBFNN controlled SAPF **a** V_s , I_s , I_L , V_{DC} **b** V_s , I_s , I_L , I_C **c** V_s , I_L , $I_{L(rms)}$, FW **d** I_L , $I_{L(rms)}$, FW , V_{DC}

current (I_{Ref}), load RMS current ($I_{L(rms)}$), and output fundamental weight are shown. From Fig. 11a–d it is observed that during the change in load, the DC link voltage is stable and the grid current is sinusoidal. The harmonic content in grid current is also maintained as per IEEE 519 standard. The fundamental gain FW shows fast convergence and reference current changes as per load changes. This indicates a satisfactory operation of RBFNN-based control technique under different loading condition. The steady-state of the system is achieved within one cycle of operation.

4.2.2 Experimental results with PV integration

Figure 12 depicts the steady-state performance of the proposed system integrated with PV. For this section, the DC link is set to 80 V, the grid supply is set to 40 V, 50 Hz and a 500 W PV array is integrated into the system through H-bridge. The single phase system operates in the MPPT mode. This set of results is obtained using the PV array simulator. Figure 12a depicts the waveform of (V_s , I_L), and Fig. 12b depicts the power demand of the load supplied by PV 78.6 kW. Figure 12c depicts the waveforms of (V_s , I_s) and Fig. 12 depicts the power delivered to the grid by PV

301.8 kW. Figure 12e–f depict THD of load current (33.61%) and THD of source current (3.98%). In this study, PV meets the load power demand and also simultaneously exports the net surplus power to the grid. During PV operation, the maximum power is drawn from PV using MPPT at the provided DC link voltage. Fig. 13a–c shows dynamic performance of proposed system with PV integrated at the DC link of the inverter. Figure 13a shows (V_s , I_s , I_L , V_{DC}) and clearly an out of phase relationship between the source voltage and source current are visible due to PV integration. Figure 13b shows (V_s , I_s , I_L , I_{ref}) and it shows the source current.

follows the reference current. The reference current is computed using the proposed RBFNN controller. Figure 13c shows (I_s , I_L , I_{FW} , V_{DC}). It is observed that during the operation of PV integrated SAPF, PV supplies power to both load and grid. The operation is stable and irrespective of highly non-linear load current, the grid current is sinusoidal and meets IEEE 519 standard. Thus, PQ improvement is obtained with the designed controller.

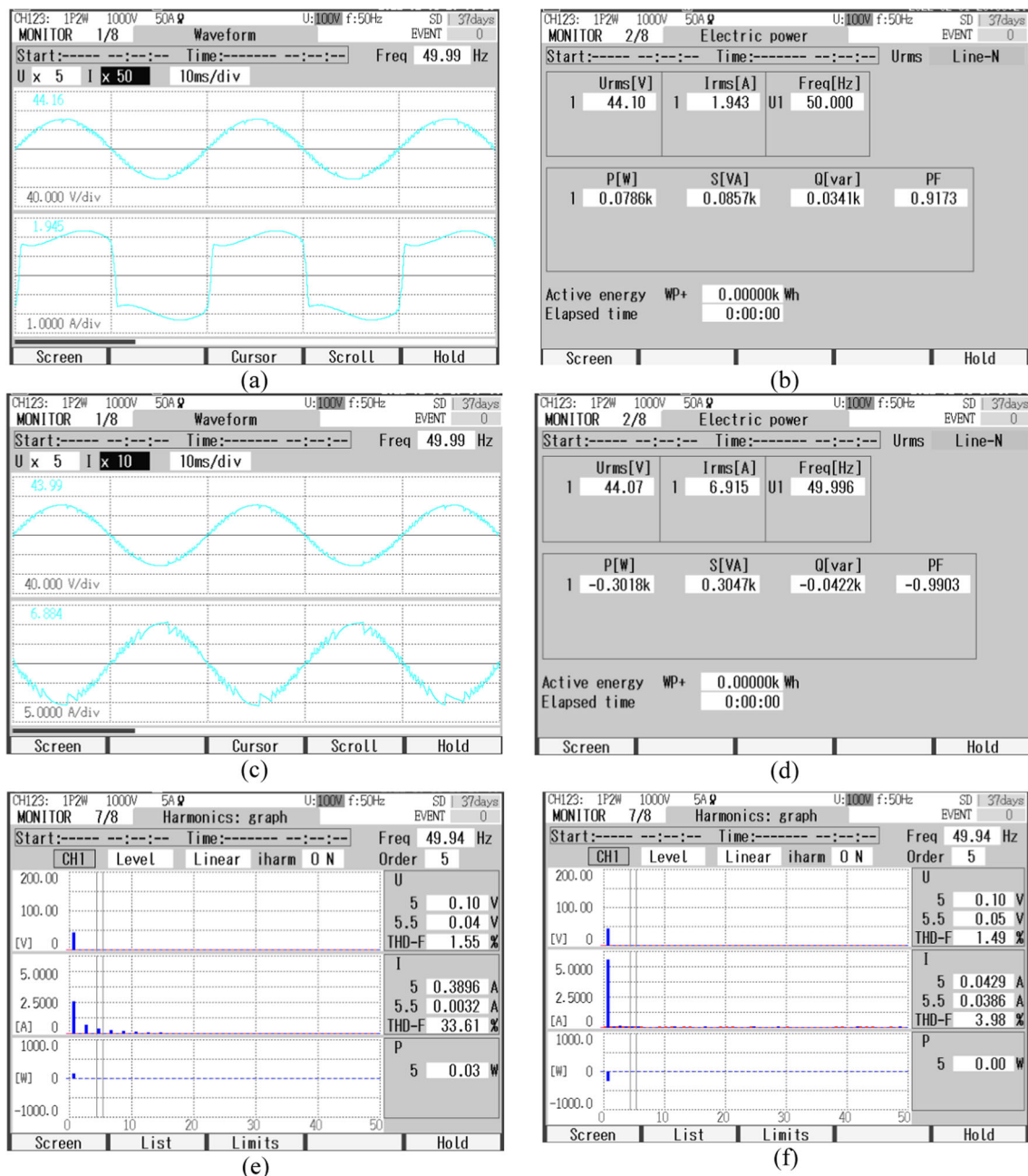


Fig. 12 Wave or ms showing **a** Source voltage (V_S) and load current (I_L) **b** Load power demand (P_L , Q_L) **c** Source voltage (V_S) and grid current (I_S) **d** Power supplied to the grid **e** THD of load current (I_L) **f** THD of source current (I_S)

5 Performance Comparison

Comparison results with a conventional SRF technique and conventional RBFNN model are presented in Table 3. It shows that SRF technique is phase locked-loop (PLL)-dependent and complex in nature. SRF and conventional RBFNN show slow convergence of weights and both techniques are non-adaptive in nature. Also, total harmonic distortion (THD) of supply currents of both techniques is

much lower with the proposed technique. The SRF technique and the conventional RBFNN needs modifications for satisfactory performance. The proposed adaptive RBFNN is simple and gives excellent results under load variations. It can be applied to SAPF system, as well as to SAPF integrated, with PV array.

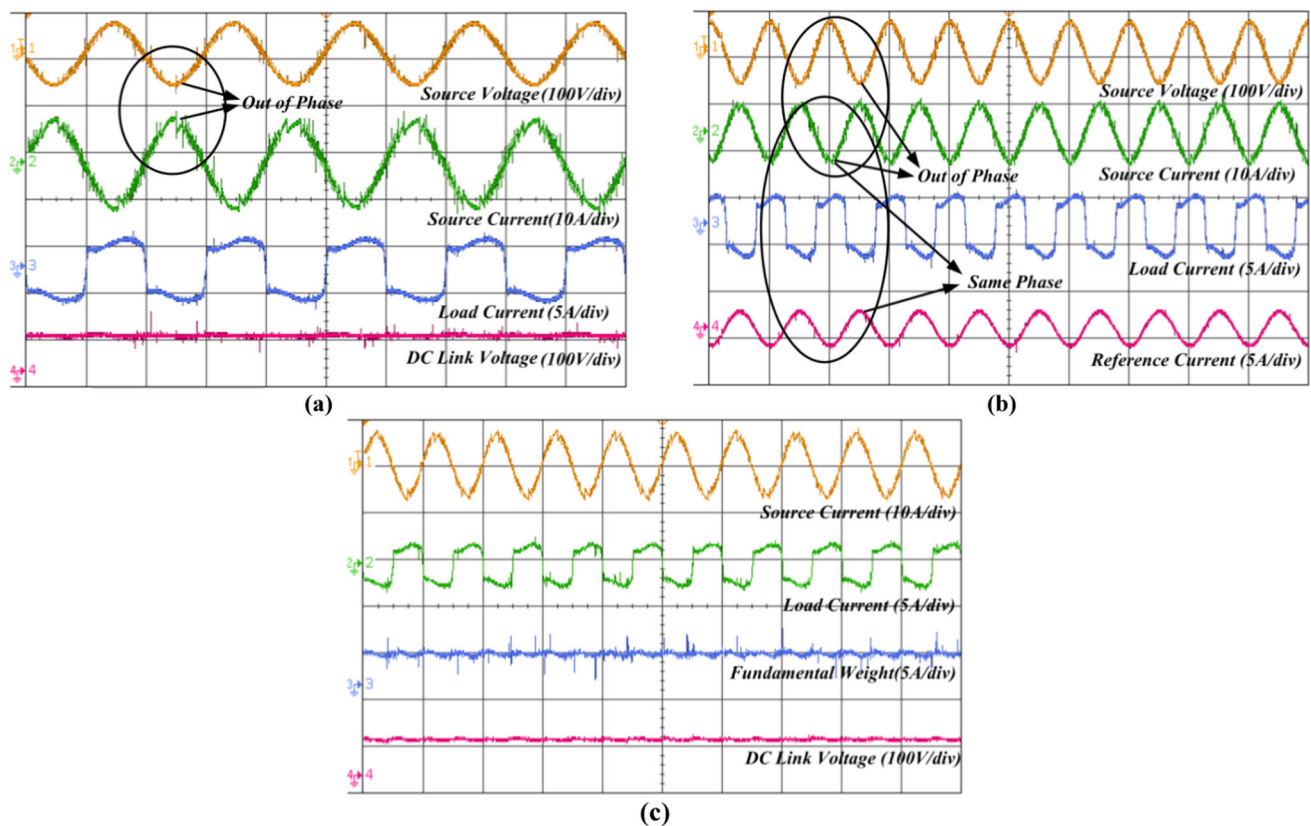


Fig. 13 Response of adaptive RBFNN controlled SAPF integrated with PV **a** V_s , I_s , I_L , V_{DC} **b** V_s , I_s , I_L , I_{ref} **c** I_s , I_L , I_{FW} , V_{DC}

Table 3 Comparative performance of different techniques

S. no	Parameter	SRF technique [11]	Conventional RBFNN [22]	Proposed RBFNN
1	PLL (requirement)	Yes	No	No
2	Technique	Non-adaptive	Non-adaptive	Self- adaptive
3	Tracking performance	Moderate	Moderate	Very good
4	Weight Convergence	Slow (> 2 cycles)	Slow(> 3cycle)	Fast (< 2 cycles)
5	% THD of I_s	4–5%	– 4%	3%
6	Dependency on control parameters	Highly dependent	Yes	No, self-adaptive
7	Complexity	Very high	Medium	Lower
8	Meets International standards	Yes but requires modifications	Yes but requires modifications	Yes

6 Conclusion

This work's major contributions include the design of self-adaptive RBFNN. This proposed structure is important for accomplishing load compensation in single-phase systems feeding a range of linear and non-linear loads. The single-phase H-bridge inverter is employed as a SAPF to mitigate many power quality issues. The use of a RBFNN-based

controller is broad in scope, and it may be utilized for grid-connected systems with or without a PV source interfaced to it. The system performance aspects are demonstrated when taking into account the influence of PV as well as at during night. The control algorithm has been designed to fulfill the entire reactive power needs of the load while also providing power quality enhancement features. Performance results show that a single- input, single- neuron NN can be trained

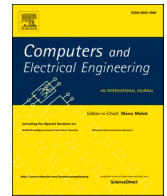
effectively using LMS technique to estimate the fundamental component of load current accurately. Unity power factor operation and harmonic elimination have been presented using the developed controller. Moreover, its comparison with SRF highlights the positive features of the developed controller which include fast convergence and estimation of weights and self-adaptive nature.

References

1. Belaidi R, Hatti M, Haddouche A, Larafi MM (2013) Shunt active power filter connected to a photovoltaic array for compensating harmonics and reactive power simultaneously. In: 4th international conference on power engineering, energy and electrical drives. pp 1482–1486. <https://doi.org/10.1109/PowerEng.2013.6635834s>
2. Haque A, Zaheeruddin (2013) Research on solar photovoltaic (PV) energy conversion system: an overview. In: Third international conference on computational intelligence and information technology (CIIT 2013). pp 605–611. <https://doi.org/10.1049/cp.2013.2653s>
3. (2020) Solar–10 predictions for 2022. BNEF – Bloomberg new energy finance. Retrieved 1 Feb 2022
4. Szemes PT, Melhem M (2020) Analyzing and modeling PV with “P&O” MPPT algorithm by MATLAB/SIMULINK. In: 2020 3rd international symposium on small-scale intelligent manufacturing systems (SIMS). pp 1–6. <https://doi.org/10.1109/SIMS49386.2020.9121579>
5. Juamperez M, Yang G, Kjær SB (2014) Voltage regulation in LV grids by coordinated volt-var control strategies. *J Mod Power Syst Clean Energy* 2:319328. <https://doi.org/10.1007/s40565-014-0072-0>
6. Singh B, Chandra A, Al-Haddad K (2015) Power quality: problems and mitigation techniques. Wiley, New Jersey
7. Ghosh A, Ledwich G (2009) Power quality enhancement using custom power devices. Springer, Delhi
8. Chittora P, Singh A, Singh M (2018) Chebyshev functional expansion based artificial neural network controller for shunt compensation. *IEEE Trans Ind Inf* 14(9):3792–3800
9. Reddy SS, Bijwe PR (2015) Real time economic dispatch considering renewable energy resources. *Renew Energy* 83:1215–1226
10. Badoni M, Singh A, Singh B (2016) Comparative performance of wiener filter and adaptive least mean square-based control for power quality improvement. *IEEE Trans Ind Electron* 63(5):3028–3037. <https://doi.org/10.1109/TIE.2016.2515558>
11. Sanjan S, Yamini NG, Gowtham N (2020) Performance comparison of single-phase SAPF using PQ theory and SRF theory. In: 2020 international conference for emerging technology (INCET). pp 1–6. <https://doi.org/10.1109/INCET49848.2020.9154126>
12. Akagi H (1996) New trends in active filters for power conditioning. *IEEE Trans Ind Appl* 32(3):1312–1322
13. Akagi H, Kanazawa Y, Nabae A (1983) Generalized theory of the instantaneous reactive power in three-phase circuits. In: Proc. IEEE int. power electron. conf., Tokyo, Japan. pp 1375–1386
14. Raj GS, Rathi K (2015) P-Q theory based shunt active power filter for power quality under ideal and non-ideal grid voltage conditions. In: 2015 international conference on power, instrumentation, control and computing (PICC). pp 1–5. <https://doi.org/10.1109/PICC.2015.7455754>
15. Li H, Zhuo F, Wang Z, Lei W, Wu L (2005) A novel time-domain current-detection algorithm for shunt active power filters. *IEEE Trans Power Syst* 20(2):644–651. <https://doi.org/10.1109/TPWRS.2005.846215>
16. Girgis AA, Chang WB, Makram EB (1991) A digital recursive measurement scheme for on-Line tracking of power system harmonics. *IEEE Trans Power Del* 3:1153–1160
17. Singh B, Arya SR (2014) Back-propagation control algorithm for power quality improvement using DSTATCOM. *IEEE Trans Ind Electron* 61(3):1204–1212
18. Arora A, Singh A (2019) Design and analysis of functional link artificial neural network controller for shunt compensation. *IET Gener Transm Distrib* 13(11):2280–2289
19. Arora A, Singh A (2019) Design and implementation of Legendre-based neural network controller in grid-connected PV systems. *IET Renew Power Gener* 13(15):2783–2792
20. Saxena H, Singh A, Rai JN (2020) Adaptive spline-based PLL for synchronisation and power quality improvement in distribution system. *IET Gener Transm Distrib* 14(7):1311–1319. <https://doi.org/10.1049/iet-gtd.2019.0662>
21. Mehrankia A, Mollakhalili Meybodi MR, Mirzaie K (2022) Prediction of heart attacks using biological signals based on recurrent GMDH neural network. *Neural Process Lett*. <https://doi.org/10.1007/s11063-021-10667-8>
22. Panda S, Panda G (2021) On the development and performance evaluation of improved radial basis function neural networks. *IEEE Trans Syst Man Cybern Syst*. <https://doi.org/10.1109/TSMC.2021.3076747>
23. Mishra S (2006) Neural-network-based adaptive UPFC for improving transient stability performance of power system. *IEEE Trans Neural Netw* 17(2):461–470
24. Dash PK, Mishra S, Panda G (2000) A radial basis function neural network controller for UPFC. *IEEE Trans Power Syst* 15(4):1293–1299
25. Jie H, Zheng G, Zou J, Xin X, Guo L (2020) Speed regulation based on adaptive control and RBFNN for PMSM considering parametric uncertainty and load fluctuation. *IEEE Access* 8:190147–190159
26. Luo Y, Zhao S, Yang D, Zhang H (2020) A new robust adaptive neural network backstepping control for single machine infinite power system with TCSC. *IEEE/CAA J Autom Sin* 7(1):48–56
27. Baghaee HR, Mirsalim M, Gharehpetan GB, Talebi HA (2018) Nonlinear load sharing and voltage compensation of microgrids based on harmonic power-flow calculations using radial basis function neural networks. *IEEE Syst J* 12(3):2749–2759
28. Jiang L, Wu C (2018) Application of the improved RBFNN based on DPC in monthly rainfall forecasting. In: 2018 IEEE international conference of safety produce informatization (IICSPI). pp 853–857
29. Lei X, Lu P (2014) The adaptive radial basis function neural network for small rotary-wing unmanned aircraft. *IEEE Trans Ind Electron* 61(9):4808–4815
30. Lyashevskiy S, Yaobin C (1997) The Lyapunov stability theory in system identification. In: Proceedings of the 1997 American control conference (Cat. No.97CH36041), vol 1. pp 617–621. <https://doi.org/10.1109/ACC.1997.611873>
31. Bajaj M, Flah A, Alowaidi M, Sharma NK, Mishra S, Sharma SK (2021) A lyapunov-function based controller for 3-phase shunt active power filter and performance assessment considering different system scenarios. *IEEE Access* 9:66079–66102. <https://doi.org/10.1109/ACCESS.2021.307524>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.



Research Article

Plant integrated proportional integrating based control design for electric vehicle charger[☆]Aakash Kumar Seth, Mukhtiar Singh^{*}

Delhi Technological University, EED, DTU, Delhi 86, India

ARTICLE INFO

Keywords:

Electric vehicle
Battery charger
Plant integrated proportional integrating control
Grid to vehicle
Reactive power compensation

ABSTRACT

The paper presents the control design of on-board two-stage electric vehicle (EV) charger. The first stage AC-DC converter of EV charger plays a very important role in supporting the grid reactively. Therefore, a reduced order plant integrated proportional integrating (PIPI) controller has been designed for first stage of EV charger. Generally, a proportional resonant (PR) regulator is used for tracking of grid current reference in control of AC-DC converter. However, conventional PR regulator is a third order system due to that the complexity of controller is increased. To reduce the complexity of controller, a first order PIPI regulator has been designed which has all the functionalities of conventional PR regulator and can track any periodic signal. This method is implemented in inner current control loop of first stage AC-DC converter. This results the reduction in controller complexity and better dynamic response as compared with conventional PR regulator.

Introduction

Due to the demolition of fossil fuel reserves and environmental conditions, the government and industries are shifting towards the electric vehicles (EVs). The internal combustion engine (ICE) based transportation is major contributor to air pollution, therefore electric vehicles (EVs) are best alternative as they offer significant saving in fuel consumption and help to improve air quality [1]. However, this paradigm shift toward EV is matter of concern as it is expected that the population of EV will increase in forthcoming.

EV mainly consist of four parts: an electric motor with gear box, energy storage system (battery, super capacitor, fuel cell), power electronics converters for battery charging & motor drive and communication system [2]. In last two decades, adequate development has been completed in these EV parts, especially in energy storage system. Most of the EVs are driven by battery and charging time of these batteries create obstacle in success of EV. The charging time of battery is always important because usually it is slower than discharging. However, with the advancement in battery technology, high value of charging current can be passed through battery, offering fast and better charging solutions and may overcome the problem of range anxiety [3]. An EV charger takes around 10 h with level-2 charging to charge the battery pack upto 100% state of charge (SOC). Whereas, level-3 off-board charger takes around half an hour to charge EV battery up to 80% from 20% SOC.

On the basis of operation, the EV charger can be unidirectional or bidirectional. The unidirectional EV charger transfers the power

[☆] This paper is for special section VSI-cccv. Reviews were processed by Guest Editor Dr. Nallapaneni Manoj Kumar and recommended for publication.

^{*} Corresponding author.

E-mail address: smukhtiar_79@yahoo.co.in (M. Singh).

Nomenclature

<i>EV</i>	Electric vehicle
<i>PR</i>	Proportional resonant
<i>I</i>	Integral
<i>PI</i>	Proportional integral
<i>PIPI</i>	Plant integrated proportional integrating
<i>SOC</i>	State of charge
<i>G2V</i>	Grid to vehicle
<i>V2G</i>	Vehicle to grid
<i>MPC</i>	Model predictive control
<i>RC</i>	Repetitive control
<i>QPR</i>	Quasi proportional resonant
<i>ICPD</i>	Integrated controller plant dynamics
<i>IGBT</i>	Insulated gate bipolar transistor
K_p	Proportional gain
K_i	Integral gain
K_r	Resonant gain
R	Virtual resistance
P	Active power
Q	Reactive power
V_g	Grid voltage
I_g	Grid current
V_α	Alpha component of grid voltage
I_α	Alpha component of grid current
V_β	Beta component of grid voltage
I_β	Beta component of grid current
I_d	Direct axis grid current component
I_q	Quadrature axis grid current component
C_{DC}	DC link capacitor
V_{DC}	DC link voltage
I_{bat}	Battery current

only in one direction i.e., from grid to vehicle (G2V). However, a Bidirectional EV charger has the flexibility to operate in both vehicle to grid (V2G) and G2V modes [4]. Moreover, on the basis of connection, EV charger can be conductive or inductive [5]. In near future, it is anticipated that the huge number of EVs will be connected to distribution grid. A single EV may not have any significant impact on grid performance due to its relatively small power rating. However, large number of EVs may create fluctuations in grid voltage or current if they are charged in inappropriate or uncontrolled manner [6]. Moreover, the large number of EVs may be very helpful in meeting out the enhanced power demand in case of emergency. Even in case of blackout when the whole system is shutdown, the V2G mode of operation may be very helpful in restoring the system. The V2G mode may further enhance the power quality by supplying the reactive power (inductive or capacitive), filtering the current harmonics and voltage regulation to certain extent by using proper control. The application of EV charger for compensation of reactive power has been widely discussed in technical literature [7–9]. EV chargers can easily support the grid reactively when battery is fully charged or charging at slower rate. This results the reduction on investment of reactive power compensation devices [10].

The first stage AC-DC converter plays very crucial role in supporting grid reactively. Here, the EV charger controller is designed to performs two operations simultaneously i.e., it charges the battery pack as well as compensating the reactive power. Generally, a two-loop structure is used for controlling the first stage AC-DC converter. The grid current and voltage/power are tracked in inner and outer loop respectively. The grid current can be tracked either in DC quantity/ dq frame or AC quantity/ $\alpha\beta$ frame. However, while tracking current reference in DC quantity, $\alpha\beta$ to dq transformation is required which is also complex in case of single phase system [11]. Therefore, a controller is required which can track periodic signal to overcome from above mentioned issues.

Motivation

Generally, *PR* controller is utilized to track the periodic signals. They are much popular, especially in case of single phase system and extensively available in literature [12–13]. Moreover, the practical implementation of *PR* controller, especially while using fixed point arithmetic and at low frequency rate is not robust [14]. A conventional *PR* compensator requires supplementary integrator to reject the DC offset. This will increase the complete controller order and makes it a third order controller because *PR* alone is a second order and integrator is first order controller. Therefore, a controller is required which has all the functionalities of *PR* regulator with reduced order.

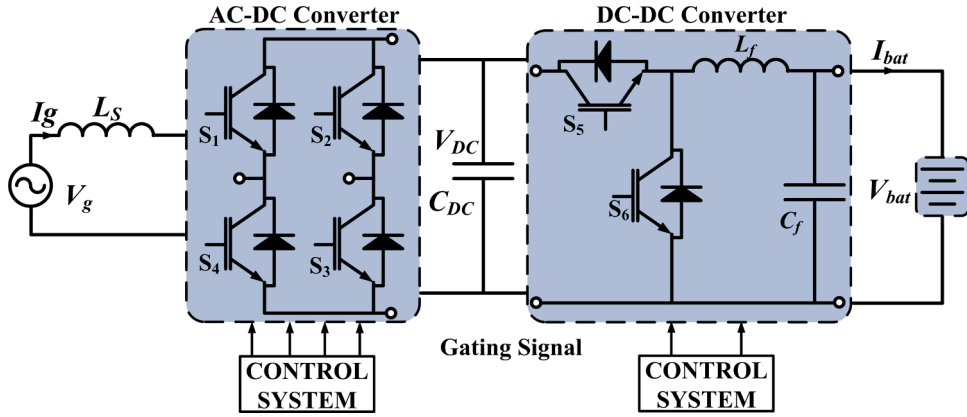


Fig. 1. On-board EV charger structure.

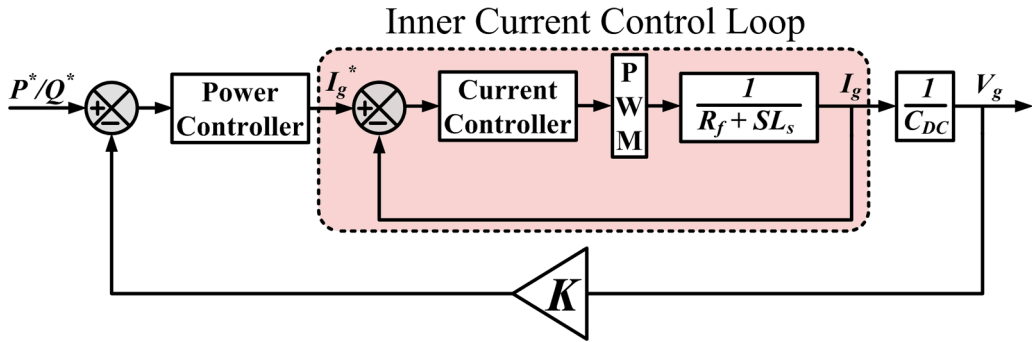


Fig. 2. Control Structure of AC-DC converter.

Literature review

Many controllers like repetitive controller (RC), proportional integral (PI), proportional resonant (PR), model predictive control (MPC) and hysteresis are used to track grid current in inner loop and widely discussed in literature [15–17]. Generally, PI controllers are used to track DC quantity as they are very popular for tracking it with zero steady state error. However, they are not very much suitable for inner current loop for aforementioned application as they have limited bandwidth and it is difficult to track command under varying AC input. MPC may also lead to high-power ripples and large distortion in grid current due to the error between reference vector and selected optimal voltage vector. The conventional PR controller may suffer from grid frequency variation, however this issue is overcome by use of quasi-PR (QPR) controller [18]. In [19], notch filter and passive damping with PR is presented to deal with harmonics of grid voltage. In [20], multiple PR controllers are implemented for single phase converter control to reject the harmonics, however the use of multiple PR controllers makes it complex and difficult to tune. Furthermore, the oscillating switching frequency may have resonance issues while using hysteresis control [21]. Some control techniques based on artificial intelligence have also been discussed in [22]. These control techniques suffer from the deficiency of computational anxiety for online adaptive tuning.

Contribution and paper organization

Hence, in proposed work a reduced order integrated controller-plant dynamics (ICPD) based plant integrated proportional integrating (PIPI) controller is designed to track the reference periodic signal for an EV charger. The PIPI controller has ability to accomplish all the purposes of conventional PR controller with reduced order. The proposed approach resolved the issues that arise during practical implementation of conventional PR controller and offers robust performance. Therefore, in this paper a controller has been designed for on-board EV charger which can track active and reactive power commands simultaneously. For this, two PI regulators are used to track active-reactive power commands in outer loop of AC-DC converter control. Furthermore, proposed PIPI regulator is used in inner grid current control loop to track periodic reference. Since, the main concentration of paper is to develop a simple and robust periodic reference tracker, therefore generalized control of DC-DC converter has been utilized which consists two regular PI controllers. Moreover, the overall performance of two-stage on-board EV charger controller has been tested in eight modes which include charging and compensation of reactive power if requested by utility. The complete system has been designed in MATLAB/Simulink toolbox and the claims are confirmed by scaled down 350 VA hardware prototype in research laboratory using

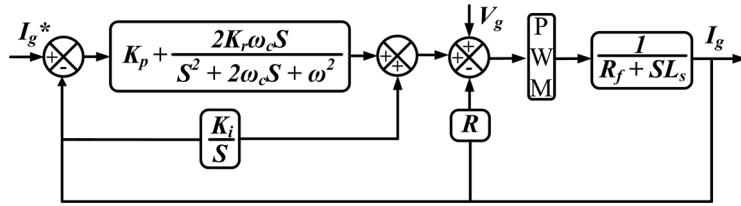


Fig. 3. Inner current control loop with conventional PR regulator.

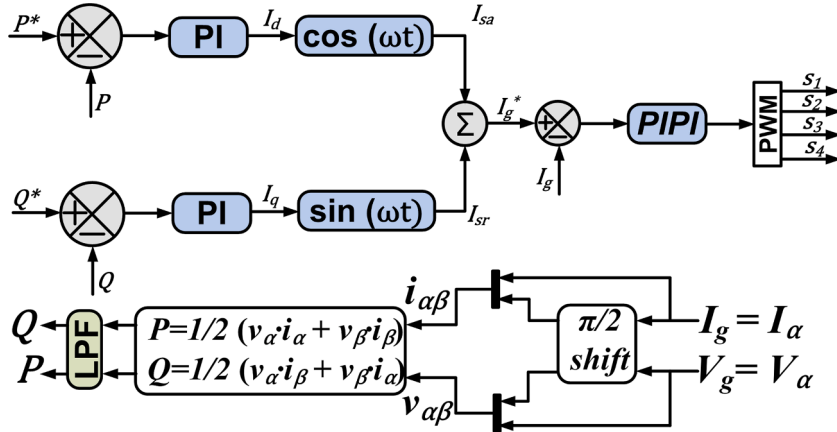


Fig. 4. Control of AC-DC Converter.

OPAL-RT (OP-4510).

The structure of the paper is as follows. The problem statement is presented in section II. The section III presented the on-board EV charger controller design. The simulation and experimental results are described in sections IV and V respectively. The result discussion and achievements are discussed in section VI. Finally, the conclusion of paper is summarized in section VII.

Problem statement

Fig. 1 shows the structure of two-stage on-board EV charger. The charger consists of single phase AC-DC converter at first stage and a bidirectional buck-boost converter at second stage. The insulated gate bipolar transistor (IGBT) switches are used in development of both converters. The DC-DC converter works in buck and boost mode in case of charging and discharging of battery respectively.

The control structure of first stage AC-DC converter is shown in Fig. 2. The control structure comprises of outer and inner loop. In present case, outer loop is used to track the active-reactive power commands and periodic grid current is tracked in inner loop. where K is the function of grid side current [23].

The conventional PR with integral (I) regulator is generally used for tracking the periodic grid current reference in inner loop as shown in Fig. 3. Sometimes, it requires additional I to eliminate the DC offset. The transfer function of PR controller alone is $K_p + \frac{2K_r\omega_c S}{S^2 + 2\omega_c S + \omega^2}$, which is a second order transfer function and integrator alone is single order transfer function i.e., K_i/S . which makes the total PR-I controller a third order system. Where, K_p , K_r and K_i are gains of controller. Moreover, it consists of internal virtual resistance (R) or damper loop and feed-forward term V_g , which is used for soft start and initial synchronism.

The above discussed PR-I controller has following downsides: i) It increases the order of controller which result increases the complexity, ii) for lower sampling frequencies, its digital implementation is not robust, iii) it requires auxiliary integrator to eliminate DC offset, iv) for fixed-point arithmetic, its digital implementation is not robust [24]. However, the issue of low sampling is overcome in [25] by expense of computations of trigonometric functions. This can be achieved by demanding of accurate trigonometric functions. Moreover, explicit design formulas are also derived for robust implementation of PR controller.

On-Board EV charger controller design

A. Control of AC-DC converter

Fig. 4 shows the control architecture for first stage AC-DC converter of EV charger. The proposed charger controller has tracked active (P) and reactive (Q) power in outer loop by PI regulators. The P and Q are comfortably tracked by regular PI compensator as they are constant in nature. The outer loop generates reference for inner current loop and here it is periodic in nature. Further, the periodic grid current reference is tracked by proposed PIPI regulator.

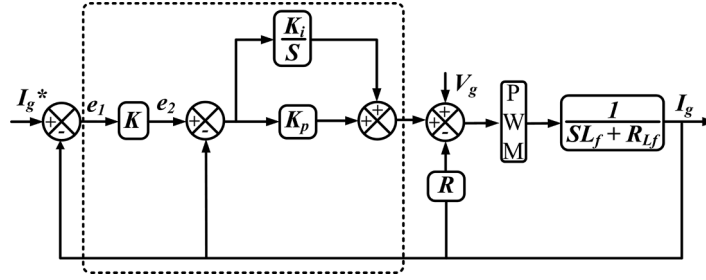


Fig. 5. Proposed PIPI controller.

The voltage and current of grid side are as follows;

$$V_g = V_a = V_m \cos \omega t \quad (1)$$

$$I_g = I_a = I_m \cos(\omega t - \varphi) \quad (2)$$

Here, V_m , I_m and φ are maximum value of grid voltage, current and phase angle between them, respectively. The β component can be find as,

$$V_\beta = V_m \cos(\omega t + \pi/2) \quad (3)$$

$$I_\beta = I_m \cos(\omega t - \varphi + \pi/2) \quad (4)$$

The active (P) and reactive (Q) power of grid side are found as,

$$P = \frac{1}{2} (V_a \cdot I_a + V_\beta \cdot I_\beta) \quad (5)$$

$$Q = \frac{1}{2} (V_a \cdot I_\beta + V_\beta \cdot I_a) \quad (6)$$

To produce the reference of grid current for inner control loop, first active (I_d) and reactive (I_q) current component found as,

$$I_d = K_{p1}(P^* - P) + K_{i1} \int (P^* - P) dt \quad (7)$$

$$I_q = K_{p2}(Q^* - Q) + K_{i2} \int (Q^* - Q) dt \quad (8)$$

Further, the periodic form of active (I_{sa}) and reactive (I_{sr}) current are found as,

$$\begin{aligned} I_{sa} &= I_d \cdot \cos \omega t \\ &= \left[K_{p1}(V_{DC}^* - V_{DC}) + K_{i1} \int (V_{DC}^* - V_{DC}) dt \right] \cdot \cos \omega t \end{aligned} \quad (9)$$

$$\begin{aligned} I_{sr} &= I_q \cdot \sin \omega t \\ &= \left[K_{p2}(Q^* - Q) + K_{i2} \int (Q^* - Q) dt \right] \cdot \sin \omega t \end{aligned} \quad (10)$$

Further, reference current is generated by adding both active and modified reactive current component.

$$I_g^* = I_{sa} + I_{sr} \quad (11)$$

B. Inner Current Controller Design.

To track the periodic reference current, a reduced order PIPI regulator is proposed here and its structure is shown in Fig. 5. It contains a proportional (P) and PI controller. The P controller having gain K where proportional and integral constant of PI controller are K_p and K_i , respectively. The open loop transfer function from e_1 to I_g is as follows,

$$\frac{I_g(s)}{e_1(s)} = K \frac{K_p s + K_i}{L s^2 + (K_p + R) s + K_i} \quad (12)$$

By choosing the $K_p = -R$ and $K_i = L\omega^2$ we get,

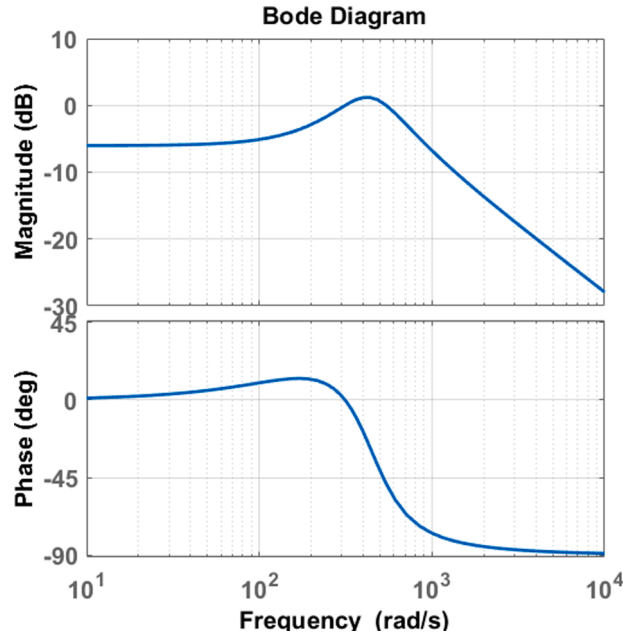


Fig. 6. Bode plot of closed loop transfer function.

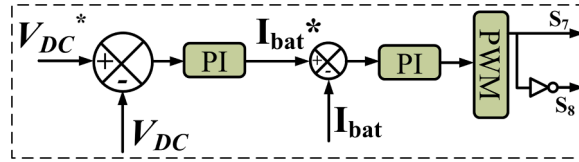


Fig. 7. DC-DC Buck Converter and its Control.

$$\frac{I_g(S)}{e_1(S)} = \frac{K L \omega^2 - RS}{L S^2 + \omega^2} \quad (13)$$

Where, ω is fundamental grid side frequency in rad/sec. The closed loop transfer function is,

$$\frac{I_g(S)}{I_g^*(S)} = \frac{K}{L} \frac{L \omega^2 - RS}{S^2 - \frac{RK}{L} + (1 + K) \omega^2} \quad (14)$$

The closed loop transfer function stability depends on value of R and K and their range is less and greater than zero for R and K , respectively. For this, the ω_n/ω ratio should be greater than 1, here ω_n is natural frequency. The parameters K and R are calculated by $(\omega_n/\omega)^2 - 1$ and $-2\zeta L \omega_n/K$, respectively. By putting these values, the closed and open loop transfer functions are,

$$\frac{I_g(S)}{E_1(S)} = \frac{2\zeta \omega_n S + \omega_n^2 - \omega^2}{S^2 + \omega^2} \quad (15)$$

$$\frac{I_g(S)}{I_g^*(S)} = \frac{2\zeta \omega_n S + \omega_n^2 - \omega^2}{S^2 + 2\zeta \omega_n S + \omega_n^2} \quad (16)$$

In present case, ω_n/ω is selected $\sqrt{2}$ and ζ is 0.45 leads to R is $-1.27L\omega$ and K is 1. For 50 Hz system frequency, the locations of closed loop poles are $-200 \pm j397$ and zero is -247 . From the frequency response shown in Fig. 6, the phase margin is 131° at crossover frequency 544 rad/sec.

From the analysis, it can be observed that proposed controller can track periodic reference signal. This is achieved due to the existence of frequency dependent $S^2 + \omega^2$ term in denominator of eq. (13). At fundamental frequency i.e., $S = j\omega_n$, the closed loop transfer function of eq. (15) tends to unity. This shows that the second order integrator term guarantees help to attain zero steady state error. Moreover, there is no DC offset in output current because if any DC current (i_d) flows through inductor, then the steady state error will be $(-i_d)$ and $e_2 = (K + 1) \cdot (-i_d)$. This is contradictory as input of PI controller cannot contain a DC term. Therefore, the proposed PIPI controller can track periodic reference signal without any DC offset while having the order of first.

Table 1
SYSTEM PARAMETER.

Parameter	Symbol	Simulation	Experimental
Charger Rating	S	6.6 KVA	350 VA
Grid Voltage	V_g	230 V	60 V
Grid Frequency	f	50 Hz	50 Hz
AC Inductors	L_s	4 mH	1 mH
DC link Capacitor	C_{DC}	1100 μ F	330 μ F
DC link Voltage	V_{DC}	400 V	150 V
Battery Voltage	V_{bat}	350 V	96 V

Table 2
SIMULATION SCENARIO.

Mode	P (kW)	Q (kVAR)	S (kVA)	Power Factor	Time (sec)
1	6.6	0	6.6	Unity	0–1.5
2	−6.6	0	6.6	−1	1.5–3
3	0	6.6	6.6	0	3–4.5
4	0	−6.6	6.6	0	4.5–6
5	4	5.25	6.6	0.6 (lag)	6–7.5
6	5	−4.31	6.6	0.75 (lead)	7.5–9
7	−5.5	3.65	6.6	0.83 (lag)	9–10.5
8	−2.5	−6.11	6.6	0.38 (lead)	10.5–12

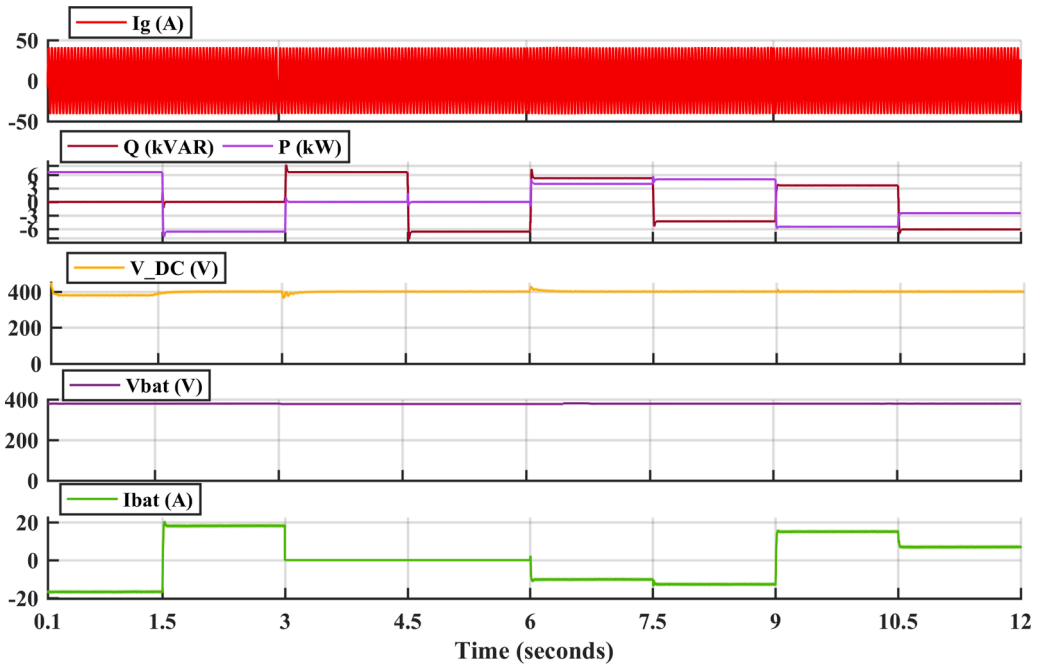


Fig. 8. Simulation outcomes of grid side current (I_g), active (P)-reactive (Q) power, DC link voltage (V_{DC}), battery voltage (V_{bat}) and battery current (I_{bat}) during all modes.

B. Control of DC-DC converter

Fig. 7 shows the controller of second stage DC-DC converter of EV charger. This is utilized to regulate the DC link voltage (V_{DC}) and battery current (I_{bat}). This also contains two control loops in which outer one is for DC link voltage control and inner one for battery current. Both the quantities are regulated by PI controller to reduce the complexity of overall EV charger controller and pulses are generated by well-known pulse width modulation technique.

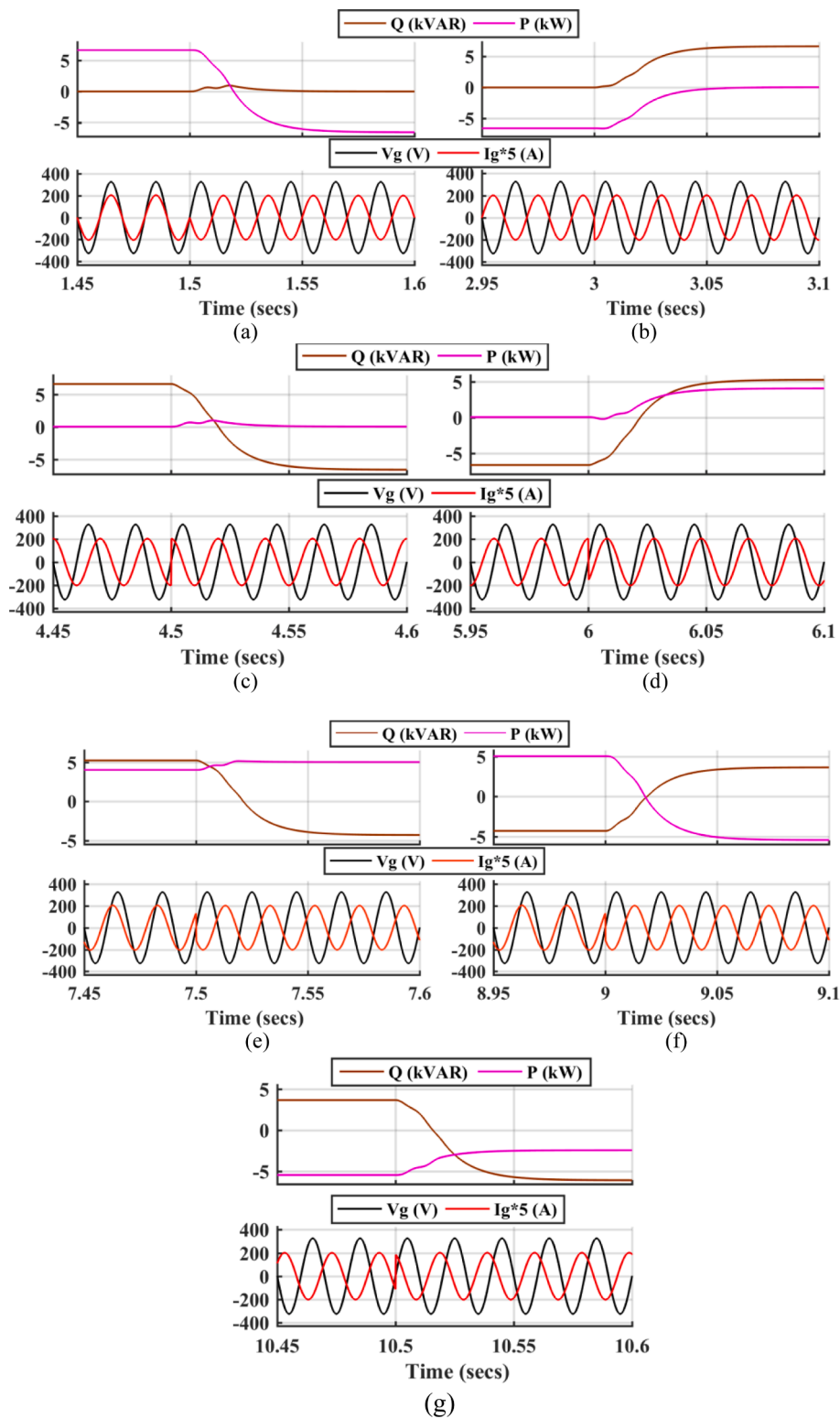


Fig. 9. Change from (a) mode 1–2, (b) mode 2–3, (c) mode 3–4, (d) mode 4–5, (e) mode 5–6, (f) mode 6–7, (g) mode 7–8.

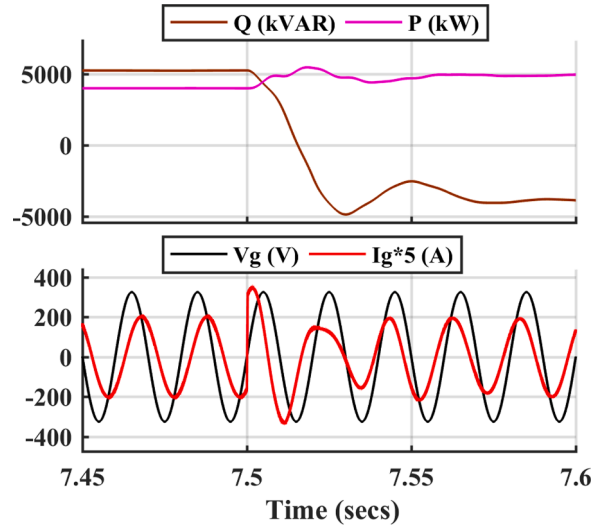
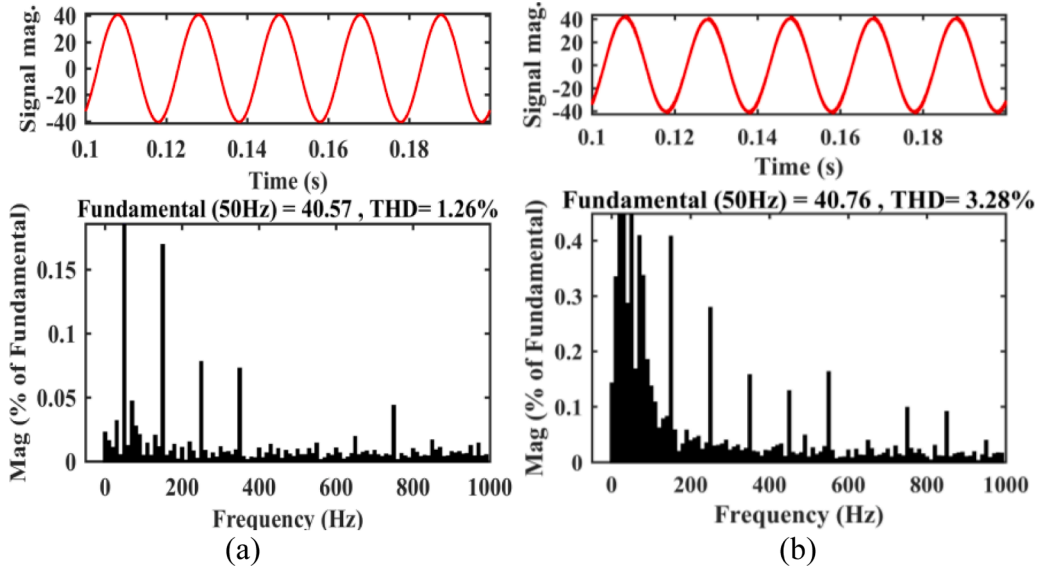


Fig. 10. Transition from mode 5-6 with conventional PR regulator.

Fig. 11. THD in grid current during mode 5, (a) Proposed *PIPI* controller and (b) conventional *PR* controller.

Simulation results

MATLAB/Simulink toolbox has been used to design above discussed 6.6 kVA on-board two stage EV charger and its proposed controller. Table 1 listed all the simulation and experimental parameters, here AC-DC converter is connected across the single phase utility grid of 230 V, 50 Hz. The reference of DC-link voltage and nominal voltage of battery pack are taken as 400 V and 350 V respectively, where battery SOC is taken as 50% in order to observe charging/discharging profile. The actual voltage of EV battery pack is greater than nominal voltage and it is directly proportional to SOC.

To confirm the behavior of proposed EV charger controller, a simulation scenario having eight operating modes has been created in MATLAB. These modes are related to different values of active-reactive power, where positive active power means utility grid is charging the battery pack. Each mode is simulated for 1.5 s in which positive and negative reactive power means compensation of inductive and capacitive reactive power, respectively. For optimal utilization of EV charger, it is suggested to utilize complete charger's rating. Therefore, during all the working modes, complete charger's rating is utilized. The mode 1 and 2 are related to battery charging/discharging respectively without any reactive power compensation. Similarly, only compensation of inductive/capacitive reactive power is shown by EV charger in mode 3 and 4 respectively without charging/discharging of battery pack. It is a case of fully charged battery pack and charger can be utilized in compensation of reactive demand of local load within its limit if

Table 3
THD with *PIPI* and *PR*.

Mode	THD (%) With <i>PIPI</i>	THD (%) With <i>PR</i>	Mode	THD (%) With <i>PIPI</i>	THD (%) With <i>PR</i>
1	1.25	3.27	5	1.26	3.28
2	1.23	3.25	6	1.57	3.36
3	1.23	3.21	7	1.53	3.30
4	1.09	3.31	8	1.35	3.26

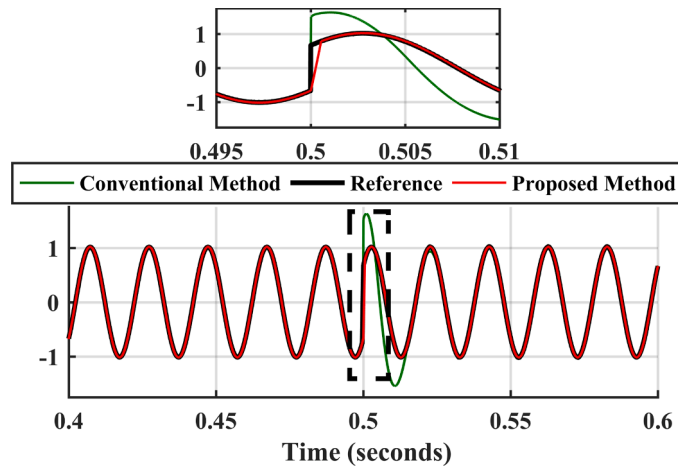


Fig. 12. Comparison between proposed and conventional method.

requested by utility grid. Further, both active as well as reactive power commands are given simultaneously to EV charger controller in last four modes. During these modes, the rate of power exchange with battery is slower rate and rest of charger's capacity is used for reactive power compensation. [Table 2](#)

[Fig. 8](#) Shows the Simulation outcomes of grid current (I_g), active (P)-reactive (Q) power, DC link voltage (V_{DC}), battery voltage (V_{bat}) and battery current (I_{bat}) during all modes. Since, EV charger is operating at constant power in all working modes, the grid is constant i. e., $6600/230 = 29$ A (rms). As, the reference of DC link voltage is taken as 400 V, it is maintained at reference value with certain variation at the time of transition. Due to the 350 V nominal voltage with 50% of SOC, the actual battery voltage is around 380 V. The battery side current directly proportional to active power command and it is negative and positive during charging and discharging. [Fig. 9\(a\)](#) and (b) shows the transition between mode 1–2 and mode 2–3, respectively. Since, only battery charging operation has been performed in mode-1, the grid current and voltage are in same phase. Similarly, in mode-2, the grid current is in opposite phase with voltage as power is taken from battery pack. During mode 3 and 4, grid current is lag and lead the voltage by exactly 90° , respectively as shown in [Fig. 9 \(c\)](#) and (d). In mode 5, the 4 kW of power is transmitted to battery pack and remaining capacity of EV charger is utilized in compensation of inductive reactive power of 5.25 kVAR, correspondingly the phase difference between voltage and current is 53° lagging as shown in [Fig. 9 \(e\)](#). Similarly, in mode 6, the 5 kW of power is transmitted to battery pack while compensating of -4.31 kVAR reactive power and now current is leading by 41° as depicted in [Fig 9 \(F\)](#). The change from mode 7 to 8 is depicted in [Fig. 9 \(g\)](#). Here, EV charger is taking 5.5 and 2.5 kW of power during mode 7 and 8 respectively and phase difference between voltage and current is 214° and 248° as shown in [Fig. 9 \(g\)](#).

From the [Fig. 9](#), it is concluded that the behavior of proposed inner current controller is very fast as the grid current settles quickly. Moreover, it has exquisite transient and steady-state performance during all possible working modes of EV charger.

[Fig. 10](#) shows the transition from mode 5 to 6 with conventional *RP* controller. The mode 5 and 6 are mostly used modes, during these modes battery is charging at slower rate and remaining capacity of EV charger is used in compensation of reactive power. From [Fig. 10](#), it can be observed that the conventional *PR* regulator has excellent steady-state response but it takes less than two cycles to settle down with some overshoot in comparison with proposed *PIPI* regulator as shown in [Fig. 9 \(e\)](#).

[Fig. 11\(a\)](#) and (b) shows the total harmonic distortion (THD) in grid side current with *PIPI* and *PR* regulator respectively during mode 5. The THD in grid side current also less than the allowable IEEE limit of 5% in conventional *PR* regulator. However, the proposed *PIPI* regulator provides better THD in grid current in comparison with conventional *PR* regulator. Furthermore, the THDs in grid current with *PIPI* and *PR* regulator during all working modes are listed in [Table 3](#).

[Fig. 12](#) shows the dynamic performance of proposed *PIPI* and *PR* inner loop current controller during transition of sinusoidal reference. It is found that the proposed *PIPI* controller is faster and more robust than conventional *PR* method.

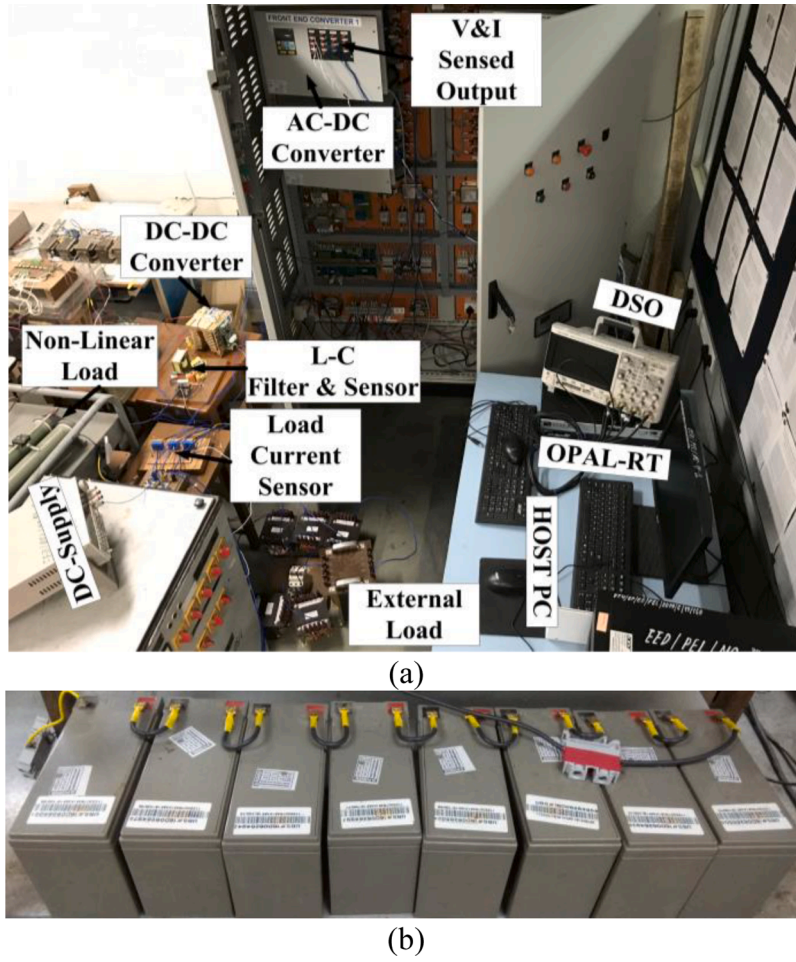


Fig. 13. (a) Experimental setup and (b) 96 V battery pack.

Table 4
Hardware Scenario.

Mode	Real Power (W)	Reactive Power (VAR)	Apparent Power (VA)	Power Factor
1	350	0	350	Unity
2	−350	0	350	−1
3	0	350	350	0
4	0	−350	350	0
5	300	180.2	350	0.85 (lag)
6	250	−245	350	0.71 (lead)
7	−200	287.2	350	0.57 (lag)
8	−325	−130	350	0.92 (lead)

Experimental results

Fig. 13 shows the laboratory 350 VA experimental setup using OPAL-RT (OP 4510) and system parameters are listed in Table 1. Both converters are made up from semikron insulated gate bipolar transistor (IGBT) legs. A battery pack of 96 V is connected across the second stage DC-DC converter. To validate the performance of proposed EV charger controller in real time, a scenario same as simulation is created as listed in Table 4.

The transient and steady state performance of proposed EV charger controller is shown in Fig. 14. The Fig. 14(a) is associated with transition from mode 1 to 2. During this, the grid side voltage and current are in same phase during mode 1 and battery current is approximately −3.5 amps. In mode 2, the charger is taking 350 W of power from battery, therefore battery current is positive i.e., 3.5 A and phase angle is 180°. In Fig. 14 (b) transition from mode 2–3 has been shown, here the charger switches the mode from active to reactive power operation and now the phase angle changed to 90° lagging. In mode 4, the charger is switched to compensation of

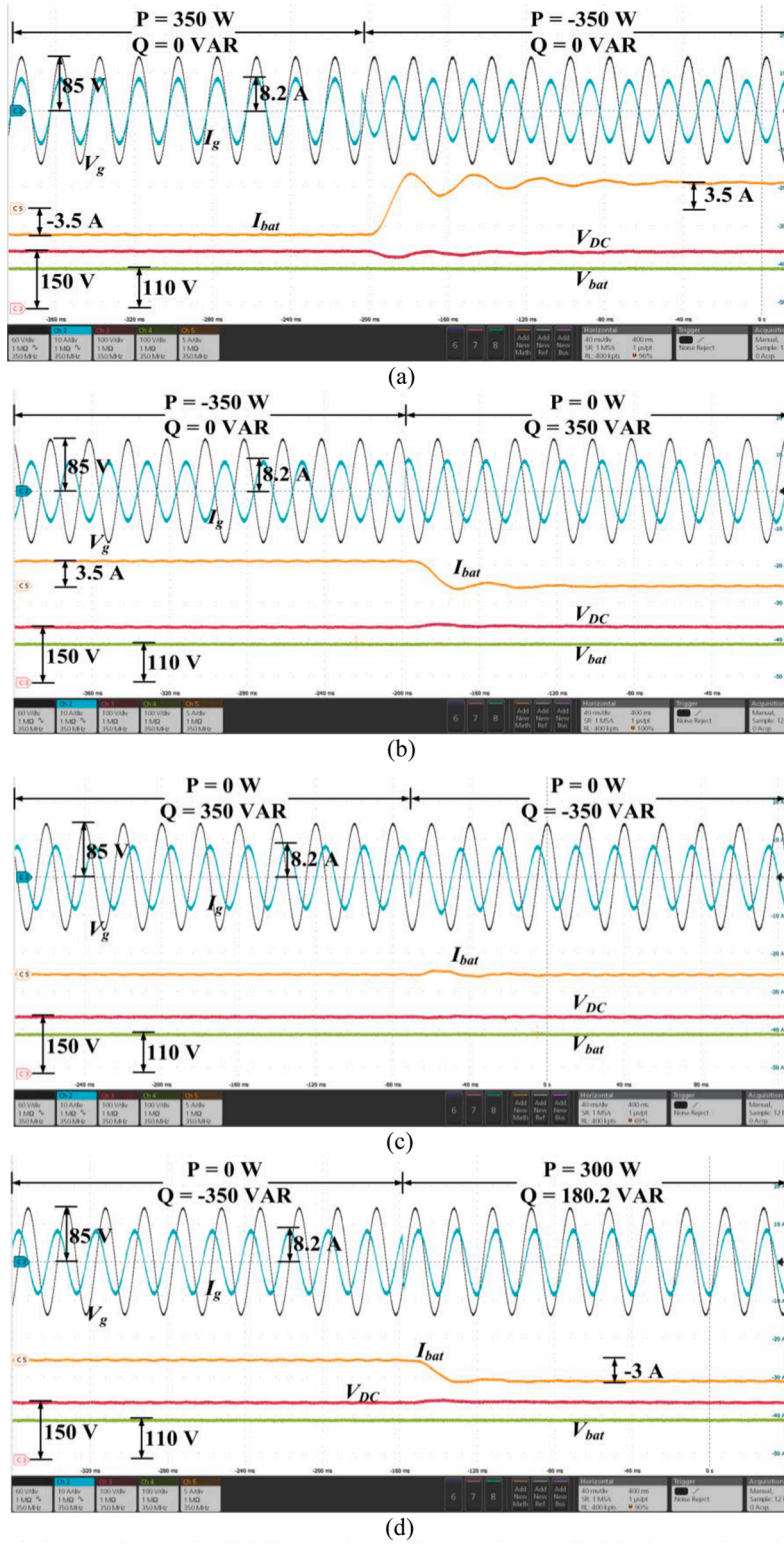
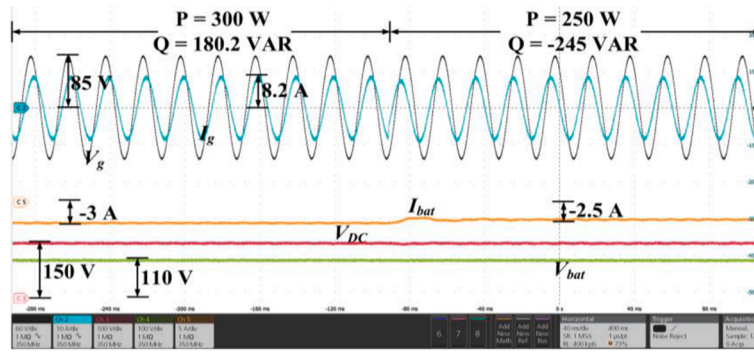
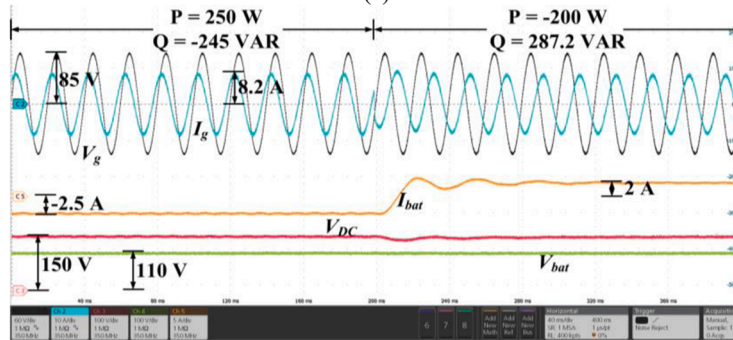


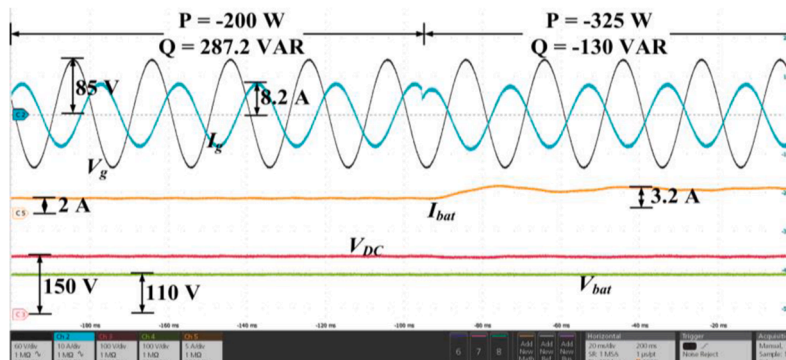
Fig. 14. Transition from (a) mode 1 to 2, (b) mode 2 to 3, (c) mode 3 to 4, (d) mode 4 to 5, (e) mode 5 to 6, (f) mode 6 to 7 and (g) mode 7 to 8.



(e)



(f)



(g)

Fig. 14. (continued).

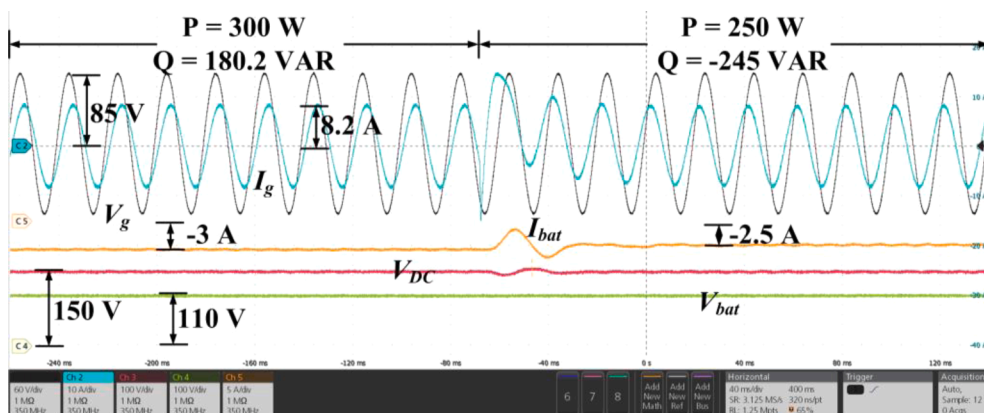


Fig. 15. Transition from mode 5 to 6 with conventional PR regulator.

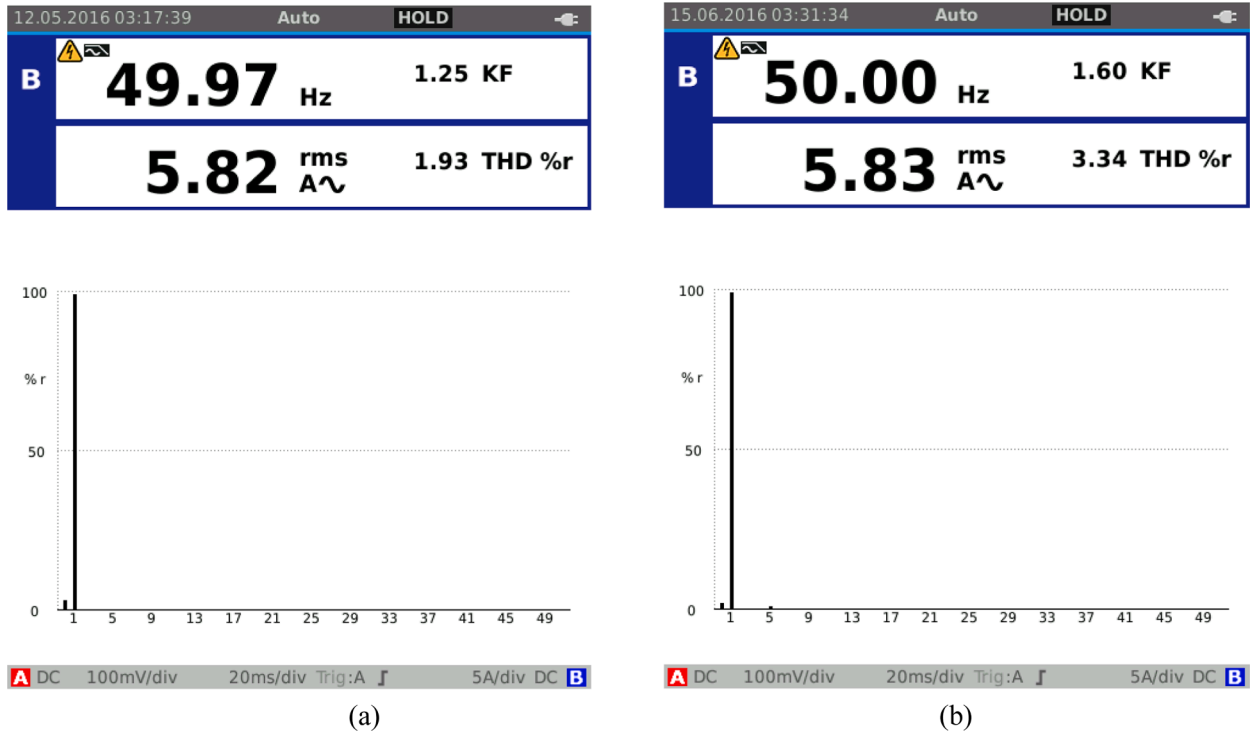


Fig. 16. THD in grid current during mode 5 (a) with proposed *PIPI* and (b) with conventional *PR* regulator.

Table 5
THD with *PIPI* and *PR*.

Mode	THD (%) With <i>PIPI</i>	THD (%) With <i>PR</i>	Mode	THD (%) With <i>PIPI</i>	THD (%) With <i>PR</i>
1	1.89	3.11	5	1.93	3.34
2	1.98	3.24	6	2.02	3.15
3	2.03	3.21	7	1.88	3.16
4	2.16	3.37	8	1.99	3.29

capacitive reactive power operation and now grid current is leading by 90° as shown in Fig. 14 (c). During this, the battery current is zero. In mode 5 and 6, the rate of active power transferred to battery is 300 and 250 W and remaining capacity of EV charger is utilized in compensation of inductive and capacitive reactive power of 180.2 and 245 VA respectively. Accordingly, the battery current is approximately 3 and 2.5 A and phase difference between grid side voltage and current is 32° lagging and 45° leading in mode 5 and 6 respectively as shown in Fig. 14 (e). Similarly, 235° lagging and 203° leading phase angle can be seen during mode 7 and 8 as the charger is taking 200 and 325 W of power from battery while compensation of 287.2 and -130 VAR of reactive power respectively as depicted in Fig. 14 (g). Moreover, the transitions from mode 4 to 5 and 6 to 7 are shown in Fig. 14 (d) and (f). Further, the DC-link voltage level is sustained at 150 V throughout all the working modes and transition does not affect it too much.

Fig. 15 shows the transition from mode 5 to 6 with conventional *PR* regulator. It is observed that at the time of transition, grid side current takes some time to settle down and an overshoot is also there in comparison with proposed *PIPI* controller as shown in Fig. 14 (e).

Fig. 16(a) and (b) depicts the real time outcome of THD in grid current with *PIPI* and *PR* regulator respectively during mode 5. It is found that the THD in grid current is better in case of *PIPI* regulator in comparison with *PR* regulator. Moreover, the THDs in grid current with *PIPI* and *PR* regulator during all working modes are listed in Table 5.

Result discussion and achievement

The proposed controller has been designed and implemented on-board EV charger. The performance of controller has been tested on all possible eight working modes and dynamic behavior have also been tested in simulation and real time experimental setup. From the results, it has been observed that the transient and steady state behavior of *PIPI* regulator is very fast and settles quickly within one grid cycle. Furthermore, the behavior of *PIPI* regulator has been compared with conventional *PR* regulator and it has been found that

proposed *PIPI* controller is faster and more robust than conventional *PR* method.

The main achievement of proposed work is to design a lesser order periodic controller which has all the functionalities of well-known *PR* regulator. The proposed controller is simple in design, robust, faster and first order system.

Conclusion

A reduced order *PIPI* controller which has the functionalities of *PR* controller has been designed and implemented on EV charger. Here, a third order conventional *PR* controller is replaced by first order *PIPI* controller. The proposed *PIPI* controller has lesser complexity, easy to design, easy to real time implementation and can track any periodic signal. This proposed technique has been implemented for controlling the first stage of on-board EV charger in inner loop. Furthermore, the performance of proposed EV charger controller has been tested by various combination of active-reactive power command in MATLAB as well as real time. From the simulation and real time outcomes, it is found that the proposed *PIPI* regulator helps in design of EV charger controller as it is simple with first order system. Moreover, the outcomes show quick and smooth response by tracking reference active-reactive power, grid current command in less than one grid cycle and provides better THD in grid side current.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mukhtiar Singh reports was provided by Delhi Technological University. Mukhtiar Singh reports a relationship with Delhi Technological University that includes: employment.

Data availability

No data was used for the research described in the article.

References

- [1] Borray Andrés Felipe Cortés, Merino Julia, Torres Esther, Mazón Javier. A review of the population-based and individual-based approaches for electric vehicles in network energy studies. *Electric Power Systems Res* 2020;189.
- [2] Bhaskar KBR, Prasanth A, Saranya P. An energy-efficient blockchain approach for secure communication in IoT-enabled electric vehicles. *Int J Commun Syst* 2022;35(11):e5189. <https://doi.org/10.1002/dac.5189>.
- [3] Sandeep V, Shastri Suchitra, Sardar Arghya, Salkuti Surender Reddy. Modeling of battery pack sizing for electric vehicles. *Int J Power Electronics Drive Syst* 2020;11(4):1987–94. <http://doi.org/10.11591/ijpeds.v11.i4.pp1987-1994>.
- [4] Lehtola TA, Zahedi A. Electric vehicle battery cell cycle aging in vehicle to grid operations: a review. *IEEE J Emerg Sel Top Power Electron* Feb. 2021;9(1):423–37. <https://doi.org/10.1109/JESTPE.2019.2959276>.
- [5] Vuddanti Sandeep, M N Shivanand, Salkuti Surender Reddy. Design of a one kilowatt wireless charging system for electric vehicle in line with Bharath EV standards. *Int J Emerg Electr Power Syst* 2021;22(3):255–67. <https://doi.org/10.1515/ijeeps-2020-0178>.
- [6] Chalia S, Seth AK, Singh M. Electric vehicle charging standards in India and safety consideration. In: Proceedings of the 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON); 2021. p. 1–6. <https://doi.org/10.1109/UPCON52273.2021.9667649>.
- [7] Gupta M, Seth AK, Singh M. Grid tied photovoltaic based electric vehicle charging infrastructure. In: Proceedings of the 2022 S International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT); 2022. p. 1–6. <https://doi.org/10.1109/ICAECT54875.2022.9808038>.
- [8] Sou W-K, et al. A deadbeat current controller of LC-hybrid active power filter for power quality improvement. *IEEE J Emerg Sel Top Power Electron* Dec. 2020;8(4):3891–905. <https://doi.org/10.1109/JESTPE.2019.2936397>.
- [9] Phan D, Lee H. Interlinking converter to improve power quality in hybrid AC–DC microgrids with nonlinear loads. *IEEE J Emerg Sel Top Power Electron* Sept. 2019;7(3):1959–68. <https://doi.org/10.1109/JESTPE.2018.2870741>.
- [10] Seth AK, Singh M. Second-order ripple minimization in single-phase single-stage onboard PEV charger. *IEEE Transactions on Transportation Electrification* Sept. 2021;7(3):1186–95. <https://doi.org/10.1109/TTE.2021.3049559>.
- [11] Quan Xiangjun, Dou Xiaobo, Wu Zaijun, Hu Minqiang, Yuan Jian. Harmonic voltage resonant compensation control of a three-phase inverter for battery energy storage systems applied in isolated microgrid. *Electric Power Systems Res* 2016;131.
- [12] Seth AK, Singh M. Resonant controller of single-stage off-board EV charger in G2V and V2G modes. *IET Power Electronics* 2020;13(5):1086–92. 8-4-.
- [13] Seifi K, Moallem M. An adaptive pr controller for synchronizing grid-connected inverters. *IEEE Trans Ind Electron* 2018;66(3):2034–43.
- [14] Gholizade-Narm H, Khajehoddin SA, Karimi-Ghartemani M. Reduced-order controllers using integrated controller-plant dynamics approach for grid-connected inverters. *IEEE Trans Ind Electron* Aug. 2021;68(8):7444–53. <https://doi.org/10.1109/TIE.2020.3007119>.
- [15] Monter Ana Rodríguez, Bueno Emilio J, García-Cerrada Aurelio, Rodríguez Francisco J, Sánchez Francisco M. Detailed analysis of the implementation of frequency-adaptive resonant and repetitive current controllers for grid-connected converters. *Electric Power Systems Research* 2014;116.
- [16] Kesler M, Ozdemir E. Synchronous-reference-frame-based control method for upqc under unbalanced and distorted load conditions. *IEEE Trans Ind Electron* Sept. 2011;58(9):3967–75. <https://doi.org/10.1109/TIE.2010.2100330>.
- [17] Suryakant MSreejeth, Singh M, Seth AK. Minimization of torque ripples in PMSM drive using PI- resonant controller-based model predictive control. *Electr Eng* 2022. <https://doi.org/10.1007/s00202-022-01660-y>.
- [18] An Q, Zhang J, An Q, Shamekov A. Quasi-proportional-resonant controller based adaptive position observer for sensorless control of pmsm drives under low carrier ratio. *IEEE Trans Ind Electron* 2019;67(4):2564–73.
- [19] Liu Y, Wu W, He Y, Lin Z, Blaabjerg F, Chung HS. An efficient and robust hybrid damper for LCL- or LLCL-based grid-tied inverter with strong grid-side harmonic voltage effect rejection. *IEEE Trans Ind Electron* 2016;63(2):926–36.
- [20] Khajehoddin SA, Karimi-Ghartemani M, Ebrahimi M. Optimal and systematic design of current controller for grid-connected inverters. *IEEE J Emerg Sel Top Power Electron* 2017;6(2):812–24.
- [21] R.M. Pindoriya, A. Yadav, B.S. Rajpurohit and R. Kumar, "A Novel Application of Random Hysteresis Current Control: acoustic Noise and Vibration Reduction of a Permanent Magnet Synchronous Motor Drive," in IEEE industry applications magazine, 2022, doi:10.1109/MIAS.2022.3160986.

- [22] Y. Gao et al., "Inverse application of artificial intelligence for the control of power converters," in IEEE transactions on power electronics, 2022, doi:[10.1109/TPEL.2022.3209093](https://doi.org/10.1109/TPEL.2022.3209093).
- [23] Sylvain Lechat Sanjuan, "Voltage Oriented Control of Three-Phase Boost PWM Converters Design, simulation and implementation of a 3-phase boost battery charger" Master of Science Thesis, CHALMERS UNIVERSITY OF TECHNOLOGY, 2010.
- [24] Bottrell N, Green TC. Comparison of current-limiting strategies during fault ride-through of inverters to prevent latch-up and wind-up. IEEE Trans Power Electron July 2014;29(7):3786–97. <https://doi.org/10.1109/TPEL.2013.2279162>.
- [25] Richter SA, De Doncker RW. Digital proportional-resonant (PR) control with anti-windup applied to a voltage-source inverter. In: Proceedings of the 2011 14th European Conference on Power Electronics and Applications; 2011. p. 1–10.

Aakash Kumar Seth received the B.Tech. in EEE from the BPIT, New Delhi, India, in 2013, and the M.Tech. in ED&C from the IET, Lucknow, in 2017. He has completed his Ph.D. in EE with DTU, New Delhi. Currently, he is working as a lecturer in government polytechnic, Uttar Pradesh. His-research interests include EV, power quality and microgrid.

Mukhtiar Singh received the B.Tech. and M.Tech. degrees in EE from NIT, Kurukshetra, India, in 1999 and 2001, respectively, and the Ph.D. degree from Ecole de Technologie Supérieure, University of Quebec, Montreal, Canada, in 2010. He is currently working as a Professor in Department of EE, DTU, Delhi, India. His-research interests include power electronics, power quality and renewable energy.

Polarization Reversal of Oblique Electromagnetic Wave in Collisional Beam-Hydrogen Plasma

Rajesh Gupta¹, Ruby Gupta², and Suresh C. Sharma^{1, *}

Abstract—Energetic ion or electron beams cause plasma instabilities. Depending on plasma and the beam parameters, an ion beam leads to change in the dispersion relation of Alfvén waves on interacting with magnetoplasmas as it can efficiently transfer its energy to the plasma. We have derived dispersion relation and the growth rates for oblique shear Alfvén wave in hydrogen plasma. The particles of the beam interact with the Shear Alfvén waves only when they counter-propagate each other and destabilize left-hand polarized mode for parallel waves and left-hand as well as right-hand polarized modes for oblique waves, via fast cyclotron interaction. The collisions between beam ions and plasma components affect the growth rate and the frequency of generated Alfvén waves, differently for right-hand (RH) and left-hand (LH) polarized oblique Alfvén modes. For $(\omega + k_z v_{bo} > \omega_{bc})$, the most unstable mode is the LH polarized oblique Alfvén mode, and it is the RH polarized oblique Alfvén mode for $(\omega + k_z v_{bo} < \omega_{bc})$, which shows a polarization reversal after resonance condition. Numerical results indicate that the growth rates increase with increase in angle of propagation. The maximum growth rate values in the presence or absence of beam increase due to obliquity of wave.

1. INTRODUCTION

Alfvén waves are a type of magneto-hydro-dynamic (MHD) waves which are low frequency waves below the ion cyclotron frequency travelling in a magnetized conducting fluid like plasma existing in space. These waves transport electromagnetic energy and communicate information concerning changes in plasma currents and magnetic field topology. At low frequencies, there are two different modes of electromagnetic propagation i) a compressional wave in which magnetic field strength and density change and ii) a shear wave in which only the direction of magnetic field varies. In the present paper, we are concerned with shear Alfvén wave only. The shear Alfvén wave propagates with the wave magnetic field vector perpendicular to background field. Gekelman et al. [1] have studied various properties related to shear Alfvén waves.

The shear Alfvén wave is nearly incompressible and hence, more readily excitable by either external perturbations, e.g., solar wind, antenna or intrinsic collective instabilities (Chen and Zonca [2]). In a cylindrical column, the dispersion of shear wave was determined by Jephcott and Stocker [3]. In last three decades, many experiments on Alfvén waves have been done where some fundamental properties of shear Alfvén wave have been explored. Oblique shear Alfvén wave, which is transverse with strong magnetic field variation perpendicular to the wave motion, can trade energy between the different frequencies which might propagate through plasma. This also means that energy can be exchanged with the particles in the plasma, in some cases, trapping particles in the troughs of the wave and carrying them along.

Plasma instabilities are caused by the energetic ion or electron beams. These beams are ever-present in the variety of space and astrophysical plasmas. Ion beam and interaction of radiation with plasma

Received 25 September 2022, Accepted 16 November 2022, Scheduled 27 November 2022

* Corresponding author: Suresh C. Sharma (suresh321sharma@gmail.in).

¹ Department of Applied Physics, Delhi Technological University, Shahbad Daultpur, Bawana Road, Delhi 110042, India. ² Department of Physics, Swami Shraddhanand College, University of Delhi, Alipur, Delhi 110036, India.

allocate a free energy source which can excite different wave modes in a plasma. Plasma waves can be stabilized or destabilized when they interact with electron or ion beams [4–11], and collisions generally have a stabilizing effect on the plasma waves [12, 13]. An ion beam can destabilize Alfvén waves and whistler waves if the beam speed is sufficiently large. The shear Alfvén wave instability depends on the free energy stored in the particle distribution function in a velocity space, and in a magnetized plasma, ion neutral collisions result in the change of its dispersion relation [14]. For the study of Alfvén waves employing parallel ion beams, several laboratory experiments have been performed. In a laboratory magnetoplasma, Tripathi et al. [15] demonstrated excitation of propagating shear Alfvén wave and established that the ambient plasma density and electron temperature were increased significantly by the ions beam. Zhang et al. [16] have experimentally explained the Doppler shifted cyclotron resonance between fast Lithium ions and shear Alfvén waves in the helium plasma of the Large Plasma Device. In the partially ionized solar chromospheres, the collisions between various particles are an efficient dissipative mechanism for Alfvén waves [16, 17].

Many physicists have shown that both Alfvén waves and magnetosonic waves can be excited nonlinearly [18–21] and have explained about interactions of Alfvén waves with ions [20]. Hollweg and Markovskii [21] have analytically studied the instabilities generated by cyclotron resonances of ions with obliquely propagating waves in coronal holes and solar wind. Li and Lu [22] have investigated the interaction between oblique propagating Alfvén waves and minor ions in the fast solar wind stream. Hellinger and Mangeney [23] studied structure of an oblique shock wave by means of numerical simulations, and they found that proton beam generates whistler waves at slightly oblique propagation. They also showed that for dense proton beams, the oblique modes have an important role in the nonlinear stage of electromagnetic instability [24]. Verscharen and Chandran [25] studied the polarization properties of both oblique and parallel propagating waves in the presence of ion beam and found that minimum beam speed required to excite such instabilities was significantly smaller for the former mode than for latter mode with $k \times B_0 = 0$. Maneva et al. [26] examined the comparative behavior of parallel vs oblique Alfvén cyclotron waves in the observed heating and acceleration alpha particles in the fast solar wind. They aimed to find which propagation angles were the most efficient in preferentially heating the alpha particles within the considered low frequency turbulent wave spectra. Gao et al. [27] found that the obliquely propagating Alfvén waves can be excited by alpha/proton instability, and background proton component & alpha component can be heated resonantly by Alfvén waves. The study of oblique modes is very important in the thermodynamics of minor ions in the solar wind and may also find application within the Earth’s bow shock as well as to basic plasma processes.

In the present paper, we study the reaction of oblique shear Alfvén waves with an ion beam forced into magnetized plasma parallel to the magnetic field. In Section 2, instability analysis is given. The dispersion relation and growth rate of Alfvén waves for fast and slow cyclotron interactions are derived in the absence as well as in the presence of plasma and beam collisions, and the results are discussed. Finally, the conclusion is given in Section 3.

2. INSTABILITY ANALYSIS

Assume a plasma in a dc background magnetic field $B_s \parallel \hat{z}$, with electron density n_{eo} , ion density n_{io} , electron mass m_e , ion mass m_i , electron charge $-e$, and ion charge e . A positively charged particle-beam having mass m_b , charge q_b , and density n_{bo} passes parallel to the magnetic field, through the plasma along the z -direction with velocity v_{bo} . Consider that an electromagnetic Alfvén wave is propagating through the plasma with electric field

$$\vec{E} = A e^{-i(\omega t - \vec{k} \cdot \vec{r})}, \quad \text{where} \quad \vec{k} = k_x \hat{x} + k_z \hat{z}.$$

The magnetic field of the electromagnetic wave is given as $\vec{B} = c \vec{k} \times \vec{E} / \omega$.

The equation of motion, governing the drift velocity of plasma ions, plasma electrons, and beam particles, is

$$m \left[\frac{\partial \vec{v}}{\partial t} + \vec{v} \cdot \nabla \vec{v} \right] = -e \vec{E} - \frac{e}{c} \vec{v} \times (\vec{B}_s + \vec{B}) - \nu m \vec{v}, \quad (1)$$

where ν is the collision frequency of beam particles and plasma components. In equilibrium (i.e., in the absence of the wave), $v_i = 0$ and $v_e = 0$. When wave is present, \vec{E} and \vec{B} of the wave are treated as

small or perturbed quantities and result in perturbed velocities v_{i1} , v_{e1} , and v_{b1} for plasma ions, plasma electrons, and beam particles, respectively.

On linearizing Eq. (1), we obtain

$$\frac{\partial \vec{v}_1}{\partial t} = -\frac{e\vec{E}}{m} - \vec{v}_1 \times \hat{z}\omega_{ce} - v\vec{v}_1, \quad (2)$$

where $\omega_{ce} = \frac{eB_s}{mc}$ is the electron cyclotron frequency.

By using equation of continuity, we have calculated the number densities of electrons and ions as

$$\frac{\partial n}{\partial t} + \vec{\nabla} \cdot (n\vec{v}) = 0, \quad (3)$$

Current densities of different components have been determined using the relation

$$\vec{J} = ne\vec{v}. \quad (4)$$

To obtain dispersion relation, the wave equation is given as

$$\nabla^2 \vec{E} - \nabla (\vec{\nabla} \cdot \vec{E}) + \frac{\omega^2}{c^2} \vec{E} = -\frac{4\pi i\omega}{c^2} \vec{J}. \quad (5)$$

2.1. Oblique Shear Alfvén Wave

We consider an elliptically polarized Alfvén wave propagating in X - Z plane with electric vectors also in X - Z plane, i.e., $\vec{E} = E_x \hat{x} + E_z \hat{z}$ and $\vec{k} = k_x \hat{x} + k_z \hat{z}$, $k_z = k \cos \theta$ and $k_x = k \sin \theta$.

Obtaining x , y , z -components of Eq. (2), we calculate the perturbed plasma electron velocities as

$$v_{e1x} = \frac{ie\omega E_x}{m_e \omega_{ce}^2}, \quad (6)$$

$$v_{e1y} = -\frac{eE_x \omega_{ce}}{m_e \omega_{ce}^2}, \quad (7)$$

$$\text{and } v_{e1z} = \frac{eE_z}{m_e i\omega}. \quad (8)$$

Substituting Eqs. (6), (7), and (8) in Eq. (4), we obtain the perturbed plasma electron current densities as

$$J_{e1x} = -n_{eo} \frac{e^2}{m_e} \frac{i\omega E_x}{\omega_{ce}^2}, \quad (9)$$

$$J_{e1y} = n_{eo} \frac{e^2}{m_e} \frac{\omega_{ce} E_x}{\omega_{ce}^2}, \quad (10)$$

$$\text{and } J_{e1z} = -\frac{n_{eo} e^2 E_z}{m_e i\omega}. \quad (11)$$

The perturbed plasma ion current densities are obtained from Eqs. (9), (10), and (11), by replacing e by $-e$, m_e by m_i , and ω_{ce} by ω_{ci} .

Similarly, the perturbed beam electron velocities are obtained using Eq. (2) as

$$v_{b1x} = \frac{q_b}{m_b \omega} \frac{(\nu - i\bar{\omega}) \bar{\omega} E_x}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]} + \frac{q_b}{m_b \omega} \frac{v_{bo} k_x (\nu - i\bar{\omega}) E_z}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \quad (12)$$

$$v_{b1y} = -\frac{q_b}{m_b \omega} \frac{\bar{\omega} \omega_{bc} E_x}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]} - \frac{q_b}{m_b \omega} \frac{v_{bo} k_x \omega_{bc} E_z}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \quad (13)$$

$$\text{and } v_{b1z} = \frac{q_b E_z}{m_b (\nu - i\bar{\omega})}, \quad (14)$$

where ω_{bc} is the cyclotron frequency of beam particles and $\bar{\omega} = \omega - k_z v_{bo}$.

The perturbed beam particle current densities calculated using Eq. (4) are

$$J_{b1x} = \frac{q_b^2}{m_b \omega} \frac{n_{bo}(\nu - i\bar{\omega})\bar{\omega}E_x}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]} + \frac{q_b^2}{m_b \omega} \frac{n_{bo}v_{bo}k_x(\nu - i\bar{\omega})E_z}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \quad (15)$$

$$J_{b1y} = -\frac{q_b^2}{m_b \omega} \frac{n_{bo}\bar{\omega}\omega_{bc}E_x}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]} - \frac{q_b^2}{m_b \omega} \frac{n_{bo}v_{bo}\omega_{bc}E_z}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \quad \text{and} \quad (16)$$

$$J_{b1z} = \frac{q_b^2}{m_b \omega} \frac{n_{bo}v_{bo}k_x(\nu - i\bar{\omega})\bar{\omega}E_x}{[(\nu - i\bar{\omega})^2 + \omega_{bc}^2]} + \frac{q_b^2 n_{bo}}{m_b} \left[\frac{\omega}{\bar{\omega}(\nu - i\bar{\omega})} + \frac{v_{bo}^2 k_x^2 (\nu - i\bar{\omega})}{\omega \bar{\omega} [(\nu - i\bar{\omega})^2 + \omega_{bc}^2]} \right] E_z \quad (17)$$

Substituting Eqs. (9), (10), (11), (15), (16), (17) and corresponding current densities of plasma ions in Eq. (5), we obtain

$$\begin{aligned} \eta E_x + \mu E_z &= 0 \\ \mu E_x + \lambda E_z &= 0 \\ \text{or } \begin{pmatrix} \eta & \mu \\ \mu & \lambda \end{pmatrix} \begin{pmatrix} E_x \\ E_z \end{pmatrix} &= 0, \end{aligned} \quad (18)$$

where

$$\begin{aligned} \eta &= -k_z^2 + \frac{\omega^2}{c^2} + \frac{\omega_{pe}^2 \omega^2}{c^2 \omega_{ce}^2} + \frac{\omega_{pi}^2 \omega^2}{c^2 \omega_{ci}^2} + \frac{\omega_{pb}^2 i(\nu - i\bar{\omega})\bar{\omega}}{c^2 [(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \\ \mu &= k_x k_z + \frac{\omega_{pb}^2 i v_{bo} k_x (\nu - i\bar{\omega})}{c^2 [(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \\ \lambda &= -k_x^2 + \frac{\omega^2}{c^2} - \frac{\omega_{pe}^2}{c^2} - \frac{\omega_{pi}^2}{c^2} + \frac{\omega_{pb}^2 i \omega^2}{c^2 \bar{\omega} (\nu - i\bar{\omega})} + \frac{\omega_{pb}^2 v_{bo}^2 k_x^2 i (\nu - i\bar{\omega})}{c^2 \bar{\omega} [(\nu - i\bar{\omega})^2 + \omega_{bc}^2]}, \\ \omega_{pe}^2 &= \frac{4\pi n_{eo} e^2}{m_e}, \quad \omega_{pi}^2 = \frac{4\pi n_{io} e^2}{m_i} \quad \text{and} \quad \omega_{pb}^2 = \frac{4\pi n_{bo} q_d^2}{m_d}. \end{aligned} \quad (19)$$

The dispersion relation can be obtained by taking

$$\eta \lambda - \mu^2 = 0. \quad (20)$$

For elliptically polarized mode of propagation, $E_x = \pm i E_z$, which gives the dispersion relation as

$$\begin{aligned} (\omega^2 - \omega_s^2) &= \frac{-\omega_{pb}^2 [(\bar{\omega}^2 - k_x v_{bo} \nu) (\omega_{bc}^2 - \bar{\omega}^2) - 2\bar{\omega}^2 \nu (k_x v_{bo} + \nu)]}{(\omega_{bc}^2 - \bar{\omega}^2)^2 (1 + A)} \\ &\quad + i \frac{\bar{\omega} [(k_x v_{bo} + \nu) (\omega_{bc}^2 - \bar{\omega}^2) + 2\nu (\bar{\omega}^2 - k_x v_{bo} \nu)]}{(\omega_{bc}^2 - \bar{\omega}^2)^2 (1 + A)}, \end{aligned} \quad (21)$$

$$\text{where } \omega_{s\mp}^2 = \frac{k_z^2 c^2 \mp i k_x k_z c^2}{(1 + A)}, \quad \text{and } A = 1 + \frac{\omega_{pe}^2}{\omega_{ce}^2} + \frac{\omega_{pi}^2}{\omega_{ci}^2}. \quad (22)$$

Equation (22) is the modified dispersion relation of RH (subscript-) and LH (subscript+) elliptically polarized oblique shear Alfvén wave in the absence of beam.

In terms of θ , the angle of propagation of shear Alfvén wave with respect to the ambient magnetic field, we can write Eq. (22) as

$$\omega_{s\mp} = \omega_A \frac{1}{\sqrt{1 - \tan^2 \theta/2}} \mp i \omega_A \frac{1}{\sqrt{\cot^2 \theta/2 - 1}}. \quad (23)$$

where $\omega_A = \frac{k_z c}{\sqrt{1+A}}$ is the Alfvén frequency. Eq. (23) indicates that the RH polarized oblique wave damps, and LH polarized oblique wave grows due to obliquity. The real frequency for both the modes

increases with angle θ , and the growth (or attenuation) of LH (or RH) waves also increases with angle θ .

From Eq. (21), we can say that both fast and slow cyclotron interactions are possible for both LH and RH polarized oblique Alfvén waves, in contrast to only slow (or fast) cyclotron interaction for RH (or LH) polarized parallel propagating Alfvén waves. In Fig. 1, the dispersion curves of Alfvén waves are plotted for different wave propagation angles θ ($= 0^\circ, 30^\circ, 45^\circ, 60^\circ$, and 80°), plotted using Eq. (23) along with the beam mode for beam velocity $v_{bo} = 1.69 \times 10^8$ cm/s. The frequencies and the corresponding parallel wave numbers of the unstable modes obtained from the points of intersection between the beam mode and plasma modes are given in Table 1. The unstable frequency of Alfvén waves ' ω ' gradually increases while the unstable wave number ' k_z ' decreases with increase in the angle of wave propagation θ .

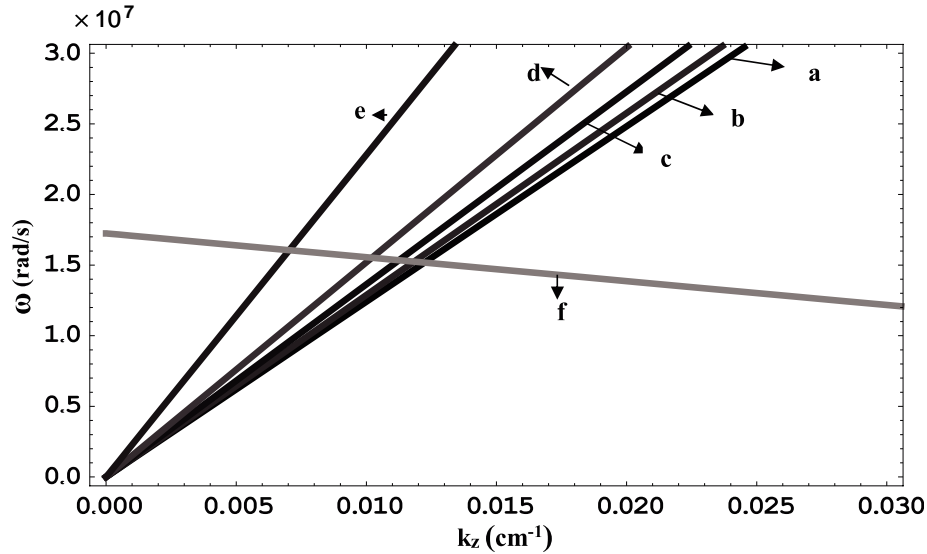


Figure 1. Dispersion curves of oblique shear Alfvén wave for different angles of propagation θ , (a) 0° , (b) 30° , (c) 45° , (d) 60° and (e) 80° and the beam mode (f) with velocity 1.69×10^8 cm/s.

Table 1. Angle of propagation, unstable wave numbers and frequencies of oblique shear Alfvén waves and maximum growth rates of wave in the absence and presence of beam-plasma collisions.

θ	0°	30°	45°	60°	80°
k_z (cm $^{-1}$)	0.01223	0.01184	0.01126	0.01021	0.00704
ω ($\times 10^7$ rads $^{-1}$)	1.5178	1.5245	1.5342	1.5514	1.6054
ω/ω_{bc}	0.8801	0.8840	0.8896	0.8996	0.9309
γ ($\times 10^5$ s $^{-1}$)	0	1.94053	3.08662	4.52084	8.08631
γ_{wc} ($\times 10^5$ s $^{-1}$) $\nu = 10^4$ s $^{-1}$	1.71571	1.75076	1.80275	1.89856	2.24340

Assuming fast cyclotron interaction and considering perturbed quantities

$$\omega = \omega_{s\mp} + \Delta \quad \text{and} \quad \bar{\omega} = \omega_{bc} + \Delta, \quad \text{we get from Eq. (21)}$$

$$\Delta^3 = \frac{-\omega_{pb}^2 \bar{\omega}^2 \nu}{4\omega_s \omega_{bc}^2 (1+A)} (\mp k_x v_{bo} + i\omega_{bc}), \quad (24)$$

$$\text{or } \Delta = \left[\frac{\omega_{pb}^2 \nu k v_{bo}}{4\omega_s (1+A)} \right]^{1/3} (\pm \sin \theta/3 + i \cos \theta/3). \quad (25)$$

In Eq. (25), a positive imaginary part or the growth rate establishes instability, while a negative imaginary part indicates damping. A positive (or negative) real part means that the frequency of wave increases (or decreases) on interaction between beam mode and wave mode. The frequency of RH polarized mode increases while the frequency of LH polarized mode decreases with the increase in angle of propagation of wave (θ) with respect to ambient magnetic field, and both the modes become unstable due to beam-wave interaction. The growth rate increases monotonically as the angle of propagation increases, but the increase in maximum growth rate in the presence of beam collisions is nominal. The frequency and growth rate of waves depend mainly on the value of $(\frac{\omega + k_z v_{bo}}{k_x v_{bo}})$ [cf. Eq. (24)]. $\omega + k_z v_{bo} = \omega_{bc} = k_x v_{bo}$ corresponds to the case of parallel propagation. The frequency increases/decreases for RH/LH polarized waves by a factor of 0.5 as θ varies from 0° to 90° due to beam-wave interaction, and the maximum growth rate at $\theta = 80^\circ$ becomes 1.3 times of maximum growth rate at $\theta = 30^\circ$.

If the collisions are ignored, Eq. (21) becomes

$$(\omega^2 - \omega_{s\mp}^2) = \frac{\omega_{pb}^2 (\omega + k_z v_{bo}) k v_{bo}}{[(\omega + k_z v_{bo})^2 - \omega_{bc}^2] (1 + A)} \exp(\mp i\theta) \quad (26)$$

Assuming perturbed quantities for fast cyclotron interaction, we get

$$\Delta = \left[\frac{\omega_{pb}^2 (\omega + k_z v_{bo}) k v_{bo}}{4\omega_{bc}\omega_s (1 + A)} \right]^{1/2} (\cos \theta/2 \mp i \sin \theta/2) \quad (27)$$

Using Eq. (26) we plot, in Fig. 2, the growth rate of oblique shear Alfvén wave with beam collision frequency γ_{wc} (in sec^{-1}), and using Eq. (27) we plot, in Fig. 3, the growth rate of oblique shear Alfvén wave without collision frequency γ , as a function of k_z , for the same parameters used for plotting dispersion curves. Fig. 2 shows that the maximum growth rate as well as the spectrum of the unstable mode increases with the increase in angle of propagation θ . Collisions have induced a growth in elliptically X-Z polarized wave for parallel propagation.

In the absence of collisions, for angle of propagation $\theta = 0^\circ$, the condition is analogous to the interaction of pure Alfvén wave, as E_z does not affect the interaction for parallel propagating waves. The frequency then increases, and growth rate is zero. The beam generates LH polarized waves and stabilizes the RH polarized wave.

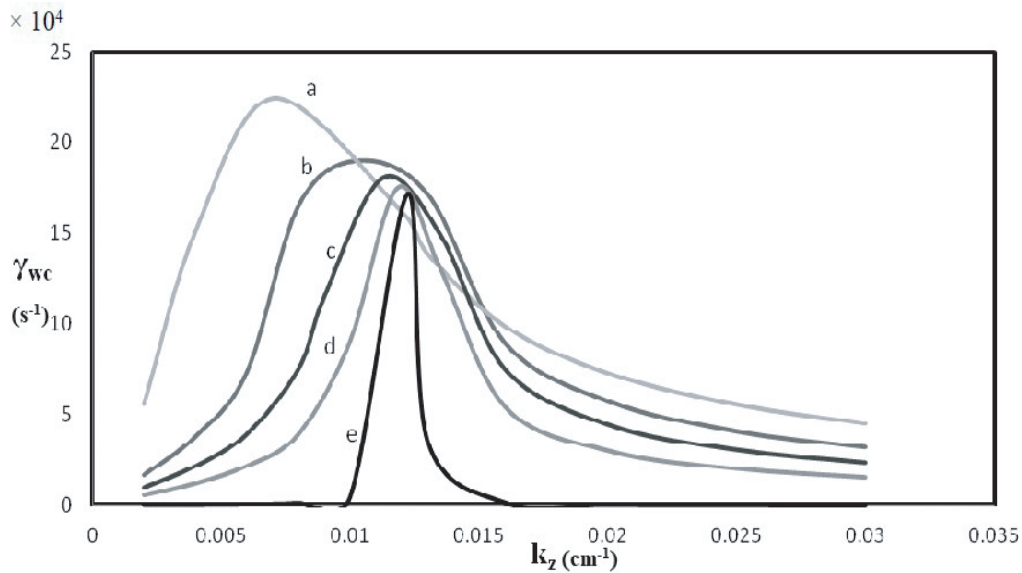


Figure 2. Growth rate γ_{wc} (in sec^{-1}) of the unstable mode as a function of k_z (in cm^{-1}) for different angles, (a) 80° , (b) 60° , (c) 45° , (d) 30° and (e) 0° of propagation of shear Alfvén wave in the presence of beam-plasma collisions.

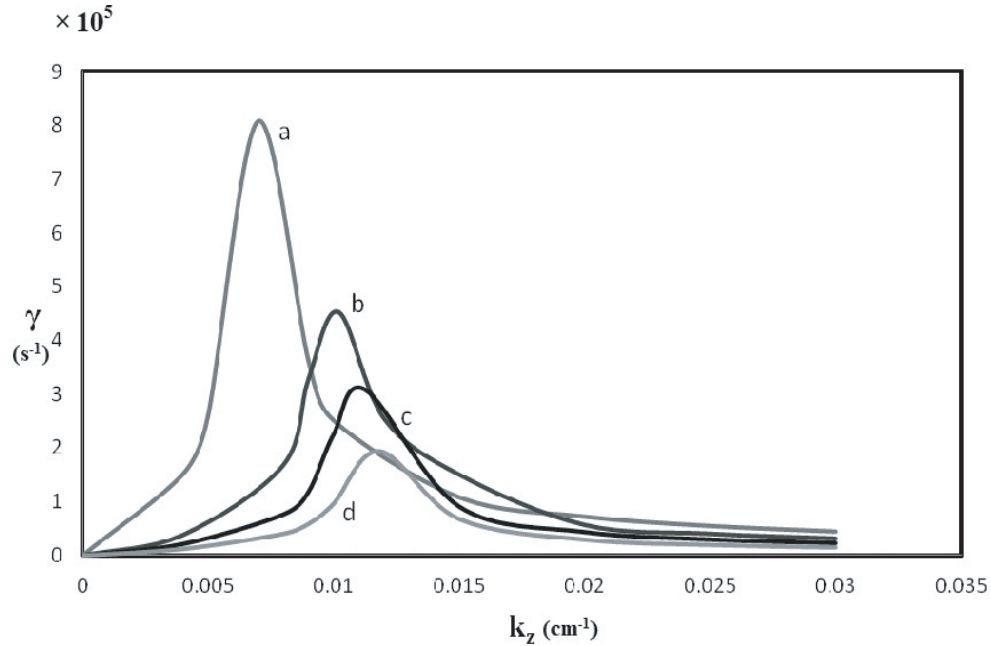


Figure 3. Growth rate γ (in sec^{-1}) of the unstable mode as a function of k_z (in cm^{-1}) for different angles, (a) 80° , (b) 60° , (c) 45° and (d) 30° of propagation of shear Alfvén wave.

Equation (27) shows that the maximum growth rate of the unstable LH polarized oblique Alfvén waves occurs at more oblique angles. The produced waves heat/accelerate the plasma ions. On the other hand, the RH polarized oblique Alfvén wave loses their energy and heat the beam ions. Li and Lu [22] have also observed similar results using hybrid simulations. The maximum growth rate becomes four-fold as θ changes from 30° to 80° , and the unstable wave spectrum decreases with θ . The maximum growth rate of the LH polarized mode excited at nearly perpendicular propagation is

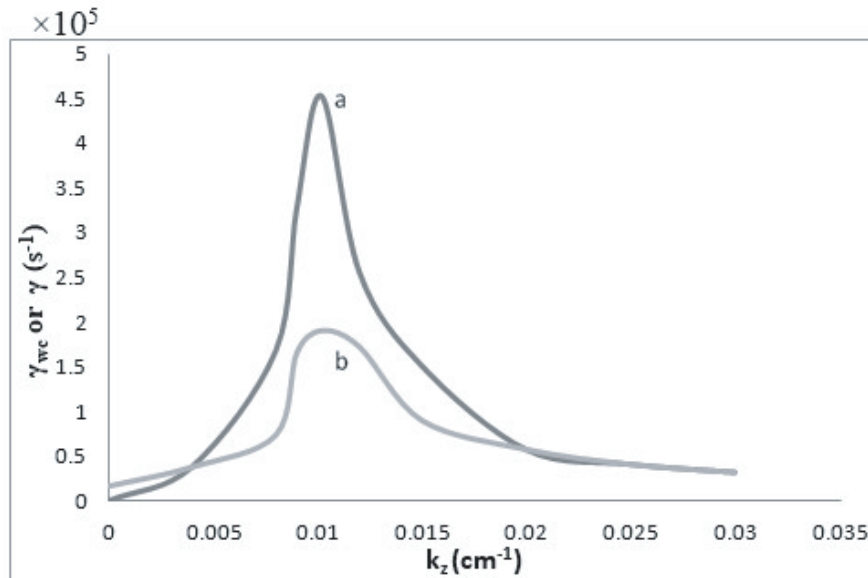


Figure 4. Growth rates, (a) γ and (b) γ_{wc} (in sec^{-1}) of the unstable mode as a function of k_z (in cm^{-1}) for an angle of propagation of shear Alfvén wave 60° in the absence and presence of beam-plasma collisions.

$0.05\omega_{bc}$ occurring at 80° . The maximum frequency of the unstable wave mode decreases slightly with an increase in propagation angle. From Table 1, we may conclude that the maximum growth rate values in the presence or absence of beam collisions increase due to obliquity of wave.

Assuming slow cyclotron interaction and considering perturbed quantities $\omega = \omega_{s\mp} + \Delta$ and $\bar{\omega} = -\omega_{bc} + \Delta$, we get from Eq. (26)

$$\Delta = \left[\frac{\omega_{pb}^2 \bar{\omega}^2 \nu k v_{bo}}{4\omega_s \omega_{bc}^2 (1 + A)} \right]^{1/3} \{ \cos(\pi/6 \mp \theta/3) + i \sin(\pi/6 \mp \theta/3) \}. \quad (28)$$

The behaviours of RH and LH polarized waves are reciprocal to each other. As the angle of propagation θ increases, the frequency of RH mode (or LH mode) increases (or decreases), while the growth rate of RH mode (or LH mode) decreases (or increases) by small factors.

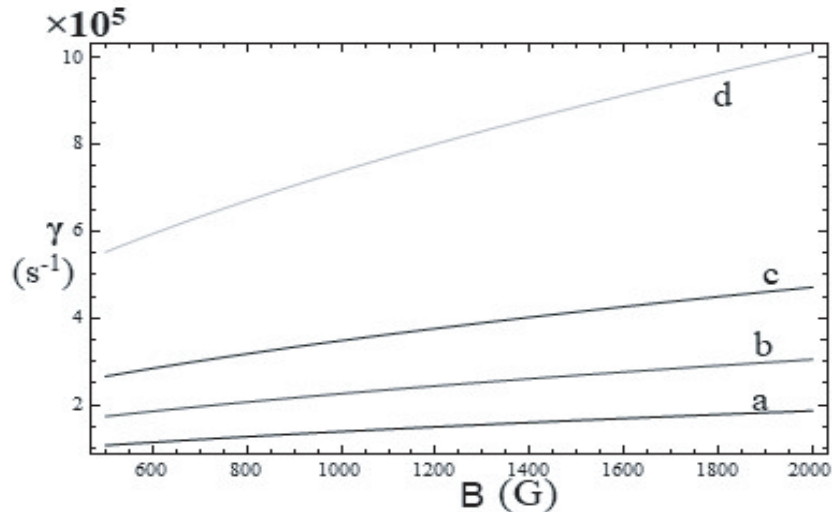


Figure 5. Growth rate γ (in sec^{-1}) of the unstable mode as a function of B (in gauss) for different angles, (a) 30° , (b) 45° , (c) 60° and (d) 80° of propagation of shear Alfvén wave.

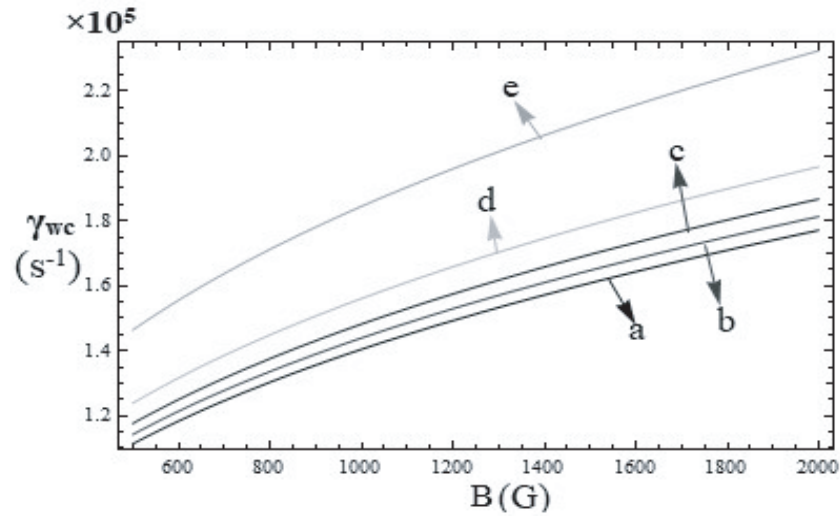


Figure 6. Growth rate γ_{wc} (in sec^{-1}) of the unstable mode as a function of B (in gauss) for different angles, (a) 0° , (b) 30° , (c) 45° , (d) 60° and (e) 80° of propagation of shear Alfvén wave in the presence of beam-plasma collisions.

In the absence of collisions, for slow cyclotron interaction, we get

$$\Delta = \left[\frac{\omega_{pb}^2 k v_{bo}}{4\omega_s (1 + A)} \right]^{1/2} (\pm \sin \theta/2 + i \cos \theta/2). \quad (29)$$

The frequency of RH mode increases, while that of LH mode decreases in this interaction, while both the modes grow with time.

Figure 4 shows the comparison of growth rates of unstable shear Alfvén modes in the presence and absence of beam plasma collisions as a function of k_z at wave propagation angle of 60° , which justifies that the growth rate decreases due to collisions. A rise in the value of ambient magnetic field raises the maximum growth rate in both the cases as shown in Fig. 5 and Fig. 6. However, the rate of increase in growth rate is slightly more in the presence of collisions. The variation of growth rate of the unstable mode with the number densities of electrons in the plasma are shown in Fig. 7 and Fig. 8 for different angles of propagation without and with beam plasma collisions. Increase in the number

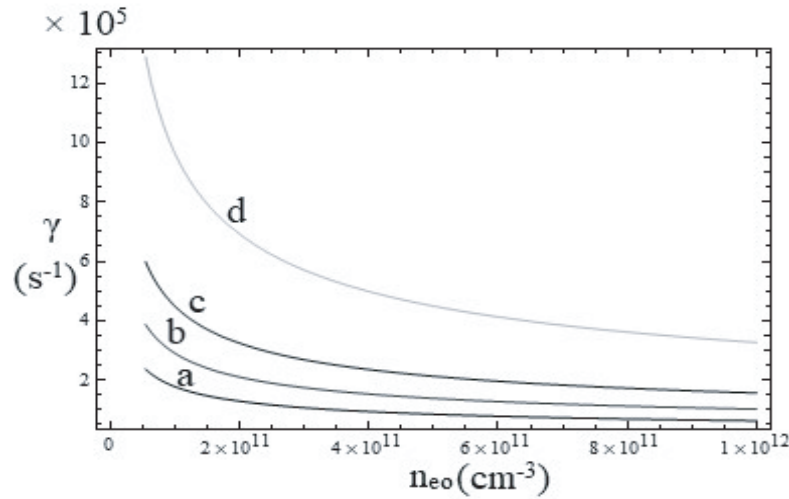


Figure 7. Growth rate γ (in sec^{-1}) of the unstable mode as a function of n_{eo} (in cm^{-3}) for different angles, (a) 30° , (b) 45° , (c) 60° and (d) 80° of propagation of shear Alfvén wave.

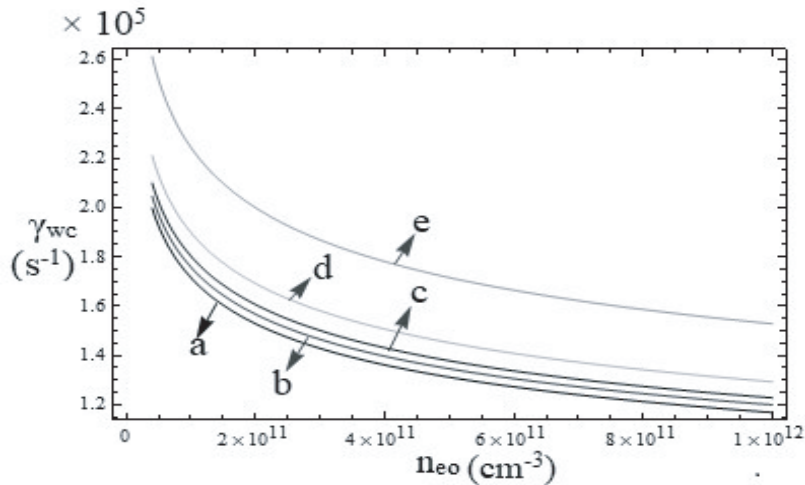


Figure 8. Growth rate γ_{ec} (in sec^{-1}) of the unstable mode as a function of n_{eo} (in cm^{-3}) for different angles, (a) 0° , (b) 30° , (c) 45° , (d) 60° and (e) 80° of propagation of shear Alfvén wave in the presence of beam-plasma collisions.

density of electrons decreases the relative density of beam ions in plasma, thus decreasing the growth rate of modes. It can also be seen from the growth rate expressions that an increase in the ion beam velocity increases the growth rate. Similar results for growth rates have also been observed by Xiang et al. [28], and they have also observed a higher growth rate for RH waves than the LH waves.

3. CONCLUSION

In this paper, we show that an ion beam can efficiently transfer its energy to plasma through wave generation. The beam ions are resonant with both the parallel and oblique shear Alfvén modes. The frequency of waves may increase or decrease depending upon the polarization of waves. For oblique shear Alfvén waves, the waves grow irrespective of their polarization via beam-wave cyclotron interaction, and the beam ions do not show Landau resonance. The numerical values of the growth rate are however much smaller in the presence of collisions, indicating a collisional damping of waves. The growth rates and frequencies of generated waves are also sensitive to the beam properties. Our results indicate that the growth rates of LH and RH Alfvén modes vary with the velocity of ion beam, number density of plasma electrons, ambient magnetic field, and collisions in plasma. A stronger magnetic field supports the Alfvén wave generation. For $(\omega + k_z v_{bo} > \omega_{bc})$, the most unstable mode is the LH polarized oblique Alfvén mode, and for $(\omega + k_z v_{bo} < \omega_{bc})$, it is the RH polarized oblique Alfvén mode, indicating a polarization reversal after resonance condition. In this case, the beam-wave interaction could take place only for propagation of Alfvén waves and beam ions in opposite direction, with respect to the ambient magnetic field, and only via fast cyclotron interaction. The growth rate is more for parallel shear Alfvén waves than oblique shear Alfvén waves, making them the dominant modes. The study of oblique modes may find application within the Earth's bow shock as well as to basic plasma processes and are important in the thermodynamics of minor ions in the solar wind.

The collision of beam ions with plasma components affects the growth rate as well as the frequency of generated waves, while the collision between plasma components damps the waves. The effect of beam collisions can be ignored for beam velocities more than $\sim 10^8$ cm/s, but as the velocity of beam decreases, the role of collisions becomes more and more significant, and they decrease the growth rate of Alfvén waves. The effect of beam and plasma collisions becomes more significant for heavy ion beams or in a complex multi-component plasma, which will be a subject of our future work.

REFERENCES

1. Gekelman, W., Vincena, D. S. Leneman, and J. Maggs, "Laboratory experiments on shear Alfvén waves and their relationship to space plasmas," *J. Geophys. Res.*, Vol. 102, No. A4, 7225–7236, 1997.
2. Chen, L. and F. Zonca, "Physics of Alfvén waves and energetic particles in burning plasmas," *Rev. Mod. Phys.*, Vol. 88, No. 1, 015008, 2016.
3. Jephcott, D. F. and P. M. Stocker, "Hydromagnetic waves in a cylindrical; plasma: An experiment," *J. Fluid Mech.*, Vol. 13, No. 4, 587–596, 1962.
4. Dwivedi, A. K., S. Kumar, and M. S. Tiwari, "Effect of ion and electron beam on kinetic Alfvén wave in an inhomogeneous magnetic field," *Astrophys. Space Sci.*, Vol. 350, No. 2, 547–556, 2014.
5. Prakash, V., R. Gupta, S. C. Sharma, and Vijayshri, "Excitation of lower hybrid wave by an ion beam in magnetized plasma," *Laser Part. Beams*, Vol. 31, No. 4, 747–752, 2013.
6. Gupta, R., V. Prakash, S. C. Sharma, and Vijayshri, "Interaction of an electron beam with whistler waves in magnetoplasmas," *Laser Part. Beams*, Vol. 33, No. 3, 455–461, 2015.
7. Gupta, R., V. Prakash, S. C. Sharma, Vijayshri, and D. N. Gupta, "Resonant ion beam interaction with Whistler waves in a magnetized dusty plasma," *J. Atomic, Molecular, Condensate and Nano Physics*, Vol. 3, No. 1, 45–53, 2016.
8. Prakash, V., R. Gupta, Vijayshri, and S. C. Sharma, "Excitation of electromagnetic surface waves at a conductor-plasma interface by an electron beam," *J. Atomic, Molecular, Condensate and Nano Physics*, Vol. 3, No. 1, 35–43, 2016.

9. Prakash, V. and S. C. Sharma, "Excitation of surface plasma waves by an electron beam in a magnetized dusty plasma," *Phys. Plasmas.*, Vol. 16, No. 9, 93703, 2009.
10. Prakash, V., S. C. Sharma, Vijayshri, and R. Gupta, "Surface wave excitation by a density modulated electron beam in a magnetized dusty plasma cylinder," *Laser Part. Beams*, Vol. 31, 411–418, 2013.
11. Shoucri, M. M. and R. R. J. Gagne, "Excitation of lower hybrid waves by electron beams in finite geometry plasmas. Part 1. Body waves," *J. Plasma Phys.*, Vol. 19, No. 2, 281–294, 1978.
12. Rubab, N. and G. Jaffer, "Excitation of dust kinetic Alfvén waves by semi-relativistic ion beams," *Phys. Plasmas.*, Vol. 23, No. 5, 053701, 2016.
13. Shevchenko, V. I., V. L. Galinsky, and S. K. Ride, "Excitation of left-hand-polarized nonlinear Alfvén waves by an ion beam in a plasma," *J. Geophys. Res.*, Vol. 107, No. A11, 1–12, 2002.
14. Amagishi, Y. and M. Tanaka, "Ion-neutral collision effect on an Alfvén wave," *Phys. Rev. Lett.*, Vol. 71, No. 3, 360–363, 1993.
15. Tripathi, S. K. P., B. V. Compernelle, W. Gekelman, P. Pribyl, and W. Heidbrink, "Excitation of shear Alfvén waves by a spiraling ion beam in a large magnetoplasma," *Phys. Rev. E*, Vol. 91, No. 1, 1–5, 2015.
16. Zhang, Y., W. W. Heidbrink, H. Boehmer, R. McWilliams, S. Vincena, T. A. Carter, W. Gekelman, D. Leneman, and P. Pribyl, "Observation of fast-ion Doppler-shifted cyclotron resonance with shear Alfvén waves," *Phys. Plasmas.*, Vol. 15, No. 10, 102112, 2008.
17. Soler, R., J. L. Ballester, and T. V. Zaqarashvili, "Overdamped Alfvén waves due to ion-neutral collisions in the solar chromospheres," *A & A*, Vol. 573, 79–91, 2014.
18. Shukla, P. K. and L. Stenflo, "Periodic structures on an ionic-plasma-vacuum interface," *Phys. Plasmas.*, Vol. 12, No. 4, 0845021, 2005.
19. Shukla, P. K., M. Y. Yu, and L. Stenflo, "Growth rates of modulationally unstable ion-cyclotron Alfvén waves," *Phys. Scr.*, Vol. 34, No. 2, 169–170, 1986.
20. Lu, X. Q., W. Z. Tang, W. Guo, and X. Y. Gong, "A study on interactions between ions and polarized Alfvén waves below cyclotron resonance frequency," *Phys. Plasmas.*, Vol. 23, No. 12, 1–5, 2016.
21. Hollweg, J. V. and S. A. Markovskii, "Cyclotron resonances of ions with oblique propagating waves in coronal holes & the fast solar wind," *J. Geophys. Res.*, Vol. 107, No. A6, 1–7, 2002.
22. Li, X. and Q. M. Lu, "Heating and deceleration of minor ions in the extended fast solar wind by oblique Alfvén waves," *J. Geophys. Res.*, Vol. 115, No. A48, A08105, 2010.
23. Hellinger, P. and A. Mangeney, "Structure of low mach number oblique shock waves," *Correlated Phenomena at the Sun, in the Heliosphere and in Geospace*, 337–340, 1997.
24. Hellinger, P. and A. Mangeney, "Electromagnetic ion beam instabilities — Oblique pulsation," *J. Geophys. Res. Atmos.*, Vol. 104, No. A3, 4669–4680, 1999.
25. Verscharen, D. and B. D. G. Chandran, "The dispersion relations and instability thresholds of oblique plasma modes in the presence of an ion beam," *Astrophys. J.*, Vol. 764, No. 1, 1–12, 2013.
26. Maneva, Y. G., A. F. Vinas, P. S. Moya, R. T. Wicks, and S. Poedts, "Dissipation of parallel & oblique Alfvén-cyclotron waves — Implications for heating of alpha particles in the solar wind," *Astrophys. J.*, Vol. 814, No. 1, 1–15, 2015.
27. Gao, X., Q. Lu, X. Li, C. Huang, and S. Wang, "Heating of the background plasma by obliquely propagating Alfvén waves excited in electromagnetic alpha/proton instability," *Phys. Plasmas.*, Vol. 19, No. 3, 1–5, 2012.
28. Xiang, L., D. J. Wu, and L. Chen, "Effect of alpha beams on low-frequency electromagnetic waves driven by proton beams," *Astrophys. J.*, Vol. 869, No. 64, 1–10, 2018.

Predictive linguistic cues for fake news: a societal artificial intelligence problem

Sandhya Aneja¹, Nagender Aneja², Ponnurangam Kumaraguru³

¹Faculty of Integrated Technologies, Universiti Brunei Darussalam, Gadong, Brunei Darussalam

²School of Digital Science, Universiti Brunei Darussalam, Gadong, Brunei Darussalam

³Department of Computer Science and Engineering, IIIT Hyderabad, New Delhi, India

Article Info

Article history:

Received Aug 22, 2021

Revised Jun 30, 2022

Accepted Jul 11, 2022

Keywords:

Fake news

Lexicon analysis

Linguistic analysis

Machine learning

Sentiment analysis

ABSTRACT

Media news are making a large part of public opinion and, therefore, must not be fake. News on web sites, blogs, and social media must be analyzed before being published. In this paper, we present linguistic characteristics of media news items to differentiate between fake news and real news using machine learning algorithms. Neural fake news generation, headlines created by machines, semantic incongruities in text and image captions generated by machine are other types of fake news problems. These problems use neural networks which mainly control distributional features rather than evidence. We propose applying correlation between features set and class, and correlation among the features to compute correlation attribute evaluation metric and covariance metric to compute variance of attributes over the news items. Features unique, negative, positive, and cardinal numbers with high values on the metrics are observed to provide a high area under the curve (AUC) and F1-score.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Nagender Aneja

School of Digital Science, Universiti Brunei Darussalam,

Academy of Brunei Studies, BE1410, Brunei Darussalam

Email: nagender.aneja@ubd.edu.bn

1. INTRODUCTION

In the media, fake news refers to news that has been fabricated and is presented to readers as being accurate. People in advanced economies are likely to see more fake content (70%) than real content. Fake news can now be created by humans or by artificial intelligence (AI) [1]. There are numerous fact-checking tools available, including NewsGuard and Hoaxy. Fact-checking websites, such as PolitiFact, GossipCop, and BuzzFeed [2], are still working on improving their ability to identify false information. However, while the quality of news content on social media is lower than that of traditional media, around 50% of Americans in 2021 get news from social media [3]. The revenue from traditional news media is also shrinking, and the online publishers are trying to earn advertising revenue by having more clicks on their content. The distrust of facts proffered by the established media is also rising. Because of the rapid dissemination, easy access, and low-cost dissemination of news on social media, the number of fake news stories is increasing all the time [4].

The goal of the linguistic analysis is to look for language leakage, also called predictive linguistic cues to detect fake news. Recent work on automatic detection captures the predictive cues or writing style using linguistic features, e.g., lexical, syntax, semantic features of the fake content [5], [6]. The news writing style captures the frequency of words accounted in content at linguistic-level, choice between noun/pronoun, writing cardinal number (CN), adjectives, using verbs at syntax level, and psycho-linguistic attributes at the semantic level. Writers of fake news prefer to use their language strategically to influence human psychology.

Rashkin *et al.* [7] presented that language stylistic cues can determine the truthfulness of text. The authors compared the language of real news (from English Gigaword corpus) with that of satire (The Onion, The Borowitz Report, Clickhole), hoaxes (American News, DC Gazette), and propaganda (The Natural News, Activist Report). The authors observed lexicon markers e.g., swear, 2nd person pronoun, modal adverb, action adverb, 1st person pronoun singular, manner adverb, gender, see, negation, strong subjective, hedge, superlatives, weak subjective were more prominent in fake news, while number, hear, money, assertive, and comparatives were more prominent in the truthful news. The fake news detection algorithm is further shown to depend on the stance classification of a news [8].

Allcott and Gentzkow [9] studied news articles of the 2016 US elections. They collected 156 news articles from which 41 were recorded as anti-Trump and 115 as anti-Clinton. Anti-Clinton articles were found 30.3 million times shared on Facebook. The sentiment analysis in [10], [11] including the positive and negative sentiment of input text for news classification seems promising. A study by Horne and Adali [10] on the headline concerning text-body of news for stance classification concluded that headline in fake news repeats the main content.

Efforts are being made to automate the process of fake news detection [12]–[16]. One such technique is Generating aRticles by Only Viewing mEtadata Records (GROVER) [17] which generates fake news and then uses nucleus sampling at each time step to sample from the most probable words whose cumulative probability comprises the top-p% of the entire vocabulary, to create fake news. It gives around 92% accuracy. However, when it is applied to human written fake news it gives 73% accuracy. Thus, it is required to create classifiers that are trained on language written by humans. Sentences created by generative models are distinguishable from human generated text due to the property of low variance and small vocabulary. This property is used by descriptors to the validity of the text [18]. The success of machine learning models depends on feature engineering since all features of a dataset might not be useful in building a machine learning (ML) model for prediction [19]–[23]. Accurate selection of effective features is a crucial step for applying ML algorithms. The automated approach given by Maronikolakis *et al.* [24] applies many recurrent neural networks (RNN) models to detect headlines created by humans or machine generated news. The paper analyses human and machine generated headlines. It was found that humans were only able to identify the fake headlines in 45% of the cases, whereas, the most accurate automatic approach of transfer learning in the paper achieved an accuracy of 94%.

Tan *et al.* [25] presented an approach to detect the semantic incongruities that are present in text and image captions generated by automated machines. The approach determines the authenticity score by using the co-occurrences of named entities in the text and captions. The word embeddings of captions and image are projected into a common visual semantic space which has a property to be built on fine-grained interactions between words in the caption and objects in the image. A semantic similarity score is computed for every possible pair of projected word and object features. The final authenticity score of an article is determined across those of its images and captions. The approach is compared with GROVER [17] model and outperformed the same. Various deep learning-based techniques are being studied to improve the correlation [26] between features through an attention mechanism. The techniques extend the feature space including multimodal features from audio, video or textual representations into the news content and apply the attention mechanism to mine the complex correlations.

In this paper, fake news detection emphasizes the technique to deeply mine the news content while using the linguistic analysis and language feature set using ML algorithms. We propose applying correlation between features set and class to compute correlation attribute evaluation metric and covariance metric to compute variance over the news items. Proposed feature set can differentiate between fake and real news with high accuracy (nearly $97 \pm 2\%$ area under curve (AUC) score) using the AdaBoost model. Main contributions of the paper are:

- a) A study of feature set comprising unique words, negative words, neutral words, positive words, compound score, noun, adjective, adverb, preposition, CN for fake news classification.
- b) We found a feature set that performed better in comparison to a set of all the features considered in the study using the Corr metric.
- c) Results show that the performance of classifiers depend on the news content i.e., linguistic characteristics of the news.
- d) Proposed methodology works for balanced, imbalanced, and small datasets.

2. METHOD

We used four datasets from Kaggle, BuzzFeed, PolitiFact, and FakeNews Challenge as shown in Table 1. Kaggle-Guardian Dataset comprises fake news from Kaggle and real news from guardian. The Kaggle data set contains text and metadata scraped from 244 different websites tagged as bullshit (BS) by the

BS detector chrome extension. We considered only English language news available in the Kaggle dataset. The total number of english language news that was found is 11439. To compare the linguistic features of fake news and real news, we downloaded 9,724 news items from the guardian using guardian application programming interface (API). The news items that we downloaded from the guardian were searched with keywords based on terms in the Kaggle dataset.

BuzzFeed news dataset [2] is collected from fact-checking platform BuzzFeed.com containing news content body text, headline, and uniform resource locator (URL) of the news posted on Twitter by the users. There are 91 real news and 91 fake news propagated through 634,750 social links by 15,257 users. PolitiFact news dataset [2] is collected from fact-checking platform PolitiFact.com similar to Buzzfeed. There are 120 real news and 120 fake news propagated through 574,744 social links by 37,259 users. Fake News Challenge Dataset includes news body text, headline, URL of the news posted with its stance correlated by the user. The dataset has news categorized into four classes agree, disagree, unrelated, and discuss. We changed four classes to two classes by taking agree class as real news while news with stances - disagree, unrelated, and discuss as fake news. There are 49,970 total news with 46,293 fake news and 3,678 real news, this is an imbalanced dataset with 1:12 ratio.

Table 1. Count of news item

	Kaggle and Guardian	BuzzFeed	PolitiFact	FakeNews
Real News	9,724	91	120	3,678
Fake News	11,439	91	120	46,293

2.1. Feature engineering

To create a feature set, the text of the news was tokenized using word tokenize function of Python nltk library. All tokens that were stop-words as per nltk corpus were removed to create clean text. SentimentIntensityAnalyzer and pos tag were used on the stem of the words from clean text to compute sentiment and parts of speech (POS) tag for each word. SnowballStemmer of Python was used to consider stem of the word to ignore different forms of the word. Frequency of POS tag and sentiment categories were computed for each news item for all the datasets to create features set.

2.1.1. Features set

Features set comprises unique words, negative words, neutral words, positive words, compound score, noun, adjective, adverb, preposition, verb in base form (VB), verb past tense (VBD), verb in gerund or present participle (VBG), verb in past participle (VBN), verb in 3rd person singular present (VBZ), and CN. Unique words represent the number of words that are unique in the given text. Unique words were observed to make 60-100% of fake news while for real news found in the range 20-80%. Positive and negative words represent a measure for identifying the sentiments in the text in terms of intensity and polarity towards emotions [27]. For example, in comparison of two sentences “the person is superb” and “the person is good,” the sentence “the person is superb” is considered more sensitive in sentiment intensity analyzer.

Valence aware dictionary for sentiment reasoning (VADAR) [28], a sentiment lexicon available in Python, was used for sentiment analysis. VADAR considers acronyms, initialism like laugh out loud (LOL), emoticons like ;), or slang like nah as crucial for sentiment analysis. The VADAR provides a compound score for intensity scale between -10 to +10. We computed a percentage of negative, positive, and neutral words in both real and fake news. We also considered other grammatical and linguistic features like VB which represents verb base form (for example take), VBD represent sverb past tense (for example took), VBG represents verb gerund/present participle (for example taking), VBN represents verb past participle (for example taken), VBZ represents verb 3rd person present (for example takes) and CN represents cardinal number.

2.1.2. Features selection

Selection of features is based on correlation attribute evaluation metric and covariance metric, which are computed using correlation between features set and class, and correlation among the features. Co-variance of features over the news items is defined by (1) and (2). Let f_1, f_2, \dots, f_n represents frequency of all linguistic n features for all m news. We computed μ_{real} mean frequency of feature f_i over m_1 real news and μ_{fake} mean frequency of feature f_i over m_2 fake news. We then calculated covariance of each feature for each real and fake news item as shown in (1) and (2) respectively. We then worked on *Corr* metric to filter the appropriate features. Correlation Attribute evaluation metrics are combined to select the features, so the approach is named as *Corr* [28].

Step 1: Correlation value shows how much one variable changes for a slight change in another variable, and covariance is the direction of the linear relationship between variables. In the proposed method, correlation attribute evaluation metric is evaluated between feature and class ($Corr_{fc}$) averaged over k features. The correlation metric is also evaluated ($Corr_{ff}$) with average over k features and the $Corr$ metric is calculated, the features with high relationship values are selected. If these values are higher than the specified threshold assign value, then the feature is effective, and list is computed in descending order. For evaluation of correlation between the features ($Corr$) correlation between the features set and feature class ($Corr_{fc}$) is calculated. If the correlation between features set and its class is strong, it indicates strong correlation between the features set and class. The wrapper technique is applied to filter the features accurately and select effective features for the selected ML algorithms.

In this technique, features are placed in ascending order with respective correlation values. Afterward, a threshold value is assigned, if feature correlation values are higher than a specified threshold assigned value the feature is put forward in the descending order. We observed that features-unique, positive, negative words and CN are having higher correlation with class and among each other rather than noun, adjective words. Here, we combine (1), (2) and (3) and define correlation and covariance attribute evaluation metrics ($CorrCov$ metric) that is presented in (4) and Table 2.

Table 2. Correlation-covariance attribute evaluation metric

Attribute	$\frac{k_{avg}Corr_{fc}}{nrCovar_{fn} + \sqrt{k + k(k-1)avgCorr_{ff}}}$
A1	= 0.6
A2	= 0.4
A3	= 0.3

Step 2: We averaged the co-variance of each feature ($nrCovar_{fn}$) over all news items in the dataset presented in (1) and (2), and after further normalization, the feature was put in a list in descending order. Step 3: Next step is to filter each feature by using the AUC metric of specific ML algorithm. However, the algorithm filters each feature one by one using AUC metric and select those features which give high AUC metric values. The ML algorithms Naive Bayes, decision tree, random forest, K-nearest neighbor, AdaBoost, and support vector machine (SVM) are used to evaluate the AUC metric. Step 4: Final step is verification phase to apply Shannon entropy (using (5)) and technique for order of preference by similarity to ideal solution (TOPSIS) [29], [30] to get desired selected effective feature set in Table 3.

$$Covar_{real_{ithfeature}} = (f_i - \mu_{real_i}) * (f_i - \mu_{real_i}) \quad (1)$$

$$Covar_{fake_{ithfeature}} = (f_i - \mu_{fake_i}) * (f_i - \mu_{fake_i}) \quad (2)$$

$$\frac{k_{avg}Corr_{fc}}{\sqrt{k + k(k-1)avgCorr_{ff}}} \quad (3)$$

$$\frac{k_{avg}Corr_{fc}}{nrCovar_{fn} + \sqrt{k + k(k-1)avgCorr_{ff}}} \quad (4)$$

$$ent = -\ln(n)^{-1} \sum_{i=1}^n A_i \ln(A_i) \quad (5)$$

Table 3. Decision matrix

Attribute	High-A1	Medium-A2	Medium-A3	Low-A1	Very High-A2	High-A3	Low-A1	Very Low-A2	Medium-A3
Writer1	0.7	0.5	0.4	0.3	0.9	0.7	0.3	0.1	0.5
Writer2	0.8	0.4	0.5	0.2	0.8	0.6	0.2	0.2	0.4
Writer3	0.6	0.4	0.5	0.1	0.8	0.7	0.1	0.3	0.5
ent	0.651	0.954	0.954	0.834	0.458	0.715	0.834	0.834	0.954
div	0.349	0.046	0.046	0.166	0.542	0.285	0.166	0.166	0.046
wgt	0.134	0.026	0.025	0.093	0.306	0.566	0.093	0.093	0.026

Different writers have their different writing style of a news while using language attributes (adjectives/adverbs) to write a news. Table 3 is the decision matrix (DM) representing possible different values of selected features. In this research, $A_1 = \text{positive}$, $A_2 = \text{negative}$, $A_3 = \text{unique}$, $A_4 = \text{cardinalnumber}$, $A_5 = \text{variance}$ features are found to be effective. Let the features sentiments, nouns, adjectives have range $high = 0.8 - 0.6$, $medium = 0.5 - 0.4$, and $low = 0.3 - 0.1$ values in various news items then the $DM = [\rho_{ij}]_{a \times b}$ for different writers is shown in Table 3 where a is the number of writers of news items and b is the number of features. Different classifiers may use different informative feature selection criteria and therefore differ in classification with different weight choices presented in Table 4.

Table 4. Representation for one classifier C1

Classifier	High-A1	Medium-A2	Medium-A3	Low-A1	Very High-A2	High-A3	Low-A1	Very Low-A2	Medium-A3	Wgt Choice
C1	1	0	0	0	0	0	0	0	1	0.16
C2	1	1	0	0	0	0	0	0	1	0.186
C3	1	1	1	1	0	0	0	1	1	0.397
wgt	0.134	0.026	0.025	0.093	0.306	0.566	0.093	0.093	0.026	

In (5) and (6) provide a quantization of the attributes. The quantization of different classifiers may be used further for training over the datasets by maximization or minimization as in Table 4. Shannon entropy (1-divergence) is the measure of uncertainty and TOPSIS is a statistical method in Table 5 to give ranking of design alternatives. TOPSIS is applied to choose the best solution on the basis of Euclidean distance, shortest distance from the ideal solution (PIS) and the farthest from the negative ideal solution (NIS) in Table 5. Each classifier measures the distance (Δ_k^*, Δ_k^+) from PIS and NIS; Therefore, the combined separation distance can be given as: $\Delta_k^* = \sqrt{\sum_{i=1}^n \Delta_{ik}^{*2}}$, $\Delta_k^+ = \sqrt{\sum_{i=1}^n \Delta_{ik}^{+2}}$, where $\Delta_k^* = (wgt_{ik} - MAX(wgt_{ik})^2)$ and $\Delta_k^+ = (wgt_{ik} - MIN(wgt_{ik})^2)$. Each classifier measure closeness of each feature to PIS as $\eta_k^* = \frac{\Delta_k^*}{\Delta_k^* + \Delta_k^+}$. Features obtained using Cov-Corr metric are listed in fSet2 in Table 6.

$$wgt = \frac{div_i}{\sum_{k=1}^n div_k} \quad (6)$$

Table 5. Distance from ideal and negative ideal solution

Expert	PIS	NIS
Model 1	0.397	0.16
Model 2	0.697	0
Model 3	0.06	0.1

Table 6. Feature sets

Feature set name	List of features
fSet1	unique, negative, neutral, positive, compound, noun, adjective, adverb, preposition, VB, VBD, VBG, VBN, VBZ, CN, negativeVar, positiveVar, cnVar
fSet2	unique, negative, positive, CN, uniqueVar, negativeVar, positiveVar, cnVar

3. RESULTS AND DISCUSSION

We used two sets fSet1 and fSet2 as presented in Table 6. The fSet1 comprises all features considered to study fake and real news items of four datasets, whereas fSet2 comprises of limited features obtained using Cov-Corr metric. We implemented Naive Bayes, decision tree, random forest, k-nearest neighbors, AdaBoost, SVM algorithms to compare the classification results using fSet1 and fSet2. The scores are obtained by randomly splitting the datasets in the ratio of 0.7:0.3 for the training and cross-validation sets. Figure 1 shows comparison of AUC of algorithms when applied on fSet1. Figure 2 shows comparison of F1 score of algorithms when applied on fSet1. Figure 3 shows comparison of AUC of algorithms when applied on fSet2. Figure 4 shows comparison of F1 score of algorithms when applied on fSet2.

The AUC score is computed from precision-recall curve. We observe high AUC scores for FNC-1 dataset which has more fake news than real news (imbalanced). In Figures 1, 2, 3 and 4, we observe less F1-score ($F1 = \frac{2 * Precision * Recall}{Precision + Recall}$) in comparison to AUC score in the classifications by all the ML algorithms. We observe that positive words, negative words, unique words, and CN are the prominent features for the

fake news detection from a linguistic analysis of text since the correlation and covariance for the features of fSet2 was higher for fake news than real news as resulted in Figures 1 and 2, however, for rest all other features correlation and covariance was similar. The Figure 1 depicts AUC score obtained using the unique words, negative words, positive words, and CN and the variance of the features (fSet2). In fSet1, we used all the features, however, we could achieve comparable performance with a reduced set of the feature set.

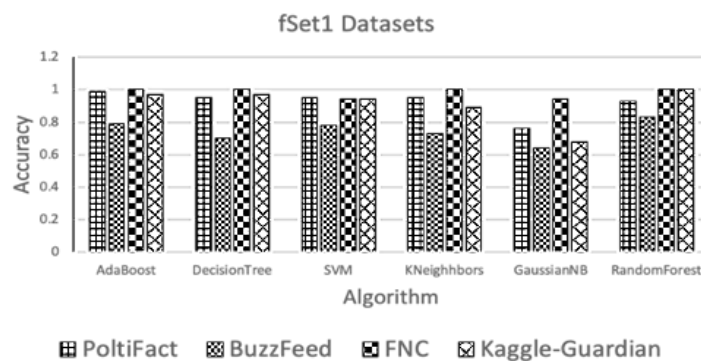


Figure 1. Comparison of AUC score of algorithms on fSet1 with varying datasets

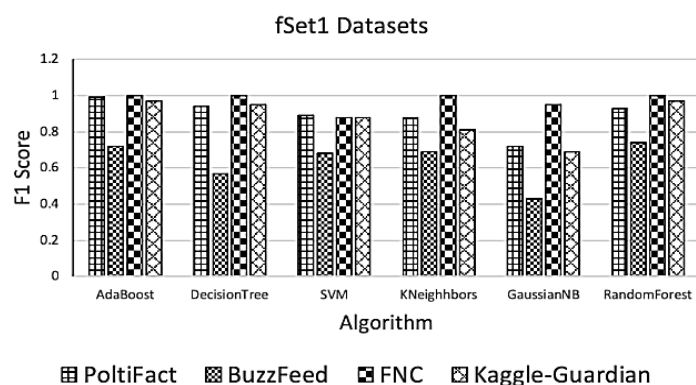


Figure 2. Comparison of F1 score of algorithms on fSet1 with varying datasets

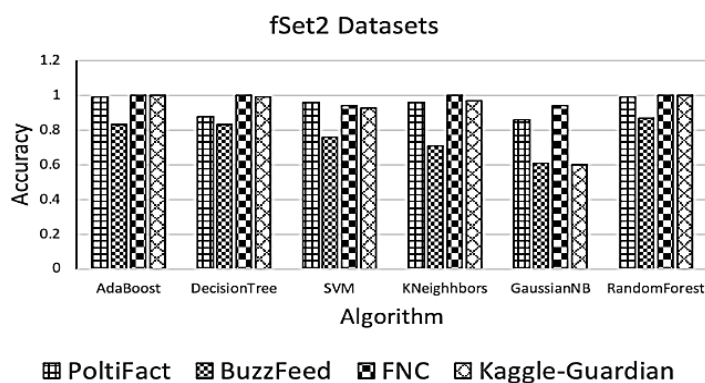


Figure 3. Comparison of AUC score of algorithms on fSet2 with varying datasets

Decision tree algorithm outperformed in comparison to other algorithms except for random forest and AdaBoost. We found that BuzzFeed dataset is the most challenging dataset for all ML algorithms. Decision tree algorithm improved AUC Score from 70% with fSet1 to 83% with fSet2 on BuzzFeed dataset. Gaussian Naive Bayes classifier did not perform well in this particular example of fake/real news

identification for all the datasets. Gaussian Naive Bayes classifier uses statistical information of mean and variance of each feature individually over the dataset and then find the joint conditional probability of all features to find the unique range of values for each class. In the FNC -1 dataset which has the highest AUC score with all algorithms, Naive Bayes classifier obtained the AUC score of 94% for fSet1 and fSet2. However, we obtained AUC score of nearly 100% with fSet2 using AdaBoost with base estimator decision tree classifier as shown in Figure 3. One of the reasons may be that naive assumption of gaussian Naive Bayes may not be true since the number of parts of speech depends on each other.

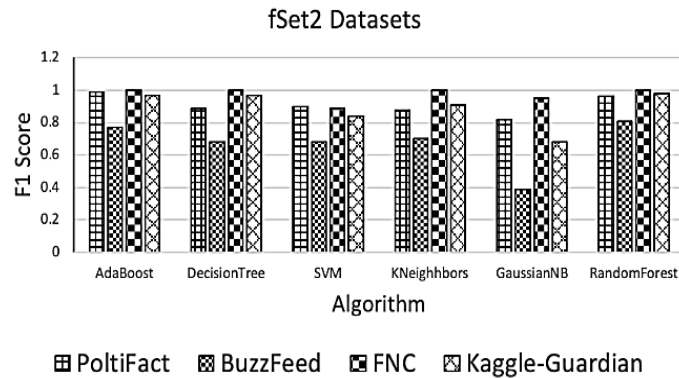


Figure 4. Comparison of F1 score of algorithms on fSet2 with varying datasets

Figure 3 and Figure 4 show that fSet2 outperforms for all the datasets with best performance by Adaboost followed by random forest. There is performance difference of the algorithms on the datasets wherein SVM classifier performed similar for all datasets, including BuzzFeed dataset shown in Figure 3. The SVM classifier is found to be the slowest classifier. We conducted an extensive study to vary all parameters, to find out the best values of hyperparameters for the performance metric. Figures 1, 2, 3 and 4 show F1 and AUC scores for different classifiers and different feature sets fSet1 and fSet2 using the best hyperparameters. We also compared the performance of algorithms when using the best hyperparameters and using the default set of hyperparameters. There was limited performance gain for the algorithms except for random forest classifier and AdaBoost classifier, which improved significantly. The AUC scores of random forest classifier and AdaBoost classifier were improved by 13% and 7% respectively.

Random forest classifier uses subsamples of the feature set to fit into decision tree classifier, and then ensemble obtained trees to predict the class. AdaBoost is a boosting algorithm and is used with weak classifiers. In our example with default parameters, it showed the AUC score of 90% with default estimators as decision tree classifier, learning rate 1 and no of estimators as 50. We increased no of estimators to 400 and improved accuracy by 7% leading to 97% AUC score. Now, we present analysis of datasets:

Imbalanced dataset (FNC-1): We observed that fSet2 outperforms in comparison to fSet1 with $97 \pm 2\%$ AUC score. Even though the dataset is skewed, the performance of ML algorithms is up to the mark for both fSet1 and fSet2. Since feature set fSet2 outperformed in comparison to fSet1, therefore we conclude that even though dataset FNC-1 is imbalanced but the frequency of features (e.g. number of unique words, number of positive sentiments words in the news items) was sufficient to perform the accurate classification. We observed that for this dataset fSet1 (other linguistic features in the fake news items) performance is also significant enough due to large numbers of news items (46,293 fake news+3,678 real news) with repeated information for ML algorithms to capture the features from real news and fake news. We observed biased predictions due to imbalance news items in few cases (e.g. the model predicted fake news items with higher accuracy than real news items) and therefore this example presents a scenario of limitation of ML algorithms in avoiding automation of bias [31].

Limited size datasets (PolitiFact and BuzzFeed): The classifiers resulted in low AUC score in comparison to other two datasets. The AUC Score with fSet2 feature set for Buzzfeed is $74 \pm 5\%$. The AUC Score for PolitiFact with fSet2 feature set is $90 \pm 10\%$. Results show that even the datasets are in limited size but the frequency of features in the news is significantly enough therefore the same feature set fSet2 outperformed for the datasets. PolitiFact dataset is better even with fSet1 even though limited in number of real and fake news items.

Balanced dataset (Kaggle-Guardian): Feature set fSet2 in comparison to fSet1 improved the performance for this dataset up to 91 ± 14 % AUC score. Results show that even the dataset is balanced but the frequency of features in the dataset is comparable, therefore, the same feature set fSet2 outperformed for the dataset but with less AUC score than FNC-1. This dataset is difficult for classifiers (less AUC score in comparison to others) though it is balanced.

4. CONCLUSION

A study on feature sets over four fake news datasets using ML algorithms conclude that feature set fSet2 is the reduced feature set over the fSet1 since random forest, AdaBoost, k-nearest neighbor, and SVM classifiers obtained high AUC score for fSet2 in comparison to fSet1. The fSet2 is computed using covariance and correlation attribute evaluation metric. The four datasets considered under study were having different proportion of real and fake news items. Thus, the proposed approach has been tested for limited size, imbalanced and balanced datasets. Fake news can be written in regional languages used across the globe to spread the distrust among the local public. Detecting fake news for the regional content is challenging since regional languages have different linguistic features with limited availability of datasets. Future work is proposed over language features for regional languages.





REFERENCES

- [1] K. D. Stephan and G. Klima, "Artificial intelligence and its natural limits," *AI & SOCIETY*, vol. 36, no. 1, pp. 9–18, May 2021, doi: 10.1007/s00146-020-00995-z.
- [2] K. Shu, S. Wang, and H. Liu, "Beyond news contents: the role of social context for fake news detection," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, Jan. 2019, pp. 312–320, doi: 10.1145/3289600.3290994.
- [3] X. Zhou and R. Zafarani, "Fake news: A survey of research, detection methods, and opportunities," *arXiv preprint*, Dec. 2018, doi: 10.1145/3395046.
- [4] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," *arXiv preprint*, Feb. 2019, doi: 10.48550/arXiv.1902.06673.
- [5] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: a theory-driven model," *arXiv preprint*, Apr. 2019.
- [6] N. Aneja and S. Aneja, "Detecting fake news with machine learning," in *International Conference on Deep Learning, Artificial Intelligence and Robotics, (ICDLAIR)*, Springer International Publishing, 2019, pp. 53–64.
- [7] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, "Truth of varying shades: analyzing language in fake news and political fact-checking," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 2931–2937, doi: 10.18653/v1/d17-1317.
- [8] W. Ferreira and A. Vlachos, "Emergent: a novel data-set for stance classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1163–1168.
- [9] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, May 2017, doi: 10.1257/jep.31.2.211.
- [10] B. D. Horne and S. Adali, "This just in: fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," *arXiv preprint*, Mar. 2017, [Online]. Available: <http://arxiv.org/abs/1703.09398>.
- [11] R. A. Bagate and R. Suguna, "Sarcasm detection of tweets without #sarcasm: data science approach," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 2, pp. 993–1001, Aug. 2021, doi: 10.11591/ijeecs.v23.i2.pp993-1001.
- [12] A. Pardamean and H. F. Pardede, "Tuned bidirectional encoder representations from transformers for fake news detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1667–1671, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1667-1671.
- [13] K. M. Fouad, S. F. Sabbeh, and W. Medhat, "Arabic Fake News Detection Using Deep Learning," *Computers, Materials, & Continua*, vol. 71, no. 2, pp. 3647–3665, 2022, doi: 10.32604/cmc.2022.021449.
- [14] P. Mookdarsanit and L. Mookdarsanit, "The covid-19 fake news detection in thai social texts," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 988–998, Apr. 2021, doi: 10.11591/eei.v10i2.2745.
- [15] G. Xiaoning, T. De Zhern, S. W. King, T. Y. Fei, and L. H. Shuan, "News reliability evaluation using latent semantic analysis," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 16, no. 4, pp. 1704–1711, Aug. 2018, doi: 10.12928/telkomnika.v16i4.9062.
- [16] S. Senhadji and R. A. S. Ahmed, "Fake news detection using naïve Bayes and long short term memory algorithms," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 2, pp. 748–754, Jun. 2022, doi: 10.11591/ijai.v11.i2.pp748-754.
- [17] R. Zellers et al., "Defending against fake news," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [18] T. Le, S. Wang, and D. Lee, "Malcom: generating malicious comments to attack neural fake news detection models," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 282–291, doi: 10.1109/icdm50108.2020.00037.
- [19] N. Aneja and S. Aneja, "Transfer learning using CNN for handwritten devanagari character recognition," in *2019 1st International Conference on Advances in Information Technology (ICAIT)*, Jul. 2019, pp. 293–296, doi: 10.1109/ICAIT47043.2019.8987286.
- [20] S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, "Transfer learning for cancer diagnosis in histopathological images," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, pp. 129–136, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp129-136.
- [21] S. Aneja, N. Aneja, B. Bhargava, and R. R. Chowdhury, "Device fingerprinting using deep convolutional neural networks," *International Journal of Communication Networks and Distributed Systems*, vol. 28, no. 2, pp. 171–198, 2022, doi: 10.1504/IJCND.2022.121197.
- [22] S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, "Defense against adversarial attacks on deep convolutional neural networks through nonlocal denoising," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 3, pp. 961–968, 2022, doi: 10.11591/ijai.v11.i3.pp961-968.
- [23] S. Aneja, M. A. X. En, and N. Aneja, "Collaborative adversary nodes learning on the logs of IoT devices in an IoT network," in





- 2022 14th International Conference on COMMunication Systems & NETworks (COMSNETS), Jan. 2022, pp. 231–235, doi: 10.1109/COMSNETS53615.2022.9668602.
- [24] A. Maronikolakis, H. Schutze, and M. Stevenson, “Transformers are better than humans at identifying generated text,” *arXiv preprint*, Sep. 2020, [Online]. Available: <http://arxiv.org/abs/2009.13375>.
- [25] R. Tan, B. Plummer, and K. Saenko, “Detecting cross-modal inconsistency to defend against neural fake news,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 2081–2106, doi: 10.18653/v1/2020.emnlp-main.163.
- [26] J. Zeng, Y. Zhang, and X. Ma, “Fake news detection for epidemic emergencies via deep correlations between text and images,” *Sustainable Cities and Society*, vol. 66, p. 102652, Mar. 2021, doi: 10.1016/j.scs.2020.102652.
- [27] T. A. Tran, J. Duangsuwan, and W. Wettayaprasit, “A new approach for extracting and scoring aspect using SentiWordNet,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 3, pp. 1731–1738, Jun. 2021, doi: 10.11591/ijeecs.v22.i3.pp1731-1738.
- [28] C. Hutto and E. Gilbert, “Vader: a parsimonious rule-based model for sentiment analysis of social media text,” in *Proceedings of the International AAAI Conference on Web and Social Media*, 2014, vol. 8, no. 1.
- [29] M. Shafiq, Z. Tian, A. K. Bashir, X. Du, and M. Guizani, “CorrAUC: a malicious Bot-IoT traffic detection method in IoT network using machine learning techniques,” *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3242–3254, Mar. 2021, doi: 10.1109/jiot.2020.3002255.
- [30] T.-C. Wang and H.-D. Lee, “Developing a fuzzy TOPSIS approach based on subjective weights and objective weights,” *Expert Systems with Applications*, vol. 36, no. 5, pp. 8980–8985, Jul. 2009, doi: 10.1016/j.eswa.2008.11.035.
- [31] D. Varona, Y. Lizama-Mue, and J. L. Suárez, “Machine learning’s limitations in avoiding automation of bias,” *AI & SOCIETY*, vol. 36, no. 1, pp. 197–203, Jun. 2020, doi: 10.1007/s00146-020-00996-y.

BIOGRAPHIES OF AUTHORS







Sandhya Aneja     is working as Assistant Professor of Information and Communication System Engineering at the Faculty of Integrated Technologies, Universiti Brunei Darussalam. Her primary areas of research interest include wireless networks, high-performance computing, internet of things, artificial intelligence technologies, machine learning, machine translation, deep learning, data science, and data analytics. Further info on her website <https://sandhyaaneja.github.io>. She can be contacted at email: sandhya.aneja@gmail.com.



Nagender Aneja     is working as Assistant Professor at School of Digital Science, Universiti Brunei Darussalam. He did his Ph.D. in Computer Engineering from J.C. Bose University of Science and Technology YMCA, and M.E. Computer Technology and Applications from Delhi College of Engineering. He is currently working in the area of deep learning, computer vision, and natural language processing. He is also founder of ResearchID.co. Further info on his website <http://naneja.github.io>. He can be contacted at email: naneja@gmail.com.



Ponnurangam Kumaraguru     is a Professor of Computer Science and Dean of Students Affairs at IIIT-Hyderabad. PK is a TEDx and an ACM Distinguished and ACM India Eminent Speaker. PK received his Ph.D. from the School of Computer Science at Carnegie Mellon University (CMU). His Ph.D. thesis work on anti-phishing research at CMU contributed in creating an award-winning startup-Wombat Security Technologies, wombatsecurity.com. Wombat was acquired in March 2018 for USD 225 Million. He can be contacted at email: pk.guru@iiit.ac.in.

Materials Advances

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: M. Chaudhary, A. Kumar, A. Devi, B. P. Singh, B. D. Malhotra, K. Singhal, S. Shukla, S. Ponnada, R. K. Sharma, C. A. Vega-Olivencia, S. Tyagi and R. Singhal, *Mater. Adv.*, 2022, DOI: 10.1039/D2MA00896C.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

Prospects of Nanostructure-based Electrochemical Sensors for Drug

Detection: A Review

Manika Chaudhary¹, Ashwani Kumar², Arti Devi¹, Beer Pal Singh^{1*}, Bansi D. Malhotra³, Kushagr Singhal⁴, Sangeeta Shukla¹, Srikanth Ponnada⁵, Rakesh K Sharma⁵, Carmen A Vega-Olivencia⁶, Shrestha Tyagi¹, Rahul Singhal^{7*}

¹Department of Physics, Chaudhary Charan Singh University, Meerut 250004, India.

²Nanoscience Laboratory, Institute Instrumentation Centre, IIT Roorkee, Roorkee, 247667, India.

³Department of Biotechnology, Delhi Technological University, Shahbad Daultpur, Main Bawana Road, Delhi, India.

⁴Rocky Hill High School, Rocky Hill, CT 06067, USA

⁵Sustainable Materials and Catalysis Research Laboratory (SMCRL), Department of Chemistry, Indian Institute of Technology Jodhpur, Karwad, Jodhpur-342037, India.

⁶Department of Chemistry, University of Puerto Rico, Mayaguez, PR 00681-9000, USA.

⁷Department of Physics and Engineering Physics, Central Connecticut State University, New Britain, Connecticut 06050, USA.

***Corresponding authors:**

1. Rahul Singhal

singhal@ccsu.edu,

2. Beer Pal Singh

drbeerpal@gmail.com



Abstract

The present study represents the advancements achieved over the last ten years towards the development of electrochemical sensors based on nanomaterials. The versatility, sensitivity, selectivity, and capability to analyze samples with minimal to no pre-treatment has created electrochemical sensors an attractive and powerful tool for detecting some analgesic and antipyretic drugs such as acetaminophen (AP), ibuprofen (IB), aspirin (ASP), and diclofenac (DCF). These analgesic and antipyretic drugs are very popular for minor pain and fever medications. The controlled doses of these drugs do not harm the human body, but their higher concentration can be hazardous for humans. These drugs are also considered as emerging chemical pollutants in the environment. Reliable and powerful analytical techniques are thus necessary for the detection of these drugs for quality control of pharmaceuticals as well as environmental control. This review emphasizes the synthesis of nanostructured materials and their electrochemical sensing of analgesic and antipyretic drugs.

Keywords: Electrochemical sensors; acetaminophen; differential pulse voltammetry; NSAIDs; nanomaterials.



Author Biographies



Manika Chaudhary has received her M.Sc. in 2016 and M.Phil. in 2017 from C.C.S University, Meerut (U.P), India. Presently, she is pursuing Ph.D from C.C.S University, Meerut (U.P), India. She has published six research papers and one book chapter in reputed journals. Her research interests comprise of nanostructured materials and metal oxide semiconducting materials for energy storage devices and sensors.



Beer Pal Singh has received his M.Sc. (1997), M.Phil. (1998) and Ph.D. (2002) from C.C.S. University, Meerut (UP), India. He is holding faculty position in Physics at C.C.S. University, Meerut since 2004. Presently, he is working as "Professor" and Head in Department of Physics, CCS University, Meerut. In addition to this, he is also holding the post of Proctor, Security Officer and Dy. Director, Centre for International Cooperation in university administration. Recently, he had worked as Visiting Professor in Tokyo University of Science, Tokyo, Japan and Visiting Scientist (Raman Fellow) in University of Puerto Rico, Mayaguez, PR, USA for one year. He has also visited Germany, France, China, and Boston and presented his research work. He has supervised 09 Ph.D. and more than 40 M.Phil. Students for their research thesis. He has published more than 60 research papers in reputed journals and serving as a reviewer of several national/international journal of repute. His research interests comprise of thin films, 2D materials, nanostructured materials, metal oxides, semiconducting materials, thin film transistors, sensors and energy storage devices.



Dr. B.D. Malhotra received his PhD from the University of Delhi, Delhi in 1980. He has published **340 papers** in refereed international journals (**Citations:25883 Research-index: 85**), has filed 11 patents (in India and overseas), and has co-authored text books on 'Nanomaterials for Biosensors: Fundamentals and Applications' and 'Biosensors: Fundamentals and Applications'. He is a recipient of the National Research Development Corporation Award 2005 for invention on 'Blood Glucose Biochemical Analyzer' and is a Fellow of the Indian National Science Academy, the National Academy of Sciences, India and an Academician of the Asia Pacific Academy of Materials (APAM).. Dr Malhotra is a former **DST-Science & Engineering Research Board (SERB, Govt of India) Distinguished Fellow**.





Dr. Rakesh K Sharma, FRSC is an Associate Professor at the Department of Chemistry at IIT Jodhpur, India. He received his BSc and MSc from the University of Rajasthan Jaipur and Ph.D. from the Indian Institute of Science Bangalore in 2008. He worked as a postdoctoral researcher from 2007 to 2010 at the Ohio State University Columbus, USA. He has nine patents and transferred five technologies in bio-fuel, energy storage, environmental technologies, and automotive applications. He has published over 100 articles in peer-reviewed journals, including *Journal of American Chemical Society*, *Chemical Science*, *ACS Sustainable and Engineering*, to name a few. He has also published ten books/book chapters. His research interest includes catalysis for biofuels and fine chemicals, natural clay catalyst, plasma catalysis for environmental remediation, and advanced materials for energy generation and storage.



Dr. Rahul Singhal is an Associate Professor at the Department of Physics and Engineering Physics at Central Connecticut State University, New Britain, CT, USA. He received his M.Sc. degree in physics from G.B. Pant University Of Agriculture And Technology, Pantnagar in 1997 and Ph.D. degree in physics from University of Delhi, India in 2003. He worked as researcher from 1997 – 2005 at National Physical Laboratory, New Delhi, India. He worked as postdoctoral fellow from 2005-2012 at various universities in USA, including University of Puerto Rico, Mayaguez; University of Puerto Rico, San Juan; University of South Florida, Tampa, FL. He published over 60 articles in peer-reviewed journals and 4 book chapters. His research interest includes biosensors based on conducting polymers, metal oxide/sulfide nanomaterials, and materials for energy storage devices such as super capacitors and rechargeable batteries.



Kushagr Singhal is currently a High-school Junior at Rocky Hill High School, CT, USA. He is actively engaged in the research related to synthesis and characterizations of transition metal oxide nanomaterials for electrochemical energy storage devices. His interests are in data analysis, finance, and materials science & engineering.



List of Abbreviations used

Abbreviations	
AP	Acetaminophen
IB	Ibuprofen
DCF	Diclofenac
ASP	Aspirin
NSAIDs	Non-steroidal anti-inflammatory drug
NPs	Nanoparticles
CV	Cyclic voltammetry
ASV	Anodic stripping voltammetry
LSV	Linear sweep voltammetry
SWV	Square wave voltammetry
DPV	Differential pulse voltammetry
AA	Ascorbic acid
DA	Dopamine
CNT	Carbon nanotubes
SWCNT	Single walled carbon nanotubes
MWCNT	Multiwalled carbon nanotubes
RT	Room temperature
LOD	Limit of detection
SEM	Scanning electron microscope
FESEM	Field emission electron microscopy



TEM	Transmission electron microscopy
XRD	X-ray diffraction
FTIR	Fourier transform infrared radiation
CVD	Chemical vapor deposition
GO	Graphene oxide
rGO	reduced Graphene oxide
GCE	Glassy Carbon Electrode
Ag	Silver
Au	Gold
GQDs	Graphene quantum dots
UA	Uric acid
NiO	Nickel oxide
ZnO	Zinc oxide
CA	Chronoamperometry

1. Introduction

In present times, drugs have become a part of our daily lives because of their therapeutic and recreational purposes [1]. Drugs can be categorized in three groups: (a) therapeutic drugs, (b) legal drugs, and (c) illicit drugs. Therapeutic drugs are those, which are generally prescribed by doctors for treatment of diseases, for example: theophylline to treat lung diseases and propofol to induce anesthesia during surgical procedures. Legal drugs such as alcohol, caffeine, and nicotine are generally consumed for recreational use in commercial products. These drugs, when consumed, provide psychoactive effects in the body. The third type of drugs; illicit drugs are consumed for



recreational purposes, which harm the central nervous system and thus invite various prolonged health issues [2, 3]. Though the last category was developed principally for pharmaceutical purposes, its therapeutic use was soon overshadowed by its potential for misuse. For instance, in 1898, Bayer first used 'Heroin' as a new constituent in a cough medicine [3, 4]. The utilization of therapeutic drugs is continually expanding, with a projected worldwide expenditure of 1.52 trillion US dollars by 2023 [5]. These sudden increasing trends reflect essentially the aging total populace and the spread of new infections and pandemics, for example; the ongoing novel COVID-19 [6, 7]. Therapeutic drugs are of various types such as anesthetics, antibiotics, analgesics, cardioactive drugs, antineoplastic drugs, and etc. There are various analgesic and antipyretic drugs which are consumed by humans in their daily life, such as acetaminophen (paracetamol), ibuprofen, aspirin, diclofenac, ascorbic acid, etc. Acetaminophen (AP) (paracetamol or N-acetyl-p-aminophenol) is a popular, safe, effective, and extensively used analgesic and antipyretic drug. Acetaminophen is used as fever reducer, and to relieve pain associated with various parts of the body, such as headache, arthralgia, cancer pain, neuralgia and pain associated to any surgical treatment [8, 9]. In spite of the fact that AP is a relatively secure medicine, overdose of any drug may have harmful effects [10]. The excess dose of AP may cause nephrotoxicity and lethal hepatotoxicity [11]. Ibuprofen (IB) falls in the category of painkillers and anti-pyretic drugs. This drug blocks the enzyme cyclooxygenase, thus inhibiting prostaglandin biosynthesis. Approximately 90% of ibuprofen is converted to hydroxyl and carboxyl metabolites of ibuprofen in the liver with the remaining 10% passed unchanged in urine and bile [12, 13]. Ibuprofen is high-selling drug worldwide, which is why it is usually the first option for different short-term non-specific indications. It is most commonly used to treat fever symptoms, headaches, arthritis, and a variety of other common aches and pains. The ease of availability and popularity of ibuprofen makes it

one of the most commonly detected and quantified drugs in pharmaceutical analysis. Novel and progressive analytical approaches with high accuracy are required for strict control of these drugs in pharmaceutical dosages and different biological fluids [14, 15]. In the family of analgesic drugs, aspirin (ASP), also known as acetylsalicylic acid, is consumed in minor pain and to reduce fever. It can also be used as blood thinner. Aspirin decomposes quickly in solutions of ammonium acetate or the acetates, citrates, carbonates, or hydroxides of alkali metals. It is stable in dry air but when it comes in contact with moisture, it slowly hydrolysis to acetic and salicylic acids. It has a adverse effect on stomach which involves stomach ulcers, stomach bleeding, and worsening asthma [16-18]. Diclofenac (DCF) is assigned as 2-(2-((2,6-dichlorophenyl) amino) phenyl) acetic acid and is a commonly prescribed drug in analgesic and antipyretic drugs because it has strong antipyretic, analgesic, and anti-inflammatory properties [19]. It is efficient in acute joint inflammation, rheumatic complaints, and mild to moderate pain [20]. If it is taken in normal therapeutic doses, it is safe and does not have toxic effects on the human body. A high dose of DCF may cause negative effects such as gastrointestinal disorders, aplastic anemia, and disturbs in renal function [21, 22]. Generally, all these drugs (except AP, because it has low anti-inflammatory properties) fall in the category of Nonsteroidal anti-inflammatory drugs (NSAIDs), which are frequently used to treat fever, pain, and control inflammation. But overdose of any drug may cause severe problems to the human body, therefore, the precise detection of pharmaceutical specimens is useful for quality control of medication, avoiding major risks to humans. In this context, various analytical techniques such as titrimetric [23], spectrofluorometric [24], chemiluminescence [25], liquid chromatography [26], spectrophotometry [27], and electrochemical analysis [28, 29] have been proposed for the determination of the concentration of different drugs. In titrimetric, spectrophotometric, and chemiluminescence techniques, extraction process is needed before the



detection, whereas liquid chromatography is a time taking process, making these strategies incongruous for routine examination. This makes the requirement of a chemical sensor to detect various drugs, which is fast, precise, reliable, and cost effective. An outline of analytical chemistry development displays that electrochemical sensors constitute the foremost and fast-developing class of chemical sensors. It delivers continuous data about the presence of chemicals in its surroundings. Ideally, a chemical sensor offers an explicit kind of response which is directly connected to the amount of a selected chemical species. Oxidation or reduction of analyte in the electrolyte is the fundamental principle of electrochemical sensors. The changes in electrical parameters resulting from the redox reaction are then measured. The cyclic voltammetry (CV), anodic stripping voltammetry (ASV), linear sweep voltammetry (LSV), and differential pulse voltammetry (DPV) are the most commonly used techniques in electrochemical sensors. Among all of these techniques, the electrochemical method has many advantages including simplicity, affordability, rapidity, ease of monitoring, and high sensitivity, which play an important role in pharmaceutical analysis. **Figure 1** shows electrochemical sensors based on nanomaterials for the detection of various drugs. These drugs are an electroactive compound that can be electrochemically oxidized. Electrochemical sensors show a captivating choice for its fast determination and measurement. On conventional materials, the electrochemical oxidation of these drugs is an irreversible process. However, it becomes reversible due to the existence of catalytic compounds, such as metallic particles, carbon-based nanomaterials and conductive polymers [29-32].

With noticeable accomplishments in nanoscience and nanotechnology, nanomaterial-based electrochemical signal amplifications have acquired incredible capability of improving both selectivity and sensibility for electrochemical sensors. It is broadly known that the electrode



materials play a crucial role in the development of superior electrochemical sensing platforms for distinguishing target molecules through different analytical principles. Furthermore, useful nanomaterials not just produce a synergic impact among conductivity, biocompatibility, and catalytic activity to speed up the signal transduction, additionally enhance bio recognition events with explicitly designed signal labels, leading to highly sensitive bio sensing [33, 34]. In 2018's report, Montaseri et al. discussed the different detection techniques, such as chromatography-mass spectrometry, spectroscopic method, capillary electrophoresis method, and electrochemical method, for acetaminophen and focused on water treatment and toxicity of acetaminophen [35]. Recently, Qian et al. published an article in which they discussed that carbon based materials and noble metal nanomaterials play a crucial role in drug sensing due to their high surface to volume ratio and high electrical conductivity [36]. Li et al. dispersed the Pd nanoparticles on the GO sheet and prepared Pd/GO nanocomposite. The prepared nanocomposite based sensor showed excellent reproducibility and stability with wider linear concentration ranges (0.005–0.5 μM and 0.5– 80 μM) and low detection limit (2.2 nM) towards the sensing of paracetamol [37]. Adekunle et al. prepared an electrochemical sensor by using edge-plane pyrolytic graphite electrode (EPPGEs) and modified it with SWCNT–iron (III) oxide (SWCNT/ Fe_2O_3) nanoparticles for the detection of dopamine [38]. Ozcelikay et al. developed a sensor for the detection of daptomycin by using integration of Au decorated Pt nanoparticles onto the nanocomposite thin film which showed the higher sensitivity [39]. So, the main motive of this report is to deliver a general analysis of electrochemical sensors and their sensing capability for some previously discussed analgesic and antipyretic drugs by using different nanostructured materials. This report also discusses some common methods to synthesize the nanomaterials and their use in electrochemical sensors for different drugs detection.



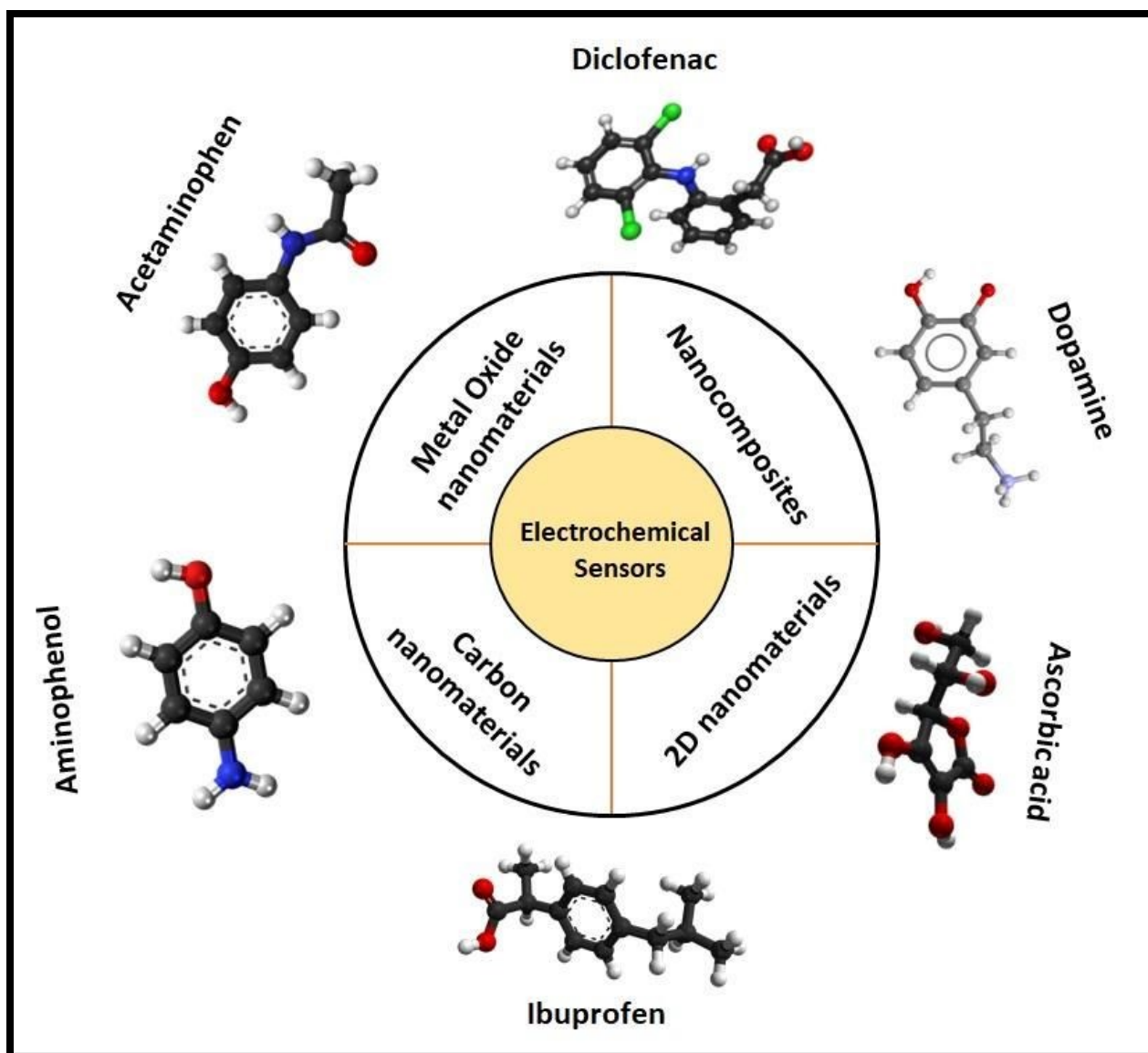


Figure: 1. Schematic of nanomaterials based electrochemical sensors for the detection of different drugs.

2. Fabrication of nanostructured materials

2.1 Hydrothermal or Solvothermal techniques

Hydrothermal method is an easy and efficient way for the fabrication of nanomaterials. With this process, single and multi-component metal oxides based nanoparticles can be produced with high purity. The crystal growth process is executed in an apparatus called autoclave, in which precursors



are supplied with water. A temperature difference is sustained at the opposite terminals of the growth chamber so that the cooler end causes seeds to take additional growth and the hotter end dissolves the nutrient. The main superiority of this technique is the synthesis of fine quality crystals with controlled structure like shape and size. But on the other hand, this process is difficult to control and there is a limitation of reliability and reproducibility [40, 41]. General synthesis procedure of this method is shown in **Figure 2**. Annadurai et al. reported their work on hydrothermally prepared nickel oxide (NiO) modified glassy carbon electrodes. The morphology reveals that the size of prepared NiO nanoparticles varies between 15 - 20nm. These NiO modified electrodes were utilized for the sensitive detection of 4-acetaminophen by DPV, CV, and chronoamperometry (CA) techniques [42]. Nurzulaikha and coworkers prepared the modified electrode of graphene/SnO₂ nanocomposite, synthesized via hydrothermal route. They used the fabricated electrodes for the detection of dopamine in the presence of ascorbic acid (AA). The electrode manifested good selectivity, sensitivity and limit of detection in the presence of AA [43]. Phosphorus-doped graphene was hydrothermally prepared by the Zhang group and was utilized as the electrode material in order to fabricate electrochemical sensors for AP sensing [44]. Xu et al. synthesized 3,4 ethylene dioxythiophene (PEDOT)-MnO₂ nanocomposite via hydrothermal method and used them for amperometric detection of paracetamol [45]. Recently, Ponnada et al. reported synthesis of Ag-Cu decorated ZnO nano flower like composite (NFLC) via single step hydrothermal method and found suitable for the detection of dopamine with high sensitivity of 0.68 $\mu\text{A mM}^{-1} \text{cm}^{-2}$ and low detection limit of 0.21 μM , detected via DPV method [46].



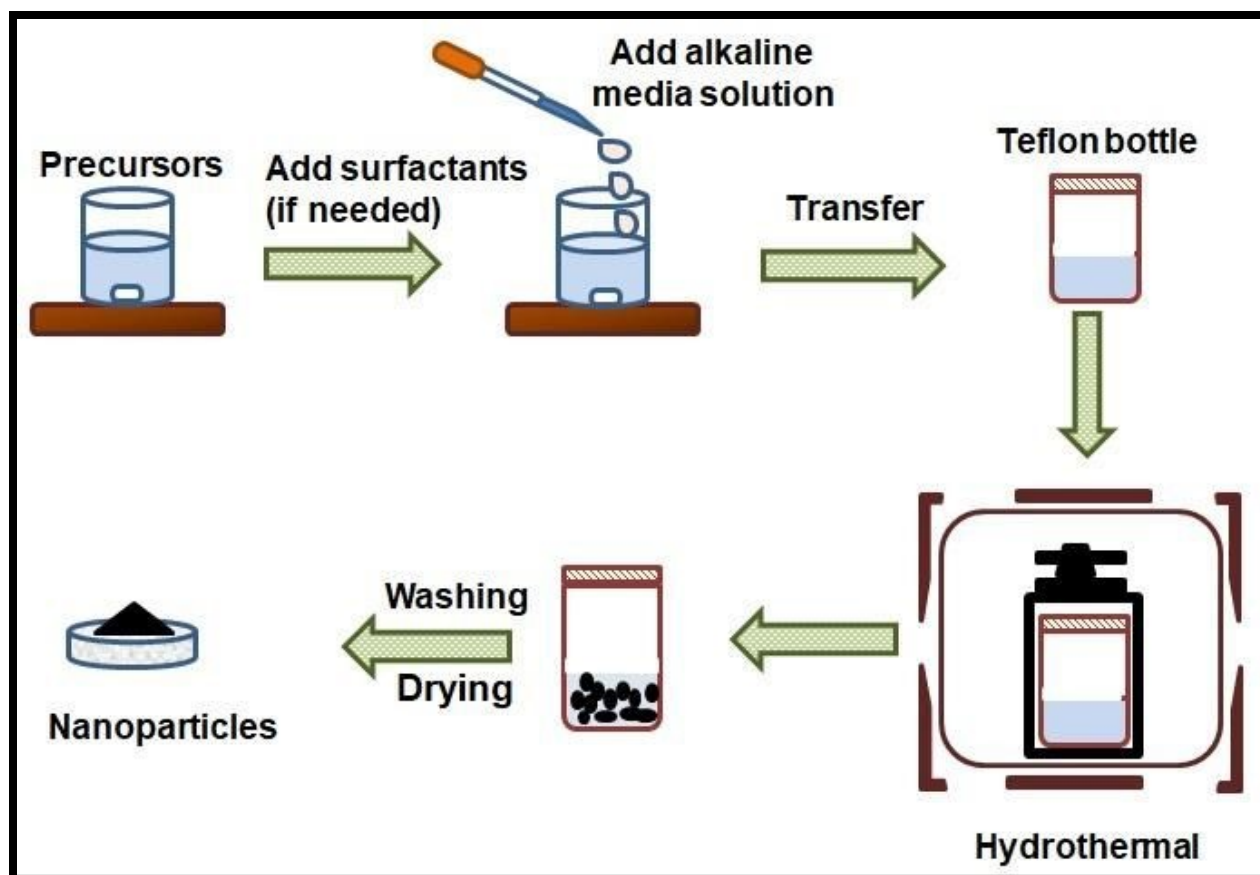


Figure 2. Schematic of a typical hydrothermal method to synthesize nanomaterials.

Lu and co-workers constructed an innovative sensing system by modifying glassy carbon electrodes with biomass carbon/metal organic frameworks derived $\text{Co}_3\text{O}_4/\text{FeCo}_2\text{O}_4(\text{BC}/\text{Co}_3\text{O}_4/\text{FeCo}_2\text{O}_4)$ composite. The $\text{BC}/\text{Co}_3\text{O}_4/\text{FeCo}_2\text{O}_4$ composite was prepared by a simple in-situ growth process and calcination treatment attached with hydrothermal treatment. First the pinecones were carbonized into BC materials and then Co-based zeolitic imidazolate framework ZIF-67 grew in situ on a porous BC matrix; the product was used as a precursor material. Afterwards, the precursor was then pyrolyzed to generate $\text{BC}/\text{Co}_3\text{O}_4$ nanocomposites, and then $\text{BC}/\text{Co}_3\text{O}_4/\text{FeCo}_2\text{O}_4$ composite was synthesized via hydrothermal method and calcinations. It was reported that ZIF-67 crystals show rhombic dodecahedral shapes [Figure 3A] and the particle sizes were found in the range of 400-600 nm. The FeCo_2O_4 powders



showed the nanorod like morphology in the diameter range of 100-150 nm [Figure 3B] while the FESEM images of biomass carbon and BC/Co₃O₄ are shown in Figure 3C & 3D, respectively. It was also observed that upon BC/Co₃O₄/FeCo₂O₄ composite formation [Figure 3E and 3F], the rhombic dodecahedral morphology of MOF-derived Co₃O₄ growing on the BC surface turned into nanosheets. The electrochemical sensor developed by the prepared composite exhibited superior electro-conductivity and vast active surface area because of the synergy effects of BC, MOF-derived Co₃O₄ and FeCo₂O₄ [47]. Razmi et al. synthesized a Fe₃O₄/MWCNT nanocomposite by using hydrothermal method and then prepared TiO₂/Fe₃O₄/MWCNT nanocomposite by following sol-gel method. They developed a sensor with the prepared nanocomposite and used it for the detection of morphine and diclofenac simultaneously [48].

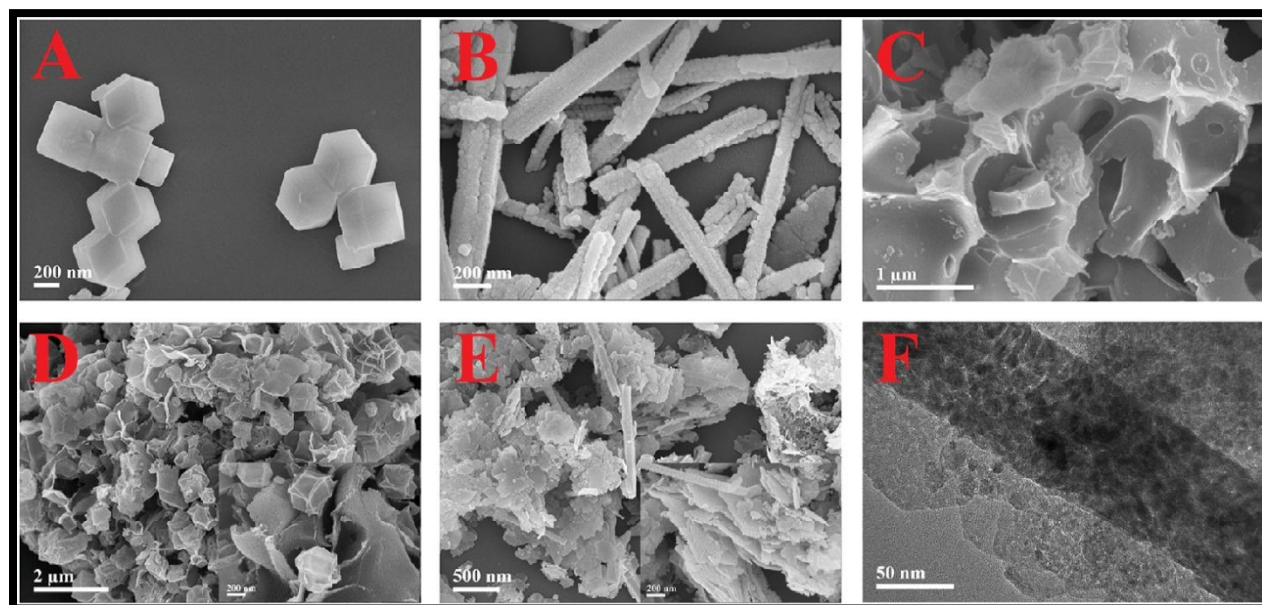


Figure 3. (A) SEM micrographs of ZIF-67 (B) FeCo₂O₄ , (C) BC (D) BC/Co₃O₄ and (E) BC/Co₃O₄/FeCo₂O₄ composite and (F) the TEM image of BC/Co₃O₄/FeCo₂O₄ composite (Reprinted with permission from ref. [47]).



2.2 Sol-gel method:

This method is very beneficial for the synthesis of composites, oxides, and hybrids of organics and inorganics. In the sol-gel method, firstly, the metal alkoxide solution undergoes hydrolysis with water or alcoholic solution in presence of acid or base, followed by polycondensation. Owing to polycondensation, the water molecules were removed, and the liquid phase was transformed into the gel phase, which increased the viscosity of the solution. Thereafter, condensation of water molecules takes place, and the gel phase changes into the powder phase. An extra heat is essential for obtaining fine crystalline powder. A sol-gel method basically uses inorganic polymerization reactions. The superiority of this method is that it's an easy process for the creation of superfine porous powder [49, 50]. Bagherinasab et al. reported synthesis of $\text{BaFe}_{12}\text{O}_9$ by sol gel method in which citric acid, $\text{Ba}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$, $\text{Fe}(\text{NO}_3)_3 \cdot 9\text{H}_2\text{O}$, and NH_3 were utilized as starting materials. The FE-SEM results showed hexagonal morphology of $\text{BaFe}_{12}\text{O}_{19}$ nanoparticles with particle mean size as 76 nm [51]. Deiminit et al. succeeded to make an electrochemical imprinted sensor by combining functionalized multiwall carbon nanotubes and lean molecularly imprinted film for the detection of dopamine. First, they functionalized multiwalled CNT using nitric acid and carboxylic acid then deposited these functionalized multiwalled carbon nanotubes (f-MWCNTs) on glassy carbon electrodes. Later, they deposited imprinted film on the assembled f-MWCNTs layer using sol-gel method. The sol solution was synthesized by mixing 75 μL of PTEOS, 75 μL of TEOS, 700 μL of water, 1100 μL of ethanol and 10 μL of TFA. All these chemicals were added to tramadol in a vial and the solution was stirred to get homogeneous sol at room temperature (RT) for 2 hrs. Thereafter, pyrrole solution (50 μL) and lithium perchlorate (5.0 mg) were added to the mixture and subsequently sonicated for 10 min. Then, the polypyrrole@sol-gel MIP/f-MWCNTs/GC electrode were fabricated by applying CV between -0.8 V and $+0.8\text{ V}$ (versus



Ag/AgCl) for 10 cycles having scan rate (50 mV s^{-1}) in imprinted sol-gel solution, which displayed a dense and uniform morphology[52]. Similarly, sol-gel imprinted polymer based electrochemical sensors for recognition and detection of paracetamol was synthesized by Zhu et al.[53]. A sol-gel fabricated Mn_2O_3 - TiO_2 decorated graphene electrode was reported for quick and selective ultra-sensitive dopamine sensing [54]. Luo and coworkers used one-pot synthesis of a graphene oxide molecularly imprinted polymer sol-gel for electrochemical sensing of paracetamol, which possessed wide detection range, high selectivity, and low LOD along with good stability [55]. Rouhani et al. fabricated an electrochemical sensor based on modified imprinted sol-gel graphite electrode with gold nanoparticles, multiwalled CNT, and Preysslerheteropolyacid [56].

2.3 Co-precipitation method: This method is widely used for the preparation of high-purity, uniform, and multicomponent ceramic precipitates with exact stoichiometry. The aqueous medium is required for this method. This method requires the mixing of two or more water soluble divalent or trivalent metal ions. These salts undergo a reaction which leads to precipitation of one or more water soluble salts. This method has a number of characteristics such as good stoichiometric control, uniform mixing, less processing time, and commercially existing chemicals [57]. **Figure 4** represents the schematic of the co-precipitation method to synthesize the nanomaterials. Singh et al. synthesized magnetite (Fe_3O_4) and hematite ($\alpha\text{-Fe}_2\text{O}_3$) iron oxide nanoparticles via facile co-precipitation method. The synthesized Fe_3O_4 material was annealed at 700°C for 2 h at room temperature to obtain $\alpha\text{-Fe}_2\text{O}_3$ phase nanoparticles. The prepared nanoparticles (Fe_3O_4 and $\alpha\text{-Fe}_2\text{O}_3$) were used for the modification of the glassy carbon electrode, which was further used for the electrochemical sensing of AP [58].



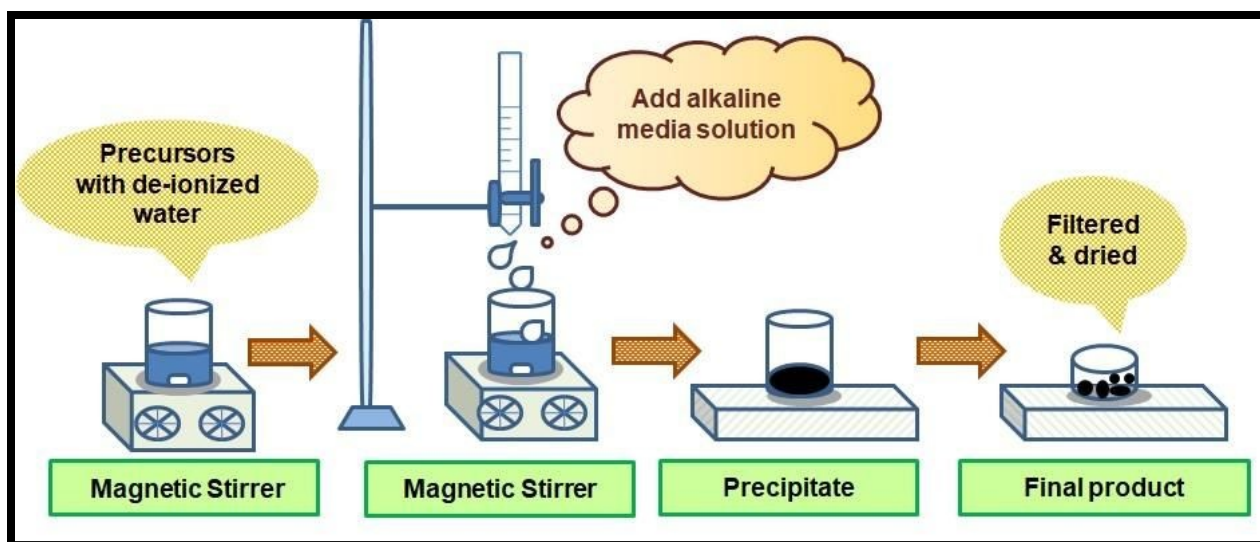


Figure 4. Schematic of co-precipitation method to synthesize nanomaterials.

Sivakumar et al. reported the synthesis of active carbon-ZnO(AC-ZnO) nanocomposites. First, they prepared ZnO via co-precipitation method using $\text{Zn}(\text{NO}_3)_2$, NaNO_3 , and NaOH as precursor materials; the 1:2 ratio of mango leaves derived activated carbon and ZnO powders were mixed in water (10 ml) to prepare ZnO-AC composite. It was observed that each ZnO nanoflake microsphere was formed by the interconnected ultrathin nanosheets [Figure 5 a, b], and the ZnO nanoflakes were successfully decorated on activated carbon. FESEM images of activated carbon at different magnifications are shown in Figure 5 c, d. The ZnO-AC modified GCE were found to be suitable to detect acetaminophen within the range from 0.05 to 1380 μM with corresponding sensitivity and LOD of about 8.33 $\mu\text{A } \mu\text{M}^{-1} \text{ cm}^{-2}$ and 0.83 μM , respectively. In this study, it was reported that AC provided large surface area and better electrochemical performance for sensor applications [59].



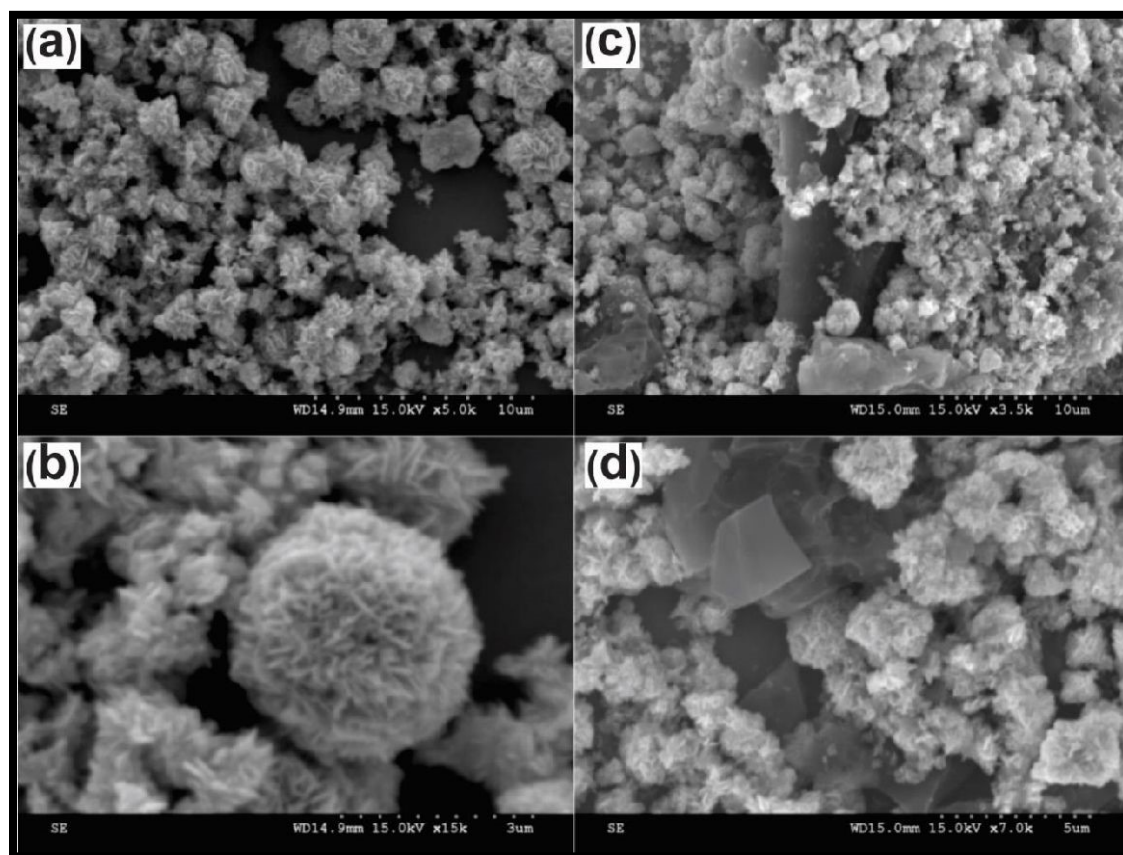


Figure 5. SEM micrographs of ZnO (a, b) and AC-ZnO (c, d) at different magnification (Reprinted with permission from ref. [59]).

Sheikhshoaie and co-workers synthesized $\text{La}^{3+}/\text{Co}_3\text{O}_4$ nanoflowers by co-precipitation method and then used them to modify graphite screen printed electrodes for sensitive detection of acetaminophen [60]. Taei et al. reported an electrochemical sensor fabricated using Fe_2O_3 (0.5)/ SnO_2 (0.5) nanocomposite for the concurrent detection of acetaminophen, epinephrine, and tryptophan [61]. Mutharani et al. synthesized 3D stone-like copper tellurate (Cu_3TeO_6) via wet chemical route and used them in electrochemical sensors for the detection of ibuprofen [62]. Zhang et al. prepared ZnO nanoflowers by using the precipitation method and then the oxygen plasma



treatment was applied for surface modification. They constructed the electrochemical sensor based on these ZnO nanoflowers for the detection of dopamine and diclofenac sodium [63].

2.4 Green synthesis

In present times, the formation of nanoparticles using the green synthesis method is an emerging shift. The main reasons behind this emergence are safety issues, reaction complications, and high cost of conventional methods. The green synthesis method involves plant products like extracts and isolates. It has a number of advantages over other methods such as its simplicity, effectiveness, strategy, rapid, and sustainability. In this procedure, fairly homogenous nanoparticles are formed and there is no requirement of toxic chemicals, high pressure, and energy, which are the main benefits of the green synthesis method. In some reactions heating is required, which slightly increases the production cost. Zamarchi et al. fabricated a biosensor using silver nanoparticles for the detection of paracetamol. Silver nanoparticles were synthesized using silver nitrate as precursor and pine nut extract as a stabilizing and reducing agent. The prepared Au NPs displayed spherical morphology with average size distribution of 91.0 ± 0.5 nm. The sensor showed linear response towards acetaminophen from 4.98 to $33.8 \mu\text{ML}^{-1}$ with detection limit of $8.50 \times 10^{-8} \text{ML}^{-1}$ and good reproducibility [64]. Iranmanesha et al. synthesized CeO_2 nanoparticles decorated with carbon nanotubes via green synthesis. They mixed CNT (50 mg) and $\text{Ce}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$ (25 mg) using mortar, pestle, and transferred this mixture in a glass vial for irradiation in the microwave. The CNT served as a microwave absorbing material and heating layer for the decomposition of facilitating $\text{Ce}(\text{NO}_3)_2 \cdot 6\text{H}_2\text{O}$. FE-SEM results revealed that the CeO_2/CNT nanocomposite had been successfully prepared and the CNT surface was completely covered by CeO_2 nanoparticles. These nanoparticles were utilized for simultaneous determination of acetaminophen (AP), uric acid (UA), ascorbic acid (AA), and dopamine (DA) in real specimens [65]. Kong et al. synthesized



rGO-TiN nanohybrid via green synthesis. They first synthesized TiO_2 nanoparticles and located them in a horizontal alumina tube furnace, raising the temperature of furnace from 0-1000°C at a rate of about 3°C min^{-1} under NH_3 gas (150 mL min^{-1}). After 6 hrs, they cooled the furnace to room temperature to obtain TiN nanoparticles. Later they mixed these nanoparticles with aqueous solution of graphene oxide and treated the resultant solution with glucose and Zn foils for 30 min at 75°C under magnetic stirring to obtain rGO-TiN hybrid nanostructures. They further reported from TEM data that TiN nanoparticles had a cuboid shape having size $\sim 50 \text{ nm}$ which were densely distributed on the surface of rGO with no free-standing TiN particles. They also found that as-synthesized rGO-TiN nanohybrid showed excellent electrocatalytic performance for the simultaneous detection of acetaminophen and 4-aminophenol within the range of $0.06\text{--}660 \mu\text{M}$ for acetaminophen and $0.05\text{--}520 \mu\text{M}$ for 4-aminophenol, respectively, along with the low detection limits of $0.02 \mu\text{M}$ for AP and $0.013 \mu\text{M}$ for 4-aminophenol, respectively [66]. Wang et al. reported green synthesis of Pd/Polyoxometalate/nitrogen-doping hollow carbon spheres tricomponent nanohybrids for selective electrochemical detection of AP [67]. Avinash et al. synthesized copper oxide nanoparticles via green synthesis using aloe vera latex as fuel. They homogeneously mixed the desired amount of cupric nitrate and aloe vera latex, and kept the blend in a preheated muffle furnace at $400 \pm 10^\circ\text{C}$. The reaction mixture bubbled to bring out a transparent gel which underwent rapid combustion throughout the volume, leaving a white colored highly porous powder, which was further calcined to get CuO nanoparticles. The TEM results displayed that the average size of the synthesized nanoparticles was 52 nm [68]. Further, the green synthesis of ZnO/Au nanoparticles [69], hematite/graphene nanocomposites [70], nitrogen doped carbon dots [71] were also reported.

2.5 Physical Vapor Deposition (PVD)



In this method, the material is deposited on a surface as a thin film or as nanoparticles. Thermal evaporation and sputtered deposition are the examples of highly controlled vacuum techniques that cause material to vaporize and then condense on a substrate. For the fabrication of thin films of various materials, physical vapor deposition techniques, such as pulsed vapor deposition, are commonly used. While for pulsed laser deposition, laser ablation is used on a solid target which results in the generation of plasma of ablated species and then deposited on a substrate to form a film. This technique is widely utilized to deposit metal nanoparticles and a thin film on carbon nanotubes. It is a very easy method for the formation of thin metal films, but it has some drawbacks such as high cost and low volume of material [72, 73]. Khoobi et al. synthesized iron oxide nanoparticles by using spray pyrolysis method and then the prepared nanoparticles were impregnated in carbon paste matrix to construct a modified sensor for the determination of acetaminophen [74].

2.6 Sputtering Technique

This method involves the vaporization of a solid through sputtering with a beam of inert gas ions. In the last few years, it was used for the preparation of nanoparticles by using magnetron sputtering of metal targets. There is a formation of collimated beams of the nanoparticles and mass nanostructured films are deposited on the silicon substrates. The whole process is carried out at relatively low pressures (1mTorr). Sputter deposition is executed in a vacuum chamber, where the molecules of sputter gas are entered and working pressure is sustained. A high voltage is applied to the target (cathode), and free electrons are pushed in a spiral direction by a magnetic system, colliding with sputtering gas (argon) atoms, resulting in gas ionization. This continuous process generates a glow discharge (plasma) that can be used to ignite. The positively charged gas ions captivated towards the target and continued to impinge. This incident occurs many times, reaching



the target's surface with energy above the surface binding energy, allowing an atom to be released. Metal atoms and gas molecules continuously collide with each other in a vacuum chamber, causing atoms to scatter and form a diffuse cloud. This technique has many advantages such as lesser impurities in deposited materials, lower cost in comparison to electron-beam lithography systems, and composition of sputtered material does not change. With the help of this technique, alloy nanoparticles can be formed with simple control on composition. In contrast to these benefits, it has some drawbacks; the nature of sputtering gas (inert gasses) can affect the texture, surface morphology, composition, and optical properties of nanocrystalline metal oxide thin films or nanoparticles [75, 76]. Soganci et al. prepared a single layer of graphene film by CVD method on copper foil and then transferred to the FTO glass slide. After this, they decorated copper nanoparticles by using the sputtering technique on it. The formed sensor showed sensitivity of $430.52 \mu\text{A mM}^{-1} \text{cm}^{-2}$ in the linear concentration range of 0.01–1.0 mM with the detection limit of 7.2 μM [77].

3. Types of electrochemical sensors

Generally, electrochemical sensors are of three types: amperometric, potentiometric and conductometric. In amperometric sensors, the current resulting from the oxidation or reduction of an electroactive species is measured. This resulting current is produced due to the potential applied between a working and reference electrode. While in potentiometric sensors, a local equilibrium is set-up on the interface of the sensor, where either the electrode or membrane potential is measured, and the concentration of analyte is acquired from the potential distinction between working and reference electrodes. On the other hand, in conductometric sensors, conductivity or resistivity is measured as a function of analyte concentration [78-81].



3.1 Potentiometric Sensors

Since early 1930's, potentiometric sensors have been developed to be the most universal practical application. These sensors have three basic device types: ion-selective electrodes (ISE), coated wire electrodes (CWES), and field effect transistors (FETS). The ion selective electrode acts as an indicator electrode, which has the potential to measure the activity of a specific ionic species selectively. In the typical layout, such electrodes are commonly membrane-based devices, containing permselective ion-conducting materials, which segregates the specimen from the inside of the electrode. Out of these, the first electrode is the working electrode whose potential is decided by its environment, while the second one is a reference electrode whose potential is determined by a solution that contains the interest in. The potential of the reference electrode is constant and potential difference value may be connected with the concentration of the dissolved ion [82, 83]. Various approaches for fabricating a cathode, particular to one species, depend on the composition and nature of the membrane material. Investigations in this field have unlocked an entire arrangement of applications to nearly an unlimited number of analytes, where the solitary limitation is the choice of ionophore matrix of the membrane and the dopant. On the basis of the nature of membrane, ISEs can be categorized into three classes: glass, liquid, or solid electrodes. It is reported that there are more than two dozen ISEs that are commercially accessible from Corning, Orion, Beckman, Hitachi, Radiometer and many more. They are broadly utilized for the investigation of organic ions and of cationic or anionic species from different effluents, also in the production and screening of drugs, utilizing selected response membrane electrodes [79].

In the mid 1970's, coated-wire electrodes (CWE) were first introduced by Freiser. In the classical CWE, a conductor is coated with a suitable ion-selective polymer membrane to make an electrode framework that is sensitive to electrolyte concentrations. The response of CWE is nearly the same



as of the classical ISE, with respect to detectability and range of concentration. Shamsipur et al. prepared a potentiometric sensor for the detection of diclofenac with the detection limit of 4.0×10^{-6} M in the concentration range of 1.0×10^{-5} to 1.0×10^{-2} M [84].

3.2 Amperometric Sensors

In reference to electroanalytical techniques, amperometric measurements are done by estimating the flow of the current in the cell at a constant potential. While a current is measured through controlled variations of the potential, it is introduced as voltammetry or voltammetric measurement. In both conditions, transfer of electrons is the main operational characteristic of the amperometric or voltammetric devices. The primary instrumentation needs controlled-potential equipment. The electrochemical cell contains two electrodes immersed in an appropriate electrolyte. While a more intricate and normal arrangement includes the utilization of a three-terminal cell containing working, counter, and reference electrodes. The main reaction occurs at the working electrode. On the other hand, the electrode which provides a constant potential in comparison to the working electrode is referred to as a reference electrode. Chemically inert conducting materials like graphite or platinum are utilized as the auxiliary (or counter) electrode. In controlled-potential experiments, a supporting electrolyte is needed to sustain the ionic strength constant, decrease in the solution's resistance, and eliminate electromigration effects [79, 85, 86]. The working electrode materials strongly affect the performance of amperometric sensors. Therefore, great efforts have been dedicated to the fabrication and maintenance of the working electrodes. The classical electrochemical measurements of analytes started after the invention of dropping mercury electrodes by Heyrovsky. In recent years, solid electrodes, fabricated using noble metals and different forms of carbon, have been found to be of great interest for the development of sensors. The effective advancements in electroanalytical chemistry have led to the



development of various electrochemical sensors. For several years, mercury was a very captivating material for electrodes due to its renewable surface, extended cathodic potential range window, and high reproducibility. But its applications were limited because of its toxicity and limited anodic potential. Alternatively, solid electrodes such as nickel, platinum, gold, carbon, and dimensionally stable anions became very famous as electrode materials. Because these materials have a low cost, multifaceted potential window, chemical inertness, low background current, and the ability for different sensing and detection usage. Currently, various nanomaterials are being developed for applications to electrochemical sensors [87]. Lima et al. fabricated an amperometric sensor based on double walled CNTs/GCE and used it for the sensing of dopamine and catechol. The sensitivity for dopamine and catechol was found to be 0.259 and 0.301 $\mu\text{A L } \mu\text{mol}^{-1}$, respectively [84].

3.3 Conductometric Sensors

This type of sensor greatly depends on variations of electrical conductivity of a film or bulk substance whose conductivity is influenced by the analyte presented in that material. Conductometric methods are fundamentally non-specific. However, there are a few practical factors that enable conductometric methods to be capitative, such as its economical perks, simplicity since no reference anodes are required, and its insensitivity to light. Besides this, these can be miniaturized as planar interdigitated electrodes, and integrated easily by the use of a thin film standard technology, which makes it inexpensive and easy to use for applications in a number of biosensors and gas sensors. Sadek and co-workers fabricated conductometric sensor for the detection of H_2 gas. They deposited doped and de-doped nanofibers onto conductometric sapphire transducers, and the various concentrations of hydrogen (H_2) gas at room temperature were used to determine the sensor characteristics. The sensitivity of the H_2 sensor was measured as 1.11 for



doped and 1.07 for de-doped polyaniline nanofiber sensors upon exposure to 1% H₂. The dedoped nanofibers exhibited better repeatability than the doped nanofibers [88].

Ghosh and co-workers have developed a low cost conductometric glucose sensor, which can detect glucose concentration as low as 10 nm [89]. Sun and co-workers have fabricated a conductometric sensor based on Tourmaline@ZnO core-shell structure for n-butanol gas detection. It was reported that the optimum sensitivity of the sensor was 120.8–100 ppm *n*-butanol at 320°C in 1% Tourmaline@ZnO sample, which was more than twice that of pure ZnO [90]. Wang and coworkers have fabricated conductometric sensor to detect ammonia gas by employing titanium dioxide nanoparticles decorated black phosphorus (BP) nanosheets as the sensing layer. First, they prepared planar interdigital electrodes (IDEs) (thickness of Au/Ti layers: 200 nm/100 nm) onto SiO₂/Si substrate by lithography and lift-off methods with an active area of 7 mm × 11 mm. Then they dropped cast aqueous BP (40 μL), TiO₂ (40 μL), and BP-TiO₂ (40 μL) solutions on different IDEs devices, followed by 60°C vacuum heating for 2 hrs. and cooling to room temperature. The titanium dioxide nanoparticles decorated with BP nanosheets electrodes were found to be suitable to detect NH₃ in the linear range of 0.5–30 ppm at RT [91].

4. Some voltammetric techniques for electrochemical sensors

In analytical chemistry, voltammetric techniques are widely used for quantitative determination of different dissolved organic and inorganic substances. In inorganic, physical, and biological chemistry, the use of voltammetric techniques is done for the different purposes such as electron transfer and reaction mechanisms, fundamental studies of oxidation and reduction processes, kinetics of electron transfer processes, thermodynamic properties of solvated species, and adsorption processes on surfaces. This technique is also of interest for the determination of compounds in pharmaceuticals.



The various voltammetric techniques that are used are differentiated from each other by the material used as working electrode and the potential function which is applied to the working electrode. Some voltammetric techniques are explained in short in this article:

4.1 Linear Sweep Voltammetry (LSV)

In this analysis method, linear potential is applied, which sweeps to the working electrode while the current flowing in the circuit is being simultaneously measured. A signal generator produces a voltage sweep from E_s to E_f , and a potentiostat applies this potential wave to the electrode under study. The direction of scan can be positive or negative, and in principle the sweep rate can possess any constant value:

$$\text{Sweep} = dE/dT$$

This technique is generally used in polarography under well-defined conditions; the limiting current derived from a redox process in the solution during LSV may be used to quantitatively determine the concentration of electroactive species in the solution [92]. Vilian and co-workers investigated the electrochemical performance of dopamine and paracetamol by using a sensor, based on ZrO_2 nanoparticles supported on graphene oxide. The prepared sensor showed two well-defined voltammetric potential peaks at 0.34V and 0.53V, for dopamine and paracetamol, respectively [93].

4.2 Square Wave Voltammetry (SWV)

This voltammetric technique was invented by Ramaley and Krause, and further developed by Oster Youngs and their co-workers [92]. In this analysis method, remarkable versatility is found. It is a differential technique in which the potential, in a form of symmetrical square wave of constant amplitude, is superimposed on a bare staircase potential [94]. In this analysis, the graph is plotted between the difference in the current measured in forward (i_f) and reverse cycle (i_r), versus average



potential of each waveform cycle. In this technique, the peak potential exists at the E_m of the redox couple, because the current function is symmetrical around the potential. High sensitivity and excellent peak separation are the main benefits of this technique. Kang et al. detected acetaminophen in pharmaceutical preparation tablets by using a graphene based sensor. They reported a recovery rate from 96.4% to 103.3% with a detection limit of 3.2×10^{-8} M [95].

4.3 Differential Pulse Polarography / Voltammetry (DPP/ DPV)

This technique was propounded by Barker and Gardner. With the help of this technique, greater sensitivity, more efficient resolutions, and differentiation of various species can be achieved. In this technique, each potential pulse is constant, which is of small amplitude (0.01 to 0.1), and the current is measured at two points from each pulse, one is just before the application of the pulse and another at the end of the pulse. The difference between the current measurements at these points, for each pulse, is calculated and plotted against the base potential. This difference in currents attains a maximum value near the redox potential while minimum (nearly zero) as the current becomes diffusion controlled. Therefore, a symmetric peak is obtained as the current resp [92, 94]. Zhang et al. prepared a sensor based on 3D-rGO/GCE nanocomposite for the detection of an antibiotic drug. They found two reduction peaks at 170 and 633 mV with detection limit of $0.15 \mu\text{mol L}^{-1}$, in the detection range of $1\text{--}113 \mu\text{mol L}^{-1}$ [96]. Sebastian et al. developed a sensor based on GO/3D hierarchical ZnO nanocomposite for the detection of chloramphenicol. The electrochemical performance was observed via DPV and CV technique. From DPV technique, the LOD of the GO/ZnO modified GCE was found to be about $0.01 \mu\text{M}$ in the linear range of $0.2\text{--}124 \mu\text{M}$, while the sensitivity was found to be about $7.27 \mu\text{A } \mu\text{M}^{-1} \text{ cm}^{-2}$ and exhibited high stability, reproducibility, and repeatability [97].

4.4 Anodic Stripping Voltammetry (ASV)



It is an electrolytic method in which mercury is kept at the negative potential to decrease the metal ions in a solution to form an amalgam with the electrode. In this method, the stirring of the solution is done to transfer the analyte metal to the electrode as possible for concentration into the amalgam. After reducing and accumulating the analyte for a duration, the potential on the electrode increases to reoxidize the analyte, and in this way a current signal is generated. The current produced by anodic stripping depends on the particular type of mercury electrode, but is directly proportional to the concentration of analyte concentrated into the electrode [92]. With the help of this method, Mohammed et al. determined the traces of timolol maleate drug by using nafion/carboxylated-MWCNTs nanocomposite GCE. The LOD was found to be 7.1×10^{-10} mol L⁻¹ in the linear range of 1.0×10^{-9} - 2.0×10^{-5} mol L⁻¹ [98].

4.5 Cyclic Voltammetry (CV)

This method was first reported in 1938 and was described by Randies. It is the most commonly used technique for achieving qualitative information about electrochemical reactions. The capacity of cyclic voltammetry results from its ability to rapidly offer considerable information on the thermodynamics of redox processes, couples' chemical reactions, kinetics of heterogeneous electron-transfer events. It is the first experimental approach used in electroanalytical investigation, because it provides rapid determination of redox potentials of the electroactive species. The CV evaluates both quantitative as well as qualitative information about electrochemical processes in working electrode active materials. This technique appeals to a potential about the working electrode in place of the fixed potential of the reference electrode; it also sweeps side to side linearly between these two predefined potentials. The operating stability of the electrolyte limits the potential capacity. The graph is plotted between a time-dependent current, which is obtained by scanning the potential range versus scanned potential (E). If IdV is



the integrated area under the CV curve, V_s is the potential scan rate and V is the measured potential range then the specific capacitance is given by:

$$C = \frac{1}{m * V_s * V} \int_0^V I(V) dV \quad (1)$$

$$C_s = \frac{C}{m} \quad (2)$$

where C is the evaluated capacitance from equation (1) and m is the active mass of the material [92, 94]. For instance, Chethana et al. investigated oxidative behavior of diclofenac by using this technique. They prepared the electrode by mixing the carbon paste with tyrosine which greatly increased the sensitivity of the CPE. The prepared sensor exhibited a satisfactory LOD of 3.28 μM and a sensitivity of 0.1905 $\mu\text{A } \mu\text{M}^{-1}$ [99].

5. Nanomaterials based electrochemical sensing platform for some analgesic and antipyretic drugs

5.1 Metal Oxide Nanomaterials

Metal oxide nanoparticles have received much consideration due to their special properties and various potential applications [100]. Different morphologies of metal oxide nanoparticles have been made through versatile synthesis techniques. These metal oxide nanoparticles exhibit several types of photochemical, electrochemical, electronic, and electrical properties because of their shape, size, stability, and high surface area. Metal oxide nanoparticles exhibit remarkable qualities especially for the noble metal nanoparticle modified electrodes, which usually show good electrocatalytic activity in the compounds with slow redox process at unmodified electrodes [101, 102]. Metal oxide nanoparticles such as TiO_2 nanoparticles, ZrO_2 nanoparticles, Fe_3O_4 nanoparticles, etc. have been successfully used for the development of electrochemical sensors because of their catalytic ability; faster electron transfer kinetics and morphology.



Ozcan et al. constructed a high performance acetaminophen sensor based on zinc (Zn)/zinc oxide (ZnO) decorated with reduced graphene oxide surfaces. They have synthesized Zinc (Zn)/zinc oxide (ZnO)/reduced graphene oxide nanohybrids by facile chemical precipitation method. It was confirmed using XRD, TEM, XPS, and TGA that Zn/ZnO nanoparticles were immobilized on the rGO surface with average particle size around 25.1 ± 6.6 nm. The response of the sensor to detect APAP was found to be on a linear range of 0.05 to 2 mM, and high sensitivity of $166.5 \pm 6 \mu\text{A.mM}^{-1}.\text{cm}^{-2}$ [103]. Kenarkob and Pourghobadi electrochemically synthesized glassy carbon electrodes modified with ZnO/Au nanoparticles for acetaminophen sensing. The characterization results displayed that Au nanoparticles were well anchored onto ZnO nanospheres. The LOD was found to be 9 nM by using SWV technique in the concentration ranges of AP between 0.05-20 μM [69].

Wang et al. prepared Co/Co₃O₄ nanoparticles, coupled with hollow nonporous carbon polyhedrons nanoparticles using pyrolysis and subsequent oxidation techniques to develop electrochemical sensors for acetaminophen detection. It was reported that pore architectures of hollow carbon polyhedrons were found to be favorable for interface features. The mesopore size and micropore size distributions of Co/Co₃O₄@HNCP were found to have pore diameter from 3 to 7 nm, and 0.6 to 1.6 nm, respectively. The mesopore size distribution was beneficial for the mass transport and adsorption of molecules; the micropore size distributions features, possible discriminating ability of the constructed Co/Co₃O₄@HNCP sensor for electrochemical sensing of AP. The constructed Co/Co₃O₄@HNCP sensor showed ultrahigh sensitivity ($157 \mu\text{A } \mu\text{M}^{-1}$) and a very low LOD (0.0083 μM) [Figure 6(a-d)] for acetaminophen sensing in the concentration range from 0.025–2.5 μM and 2.5–50 μM , respectively [67]. Sheikshoaie and coworkers have synthesized La³⁺/Co₃O₄ nanoflowers by co-precipitation method and used them to modify graphite screen



printed electrodes for acetaminophen detection. The $\text{La}^{3+}/\text{Co}_3\text{O}_4$ modified graphite screen printed electrode was found to be electrocatalytic active toward the detection of acetaminophen with a wide linear range of concentrations from $0.5\mu\text{M}$ to $250.0\mu\text{M}$, and detection limit of $0.09\mu\text{M}$ [60]. Annadurai et al. prepared nickel oxide nanoparticles by the hydrothermal route to modify GCE for the determination of acetaminophen. CV, DPV, and amperometry were employed to investigate the electrochemical behavior of NiO modified glassy carbon electrodes (3 mm diameter). The size of NiO nanoparticles was obtained to be between 15 and 20 nm. It was found from the electrochemical studies that the sensor exhibited linear detection range from 7.5 to 3000 mM with high sensitivity of $91.0\mu\text{A cm}^{-2}\text{mM}^{-1}$, and low detection limit of 0.23 mM. The repeatability and dynamic stability of the constructed sensor is shown in **Figure 6(e)** and **Figure 6(f)**, respectively [42]. The electrodeposition preparation technique was adopted by Liu et al. to synthesize nickel and copper oxide nanoparticles. These prepared oxides were further used to modify graphene electrodes to detect dopamine, acetaminophen, and tryptophan. The modified electrode displayed linear response ranges of 0.5–20 μM , 4–400 μM , 0.3–40 μM , and the detection limits of 0.17 μM , 1.33 μM , 0.1 μM , for detecting dopamine, acetaminophen, and tryptophan, respectively[104]. Manikandan and Dharuman prepared un-doped $\alpha\text{-Fe}_2\text{O}_3$, platinum doped Fe_2O_3 , (dPtFe_2O_3), Pt decorated Fe_2O_3 (sPtFe_2O_3) and doped and decorated Fe_2O_3 ($\text{sdPtFe}_2\text{O}_3$) nanoparticles by co-precipitation method in presence of polyethylene glycol and modified glassy carbon electrode surface. These modified electrodes were utilized for the simultaneous detection of melatonin, dopamine, and acetaminophen. The experimental results concluded that $\text{sdPtFe}_2\text{O}_3$ has higher catalytic activity than the other modified electrodes such as, dPtFe_2O_3 , sPtFe_2O_3 and un-doped $\alpha\text{-Fe}_2\text{O}_3$ [105a].



Cao and coworkers have synthesized CeBiO_x nanofibers via a simple two-step procedure which includes electrospinning and calcination with Ce:Bi ratio of 0.25:0.75, 0.5:0.5, and 0.75:0.25. The SEM studies showed uniform, long, and continuous Ce_{0.75}Bi_{0.25}O_x NFs with an average diameter of ~200 nm. They prepared CeBiO_x nanofibers with modified screen printed carbon electrodes for the detection of acetaminophen. From the DPV studies, it was reported that Ce_{0.75}Bi_{0.25}O_x NFs modified SPE showed linear detection range for AP between 2.5 μM to 130 μM with high sensitivity of 360 μA mM⁻¹cm⁻² and a low detection limit of 0.2 μM [106]. Khoobi et al. constructed a modified sensor by using iron oxide nanoparticles by impregnating them in a carbon paste matrix for the detection of diclofenac. The developed sensor showed a very low detection limit of 2.45 nM in the linear range of 0.01-100.0 μM; it also exhibited long-term stability, good sensitivity, and repeatability [74]. Mutharani et al. developed an electrochemical sensor based on (Cu₃TeO₆) for the detection of anti-inflammatory agent ibuprofen in the fluids of the human body. The constructed sensor showed the detection limit of 0.017 μM in the linear range of 0.02–5 μM and 9–246 μM, respectively. This designed sensor also demonstrated high sensitivity, good selectivity, and long term stability [62]. Diouf et al. fabricated an electrochemical sensor by self-assembling chitosan capped with Au NPs on a screen-printed carbon electrode, for the detection of aspirin in tablets and human physiological fluids. The designed sensor demonstrated good selectivity, sensitivity, and reproducibility with satisfactory analytical parameters ($R^2 = 0.97$) and a low LOD of about 0.03 pg/mL estimated from DPV [107].



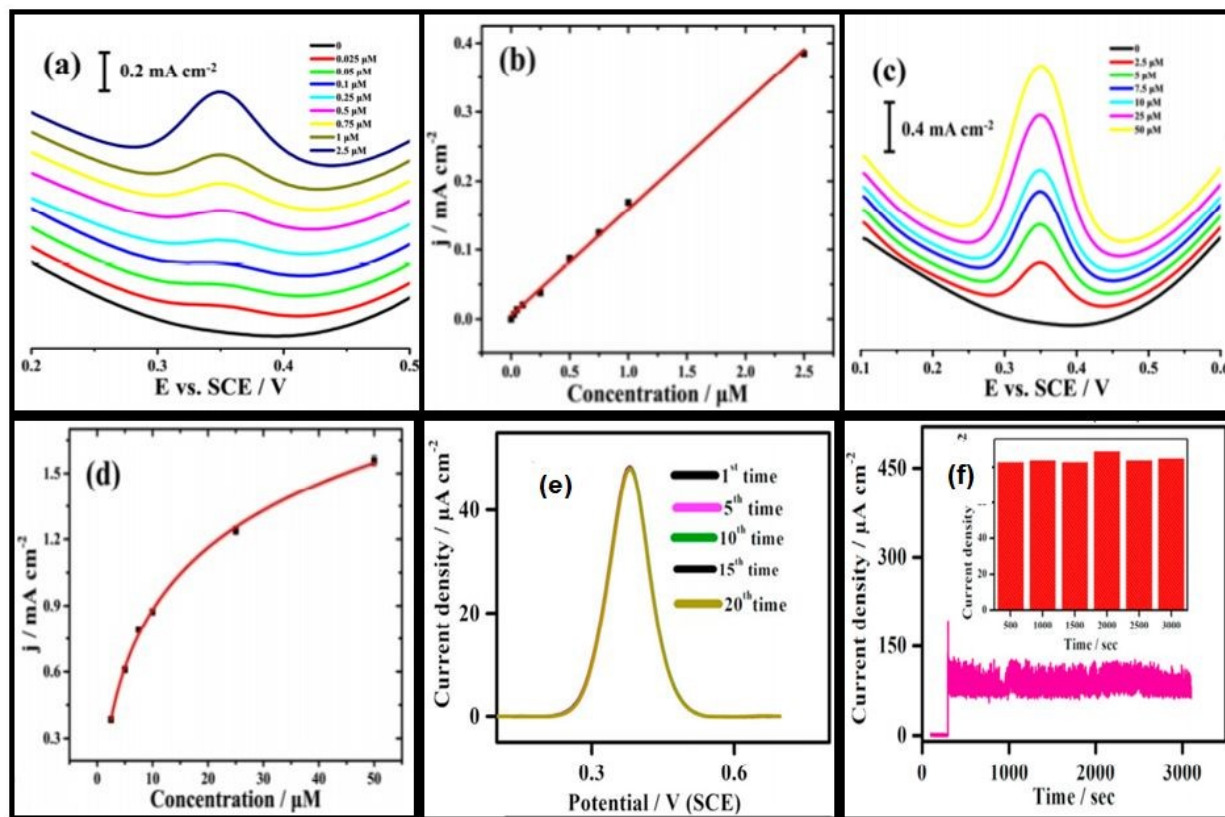


Figure 6. DPV curves for Co/Co₃O₄@HNCP sensor for sensing of AP in concentration windows of (a) 0.025–2.5 μM and (c) 2.5–50 μM . (b) linear calibration curves and (d) logarithmic calibration curves (Reprinted with permission from ref.[67]). (e) Repeatability test using DPV and (f) Amperometric stability study for the NiO/GCE with 0.25 mM PA addition. Insert shows flow chart plot of amperometric current density vs time (Reprinted with permission from ref. [42]).

5.2 Carbon Materials

Carbon, one of the most versatile element found on earth, attracted a lot of attention due to its special type of hybridization states (sp , sp^2 and sp^3), which makes it capable of forming a wide range of allotropes from diamond (hardest material) to graphite (softest materials) [108, 109]. Recently, various type of carbon based nanomaterials, such as single or multiwalled carbon nanotubes (SWCNT or MWCNT) [110], carbon nanostructures, graphene [111], Graphene



oxide[112], and reduced graphene oxide [113] have been synthesized, gaining attention due to their physical, optical, magnetic, and super electronic properties; these make carbon suitable for wide applications such as sensing devices [114], energy storage [115], and drug delivery [116].

There are various studies which confirm the use of carbon-based materials for the detection of acetaminophen. Cernat and team reported, in their review article, that different configurations of carbon such as modified/unmodified carbon nanotubes and graphene could become a good candidate for electrochemical sensors and biosensors to detect acetaminophen [117]. Gopal and coworkers fabricated eco-friendly and bio-waste-based hydroxyapatite/rGO hybrid modified carbon paste electrodes for the simultaneous detection of dopamine, acetaminophen, and ascorbic acid. They used eggshell bio-waste-based hydroxyapatite materials for the sensing applications. The electrochemical performance results showed a good linear range for the detection of acetaminophen from 20 μM - 160 μM [118]. Pham and coworkers have prepared platinum decorated rGO modified glassy carbon electrodes via environment friendly electrodeposition technique for the detection of acetaminophen. The morphology characterization techniques revealed the cauliflower-like structure of Pt particles. The prepared electrochemical sensor was able to detect acetaminophen in a linear concentration range from 0.01 to 350 μM , with a detection limit of 2.2 nM [119].

Alam and coworkers coated glassy carbon electrode by drop cast method with multi-walled carbon nanotubes- β -cyclodextrin composites for low level detection of acetaminophen in water. It was reported that the sensor responded to acetaminophen in a linear range of 50 nM - 300 μM with the detection limit of 11.5 nM, good reproducibility, high stability, and exclusive selectivity. They also mentioned that this improvement was because of electron transfer capability and high conductivity of MWCNT, high surface-to-volume ratio of the MWCNT, and higher surface area



of the sensor due to the porous structure of CD [110]. Berto et al. designed and tested electro-activated glassy-carbon electrode for acetaminophen detection in surface water. The sensor showed a linear range of $13.3 \mu\text{g L}^{-1}$ to $33 \mu\text{g L}^{-1}$ for acetaminophen. The sensor was able to detect acetaminophen concentrations higher than $4.4 \mu\text{g L}^{-1}$ in untreated samples [120].

Liang and coworkers developed a sensor based on glassy carbon electrodes modified with nitrogen-rich porous carbon for acetaminophen detection. They have synthesized nitrogen rich porous carbon nanotubes by assisted carbonization of the zeolitic imidazolate framework ZIF-8 using poly vinylpyrrolidone. They found from TEM studies that the average size of the ZIF-8 polyhedra was 70 nm. The pore size of P-NC was found to be distributed in 4–4.5 and 30–50 nm range, which confirmed the efficient preparation of a meso-microporous hierarchical structure. According to them, the highly porous structure of prepared electrodes provided an interweaving network which facilitated more active sites and helped in better transportation of reactants and products, contributing to better electrochemical performance. The sensor showed good linearity for acetaminophen in the linear range of 3–110 μM , with a minimum LOD $\sim 0.5 \mu\text{M}$ (S/N $\frac{1}{4}$ 3) and the sensor was efficiently applied to the detection of acetaminophen in urine samples. **Figure 7(a)** displays DPV voltammograms while **Figure 7(b)** shows the calibration plot of the synthesized sample [121].

Amiri et al. reported carbon nanoparticle modified carbon paste electrodes for detection of paracetamol, phenylephrine, and dextromethorphan. The paste design consisted of an hydrophobic binder, hydrophobic graphite as a conducting component, and a nanoparticulate thin film with hydrophilic surface to provide sensitivity and selectivity. The sensor was found to be suitable for detection of acetaminophen in the linear range of $1 \times 10^{-7} \text{ M}$ to $1.0 \times 10^{-3} \text{ M}$, with a detection limit of $1.5 \times 10^{-8} \text{ M}$ [9].



Tsierkezos studied the effect of the incorporation of nitrogen on electrocatalytic activity of MWCNTs. They synthesized vertically aligned MWCNTs onto oxidized porous silicon wafer by means of catalytic CVD technique using acetonitrile as carbon source material and ferrocene as catalyst. The solution of ferrocene in acetonitrile (1% w/w) was introduced to a furnace at 900 °C through a syringe with a flow rate of 0.2 ml min⁻¹ to fabricate nitrogen-doped multiwalled carbon nanotubes. The modified carbon nanotubes were utilized to develop the electrochemical sensor for acetaminophen sensing. It was investigated that the sensor detection ability was enhanced by the nitrogen doping in CNTs. The sensor based on the pristine MWCNTs recorded very low sensitivity (0.6010 A M⁻¹ cm⁻²) and detection limit (0.950 μM), but nitrogen doped MWCNT sensor's detection limit; sensitivity increased significantly to 0.485 μM and 0.8406 A M⁻¹ cm⁻², respectively [122]. Barsan and coworkers have developed and characterized acetaminophen sensors in two different architectures. In first configuration, they electropolymerized PMG onto graphite-epoxy composite electrode (CE) and then coated these electrodes with fCNT (fCNT/PMG/CE); in second architecture they coated CE with fCNT and then electropolymerized PMG onto these electrodes (PMG/fCNT/CE). It was observed that on fCNT/CE, the polymer was better formed compared to CE because of the higher surface area of fCNTs. The sensors were used to detect pyridoxine and acetaminophen, observing that fCNT-PMG-CE possessed higher sensitivity than PMG-fCNT-CE [123]. Li and coworkers described the fabrication of layer-by-layer (LBL) carboxylic acid functionalized multiwalled CNT on glassy carbon electrodes to develop an electrochemical sensor for paracetamol (Acetaminophen). The covalent LBL assembly was confirmed using SEM. It was reported that the modified electrode with six layers exhibited good sensitivity of 2.293 μA M⁻¹ cm⁻² in the linear range and detection limit of 1-200 μM and 0.092 μM [93]. Sarhangzadeh et al. constructed a sensor for the simultaneous detection of diclofenac and indomethacin by using

MWCNT and ionic liquid modified carbon ceramic electrodes. By using the DPV technique, the prepared electrode showed linear calibration curves in the concentration range of 0.05–50 $\mu\text{mol L}^{-1}$ for diclofenac and 1–50 $\mu\text{mol L}^{-1}$ for indomethacin. The developed sensor showed the limit of detection 18 and 260 nM for diclofenac and indomethacin, respectively [124]. Roushani et al. fabricated an ultrasensitive sensor by using gold nanoparticles which were electrochemically deposited on glassy carbon electrode surface for the detection of ibuprofen in spiked human serum. In this study, it was observed that a layer of Au nanoparticles can improve the electrochemical performance as well as electron transfer due to its large surface area. The designed sensor demonstrated good reproducibility and long-term stability. The detection limit was found to be 0.5 p mol^{-1} in the concentration range from 0.005 nmol^{-1} to 7 nmol^{-1} [125]. Suresh et al. simultaneously detected paracetamol and ibuprofen in different tablets by using glassy carbon electrodes with the help of stripping SWV and DPV. The constructed sensor showed good repeatability and recorded a detection limit (LOD) of 0.96 $\mu\text{mol L}^{-1}$ in a linear concentration range between 1.45 and 3.87 $\mu\text{mol L}^{-1}$. It was reported that this type of sensor can be successfully used for both drugs (AP and IB) in commercial tablets. This obtained data was found in agreement with the data of many manufacturing companies [126].



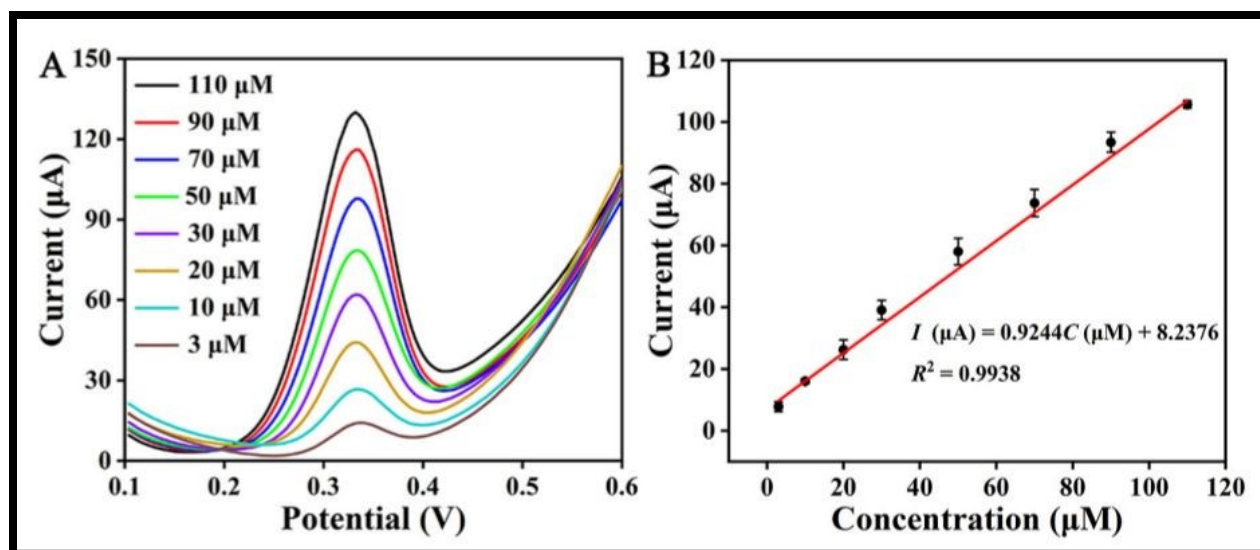


Figure 7. (A) DPV curves of AP at different concentrations on P-NC/GCE in 0.1 M PBS (pH 7.0). (B) Calibration plot of peak current as a function of concentration of AP (Reprinted with permission from ref. [121]).

5.3 Nanocomposites:

Nanocomposites have the properties of their components including a new characteristic due to the synergistic effect. Because of their new features, nanocomposites are widely employed in electrochemical sensors to detect different drugs. Various investigations have been done on nanocomposite based electrochemical sensors for the detection of different drugs. Hao et al. synthesized copper nanowires and used them to prepare Cu Nanowire/Graphene oxide nanocomposite (Cu-NW/GO). Later they modified glassy carbon electrodes using this Cu-NW/GO nanocomposite for the simultaneous detection of acetaminophen, dopamine, and ascorbic acid. They reported that the sensor presented a wide linear range from 1–60 μM, 1–100 μM, and 1–100 μM with detection limit of 50, 410, and 40 nM, for ascorbic acid, dopamine and acetaminophen, respectively [127]. Hasanpour et al. prepared a semiconductor composite CuO/CuFe₂O₄ with p-n junction. The prepared nanocomposite was to prepare carbon paste electrodes for the detection of acetaminophen and codeine in the biological fluids. The electrode surface area of



CuO/CuFe₂O₄/CPE was observed to be 0.85 cm² which was 5.21 times greater than the surface area of the unmodified carbon paste electrode. It was reported that CuO/CuFe₂O₄/CPE electrode was found suitable for acetaminophen detection with a very low detection limit of 0.007 μmol L⁻¹ in the linear range of 0.01 - 1.5 μmol L⁻¹ [128]. Lin et al. synthesized graphene/ZrO₂ nanocomposite to modify screen printed carbon electrodes (SPCE), and it also studies their sensing properties for acetaminophen detection. Graphene/ZrO₂ modified screen printed carbon electrodes showed the acetaminophen detection limit of 75.5 nM, in the linear range of 10-100 μM [129]. Tamilalagan et al. fabricated (Ni/Zn)O@rGO p-n heterojunction semiconductor nanocomposite which was used for acetaminophen detection using the DPV technique. The DPV response curve shows the linear relationship between AAP concentration and anodic current. This excellent response behavior of synthesized nanocomposite was observed with the detection limit of up to 2.2 nM in the wider range from 0.009 to 413 μM. On the other hand, (Ni/Zn)O@rGO/GCE exhibited high sensitivity of 19.1 μA μM⁻¹ cm⁻². This might occur because of larger surface area of prepared nanocomposite, where nano-sized spherical particles were found to be decorated on the rGO sheets. It was also noticed that (Ni/Zn)O@rGO modified GCE demonstrated very good selectivity towards acetaminophen with good repeatability and reproducibility [130]. Nikpanje et al. developed an electrochemical sensor based on a carbon paste electrode (CPE), modified with ZnO-Zn₂SnO₄-SnO₂ and graphene (ZnO-Zn₂SnO₄-SnO₂/Gr/CPE) for the detection of acetaminophen, ascorbic acid, and caffeine. It was shown that the modified electrodes exhibited high electrical conductivity which made it a good candidate for electrochemical applications. The amperometric response of the electrodes confirmed the detection limits as 0.00364, 0.00385, and 0.00628 μM for acetaminophen, caffeine, and ascorbic acid in the linear range from 0.008-12 μM, 0.01-14 μM, and 0.013-16 μM, respectively. On the other hand, ZnO-Zn₂SnO₄-SnO₂/Gr/CPE



based sensor showed superb selectivity, repeatability, stability, and reproducibility for the determination of acetaminophen, ascorbic acid, and caffeine [131]. Iranmanesh et al. modified glassy carbon electrode with CeO_2 -CNTs nanocomposites and used them for the simultaneous detection of ascorbic acid (AA), dopamine (DA), uric acid (UA), and acetaminophen (AP). The detection limits for AA, DA, UA and AP were found to be 3.1 nM, 2.6 nM, 2.4 nM, and 4.4 nM in the linear range from 0.01–900.0 μM , 0.01–700.0 μM , 0.01–900.0 μM , and 0.01–900.0 μM , respectively [65].

Afkhami and team developed a novel electrochemical sensor based on carbon paste electrode modified with NiFe_2O_4 /graphene nanocomposites for effective and simultaneous detection of acetaminophen and tramadol. The morphology and electronic composition of the prepared NiFe_2O_4 /graphene nanocomposite was confirmed by the SEM, XRD, and FT-IR spectrometry. The limit of detection for acetaminophen and tramadol was confirmed to be 0.0036 and 0.0030 $\mu\text{mol L}^{-1}$, respectively. It was found that the sensitivity of the sensor was enhanced by using the combination of graphene and NiFe_2O_4 nanocomposites. The fabricated sensor possessed high sensitivity and good stability for clinical assay of tramadol and acetaminophen [132]. Demir and group designed an electrochemical sensor based on screen printed electrode modified with molybdenum disulphide (MoS_2)-titanium dioxide (TiO_2)/reduced graphene oxide (rGO) (MoS_2 - TiO_2 /rGO/SPE) nanocomposite for paracetamol detection. The sensor was examined for the effect of rGO: MoS_2 - TiO_2 ratio, MoS_2 - TiO_2 /rGO composite amount on the screen-printed electrode for sensitive and selectivity detection of paracetamol. It was reported that the sensor showed high electrocatalytic performance for oxidation of acetaminophen with a low detection limit of 0.046 μM Ac and a wide linear response in the range of 0.1–125 μM . The sensor was also used for the detection of acetaminophen in urine and drug samples with acceptable recovery values [133].



Anuar et al. fabricated modified glassy carbon electrodes with platinum nitrogen-doped graphene (Pt/NGr) nanocomposite for sensitive determination of acetaminophen. They reported from TEM images that Pt nanoparticles were fairly distributed on the NGr sheets. It was also reported that the synergy between nitrogen-doped graphene and platinum nanoparticles enhanced the interfacial electron transfer process and showed higher catalytic performance towards the electrochemical oxidation of acetaminophen. The sensor was found to be suitable for the detection of acetaminophen in a linear range of 0.05–90 $\mu\text{mol L}^{-1}$ with a detection limit of 0.008 $\mu\text{mol L}^{-1}$. FESEM images are shown in **Figure 8(a & b)**, while **Figure 8c** and **8d** represent the electrochemical results of as prepared sensor [134].

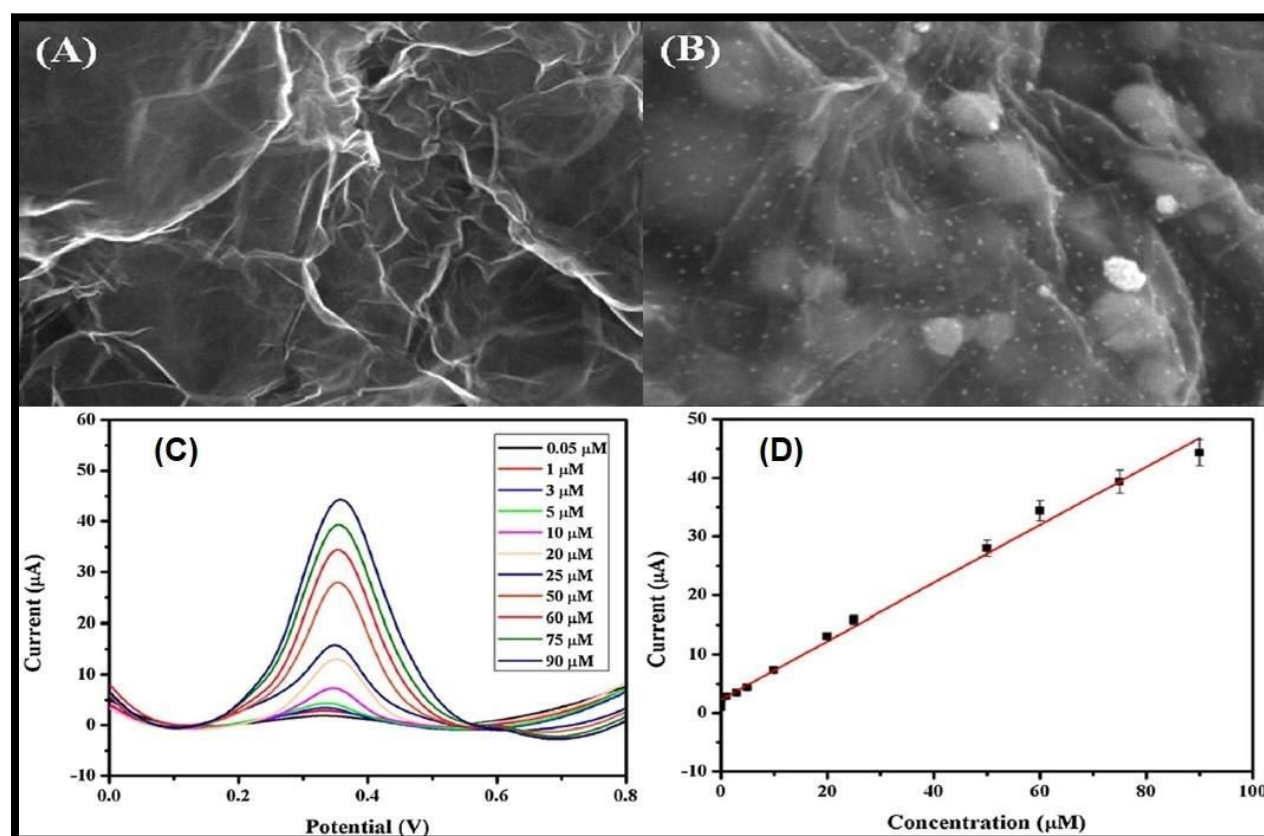


Figure 8. FESEM images of (A) NGr, (B) NGr deposited with Pt NPs., (C) Square-wave voltammograms with different AC concentrations at Pt/NGr/GCE in 0.1 mol L⁻¹ phosphate buffer



solution at pH 7.0. (D) A plot of peak current against concentration of AC (Reprinted with permission from ref.[134]).

Shaikshavali et al. fabricated an electrochemical sensor based on MWCNTs decorated with CuO-Au composite for sensitive determination of acetaminophen and 4-Aminophenol. It was reported that the electrochemical activity of CuO-Au/MWCNTs/GCE was higher than bare GCE, CuO-Au/GCE and MWCNTs/GCE. Good linear response from 0.2 μM to 6.0 μM and 0.5 μM to 1.6 μM with detection limits of 0.016 μM and 0.105 μM recorded for acetaminophen and aminophenol, respectively[135]. Huang and co-workers reported that modified glassy carbon electrodes with layered MoS_2 -graphene composites could be used for acetaminophen detection in the linear range of 0.1–100 μM with a detection limit of $2.0 \times 10^{-8} \text{M}$. The superior electrochemical performances resulted because of the robust composite structure and the synergistic effects between layered MoS_2 and graphene [136]. Kimuam et al. prepared platinum nanoflowers and reduced graphene oxide nanocomposite (PtNFs/rGO), using them for the determination of diclofenac in urine samples. With the help of the DPV technique, a linear range was observed between 0.1–100 μM , with a detection limit of 40 nM. Recovery rate was found to be in the range of 85-100%; they proposed that this system might be used in various applications [137]. Goyal et al. used SWCNT modified pyrolytic graphite electrode for the determination of diclofenac. Diclofenac oxidized at 439 mV and 854 mV, at pH 7.2. It was observed that the modified electrode demonstrated excellent catalytic activity as compared to the bare electrode. The calibration curves are found to be linear in the concentration range of 1×10^{-9} –500 $\times 10^{-9}$ M and 25×10^{-9} –1500 $\times 10^{-9}$ M for peaks I and II, respectively. In this investigation, they used the SWV technique to determine diclofenac in biological and pharmaceutical samples [138]. Nasiri et al. constructed a sensor by using graphene oxide/CNT nanocomposite and gold nanoparticles for the determination of



diclofenac molecules. They used an electrochemical deposition method to deposit AuNPs at the surface of MWCNT-GO nanocomposite films. The developed sensor showed good results which may be attributed to large surface area of prepared nanocomposite and fast electron transfer rate of AuNPs. The detection limit was found to be $0.09 \mu\text{mol L}^{-1}$ in the linear range of $0.4\text{--}1000 \mu\text{mol L}^{-1}$ [139].

For the detection of acetaminophen, Charithra et al. fabricated poly asparagine that occurred at the modified electrode were pH dependent. The prepared electrode showed a very low detection limit of $4.10 \times 10^{-8} \text{ M}$ in the linear range of $20\text{--}100 \mu\text{M}$ [140]. Chen et al. attached CuO, Cu₂O, and CuS on g-C₃N₄ to prepare CuX/g-C₃N₄ composite and used these composite electrodes for acetaminophen detection. **Figure 9(a)** shows the detection of acetaminophen by these synthesized nanocomposites. It was found that CuS/g-C₃N₄, CuO/g-C₃N₄, and Cu₂O/g-C₃N₄ sensors have wide linear ranges of $5\text{ to }500 \mu\text{M}$ with LOD of $0.26 \mu\text{M}$, $5\text{ to }300 \mu\text{M}$ with LOD of $0.32 \mu\text{M}$, and $5\text{ to }250 \mu\text{M}$ with LOD of $0.47 \mu\text{M}$, respectively as shown in **Figure 9(b-d)** [36]. Shi et al. prepared a zinc tetra hydroxy phthalocyanine-reduced graphene oxide nanocomposite (rGO-ZnPc-OH) as an electrode material for the acetaminophen sensing in human urine samples including drug formulation. It was observed that the formed nanocomposite provided effective electroactive surface area. The synergistic enrichment was also seen because of the adsorption of $\pi\text{--}\pi$ stacking and hydrogen bonding of hydroxyl groups [141]. Yigit et al. fabricated graphene-Nafion composite film on GCE; they used it for the simultaneous detection of acetaminophen, aspirin and caffeine in commercial tablets. The CV and ASV techniques were used for the determination of electrochemical behavior of all the mentioned drugs. The oxidation peaks of the designed electrochemical sensor were observed at 0.64 , 1.04 and 1.44 V and also demonstrated good linear current responses with detection limits of $1.2 \times 10^{-9} \text{ M}$, $6.5 \times 10^{-8} \text{ M}$ and $3.8 \times 10^{-8} \text{ M}$, respectively



[142]. Roushani et al. fabricated a low-cost electrochemical aptasensor for the detection of ibuprofen. They formed a nanocomposite with nitrogen doped GQDs and gold nanoparticles which have a unique matrix for covalently attaching the Apt molecules. The modified GCE with prepared nanocomposite provided higher surface area and electrical conductivity. In this study, riboflavin was first used for electrochemical detection of ibuprofen. The detection limit was found to be 33.33 aM. The obtained results revealed that this type of strategy can be implicated in the design of biosensors and electrochemical sensors for the detection of different targets [143]. The literature survey mentioned above is summarized in Table 1.

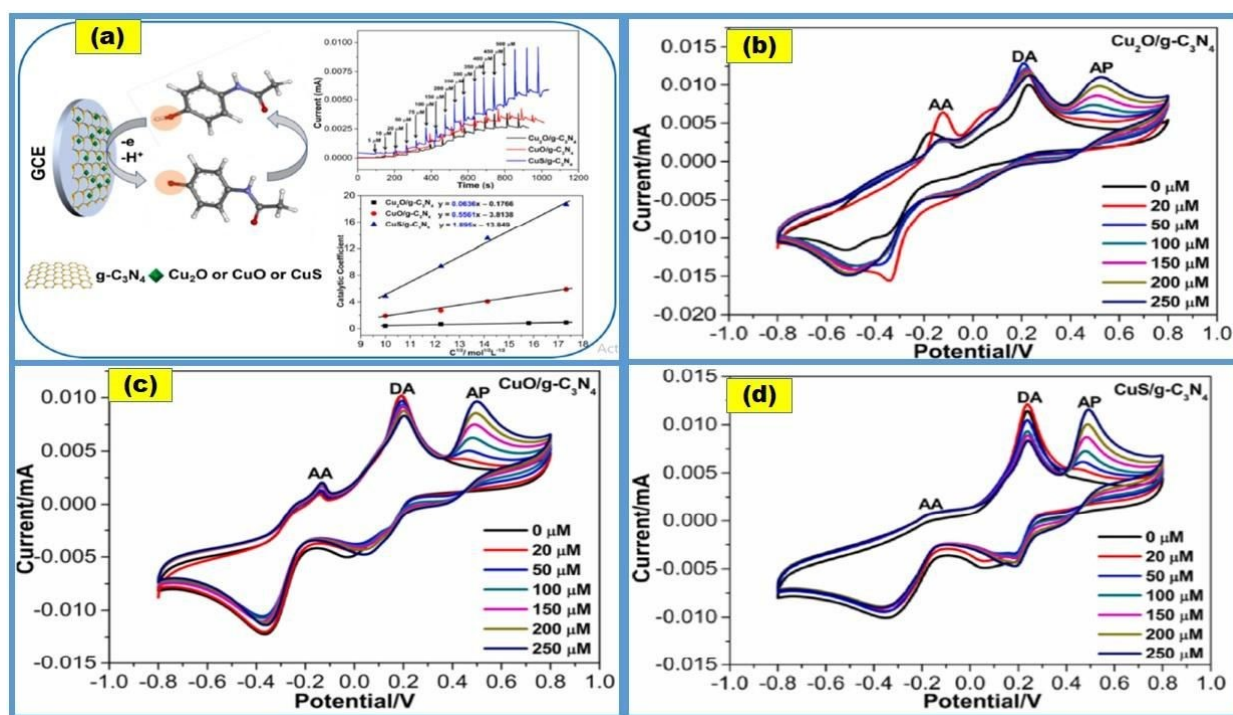


Figure 9. (a) Detection of AP by CuX/g-C₃N₄ nanocomposites-based electrochemical sensors, (b) (c), and (d) CV curves of paracetamol (AP) on different CuX-C₃N₄ modified electrodes in the presence of dopamine (DA) and ascorbic acid (AA) (Reprinted with permission from ref. [36]).



Table 1. Comparison data of nanomaterials based electrochemical sensor for the detection of drugs.

Material	Synthesis Method	Morphology	Detection Limit	Linear Range	Drug	Ref.
MWCNTs	Layer by layer	-	0.5 μM ($5 \times 10^{-7} \text{mol/L}$)	25–400 μM	AP	[144]
f-MWCNTs modified glassy carbon electrodes (GCEs)	-	densely packed	0.6 μM ($0.6 \mu\text{mol L}^{-1}$)	3–300 μM ($3\text{--}300 \mu\text{mol L}^{-1}$)	AP	[145]
SWCNT graphene nanosheet hybrid film	modified Hummers method	uniformly dispersed	$38 \times 10^{-3} \mu\text{M}$ (38 nM)	0.05–64.5 μM .	AP	[146]
Iron oxide NPs			$2.45 \times 10^{-3} \mu\text{M}$ (2.45 nM)	0.01–100.0 μM	DC F	[74]
Cu_3TeO_6	Wet chemical route	Stone like	0.017 μM	0.02–5 μM	IB	[62]
Nafion/TiO ₂ –graphene	–	–	0.21 μM ($2.1 \times 10^{-7} \text{M}$)	1–100 μM	AP	[147]



nanocomposite						
Fe ₃ O ₄ @Au–S–Fc/GS	Co-precipitation		0.05 μM		AP	[148]
MWNTs	Molecular imprinting and sol-gel	3-D network	0.04 μM (4.0 ×10 ⁻⁸ mol/L)	8.0 ×10 ⁻² –5.0 ×10 ¹¹ μM (8.0 ×10 ⁻⁸ –5.0 ×10 ⁵ mol/L)	AP	[53]
ZnO		Nanoflowers		0.1–300 μM	DA	[63]
GO/MIPs	modified Hummer’s method	Curved and layer like structure	20×10 ⁻³ μM (20 nM)	0.1 μM to 80 μM	AP	[55]
Fe ₂ O ₃ (0.5)/SnO ₂ (0.5)	Solid phase reaction	-	0.2 μM (0.2 μmol L ⁻¹)	4.5-876.0 μM (4.5–876.0 μmol L ⁻¹)	AP	[61]
MoS ₂ - Gr/GCE	Solution-phase method	3-D sphere-like	0.02 μM	0.1-100 μM	AP	[136]
RGO-Tin nanohybrid	Green synthesis	Wrinkled and flake like	0.02 μM	0.06-660 μM	AP	[66]

grapheme-Nafion composite film			6.5×10^{-8}		AS P	[142]
Graphene/SnO ₂ nanocomposite	Hydrothermal method	irregular	1 μ M		AP	[43]
ERG/GCE	-	-	0.0021/1.2 μ M	0.005–4/5–800 μ M	AP	[111]
Activated carbon-ZnO composite	Co-precipitation	Nanoflakes	0.02 μ M (0.02 μ ML ⁻¹)	0.05–1380 μ M (0.05–1380 μ mol L ⁻¹)	AP	[59]
Cd(OH) ₂ -rGO	Co-precipitation	nanorods-like	0.08 μ M	0.1 to 102 μ M	AP	[149]
CuO-CuFe ₂ O ₄	Co-precipitation	spherical	0.007 μ M (0.007 μ mol L ⁻¹)	0.01–1.5 μ M (0.01–1.5 μ mol L ⁻¹)	AP	[128]
γ -Fe ₂ O ₃	Co-precipitation	-	75 μ M (0.075 mM)	31–1000 μ M (3.1 $\times 10^{-5}$ M to 1.0 $\times 10^{-3}$ M)	AP	[150]



P-RGO	Hydrothermal method	-	0.36 μM	1.5–120 μM	AP	[44]
Graphene/ZrO ₂	Green synthesis	Sheets	$75.5 \times 10^{-3} \mu\text{M}$ (75.5 nM)	10 - 100 μM	AP	[129]
Chitosan-Au NPs	Chemical route	Dispersed	0.03 pg/mL	-	AS P	[107]
Pd/POMs/NHCSs	Green synthesis	Nanospheres	$3 \times 10^{-3} \mu\text{M}$ (3 nM)	0.63 μM to 0.083 mM	AP	[151]
La ³⁺ /Co ₃ O ₄	Co-precipitation	nanoflower	0.09 μM	0.5 μM to 250.0 μM	AP	[60]
SI-CPE	-	-	$21 \times 10^{-3} \mu\text{M}$ (21 nm)	1–160 μM	AP	[152]
Ag NPs/GCE	Green synthesis	–	$8.50 \times 10^{-2} \mu\text{M}$ ($8.50 \times 10^{-8} \text{mol L}^{-1}$)	-	AP	[64]
NiO/GCE	Hydrothermal method	asymmetric	0.23 μM	7.5-3000 μM	AP	[42]
CeO ₂ -CNT	Green synthesis		$4.4 \times 10^{-3} \mu\text{M}$ (4.4 nM)	0.01–900.0 μM	AP	[65]
rGO/AuNPs/MWCNTs	Chemical route	Tubular/ flaky	$42 \times 10^{-3} \mu\text{M}$ (42 nM)	0.12~12 μM	AP	[153]

Cu ₂ O– CuO/rGO/CP E	Hummers' method/ chemical reduction	-	0.003 μM	0.008–13 μM	AP	[154]
SrP/g-CN NSs	Thermal polymerizat ion technique	Nanosheets	2×10 ⁻³ μM (2.0 nM)	0.01 to 370 μM	AP	[155]
Ni/C- 400/GCE	Direct calcinations method	Spherical	4.04×10 ⁻² μM	0.20–53.75 μM	AP	[156]
GCE	-	-	0.96 μM	1.45- 3.87 μmol L ⁻¹	IB	[126]
PASMCNTM GPE	-	Thick layer	0.04 μM	20–100 μM	AP	[140]
BC/Co ₃ O ₄ / FeCo ₂ O ₄	Calcination treatment with hydrotherm al	Nanosheets	0.02886 μM	0.1- 220 μM	AP	[47]



rGO-ZnPc-OH nanocomposite	Hummers method/chemical route	Gauzy crinkled and folded nanosheets	$10^{-2}\mu\text{M}$ (10 nM)	100 to 800 μM	AP	[141]
(WP6- Pd-COF) nanocomposite	Chemical route	-	$3\times 10^4\mu\text{M}$ (0.03M)	0.1 - 7.5 μM	AP	[157]

6. Challenges and Future Prospects

Over the past few years, there has been a significant increase in the need for precise and affordable methods for the detection and quality of pharmaceutical compounds. As a result, there has been a lot of interest in and development of specialized nanomaterial-based electrochemical sensors. For the full utilization of the capabilities of advanced electrochemical sensors, a thorough understanding of the physicochemical and electronic interactions that take place at the interfaces between nanomaterials and relevant analytes is required. In this review, the most recent developments in the creation of electrochemical sensors based on nanomaterials for the detection of important pharmaceutical drugs such as acetaminophen, Ibuprofen, aminophenol, Diclofenac, Dopamine, and others are covered in-depth. Noteworthy progress has been made in combining carbon-based nanomaterials with metal nanoparticles, quantum dots, organic functional groups, and conductive polymers to produce powerful synergistic effects that will make it easier to catalyze reactions with target analytes. Although sensitivity and LOD were often improved with these methods, it also causes other problems like fouling and the non-specific adsorption of other species



that make it more challenging to commercialize the proposed electrochemical sensors. In fact, few researchers have questioned whether adding carbon-based nanomaterials like graphene, GO, and CNTs had a synergistic effect on the electrochemical detection of analytes. It was not always confirmed that carbon nanomaterials or the modified materials themselves would produce a signal that was comparable. Moreover, the improved performance observed after using carbon nanomaterials may simply be the result of a rise in the active surface area. For tailoring the design of sensing platforms, a thorough understanding of the electrochemical reactions at electrode/electrolyte interfaces remains a foremost challenge. Despite the fact that most proposed electrochemical sensors have been tested with real samples but there is a significant gap between laboratory tests and the commercialization of these sensing devices. To develop and commercialize electrochemical sensors and biosensors for the efficient detection of pharmaceutical drugs, the industry, as well as multidisciplinary research groups, must collaborate closely. Electrochemical sensors and biosensors must be carefully analyzed with regard to their costs and stability in addition to their sensitivity and selectivity. For the analysis of pharmaceuticals, chromatographic techniques such as HPLC, GC-MS, and LC-MS/MS as well as colorimetric methods are currently widely employed in hospitals and laboratories. However, these techniques are often constrained in terms of their portability. Electrochemical sensors, on the other hand, have advantages that allow them to surpass the limitations of those traditional methods. Using liquid chromatography or mass spectrometry in conjunction with electrochemical sensing for effective drug detection is one of the newest trends. Over the past few decades, remarkable work has been made in the advancement of sophisticated electrochemical sensors and biosensors for the sensing of pharmaceutical compounds. An encouraging trend for further improvement in the sensitivity and reduction in the detection limits for target analytes through synergistic effects is the use of nanocomposites made



of carbon and other nanomaterials like metal nanoparticles, polymers, and functionalized nanostructures.

7. Conclusions:

This review summarizes the recent advances in the fabrication of electrochemical sensors used for analgesic antipyretic drug detection. We have discussed some commonly used drugs such as acetaminophen, ibuprofen, aspirin, and diclofenac. We deliberately discussed synthesis methods of nanomaterials and their utilization in electrochemical sensors for the detection of drugs. Generally, the special properties of metals and carbon-based nanomaterials have considerably contributed to the advancement of electrochemical sensors. Both the novel and adjusted metal-based probes often show improved analytical performance over traditional non-nanostructured electrochemical frameworks. Electroanalytical techniques utilizing sensing and biosensing devices, including carbon and metal oxide-based nanostructure modified electrodes, are promising for real-life analytical detection applications. Specifically, diamonds and CNTs have been used as electrode materials for electrochemical sensing. Albeit a few challenges still stay, for instance, scalability and reproducibility of current "nano" gadgets, the sensing frameworks are especially influenced by the properties of the nanostructures utilized. However, greater suitable assessments of the few performance properties, including their application for detecting analytes in real world samples, are essential before potential commercialization.

CRedit authorship contribution statement

Manika Chaudhary: Role/Writing – original draft, Data curation, Validation; **Ashwani Kumar:** Review editing & Software; **Arti Devi:** Role/Writing – original draft, Investigation; **Beer Pal Singh:** Supervision, Resources, Conceptualization; **Bansi Dhar Malhotra:** Supervision, Writing – review & editing; **Kushagr Singhal:** Review, editing, **Sangeeta Shukla:** Resources, Project



administration; **Carmen A Vega-Olivencia:** Conceptualization, **Srikanth Ponnada:** Formal analysis and review editing; **Rakesh K. Sharma:** Supervision & review editing, **Shrestha Tyagi:** Software; **Rahul Singhal:** Writing – review & editing, Conceptualization, Visualization, Methodology, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships.

Acknowledgement:

One of us (BDM) Acknowledges the Science and Engineering Board (DST-SERB, Govt. of India) for the award of Distinguished Fellowship (SB/DF/011/2019).

References

1. Brownstein, H.H., *Drugs and Society*, in *The Handbook of Drugs and Society*. 2015. p. 1-13.
2. JANE-LLOPIS, E. and I. MATYTSINA, *Mental health and alcohol, drugs and tobacco: a review of the comorbidity between mental disorders and the use of alcohol, tobacco and illicit drugs*. *Drug and Alcohol Review*, 2006. **25**(6): p. 515-536, <https://doi.org/10.1080/09595230600944461>.
3. van Amsterdam, J., A. Opperhuizen, M. Koeter, and W. van den Brink, *Ranking the harm of alcohol, tobacco and illicit drugs for the individual and the population*. *Eur Addict Res*, 2010. **16**(4): p. 202-7, 10.1159/000317249.
4. Peppin, J., K. Dineen, A. Ruggles, and J. Coleman, *Prescription Drug Diversion and Pain: History, Policy, and Treatment*. 2018: Oxford University Press.



5. Agnieszka, S., *Global Pharmaceutical Industry: Characteristics and Trends*, in *Global Supply Chains in the Pharmaceutical Industry*, N. Hamed and S. Agnieszka, Editors. 2019, IGI Global: Hershey, PA, USA. p. 57-85.
6. Lunenfeld, B. and P. Stratton, *The clinical consequences of an aging world and preventive strategies*. Best Practice & Research Clinical Obstetrics & Gynaecology, 2013. **27**(5): p. 643-659, <https://doi.org/10.1016/j.bpobgyn.2013.02.005>.
7. Kaushik, A. and E. Mostafavi, *To manage long COVID by selective SARS-CoV-2 infection biosensing*. Innovation (Camb), 2022. **3**(5): p. 100303, 10.1016/j.xinn.2022.100303.
8. Keyvanfard, M., R. Shakeri, H. Karimi-Maleh, and K. Alizad, *Highly selective and sensitive voltammetric sensor based on modified multiwall carbon nanotube paste electrode for simultaneous determination of ascorbic acid, acetaminophen and tryptophan*. Materials Science and Engineering: C, 2013. **33**(2): p. 811-816, <https://doi.org/10.1016/j.msec.2012.11.005>.
9. Amiri, M., F. Rezapour, and A. Bezaatpour, *Hydrophilic carbon nanoparticulates at the surface of carbon paste electrode improve determination of paracetamol, phenylephrine and dextromethorphan*. Journal of Electroanalytical Chemistry, 2014. **735**: p. 10-18, 10.1016/j.jelechem.2014.10.006.
10. Li, J., J. Liu, G. Tan, J. Jiang, S. Peng, M. Deng, D. Qian, Y. Feng, and Y. Liu, *High-sensitivity paracetamol sensor based on Pd/graphene oxide nanocomposite as an enhanced electrochemical sensing platform*. Biosensors and Bioelectronics, 2014. **54**: p. 468-475, <https://doi.org/10.1016/j.bios.2013.11.001>.



11. Olaleye, M.T. and B.T.J. Rocha, *Acetaminophen-induced liver damage in mice: Effects of some medicinal plants on the oxidative defense system*. Experimental and Toxicologic Pathology, 2008. **59**(5): p. 319-327, <https://doi.org/10.1016/j.etp.2007.10.003>.
12. Wu, X., Y. Wu, H. Dong, J. Zhao, C. Wang, S. Zhou, J. Lu, Y. Yan, and H. Li, *Accelerating the design of molecularly imprinted nanocomposite membranes modified by Au@ polyaniline for selective enrichment and separation of ibuprofen*. Applied Surface Science, 2018. **428**: p. 555-565, <https://doi.org/10.1016/j.apsusc.2017.09.104>.
13. Parlak, C. and Ö. Alver, *Adsorption of ibuprofen on silicon decorated fullerenes and single walled carbon nanotubes: A comparative DFT study*. Journal of Molecular Structure, 2019. **1184**: p. 110-113, <https://doi.org/10.1016/j.molstruc.2019.02.023>.
14. Dinç-Zor, Ş. and Ö. Aksu Dönmez, *Box-Behnken Design-Desirability Function Approach in Optimization of HPLC Method for Simultaneous Determination of Ibuprofen Along with Additives in Syrup Formulation*. Journal of AOAC International, 2021. **104**(1): p. 78-83, <https://doi.org/10.1093/jaoacint/qsaa096>.
15. Mahmoud, E.-S., A. Omar, A.M. Bayoumy, and M. Ibrahim, *Chitosan ibuprofen interaction: modeling approach*. Sensor Letters, 2018. **16**(5): p. 347-355, <https://doi.org/10.1166/sl.2018.3956>.
16. Cao, C., R. Jin, H. Wei, Z. Liu, S. Ni, G.-J. Liu, H.A. Young, X. Chen, and G. Liu, *Adaptive in vivo device for theranostics of inflammation: Real-time monitoring of interferon- γ and aspirin*. Acta biomaterialia, 2020. **101**: p. 372-383, <https://doi.org/10.1016/j.actbio.2019.10.021>.
17. Han, X.-J., X.-F. Ji, Q. Zhang, J.-W. Sun, P.-X. Sun, W.-J. Pan, J. Wang, and C. Yang, *Giant “molecular capacitor” arrays - portable sensors to determine ionizable compounds*.



- Journal of Electroanalytical Chemistry, 2020. **865**: p. 114108, 10.1016/j.jelechem.2020.114108.
18. Abdel-Haleem, F.M. and E.M. Zahran, *Miniaturization overcomes macro sample analysis limitations: Salicylate-selective polystyrene nanoparticle-modified optical sensor*. Talanta, 2019. **196**: p. 436-441, 10.1016/j.talanta.2018.12.073.
19. Altman, R., B. Bosch, K. Brune, P. Patrignani, and C. Young, *Advances in NSAID development: evolution of diclofenac products using pharmaceutical technology*. Drugs, 2015. **75**(8): p. 859-77, 10.1007/s40265-015-0392-z.
20. Qian, L., S. Durairaj, S. Prins, and A. Chen, *Nanomaterial-based electrochemical sensors and biosensors for the detection of pharmaceutical compounds*. Biosensors and Bioelectronics, 2021. **175**: p. 112836, 10.1016/j.bios.2020.112836.
21. Daneshgar, P., P. Norouzi, M.R. Ganjali, R. Dinarvand, and A.A. Moosavi-Movahedi, *Determination of diclofenac on a dysprosium nanowire- modified carbon paste electrode accomplished in a flow injection system by advanced filtering*. Sensors (Basel), 2009. **9**(10): p. 7903-18, 10.3390/s91007903.
22. Dowling, G., P. Gallo, S. Fabbrocino, L. Serpe, and L. Regan, *Determination of ibuprofen, ketoprofen, diclofenac and phenylbutazone in bovine milk by gas chromatography-tandem mass spectrometry*. Food Addit Contam Part A Chem Anal Control Expo Risk Assess, 2008. **25**(12): p. 1497-508, 10.1080/02652030802383160.
23. Burgot, G., F. Auffret, and J.L. Burgot, *Determination of acetaminophen by thermometric titrimetry*. Analytica Chimica Acta, 1997. **343**(1): p. 125-128, [https://doi.org/10.1016/S0003-2670\(96\)00613-7](https://doi.org/10.1016/S0003-2670(96)00613-7).



24. Ni, Y., C. Liu, and S. Kokot, *Simultaneous kinetic spectrophotometric determination of acetaminophen and phenobarbital by artificial neural networks and partial least squares*. *Analytica Chimica Acta*, 2000. **419**: p. 185-196, 10.1016/S0003-2670(00)00978-8.
25. Easwaramoorthy, D., Y.-C. Yu, and H.-J. Huang, *Chemiluminescence detection of paracetamol by a luminol-permanganate based reaction*. *Analytica Chimica Acta*, 2001. **439**(1): p. 95-100, [https://doi.org/10.1016/S0003-2670\(01\)00968-0](https://doi.org/10.1016/S0003-2670(01)00968-0).
26. Nebot, C., S.W. Gibb, and K.G. Boyd, *Quantification of human pharmaceuticals in water samples by high performance liquid chromatography–tandem mass spectrometry*. *Analytica Chimica Acta*, 2007. **598**(1): p. 87-94, <https://doi.org/10.1016/j.aca.2007.07.029>.
27. Moreira, A.B., H.P.M. Oliveira, T.D.Z. Atvars, I.L.T. Dias, G.O. Neto, E.A.G. Zagatto, and L.T. Kubota, *Direct determination of paracetamol in powdered pharmaceutical samples by fluorescence spectroscopy*. *Analytica Chimica Acta*, 2005. **539**(1): p. 257-261, <https://doi.org/10.1016/j.aca.2005.03.012>.
28. Goyal, R.N., V.K. Gupta, and S. Chatterjee, *Voltammetric biosensors for the determination of paracetamol at carbon nanotube modified pyrolytic graphite electrode*. *Sensors and Actuators B: Chemical*, 2010. **149**(1): p. 252-258, <https://doi.org/10.1016/j.snb.2010.05.019>.
29. Dalmaso, P.R., M.L. Pedano, and G.A. Rivas, *Electrochemical determination of ascorbic acid and paracetamol in pharmaceutical formulations using a glassy carbon electrode modified with multi-wall carbon nanotubes dispersed in polyhistidine*. *Sensors and Actuators B: Chemical*, 2012. **173**: p. 732-736, <https://doi.org/10.1016/j.snb.2012.07.087>.



30. Kachoosangi, R.T., G.G. Wildgoose, and R.G. Compton, *Sensitive adsorptive stripping voltammetric determination of paracetamol at multiwalled carbon nanotube modified basal plane pyrolytic graphite electrode*. *Analytica Chimica Acta*, 2008. **618**(1): p. 54-60, <https://doi.org/10.1016/j.aca.2008.04.053>.
31. Kaya, S.I., S. Kurbanoglu, and S.A. Ozkan, *Nanomaterials-Based Nanosensors for the Simultaneous Electrochemical Determination of Biologically Important Compounds: Ascorbic Acid, Uric Acid, and Dopamine*. *Critical Reviews in Analytical Chemistry*, 2019. **49**(2): p. 101-125, 10.1080/10408347.2018.1489217.
32. Chaudhary, V., N. Ashraf, M. Khalid, R. Walvekar, Y. Yang, A. Kaushik, and Y.K. Mishra, *Emergence of MXene–Polymer Hybrid Nanocomposites as High-Performance Next-Generation Chemiresistors for Efficient Air Quality Monitoring*. *Advanced Functional Materials*, 2022. **32**(33): p. 2112913, 10.1002/adfm.202112913.
33. Manickam, P., S.A. Mariappan, S.M. Murugesan, S. Hansda, A. Kaushik, R. Shinde, and S.P. Thipperudraswamy, *Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare*. *Biosensors (Basel)*, 2022. **12**(8), 10.3390/bios12080562.
34. Chaudhary, V., A. Kaushik, H. Furukawa, and A. Khosla, *Review—Towards 5th Generation AI and IoT Driven Sustainable Intelligent Sensors Based on 2D MXenes and Borophene*. *ECS Sensors Plus*, 2022. **1**(1): p. 013601, 10.1149/2754-2726/ac5ac6.
35. Montaseri, H. and P.B.C. Forbes, *Analytical techniques for the determination of acetaminophen: A review*. *TrAC Trends in Analytical Chemistry*, 2018. **108**: p. 122-134, 10.1016/j.trac.2018.08.023.



36. Qian, L., S. Durairaj, S. Prins, and A. Chen, *Nanomaterial-based electrochemical sensors and biosensors for the detection of pharmaceutical compounds*. Biosens Bioelectron, 2021. **175**: p. 112836, 10.1016/j.bios.2020.112836.
37. Li, J., J. Liu, G. Tan, J. Jiang, S. Peng, M. Deng, D. Qian, Y. Feng, and Y. Liu, *High-sensitivity paracetamol sensor based on Pd/graphene oxide nanocomposite as an enhanced electrochemical sensing platform*. Biosens Bioelectron, 2014. **54**: p. 468-75, 10.1016/j.bios.2013.11.001.
38. Adekunle, A.S., B.O. Agboola, J. Pillay, and K.I. Ozoemena, *Electrocatalytic detection of dopamine at single-walled carbon nanotubes–iron (III) oxide nanoparticles platform*. Sensors and Actuators B: Chemical, 2010. **148**(1): p. 93-102, 10.1016/j.snb.2010.03.088.
39. Ozcelikay, G., S. Kurbanoglu, A. Yarman, F.W. Scheller, and S.A. Ozkan, *Au-Pt nanoparticles based molecularly imprinted nanosensor for electrochemical detection of the lipopeptide antibiotic drug Daptomycin*. Sensors and Actuators B: Chemical, 2020. **320**: p. 128285, 10.1016/j.snb.2020.128285.
40. Gan, Y.X., A.H. Jayatissa, Z. Yu, X. Chen, and M. Li, *Hydrothermal Synthesis of Nanomaterials*. Journal of Nanomaterials, 2020. **2020**: p. 8917013, 10.1155/2020/8917013.
41. Khan, F.A., *Synthesis of nanomaterials: methods & technology*, in *Applications of Nanomaterials in Human Health*. 2020, Springer. p. 15-21.
42. Annadurai, K., V. Sudha, G. Murugadoss, and R. Thangamuthu, *Electrochemical sensor based on hydrothermally prepared nickel oxide for the determination of 4-acetaminophen in paracetamol tablets and human blood serum samples*. Journal of Alloys and Compounds, 2021. **852**: p. 156911, 10.1016/j.jallcom.2020.156911.



43. Nurzulaikha, R., H.N. Lim, I. Harrison, S.S. Lim, A. Pandikumar, N.M. Huang, S.P. Lim, G.S.H. Thien, N. Yusoff, and I. Ibrahim, *Graphene/SnO₂ nanocomposite-modified electrode for electrochemical detection of dopamine*. Sensing and Bio-Sensing Research, 2015. **5**: p. 42-49, 10.1016/j.sbsr.2015.06.002.
44. Zhang, X., K.P. Wang, L.N. Zhang, Y.C. Zhang, and L. Shen, *Phosphorus-doped graphene-based electrochemical sensor for sensitive detection of acetaminophen*. Anal Chim Acta, 2018. **1036**: p. 26-32, 10.1016/j.aca.2018.06.079.
45. Xu, Z., H. Teng, J. Song, F. Gao, L. Ma, G. Xu, and X. Luo, *A nanocomposite consisting of MnO₂ nanoflowers and the conducting polymer PEDOT for highly sensitive amperometric detection of paracetamol*. Microchimica Acta, 2019. **186**(8): p. 499, 10.1007/s00604-019-3614-3.
46. Ponnada, S., D.B. Gorle, M.S. Kiai, S. Rajagopal, R.K. Sharma, and A. Nowduri, *A facile, cost-effective, rapid, single-step synthesis of Ag–Cu decorated ZnO nanoflower-like composites (NFLCs) for electrochemical sensing of dopamine*. Materials Advances, 2021. **2**(18): p. 5986-5996, 10.1039/d1ma00319d.
47. Lu, Z., J. Zhong, Y. Zhang, M. Sun, P. Zou, H. Du, X. Wang, H. Rao, and Y. Wang, *MOF-derived Co₃O₄/FeCo₂O₄ incorporated porous biomass carbon: Simultaneous electrochemical determination of dopamine, acetaminophen and xanthine*. Journal of Alloys and Compounds, 2021. **858**: p. 157701, 10.1016/j.jallcom.2020.157701.
48. Razmi, E.D., H. Beitollahi, M.T. Mahani, and M. Anjomshoa, *TiO₂/Fe₃O₄/Multiwalled Carbon Nanotubes Nanocomposite as Sensing Platform for Simultaneous Determination of Morphine and Diclofenac at a Carbon Paste Electrode*. Russian Journal of Electrochemistry, 2018. **54**(12): p. 1132-1140, 10.1134/S1023193518140057.



49. Sajjadi, S.P., *Sol-gel process and its application in Nanotechnology*. J. Polym. Eng. Technol, 2005. **13**: p. 38-41.
50. Sakka, S., *Sol-Gel Process and Applications*. 2013: p. 883-910, 10.1016/b978-0-12-385469-8.00048-4.
51. Bagherinasab, Z., H. Beitollahi, M. Yousefi, M. Bagherzadeh, and M. Hekmati, *Rapid sol gel synthesis of BaFe12O19 nanoparticles: An excellent catalytic application in the electrochemical detection of tramadol in the presence of acetaminophen*. Microchemical Journal, 2020. **156**: p. 104803, <https://doi.org/10.1016/j.microc.2020.104803>.
52. Deiminiat, B., G.H. Rounaghi, and M.H. Arbab-Zavar, *Development of a new electrochemical imprinted sensor based on poly-pyrrole, sol-gel and multiwall carbon nanotubes for determination of tramadol*. Sensors and Actuators B: Chemical, 2017. **238**: p. 651-659, 10.1016/j.snb.2016.07.110.
53. Zhu, A., G. Xu, L. Li, L. Yang, H. Zhou, and X. Kan, *Sol-Gel Imprinted Polymers Based Electrochemical Sensor for Paracetamol Recognition and Detection*. Analytical Letters, 2013. **46**(7): p. 1132-1144, 10.1080/00032719.2012.753607.
54. Anancia Grace, A., K.P. Divya, V. Dharuman, and J.H. Hahn, *Single step sol-gel synthesized Mn2O3-TiO2 decorated graphene for the rapid and selective ultra sensitive electrochemical sensing of dopamine*. Electrochimica Acta, 2019. **302**: p. 291-300, <https://doi.org/10.1016/j.electacta.2019.02.053>.
55. Luo, J., J. Cong, R. Fang, X. Fei, and X. Liu, *One-pot synthesis of a graphene oxide coated with an imprinted sol-gel for use in electrochemical sensing of paracetamol*. Microchimica Acta, 2014. **181**(11-12): p. 1257-1266, 10.1007/s00604-014-1237-2.



56. Rouhani, M. and A. Soleymanpour, *Molecularly imprinted sol-gel electrochemical sensor for sildenafil based on a pencil graphite electrode modified by Preyssler heteropolyacid/gold nanoparticles/MWCNT nanocomposite*. *Mikrochim Acta*, 2020. **187**(9): p. 512, 10.1007/s00604-020-04482-6.
57. Petcharoen, K. and A. Sirivat, *Synthesis and characterization of magnetite nanoparticles via the chemical co-precipitation method*. *Materials Science and Engineering B, Solid-State Materials for Advanced Technology*, 2012. **177**(5): p. 421-427, DOI:10.1016/j.jmseb.2012.01.003.
58. Singh, B.P., A. Kumar, A.P. Duarte, S.J. Rojas, M. Crespo-Medina, H.I. Areizaga-Martinez, C.A. Vega-Olivencia, and M.S. Tomar, *Synthesis, characterization, and electrochemical response of iron oxide nanoparticles for sensing acetaminophen*. *Materials Research Express*, 2016. **3**(10): p. 106105, 10.1088/2053-1591/3/10/106105.
59. Sivakumar, M., *Activated Carbon -ZnO Nanocomposite for Electrochemical Sensing of Acetaminophen*. *International Journal of Electrochemical Science*, 2016: p. 8363-8373, 10.20964/2016.10.51.
60. Sheikhshoaie, I., F. Garakani Nejad, and H. Beitollahi, *An electrochemical acetaminophen sensor based on La₃+/Co₃O₄ nanoflowers modified graphite screen printed electrode architecture*. *International Journal of Nano Dimension*, 2019. **10**(2): p. 154-162, 20.1001.1.20088868.2019.10.2.3.6.
61. Taei, M., M. Shavakhi, H. Hadadzadeh, M. Movahedi, M. Rahimi, and S. Habibollahi, *Simultaneous determination of epinephrine, acetaminophen, and tryptophan using Fe₂O₃(0.5)/SnO₂(0.5) nanocomposite sensor*. *Journal of Applied Electrochemistry*, 2014. **45**(2): p. 185-195, 10.1007/s10800-014-0756-1.



62. Mutharani, B., R. Rajakumaran, S.-M. Chen, P. Ranganathan, T.-W. Chen, D.A. Al Farraj, M. Ajmal Ali, and F.M.A. Al-Hemaid, *Facile synthesis of 3D stone-like copper tellurate (Cu₃TeO₆) as a new platform for anti-inflammatory drug ibuprofen sensor in human blood serum and urine samples*. Microchemical Journal, 2020. **159**: p. 105378, 10.1016/j.microc.2020.105378.
63. Zhang, C., Z. Cao, G. Zhang, Y. Yan, X. Yang, J. Chang, Y. Song, Y. Jia, P. Pan, W. Mi, Z. Yang, J. Zhao, and J. Wei, *An electrochemical sensor based on plasma-treated zinc oxide nanoflowers for the simultaneous detection of dopamine and diclofenac sodium*. Microchemical Journal, 2020. **158**: p. 105237, <https://doi.org/10.1016/j.microc.2020.105237>.
64. Zamarchi, F. and I.C. Vieira, *Determination of paracetamol using a sensor based on green synthesis of silver nanoparticles in plant extract*. J Pharm Biomed Anal, 2021. **196**: p. 113912, 10.1016/j.jpba.2021.113912.
65. Iranmanesh, T., M.M. Foroughi, S. Jahani, M. Shahidi Zandi, and H. Hassani Nadiki, *Green and facile microwave solvent-free synthesis of CeO₂ nanoparticle-decorated CNTs as a quadruplet electrochemical platform for ultrasensitive and simultaneous detection of ascorbic acid, dopamine, uric acid and acetaminophen*. Talanta, 2020. **207**: p. 120318, 10.1016/j.talanta.2019.120318.
66. Kong, F.-Y., S.-X. Gu, J.-Y. Wang, H.-L. Fang, and W. Wang, *Facile green synthesis of graphene–titanium nitride hybrid nanostructure for the simultaneous determination of acetaminophen and 4-aminophenol*. Sensors and Actuators B: Chemical, 2015. **213**: p. 397-403, 10.1016/j.snb.2015.02.120.



67. Wang, K., C. Wu, F. Wang, N. Jing, and G. Jiang, *Co/Co₃O₄ Nanoparticles Coupled with Hollow Nanoporous Carbon Polyhedrons for the Enhanced Electrochemical Sensing of Acetaminophen*. ACS Sustainable Chemistry & Engineering, 2019. **7**(22): p. 18582-18592, 10.1021/acssuschemeng.9b04813.
68. Avinash, B., C.R. Ravikumar, M.R.A. Kumar, H.P. Nagaswarupa, M.S. Santosh, A.S. Bhatt, and D. Kuznetsov, *Nano CuO: Electrochemical sensor for the determination of paracetamol and d-glucose*. Journal of Physics and Chemistry of Solids, 2019. **134**: p. 193-200, 10.1016/j.jpcs.2019.06.012.
69. Kenarkob, M. and Z. Pourghobadi, *Electrochemical sensor for acetaminophen based on a glassy carbon electrode modified with ZnO/Au nanoparticles on functionalized multi-walled carbon nano-tubes*. Microchemical Journal, 2019. **146**: p. 1019-1025, 10.1016/j.microc.2019.02.038.
70. Haridas, V., Z. Yaakob, R.N. K, S. Sugunan, and B.N. Narayanan, *Selective electrochemical determination of paracetamol using hematite/graphene nanocomposite modified electrode prepared in a green chemical route*. Materials Chemistry and Physics, 2021. **263**: p. 124379, <https://doi.org/10.1016/j.matchemphys.2021.124379>.
71. Fu, L., A. Wang, G. Lai, C.-T. Lin, J. Yu, A. Yu, Z. Liu, K. Xie, and W. Su, *A glassy carbon electrode modified with N-doped carbon dots for improved detection of hydrogen peroxide and paracetamol*. Microchimica Acta, 2018. **185**(2): p. 87, 10.1007/s00604-017-2646-9.
72. Pandey, P.A., G.R. Bell, J.P. Rourke, A.M. Sanchez, M.D. Elkin, B.J. Hickey, and N.R. Wilson, *Physical Vapor Deposition of Metal Nanoparticles on Chemically Modified*



- Graphene: Observations on Metal–Graphene Interactions*. Small, 2011. **7**(22): p. 3202-3210, <https://doi.org/10.1002/sml.201101430>.
73. Park, J.-S., W.-H. Chung, H.-S. Kim, and Y.-B. Kim, *Rapid fabrication of chemical-solution-deposited $\text{La}_{0.6}\text{Sr}_{0.4}\text{CoO}_{3-\delta}$ thin films via flashlight sintering*. Journal of Alloys and Compounds, 2017. **696**: p. 102-108, <https://doi.org/10.1016/j.jallcom.2016.11.074>.
74. Khoobi, A., N. Soltani, and M. Aghaei, *Computational design and multivariate statistical analysis for electrochemical sensing platform of iron oxide nanoparticles in sensitive detection of anti-inflammatory drug diclofenac in biological fluids*. Journal of Alloys and Compounds, 2020. **831**: p. 154715, <https://doi.org/10.1016/j.jallcom.2020.154715>.
75. Swihart, M.T., *Vapor-phase synthesis of nanoparticles*. Current Opinion in Colloid & Interface Science, 2003. **8**(1): p. 127-133, [https://doi.org/10.1016/S1359-0294\(03\)00007-4](https://doi.org/10.1016/S1359-0294(03)00007-4).
76. Vanecht, E., *Gold Nanoparticles in Ionic Liquids Prepared by Sputter Deposition*. 2012.
77. Soganci, T., R. Ayranci, E. Harputlu, K. Ocakoglu, M. Acet, M. Farle, C.G. Unlu, and M. Ak, *An effective non-enzymatic biosensor platform based on copper nanoparticles decorated by sputtering on CVD graphene*. Sensors and Actuators B: Chemical, 2018. **273**: p. 1501-1507, <https://doi.org/10.1016/j.snb.2018.07.064>.
78. Woermann, D., *J. Janata: Principles of Chemical Sensors*. Plenum Press, New York and London 1989. 317 Seiten, Preis in Europa: US \$47.40. 1990, Wiley Online Library.
79. Wang, J., *Electroanalytical techniques in clinical chemistry and laboratory medicine*. 1988: John Wiley & Sons.
80. Janata, J., *Chemical sensors*. Analytical Chemistry, 1992. **64**(12): p. 196-219, [10.1021/ac00036a012](https://doi.org/10.1021/ac00036a012).



81. Widrig, C.A., M.D. Porter, M.D. Ryan, T.G. Strein, and A.G. Ewing, *Dynamic electrochemistry: methodology and application*. Analytical Chemistry, 1990. **62**(12): p. 1-20, 10.1021/ac00211a001.
82. Bowers, L.D. and P.W. Carr, *Applications of immobilized enzymes in analytical chemistry*. Analytical Chemistry, 1976. **48**(7): p. 544A-559a, 10.1021/ac60371a033.
83. Weetall, H.H., *Immobilized enzymes. Analytical applications*. Analytical Chemistry, 1974. **46**(7): p. 602A-615a, 10.1021/ac60343a035.
84. Lima, A.P., A.C. Catto, E. Longo, E. Nossol, E.M. Richter, and R.A.A. Munoz, *Investigation on acid functionalization of double-walled carbon nanotubes of different lengths on the development of amperometric sensors*. Electrochimica Acta, 2019. **299**: p. 762-771, 10.1016/j.electacta.2019.01.042.
85. Wang, J., *Analytical Electrochemistry*, VCH, Publishers. Inc., New York, 1994.
86. Lagowski, J.J., *Ion-Selective Electrode Methodology, Vols. I and II* (Covington, Arthur K., ed.). Journal of Chemical Education, 1981. **58**(1): p. A30, 10.1021/ed058pA30.1.
87. Stradiotto, N.R., H. Yamanaka, and M.V.B. Zanoni, *Electrochemical sensors: A powerful tool in analytical chemistry*. 2003, SciELO Brasil. p. pp. 159-173.
88. Sadek, A.Z., W. Wlodarski, K. Kalantar-Zadeh, C. Baker, and R.B. Kaner, *Doped and dedoped polyaniline nanofiber based conductometric hydrogen gas sensors*. Sensors and Actuators A: Physical, 2007. **139**(1-2): p. 53-57, 10.1016/j.sna.2006.11.033.
89. Ghosh, P., S. Biswas, and A. Kushagra, *Development of conductometric glucose sensor in nanomolar (nM) range from phantom blood serum*. Materials Today: Proceedings, 2021, 10.1016/j.matpr.2021.05.627.



90. Sun, L., Y. Guo, Y. Hu, S. Pan, and Z. Jiao, *Conductometric n-butanol gas sensor based on Tourmaline@ZnO hierarchical micro-nanostructures*. *Sensors and Actuators B: Chemical*, 2021. **337**: p. 129793, 10.1016/j.snb.2021.129793.
91. Wang, Y., Y. Zhou, Y. Wang, R. Zhang, J. Li, X. Li, and Z. Zang, *Conductometric room temperature ammonia sensors based on titanium dioxide nanoparticles decorated thin black phosphorus nanosheets*. *Sensors and Actuators B: Chemical*, 2021. **349**: p. 130770, 10.1016/j.snb.2021.130770.
92. Bard, A.J., L.R. Faulkner, and H.S. White, *Electrochemical methods: fundamentals and applications*. 2022: John Wiley & Sons.
93. Ezhil Vilian, A.T., M. Rajkumar, and S.M. Chen, *In situ electrochemical synthesis of highly loaded zirconium nanoparticles decorated reduced graphene oxide for the selective determination of dopamine and paracetamol in presence of ascorbic acid*. *Colloids Surf B Biointerfaces*, 2014. **115**: p. 295-301, 10.1016/j.colsurfb.2013.12.014.
94. Wang, J., D.B. Luo, P.A. Farias, and J.S. Mahmoud, *Adsorptive stripping voltammetry of riboflavin and other flavin analogs at the static mercury drop electrode*. *Analytical chemistry*, 1985. **57**(1): p. 158-162, <https://doi.org/10.1021/ac00279a039>.
95. Kang, X., J. Wang, H. Wu, J. Liu, I.A. Aksay, and Y. Lin, *A graphene-based electrochemical sensor for sensitive detection of paracetamol*. *Talanta*, 2010. **81**(3): p. 754-9, 10.1016/j.talanta.2010.01.009.
96. Zhang, X., Y.C. Zhang, and J.W. Zhang, *A highly selective electrochemical sensor for chloramphenicol based on three-dimensional reduced graphene oxide architectures*. *Talanta*, 2016. **161**: p. 567-573, 10.1016/j.talanta.2016.09.013.



97. Sebastian, N., W.-C. Yu, and D. Balram, *Electrochemical detection of an antibiotic drug chloramphenicol based on a graphene oxide/hierarchical zinc oxide nanocomposite*. Inorganic Chemistry Frontiers, 2019. **6**(1): p. 82-93, 10.1039/c8qi01000e.
98. Mohammed, G.I., N.H. Khraibah, A.S. Bashammakh, and M.S. El-Shahawi, *Electrochemical sensor for trace determination of timolol maleate drug in real samples and drug residues using Nafion/carboxylated-MWCNTs nanocomposite modified glassy carbon electrode*. Microchemical Journal, 2018. **143**: p. 474-483, 10.1016/j.microc.2018.08.011.
99. Chethana, B.K., S. Basavanna, and Y. Arthoba Naik, *Voltammetric Determination of Diclofenac Sodium Using Tyrosine-Modified Carbon Paste Electrode*. Industrial & Engineering Chemistry Research, 2012. **51**(31): p. 10287-10295, 10.1021/ie202921e.
100. Song, X.C., X. Wang, Y.F. Zheng, R. Ma, and H.Y. Yin, *A hydrogen peroxide electrochemical sensor based on Ag nanoparticles grown on ITO substrate*. Journal of Nanoparticle Research, 2011. **13**(10): p. 5449, 10.1007/s11051-011-0532-7.
101. Burda, C., X. Chen, R. Narayanan, and M.A. El-Sayed, *Chemistry and Properties of Nanocrystals of Different Shapes*. Chemical Reviews, 2005. **105**(4): p. 1025-1102, 10.1021/cr030063a.
102. Linting, Z., L. Ruiyi, L. Zaijun, X. Qianfang, F. Yinjun, and L. Junkang, *An immunosensor for ultrasensitive detection of aflatoxin B1 with an enhanced electrochemical performance based on graphene/conducting polymer/gold nanoparticles/the ionic liquid composite film on modified gold electrode with electrodeposition*. Sensors and Actuators B: Chemical, 2012. **174**: p. 359-365, <https://doi.org/10.1016/j.snb.2012.06.051>.



103. Ozcan, M., A. Basak, and A. Uzunoglu, *Construction of High-Performance Amperometric Acetaminophen Sensors Using Zn/ZnO-Decorated Reduced Graphene Oxide Surfaces*. ECS Journal of Solid State Science and Technology, 2020. **9**(9): p. 093003, 10.1149/2162-8777/ab951b.
104. Liu, B., X. Ouyang, Y. Ding, L. Luo, D. Xu, and Y. Ning, *Electrochemical preparation of nickel and copper oxides-decorated graphene composite for simultaneous determination of dopamine, acetaminophen and tryptophan*. Talanta, 2016. **146**: p. 114-121, <https://doi.org/10.1016/j.talanta.2015.08.034>.
105. Manikandan, P.N. and V. Dharuman, *Electrochemical Simultaneous Sensing of Melatonin, Dopamine and Acetaminophen at Platinum Doped and Decorated Alpha Iron Oxide*. Electroanalysis, 2017. **29**(6): p. 1524-1531, 10.1002/elan.201700054.
106. Cao, F., Q. Dong, C. Li, J. Chen, X. Ma, Y. Huang, D. Song, C. Ji, and Y. Lei, *Electrochemical sensor for detecting pain reliever/fever reducer drug acetaminophen based on electrospun CeBiO nanofibers modified screen-printed electrode*. Sensors and Actuators B: Chemical, 2018. **256**: p. 143-150, 10.1016/j.snb.2017.09.204.
107. Diouf, A., M. Moufid, D. Bouyahya, L. Österlund, N. El Bari, and B. Bouchikhi, *An electrochemical sensor based on chitosan capped with gold nanoparticles combined with a voltammetric electronic tongue for quantitative aspirin detection in human physiological fluids and tablets*. Materials Science and Engineering: C, 2020. **110**: p. 110665, <https://doi.org/10.1016/j.msec.2020.110665>.
108. Yang, W., K.R. Ratinac, S.P. Ringer, P. Thordarson, J.J. Gooding, and F. Braet, *Carbon Nanomaterials in Biosensors: Should You Use Nanotubes or Graphene?* Angewandte



- Chemie International Edition, 2010. **49**(12): p. 2114-2138, <https://doi.org/10.1002/anie.200903463>.
109. Alavi-Tabari, S.A.R., M.A. Khalilzadeh, and H. Karimi-Maleh, *Simultaneous determination of doxorubicin and dasatinib as two breast anticancer drugs uses an amplified sensor with ionic liquid and ZnO nanoparticle*. Journal of Electroanalytical Chemistry, 2018. **811**: p. 84-88, <https://doi.org/10.1016/j.jelechem.2018.01.034>.
 110. Alam, A.U., Y. Qin, M.M.R. Howlader, N.-X. Hu, and M.J. Deen, *Electrochemical sensing of acetaminophen using multi-walled carbon nanotube and β -cyclodextrin*. Sensors and Actuators B: Chemical, 2018. **254**: p. 896-909, [10.1016/j.snb.2017.07.127](https://doi.org/10.1016/j.snb.2017.07.127).
 111. Adhikari, B.-R., M. Govindhan, and A. Chen, *Sensitive Detection of Acetaminophen with Graphene-Based Electrochemical Sensor*. Electrochimica Acta, 2015. **162**: p. 198-204, [10.1016/j.electacta.2014.10.028](https://doi.org/10.1016/j.electacta.2014.10.028).
 112. Dou, N., S. Zhang, and J. Qu, *Simultaneous detection of acetaminophen and 4-aminophenol with an electrochemical sensor based on silver–palladium bimetal nanoparticles and reduced graphene oxide*. RSC Advances, 2019. **9**(54): p. 31440-31446, [10.1039/c9ra05987c](https://doi.org/10.1039/c9ra05987c).
 113. Wu, C., J. Li, X. Liu, H. Zhang, R. Li, G. Wang, Z. Wang, Q. Li, and E. Shangguan, *Simultaneous voltammetric determination of epinephrine and acetaminophen using a highly sensitive CoAl-OOH/reduced graphene oxide sensor in pharmaceutical samples and biological fluids*. Mater Sci Eng C Mater Biol Appl, 2021. **119**: p. 111557, [10.1016/j.msec.2020.111557](https://doi.org/10.1016/j.msec.2020.111557).
 114. Qian, Z., X. Shan, L. Chai, J. Ma, J. Chen, and H. Feng, *Si-Doped Carbon Quantum Dots: A Facile and General Preparation Strategy, Bioimaging Application, and Multifunctional*



- Sensor*. ACS Applied Materials & Interfaces, 2014. **6**(9): p. 6797-6805, 10.1021/am500403n.
115. Li, X., M. Rui, J. Song, Z. Shen, and H. Zeng, *Carbon and Graphene Quantum Dots for Optoelectronic and Energy Devices: A Review*. Advanced Functional Materials, 2015. **25**(31): p. 4929-4947, <https://doi.org/10.1002/adfm.201501250>.
116. Kumar, V., G. Toffoli, and F. Rizzolio, *Fluorescent Carbon Nanoparticles in Medicine for Cancer Therapy*. ACS Medicinal Chemistry Letters, 2013. **4**(11): p. 1012-1013, 10.1021/ml400394a.
117. Cernat, A., M. Tertis, R. Sandulescu, F. Bedioui, A. Cristea, and C. Cristea, *Electrochemical sensors based on carbon nanomaterials for acetaminophen detection: A review*. Anal Chim Acta, 2015. **886**: p. 16-28, 10.1016/j.aca.2015.05.044.
118. Gopal, T.V., T.M. Reddy, P. Shaikshavali, and G. Venkataprasad, *Eco-friendly and bio-waste based hydroxyapatite/reduced graphene oxide hybrid material for synergic electrocatalytic detection of dopamine and study of its simultaneous performance with acetaminophen and uric acid*. Surfaces and Interfaces, 2021. **24**: p. 101145, 10.1016/j.surfin.2021.101145.
119. Pham, T.S.H., P.J. Mahon, G. Lai, L. Fu, C.T. Lin, and A. Yu, *Cauliflower-like Platinum Particles Decorated Reduced Graphene Oxide for Sensitive Determination of Acetaminophen*. Electroanalysis, 2019. **31**(9): p. 1758-1768, 10.1002/elan.201900138.
120. Berto, S., L. Carena, F. Valmacco, C. Barolo, E. Conca, D. Vione, R. Buscaino, M. Fiorito, C. Bussi, O. Abollino, and M. Malandrino, *Application of an electro-activated glassy-carbon electrode to the determination of acetaminophen (paracetamol) in surface waters*. Electrochimica Acta, 2018. **284**: p. 279-286, 10.1016/j.electacta.2018.07.145.



121. Liang, W., L. Liu, Y. Li, H. Ren, T. Zhu, Y. Xu, and B.-C. Ye, *Nitrogen-rich porous carbon modified electrochemical sensor for the detection of acetaminophen*. Journal of Electroanalytical Chemistry, 2019. **855**: p. 113496, 10.1016/j.jelechem.2019.113496.
122. Tsierkezos, N.G., S.H. Othman, and U. Ritter, *Nitrogen-doped multi-walled carbon nanotubes for paracetamol sensing*. Ionics, 2013. **19**(12): p. 1897-1905, 10.1007/s11581-013-0930-1.
123. Barsan, M.M., C.T. Toledo, and C.M.A. Brett, *New electrode architectures based on poly(methylene green) and functionalized carbon nanotubes: Characterization and application to detection of acetaminophen and pyridoxine*. Journal of Electroanalytical Chemistry, 2015. **736**: p. 8-15, <https://doi.org/10.1016/j.jelechem.2014.10.026>.
124. Sarhangzadeh, K., A.A. Khatami, M. Jabbari, and S. Bahari, *Simultaneous determination of diclofenac and indomethacin using a sensitive electrochemical sensor based on multiwalled carbon nanotube and ionic liquid nanocomposite*. Journal of Applied Electrochemistry, 2013. **43**(12): p. 1217-1224, 10.1007/s10800-013-0609-3.
125. Roushani, M., Z. Rahmati, S. Farokhi, S.J. Hoseini, and R.H. Fath, *The development of an electrochemical nanoaptasensor to sensing chloramphenicol using a nanocomposite consisting of graphene oxide functionalized with (3-Aminopropyl) triethoxysilane and silver nanoparticles*. Materials Science and Engineering: C, 2020. **108**: p. 110388, <https://doi.org/10.1016/j.msec.2019.110388>.
126. Suresh, E., K. Sundaram, B. Kavitha, S.M. Rayappan, and N.S. Kumar, *Simultaneous Electrochemical Determination of Paracetamol and Ibuprofen at The Glassy Carbon Electrode*. Journal of Advanced Chemical Sciences, 2016: p. 369-372.



127. Hao, W., Y. Zhang, J. Fan, H. Liu, Q. Shi, W. Liu, Q. Peng, and G. Zang, *Copper Nanowires Modified with Graphene Oxide Nanosheets for Simultaneous Voltammetric Determination of Ascorbic Acid, Dopamine and Acetaminophen*. *Molecules*, 2019. **24**(12), 10.3390/molecules24122320.
128. Hasanpour, F., M. Taei, and S. Tahmasebi, *Ultra-sensitive electrochemical sensing of acetaminophen and codeine in biological fluids using CuO/CuFe₂O₄ nanoparticles as a novel electrocatalyst*. *J Food Drug Anal*, 2018. **26**(2): p. 879-886, 10.1016/j.jfda.2017.10.001.
129. Lin, L.P., P.S. Khiew, W.S. Chiu, and M.T.T. Tan, *A Disposable Electrochemical Sensing Platform for Acetaminophen Based on Graphene/ZrO₂ Nanocomposite Produced via a Facile, Green Synthesis Method*. *IEEE Sensors Journal*, 2018. **18**(19): p. 7907-7916, 10.1109/jsen.2018.2864326.
130. Tamilalagan, E., S. Vetri Selvi, S.-M. Chen, M. Akilarasan, S. Maheshwaran, T.-W. Chen, A.M. Al-Mohaimeed, W.A. Al-onazi, M. Soliman Elshikh, and X. Liu, *Fabrication of p-n Junction (Ni/Zn)O and Reduced Graphene Oxide (rGO) Nanocomposites for the Electrocatalysis of Analgesic Drug (Acetaminophen) Detection in Pharmaceutical and Biological Samples*. *Journal of The Electrochemical Society*, 2021. **168**(3): p. 036501, 10.1149/1945-7111/abe6eb.
131. Nikpanje, E., M. Bahmaei, and A.M. Sharif, *Determination of Ascorbic Acid, Acetaminophen, and Caffeine in Urine, Blood Serum by Electrochemical Sensor Based on ZnO-Zn₂SnO₄-SnO₂ Nanocomposite and Graphene*. *Journal of Electrochemical Science and Technology*, 2021. **12**(2): p. 173-187, 10.33961/jecst.2020.00724.



132. Afkhami, A., H. Khoshsafar, H. Bagheri, and T. Madrakian, *Preparation of NiFe(2)O(4)/graphene nanocomposite and its application as a modifier for the fabrication of an electrochemical sensor for the simultaneous determination of tramadol and acetaminophen*. Anal Chim Acta, 2014. **831**: p. 50-9, 10.1016/j.aca.2014.04.061.
133. Demir, N., K. Atacan, M. Ozmen, and S.Z. Bas, *Design of a new electrochemical sensing system based on MoS₂-TiO₂/reduced graphene oxide nanocomposite for the detection of paracetamol*. New Journal of Chemistry, 2020. **44**(27): p. 11759-11767, 10.1039/d0nj02298e.
134. Anuar, N.S., W.J. Basirun, M. Ladan, M. Shalauddin, and M.S. Mehmood, *Fabrication of platinum nitrogen-doped graphene nanocomposite modified electrode for the electrochemical detection of acetaminophen*. Sensors and Actuators B: Chemical, 2018. **266**: p. 375-383, 10.1016/j.snb.2018.03.138.
135. Shaikshavali, P., T. Madhusudana Reddy, V.N. Palakollu, R. Karpoormath, Y. Subba Rao, G. Venkataprasad, T.V. Gopal, and P. Gopal, *Multi walled carbon nanotubes supported CuO-Au hybrid nanocomposite for the effective application towards the electrochemical determination of Acetaminophen and 4-Aminophenol*. Synthetic Metals, 2019. **252**: p. 29-39, 10.1016/j.synthmet.2019.04.009.
136. Huang, K.-J., L. Wang, J. Li, and Y.-M. Liu, *Electrochemical sensing based on layered MoS₂-graphene composites*. Sensors and Actuators B: Chemical, 2013. **178**: p. 671-677, 10.1016/j.snb.2013.01.028.
137. Kimuam, K., N. Rodthongkum, N. Ngamrojanavanich, O. Chailapakul, and N. Ruecha, *Single step preparation of platinum nanoflowers/reduced graphene oxide electrode as a*



- novel platform for diclofenac sensor*. Microchemical Journal, 2020. **155**: p. 104744, 10.1016/j.microc.2020.104744
138. Goyal, R.N., S. Chatterjee, and A.R.S. Rana, *The effect of modifying an edge-plane pyrolytic graphite electrode with single-wall carbon nanotubes on its use for sensing diclofenac*. Carbon, 2010. **48**(14): p. 4136-4144, <https://doi.org/10.1016/j.carbon.2010.07.024>.
139. Nasiri, F., G.H. Rounaghi, N. Ashraf, and B. Deiminiat, *A new electrochemical sensing platform for quantitative determination of diclofenac based on gold nanoparticles decorated multiwalled carbon nanotubes/graphene oxide nanocomposite film*. International Journal of Environmental Analytical Chemistry, 2019. **101**(2): p. 153-166, 10.1080/03067319.2019.1661396.
140. Charithra, M.M. and J.G. Manjunatha, *Electroanalytical determination of acetaminophen using polymerized carbon nanocomposite based sensor*. Chemical Data Collections, 2021. **33**: p. 100718, 10.1016/j.cdc.2021.100718.
141. Shi, Y.-m., X. Zhang, L. Mei, D.-e. Han, K. Hu, L.-Q. Chao, X.-m. Li, and M.-s. Miao, *Sensitive acetaminophen electrochemical sensor with amplified signal strategy via non-covalent functionalization of soluble tetrahydroxyphthalocyanine and graphene*. Microchemical Journal, 2021. **160**: p. 105609, 10.1016/j.microc.2020.105609.
142. Yiğit, A., Y. Yardım, M. Çelebi, A. Levent, and Z. Şentürk, *Graphene/Nafion composite film modified glassy carbon electrode for simultaneous determination of paracetamol, aspirin and caffeine in pharmaceutical formulations*. Talanta, 2016. **158**: p. 21-29, <https://doi.org/10.1016/j.talanta.2016.05.046>.



143. Roushani, M. and F. Shahdost-fard, *Applicability of AuNPs@N-GQDs nanocomposite in the modeling of the amplified electrochemical Ibuprofen aptasensing assay by monitoring of riboflavin*. Bioelectrochemistry, 2019. **126**: p. 38-47, <https://doi.org/10.1016/j.bioelechem.2018.11.005>.
144. Manjunatha, R., D.H. Nagaraju, G.S. Suresh, J.S. Melo, S.F. D'Souza, and T.V. Venkatesha, *Electrochemical detection of acetaminophen on the functionalized MWCNTs modified electrode using layer-by-layer technique*. Electrochimica Acta, 2011. **56**(19): p. 6619-6627, 10.1016/j.electacta.2011.05.018.
145. Alothman, Z.A., N. Bukhari, S.M. Wabaidur, and S. Haider, *Simultaneous electrochemical determination of dopamine and acetaminophen using multiwall carbon nanotubes modified glassy carbon electrode*. Sensors and Actuators B: Chemical, 2010. **146**(1): p. 314-320, 10.1016/j.snb.2010.02.024.
146. Chen, X., J. Zhu, Q. Xi, and W. Yang, *A high performance electrochemical sensor for acetaminophen based on single-walled carbon nanotube-graphene nanosheet hybrid films*. Sensors and Actuators B: Chemical, 2012. **161**(1): p. 648-654, 10.1016/j.snb.2011.10.085.
147. Fan, Y., J.H. Liu, H.T. Lu, and Q. Zhang, *Electrochemical behavior and voltammetric determination of paracetamol on Nafion/TiO₂-graphene modified glassy carbon electrode*. Colloids Surf B Biointerfaces, 2011. **85**(2): p. 289-92, 10.1016/j.colsurfb.2011.02.041.
148. Liu, M., Q. Chen, C. Lai, Y. Zhang, J. Deng, H. Li, and S. Yao, *A double signal amplification platform for ultrasensitive and simultaneous detection of ascorbic acid, dopamine, uric acid and acetaminophen based on a nanocomposite of ferrocene thiolate*



- stabilized Fe(3)O(4)@Au nanoparticles with graphene sheet*. Biosens Bioelectron, 2013. **48**: p. 75-81, 10.1016/j.bios.2013.03.070.
149. Sakthivel, M., M. Sivakumar, S.-M. Chen, Y.-S. Hou, V. Veeramani, R. Madhu, and N. Miyamoto, *A Facile Synthesis of Cd(OH)2-rGO Nanocomposites for the Practical Electrochemical Detection of Acetaminophen*. Electroanalysis, 2017. **29**(1): p. 280-286, 10.1002/elan.201600351.
150. Yakowitz, H., *Methods of quantitative X-ray analysis used in electron probe microanalysis and scanning electron microscopy*, in *Practical Scanning Electron Microscopy*. 1975, Springer. p. 327-372.
151. Wang, L., T. Meng, J. Sun, S. Wu, M. Zhang, H. Wang, and Y. Zhang, *Development of Pd/Polyoxometalate/nitrogen-doping hollow carbon spheres tricomponent nanohybrids: A selective electrochemical sensor for acetaminophen*. Analytica Chimica Acta, 2019. **1047**: p. 28-35, <https://doi.org/10.1016/j.aca.2018.09.042>.
152. Shetti, N.P., S.J. Malode, D.S. Nayak, K.R. Reddy, C.V. Reddy, and K. Ravindranadh, *Silica gel-modified electrode as an electrochemical sensor for the detection of acetaminophen*. Microchemical Journal, 2019. **150**: p. 104206, 10.1016/j.microc.2019.104206.
153. Dou, N. and J. Qu, *Rapid synthesis of a hybrid of rGO/AuNPs/MWCNTs for sensitive sensing of 4-aminophenol and acetaminophen simultaneously*. Anal Bioanal Chem, 2021. **413**(3): p. 813-820, 10.1007/s00216-020-02856-6.
154. Farzad, H., M. Bahmaei, and M. Davallo, *Electrochemical Determination of Propranolol, Acetaminophen and Folic Acid in Urine, and Human Plasma Using Cu2O-*



- CuO/rGO/CPE*. Russian Journal of Electrochemistry, 2021. **57**(4): p. 357-374, 10.1134/s1023193521040054.
155. Tseng, T.W., T.W. Chen, S.M. Chen, T. Kokulnathan, F. Ahmed, P.M.Z. Hasan, A.L. Bilgrami, and S. Kumar, *Construction of strontium phosphate/graphitic-carbon nitride: A flexible and disposable strip for acetaminophen detection*. J Hazard Mater, 2021. **410**: p. 124542, 10.1016/j.jhazmat.2020.124542.
156. Guo, L., L. Hao, Y. Zhang, X. Yang, Q. Wang, Z. Wang, and C. Wang, *Metal-organic framework precursors derived Ni-doping porous carbon spheres for sensitive electrochemical detection of acetaminophen*. Talanta, 2021. **228**: p. 122228, 10.1016/j.talanta.2021.122228.
157. Tan, X., T. Mu, S. Wang, J. Li, J. Huang, H. Huang, Y. Pu, and G. Zhao, *Simultaneous determination of Acetaminophen and dopamine based on a water-soluble pillar[6]arene and ultrafine Pd nanoparticle-modified covalent organic framework nanocomposite*. Analyst, 2021. **146**(1): p. 262-269, 10.1039/d0an01717e.



Radius of γ -Spirallikeness of order α for some Special functions

Sercan Kazımoğlu · Kamaljeet Gangania*

Received: date / Accepted: date

Abstract In this paper, we establish the radius of γ -Spirallike of order α of certain well-known special functions. The main results of the paper are new and natural extensions of some known results.

Keywords γ -Spirallike functions · Radii of starlikeness and convexity · Wright and Mittag-Leffler functions · Legendre polynomials · Lommel and Struve functions · Ramanujan type entire functions

Mathematics Subject Classification (2010) 30C45 · 30C80 · 30C15

1 Introduction

Let \mathcal{A} be the class of analytic functions normalized by the condition $f(0) = 0 = f'(0) - 1$ in the unit disk $\mathbb{D} := \mathbb{D}_1$, where $\mathbb{D}_r := \{z \in \mathbb{C} : |z| < r\}$. We say that a function $f \in \mathcal{A}$ is γ -Spirallike of order α if and only if

$$\operatorname{Re} \left(e^{-i\gamma} \frac{zf'(z)}{f(z)} \right) > \alpha \cos \gamma,$$

where $\gamma \in (-\frac{\pi}{2}, \frac{\pi}{2})$ and $0 \leq \alpha < 1$. We denote the class of such functions by $\mathcal{S}_p^\gamma(\alpha)$. We also denote its convex analog, that is the class $\mathcal{CS}_p^\gamma(\alpha)$ of convex γ -spirallike functions of order α , which is defined below

$$\operatorname{Re} \left(e^{-i\gamma} \left(1 + \frac{zf''(z)}{f'(z)} \right) \right) > \alpha \cos \gamma.$$

Sercan Kazımoğlu

E-mail: srcnkzmglu@gmail.com

Department of Mathematics, Faculty of Science and Literature, Kafkas University, Campus, 36100, Kars-Turkey

Kamaljeet Gangania

E-mail: gangania.m1991@gmail.com

Department of Applied Mathematics, Delhi Technological University, Delhi-110042, India

* Corresponding author

The class $\mathcal{S}_p^\gamma(0)$ was introduced by Spacek [24]. Each function in $\mathcal{CS}_p^\gamma(\alpha)$ is univalent in \mathbb{D} , but they do not necessarily be starlike. Further, it is worth to mention that for general values of γ ($|\gamma| < \pi/2$), a function in $\mathcal{CS}_p^\gamma(0)$ need not be univalent in \mathbb{D} . For example: $f(z) = i(1-z)^i - i \in \mathcal{CS}_p^{\pi/4}(0)$, but not univalent. Indeed, $f \in \mathcal{CS}_p^\gamma(0)$ is univalent if $0 < \cos \gamma < 1/2$, see Robertson [23] and Pfaltzgraff [20]. Note that for $\gamma = 0$, the classes $\mathcal{S}_p^\gamma(\alpha)$ and $\mathcal{CS}_p^\gamma(\alpha)$ reduce to the classes of starlike and convex functions of order α , given by

$$\operatorname{Re} \left(\frac{zf'(z)}{f(z)} \right) > \alpha \quad \text{and} \quad \operatorname{Re} \left(1 + \frac{zf''(z)}{f'(z)} \right) > \alpha,$$

which we denote by $\mathcal{S}^*(\alpha)$ and $\mathcal{C}(\alpha)$, respectively.

In the recent past, connections between the special functions and their geometrical properties have been established in terms of radius problems [1, 2, 3, 5, 6, 7, 8, 9, 10, 25]. In this direction, behavior of the positive roots of a special function and the Laguerre-Pólya class play an evident role. A real entire function L maps real line into itself is said to be in the Laguerre-Pólya class \mathcal{LP} , if it can be expressed as follows:

$$L(x) = cx^m e^{-ax^2 + \beta x} \prod_{k \geq 1} \left(1 + \frac{x}{x_k} \right) e^{-\frac{x}{x_k}},$$

where $c, \beta, x_k \in \mathbb{R}$, $a \geq 0$, $m \in \mathbb{N} \cup \{0\}$ and $\sum x_k^{-2} < \infty$, see [2], [12, p. 703], [18] and the references therein. The class \mathcal{LP} consists of entire functions which can be approximated by polynomials with only real zeros, uniformly on the compact sets of the complex plane and it is closed under differentiation.

The $\mathcal{S}^*(\alpha)$ -radius, which is given below

$$\sup \{ r \in \mathbb{R}^+ : \operatorname{Re} \left(\frac{zg'(z)}{g(z)} \right) > \alpha, z \in \mathbb{D}_r \}$$

and similarly, $\mathcal{C}(\alpha)$ -radius has recently been obtained for some normalized forms of Bessel functions [1, 3, 6] (see Watson's treatise [26] for more on Bessel function), Struve functions [1, 2], Wright functions [7], Lommel functions [1, 2], Legendre polynomials of odd degree [9] and Ramanujan type entire functions [10]. For their generalization to Ma-Minda classes [19] of starlike and convex functions, we refer to see [13, 16].

With the best of our knowledge, $\mathcal{S}_p^\gamma(\alpha)$ -radius and $\mathcal{CS}_p^\gamma(\alpha)$ -radius for special functions are not handled till date. Therefore, in this paper, we now aim to derive the radius of γ -Spirallike of order α , which is given below

$$R_{sp}(g) = \sup \left\{ r \in \mathbb{R}^+ : \operatorname{Re} \left(e^{-i\gamma} \frac{zg'(z)}{g(z)} \right) > \alpha \cos \gamma, z \in \mathbb{D}_r \right\}$$

and also the radius of convex γ -Spirallike of order α , which is

$$R_{sp}^c(g) = \sup \left\{ r \in \mathbb{R}^+ : \operatorname{Re} \left(e^{-i\gamma} \left(1 + \frac{zg''(z)}{g'(z)} \right) \right) > \alpha \cos \gamma, z \in \mathbb{D}_r \right\}.$$

for the function g in \mathcal{A} to be a special function.

2 Wright functions

Let us consider the generalized Bessel function given by

$$\Phi(\kappa, \delta, z) = \sum_{n \geq 0} \frac{z^n}{n! \Gamma(n\kappa + \delta)},$$

where $\kappa > -1$ and $z, \delta \in \mathbb{C}$, named after E. M. Wright. The function Φ is entire for $\kappa > -1$. From [7, Lemma 1, p. 100], we have the Hadamard factorization

$$\Gamma(\delta) \Phi(\kappa, \delta, -z^2) = \prod_{n \geq 1} \left(1 - \frac{z^2}{\zeta_{\kappa, \delta, n}^2} \right), \quad (2.1)$$

where $\kappa, \delta > 0$ and $\zeta_{\kappa, \delta, n}$ is the n -th positive root of $\Phi(\kappa, \delta, -z^2)$ and satisfies the interlacing property:

$$\check{\zeta}_{\kappa, \delta, n} < \zeta_{\kappa, \delta, n} < \check{\zeta}_{\kappa, \delta, n+1} < \zeta_{\kappa, \delta, n+1}, \quad (n \geq 1) \quad (2.2)$$

where $\check{\zeta}_{\kappa, \delta, n}$ is the n -th positive root of the derivative of the function

$$\Psi_{\kappa, \delta}(z) = z^\delta \Phi(\kappa, \delta, -z^2).$$

Since $\Phi(\kappa, \delta, -z^2) \notin \mathcal{A}$, therefore we choose the normalized Wright functions:

$$\begin{cases} f_{\kappa, \delta}(z) = [z^\delta \Gamma(\delta) \Phi(\kappa, \delta, -z^2)]^{1/\delta} \\ g_{\kappa, \delta}(z) = z \Gamma(\delta) \Phi(\kappa, \delta, -z^2) \\ h_{\kappa, \delta}(z) = z \Gamma(\delta) \Phi(\kappa, \delta, -z). \end{cases} \quad (2.3)$$

For brevity, we write $W_{\kappa, \delta}(z) := \Phi(\kappa, \delta, -z^2)$.

Theorem 1 *Let $\kappa, \delta > 0$. The radius of γ -Spirallikeness for the functions $f_{\kappa, \delta}$, $g_{\kappa, \delta}$ and $h_{\kappa, \delta}$ are the smallest positive roots of the following equations:*

- (i) $r W'_{\kappa, \delta}(r) + \delta(1 - \alpha) \cos \gamma W_{\kappa, \delta}(r) = 0$
- (ii) $r W'_{\kappa, \delta}(r) + (1 - \alpha) \cos \gamma W_{\kappa, \delta}(r) = 0$
- (iii) $\sqrt{r} W'_{\kappa, \delta}(\sqrt{r}) + 2(1 - \alpha) \cos \gamma W_{\kappa, \delta}(\sqrt{r}) = 0$

in $|z| < (0, \zeta_{\kappa, \delta, 1})$, $(0, \zeta_{\kappa, \delta, 1})$ and $(0, \zeta_{\kappa, \delta, 1}^2)$, respectively.

Proof Using (2.1), we obtain the following by the logarithmic differentiation of (2.3):

$$\begin{cases} \frac{z f'_{\kappa, \delta}(z)}{f_{\kappa, \delta}(z)} = 1 + \frac{1}{\delta} \frac{z W'_{\kappa, \delta}(z)}{W_{\kappa, \delta}(z)} = 1 - \frac{1}{\delta} \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa, \delta, n}^2 - z^2} \\ \frac{z g'_{\kappa, \delta}(z)}{g_{\kappa, \delta}(z)} = 1 + \frac{z W'_{\kappa, \delta}(z)}{W_{\kappa, \delta}(z)} = 1 - \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa, \delta, n}^2 - z^2} \\ \frac{z h'_{\kappa, \delta}(z)}{h_{\kappa, \delta}(z)} = 1 + \frac{1}{2} \frac{\sqrt{z} W'_{\kappa, \delta}(\sqrt{z})}{W_{\kappa, \delta}(\sqrt{z})} = 1 - \sum_{n \geq 1} \frac{z}{\zeta_{\kappa, \delta, n}^2 - z}. \end{cases} \quad (2.4)$$

We need to show that the following inequalities for $\alpha \in [0, 1)$ and $\gamma \in (-\frac{\pi}{2}, \frac{\pi}{2})$,

$$\operatorname{Re} \left(e^{-i\gamma} \frac{z f'_{\kappa, \delta}(z)}{f_{\kappa, \delta}(z)} \right) > \alpha \cos \gamma, \quad \operatorname{Re} \left(e^{-i\gamma} \frac{z g'_{\kappa, \delta}(z)}{g_{\kappa, \delta}(z)} \right) > \alpha \cos \gamma \quad (2.5)$$

and

$$\operatorname{Re} \left(e^{-i\gamma} \frac{zh'_{\kappa,\delta}(z)}{h_{\kappa,\delta}(z)} \right) > \alpha \cos \gamma$$

are valid for $z \in \mathbb{D}_{r_{sp}(f_{\kappa,\delta})}$, $z \in \mathbb{D}_{r_{sp}(g_{\kappa,\delta})}$ and $z \in \mathbb{D}_{r_{sp}(h_{\kappa,\delta})}$ respectively, and each of the above inequalities does not hold in larger disks. It is known [11] that if $z \in \mathbb{C}$ and $\lambda \in \mathbb{R}$ are such that $|z| \leq r < \lambda$, then

$$\operatorname{Re} \left(\frac{z}{\lambda - z} \right) \leq \left| \frac{z}{\lambda - z} \right| \leq \frac{|z|}{\lambda - |z|}. \quad (2.6)$$

Then the inequality

$$\operatorname{Re} \left(\frac{z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \right) \leq \left| \frac{z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \right| \leq \frac{|z|^2}{\zeta_{\kappa,\delta,n}^2 - |z|^2}$$

holds for every $|z| < \zeta_{\kappa,\delta,1}$. Therefore, from (2.4) and (2.6), we have

$$\begin{aligned} \operatorname{Re} \left(e^{-i\gamma} \frac{zf'_{\kappa,\delta}(z)}{f_{\kappa,\delta}(z)} \right) &= \operatorname{Re} (e^{-i\gamma}) - \frac{1}{\delta} \operatorname{Re} \left(e^{-i\gamma} \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \right) \\ &\geq \cos \gamma - \frac{1}{\delta} \left| e^{-i\gamma} \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \right| \geq \cos \gamma - \frac{1}{\delta} \sum_{n \geq 1} \frac{2|z|^2}{\zeta_{\kappa,\delta,n}^2 - |z|^2} \\ &= \frac{|zf_{\kappa,\delta}(|z|)|}{f_{\kappa,\delta}(|z|)} + \cos \gamma - 1. \end{aligned} \quad (2.7)$$

Equality in the each of the above inequalities (2.9) holds when $z = r$. Thus, for $r \in (0, \zeta_{\kappa,\delta,1})$ it follows that

$$\inf_{z \in \mathbb{D}_r} \left\{ \operatorname{Re} \left(e^{-i\gamma} \frac{zf'_{\kappa,\delta}(z)}{f_{\kappa,\delta}(z)} - \alpha \cos \gamma \right) \right\} = \frac{|zf'_{\kappa,\delta}(|z|)|}{f_{\kappa,\delta}(|z|)} + (1 - \alpha) \cos \gamma - 1.$$

Now, the mapping $\Theta : (0, \zeta_{\kappa,\delta,1}) \longrightarrow \mathbb{R}$ defined by

$$\Theta(r) = \frac{rf'_{\kappa,\delta}(r)}{f_{\kappa,\delta}(r)} + (1 - \alpha) \cos \gamma - 1 = (1 - \alpha) \cos \gamma - \frac{1}{\delta} \sum_{n \geq 1} \left(\frac{2r^2}{\zeta_{\kappa,\delta,n}^2 - r^2} \right).$$

is strictly decreasing since

$$\Theta'(r) = -\frac{1}{\delta} \sum_{n \geq 1} \left(\frac{4r\zeta_{\kappa,\delta,n}}{(\zeta_{\kappa,\delta,n}^2 - r^2)^2} \right) < 0$$

for all $\delta > 0$. On the other hand, since

$$\lim_{r \searrow 0} \Theta(r) = (1 - \alpha) \cos \gamma > 0 \quad \text{and} \quad \lim_{r \nearrow \zeta_{\kappa,\delta,1}} \Theta(r) = -\infty,$$

in view of the minimum principle for harmonic functions imply that the corresponding inequality for $f_{\kappa,\delta}$ in (2) for $\delta > 0$ holds if and only if $z \in \mathbb{D}_{r_{sp}(f_{\kappa,\delta})}$, where $r_{sp}(f_{\kappa,\delta})$ is the smallest positive root of equation

$$\frac{r f'_{\kappa,\delta}(r)}{f_{\kappa,\delta}(r)} = 1 - (1 - \alpha) \cos \gamma$$

which is equivalent to

$$\frac{1}{\delta} \frac{z W'_{\kappa,\delta}(z)}{W_{\kappa,\delta}(z)} = -(1 - \alpha) \cos \gamma,$$

situated in $(0, \zeta_{\kappa,\delta,1})$. Reasoning along the same lines, proofs of the other parts follows. \square

Remark 1 Taking $\gamma = 0$ in Theorem 1 yields [7, Theorem 1].

In the following, we deal with convex analogue of the class of γ -spirallike functions of order α .

Theorem 2 Let $\kappa, \delta > 0$ and the functions $f_{\kappa,\delta}$, $g_{\kappa,\delta}$ and $h_{\kappa,\delta}$ as given in (2.3). Then

(i) the radius $R_{sp}^c(f_{\kappa,\delta})$ is the smallest positive root of the equation

$$\frac{r \Psi''_{\kappa,\delta}(r)}{\Psi'_{\kappa,\delta}(r)} + \left(\frac{1}{\delta} - 1 \right) \frac{r \Psi'_{\kappa,\delta}(r)}{\Psi_{\kappa,\delta}(r)} + (1 - \alpha) \cos \gamma = 0.$$

(ii) the radius $R_{sp}^c(g_{\kappa,\delta})$ is the smallest positive root of the equation

$$r g''_{\kappa,\delta}(r) + (1 - \alpha) \cos \gamma g'_{\kappa,\delta}(r) = 0.$$

(iii) the radius $R_{sp}^c(h_{\kappa,\delta})$ is the smallest positive root of the equation

$$r h''_{\kappa,\delta}(r) + (1 - \alpha) \cos \gamma h'_{\kappa,\delta}(r) = 0.$$

Proof We first prove the part (i). From (2.1), (2.3) and using the Hadamard representation $\Gamma(\delta) \Psi'_{\kappa,\delta}(z) = \delta z^{\delta-1} \prod_{n \geq 1} \left(1 - \frac{z^2}{\zeta_{\kappa,\delta,n}^2} \right)$, (see [7, Eq. 7]), we have

$$\begin{aligned} 1 + \frac{z f''_{\kappa,\delta}(z)}{f'_{\kappa,\delta}(z)} &= 1 + \frac{z \Psi''_{\kappa,\delta}(z)}{\Psi'_{\kappa,\delta}(z)} + \left(\frac{1}{\delta} - 1 \right) \frac{z \Psi'_{\kappa,\delta}(z)}{\Psi_{\kappa,\delta}(z)} \\ &= 1 - \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} - \left(\frac{1}{\delta} - 1 \right) \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \end{aligned}$$

and for $\delta > 1$, using the following inequality of [11]:

$$\left| \frac{z}{y-z} - \lambda \frac{z}{x-z} \right| \leq \frac{|z|}{y-|z|} - \lambda \frac{|z|}{x-|z|}, \quad (x > y > r \geq |z|) \quad (2.8)$$

with $\lambda = 1 - 1/\delta$, we get

$$\left| \frac{zf''_{\kappa,\delta}(z)}{f'_{\kappa,\delta}(z)} \right| \leq -\frac{rf''_{\kappa,\delta}(r)}{f'_{\kappa,\delta}(r)} = -\frac{r\Psi''_{\kappa,\delta}(r)}{\Psi'_{\kappa,\delta}(r)} - \left(\frac{1}{\delta} - 1\right) \frac{r\Psi'_{\kappa,\delta}(r)}{\Psi_{\kappa,\delta}(r)}.$$

Also, using the inequality $\|x\| - \|y\| \leq \|x - y\|$ and the relation in (2.2), we see that for $\delta > 0$

$$\left| \frac{zf''_{\kappa,\delta}(z)}{f'_{\kappa,\delta}(z)} \right| \leq -\frac{rf''_{\kappa,\delta}(r)}{f'_{\kappa,\delta}(r)},$$

holds in $|z| = r < \check{\zeta}_{\kappa,\delta,1}$. Therefore, we have

$$\begin{aligned} & \operatorname{Re} \left(e^{-i\gamma} \left(1 + \frac{zf''_{\kappa,\delta}(z)}{f'_{\kappa,\delta}(z)} \right) \right) \\ &= \operatorname{Re}(e^{-i\gamma}) - \operatorname{Re} \left(e^{-i\gamma} \left(\sum_{n \geq 1} \frac{2z^2}{\check{\zeta}_{\kappa,\delta,n}^2 - z^2} + \left(\frac{1}{\delta} - 1\right) \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \right) \right) \\ &\geq \cos \gamma - \left| \sum_{n \geq 1} \frac{2z^2}{\check{\zeta}_{\kappa,\delta,n}^2 - z^2} + \left(\frac{1}{\delta} - 1\right) \sum_{n \geq 1} \frac{2z^2}{\zeta_{\kappa,\delta,n}^2 - z^2} \right| \\ &\geq \cos \gamma + \frac{rf''_{\kappa,\delta}(r)}{f'_{\kappa,\delta}(r)} \end{aligned} \quad (2.9)$$

hold for $\delta > 1$. Observe that these inequalities also hold for $\delta > 0$. Equality in the each of the above inequalities (2.9) holds when $z = r$. Thus, for $r \in (0, \check{\zeta}_{\kappa,\delta,1})$ it follows that

$$\inf_{z \in \mathbb{D}_r} \left\{ \operatorname{Re} \left(e^{-i\gamma} \left(1 + \frac{zf''_{\kappa,\delta}(z)}{f'_{\kappa,\delta}(z)} \right) - \alpha \cos \gamma \right) \right\} = (1 - \alpha) \cos \gamma + \frac{|z| f''_{\kappa,\delta}(|z|)}{f'_{\kappa,\delta}(|z|)}.$$

Now, the proof of part (i) follows on similar lines as of Theorem 1.

For the other parts, note that the functions $g_{\kappa,\delta}$ and $h_{\kappa,\delta}$ belong to the Laguerre-Pólya class \mathcal{LP} , which is closed under differentiation, their derivatives $g'_{\kappa,\delta}$ and $h'_{\kappa,\delta}$ also belong to \mathcal{LP} and the zeros are real. Thus assuming $\tau_{\kappa,\delta,n}$ and $\eta_{\kappa,\delta,n}$ are the positive zeros of $g'_{\kappa,\delta}$ and $h'_{\kappa,\delta}$, respectively, we have the following representations:

$$g'_{\kappa,\delta}(z) = \prod_{n \geq 1} \left(1 - \frac{z^2}{\tau_{\kappa,\delta,n}^2} \right) \quad \text{and} \quad h'_{\kappa,\delta}(z) = \prod_{n \geq 1} \left(1 - \frac{z}{\eta_{\kappa,\delta,n}} \right),$$

which yield

$$1 + \frac{zg''_{\kappa,\delta}(z)}{g'_{\kappa,\delta}(z)} = 1 - \sum_{n \geq 1} \frac{2z^2}{\tau_{\kappa,\delta,n}^2 - z^2} \quad \text{and} \quad 1 + \frac{zh''_{\kappa,\delta}(z)}{h'_{\kappa,\delta}(z)} = 1 - \sum_{n \geq 1} \frac{z}{\eta_{\kappa,\delta,n} - z}.$$

Further, reasoning along the same lines as in Theorem 1, the result follows at once. \square

Remark 2 Taking $\gamma = 0$ in Theorem 2 yields [7, Theorem 5].

3 Mittag-Leffler functions

In 1971, Prabhakar [21] introduced the following function

$$M(\mu, \nu, a, z) := \sum_{n \geq 0} \frac{(a)_n z^n}{n! \Gamma(\mu n + \nu)},$$

where $(a)_n = \Gamma(a+n)/\Gamma(a)$ denotes the Pochhammer symbol and $\mu, \nu, a > 0$. The functions $M(\mu, \nu, 1, z)$ and $M(\mu, 1, 1, z)$ were introduced and studied by Wiman and Mittag-Leffler, respectively. Now let us consider the set $W_b = A(W_c) \cup B(W_c)$, where

$$W_c := \left\{ \left(\frac{1}{\mu}, \nu \right) : 1 < \mu < 2, \nu \in [\mu - 1, 1] \cup [\mu, 2] \right\}$$

and denote by W_i , the smallest set containing W_b and invariant under the transformations A , B and C mapping the set $\{(\frac{1}{\mu}, \nu) : \mu > 1, \nu > 0\}$ into itself and are defined as:

$$\begin{aligned} A : \left(\frac{1}{\mu}, \nu \right) &\rightarrow \left(\frac{1}{2\mu}, \nu \right), & B : \left(\frac{1}{\mu}, \nu \right) &\rightarrow \left(\frac{1}{2\mu}, \mu + \nu \right), \\ C : \left(\frac{1}{\mu}, \nu \right) &\rightarrow \begin{cases} \left(\frac{1}{\mu}, \nu - 1 \right), & \text{if } \nu > 1; \\ \left(\frac{1}{\mu}, \nu \right), & \text{if } 0 < \nu \leq 1. \end{cases} \end{aligned}$$

Kumar and Pathan [17] proved that if $(\frac{1}{\mu}, \nu) \in W_i$ and $a > 0$, then all zeros of $M(\mu, \nu, a, z)$ are real and negative. From [5, Lemma 1, p. 121], we see that if $(\frac{1}{\mu}, \nu) \in W_i$ and $a > 0$, then the function $M(\mu, \nu, a, -z^2)$ has infinitely many zeros, which are all real and have the following representation:

$$\Gamma(\nu) M(\mu, \nu, a, -z^2) = \prod_{n \geq 1} \left(1 - \frac{z^2}{\lambda_{\mu, \nu, a, n}^2} \right),$$

where $\lambda_{\mu, \nu, a, n}$ is the n -th positive zero of $M(\mu, \nu, a, -z^2)$ and satisfy the interlacing relation

$$\xi_{\mu, \nu, a, n} < \lambda_{\mu, \nu, a, n} < \xi_{\mu, \nu, a, n+1} < \lambda_{\mu, \nu, a, n+1} \quad (n \geq 1),$$

where $\xi_{\mu, \nu, a, n}$ is the n -th positive zero of the derivative of $z^\nu M(\mu, \nu, a, -z^2)$. Since $M(\mu, \nu, a, -z^2) \notin \mathcal{A}$, therefore we consider the following normalized forms (belong to the Laguerre-Pólya class):

$$\begin{cases} f_{\mu, \nu, a}(z) = [z^\nu \Gamma(\nu) M(\mu, \nu, a, -z^2)]^{1/\nu}, \\ g_{\mu, \nu, a}(z) = z \Gamma(\nu) M(\mu, \nu, a, -z^2) \\ h_{\mu, \nu, a}(z) = z \Gamma(\nu) M(\mu, \nu, a, -z). \end{cases} \quad (3.1)$$

For brevity, write $L_{\mu, \nu, a}(z) := M(\mu, \nu, a, -z^2)$. Now proceeding similarly as in Section 2, we obtain the following results:

Theorem 3 Let $(\frac{1}{\mu}, \nu) \in W_i$, $a > 0$. Then the radius of γ -Spirallikeness of order α for the functions $f_{\mu,\nu,a}$, $g_{\mu,\nu,a}$ and $h_{\mu,\nu,a}$ given by (3.1) are the smallest positive roots of the following equations:

- (i) $rL'_{\mu,\nu,a}(r) + \delta(1 - \alpha) \cos \gamma L_{\mu,\nu,a}(r) = 0$
 - (ii) $rL'_{\mu,\nu,a}(r) + (1 - \alpha) \cos \gamma L_{\mu,\nu,a}(r) = 0$
 - (iii) $\sqrt{r}L'_{\mu,\nu,a}(\sqrt{r}) + 2(1 - \alpha) \cos \gamma L'_{\mu,\nu,a}(\sqrt{r}) = 0$
- in $|z| < (0, \lambda_{\mu,\nu,a,1})$, $(0, \lambda_{\mu,\nu,a,1})$ and $(0, \lambda_{\mu,\nu,a,1}^2)$, respectively.

Proof Using (3.1), we obtain after the logarithmic differentiation:

$$\begin{cases} \frac{zf'_{\mu,\nu,a}(z)}{f_{\mu,\nu,a}(z)} = 1 - \frac{1}{\nu} \sum_{n \geq 1} \frac{2z^2}{\lambda_{\mu,\nu,a,n}^2 - z^2} \\ \frac{zg'_{\mu,\nu,a}(z)}{g_{\mu,\nu,a}(z)} = 1 - \sum_{n \geq 1} \frac{2z^2}{\lambda_{\mu,\nu,a,n}^2 - z^2} \\ \frac{zh'_{\mu,\nu,a}(z)}{h_{\mu,\nu,a}(z)} = 1 - \sum_{n \geq 1} \frac{z}{\lambda_{\mu,\nu,a,n}^2 - z} \end{cases} \quad (3.2)$$

We need to show that the following inequalities for $\alpha \in [0, 1)$ and $\gamma \in (-\frac{\pi}{2}, \frac{\pi}{2})$,

$$\operatorname{Re} \left(e^{-i\gamma} \frac{zf'_{\mu,\nu,a}(z)}{f_{\mu,\nu,a}(z)} \right) > \alpha \cos \gamma, \quad \operatorname{Re} \left(e^{-i\gamma} \frac{zg'_{\mu,\nu,a}(z)}{g_{\mu,\nu,a}(z)} \right) > \alpha \cos \gamma \quad (3.3)$$

and

$$\operatorname{Re} \left(e^{-i\gamma} \frac{zh'_{\mu,\nu,a}(z)}{h_{\mu,\nu,a}(z)} \right) > \alpha \cos \gamma$$

are valid for $z \in \mathbb{D}_{r_{sp}(f_{\mu,\nu,a})}$, $z \in \mathbb{D}_{r_{sp}(g_{\mu,\nu,a})}$ and $z \in \mathbb{D}_{r_{sp}(h_{\mu,\nu,a})}$ respectively, and each of the above inequalities does not hold in larger disks. Since using (2.6)

$$\operatorname{Re} \left(\frac{z^2}{\lambda_{\mu,\nu,a,n}^2 - z^2} \right) \leq \left| \frac{z^2}{\lambda_{\mu,\nu,a,n}^2 - z^2} \right| \leq \frac{|z|^2}{\lambda_{\mu,\nu,a,n}^2 - |z|^2} \quad (3.4)$$

holds for every $|z| < \lambda_{\mu,\nu,a,1}$. Therefore, from (3.2) and (3.4), we have

$$\begin{aligned} \operatorname{Re} \left(e^{-i\gamma} \frac{zf'_{\mu,\nu,a}(z)}{f_{\mu,\nu,a}(z)} \right) &= \operatorname{Re} (e^{-i\gamma}) - \frac{1}{\nu} \operatorname{Re} \left(e^{-i\gamma} \sum_{n \geq 1} \frac{2z^2}{\lambda_{\mu,\nu,a,n}^2 - z^2} \right) \\ &\geq \cos \gamma - \frac{1}{\nu} \left| e^{-i\gamma} \sum_{n \geq 1} \frac{2z^2}{\lambda_{\mu,\nu,a,n}^2 - z^2} \right| \\ &\geq \cos \gamma - \frac{1}{\nu} \sum_{n \geq 1} \frac{2|z|^2}{\lambda_{\mu,\nu,a,n}^2 - |z|^2} \\ &= \frac{|z|f'_{\mu,\nu,a}(|z|)}{f_{\mu,\nu,a}(|z|)} + \cos \gamma - 1. \end{aligned} \quad (3.5)$$

Equality in the each of the above inequalities (3.5) holds when $z = r$. Thus, for $r \in (0, \lambda_{\mu,\nu,a,1})$ it follows that

$$\inf_{z \in \mathbb{D}_r} \left\{ \operatorname{Re} \left(e^{-i\gamma} \frac{z f'_{\mu,\nu,a}(z)}{f_{\mu,\nu,a}(z)} - \alpha \cos \gamma \right) \right\} = \frac{|z| f'_{\mu,\nu,a}(|z|)}{f_{\mu,\nu,a}(|z|)} + (1 - \alpha) \cos \gamma - 1.$$

Now, the mapping $\Theta : (0, \lambda_{\mu,\nu,a,1}) \rightarrow \mathbb{R}$ defined by

$$\Theta(r) = \frac{r f'_{\mu,\nu,a}(r)}{f_{\mu,\nu,a}(r)} + (1 - \alpha) \cos \gamma - 1 = (1 - \alpha) \cos \gamma - \frac{1}{\nu} \sum_{n \geq 1} \left(\frac{2r^2}{\lambda_{\mu,\nu,a,n}^2 - r^2} \right).$$

is strictly decreasing since

$$\Theta'(r) = -\frac{1}{\nu} \sum_{n \geq 1} \left(\frac{4r \lambda_{\mu,\nu,a,n}}{(\lambda_{\mu,\nu,a,n}^2 - r^2)^2} \right) < 0$$

for all $\nu > 0$. On the other hand, since

$$\lim_{r \searrow 0} \Theta(r) = (1 - \alpha) \cos \gamma > 0 \quad \text{and} \quad \lim_{r \nearrow \lambda_{\mu,\nu,a,1}} \Theta(r) = -\infty,$$

in view of the minimum principle for harmonic functions imply that the corresponding inequality for $f_{\mu,\nu,a}$ in (3.3) for $\nu > 0$ holds if and only if $z \in \mathbb{D}_{r_{sp}(f_{\mu,\nu,a})}$, where $r_{sp}(f_{\mu,\nu,a})$ is the smallest positive root of equation

$$\frac{r f'_{\mu,\nu,a}(r)}{f_{\mu,\nu,a}(r)} = 1 - (1 - \alpha) \cos \gamma$$

which is equivalent to

$$\frac{1}{\nu} \frac{z L'_{\mu,\nu,a}(z)}{L_{\mu,\nu,a}(z)} = -(1 - \alpha) \cos \gamma,$$

situated in $(0, \lambda_{\mu,\nu,1})$. Reasoning along the same lines, proofs of the other parts follows. \square

Remark 3 Taking $\gamma = 0$ in Theorem 3 yields [5, Theorem 1].

In the following, we derive the result for the convex analog proceedings on similar lines as Theorem 2.

Theorem 4 Let $(\frac{1}{\mu}, \nu) \in W_i$, $a > 0$. Let the functions $f_{\mu,\nu,a}$, $g_{\mu,\nu,a}$ and $h_{\mu,\nu,a}$ be given by (3.1). Then

(i) the radius $R_{sp}^c(f_{\mu,\nu,a})$ is the smallest positive root of the equation

$$r f''_{\mu,\nu,a}(r) + (1 - \alpha) \cos \gamma f'_{\mu,\nu,a}(r) = 0.$$

(ii) the radius $R_{sp}^c(g_{\mu,\nu,a})$ is the smallest positive root of the equation

$$r g''_{\mu,\nu,a}(r) + (1 - \alpha) \cos \gamma g'_{\mu,\nu,a}(r) = 0.$$

(iii) the radius $R_{sp}^c(h_{\mu,\nu,a})$ is the smallest positive root of the equation

$$r h''_{\mu,\nu,a}(r) - (1 - \alpha) \cos \gamma h'_{\mu,\nu,a}(r) = 0.$$

Remark 4 Taking $\gamma = 0$ in Theorem 4 yields [5, Theorem 3].

4 Legendre polynomials

The Legendre polynomials P_n are the solutions of the Legendre differential equation

$$((1 - z^2)P'_n(z))' + n(n+1)P_n(z) = 0,$$

where $n \in \mathbb{Z}^+$ and using Rodrigues formula, P_n can be represented in the form:

$$P_n(z) = \frac{1}{2^n n!} \frac{d^n (z^2 - 1)^n}{dz^n}$$

and it also satisfies the geometric condition $P_n(-z) = (-1)^n P_n(z)$. Moreover, the odd degree Legendre polynomials $P_{2n-1}(z)$ have only real roots which satisfy

$$0 = z_0 < z_1 < \cdots < z_{n-1} \quad \text{or} \quad -z_1 > \cdots > -z_{n-1}. \quad (4.1)$$

Thus the normalized form is as follows:

$$\mathcal{P}_{2n-1}(z) := \frac{P_{2n-1}(z)}{P'_{2n-1}(0)} = z + \sum_{k=2}^{2n-1} a_k z^k = a_{2n-1} z \prod_{k=1}^{n-1} (z^2 - z_k^2). \quad (4.2)$$

Theorem 5 *Let \mathcal{P}_{2n-1} be given by (4.2). Then*

(i) *the radius $R_{sp}^c(\mathcal{P}_{2n-1})$ is the smallest positive root of the equation*

$$r\mathcal{P}_{2n-1}''(r) + (1 - \alpha)\mathcal{P}_{2n-1}'(r) = 0.$$

(ii) *the radius of γ -Spirallikeness of order α for the normalized Legendre polynomial of odd degree is given by the smallest positive root of the equation*

$$r\mathcal{P}_{2n-1}'(r) + (1 - \alpha)\mathcal{P}_{2n-1}(r) = 0.$$

Proof We prove first part and second part follows on same lines. From (4.2), upon the logarithmic differentiation, we have

$$1 + \frac{z\mathcal{P}_{2n-1}''(z)}{\mathcal{P}_{2n-1}'(z)} = \frac{z\mathcal{P}_{2n-1}'(z)}{\mathcal{P}_{2n-1}(z)} - \frac{\sum_{k=1}^{n-1} \frac{4z_k^2 z^2}{(z_k^2 - z^2)^2}}{\frac{z\mathcal{P}_{2n-1}'(z)}{\mathcal{P}_{2n-1}(z)}},$$

where

$$\frac{z\mathcal{P}_{2n-1}'(z)}{\mathcal{P}_{2n-1}(z)} = 1 - \sum_{k=1}^{n-1} \frac{2z^2}{z_k^2 - z^2}.$$

Further, after using the inequality $||x| - |y|| \leq |x - y|$ and (4.1) for $|z| = r < z_1$, we see that

$$\begin{aligned}
& \operatorname{Re} \left(e^{-i\gamma} \left(1 + \frac{z \mathcal{P}_{2n-1}''(z)}{\mathcal{P}_{2n-1}'(z)} \right) \right) \\
&= \operatorname{Re} (e^{-i\gamma}) - \operatorname{Re} \left(e^{-i\gamma} \sum_{k=1}^{n-1} \frac{2z^2}{z_k^2 - z^2} \right) - \operatorname{Re} \left(e^{-i\gamma} \frac{\sum_{k=1}^{n-1} \frac{4z_k^2 z^2}{(z_k^2 - z^2)^2}}{1 - \sum_{k=1}^{n-1} \frac{2z^2}{z_k^2 - z^2}} \right) \\
&\geq \cos \gamma - \left| e^{-i\gamma} \sum_{k=1}^{n-1} \frac{2z^2}{z_k^2 - z^2} \right| - \left| e^{-i\gamma} \frac{\sum_{k=1}^{n-1} \frac{4z_k^2 z^2}{(z_k^2 - z^2)^2}}{1 - \sum_{k=1}^{n-1} \frac{2z^2}{z_k^2 - z^2}} \right| \\
&= \cos \gamma - \sum_{k=1}^{n-1} \frac{2r^2}{z_k^2 - r^2} - \frac{\sum_{k=1}^{n-1} \frac{4z_k^2 r^2}{(z_k^2 - r^2)^2}}{1 - \sum_{k=1}^{n-1} \frac{2r^2}{z_k^2 - r^2}} = \cos \gamma + \frac{r \mathcal{P}_{2n-1}''(r)}{\mathcal{P}_{2n-1}'(r)}.
\end{aligned}$$

Further, with similar reasoning as Theorem 3, result follows. \square

Remark 5 Taking $\gamma = 0$ in Theorem 5 yields [9, Theorem 2.2] and [9, Theorem 2.1].

5 Lommel functions

The Lommel function $\mathcal{L}_{u,v}$ of first kind is a particular solution of the second-order inhomogeneous Bessel differential equation

$$z^2 w''(z) + z w'(z) + (z^2 - v^2) w(z) = z^{u+1},$$

where $u \pm v \notin \mathbb{Z}^-$ and is given by

$$\mathcal{L}_{u,v} = \frac{z^{u+1}}{(u-v+1)(u+v+1)} {}_1F_2 \left(1; \frac{u-v+3}{2}, \frac{u+v+3}{2}; -\frac{z^2}{4} \right),$$

where $\frac{1}{2}(-u \pm v - 3) \notin \mathbb{N}$ and ${}_1F_2$ is a hypergeometric function. Since it is not normalized, therefore we consider the following three normalized functions involving $\mathcal{L}_{u,v}$:

$$\begin{cases} f_{u,v}(z) = ((u-v+1)(u+v+1)\mathcal{L}_{u,v}(z))^{\frac{1}{u+1}}, \\ g_{u,v}(z) = (u-v+1)(u+v+1)z^{-u}\mathcal{L}_{u,v}(z), \\ h_{u,v}(z) = (u-v+1)(u+v+1)z^{(1-u)/2}\mathcal{L}_{u,v}(\sqrt{z}). \end{cases} \quad (5.1)$$

Authors in [1, 2] and [8] proved the radius of starlikeness and convexity for the following normalized functions expressed in terms of $\mathcal{L}_{u-\frac{1}{2}, \frac{1}{2}}$:

$$f_{u-\frac{1}{2}, \frac{1}{2}}(z), \quad g_{u-\frac{1}{2}, \frac{1}{2}}(z) \quad \text{and} \quad h_{u-\frac{1}{2}, \frac{1}{2}}(z), \quad (5.2)$$

where $0 \neq u \in (-1, 1)$.

For brevity, we write these as f_u, g_u and h_u , respectively and $\mathcal{L}_{u-\frac{1}{2}, \frac{1}{2}} = \mathcal{L}_u$.

Theorem 6 Let $u \in (-1, 1)$, $u \neq 0$. Let the functions f_u, g_u and h_u be given by (5.2). Then

(i) the radius $R_{sp}(f_u)$ is the smallest positive root of the equation

$$rf_u''(r) + (1 - \alpha) \cos \gamma f_u'(r) = 0, \quad \text{if } u \neq -1/2.$$

(ii) the radius $R_{sp}(g_u)$ is the smallest positive root of the equation

$$rg_u''(r) + (1 - \alpha) \cos \gamma g_u'(r) = 0.$$

(iii) the radius $R_{sp}^c(h_u)$ is the smallest positive root of the equation

$$rh_u''(r) + (1 - \alpha) \cos \gamma h_u'(r) = 0.$$

Proof We begin with the first part. From (5.1), we have

$$1 + \frac{zf_u''(z)}{f_u'(z)} = 1 + \frac{z\mathcal{L}_u''(z)}{\mathcal{L}_u'(z)} + \left(\frac{1}{u + \frac{1}{2}} - 1 \right) \frac{z\mathcal{L}_u'(z)}{\mathcal{L}_u(z)}. \quad (5.3)$$

Also using the result [8, Lemma 1], we have

$$\mathcal{L}_u(z) = \frac{z^{u+\frac{1}{2}}}{u(u+1)} \Phi_0(z) = \frac{z^{u+\frac{1}{2}}}{u(u+1)} \prod_{n \geq 1} \left(1 - \frac{z^2}{\tau_{u,n}^2} \right),$$

where $\Phi_k(z) := {}_1F_2 \left(1; \frac{u-k+2}{2}, \frac{u-k+3}{2}; -\frac{z^2}{4} \right)$ with conditions as mentioned in [8, Lemma 1], and from the proof of [8, Theorem 3], we see that the entire function $\frac{u(u+1)}{u+\frac{1}{2}} z^{-u+\frac{1}{2}} \mathcal{L}_u'(z)$ is of order $1/2$ and therefore, has the following Hadamard factorization:

$$\mathcal{L}_u'(z) = \frac{u + \frac{1}{2}}{u(u+1)} z^{u-\frac{1}{2}} \prod_{n \geq 1} \left(1 - \frac{z^2}{\check{\tau}_{u,n}^2} \right),$$

where $\tau_{u,n}$ and $\check{\tau}_{u,n}$ are the n -th positive zeros of \mathcal{L}_u and \mathcal{L}_u' , respectively and interlace for $0 \neq u \in (-1, 1)$ (see [8, Theorem 1]). Now we can rewrite (5.3) as follows:

$$1 + \frac{zf_u''(z)}{f_u'(z)} = 1 - \left(\frac{1}{u + \frac{1}{2}} - 1 \right) \sum_{n \geq 1} \frac{2z^2}{\tau_{u,n}^2 - z^2} - \sum_{n \geq 1} \frac{2z^2}{\check{\tau}_{u,n}^2 - z^2}.$$

Let us now consider the case $u \in (0, 1/2]$. Then using the inequality $||x| - |y|| \leq |x - y|$ for $|z| = r < \check{\tau}_{u,1} < \tau_{u,1}$ we get

$$\left| \frac{zf_u''(z)}{f_u'(z)} \right| \leq \left(\frac{1}{u + \frac{1}{2}} - 1 \right) \sum_{n \geq 1} \frac{2r^2}{\tau_{u,n}^2 - r^2} + \sum_{n \geq 1} \frac{2r^2}{\check{\tau}_{u,n}^2 - r^2} = -\frac{rf_u''(r)}{f_u'(r)} \quad (5.4)$$

and for the case $u \in (1/2, 1)$, using the inequality (2.8) with $\lambda = 1 - 1/(u + 1/2)$, we also get

$$\left| \frac{zf_u''(z)}{f_u'(z)} \right| \leq -\frac{rf_u''(r)}{f_u'(r)}, \quad (5.5)$$

which is same as (5.4). When $u \in (-1, 0)$, then we proceed similarly substituting u by $u-1$, Φ_0 by Φ_1 , where Φ_1 belongs to the Laguerre-Pólya class \mathcal{LP} and the n -th positive zeros $\xi_{u,n}$ and $\xi_{u,n}'$ of Φ_1 and its derivative Φ_1' , respectively interlace. Finally, replacing u by $u+1$, we obtain the required inequality.

For $0 \neq u \in (-1, 1)$, the Hadamard factorization for the entire functions g'_u and h'_u of order $1/2$ [8, Theorem 3] is given by

$$g'_u(z) = \prod_{n \geq 1} \left(1 - \frac{z^2}{\gamma_{u,n}^2}\right) \quad \text{and} \quad h'_u(z) = \prod_{n \geq 1} \left(1 - \frac{z}{\delta_{u,n}^2}\right), \quad (5.6)$$

where $\gamma_{u,n}$ and $\delta_{u,n}$ are n -th positive zeros of g'_u and h'_u , respectively and $\gamma_{u,1}, \delta_{u,1} < \tau_{u,1}$. Now from (5.1) and (5.6), we have

$$\begin{cases} 1 + \frac{zg''_u(z)}{g'_u(z)} = \frac{1}{2} - u + z \frac{(\frac{3}{2} - u)\mathcal{L}'_u(z) + z\mathcal{L}''_u(z)}{(\frac{1}{2} - u)\mathcal{L}_u(z) + z\mathcal{L}'_u(z)} = 1 - \sum_{n \geq 1} \frac{2z^2}{\gamma_{u,n}^2 - z^2} \\ 1 + \frac{zh''_u(z)}{h'_u(z)} = \frac{1}{2} \left(\frac{3}{2} - u + \sqrt{z} \frac{(\frac{5}{2} - u)\mathcal{L}'_u(\sqrt{z}) + \sqrt{z}\mathcal{L}''_u(\sqrt{z})}{(\frac{3}{2} - u)\mathcal{L}_u(\sqrt{z}) + \sqrt{z}\mathcal{L}'_u(\sqrt{z})} \right) = 1 - \sum_{n \geq 1} \frac{z}{\delta_{u,n}^2 - z}. \end{cases} \quad (5.7)$$

Using the inequality $||x| - |y|| \leq |x - y|$ in (5.7) for $|z| = r < \gamma_{u,1}$ and $|z| = r < \delta_{u,1}$, we get

$$\begin{cases} \left| \frac{zg''_u(z)}{g'_u(z)} \right| \leq \sum_{n \geq 1} \frac{2r^2}{\gamma_{u,n}^2 - r^2} = -\frac{rg''_u(r)}{g'_u(r)} \\ \left| \frac{zh''_u(z)}{h'_u(z)} \right| \leq \sum_{n \geq 1} \frac{r}{\delta_{u,n}^2 - r} = -\frac{rh''_u(r)}{h'_u(r)}. \end{cases} \quad (5.8)$$

Further, proceeding with the similar method as in Theorem 1, result follows. \square

With similar reasoning as Theorem 1, the proof of the following holds.

Theorem 7 Let $u \in (-1, 1)$, $u \neq 0$. Then the radius of γ -Spirallikeness of order α for the functions Let the functions f_u, g_u and h_u given by (5.2) are the smallest positive roots of the following equations:

- (i) $rf'_u(r) + ((1 - \alpha) \cos \gamma - 1)f_u(r) = 0$
- (ii) $rg'_u(r) + ((1 - \alpha) \cos \gamma - 1)g_u(r) = 0$
- (iii) $rh'_u(r) + ((1 - \alpha) \cos \gamma - 1)h_u(r) = 0$

in $(0, \tau_{u,1})$, $(0, \tau_{u,1})$ and $(0, \tau_{u,1}^2)$, respectively.

Remark 6 Taking $\gamma = 0$, Theorem 7 reduces to [8, Theorem 3].

6 Struve functions

The Struve function \mathbf{H}_β of first kind is a particular solution of the second-order inhomogeneous Bessel differential equation

$$z^2 w''(z) + zw'(z) + (z^2 - \beta^2)w(z) = \frac{4 \left(\frac{z}{2}\right)^{\beta+1}}{\sqrt{\pi} \Gamma(\beta + \frac{1}{2})}$$

and have the following form:

$$\mathbf{H}_\beta(z) := \frac{\left(\frac{z}{2}\right)^{\beta+1}}{\sqrt{\frac{\pi}{4}}\Gamma\left(\beta + \frac{1}{2}\right)} {}_1F_2\left(1; \frac{3}{2}, \beta + \frac{3}{2}; -\frac{z^2}{4}\right),$$

where $-\beta - \frac{3}{2} \notin \mathbb{N}$ and ${}_1F_2$ is a hypergeometric function. Since it is not normalized, therefore we take the normalized functions:

$$\begin{cases} U_\beta(z) = \left(\sqrt{\pi}2^\beta \left(\beta + \frac{3}{2}\right) \mathbf{H}_\beta(z)\right)^{\frac{1}{\beta+1}}, \\ V_\beta(z) = \sqrt{\pi}2^\beta z^{-\beta} \Gamma\left(\beta + \frac{3}{2}\right) \mathbf{H}_\beta(z), \\ W_\beta(z) = \sqrt{\pi}2^\beta z^{\frac{1-\beta}{2}} \Gamma\left(\beta + \frac{3}{2}\right) \mathbf{H}_\beta(\sqrt{z}). \end{cases} \quad (6.1)$$

Moreover, for $|\beta| \leq 1/2$, \mathbf{H}_β (see [4, Lemma 1]) and \mathbf{H}'_β have the Hadamard factorizations [8, Theorem 4] given by

$$\mathbf{H}_\beta(z) = \frac{z^{\beta+1}}{\sqrt{\pi}2^\beta \Gamma\left(\beta + \frac{3}{2}\right)} \prod_{n \geq 1} \left(1 - \frac{z^2}{z_{\beta,n}^2}\right)$$

and

$$\mathbf{H}'_\beta(z) = \frac{(\beta+1)z^\beta}{\sqrt{\pi}2^\beta \Gamma\left(\beta + \frac{3}{2}\right)} \prod_{n \geq 1} \left(1 - \frac{z^2}{z_{\beta,n}^2}\right) \quad (6.2)$$

where $z_{\beta,n}$ and $\check{z}_{\beta,n}$ are the n -th positive zeros of \mathbf{H}_β and \mathbf{H}'_β , respectively and interlace [8, Theorem 2]. Thus from (6.2) with logarithmic differentiation, we obtain respectively

$$\frac{z\mathbf{H}'_\beta(z)}{\mathbf{H}_\beta(z)} = (\beta+1) - \sum_{n \geq 1} \frac{2z^2}{z_{\beta,n}^2 - z^2}$$

and

$$1 + \frac{z\mathbf{H}''_\beta(z)}{\mathbf{H}'_\beta(z)} = (\beta+1) - \sum_{n \geq 1} \frac{2z^2}{z_{\beta,n}^2 - z^2}. \quad (6.3)$$

Also for $|\beta| \leq 1/2$, the Hadamard factorization for the entire functions V'_β and W'_β of order $1/2$ [8, Theorem 4] is given by

$$V'_\beta(z) = \prod_{n \geq 1} \left(1 - \frac{z^2}{\eta_{\beta,n}^2}\right) \quad \text{and} \quad W'_\beta(z) = \prod_{n \geq 1} \left(1 - \frac{z}{\sigma_{\beta,n}^2}\right), \quad (6.4)$$

where $\eta_{\beta,n}$ and $\sigma_{\beta,n}$ are n -th positive zeros of V'_β and W'_β , respectively. V'_β and W'_β belong to the Laguerre-Pólya class and zeros satisfy $\eta_{\beta,1}, \sigma_{\beta,1} < z_{\beta,1}$. Now proceeding as in Theorem 6 using (6.1), (6.2), (6.3) and (6.4), we obtain the following results:

Theorem 8 *Let $|\beta| \leq 1/2$. Then the radii of γ -Spirallikeness of order α for the functions U_β, V_β and W_β given by (6.1) are the smallest positive roots of the following equations:*

- (i) $rU'_\beta(r) + ((1 - \alpha) \cos \gamma - 1)U_\beta(r) = 0$
- (ii) $rV'_\beta(r) + ((1 - \alpha) \cos \gamma - 1)V_\beta(r) = 0$
- (iii) $rW'_\beta(r) + ((1 - \alpha) \cos \gamma - 1)W_\beta(r) = 0$

in $(0, z_{\beta,1})$, $(0, z_{\beta,1})$ and $(0, z_{\beta,1}^2)$, respectively.

Remark 7 Taking $\gamma = 0$ in Theorem 8 gives [2, Theorem 2].

Theorem 9 Let $|\beta| \leq 1/2$. Let the functions U_β , V_β and W_β be given by (6.1). Then

- (i) the radius $R_{sp}^c(U_\beta)$ is the smallest positive root of the equation

$$rU''_\beta(r) + (1 - \alpha) \cos \gamma U'_\beta(r) = 0.$$

- (ii) the radius $R_{sp}^c(V_\beta)$ is the smallest positive root of the equation

$$rV''_\beta(r) + (1 - \alpha) \cos \gamma V'_\beta(r) = 0.$$

- (iii) the radius $R_{sp}^c(W_\beta)$ is the smallest positive root of the equation

$$rW''_\beta(r) + (1 - \alpha) \cos \gamma W'_\beta(r) = 0.$$

Remark 8 Taking $\gamma = 0$ in Theorem 9 gives [8, Theorem 4].

7 On Ramanujan type entire functions

Ismail and Zhang [14] defined the following entire function of growth order zero for $\beta > 0$, called Ramanujan type entire function

$$A_p^{(\beta)}(c, z) = \sum_{n \geq 0} \frac{(c; p)_n p^{\beta n^2}}{(p; p)_n} z^n,$$

where $\beta > 0$, $0 < p < 1$, $c \in \mathbb{C}$, $(c; p)_0 = 1$ and $(c; p)_k = \prod_{j=0}^{k-1} (1 - cp^j)$ for $k \geq 1$, which is the generalization of both the Ramanujan entire function $A_p(z)$ and Stieltjes-Wigert polynomial $S_n(z; p)$ defined as (see [15, 22]):

$$A_p(-z) = A_p^{(1)}(0, z) = \sum_{n=0}^{\infty} \frac{p^{n^2}}{(p; p)_n} z^n$$

and

$$A_p^{(1/2)}(p^{-n}, z) = \sum_{m=0}^{\infty} \frac{(p^{-n}; p)_m p^{m^2/2}}{(p; p)_m} z^m = (p; p)_n S_n(z p^{(1/2)-n}; p).$$

Since $A_p^{(\beta)}(c, z) \notin \mathcal{A}$, therefore consider the following three normalized functions in \mathcal{A} :

$$\begin{cases} f_{\beta,p,c}(z) := \left(z^\beta A_p^{(\beta)}(-c, -z^2) \right)^{1/\beta} \\ g_{\beta,p,c}(z) := z A_p^{(\beta)}(-c, -z^2) \\ h_{\beta,p,c}(z) := z A_p^{(\beta)}(-c, -z), \end{cases} \quad (7.1)$$

where $\beta > 0$, $c \geq 0$ and $0 < p < 1$. From [10, Lemma 2.1, p. 4-5], we see that the function

$$z \rightarrow \Psi_{\beta,p,c}(z) := A_p^{(\beta)}(-c, -z^2)$$

has infinitely many zeros (all are positive) for $\beta > 0$, $c \geq 0$ and $0 < p < 1$. Let $\psi_{\beta,p,n}(c)$ be the n -th positive zero of $\Psi_{\beta,p,c}(z)$. Then it has the following Weiersstrass decomposition:

$$\Psi_{\beta,p,c}(z) = \prod_{n \geq 1} \left(1 - \frac{z^2}{\psi_{\beta,p,n}^2(c)} \right). \quad (7.2)$$

Moreover, the n -th positive zero $\Xi_{\beta,p,n}(c)$ of the derivative of the following function

$$\Phi_{\beta,p,c}(z) := z^\beta \Psi_{\beta,p,c}(z) \quad (7.3)$$

interlace with $\psi_{\beta,p,n}(c)$ and satisfy the relation

$$\Xi_{\beta,p,n}(c) < \psi_{\beta,p,n}(c) < \Xi_{\beta,p,n+1}(c) < \psi_{\beta,p,n+1}(c)$$

for $n \geq 1$. Now using (7.1) and (7.2), we have

$$\begin{aligned} \frac{zf'_{\beta,p,c}(z)}{f_{\beta,p,c}(z)} &= 1 + \frac{1}{\beta} \frac{z\Psi'_{\beta,p,c}(z)}{\Psi_{\beta,p,c}(z)} = 1 - \frac{1}{\beta} \sum_{n \geq 1} \frac{2z^2}{\psi_{\beta,p,n}^2(c) - z^2}; \quad (c > 0) \\ \frac{zg'_{\beta,p,c}(z)}{g_{\beta,p,c}(z)} &= 1 + \frac{z\Psi'_{\beta,p,c}(z)}{\Psi_{\beta,p,c}(z)} = 1 - \sum_{n \geq 1} \frac{2z^2}{\psi_{\beta,p,n}^2(c) - z^2}; \\ \frac{zh'_{\beta,p,c}(z)}{h_{\beta,p,c}(z)} &= 1 + \frac{1}{2} \frac{\sqrt{z}\Psi'_{\beta,p,c}(\sqrt{z})}{\Psi_{\beta,p,c}(\sqrt{z})} = 1 - \sum_{n \geq 1} \frac{z}{\psi_{\beta,p,n}^2(c) - z}, \end{aligned}$$

where $\beta > 0$, $c \geq 0$ and $0 < p < 1$. Also, using (7.3) and the infinite product representation of Φ' [10, p. 14-15, Also see Eq. 4.6], we have

$$\begin{aligned} 1 + \frac{zf''_{\beta,p,c}(z)}{f'_{\beta,p,c}(z)} &= 1 + \frac{z\Phi''_{\beta,p,c}(z)}{\Phi'_{\beta,p,c}(z)} + \left(\frac{1}{\beta} - 1 \right) \frac{z\Phi'_{\beta,p,c}(z)}{\Phi_{\beta,p,c}(z)} \\ &= 1 - \sum_{n \geq 1} \frac{2z^2}{\Xi_{\beta,p,n}^2(c) - z^2} - \left(\frac{1}{\beta} - 1 \right) \sum_{n \geq 1} \frac{2z^2}{\psi_{\beta,p,n}^2(c) - z^2}. \end{aligned}$$

As $(z\Psi_{\beta,p,c}(z))'$ and $h'_{\beta,p,c}(z)$ belong to \mathcal{LP} . So suppose $\gamma_{\beta,p,n}(c)$ be the positive zeros of $g'_{\beta,p,c}(z)$ (growth order is same as $\Psi_{\beta,p,c}(z)$) and $\delta_{\beta,p,n}(c)$ be the positive zeros of $h'_{\beta,p,c}(z)$. Thus using their infinite product representations, we have

$$\begin{aligned} 1 + \frac{zg''_{\beta,p,c}(z)}{g'_{\beta,p,c}(z)} &= 1 - \sum_{n \geq 1} \frac{2z^2}{\gamma_{\beta,p,n}^2(c) - z^2} \\ 1 + \frac{zh''_{\beta,p,c}(z)}{h'_{\beta,p,c}(z)} &= 1 - \sum_{n \geq 1} \frac{z}{\delta_{\beta,p,n}^2(c) - z}. \end{aligned}$$

Now proceeding similarly as done in the above sections, we obtain the following results:

Theorem 10 Let $\beta > 0$, $c \geq 0$ and $0 < p < 1$. Then the radii of γ -Spirallikeness of order α for the functions $f_{\beta,p,c}(z)$, $g_{\beta,p,c}(z)$ and $h_{\beta,p,c}(z)$ given by (7.1) are the smallest positive roots of the following equations:

- (i) $rf'_{\beta,p,c}(r) + ((1 - \alpha) \cos \gamma - 1)f_{\beta,p,c}(r) = 0$
 - (ii) $rg'_{\beta,p,c}(r) + ((1 - \alpha) \cos \gamma - 1)g_{\beta,p,c}(r) = 0$
 - (iii) $rh'_{\beta,p,c}(r) + ((1 - \alpha) \cos \gamma - 1)h_{\beta,p,c}(r) = 0$
- in $(0, \psi_{\beta,p,1}(c))$, $(0, \psi_{\beta,p,1}(c))$ and $(0, \psi_{\beta,p,1}^2(c))$, respectively.

We now conclude this section with the convex analog of Theorem 10.

Theorem 11 Let $\beta > 0$, $c \geq 0$ and $0 < p < 1$. Let the functions $f_{\beta,p,c}(z)$, $g_{\beta,p,c}(z)$ and $h_{\beta,p,c}(z)$ be given by (7.1). Then

- (i) the radius $R_{sp}^c(f_{\beta,p,c}(z))$ is the smallest positive root of the equation

$$rf''_{\beta,p,c}(r) + (1 - \alpha) \cos \gamma f'_{\beta,p,c}(r) = 0.$$

- (ii) the radius $R_{sp}^c(g_{\beta,p,c}(z))$ is the smallest positive root of the equation

$$rg''_{\beta,p,c}(r) + (1 - \alpha) \cos \gamma g'_{\beta,p,c}(r) = 0.$$

- (iii) the radius $R_{sp}^c(h_{\beta,p,c}(z))$ is the smallest positive root of the equation

$$rh''_{\beta,p,c}(r) + (1 - \alpha) \cos \gamma h'_{\beta,p,c}(r) = 0.$$

Statements and Declarations

- **Conflict of interest:** The authors declare that they have no conflict of interest
- **Availability of data and materials :** None
- **Authors' contributions :** All authors contributed Equally.

References

1. Aktaş, İ. Baricz, Á. and Orhan, H.: Bounds for radii of starlikeness and convexity of some special functions. Turkish J. Math. **42**, 211–226 (2018) doi: 10.3906/mat-1610-41
2. Baricz, Á. Dimitrov, D.K., Orhan, H. and Yağmur, N.: Radii of starlikeness of some special functions. Proc. Amer. Math. Soc. **144**, 3355–3367 (2016). doi: 10.1090/proc/13120
3. Baricz, Á. Kupán, P.A. and Szász, R.: The radius of starlikeness of normalized Bessel functions of the first kind. Proc. Amer. Math. Soc. **142**, 2019–2025 (2014). doi: 10.1090/S0002-9939-2014-11902-2
4. Baricz, Á. Ponnusamy, S. and Singh, S.: Turán type inequalities for Struve functions. J. Math. Anal. Appl. **445**, 971–984 (2017). doi: 10.1016/j.jmaa.2016.08.026
5. Baricz, Á. and Prajapati, A.: Radii of starlikeness and convexity of generalized Mittag-Leffler functions. Math. Commun. **25**, 117–135 (2020).
6. Baricz, Á. and Szász, R.: The radius of convexity of normalized Bessel functions. Anal. Math. **41**, 141–151 (2015). doi: 10.1007/s10476-015-0202-6
7. Baricz, Á. Toklu, E. and Kadioğlu, E.: Radii of starlikeness and convexity of Wright functions. Math. Commun. **23**, 97–117 (2018).

8. Baricz, Á. and Yağmur, N.: Geometric properties of some Lommel and Struve functions. *Ramanujan J.* **42**, 325–346 (2017). doi: 10.1007/s11139-015-9724-6
9. Bulut, S. and Engel, O.: The radius of starlikeness, convexity and uniform convexity of the Legendre polynomials of odd degree. *Results Math.* **74**, Paper No. 48, 9 pp (2019). doi: 10.1007/s00025-019-0975-1
10. Deniz, E.: Geometric and monotonic properties of Ramanujan type entire functions. *Ramanujan J.* **55**, 103–130 (2020). doi: 10.1007/s11139-020-00267-w
11. Deniz, E. and Szász, R.: The radius of uniform convexity of Bessel functions. *J. Math. Anal. Appl.* **453**, 572–588 (2017). doi: 10.1016/j.jmaa.2017.03.079
12. Dimitrov, D.K. and Ben Cheikh, Y.: Laguerre polynomials as Jensen polynomials of Laguerre-Pólya entire functions. *J. Comput. Appl. Math.* **233**, 703–707 (2009). doi: 10.1016/j.cam.2009.02.039
13. Gangania, K. and Kumar, S.S.: $S^*(\psi)$ and $\mathcal{C}(\psi)$ -radii for some special functions. *Iran. J. Sci. Technol. Trans. A Sci.* **46**, 955–966 (2022).
14. Ismail, M.E.H. and Zhang, R.: q -Bessel functions and Rogers-Ramanujan type identities. *Proc. Amer. Math. Soc.* **146**, 3633–3646 (2018). doi: 10.1090/proc/13078
15. Ismail, M.E.H.: Classical and quantum orthogonal polynomials in one variable. *Encyclopedia of Mathematics and its Applications*, 98, Cambridge University Press, Cambridge, 2005.
16. Kumar, S.S. and Gangania, K.: Subordination and radius problems for certain starlike functions. arXiv:2007.07816 (2020)
17. Kumar H and Pathan A.M. On the distribution of non-zero zeros of generalized Mittag-Leffler functions. *International Journal of Engineering Research and Application*, **6**, 66–71 (2016).
18. Ya. Levin B. Lectures on Entire Functions. Translation of Mathematics Monographs, American Mathematical Society, Providence 150 (1996).
19. Ma, W.C. and Minda, D.: A unified treatment of some special classes of univalent functions. *Proceedings of the Conference on Complex Analysis, Tianjin, Conf Proc Lecture Notes Anal, I Int Press. Cambridge, MA.* 157–169 (1992).
20. Pfaltzgraff, J.A.: Univalence of the integral of $f'(z)^\lambda$. *Bull. London Math. Soc.* **7**, no. 3, 254–256 (1975).
21. Prabhakar, T.R.: A singular integral equation with a generalized Mittag Leffler function in the kernel. *Yokohama Math. J.* **19**: 7–15 (1971).
22. Ramanujan, S.: The lost notebook and other unpublished papers, Springer-Verlag, Berlin, 1988.
23. Robertson, M.S.: Univalent functions $f(z)$ for which $zf'(z)$ is spirallike. *Michigan Math. J.* **16**, 97–101 (1969).
24. Spacek, L.: Contribution á la théorie des fonctions univalentes, *Casop Pest. Mat.-Fys.* **62**, 12-19 (1933).
25. Kazımoğlu, S. and Deniz, E.: Radius Problems for Functions Containing Derivatives of Bessel Functions. *Comput. Methods Funct. Theory* (2022). <https://doi.org/10.1007/s40315-022-00455-3>
26. Watson, G.N. : A Treatise on the Theory of Bessel Functions, Cambridge University Press, 1944.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/365978020>

Licensed Under Creative Commons Attribution CC BY Recent Advances in Various Types of Forging –A Research Review

Conference Paper · November 2021

DOI: 10.21275/SR21111618554

CITATIONS

0

READS

14

2 authors, including:



[Singye Wangchuk](#)

Delhi Technological University(DTU)

4 PUBLICATIONS 0 CITATIONS

[SEE PROFILE](#)

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



Production and Industrial Engineering [View project](#)

Recent Advances in Various Types of Forging - A Research Review

S. Wangchuk¹, Dr. AK Madan²

¹Production and Industrial Engineering, Department of Mechanical Engineering, Delhi Technological University, Bawana Road, Shahbad Daulatpur, Rohini, Delhi 110042, India

²Professor, Department of Mechanical Engineering, Delhi Technological University, Delhi, India

¹Corresponding Author Email: [singyewangchuk_pe20a2_19\[at\]dtu.ac.in](mailto:singyewangchuk_pe20a2_19[at]dtu.ac.in)

Abstract: Forging is one of the famous and primitive sorts of manufacturing processes. The history of forging dates to even earlier than 4000BC which involved a couple of hand tools and anvils called smith forging which relied on hits and trials requiring huge skills and high precisions. These days, the forging tools, operations, and methods are automated and power-driven. In this research, the recent developments and advancements in various forging processes are analyzed by reading various research papers and journals of the last two decades. The papers were evaluated for substantial upgrades in precision and technologies. The use of power hammers, artificial intelligence such as the use of robotics, computer numerical controls (CNCs) has made forging a drastic shift towards mechanization. This review will help to highlight the recent progress of the forging processes and act as an easy guide to encapsulate the recent optimization of various processes.

Keywords: Forging, power hammers, artificial intelligence, computer numerical controls, optimization.

1. Introduction

The manufacturing process which involves the shaping of metals by use of compressive power is termed forging. Warm, cold, and hot are the three types of forging that are named based on the temperature which is done [1]. It is an effective method of producing workpieces of various shapes of discrete parts. It includes bolts, rivets, crane hooks, connecting rods, gears, turbine shafts, hand tools, railroads [2]. Forging parts are used from early times but their demand drastically increased during the industrial revolution for developing new technologies and improvements of mechanical properties of the material [1]. Forged parts are considered better than casting due to grain flow [1, 2].

2. Importance of forging

Forging techniques have various significances. The forging process is an economical manufacturing process and increases mechanical properties like strength, toughness, and hardness [5]. Moreover, can reduce defects of casting (porosities, cavities) thus obtaining almost perfect workpieces by ensuring uniform plastic strain throughout the job [14, 15]. Forging allows varieties of metals and alloys to be forged such as Aluminium, Copper, Magnesium, Carbon and low steels, Nickel, and Titanium alloys [1]. It produces heavy-duty components and versatile dimensions of work pieces [3].

3. History of Forging Process

The art of forging dates to as early as 4000BC. Blacksmiths are early pioneers of forging processes. It involved self-skills and manual work. Hammers are lifted by hands placing work pieces on anvil and hydraulic energies were used by some blacksmiths at their workplace. Some of the forging that time constituted of set hammer, anvil, swage

block, tongs, hammer, hot set, heading tools, fullers, flatters, punches, and drifts. The energy sources used in early forging processes included the charcoal to heat metal piece up to certain forging temperature [1]. Toward the end of the 19th century the simultaneous development of the open-hearth steel making processes were employed making forging industry now had a reliable, low-cost volume raw material [12]. Some of furnace types are continuous and batch type, box type, muffle types and electric resistance heated furnaces [1].

1) Open die-forging process

Open die forging is an essential forging technology mainly employed to produce large components with improved tensile properties and toughness behavior together with reliability of the forged parts [6]. It involves pressing of workpiece using dies of various shapes: V-shape or concave. The job undergoes a plastic deformation at high temperature followed by presses of multiple strokes along feed direction. It leads to change internal and geometric properties of workpiece [14]. It is mainly used to produce large parts with good mechanical properties and reliability [5].

In the past, the common idea of the work was to develop a process model which can merge data from online measurements and plasto-mechanical model for determination of equivalent strain, strain rate and core temperature of the workpiece [14]. The Artificial Neural Network (ANN), an algorithm developed are used mainly to calculate optimum number of passes for reduction of forging cycle and economization of power. A new formulation of ANN is employed to quickly evaluate plastic strain at the core of job. The correct evaluation of plastic strain can improve the internal integrity of material and optimization of microstructures [7].

2) Impression die forging

It uses slapped die for controlling metal flow. The heated metal is located at the lower cavity and one or more blows are attached on upper part of die. The metal flows when hammered and die cavity is filled completely. Excess metal is squeezed out around the periphery of the cavity. The flash formed is cleared out (by trimming) with trimming die. It contains different die cavities. The final shape to job is obtained by subjecting to series of cavities in the die set. The die cavities are designed in such a way that the metal flows evenly so that desired shape is obtained according to die cavities [16].

It is modified to auto forging. In this the metal piece is removed from the mold while it is hot. It is trimmed later like old techniques. The transferring from mold to die, forging and trimming are highly mechanized perhaps called Auto forging [19].

3) Isothermal forging processes

Die chilling involves the flow of metals from workpiece to die surfaces resulting in thermal gradients in the job. The plastic deformation is not uniform as the colder places of die areas has flow as compared to the hotter core areas. It is commonly heated to maximum temperature of (400 to 500)°F [205-260°F]. The chilling is reduced by using speed forging machines, hammers, screw, and mechanical presses. The use of glass lubricants also reduces thermal chilling processes. The die is heated to certain temperature equal to that of workpiece hence reducing chilling. It is called isothermal forging [17].

Defense Research and Development Organization (DRDO) has established isothermal forging processes such as high-pressure compressors (HPC) for deforming titanium alloy. The technology developed by Defense Metallurgy Research Laboratory (DMRL) a premier metallurgical Laboratory of DRDO at Hyderabad. It established self-reliance in the aerospace industry [8, 10].

4) Press forging process

Press forging is the technique of shaping of metal by placing between two dies through mechanical or hydraulic pressure. It is done in forge press, a machine which applies gradual pressure in each die. The shape of workpiece is obtained by single stroke of pressure in each die among series of successive aligned dies [11].

It is mostly used for carrying out heavy forging of large sections of metal by using hydraulic presses. A continuous pressure by series of hydraulic presses makes deformation of job uniform. Hydraulic presses are available in capacity of 5-500MN but 10-100MN range are commonly used [9]. Press forging needs less flash and draft compared to open forging. Its applications are coining and hubbing [11].

In last two decades, it has emerged as alternative technique in manufacturing thin-walled electronic components of magnesium alloys. The process squeezes a thick sheet along pressure direction which is different from traditional ways of using thick sheets [18].

5) Upset forging

The upset forging involves increasing the cross section of metal piece in expense of its length. It was developed initially for producing continuous bolt heads. Parts are upset-forged from bars and rods of up to 200mm in both cold and hot conditions. Some of parts forged by this technique includes nails, valves, fasteners, and couplings [9].

Recently, upset forging is done for producing cylindrical billets having different frictional conditions at two die surfaces [1].

6) Roll forging process

Roll forming is mainly performed under hot conditions. It is of two types depending upon types of shape and tool set up and motions. They are discussed as below:

- **Longitudinal rolling:** the job undergoes translational motion between two spinning tools. The translational motion takes place parallel to the axis of the work;
- **Cross rolling:** the work rotates between two tools which is set to same rotary motion as job. The point contact between tools and workpiece makes job to move along the plane perpendicular to axis of the work;
- **Helical rolling:** the work piece is subjected to both translational and rotary motions where rolls also rotating at same direction. The points if contact between work and tools results in central motion [4].

Roll forging is now used to make reductions in cross-sections and distribution of a metal of billet which reduces extensive work using forging hammers or presses [20].

7) Net shape or near net-shape forging

In this forging, the metal deformation takes place in the cavity where no flash is formed, and final dimensions are very accurate. Its produces parts that would require least or no machining process to complete. It can produce stronger components which can improve performance of engine [17]. The process uses dies with greater dimensional accuracy than any other dies which also requires high driving power. Typical parts forged by this method includes gears, turbine blades, fuel injection nozzles, and barring castings [19].

It is an alternative for open die forging [19].

In the last decade, CAD, CNC technologies and innovation in materials have contributed enormously on the development of NNS (Near net shape forging) technologies [13].

Meleform Flo- forge is an automated near net shape forging technology which requires no applied force but instead only one hydraulic cylinder. It also uses modular furnaces to maintain uniform die temperature [21].

4. Conclusion

From this short review study, it can be summarized that various advances in different categories of forging technologies have been employed through various advances in science and technologies. In each technique starting from forging at early age, it has been observed that the

methodologies have been enhanced requiring minimal manual human work force. Moreover, it should be noted that forging time and quality of forged parts are being drastically optimized for greater productivity with minimal defects in finished forged components.

References

- [1] Shalok Bharti, "Advancement in Forging Process: A Review", International Journal of Science and Research (IJSR), https://www.ijsr.net/get_abstract.php?paper_id=ART20178736, Volume 6 Issue 12, December 2017, 465 - 468
- [2] Mr.Praveena R, Mr.Abhishek C R, Mr. MujiburRehaman, Mr. Bharath Gowda G, 2019, Study on Development of an Automated Open Die Forging Machine, International Journal Of Engineering Research & Technology (Ijert) Ncmpe - 2019 (Volume 7, Issue 07)
- [3] Gontarz A. Forming process of valve drop forging with three cavities. Journal of Materials Processing Technology. 2006;177(1-3):228–32.
- [4] Pater Z, Tofil A. Overview of the research on roll forging processes. Advances in Science and Technology Research Journal. 2017Jun1;11(2):72–86.
- [5] Mancini S, Langellotto L, Zangari G, Maccaglia R, Schino A. Optimization of open die ironing process through artificial neural network for rapid process simulation. Metals. 2020Oct21;10(10):1397–411.
- [6] Di Schino A. Open die forging process simulation: A simplified industrial approach based on Artificial Neural Network. AIMS Materials Science. 2021Sep9;8(5):685–97.
- [7] Kim, P.H.; Chun, M.S.; Yi, J.J.; Moon, Y.H. Pass schedule algorithms for hot open die forging. J. Mater. Process. Technol. 2002, 130, 516–523.
- [8] Raghu T, I Balasundar I, Rao M. Isothermal and near isothermal processing of titanium alloys. Defence Science Journal. 2011Jan6;61(1):72–80.
- [9] Prashant. FORGING. Lucknow, Uttar Pradesh: Lucknow University.
- [10] Ministry of Defense. (2021, May 28). *DRDO develops critical near isothermal forging technology for aeroengines: Defence research and development organisation - DRDO, Ministry of Defence, Government of India*. DRDO develops Critical Near Isothermal Forging Technology for aeroengines | Defence Research and Development Organisation - DRDO, Ministry of Defence, Government of India. Retrieved October 11, 2021, from <https://www.drdo.gov.in/press-release/drdo-develops-critical-near-isothermal-forging-technology-aeroengines>.
- [11] Analytik, F. I. O. N. I. C. O. N., Instruments, F. T. A., Technologies, F. Y. F. I., Solutions, F. H. C. S., Corporation, F. S., Ltd., F. B. I., & Radleys, F. (2019, August 1). *Press forging – metallurgical processes*. AZoM.com. Retrieved October 11, 2021, from <https://www.azom.com/article.aspx?ArticleID=9696>.
- [12] Analytik, F. I. O. N. I. C. O. N., Instruments, F. T. A., Technologies, F. Y. F. I., Solutions, F. H. C. S., Corporation, F. S., Ltd., F. B. I., & Radleys, F. (2017, August 1). *Forging - history and key developments in the metals forging industry*. AZoM.com. Retrieved October 11, 2021, from <https://www.azom.com/article.aspx?ArticleID=2195>.
- [13] Marini, D., Cunningham, D., & Corney, J. R. (2017). Near net shape manufacturing of metal: A review of approaches and their evolutions. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 232(4), 650–669. <https://doi.org/10.1177/0954405417708220>.
- [14] Sharma, D.K.; Filippini, M.; Di Schino, A.; Rossi, F.; Castaldi, J. Corrosion behavior of high temperature fuel cells: Issues for materials selection. *Metalurgia* 2019, 58, 347–351.
- [15] Wolfgarten, M.; Rosenstock, D.; Schaeffer, L.; Hirt, G. Implementation of an open-die forging process for large hollow shafts for wind power plants with respect to an optimized microstructure. *AIM* 2015, 107, 43–49.
- [16] Virginia Polytechnic Institute and State University. (2011). *Hot-rolling, cold-forming; extrusion, and forging*. Structure And Form Analysis System (SAFAS). Retrieved October 2, 2021, from https://www.setareh.arch.vt.edu/safas/007_fdmtl_34_rolling_extrusion_forging.html.
- [17] Forging Industry Educational and Research Foundation. (2020, January 12). *5.2.2.4 hot die and isothermal forging: Forging Industry Association*. 5.2.2.4 Hot Die and Isothermal Forging | Forging Industry Association. Retrieved October 5, 2021, from <https://www.forging.org/forging/design/5224-hot-die-and-isothermal-forging.html>.
- [18] Chen, F.-K., Huang, T.-B., & Wang, S.-J. (2007). A study of flow-through phenomenon in the press forging of magnesium-alloy sheets. *Journal of Materials Processing Technology*, 187-188, 770–774. <https://doi.org/10.1016/j.jmatprotec.2006.11.192>
- [19] Shan, H. S. Prof. (2009, December 31). *Mechanical Engineering - Manufacturing Processes I*. NPTEL (The National Programme on Technology Enhanced Learning). Retrieved October 7, 2021, from <https://nptel.ac.in/courses/112/107/112107144/>.
- [20] A Text Book of Production Technology, O.P. Khanna, M. Lal; Dhanpat Rai Publication, pp. 16.24 - 16.31, 16.49, 16.53.
- [21] Di Schino, A. (2021). Open die forging process simulation: A simplified industrial approach based on Artificial Neural Network.

Author Profile



Singye Wangchuk is currently pursuing second year B.Tech. Production and Industrial Engineering (PIE) in Delhi Technological University (DTU). He is from Bhutan.



Retraction Note: Indian smart city ranking model using taxicab distance-based approach

Kapil Sharma¹ · Sandeep Tayal^{1,2}

© Springer-Verlag GmbH Germany, part of Springer Nature 2022

Retraction Note: *Energy Systems* (2019) 13:625–642

<https://doi.org/10.1007/s12667-019-00365-9>

The Editor-in-Chief and the publisher have retracted this article. The article was submitted to be part of a guest-edited issue. An investigation by the publisher found a number of articles, including this one, with a number of concerns, including but not limited to compromised editorial handling and peer review process, inappropriate or irrelevant references or not being in scope of the journal or guest-edited issue. Based on the investigation's findings the Editor-in-Chief therefore no longer has confidence in the results and conclusions of this article. Kapil Sharma & Sandeep Tayal have not responded to correspondence regarding this retraction.

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

The online version of the original article can be found at <https://doi.org/10.1007/s12667-019-00365-9>.

✉ Kapil Sharma
kapil@ieee.org

¹ Delhi Technological University, Delhi, India

² Maharaja Agrasen Institute of Technology, Delhi, India

Review on Chloride Ingress in Concrete: Chloride Diffusion and Predicting Corrosion Initiation Time

Pradeep K. Goyal¹, Andualem E. Yadeta²

^{1,2}Department of Civil Engineering, Delhi Technological University, Delhi - 110042, India

Received Date: 10 March 2022

Revised Date: 18 April 2022

Accepted Date: 25 April 2022

Abstract - Corrosion is a common cause of degradation of reinforced concrete (RC) structures. Even though various reasons cause the phenomenon, chloride-induced corrosion in concrete is an important issue in current research. The maintenance of corroded RC increases the cost of facilities; durability assessment is helpful for economic savings and the performance of structures. Therefore, this paper reviews the chloride diffusion process and the time it takes for corrosion to start in RC structures. In this paper, Fick's second law is considered for the transport mechanism of chlorides into concrete structures. Different models for predicting corrosion initiation time by previous scholars are critically reviewed.

Keywords - Concrete structures, Durability, Chloride diffusion, Reinforcement bar, Corrosion.

I. INTRODUCTION

Concrete durability is vital for the economy and the safety of society. However, the performance of constructed facilities can be affected by strength degradation over time. Among the common degradation mechanisms in concrete structures, rebar corrosion is a significant issue. Corrosion may occur by transporting different agents like chloride ions into concrete. Chlorides can present in the fresh concrete mix when chloride-contaminated materials are used during concrete mixing, and they can also diffuse into the hardened concrete at a later stage. When RC structures are exposed to seawater or de-icing salts such as calcium chloride (CaCl_2) and sodium chloride (NaCl), steel passivity will be broken, risk steel reinforcement cross-section loss. Corrosion risk increases as chloride concentration rise; however, it can be mitigated by using supplemental cementitious elements in concrete (Zhou et al., 2014). It is also further stated that including silica fume and corrosion inhibitors in the concrete mix helps speed up the onset of corrosion in concrete (Dinh 2017). Even though structures are designed for a specific service period, corrosion in concrete degrades the durability of concrete structures, and they fail to provide the required service. To protect against rebar corrosion in concrete,

identifying the transport mechanism of chlorides is helpful (Liu, Easterbrook, and Li, 2017).

Chlorides can penetrate the concrete by the diffusion transport mechanism, assuming that the concrete cover is in a fully saturated environmental condition. This general diffusion process in materials is described using Fick's second law (Que 2007). Among the previous works done to predict corrosion initiation time in concrete reported by different authors, most models were developed based on concrete microstructure (Jaturapitakkul, Cheewaket, and Chalee 2014). The diffusion process in concrete structures can be affected by different environmental factors, and there is a need to review the available models for the corrosion process in RC structures.

Therefore, the objective of this study is to review the process of chloride diffusion and determine the appropriate model to predict the corrosion initiation time in concrete. First, the chloride diffusion mechanisms are investigated using Fick's second law. Secondly, the most suitable model for predicting corrosion initiation time in concrete was determined after a thorough assessment of several models proposed by previous scholars.

A. Steel Corrosion

Corrosion is a gradual disintegration of materials (mostly metals) caused by chemical and/or electrochemical reactions with their environment, resulting in mass loss and dimension changes (ASTM G193-12d 2012). Corrosion requires water, oxygen, and ions which are all present in the environment (Palanisamy 2019). The general steel corrosion process can be briefly explained using a simple model (see Fig. 1).



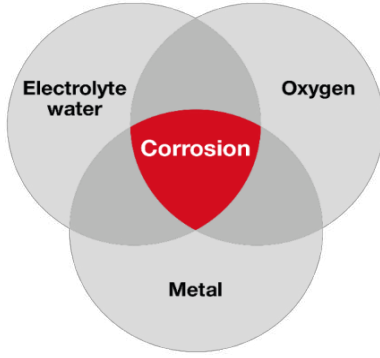


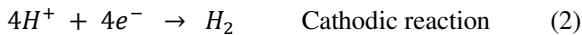
Fig. 1 Corrosion process

The reinforcing steels used in concrete structures have a passive coating on their surface that protects them from corrosion. Because the pH of concrete is very high, around 13, in mild ambient circumstances and the passive protective layer inhibits corrosion initiation. The change of steel into rust is an electrochemical process in nature, commonly called electrochemical reaction since it is different from the chemical reaction as it involves the passage of electrical charges.

Corrosion is a two-stage electrochemical process in general: primary and secondary process. In the primary process, the steel dissolves in the pore water solution:



To keep the electrical neutrality of the reaction process, the electrons released in the primary process should be utilized, and then the secondary electrochemical process occurs on metals:



Therefore, corrosion starts in the secondary stage of the electrochemical process. Many different products can be formed based on influencing factors like the pH of the solution and the environmental condition. The secondary reaction process consumes the Fe^{2+} produced in the primary process. Then it initiates corrosion by taking oxygen and hydrogen from pore water, as illustrated in Fig. 2 below:

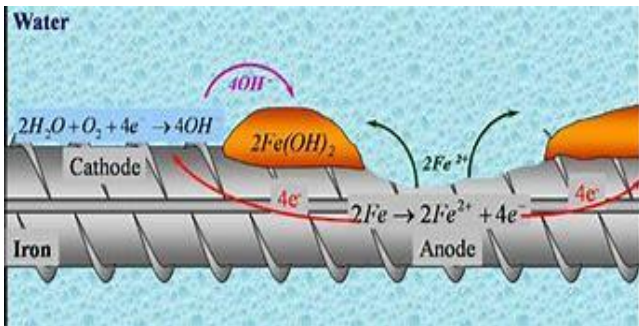
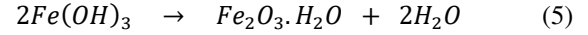
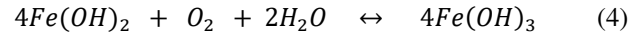
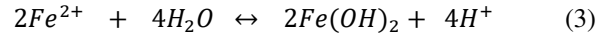


Fig. 2 Corrosion cell (Portland Cement Association 2019)



The corrosion product $Fe(OH)_2$ will be further oxidized to Fe_2O_4 or Fe_2O_3 in the usual circumstances of steel embedded in concrete. The equations in 1-5 are universal equations that can be applied to any metal corrosion process. In this study, chloride-induced steel corrosion is a specific scenario in which chlorides penetrate concrete from the environment and cause corrosion by lowering the pH and breaking the passive layer of the bar.

B. Corrosion influencing factors in concrete

Corrosion can be influenced by different factors that determine structures' performance against a corrosive environment. These can be categorized as practices during design, construction, and environmental factors. The construction practice determines concrete's durability properties (permeability, porosity, etc.). For example, appropriate curing decreases the porosity, which determines the durability of concrete against aggressive environmental actions (Koleva et al. 2008). The design practice significantly influences the durability of concrete in a corrosive environment (Alonso and Sanchez 2009). The penetration of chlorides into concrete can be minimized with a lower w/c ratio and increased concrete strength (Liu, Hu et al., 2020). According to (Kristawan et al. 2017), the larger concrete cover increases the time for rebar in structures to start rusting. The performance of concrete structures against corrosion also depends on environmental conditions such as temperature and humidity.

II. CHLORIDE-INDUCED CONCRETE

Even though there are many aggressive agents in the environment, the exposure of RC structures to chlorides is the principal cause of reinforcement corrosion. Chlorides can be dissolved in water and diffuse through concrete or reach the steel via pore structures and micro-cracks (Tang and Gan 2015). Chloride-containing additives can also cause corrosion. The chloride ions contained in de-icing salts and seawater can cause corrosion when oxygen and moisture are present, as illustrated in Fig. 3 below. Chlorides in concrete can also be facilitated by cement, water, aggregate, and admixtures.



Fig. 3 An RC structure exposed to chlorides (Legault 2011)

Chloride ions develop an oxide film over the reinforcing steel, which speeds up corrosion reactions and reduces structural durability (Valipour et al., 2013). Even though the chlorides penetration process is complex, diffusion is the most common transport mechanism used to assess the durability of concrete (Liu, Easterbrook, and Li, 2017). The mechanism by which chlorides cause corrosion is unknown. Still, the most widely accepted explanation is that chlorides penetrate the protective layer of the reinforcement more easily than other ions, exposing the rebar to corrosion. Then the corrosion products continue to expand, causing component cracking and failure (Chen, Baji, and Li, 2018). It is shown in Fig. 4 below that chloride-induced corrosion is occurred due to exposure of an RC structure to seawater.



Fig. 4 Chlorides-induced corrosion (Graham 2017)

The chloride-induced reinforcement corrosion in RC structures is essentially an electrochemical reaction process where the passive protective layer of steel is lost due to chloride ions forming microcells on the reinforcement surface. The electrochemical process is initiated when moisture in the pores of concrete functions as an electrolyte and the area next to the chloride ion concentration acts as a cathode (Zhang, Zhang, and Ye, 2018).

III. CHLORIDE DIFFUSION

Diffusion is the movement of any substance from a region of a high concentration area to a low concentration area along a concentration gradient (Pal, Paulson, and Rousseau 2013). It is a very important phenomenon as many reactions are diffusion-dependent processes. In the case of diffusion of chlorides in concrete, only the free chlorides contribute to the reaction potential. In this process, free chlorides initiate corrosion while the bound chlorides can be released (Luque et al., 2014). The concentration gradient and chlorides diffusion result from a high concentration region of chlorides. As a result, chloride diffusion is determined by absorption for a particular depth into concrete, and diffusion becomes the major transport mechanism in concrete (Hunkeler 2005). The typical chloride diffusion process in concrete is illustrated in Fig. 5 below.

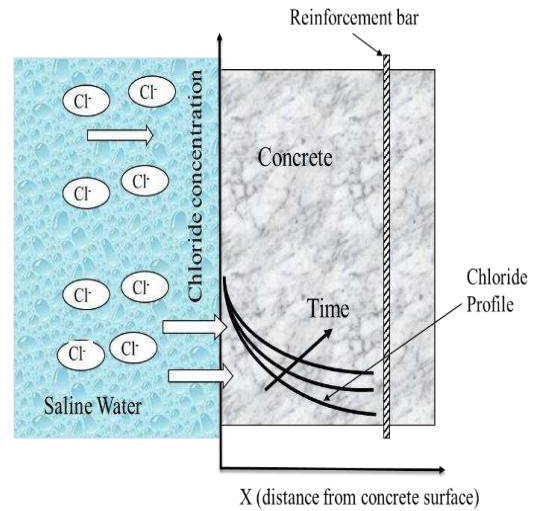


Fig. 5 Typical chloride ions diffusion process in RC

The diffusion process generally occurs under the steady-state and the non-steady-state conditions. Under the steady-state condition, the chloride concentration in the diffusion process is constant, whereas it is time-dependent under the non-steady-state condition. Under the steady-state conditions, diffusion responds to the concentration gradient, which is expressed as the changes in concentration due to the changes in position, $\partial c / \partial x$. This diffusion process is commonly expressed using Fick's first law as follows:

$$J = -D_e \frac{\partial c}{\partial x} \quad (6)$$

Where:

J is the diffusion flux (flow rate of ions) of the system (cm^2/s);

D_e is the coefficient of diffusion or diffusivity for Fick's 1st law (cm^2/s), and

$\partial c / \partial x$ is the gradient concentration. If linear $\partial c / \partial x = \Delta C / \Delta X = C_2 - C_1 / X_2 - X_1$.

Flow occurs at the negative concentration gradient, as shown by the negative signal in the equation. In Fick's first law, the diffusion coefficient, D_e is called the 'effective' diffusion coefficient. This coefficient measures concrete penetration in relation to the flow rate of chlorides in concrete. (Tang, Nilsson and Basheer 2011). Under the steady conditions, diffusion is calculated using Fick's second law. Fick's second law is found in the first law to determine how scattering causes a change in the focus of an object in relation to time. Since concrete is an equally permanent method in the filling area, Fick's second law is often used to describe the process of chloride diffusion in concrete (Martin-Perez et al., 2000, Yuan et al., 2009, Wang, Gong, and Wu, 2019):

$$\frac{\partial c}{\partial t} = -D_a \frac{\partial^2 c}{\partial x^2} \quad (7)$$

where:

c is the concentration of chlorides (cm^3)

t is the exposure duration (s);

x is the depth (cm); and

D_a is the diffusivity for Fick's second law, also called the 'apparent diffusivity' (cm^2/s).

Several attempts were made to determine the apparent diffusion coefficient in Equation 7 to determine the corrosion initiation time (Garboczi and Bentz 1992, Ababneh, Benboudjema and Xi 2003, Oh and Jang 2004, Bentz 2007, Han 2007, Wang and Ueda 2011, Li, Xia and Lin 2012, Zhang, Dong and Jiang 2013). Most of them reported that the apparent difference could be obtained by inserting the error line function of Fick's second law into the chloride profile of the particular concrete being investigated. (Wang, Gong, and Wu 2019). The apparent diffusivity cannot be considered a parameter, but it is a regression coefficient that measures the permeability of a particular

concrete exposed to a chloride-rich environment (Tang, Nilsson, and Basheer 2011). The difference between the apparent diffusion and the effective diffusion coefficient is that the apparent diffusivity includes the binding capacity of concrete in Fick's second law, but the first law does not. Chloride binding plays a crucial role in the chloride diffusion process in concrete. The relationship between the two coefficients can be determined from Fick's second law as given in equation (8) below:

$$D_a = \frac{D_e}{p_{sol} (1 + \frac{\partial c_b}{\partial c})} \quad (8)$$

Where p_{sol} is the porosity of the concrete.

IV. PREDICTION OF CORROSION INITIATION TIME IN CONCRETE

Mathematically, the time to start corrosion in concrete can be calculated based on the diffusion parameters, especially can be taken from Fick's second law. This is because of the difference in the concentration of the chlorides from the structure's surface penetration through the concrete into the reinforcement surface. If the rebar area's chloride ions exceed the concrete's critical threshold and destroy the resulting layer, Fick's second rule applies, and then corrosion begins. The corrosion rate of the reinforcement is heavily employed in RC structural evaluations, maintenance decisions, and residual life forecasts. (Yu et al. 2014). Therefore, most of the mathematical equations developed for durability assessment of RC structures are related to steel corrosion following a simple model that was developed by (Tuutti 1980) in the very back period in which corrosion mechanism is divided into two phases, as shown in Fig. 6 below.

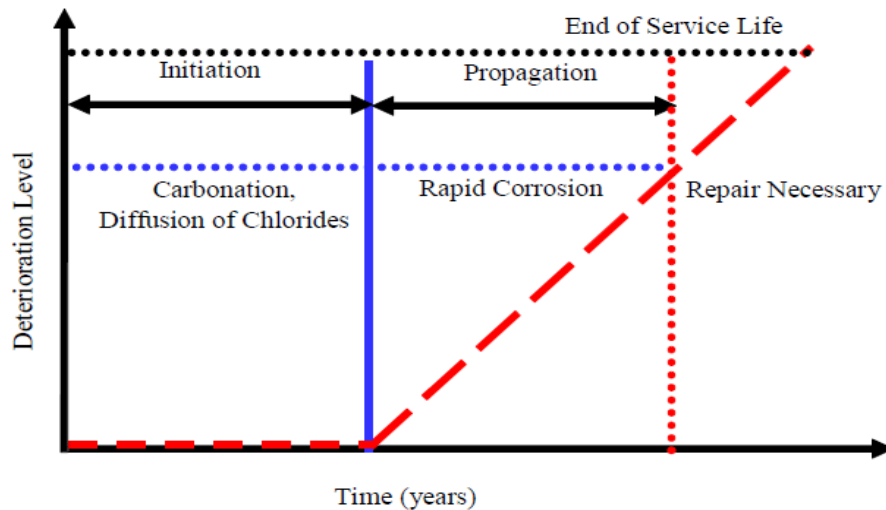


Fig. 6 Tutti's model (Andrade 2020)

$$t = t_i + t_p \quad (9)$$

Where t_i is the initiation period for corrosion
 t_p is the propagation period for corrosion

The initiation time is the amount of time it takes for chlorides on the surface of the concrete to permeate into the micropores. When the concentration of chlorides reaches a threshold level, steel reinforcement de-passivates, and the corrosion start phase terminates. The propagation period then begins, during which a chemical reaction occurs that causes the reinforcing steel to rust.

A. Fick's second law of diffusion

The passage of chloride ions into structures is the primary cause of rebar corrosion in concrete. As a result, based on Equation 7 of Fick's second law, the following formula can be used to calculate the rebar corrosion initiation time in concrete (Bazant 1979):

$$t_{ic} = \frac{x_c^2}{4D_c} [erf^{-1}(1 - \frac{C_{cr}}{C_o})]^{-2} \quad (10)$$

From previous studies, the mathematical equations to determine chlorides diffusion into concrete as a depth and time function are computed using Fick's second law (Equation 7), as discussed in the previous section. From this equation, the corrosion initiation time can be calculated. The solution to this one-dimensional diffusion problem of a specific material depends on boundary conditions and certain assumptions.

B. Constant Chloride Diffusion

Chloride diffusion is a complex process as the transport of chlorides is specifically affected by other ions present in concrete pore solution. The following general solution was proposed by most studies, assuming constant chloride diffusivity and chloride concentration at the surface of RC structures. Fick's second law leads to the following equation when these parameters are treated as constants (Anacta 2009):

$$C(x, t) = C_i + (C_{sa} - C_i)erfc \left[\frac{x}{\sqrt{4(t - t_{ex})D_a}} \right] \quad (11)$$

where:

$C(x, t)$ is the chloride concentration at time t and depth x

C_i is the initial chloride concentration (constant depth).

C_{sa} is the chloride concentration on concrete surface

D_a is the apparent coefficient of diffusion (constant)

x is the depth from the surface

t is the inspection time

We can solve (Equation 11) by taking D as a constant over time, and the surface chloride concentration is constant; we have (Crank 2004):

$$C(x, t) = C_s [1 - erf(\frac{x}{2\sqrt{Dt}})] \quad (12)$$

where:

$C_{x,t}$ is the concentration of chlorides at depth x and time t ,

C_s is the concentration of chlorides on concrete surfaces,

C_o is the initial uniform concentration of chlorides at the surface,

x is the depth from the surface, t is time, and

D is the apparent coefficient of chloride diffusion.

When it comes to concrete, applying the constant chloride diffusivity concept is limited to a particular structure with long exposure to environmental actions. This is because, at significantly longer exposure of structures to chlorides, the chloride diffusivity of the concrete exhibits a constant character.

C. Time-dependent chloride diffusion

The chloride concentration varies with time because the diffusion rate determines the amount of diffusing chloride ions into concrete at a given depth. As the exposure duration increases, the chlorides ion concentration in concrete drops. According to (Poulsen 1993), an equation for the determination of a time-dependent diffusion coefficient (D_a) is given by:

$$D_a = \frac{1}{t} \int_0^t D(t) dt \quad (13)$$

Different scholars previously developed many models. Among the different models, the model created by (Anacta 2009) for corrosion initiation time prediction considers some environmental factors such as temperature, humidity, and rainfall in his assumption. This model also further considered the influence of the duration of concrete exposure to environmental actions. In this model, the equation to calculate the depth of chloride diffusion, which helps to compute corrosion initiation time, was formulated as shown in (Equation 14) below:

$$x_c = 2s\sqrt{D_c t} \quad (14)$$

The chloride diffusion coefficient (D_c), which is taken as the function of environmental factors, is given by:

$$D_c = D_{c,rmt} x f_1(t) x f_2(T) x f_3(RH) x f_4(R) \quad (15)$$

In this expression, $D_{c,rmt}$ is called the reference coefficient of chloride diffusion, which can be obtained from the laboratory test. To critically review the previous models reported by different authors to predict corrosion initiation time are summarized (see Table1).

Table 1. Comparison of different models for prediction of corrosion initiation time in concrete

Model	Basis	Equation
Clear's Model (Clear 1976)	Empirical	$t_{ic} = \frac{129 \cdot x_c^{1.22}}{\left(\frac{w}{c}\right) \cdot [C_s]^{0.42}}$
Bazant's Model (Bazant 1979)	Fick's second law	$t_{ic} = \frac{x_c^2}{4D_c} [\operatorname{erf}^{-1} (1 - \frac{C_{cr}}{C_o})]^{-2}$
Poulsen-Mejlborg's Model (Poulsen 1993)	Fick's second law	$t_{ic} = t_{ex} X \left(\frac{0.5x_c}{\sqrt{t_{ex} D_{aex}}} \right)^{\frac{2}{1-\alpha}}$ $X \left(\left(\frac{1}{\operatorname{inv} \Lambda p (y_{cr})} \right)^{\frac{2}{1-\alpha}} \right)$
Yamamoto-Hosoya's Model (Yamamoto and Hosoya 1995)	Fick's second law	$t_{ic} = \frac{1}{D_c} \left[\frac{x_c}{2 \operatorname{erf}^{-1} (1 - C_{cr}/C_o)} \right]^2$
Anacta's Model (Anacta 2009)	Fick's second law	$t_{ic} = \frac{f_s}{D_c} \left[\frac{x_c}{2S} \right]^2$
Tang-Nilsson's Model (Tang and Nilsson 1992)	Numerical	$Q_{i,j}(\text{total}) = Q_{i,j-1}(\text{total}) + Q_{i,j}(\text{diff})$

The chloride diffusion and the associated process in concrete depend on the environmental condition that the structure is exposed to (Bester 2014). The previous (Equation 14) is more suitable for concrete structures when the depth of chloride diffusion is equivalent to the depth of concrete cover to predict when chloride ions reach the rebar surface. The calculated time, t , as shown in the following (Equation 16), also helps to determine the shape factor, S , and the diffusion coefficient, D_c , which are significant parameters in the computation of corrosion initiation time in concrete.

$$t_{ic} = \frac{f_s}{D_c} \left[\frac{x_c}{2S} \right]^2 \quad (16)$$

$$f_s = 0.316\xi \sqrt{t_{ic,ref}} \quad (17)$$

where:

t_{ic} is the corrosion initiation time (years)

$t_{ic, the ref}$ is the reference corrosion initiation time which will be obtained from experiments (days)

x_c is the thickness of concrete cover (mm)

f_s is the reinforcement factor

ξ is the curve-fitting parameter

From previous works of different authors, Equation 16 for the determination of corrosion initiation time in concrete is found as the best applicable equation due to the fact that

the performance of RC structures depends on environmental conditions. The big advantage of using this formula is the application of local data for validation, which considers the environmental factors, including temperature, humidity, and rainfall, as these are the significant factors for corrosion.

V. CONCLUSION

The required parameters for predicting chloride diffusion and the corrosion initiation time in concrete structures are all given in the critical review and can be determined experimentally. The chloride binding capacity inside the concrete and the chloride diffusion coefficient decreases over time. This implies that the equation that considers the influence of chloride binding and the chloride diffusion coefficient forecasts more accurate performance for structures than Fick's second law equation. Compared to other models, the equation for corrosion initiation time computation that considers the local environmental conditions and materials (Equation 16) is more appropriate. Therefore, the model formulated in Equation 16 is recommended to determine the corrosion initiation time of reinforcement bars with very high accuracy in concrete structures.

VI. DECLARATION OF COMPETING INTEREST

The authors declare that they have no known personal or financial conflicts of interest that could influence their work.

REFERENCES

- [1] Ababneh, A, F Benboudjema, and Y Xi, Chloride Penetration in Nonsaturated Concrete, *Journal of Materials in Civil Engineering*. 15(2) (2003) 183-191.
- [2] Alonso M, and M Sanchez, Analysis of the Variability of Chloride Threshold Values in the Literature, *Materials and Corrosion* (Wiley Online Library). 60(8) (2009) 631-637.
- [3] Ananta E, Modeling the Depth of Chloride Ingress and Time to Initiate Corrosion of RC Exposed to Marine Environment, Ph.D. Dissertation. (2009).
- [4] Andrade C, Rebar Corrosion Modeling and Deterioration Limit State, *Revista Alconpat*. 10(2) (2020) 165-179.
- [5] ASTM G193-12d, Standard Terminology and Acronyms Relating to Corrosion, West Conshohocken, PA: ASTM International. (2012).
- [6] Bazant P, Physical Model for Steel Corrosion in Concrete Sea Structures - Application, *Journal of Structural Division*. 105(6) (1979) 1155-1166.
- [7] Bentz P, A Virtual Rapid Chloride Permeability Test, *Cement and Concrete Composites*. 29(10) (2007) 723-731.
- [8] Bester N, Mechanisms and Modeling of Chloride Ingress in Concrete, *Durability and Condition Assessment of Concrete Structures*. (2014) 1-6.
- [9] Chen F, H Baji, and C Li, A Comparative Study on Factors Affecting Time to Cover Cracking as a Service Life Indicator, *Construction and Building Materials*. 163(2) (2018) 681-694.
- [10] Clear C, Time-to-Corrosion of Reinforcing Steel in Concrete Slabs, Washington, DC: Federal Highway Administration. (1976).
- [11] Crank J, The Mathematics of Diffusion, 2nd Edition, Oxford: Oxford University Press. (2004).
- [12] Dinh D, Initiation Time of Corrosion in Reinforced Concrete Structures Exposed to Chloride in Marine Environment, *International Journal of Civil Engineering and Technology*. (2017) 564-571.
- [13] Garboczi E, and D Bentz, Computer Simulation of the Diffusivity of Cement-Based Materials, *Journal of Materials Science*. 27(8) (1992) 2083-2092.
- [14] (2017). Graham, What is Concrete Cancer & How Can it be Prevented?. [Online]. Available: <https://freyssinet.co.uk/concrete-cancer-can-prevented-concrete-repairs/>.
- [15] Han H, Influence of Diffusion Coefficient on Chloride Ion Penetration of Concrete Structure, *Construction and Building Materials*. 21(2) (2007) 370-378.
- [16] Hunkeler F, Corrosion in Reinforced Concrete Structures. Cambridge: Woodhead Publishing Ltd. (2005).
- [17] Jaturapitakkul C.T Cheewaket, and W Chalee, Concrete Durability Presented by Acceptable Chloride Level and Chloride Diffusion Coefficient in Concrete: 10-Year Results in Marine Site, *Materials and Structures*. (2014) 1501-1511.
- [18] Koleva D, et al., Correlation of Microstructure, Electrical Properties and Electrochemical Phenomena in Reinforced Mortar, Breakdown to Multi-Phase Interphase Structures, Part I: Microstructural Observations and Electrical Properties, *Materials Characterization*. 59(3) (2008) 290-300.
- [19] Kristawan S, W Sunarmasto, B Gan, and S Nurrohmah, Estimating Initiation Period Due to Chloride Ingress into Reinforced Self-Compacting Concrete Incorporating High Volume Fly Ash, *ISCEE, MATEC Web of Conferences*. (2017).
- [20] (2011). Legault M, Composite vs. Corrosion: Battling for Marketshare. [Online]. Available: <https://www.compositesworld.com/articles/composite-vs-corrosion-battling-for-marketshare>.
- [21] Li, Y, J Xia, and S Lin, A Multi-Phase Model for Predicting the Effective Diffusion Coefficient of Chlorides in Concrete, *Construction and Building Materials*. 26(1) (2012) 295-301.
- [22] Liu, Q, Y Easterbrook, and D Li, Prediction of Chloride Diffusion Coefficients Using Multi-Phase Models, *Magazine of Concrete Research*. 69(3) (2017) 134-144.
- [23] Liu Q, Z Hu, X Lu, J Yang, I Azim, and W Sun, Prediction of Chloride Distribution for Offshore Concrete Based on Statistical Analysis, *Materials*. (2020) 1-6.
- [24] Luque M, E Arteaga, F Schoefs, and M Silva, Non-Destructive Methods for Measuring Chloride Ingress into Concrete: State-of-the-Art and Future Challenges, *Construction and Building Materials* (Elsevier). 68 (2014) 68-81.
- [25] Martin-Perez B, H Zibara, D Hooton, and A Thomas, A Study of the Effect of Chloride Binding on Service Life Predictions, *Cement and Concrete Research*. 30(8) (2000) 1215-1223.
- [26] Oh, H, and Y Jang, Prediction of Diffusivity of Concrete Based on Simple Analytic Equations, *Cement and Concrete Research*. 34(3) (2004) 463-480.
- [27] Pal, K, A Paulson, and D Rousseau, *Handbook of Biopolymers and Biodegradable Plastics*, Elsevier Inc. (2013).
- [28] Palanisamy G, *Corrosion Inhibitors*, London, SW7 2QJ: IntechOpen Limited. (2019).
- [29] (2019). Portland Cement Association. [Online]. Available: <https://www.cement.org/learn/concrete-technology/durability/corrosion-of-embedded-materials>.
- [30] Poulsen E, A Model of Chloride Ingress Into Concrete Having Time-Dependent Diffusion Coefficient, *Nordic Mini-Seminar, Sweden*. (1993) 298-309.
- [31] Que N, History and Development of Prediction Models of Time to Initiate Corrosion in Reinforced Concrete Structures in Marine Environment, *Philippine Engineering Journal*. 28(2) (2007) 29-44.
- [32] Tang C, and W Gan, The Analysis of Reinforcement Corrosion in Concrete Under the Non-Longitudinal Cracks in Marine Environment, *International Forum on Energy, Environment Science, and Materials (IFEESM)*. (2015) 63-69.
- [33] Tang L, L Nilsson, and P Basheer, *Resistance of Concrete to Chloride Ingress: Testing and Modeling*, CRC Press. (2011).
- [34] Tang P, and O Nilsson, Rapid Determination of Chloride Diffusivity in Concrete by Applying an Electrical Field, *ACI Materials Journal*. (1992) 49-53.
- [35] Tuutti K, Service Life of Structures with Regard to Corrosion of Embedded Steel, *Proceedings of the International Conference on Performance of Concrete in Marine Environment*. (1980) 223-236.
- [36] Valipour M, F Pargar, M Shekarchi, S Khani, and M Moradian, In Situ Study of Chloride Ingress in Concretes Containing Natural Zeolite, Metakaolin and Silica Fume Exposed to Various Exposure Conditions in a Harsh Marine Environment, *Construction and Building Materials*. (2013) 63-70.
- [37] Wang L, and T Ueda, Mesoscale Simulation of Chloride Diffusion in Concrete Considering the Binding Capacity and Concentration Dependence, *Computers and Concrete*. 8(3) (2011) 125-142.
- [38] Wang Y, X Gong, and L Wu, Prediction Model of Chloride Diffusion in Concrete Considering the Coupling Effects of Course Aggregate and Steel Reinforcement Exposed to Marine Tidal Environment, *Construction and Building Materials*. (2019).
- [39] Yamamoto K, and K Hosoya, Corrosivity of Bromine and Chloride on Duplex Stainless Steel, *Materials Science and Engineering*. (1995) 239-243.
- [40] Yu B, L Yang, M Wu, and B Li, Practical Model for Predicting Corrosion Rate of Steel Reinforcement in Concrete Structures, *Construction and Building Materials*. (2014) 385-401.
- [41] Yuan Q, C Shi, G De Schutter, K Audenaert, and D Deng, Chloride Binding of Cement-Based Materials Subjected to External Chloride Environment - A Review, *Construction and Building Materials*. 23(1) (2009) 1-13.
- [42] Zhang P, X Dong, and Y Jiang, Effect of Measurement Method and Cracking on Chloride Transport in Concrete, *Computers and Concrete*. 11(4) (2013) 305-316.
- [43] Zhang Y, M Zhang, and G Ye, Influence of Moisture Condition on Chloride Diffusion in Partially Saturated Ordinary Portland Cement Mortar, *Materials and Structures*. (2018).
- [44] Zhou Y, B Gencturk, K William, and A Attar, Carbonation-Induced and Chloride-Induced Corrosion in Reinforced Concrete Structures, *American Society of Civil Engineers*. (2014) 1-17.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/366464403>

Role of bioactive compounds in the treatment of hepatitis: A review

Article in *Frontiers in Pharmacology* · December 2022

DOI: 10.3389/fphar.2022.1051751

CITATIONS

0

READS

31

9 authors, including:



Arpita Roy

Sharda University

153 PUBLICATIONS 1,503 CITATIONS

[SEE PROFILE](#)



Shreeja Datta

Delhi Technological University

14 PUBLICATIONS 56 CITATIONS

[SEE PROFILE](#)



Thoraya A. Farghaly

Umm Al-Qura University

219 PUBLICATIONS 2,761 CITATIONS

[SEE PROFILE](#)



Magda H. Abdellattif

Taif University

137 PUBLICATIONS 670 CITATIONS

[SEE PROFILE](#)

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



Biosurfactant Production [View project](#)



Pyridone discovery as potent pharmacological compound [View project](#)



OPEN ACCESS

EDITED BY

Talha Bin Emran,
Begum Gulchemonara Trust University,
Bangladesh

REVIEWED BY

Mohsina Patwekar,
Luqman College of Pharmacy, India
Borehalli Mayegowda Shilpa,
Christ University, India
Syed Mahmood,
University of Malaya, Malaysia
Ashok K. Dubey,
Netaji Subhas University of Technology,
India
Carla Pereira,
Centro de Investigação de Montanha
(CIMO), Portugal
Flavia Tonelli,
Estácio de Sá University, Brazil

*CORRESPONDENCE

Arpita Roy,
arbt2014@gmail.com
Jesus Simal-Gandara,
jsimal@uvigo.es

SPECIALTY SECTION

This article was submitted to
Gastrointestinal and Hepatic
Pharmacology,
a section of the journal
Frontiers in Pharmacology

RECEIVED 23 September 2022

ACCEPTED 24 November 2022

PUBLISHED 21 December 2022

CITATION

Roy A, Roy M, Gacem A, Datta S,
Zeyaulah M, Muzammil K, Farghaly TA,
Abdellattif MH, Yadav KK and
Simal-Gandara J (2022), Role of
bioactive compounds in the treatment
of hepatitis: A review.
Front. Pharmacol. 13:1051751.
doi: 10.3389/fphar.2022.1051751

COPYRIGHT

© 2022 Roy, Roy, Gacem, Datta, Zeyaulah,
Muzammil, Farghaly, Abdellattif, Yadav and
Simal-Gandara. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that
the original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Role of bioactive compounds in the treatment of hepatitis: A review

Arpita Roy^{1*}, Madhura Roy², Amel Gacem³, Shreeja Datta⁴,
Md. Zeyaulah⁵, Khursheed Muzammil⁶, Thoraya A. Farghaly⁷,
Magda H. Abdellattif⁸, Krishna Kumar Yadav⁹ and
Jesus Simal-Gandara^{10*}

¹Department of Biotechnology, School of Engineering and Technology, Sharda University, Greater Noida, India, ²Centre for Translational and Clinical Research, School of Chemical and Life Sciences, Jamia Hamdard University, New Delhi, India, ³Department of Physics, Faculty of Sciences, University 20 Août 1955, Skikda, Algeria, ⁴Biotechnology Department, Delhi Technological University, Rohini, India, ⁵Department of Basic Medical Science, College of Applied Medical Sciences, Khamis Mushait Campus, King Khalid University, Abha, Saudi Arabia, ⁶Department of Public Health, College of Applied Medical Sciences, Khamis Mushait Campus, King Khalid University, Abha, Saudi Arabia, ⁷Department of Chemistry, Faculty of Applied Science, Umm Al-Qura University, Makkah, Saudi Arabia, ⁸Department of Chemistry, College of Science, Taif University, Taif, Saudi Arabia, ⁹Faculty of Science and Technology, Madhyanchal Professional University, Bhopal, India, ¹⁰Nutrition and Bromatology Group, Analytical and Food Chemistry Department, Faculty of Science, Universidade de Vigo, Ourense, Spain

Hepatitis causes liver infection leading to inflammation that is swelling of the liver. They are of various types and detrimental to human beings. Natural products have recently been used to develop antiviral drugs against severe viral infections like viral hepatitis. They are usually extracted from herbs or plants and animals. The naturally derived compounds have demonstrated significant antiviral effects against the hepatitis virus and they interfere with different stages of the life cycle of the virus, viral release, replication, and its host-specific interactions. Antiviral activities have been demonstrated by natural products such as phenylpropanoids, flavonoids, xanthenes, anthraquinones, terpenoids, alkaloids, aromatics, etc., against hepatitis B and hepatitis C viruses. The recent studies conducted to understand the viral hepatitis life cycle, more effective naturally derived drugs are being produced with a promising future for the treatment of the infection. This review emphasizes the current strategies for treating hepatitis, their shortcomings, the properties of natural products and their numerous types, clinical trials, and future prospects as potential drugs.

KEYWORDS

bioactive compounds, hepatitis, infection, treatment strategies, clinical trials

Abbreviations: HAV, Hepatitis A virus; HBV, Hepatitis B virus; HCV, Hepatitis C virus; HDV, Hepatitis D virus; HEV, Hepatitis E virus; ALT, Alanine transaminase; AST, Aspartate transaminase; IFN, Interferons; PEG-IFN- α , Pegylated interferon alfa; EMA, European medicines agency; DHCH, Dehydrocheilanthifoline; PHAP, p-hydroxy acetophenone; PFU, Plaque forming unit; CTL, Cytotoxic T-lymphocyte; EO, Essential oils; GTC, Green tea catechins; EGCG, Epigallocatechin 3-gallate; PBMC, Peripheral blood mononuclear cells.

Introduction

Hepatitis or as it is most commonly called, viral hepatitis is a severe/fatal disease. According to an estimated value, the complications of viral hepatitis approximately led to around 1–4 million deaths per year, throughout the world (Zarrin and Akhondi, 2021). Usually, various viruses could lead to inflammation of the liver like Epstein-Barr virus or Herpes simplex virus, or Cytomegalovirus, to name a few. But, all types of hepatitis viruses (A, B, C, D, and E), are causative agent references. Most hepatitis viruses lead to acute conditions and are self-limiting, but types B, C, and E can lead to chronic conditions. Chronic hepatitis can cause life-threatening conditions such as liver cirrhosis or hepatocellular carcinoma (González et al., 2017). Each year 1.5 million infections are reported due to Hepatitis-B and Hepatitis-C (Refer:<https://www.who.int/news-room/fact-sheets/detail/hepatitis-b>; <https://www.who.int/news-room/fact-sheets/detail/hepatitis-c>). Hepatitis viruses like A, C, D, and E are composed of RNA while the hepatitis-B virus is composed of DNA (Sinn et al., 2017; Shabanah et al., 2019; AlMalki et al., 2021). Hepatitis A virus (HAV) spreads through the contamination of food and water, in the feces of an individual. Hepatitis-B virus (HBV) exhibits vertical and distinctive transmission and can be transmitted through sexual passages (through secretions of vagina and semen), blood (through injections, drug abuse, etc) as well as by close human-to-human contact (MacLachlan and Cowie, 2015). Hepatitis C virus (HCV) spreads mainly *via* blood transfusions, but can also be transmitted *via* sexual contact, contaminated healthcare injections, and the use

of drugs (like intravenous). Hepatitis-D virus (HDV) is transmitted through sexual or blood contact, like in the case of HBV and HCV. It is dependent on HBV as it needs HBsAg (HBV surface antigen) for its replication (Rizzetto, 2015). Hepatitis-E virus (HEV) spreads *via* contamination of food and water, and also *via* zoonotic route and transfusion (Figure 1).

Natural products have been shown to be extremely useful for curative and prophylaxis as well as palliative treatment of myriad conditions caused by bacteria, fungi, and viruses (AlMalki et al., 2021; Alghamdi et al., 2021; Shahid et al., 2021). Novel antiviral drugs have been designed and developed using natural compounds and modified into useful compounds for preventive and curative actions (Lin et al., 2014). Antiviral drugs that are derived from natural products are usually extracted from medicinal microbes, plants and herbs, animals, and humans. The most economically beneficial option for treatment involves the use of medicinal plants with minimum side effects (Roy and Datta, 2021). It has been seen that herbal drugs showed lower complications in comparison to chemical drugs, leading to few or no after-effects. In the last decade, the use of herbal products has increased because of the problems with chemical drugs. Additionally, herbal remedies show better efficiency in the case of some diseases, and for some others, only herbal remedies are present (Rabiei et al., 2015). Additionally, the currently available drugs have higher toxicity levels which pose the requirement of replacing these drugs with medications having a lower toxicity range (Zeinab and Kopaei, 2018). As a result, this review highlights the effects of different types of natural products in treating hepatitis, involving the

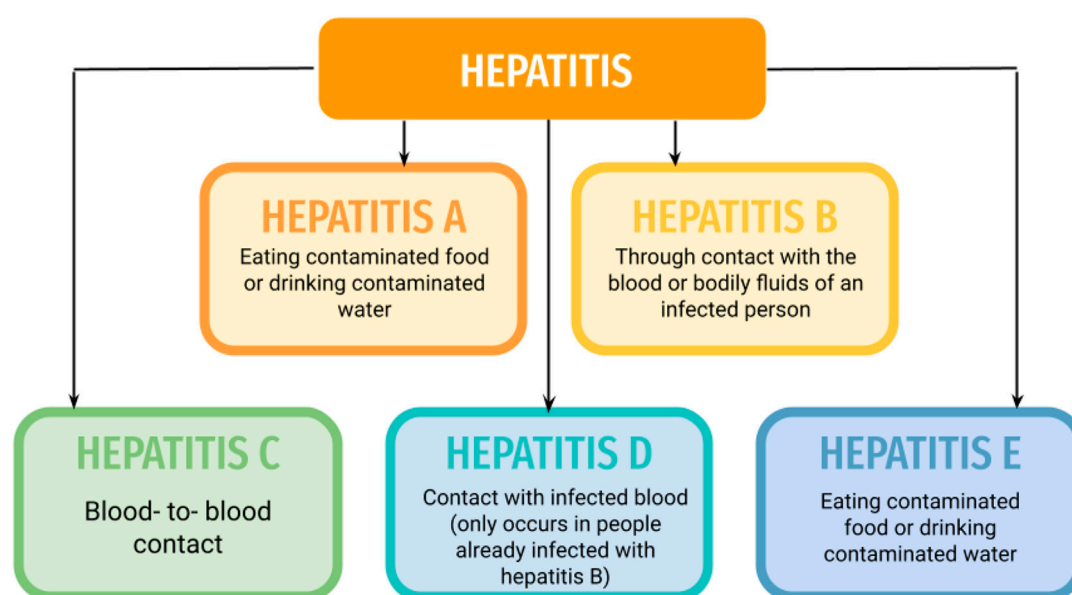


FIGURE 1
Transmission routes of Hepatitis types.

current treatment options and the properties of natural compounds, and various studies proving their properties in the management of the disease, along with the future perspectives and clinical trials of the same.

Hepatitis infection

HAV is a non-enveloped virus and belongs to the genus *Hepatitisvirus*. It can easily live outside the human body as well as in adverse conditions for a long time, which causes easy transmission of the virus and is the prime reason for past epidemics of jaundice. Its mode of transmission involves the fecal-oral route. HAV exhibits variable and nonspecific symptoms like jaundice, anorexia, fever, nausea, abdominal pain, fatigue, vomiting, fever, and nausea (Centre for Disease Control and Prevention, 2016). The severity of the HAV infection can be from asymptomatic to severe liver failure. Infants commonly suffer from asymptomatic viral infection, while adults experience symptoms such as jaundice.

HBV is a DNA (double-stranded) virus that belongs to the *Hepadnaviridae* family and is enveloped. The viral DNA is circular, partly double-stranded, and is the smallest known DNA virus. It exhibits replication in the hepatocytes and leads to liver dysfunction. The mode of transmission involves per mucosal or percutaneous route and blood transfusion, and also through vaginal fluid or semen to the uninfected person. The transmission of HBV can occur during childbirth (from mother to child), through sharing of needles, sexual intercourse, and transmission of blood through a mucosal surface or an open wound. HBV can lead to acute hepatitis with similar clinical manifestations as that of the Hepatitis A virus. But, the infection caused due to HBV is asymptomatic in 50% of the affected people (Centre for disease control and prevention, 2016). Less than 10% of the acute cases of jaundice prevail in the younger age group. HBV exhibits similar symptoms that HAV like nausea, fever, vomiting, abdominal pain, malaise, and flu-like symptoms (Centre for Disease Control and Prevention, 2016). The laboratory outcomes of HAV and HBV are identical as it demonstrates indistinguishable variations in the liver transaminases- ALT and AST are notably increased, during acute infection, sometimes greater than 1000 (Kim, 2009).

HCV is enveloped and transmitted through the blood of an infected person or through sexual transmission. It primarily affects the hepatocytes and if untreated/not treated properly, then it leads to liver cirrhosis in 20% of patients, which may cause hepatocellular carcinoma in later stages (Ringelhan et al., 2017). The immune systems (innate and adaptive) are evaded by HCV and chronic infections are caused in 70% of the patients (Ringelhan et al., 2017). Regular blood tests have significantly reduced transmission rates. Chronic HCV infection leads to severe conditions like liver cirrhosis, hepatocellular carcinoma, and fibrosis. Cryoglobulinemia vasculitis is one of the

extrahepatic manifestations that is developed in about 2/3rd of the patients (Davuluri and Bansal, 2021). HCV replication can significantly affect the metabolism, which leads to inflammation and steatosis of the liver.

Hepatitis delta virus (HDV) belongs to the genus *Deltavirus*. The extracellular virions of HDV possess the single-stranded genomic RNA, covalently and circularly closed in a negative sense. It is a satellite virus and regulates packaging, release and transmission, depending on HBV (Netter et al., 2021). Acute viral infection can be caused either due to co-infection (when both HBV and HDV infection occurs simultaneously due to the same exposure) or superinfection (occurrence of HDV infection after the HBV infection like in the case of HBsAg positive patients). The clinical manifestation of the simultaneous infection corresponds to an acute infection caused due to HBV. Moreover, concomitant infection causes a high risk of acute hepatic failure. In addition to that, co-infection occurs in a biphasic manner in which two peaks of ALT levels are observed within a span of a short time, as HBV infection should occur first in order to begin the spread of HDV infection. Chronic HDV infection leads to higher levels of transaminases, especially in patients who are infected due to HBV. Chronic HDV infection causes serious liver disease with higher fibrosis progression rates as compared to HCV or HBV-affected patients.

The HEV is an RNA virus (positive sense) that is single-stranded. It is transmitted when the drinking water is contaminated by faeces or the meat of infected animals is consumed (Doceul et al., 2016) or also through iatrogenic transmission. The infection caused by HEV is asymptomatic or causes minor symptoms without affecting the liver in most individuals. Acute icteric hepatitis is a classic example of HEV infection that lasts for more than 2–6 weeks in about 5%–30% of patients. It includes a prodromal phase that leads to common manifestations such as fever, vomiting, body pain, malaise, and nausea and continues for up to 1 week. Jaundice and dark-coloured urine are primary signs of the icteric phase. In the convalescent phase, jaundice and other symptoms related to it resolve within a week or a few days. At the beginning of the prodromal phase and the initial icteric phase, the serum levels of alanine aminotransferase are significantly increased, and throughout the icteric phase, the levels of bilirubin are markedly high. About 0.5%–4% of patients affected with Hepatitis-E infection, suffer from acute liver failure (Blasco-Perrin et al., 2015).

Current treatment strategies and their limitations

The treatment of HAV involves supportive management and prevention of the infection. Hepatitis virus infection can be effectively prevented by vaccination (Nelson et al., 2018;

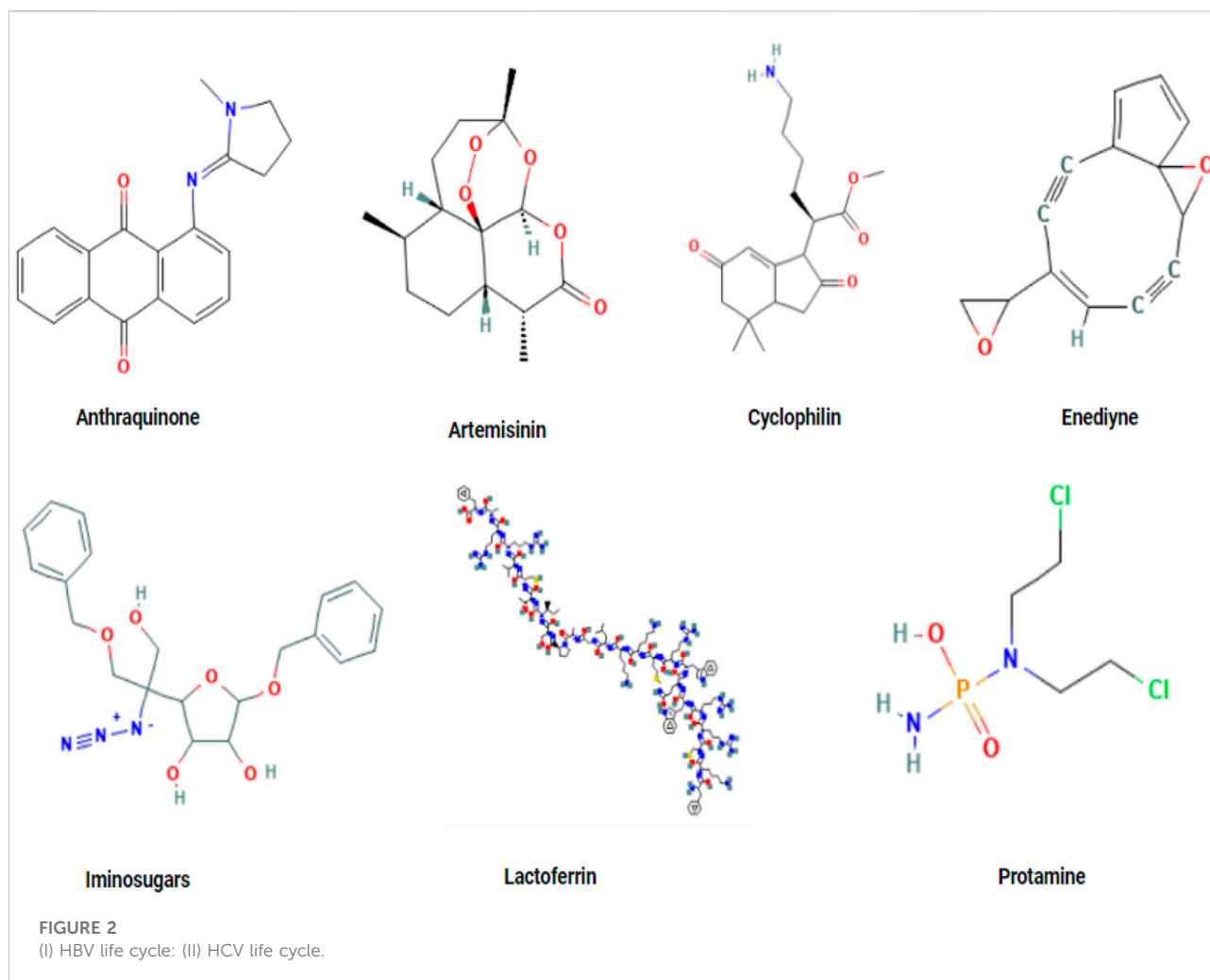
Centre for Disease Control and Prevention, 2020). At present, there are two single-antigen inactivated vaccines that are marketed in the United States- Vaxta, and Havrix. However, Vaxta demonstrates certain adverse effects like nausea, abdominal pain, appetite loss, diarrhea, joint pain, sore throat, etc (Refer: <https://www.rxlist.com/vaxta-drug.htm#description>).

Also, Havrix has also displayed some adverse events like complex regional pain syndrome, impaired work ability, oral discomfort, paresis, pelvic pain, etc. Twinrix is a single inactivated combination of both the vaccines Vaxta and Havrix, that is licensed for use (Centre for Disease Control and Prevention, 2019). However, it also has shown adverse effects like facial paresis, hypoesthesia facial, impaired driving ability, impaired work ability, monoparesis, paraparesis, paresis, pelvic pain, etc. Besides, nine approved drugs are available which can be used for treating chronic HBV, which include, two formulations of IFN (interferons)- one is conventional and the other is pegylated interferon, and there are 7 NAs (Nucleot(s) ide analogues)- tenofovir alafenamide fumarate, lamivudine, tenofovir disoproxil fumarate, telbivudine, entecavir, adefovir and besifovir dipivoxil (available only in Korea). The main objectives of therapy involve-the restriction of the advancement of the disease and upscale the rates of survival. PEG-IFN (pegylated-interferon) alpha along with ribavirin was conventionally used for 3–4 months for the therapeutic care of HCV infections (Ray et al., 2015). However, it showcased a broad spectrum of neuropsychiatric adverse effects like anorexia, depression, and sleep disturbances. PEG-IFN- α is also effective against HBV and HDV infection (Wedemeyer et al., 2011, Wedemeyer et al., 2019). But, its use can cause conditions like leukopenia and thrombocytopenia, which may lead to the discontinuation of the medication or modification in dose (Perrillo, 2009). Bulevirtide, which is an entry-inhibitor, has acquired conditional based approval from the European Medicines Agency (EMA) but the outcomes of the phase-3 trial are yet to be obtained. The phase-3 trials of Lonafarnib (prenylation inhibitor) are still ongoing. Additionally, other potential agents like RNA interference substances, nucleic acid polymers, and IFN-lambda are currently being examined. NA (Nucleot(s) ide analogues) exhibit inhibition of the reverse transcription of the HBV, but it does affect the replication of HDV. But, NA is effective in case of severe liver disease and against the viral DNA in case of co-infection with HBV and HDV (Lampertico et al., 2017). However, the disadvantages that are associated with nucleotide analogues (NAs) include the decreased rates of seroconversion of HBsAg and HBeAg and long-term therapy in most of the infected people (Zoulim and Durantel, 2015). Additionally, NAs do not affect the activity and level of cccDNA that exists in the liver of the infected person, even after the treatment with antivirals. Therefore, it takes time for NAs to exhibit their therapeutic actions and requires unspecified therapy.

Natural products

In recent times, natural products have been shown to be of great use in the therapy of hepatitis, with less drug resistance and adverse effects (Cai and Qin, 2019; Xu et al., 2019). Consequently, more studies are being conducted to comprehend the actions of natural products (Duan and Chen, 2016; Ahmed et al., 2017; Yao et al., 2019). Some reports have proven that numerous natural medicines having novel structures as well as anti-HBV properties might be good drug candidates for hepatitis B infection. Although such studies were primarily included in the recognition of products displaying anti-viral effects against HBV; the mechanism and targets of the products were fewer. The mechanism of action of therapeutic drugs like NAs as well as interferons, on anti-Hepatitis B Virus, is evident, but the unfolding of the drug-resistant mutants of HBV usually reduces the therapeutic activities. Therefore, the production of safe and efficient anti-HBV medications having unconventional mechanisms is the main target of today's research (Li Y. T. et al, 2012; Cai and Qin, 2019). Some of the different kinds of natural compounds having anti-HBV properties are flavonoids, phenylpropanoids, alkaloids, glycosides, terpenes, lactones as well as organic acids. Because of the wide popularity of natural products for treating and preventing diseases, recently, pharmaceutical companies have been developing new antimicrobial formulations which are derived from such compounds. For instance, phytotherapy or use of medicinal plants is practised all over the world, specifically in developed nations like some of the European nations as well as the United States (Solati et al., 2017). Around 45% of the marketed natural products that are utilized for the management of infections caused by the hepatitis virus are obtained from medicinal extracts of plants or their derivatives (Lahlou, 2013). In addition to this, a global upsurge in the isolation of active compounds from medicinal plants had emerged in health care.

The hunt for new bioactive compounds is still prevalent in prime therapeutic areas like immunosuppression, metabolic and infectious diseases, as well as oncology. It has been an enormously studied part of pharmaceutical research for many years (Newman and Cragg, 2012). Around 40 new drugs have been launched between the time period of 2000- 2010, derived from plants, microorganisms, marine organisms, and a few chordates (Brahmachari, 2011). Furthermore, the WHO estimated that around 80% of the global population depends on conventional medications, mostly derived from plants, for their primary healthcare. The active compounds originating from herbs or plants are utilized either for therapeutic treatments or are administered through the oral route to infected patients in the form of powders, teas as well as other



herbal formulations. Further, even phenolic products cause the bioactivity of the unrefined extracts of plants. In recent decades, researchers have attempted to recognize the bioactive compounds of these traditional medicines by systematically screening the natural products obtained from the extracts of herbs or plants and then testing their effectiveness using appropriate assays (on the basis of the studied pathology). One of the primary benefits of natural compounds extracted from plants is the lower manufacturing costs, due to the absence of the requirement for chemical synthesis. Such production leads to lesser expensive treatments and is available for low-income populations. Apart from this, different natural compounds have proven to exhibit antiviral effects against the influenza virus, HIV, herpes simplex virus, influenza virus as well as HBV and HCV. Further, the screening and development of natural compounds have resulted in the detection of effective inhibitors that inhibits the growth of the virus. However, there are various limitations as well

associated with natural products. For instance, the extraction process of natural products from organisms is a cumbersome task. Along with this, the mode of extraction process is dependent on the type of compound to be extracted. To increase the varieties of the extracted natural products, the bioactive compound can be extracted in the presence of several solvents of varying polarity. Apart from this, another limitation is to identify novel natural products as some of the potential source organisms are difficult to be produced or culture in the laboratories as they only survive in their ecosystem. Such challenges are now being tackled by establishing novel methodologies for culturing for natural product synthesis induction. Therefore, naturally derived compounds of different origins have been shown to be of therapeutic use for hepatitis infection. The detailed studies of numerous natural products have been explained in the following sections, with specific emphasis on their role in inhibiting viral infection.

List of natural products

Natural products have the potential treat various diseases one of them is hepatitis. Various natural compounds have been studied for effective results against hepatitis (Figure 2).

Alkaloids

Alkaloids are natural compounds that possess a complex ring and nitrogen heterocyclic structure, which is responsible for most of their physiological effects. They also exhibit anti-microbial, anti-inflammatory, anti-cancer, and antioxidant activities. Jiang et al., 2013, reported that ethanol extract which was derived from the fruit of *Piper longum* L. fruit exhibited effective antiviral activity, and its specific derivatives demonstrated significant activities against the production of HBeAg and HBsAg on the cell line of HepG 2.2.15. This study also demonstrated that one of such compounds demonstrated the inhibition of the production of HBeAg and HBsAg at IC₅₀ values of 0.21 and 1.80 mM. One more experiment by Zeng et al., 2013 exhibited that one of the alkaloids, namely DHCH, extracted from the *Corydalis saxicola* plant, caused effective inhibition of HBeAg and HBsAg production in HepG2.2.15 cells, along with TI of around 6.77 and 7.32. Furthermore, DHCH was shown to decrease DNA and cccDNA levels in time and dose-dependent ways, at IC₅₀ values of 7.62, 8.25, and 15.08 μM.

Anthraquinones

Anthraquinones are derived from the metabolites of fungi and lichens. They exhibit purgation, immunoregulatory, anti-cancer, and anti-inflammatory actions (Wang et al., 2019). Recent studies demonstrated the anti-HBV effect of anthraquinones (Bu et al., 2019). Sulochrin, (–)-2' R-1-hydroxyisorhodoptilometrin, questinol, monochlorsulochrin, endo crocin, dihydrocodeine, astringic acid and (+)-2' S-isorho-doptilometrin are extracted from the aciduric fungus *Penicillium* sp. Which is mangrove-derived, significantly inhibits the secretion of HBsAg than 3TC (positive control) at a specific dosage (Qin et al., 2016). Discovered that hypericin effectively decreased the viral DNA expression and also the HBeAg and HBsAg expression, like lamivudine (3 TC).

Another study by Peng et al. (1970) reported that anthraquinones, anthraquinone bile acid conjugates, and rubiadin exhibited activities against HBV infections on the cell line of HepG2.2.15 at IC₅₀ of 12.41, 8.03, 17.05, and 8.13 g/ml. Rubiadin is a compound that not only effectively decreases the production level of HBsAg and HBeAg, inhibits the HBV DNA replication, as well as prevents the activities of the HBx protein and the growth of cells in a dose-dependent method, but may also result in an unconventional candidate of the anti-HBV agent.

Parvez et al. (2019), first time demonstrated the anti-HBV property of anthraquinones which are AV-derived, probably through inhibition of HBV-DNA polymerase.

Aromatics

The six phenols namely, ethyl 2,5-dihydroxybenzoate, m-hydroxybenzoic acid, ethyl 3,4-dihydroxy-benzoate, p-hydroxybenzoic acid, 3,4-dihydroxybenzoic acid, and m-hydroxy benzenmethanol. Aromatics demonstrate anti-microbial, anti-pyretic, analgesic, and anti-inflammatory properties. They demonstrated anti-Hepatitis B virus activities by causing inhibitory actions against the production of HBsAg as well as HBeAg at IC₅₀ levels of 0.23–5.18 mM, and replication of HBV-DNA at IC₅₀ of 0.06–2.62 mM. In addition to this, p-hydroxybenzoic acid, m-hydroxybenzoic acid, and m-hydroxy benzenmethanol exhibited anti-HBV actions at IC₅₀ levels of 5.18, 3.76, and 4.55 mM against the secretion of HBsAg and 2.54, 2.36, and 2.62 mM for the inhibition replication of HBV-DNA (Cao et al., 2015). Zhou et al. (2014) demonstrated that compounds extracted from the plant *Tarphochlamys affinis* (Griff.) significantly caused the inhibition of the production of HBeAg and HBsAg. Huang et al. (2014) exhibited that PHAP (p-hydroxy acetophenone) extracted from the plant *A. morrisonensis*, effectively inhibited the replication of HBV-DNA. Zhao et al. (2015) discovered that PHAP and its derivatives exhibited activities against the viral DNA. A sequence of derivatives was obtained after the structural changes of p-HAP and its derivatives, amidst them, p-HAP derivative 2f demonstrated the most effective inhibition of the HBV-DNA replication (SI = 160:3, IC₅₀ = 5:8 μM). The relationships between the primary structure and its activity indicated that the substituted cinnamic acids and the conjugated derivatives of p-HAP glycoside increased their actions against the replication of HBV-DNA.

Artemisinin

Artemisinin is a plant-based product that is extracted from *Artemisia annua* and is a widely known antimalarial agent. It has been reported that Artemisinin also exhibits anti-HBV activities. Demonstrated that Artesunate, which is the semisynthetic derivative of Artemisinin demonstrated more effective activities by decreasing the amount of HBV-DNA at IC₅₀ of 0.5 μmol/L and led to the inhibition of the production of HBsAg at IC₅₀ value of 2.3 μmol/L. The values were not observed to be better compared to Lamivudine (IC₅₀ of 0.3 μmol/L and 0.3 μmol/L), although a combination of both compounds produced a significantly effective result. As Lamivudine was subjected to drug resistance, thus, this combination effectively reduced the emergence of drug

resistance against Lamivudine. Also, Artemisinin and artesunate do not cause serious side effects, regarding their anti-HBV effects.

2-Arylbenzofuran derivatives

They are obtained from MCR or Mori cortex radices and exhibited anti-HCV effects in a system of HCV replicon (Lee et al., 2007). They also demonstrate anti-oxidative properties. The assay of NS3 helicase revealed that the two compounds demonstrated effective inhibitory effects (IC_{50} of 42.9 and 27.0 $\mu\text{mol/L}$). The NS3 viral helicase unwinds the duplexes of RNA \times DNA as well as RNA \times RNA, hence, is significant for viral replication. Therefore, targeting the NS3 helicase enzyme is ideally considered, and the by-products of this enzyme can be used for the development of effective inhibitors of helicase in the future (Lee et al., 2007). Dai et al. (2001) demonstrated that mellenin, a fungus-based compound, extracted from *Aspergillus ochraceus*, demonstrated anti-Hepatitis C virus protease activity with 35 $\mu\text{mol/L}$, as an IC_{50} value. Another study by Hu et al. (2007) included the screening of numerous compounds of pseudo guaianolides derived from *Parthenium hispidum*. The results revealed antiviral effects in a HCV subgenomic replicon system as three out of all the compounds caused inhibition up to around 90%, at a concentration of 2 $\mu\text{mol/L}$. Also, the other compounds demonstrated inhibition up to 50% without any cytotoxicity, thus, indicating further research for their higher potentials.

Blueberry proanthocyanidins

According to the USDA database, every 100 g of the edible portion of blueberries contains about 88–261 mg of proanthocyanidin. They are structurally similar to polyphenols like flavonoids and anthocyanins (Huang et al., 2012). Proanthocyanidins exhibit anti-oxidative, anti-tumor, and anti-inflammatory effects (Yang et al., 2014). Blueberry as well as polyphenols demonstrate some important biological properties as antibacterial, neuroprotective, antiviral, anticarcinogenic, and cardioprotective agent (Joshi et al., 2016). Additionally, Takeshita et al. (2009) demonstrated that the fraction of methanol extract from blueberry leaves (0.112–2200 lg/ml) led to the reduction in the activities of subgenomic HCV in a replicon cell system of HCV after 72 h at a temperature of 37°C. Joshi et al. (2016) reported that blueberry juice and its proanthocyanidins (B type) exhibited anti-viral effects against HAV and are also effective against human norovirus. The study also showed that the HAV titers in the suspension decreased significantly to undetectable levels by proanthocyanidins at concentrations of 2 as well as 5 mg/ml in half an hour and by 1 mg/ml proanthocyanidins after 3 h.

Furthermore, it was observed that within 24 h, blueberry juice (37°C and pH 2.8) significantly reduced HAV level (2 log PFU/ml). The blueberry juice and isolated proanthocyanidins were evaluated for their activity against HAV adsorption and replication in FRhK4 cells, after being pre- and post-infected with the HAV HM175 strain. The blueberry juice and isolated proanthocyanidins were significantly able to decrease the HAV level in the pre-infected cells, however, in the post-infected cells, there was no inhibition of the viral replication (Joshi et al., 2016).

Cyclophilins

Cyclosporin A (CsA) exhibits immunosuppressive activity. CsA is a significantly effective anti-HCV compound based on the screening of cell cultures for anti-HCV products. It is a fungus-based compound that is produced by *Tolypocladium inflatum* Gams and exhibited significant biological applications in organ transplantation and immunosuppressive effects (Doutre, 2002). CyPB (cyclophilin PB), stimulates the anti-viral effects of CsA against HCV (Watashi and Shimotohno, 2007) and is primarily found in ER membrane's cytoplasm. It is exactly similar to the NS5B polymerase of HCV. Additionally, both of these compounds establish complexes with HCV-RNA. Watashi et al. (2005) demonstrated that cyclophilin PB enhanced the RNA binding via NS5B, and the reduced levels of CyPB resulted in the lack of viral replication (Watashi et al., 2005). Further, the complex of NS5B and CyPB was disrupted by CsA, as a result of which, the replication of the HCV genome was reduced. As CsA exhibited immunosuppressive effects, therefore, it is not advisable for the treatment of HCV infections. However, NIM811 (another derivative having one substituted amino acid), demonstrates two-fold effective binding affinity to CyPB and lacks immunosuppressive activity. Five out of eight patients, who'd undergone liver transplantation and also suffered from HCV recurrence, did not respond to the standard HCV therapy, but the effective results of cyclosporin gave hope due to the reduced HCV-RNA below the level of detection (Watashi and Shimotohno, 2007).

Ellagic acid

Ellagic acid, is a flavonoid product obtained from *Phyllanthus urinaria* and it demonstrates anti-oxidative, anti-inflammatory, and neuroprotective activities. It also demonstrated an interesting anti-HBV effect. Shin et al. (2005) exhibited that it hindered the secretion of HBeAg in cell culture with 0.07 $\mu\text{g/mL}$ as an IC_{50} value, but it did not cause the inhibition of production of HBsAg, polymerase activity, and HBV replication. Another report by Kang et al. (2006) illustrated an experiment on the HBeAg-producing transgenic mice and showed that they exhibited significant effective tolerance towards HBeAg. As a

result, there were reduced levels of CTL (cytotoxic T-lymphocyte) responses, minimal levels of cytokines production, and no secretion of antibodies to the antigen. But, when the mice were fed with ellagic acid, there was an inhibition of this immune tolerance and thus it was concluded that ellagic acid is considered an effective agent to overcome this essential mechanism against the chronic infection of HBV.

Enediynes

Enediynes derived from the plant *A. capillaris* (Yin-Chen) is used as a therapeutic agent for hepatitis predominantly in China (Liu et al., 2019). They also demonstrate anti-tumor actions. Subsequently, Geng et al. (2015) exhibited that two glucopyranoside derivatives of enediynes caused effective inhibition of HBV DNA replication and the production of HBsAg along with HBeAg. These compounds exhibited inhibitory actions against the HBV-DNA replication with SI values equal to 17.1 and 23.6, as well as IC_{50} levels of 0.0127 ± 0.05 and 0.077 ± 0.04 mM. Also, a specific pair of isomers of enediynes demonstrated inhibitory activities of the production of HBsAg at IC_{50} of 0.887 ± 0.20 mM (SI = 2.3) and 0.797 ± 0.23 mM (SI = 2.1). This study even showed that one of the compounds exhibited most effective inhibitory actions against the replication of the viral DNA with SI = 23.6 and IC_{50} of 0.077 ± 0.04 mM, while the similar derivative with $-(2'-O\text{-caffeoyl})$, exhibited slightly reduced actions against HBV-DNA replication with SI = 17.1 and IC_{50} level of 0.127 ± 0.05 mM. Another study by Geng et al., 2018, extracted fourteen compounds from *A. capillaris* which were essayed for their structure-activity relationship, and their anti-HBV properties were summed up on the basis of their biological actions. Out of these, two of the compounds effectively caused the inhibitory actions against the productions of HBeAg, HBsAg and HBV DNA replication at IC_{50} values of 48.7 (SI > 20.5), 197.2 (SI > 5: 1), and 9.8 (SI > 102) μ M.

Essential oils

Essential oils (EO) are plant-based aromatic oils that are extracted from roots, grass, fruit, branches, flowers, bark, leaves, buds, wood, and seeds. Some EOs derived from sweet orange, rosemary cineole, lemon, and grapefruit (common names of *Citrus sinensis*, *Rosmarinus officinalis*, *Citrus limon*, and *Citrus paradisi*, respectively) exhibit anti-HAV effects (Battistini et al., 2019). Essential oils like sesquiterpenes, hydrocarbons, and limonene, which are produced by the genus *Citrus*, are 85%–99% volatile, with their oxygenated agents like esters, ketones, aldehydes (citral), alcohols (linalool) and acids (Fisher and Phillips, 2008). EOs exhibit antimicrobial, anti-cancer, anti-fungal, and anti-spasmodic activities. Battistini et al. (2019)

conducted an experiment in which the Frp3 cells were inoculated with ATCC or HM-175 hepatitis A strain and then treated with rosemary cineole essential oil, after 30 min of incubation at room temperature. It was observed that the rosemary cineole essential oil led to an effective decrease of cell infectivity followed by lemon and grapefruit essential oils. Apart from this, it is essential to estimate the least time taken by essential oils to cause maximum effectiveness against the viral loads in food products in order to make people more aware about food safety (Battistini et al., 2019).

Flavonoids

Flavonoids are plant-based products that exhibit numerous clinical functions, it acts as an anti-bacterial anticancer, and anti-inflammatory. It demonstrates a major function in the protection of the liver, like, Silymarin is an effective medicine that is developed for protecting the liver (Yi, 2012). Wang et al. (2013) exhibited that flavonoids are effective against HBV, wherein Luteolin caused the inhibition of the production HBsAg as well as HBeAg on the cells of HepG 2.2.15 *in vitro*, at IC_{50} value of 0.02 mM. Another plant-based flavonoid, Isovitexin, derived from *S. yunnanensis* also exhibited effective anti-HBV properties. Cao et al. (2013a) showed how it has not just prevented the HBsAg and HBeAg secretion (at IC_{50} levels of 0.04, less than 0.03, and 0.23 mM), and also inhibited the replication of HBV- DNA (at IC_{50} of 0.09, less than 0.01 as well as 0.05 mM). Cao et al. (2015) reported that Isoorientin (derived from *S. musotie*) possessed anti-HBV functions in opposition to the production of HBeAg and HBsAg at IC_{50} of 1.12 and 0.79 mM, and the viral replication at IC_{50} of 0.02 mM. Another study by Xiao. (2018) proved the anti-HBV effects demonstrated by Epimedium Hyde II (a Chinese herbal compound). It also prevented the HBV-DNA replication and the activities of HBeAg and HBsAg in the HBV-replicated serum of the C57BL/6 mice. Additionally, various researchers discovered that isopongachromene and glabaarachalcone, which are plant-based compounds and are extracted from *P. pinnata* can be linked to the viral DNA polymerase protein target (Mathayan et al., 2019).

Ginsenosides

Ginseng which is also known as *Panax ginseng* Meyer is popularly used in Korea and China as a medicinal herb for over 5000 years (Yun, 2001). They contain numerous bioactive compounds like peptides, polysaccharides, ginsenosides, fatty acids, phytosterols, poly acetylenic acids, and polyacetylenes. There are various studies on the biological activities of Ginseng like they are used as an antifungal, anti-stress, anti-inflammatory, anti-bacterial, anti-carcinogenic, antiviral, and

anti-oxidant agent (Lee et al., 2011). Its accumulation is primarily reported in its plant's roots and traditionally, isolation of the same usually takes an extended period of time (Luthra et al., 2021). Lee DY. et al (2013) exhibited that the Korean red ginseng (or KRG), as well as purified ginsenosides (Rg1 and Rb1), were used at different concentrations for the pre-treatment and co-treatment on FRhK-4 cells, after the inoculation of Hepatitis-A virus ATCC strain on the cell line. The outcomes demonstrated that both of the above-mentioned compounds effectively reduced the HAV levels. This research also exhibited that the KRG compound exhibited cytotoxicity exceeding 10 µg/ml concentration, but ginsenosides did not demonstrate any cytotoxicity up to the concentration of 40 µg/ml. Furthermore, even though KRG and the purified ginsenosides were used for the co-treatment of the cell lines, they effectively decreased the viral concentration and the pretreated cells exhibited significant anti-HAV effects. Therefore, pretreatment with ginseng significantly prevents HAV infection.

Green tea catechins

Green tea catechins (GTCs) are natural and herbal compounds that are highly beneficial to human health. As the name suggests, they are the components of *Camellia sinensis*, of the family *Theaceae*. One of the most studied and abundant catechins is EGCG or (–)-epigallocatechin 3-gallate (EGCG), while others include EC or epicatechin, (–)-epigallocatechin (EGC) and (–)-epicatechin gallate. They comprise anti-cancerous, anti-oxidative, anti-infectious, and anti-inflammatory functions, according to *in vitro* as well as *in vivo* examinations (Cao et al., 2016). Ye et al. (2009) analyzed the anti-HIV actions of tea catechin mixtures on the production of Hepatitis-B-antigen and the production of DNA in a stable cell line of HBV-transfected HepG2. Other catechins and EGCG inhibits the production of the HBeAg as well as HBV DNA at a specific dosage with IC₅₀ of 7.34 µg/ml and 2.54 µg/ml. He et al. (2011) showed that when HepG2.117 cells were made to grow when EGCG, HBeAg expression was suppressed, undisturbing the HBsAg expression. Chen C. et al (2012) proved the reduction of HCV infection *via* EGCG in the cells of Huh7.5.1 cells, with the help of the JFH1 strain of hepatitis C genotype 2a which produced the contagious viruses in the cell culture. This treatment for HCV with EGCG at EC₅₀ was around 17.9 µM.

Iminosugars

The potential ER (or endoplasmic reticulum) inhibitors α-glucosidases are basically the by-products of DNJ or deoxynojirimycin iminosugars. Such iminosugars are naturally present, for instance in silkworms or *Bombyx mori* (Jacob et al., 2007). They are potent immunomodulators and demonstrate

anti-microbial properties. Iminosugars significantly reduce the virality of infectious viral particles. This suppression might be sourced by the incorporated envelope proteins (misfolded) in the secreted particles (Chapel et al., 2007). The derivatives of iminosugars have been proven to show antiviral effects against infections caused due to HBV as well as HCV. A product namely, 1-DNJ, extracted from the plant of *Morus alba* L, suppressed the hepatitis B viral particle secretion in a dose-dependent way (Jacob et al., 2007). Steinmann et al. (2007) reported that compounds having a long alkyl side chain significant for inhibitory effects on p7 (an ion channel of HCV) as well as a DNJ head group, lead to an advantage for susceptibility to resistance. Misumi et al. (2021) showed that Miglustat was used for treating lipid storage illnesses in humans, as well as UV-4 inhibited the replication of HAV in cell culture, at IC₅₀ 32.13 µM as well as 8.05 µM, respectively *via* blockage of ganglioside synthesis (crucial for the HAV cell entry).

Japanese rice-koji miso extracts

Koji (or *Aspergillus oryzae*) has been predominantly used by the Japanese for the fermentation of various food items like rice, soybean, grains, and potatoes. Miso is obtained as a by-product when Japanese rice is fermented by Koji. It is used as a seasoning in the preparation of Miso soup (Win et al., 2018). It exhibits antioxidant and anti-aging effects (Lee MH. et al., 2013). Jiang et al. (2011) demonstrated that Miso enhanced the effects of GRP78- also known as glucose-controlled protein 78, which is basically a heat-shock protein that resulted in the suppression of Ultraviolet C mutagenesis. Some researchers also demonstrated that the expression of GRP78 retarded the HAV replication (Win et al., 2018). Therefore, GRP78 acts against HAV infection as an effective antiviral agent (Jiang et al., 2017). Another report showed how Win et al. (2018) conducted a post-infection assay demonstrating that the Miso extract synergistically functioned as an antiviral against HAV infection by partially enhancing the effect of GRP78. It even stimulated the effects of GRP78 in PXB cells and Huh 7 (human hepatocytes) and suppressed the replication of HAV (Win et al., 2018). As a result, the Miso extract has been used as an important dietary product for controlling HAV infection.

Lactoferrin

Lactoferrin is derived from cattle as well as camel milk and has been reported as a combinational therapy along with conventional hepatitis C drugs. It also demonstrates antimicrobial, immunomodulatory, and, anti-cancer activities. Many trials have confirmed the effective functions of camel milk-derived lactoferrin, regarding its therapeutic value against hepatitis (Adinolfi et al., 2001; Redwan and Tabll, 2007). It

was mentioned in a study that lactoferrin derived from camel milk inhibited hepatitis C genotype 4 *via* the prevention of virus from the entering cells (Gader and Alhaider, 2016). The compound also showed antiviral, antifungal as well as antiparasitic activities, toward a wide spectrum of species (Jenssen and Hancock, 2009). EL-Fakharany et al. (2013), demonstrated significant activities of lactoferrin against hepatitis virus, isolated from camel milk was reported on the PBMCs or peripheral blood mononuclear white blood cells as well as HepG2 or human hepatoma infected cells having HCV. Redwan et al. (2014) reported the activities of lactoferrin on the Huh-7 cell line in a cell culture medium that was inoculated with HCV and further noted the dismantling of viral peptides and inhibition of the virus's growth.

Lignans

They usually extracted from plants. One of the primary groups of phytoestrogen, they play a role as antioxidants (Wohlfarth and Efferth, 2009). For instance, Honokiol is a lignan isolated from leaves, barks, and cones of *Magnolia officinalis*. Lan et al. (2012) assessed the effects of honokiol HCV infection; its entry, replications as well as translation, in the cell line Huh-7 using HCVcc, HCVpp as well as subgenomic replicons. The results showed that it strongly reduced the HCVcc infection (at EC₅₀ of around 1.2 µg/ml, with respect to 4.5 µM, and EC₉₀ of 6.5 µg/ml) at non-toxicity (median lethal dose = 35 µg/ml). Wu et al. (2012) proved the anti-HCV effects of another lignan namely, 3-hydroxy caruillignan C (or 3-HCL-C) which was isolated from the plant *Swietenia macrophylla* (stems). The results also included that 3-HCL-C decreased NS3 proteins and the levels of HCV-RNA at EC₅₀ of around 10.5 µg/ml (37.5 µM). Apart from this, 3-HCL-C is hindered with the replication of HCV as well, *via* induction of IFN-induced response element transcription as well as IFN-dependent anti-viral gene expression. Hence, 3-HCL-C is a potent adjuvant for the therapy of HCV.

4-Phenylcoumarin derivatives

Coumarin is a plant-based natural product that was first derived from *Dipteryx odoranta* and tonka beans. It is also known as Coumarou and there are numerous natural coumarins that are isolated from plants, fungi, bacteria, and chemical synthesis (Kassem et al., 2019). Coumarin along with its derivatives is used to synthesize antiviral agents (Garro and Pungitore, 2015). (2H-chromen-2-ones) are known to be superior bioactive agents for the synthesis of novel agents which possess high specificity and affinity to numerous molecular targets (Batan et al., 2018). Coumarin derivatives exhibit antioxidant, anti-inflammatory, neuroprotective and anti-cancer effects. Recently, it was

reported that various derivatives of Coumarin exhibit anti-HAV activities (Kassem et al., 2019). Like picornaviruses, the Hepatitis-A virus genome encodes HAV 3C pro, also known as HAV three protease (an essential processing protease that is responsible to enhance viral replication by transcription, translation, and nucleo-cytoplasmic trafficking). 4-Phenylcoumarin-based compounds, which are recently modified antiviral compounds, target the 3C proteases and inhibit them (Kassem et al., 2019). There are various derivatives that demonstrate anti-HAV activity, which has been reported to exhibit the strongest virucidal activity and also inhibit the adsorption and replication of HAV, therefore, it possesses effective virustatic properties.

Phenylalanine dipeptides

Dipeptide derivatives exhibit anti-inflammatory and antimalarial effects. A study by Yang et al. (2014) was conducted which extracted and altered the Matijun-Su (phenylalanine dipeptide) with anti-HBV actions from *Dichondra repens*. Forst, as well as four by-products, were tested with effective anti-HBV properties *in vitro*.

Another report by Meng et al. (2018) examined the twenty species of natural marine small molecules *via* the cells of HepG 2.2.15; three types of agents namely, 4-hydroxy proline-phenylalanine, glycine-L-proline, and L-2-hydroxy proline-phenylalanine demonstrated effects against the HBV by the hindrance of HBV-DNA, HBeAg and HBsAg. Wang et al. (2019) showed that N-acetyl phenylalanine demonstrated inhibitory effects on HBsAg as well as HBeAg at IC₅₀ of 55.5 and 69.5 µg/ml, respectively. Kuang et al., 2019, utilized Matijun-Su (MTS) as a primary compound; and synthesized a novel derivative of MTS which demonstrated anti-HBV effects. Further, a series of derivatives of MTS were synthesized with Matijun-su as the primary compound, *via* incorporating chlorine or fluorine substitution, as well as the acquired derivatives of MTS, and were evaluated for anti-HBV actions *in vitro*. These outcomes demonstrate that the extracted compounds exhibited anti-HBV effects at IC₅₀ values of 10.53, 12.61 12.61 mol/L.

Phenylpropanoids

Phenylpropanoids are usually derived by plants from tyrosine and phenylalanine amino acids and comprise a broad spectrum of biological activities like antioxidation, antitumor, liver protection as well as antiviral. Chen H. et al. (2012) demonstrated that the extraction of a sequence of phenylpropanoids from the roots or core materials or bark of *S. asper* caused anti-HBV effects. Compounds like Magnatriol B displayed mild anti-HBV activity and inhibited both HBsAg as well as HBeAg secretions with lower cytotoxicities. Honokiol, on

the other hand, showed strong inhibition on HBeAg as well as HBsAg with IC_{50} of 4.74 μ M (SI = 14:22) and 3.14 μ M (SI = 21:47), respectively. Isomagnolol and isocarpine, isolated from the roots or bark of the plant *S. asper* displayed prominent anti-HBV effects via the cell assay of HepG 2.2.15 and reduced the HBsAg production at IC_{50} values 10.34 μ M as well as 3.67 μ M. Also, for inhibition of HBeAg production, the IC_{50} was around 8.83 μ M as well as 14.67 μ M, at non-toxic levels. Another compound, Coumarin lignan, was derived from the plant of *Kadsura heteroclita* (stems), and caused inhibition of HBsAg as well as HBeAg production at a concentration of 25 μ g/ml. Niranthin, derived from *Phyllanthus niruri*, also suppressed the HBsAg as well as HBeAg secretion at specific dosages, with IC_{50} of 16.5 μ M as well as 25.1 μ M, respectively.

Polyphenols

Polyphenols demonstrate antioxidant, neuroprotective and anti-inflammatory effects. Many polyphenols exhibit anti-HCV activities. Nobiletin or 3',4',5,6,7,8-hexamethoxyflavone derived from the extract of *Citrus unshiu* was responsible for demonstrating anti-HCV effects (Suzuki et al., 2005). Hegde et al. (2003) proved that nobiletin showed activity against hepatitis C infection at 10 μ g/ml in the MOLT-4 cells. Another study involved the isolation and characterization of two novel compounds-oligophenolic in nature, namely SCH 644343 as well as SCH 644342 from the plant of *Stylogne cauliflora* wherein they demonstrated inhibitory actions against HCV NS3 protease activity *in vitro*, with IC_{50} values of 0.3 μ M as well as 0.8 μ M (Hegde et al., 2003). Duan et al. (2004) recognized three polyphenol compounds from ethyl acetate fraction of Galla Chinese which is traditional Chinese medicine and showed that they inhibited NS3 protease *in vitro* at IC_{50} values of 0.75, 1.60, and 1.89 μ M.

Zuo et al. (2005) reported one of the compounds inhibited the NS3 protease, derived from *Saxifraga melanocentra* Franch. The results showed that the IC_{50} value was 0.68 μ M as well as the compound was safe till 6 mg/ml (or 6.4 mM) on the COS cells. Li Y. et al. (2012) characterized four polyphenolic compounds isolated from *Excoecaria agallocha* L. which inhibited NS3 protease *in vitro*. Out of these, two compounds namely, excoecariphenol D as well as corilagin displayed a prominent inhibitory effect in the replicon assay, with IC_{50} of 12.6 as well as 13.5 μ M.

Protamine, taxifolin and atropine

Taxifolin which is also known as Dihydroquercetin is a plant based product that is obtained from onions, grapes, citrus fruits, and olive oil. It takes a major part in the prevention of Alzheimer's disease and was known for its effective

pharmacological actions, which included, anti-diabetic, antitumor, antioxidative, hepatoprotective, cardioprotective as well as neuroprotective effects.

Protamine is an animal-based compound that is derived from fish milt. It is a cationic peptide and possesses numerous properties. It was used in the form of an antibacterial agent in food items and apart from that, it was used as a heparin antagonist and as an injectable-insulin carrier (Gill et al., 2006). Atropine is also a plant-based product that is derived from *Belladonna*. It is an anticholinergic agent (muscarinic receptor antagonist) which is administered to regulate the contractions as well as dilations of muscles in order to maintain the blood flow in cells (Behcet, 2014). Earlier, a study was conducted that demonstrated the significant inhibitory actions of Taxifolin, Protamine, and Atropine against the replication of HAV in the cells of PLC/PRF/5. Atropine resulted in a dose-dependent decrease in HAV infectivity. Taxifolin, Protamine, and Atropine, at a maximum concentration of 59, 50 as well as 50 μ g/ml, respectively, decreased the HAV titer.

Resveratrol

Resveratrol, which is also known as 3,5,40-trihydroxystilbene, is a naturally derived phytoalexin. It is commonly found in plants like grapes, cranberries, peanuts, etc. It exhibits numerous biological activities as it is administered as a vasoprotective, chemopreventive, anti-inflammatory, and antioxidant compound (Ungvari et al., 2007). It has been studied that Resveratrol causes inhibition of liver steatosis (ethanol-induced) in rats (Kasdallah-Grissa et al., 2006). Bujanda et al. (2008) reported that in rats with fatty liver infection (non-alcoholic) and found that it inhibited *de novo* lipogenesis of adipocytes, adipogenic differentiation, and reduced hepatic steatosis. Another report by Jiang et al. (2012) showed that Resveratrol exhibited effective activity against HCV core protein-stimulated hepatic steatosis by enhancing the PPAR- α levels, which was inhibited through the HCV core protein, *in vivo* and *in vitro*.

Silverstrol

It is extracted from *Aglaia foveolata* (Kim et al., 2007). Silverstrol exhibits anti-leukemia effects. It causes effective inhibition of the DEAD-box RNA helicase eIF4A (Bordeleau et al., 2008), which is a member of the eIF4F complex that is responsible for the cap-dependent initiation of eukaryotic translation (Silvera et al., 2010). Todt et al. (2018) exhibited that silvestrol has inhibitory actions against the HEV genotypes replication at a specific dosage. Zhou et al. (2015) examined the actions of the eIF4F complex in the HEV replication as well as

reported that with respect to the silverstrol's activity, effective HEV replication required the machinery eIF4A, eIF4G as well as eIF4E. Further, the study also reported that programmed cell death 4 (or PDCD4) as well as eIF4E-binding protein 1 (or 4E-BP1)- the negative regulatory factors with respect to the complex, displayed anti-HEV effects, thus proving the necessity of both eIF4A and eIF4E in the replication of HEV. As a result, silverstrol targeted these mRNA translation host factors for its antiviral effects. Another study by Glitscher et al. (2018) identified silverstrol as an efficient inhibitor of the viral particle release of HEV. The results showed a highly decreased HEV capsid protein translation as well as control of the viral RNA inside the cytoplasm, in the absence of any prime cytotoxic effects.

Terpenoids

Terpenoids are natural compounds with a basic structural unit as isoprene. They exhibit effective biological actions, which mainly comprise antivirals and anti-inflammatory actions (Zhang et al., 2018). Li L. Q. et al. (2012) extracted ursolic acid from the core of *S. asper*. Ursolic acid possesses effective anti-HBV properties by inhibition of HBeAg as well as HBsAg secretion, at IC₅₀ 97.61 and 89.91 μ M. Another triterpenoid, MH, is a plant-based compound and is extracted from *Vicia tenuifolia* Roth and it demonstrated inhibitory action on the production of HBeAg as well as HBsAg at a specific dosage (Huang et al., 2013). Zhou et al. (2013), isolated heptane terpenoids from the rhizomes and roots of the plant *Aster tataricus*- Andepishionol and Astartaricusones B, which caused inhibition of the production of HBeAg, at IC₅₀ values of 18.6, 40.5 μ M. It also inhibited the replication of HBV-DNA at IC₅₀ of 2.7, 30.7 μ M. Another report by Liu and Wu. (2013) proved that Diosgenin significantly led to the suppression of HBsAg along with HBeAg secretion at the rate of inhibition of 40%–50%. Zhao et al. (2014) showed that Pumila Side A and 7-Eudesma-4 (15)-ene-1 β ,6 α -diol, extracted from *Artemisia capillaris*, demonstrated effective actions against the replication of HBV-DNA at IC₅₀ of 19.70, 12.01 μ M (high SI values equivalent to 105.5, 139.2). Moreover, Pumila Side A suppressed the production of HBsAg as well as HBeAg at IC₅₀ of 15.02 μ M (SI of 111:3) and 9.00 μ M (SI of 185:9). Swertia Side is a plant-based product which not only revealed the most effective pursuit against replication of HBV-DNA at IC₅₀ of 0.05 mM (SI of 29:1), but also acted against the production of HBsAg (IC₅₀ equivalent to 0:79 mM) (Jie et al., 2015). In addition to this, Laurifolioside and Genkwanin, extracted from *Wikstroemia chamaedaphne* Meisn, demonstrated significant anti-HBV actions at IC₅₀ of 46.5, 88.3 mg/ml. Further, 2-epi-laurifolioside, Wikstroelide W, laurifolioside, 2-epi-laurifolioside A, laurifolioside B, and 2-epi-laurifolioside B exhibited inhibitory activities against the replication viral DNA to a certain extent, within the range of

levels from 0.39–6.25 mg/ml and the ratios of inhibition ranging between 2.0% and 33.0% (Zhang et al., 2017).

Wogonin

Wogonin is a plant-based compound that is extracted from *Scutellaria radix* and is a known mono-flavonoid. This herb exhibited biological activities against hepatitis and inflammatory diseases and has been administered in Asia for over a thousand years (Wohlfarth and Efferth, 2009). Huang et al. (2000) exhibited antiviral activities against HBV of this compound by illustrating its activities against the HBsAg secretion in the cell suspension. Guo et al. (2007) noted that it led to the suppression of HBsAg and HBeAg secretions at IC₅₀ equivalent to 4 μ g/ml. Furthermore, the study showed that Hepatitis-B virus DNA was decreased at a dose-dependent way. The results were tested on a bunch of ducks infected with DHBV (or duck-hepatitis B virus), proving the decreased rates of DHBV-DNA and plasma HBsAg and thus an improvement in the functions of the liver after the histopathological evaluations. Other results were seen on the human HBV-transgenic mouse livers treated with Wogonin, which led to the reduction of HBsAg after the immunological staining.

Xanthones

Xanthones demonstrate anti-inflammatory, anti-cancer, and antioxidant activities (Shagufta., 2016). There are various types/derivatives of Xanthones which exhibited effective inhibition of the replication of HBV-DNA between IC₅₀ from 0.01 to 0.13 mM (Cao et al., 2013b). This report also noted that the compounds which possessed hydroxy groups (3 or more) like 1,5,8-trihydroxy- 3-methylxanthine, Norbellidifolin as well as 2-C- β -D-glucopyranosyl-1,3,7-trihydroxy xanthone demonstrated effective inhibitory action against HBV at IC₅₀ of <0.62, 0.35, and 0.04 mM for HBeAg as well as 0.77, >0.98, and 0.21 mM for HBsAg. Apart from this, Norswertianolin, Neolancerin, and 1,8-Dihydroxy-3,5-dimethoxy xanthone were derived from *S. yunnanensis* and demonstrated effective anti-HBV activity. Amongst the three compounds, Neolancerin not only prevented the HBsAg as well as HBeAg secretions at IC₅₀ of 0.21, 0.10, and 1.51 but also hindered the viral DNA replication at IC₅₀ of 0.09 mM, less than 0.01 mM, and 0.05 mM (Cao et al., 2013a). Cao et al. (2015) showed that 1,5,8-Trihydroxy-3-methylxanthine exhibited the inhibition of replication of the viral DNA at IC₅₀ of 0.09 as well as 0.05 mmol-L-1 (SI = 10.89) and demonstrated a significant action against the HBeAg secretions at IC₅₀ of 0.35 (SI of \geq 2.80). Qin et al. (2016) reported that a natural compound extracted from *Penicillium*

TABLE 1 Natural products for hepatitis A treatment.

Natural product	Mechanism of action	Concentration	Result of action	Reference
Blueberry Proanthocyanidins	Interruption of binding of HAV and its entry into the cell	2–5 mg/ml at a temperature of 37°C for half-hour	Decrease in HAV levels to undetectable measure in the medium	Joshi et al. (2016)
Essential Oil (or EO) derived from rosemary cineole, lemon and grapefruit	-	0.05% of rosemary cineole EO, 0.5% of lemon EO, and 0.1% of grapefruit EO	Decrease in cell infectivity (rosemary cineole > grapefruit > lemon)	Battistini et al. (2019)
Grape seed extract	Interruption of HAV binding to cell membrane receptors; prevention of adsorption	2 mg/ml at 37°C for 6 h	Suppression of HAV titer to undetectable levels	Joshi et al. (2015)
Green tea essence/extract	Binds to them and interferes with the viral attachment with the cell membrane receptors	5 mg/ml at 37°C for 2 h with pH 7.2	Total HAV inactivation in the suspension culture	Steinmann et al. (2013)
Japanese rice-koji miso extracts	Reduction in replication of HAV <i>via</i> improvement of GRP78 levels in human hepatocytes	-	Reduction in HAV replication	Win et al. (2018)
KRG or Korean Red Ginseng Extract and Ginsenosides	Activating the RNaseL pathway and boosting the cytokine production	5–10 µg/ml at a temperature of 37°C for 24 h	Suppression of HAV in a dose-dependent manner	Lee et al. (2013a)
4-phenylcoumarin derivatives	Interruption of adsorption of HAV on cell surface	10 µl at a temperature 37°C	Inhibition of HAV's 3C protease activity	Kassem et al. (2019)

TABLE 2 Natural products for hepatitis B treatment.

Natural compound	Plant source	Target	Concentration	Result of action	Reference
Aloe-emodin	<i>Aloe vera</i>	CYP3A4	10 µg/ml	Reduction in HBeAg expression in the cells of HepG2.2.15	Parvez et al. (2019)
Aloin B	<i>Aloe vera</i>	HBV-DNA polymerase	50 µg/ml	Suppression of HBV antigen	Parvez et al. (2019)
Diterpenoids	<i>W. chamaedaphne</i>	HBsAg	0.016 µg/ml	Exhibited strong anti-HBV activity	Zhang et al. (2017)
Esculetin	<i>M. fortunei</i>	HBV DNA, HBsAg and HBeAg	-	Inhibition of antigens of HBV as well as HBV DNA expression <i>in vitro</i> along with HBV replication	Huang et al. (2019)
Glabaarachalcone	<i>P. pinnata</i>	HBV-DNA	5 mg/ml	Inhibition of virus binding	Mathayan et al. (2019)
Hypericin	-	HBV DNA, HBsAg and HBeAg	-	Notable inhibition of HBV DNA, HBsAg, and HBeAg	-
Rosmarinic acid	-	ε-Pol binding	-	Inhibition of HBV replication	Tsukamoto et al. (2018)
Rubiadin	<i>P. connata</i>	HBeAg as well as HBsAg	20 µg/ml for 10 min	Reduction of HepG2.2.15 cells and reduction of HBx expression levels	Peng et al. (1970)

sp. Effectively reduced the HBsAg secretions than 3TC (positive control) at a dose-dependent way. [Tables 1, 2, 3](#).

Future perspectives

In recent times, many studies have explored the potential of natural compounds as an emerging treatment option for hepatitis. Even though vaccines have been developed for mitigating the spread of viral infection for a considerable time period, there is an

urgent requirement for developing efficient anti-hepatitis adjunctive, therapeutic as well as prophylactic agents. Moreover, naturally derived compounds with beneficial potential against hepatitis have been investigated; some of which have even displayed prominent potential for control of hepatitis. Furthermore, the studies should be focused on imitating the characteristics of the hepatitis virus *in vivo* rather than *in vitro* in order to properly demonstrate the foundation for the applications of such natural compounds in a clinical setup. There is a requirement for the development of more

TABLE 3 Natural products for Hepatitis C treatment.

Natural compounds	Source	Target	IC ₅₀ (<i>in vitro</i>)	Result of action	Reference
EGCG	<i>Camellia sinensis</i>	HCV virion	5–21 μ M	Inhibition of cell-free virus transmission, as well as HCV cell-to-cell, spread, leading to undetectable infection levels	Calland et al. (2012)
Honokiol	<i>Magnolia grandiflora</i>	NS5B polymerase, NS3 protease and NS5A	4.5 μ M	Active against HCVcc infection at non-toxic concentration; reduction of HCV replication at a dose-dependent way in both 1b and 2a subgenomic replicons	Lan et al. (2012)
3-hydroxy caruillignan C	<i>S. macrophylla</i>	HCV RNA and NS3 protease	37.5 μ M	Expressed anti-viral activities against HCV at both protein as well as RNA levels (in safe concentrations)	Wu et al. (2012)
Ladanein-BJ486K	<i>M. peregrinum</i>	HCV RNA	2.5–10 μ M	Efficient against all major HCV genotypes and inhibited HCV infection	Haid et al. (2012)
Luteolin–Apigenin	-	NS5B polymerase	1.1–7.9 μ M	Inhibition of HCV activity and against the NS5B enzyme	Liu et al. (2012)
Naringenin	Grapefruit	HCV life cycle	109 μ M	Non-toxic inhibition of HCV	Goldwasser et al. (2011)
Quercetin	<i>Embelia ribes</i>	NS3 protease	-	Inhibitory effect on HCV NS3 protease, reduction of viral production <i>via</i> suppression of NS3 required for HCV replication	Bachmetov et al. (2012)
Silymarin or Silibinin	<i>Silybum marianum</i>	NS5B polymerase	40–100 μ M	Inhibition of HCV replicon as well as JFH1 replication in cell culture, suppression of HCV RNA-dependent RNA polymerase activity	Ahmed-Belkacem et al. (2010)

suitable animal models which would display more or less the same clinical demonstrations as seen in humans suffering from hepatitis, for more precise elucidation as well as the correlation of results from pre-clinical trials including natural agents therapy. Also, more drug combinations with different targets, may have better therapeutic intervention. As a result, natural products should be included in the development of newer therapies, since they show promising and potent effects and may hence, replace the current standard and aggressive therapies.

Conclusion

In recent times, because of the after-effects plus the appearance of medicinal residues in chemical drugs in hepatitis treatment, more attention is being paid to naturally derived medicines. As a result, it is extremely important to study as well as address efficient natural compounds for curing hepatitis and its types, as also their action mechanisms. These compounds are a viable source for the synthesis of novel drugs for the treatment of hepatitis. They range in different bioactivities which can be directly developed or administered as initial points for novel drug optimization. In addition to this, clinical trials have shown that bioactive products have the potential to treat hepatitis, mostly HBV and HCV infections. Therefore, this review can provide a strong foundation for bioactive compounds to be used for the treatment of hepatitis.

Author contributions

AR, SD, and MR conceptualized, designed and written the initial manuscript draft the manuscript, SD, MR, AR, GA, MZ, KM, TF, MA, KY, and JS-G prepared the figures and tables, edited and revised the manuscript critically. Final manuscript has been approved by all the authors. The authors declare no conflict of interest.

Funding

The authors extend their appreciation to the Deanship of Scientific Research at King Khalid University, KSA, for funding this work through a research group program under grant number RGP. 2/181/43. This work was supported by the Deanship of scientific research at Umm Al-Qura University for supporting this work by grant code 22UQU4350477DSR06. Funding by Universidade de Vigo/CISUG.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated

References

- Adinolfi, L. E., Gambardella, M., Andreana, A., Utili, R., and RuGGiero, G. (2001). Steatosis accelerates the progression of liver damage of chronic hepatitis C patients and correlates with specific HCV genotype and visceral obesity. *Hepatology* 33, 1358–1364. doi:10.1053/jhep.2001.24432
- Ahmed, H., Arbab, M. K. P., Mohammed, S. A. D., and Adnan, J. A. R. (2017). *In vitro* evaluation of novel antiviral activities of 60 medicinal plants extracts against Hepatitis B virus. *Exp. Ther. Med.* 14, 626–634. doi:10.3892/etm.2017.4530
- Ahmed-Belkacem, A., Ahnou, N., Barbotte, L., Wychowski, C., Pallier, C., Brillet, R., et al. (2010). Silibinin and related compounds are direct inhibitors of hepatitis C virus RNA-dependent RNA polymerase. *Gastroenterology* 138 (3), 1112–1122. doi:10.1053/j.gastro.2009.11.053
- Alghamdi, M., Alotaibi, F., Ahmed, H., Alharbi, F., Bukhari, O., and Youssef, A. (2021). Effect of medical education on the knowledge, attitude and compliance regarding infection control measures among dental students in Makkah. *J. UMM AL-QURA Univ.* 7, 14–17. doi:10.3390/medsci9010014
- AlMalki, W. H., Shahid, I., Abdalla, A. N., Johargy, A. K., Ahmed, M., and Hassan, S. (2021). Virological surveillance, molecular phylogeny, and evolutionary dynamics of hepatitis C virus subtypes 1a and 4a isolates in patients from Saudi Arabia. *Saudi J. Biol. Sci.* 28, 1664–1677.
- Bachmetov, L., Gal-Tanamy, M., Shapira, A., Vorobeychik, M., Giterman-Galam, T., Sathiyamoorthy, P., et al. (2012). Suppression of hepatitis C virus by the flavonoid quercetin is mediated by inhibition of NS3 protease activity. *J. Viral Hepat.* 19 (2), e81–e88. doi:10.1111/j.1365-2893.2011.01507.x
- Batran, R. Z., Kassem, A. F., Abbas, E. M. H., Elseginy, S. A., and Mounier, M. M. (2018). Design, synthesis and molecular modeling of new 4-phenylcoumarin derivatives astubulin polymerization inhibitors targeting MCF-7 breast cancer cells. *Bioorg. Med. Chem.* 26, 3474–3490. doi:10.1016/j.bmc.2018.05.022
- Battistini, R., Rossini, I., Ercolini, C., Gorla, M., Callipo, M. R., Maurella, C., et al. (2019). Antiviral activity of essential oils against hepatitis A virus in soft fruits. *Food Environ. Virol.* 11 (1), 90–95. doi:10.1007/s12560-019-09367-3
- Behçet, A. (2014). The source-synthesis-history and use of atropine. *J. Acad. Emerg. Med.* 13, 2–3. doi:10.5152/jaem.2014.1120141
- Blasco-Perrin, H., Madden, R. G., Stanley, A., Crossan, C., Hunter, J. G., Vine, L., et al. (2015). Hepatitis E virus in patients with decompensated chronic liver disease: A prospective UK/French study. *Aliment. Pharmacol. Ther.* 42, 574–581. doi:10.1111/apt.13309
- Bordeleau, M. E., Robert, F., Gerard, B., Lindqvist, L., Chen, S. M., Wendel, H. G., et al. (2008). Therapeutic suppression of translation initiation modulates chemosensitivity in a mouse lymphoma model. *J. Clin. Invest.* 118 (7), 2651–2660. doi:10.1172/JCI34753
- Brahmachari, G. (2011). "Natural products in drug discovery: Impacts and opportunities—an assessment," in *Bioactive natural products*. Editor G. Brahmachari (Singapore: World Scientific Publishing Company), 1–199.
- Bu, Z. L., Yu, C. M., Lin, W. Y., Hong, P. Z., and Li, Y. (2019). Research progress on the synthesis of anthraquinones. *Chin. J. Synthetic Chem.* 9, 747–762.
- Bujanda, L., Hijona, E., Larzabal, M., Beraza, M., Aldazabal, P., Garcia-Urkia, N., et al. (2008). Resveratrol inhibits nonalcoholic fatty liver disease in rats. *BMC Gastroenterol.* 8, 40. doi:10.1186/1471-230X-8-40
- Cai, M. Z., and Qin, G. (2019). Research advances in anti-Hepatitis B virus drugs. *Clin. Gastroenterology Hepatology* 35 (10), 2302–2307. doi:10.3390/ph14050417
- Calland, N., Albecka, A., Belouzard, S., Wychowski, C., Duverlie, G., Descamps, V., et al. (2012). (–)-Epigallocatechin-3-gallate is a new inhibitor of hepatitis C virus entry. *Hepatology* 55, 720–729. doi:10.1002/hep.24803
- Cao, J., Han, J., Xiao, H., Qiao, J., and Han, M. (2016). Effect of tea polyphenol compounds on anticancer drugs in terms of anti-tumor activity, toxicology, and pharmacokinetics. *Nutrients* 8, E762. doi:10.3390/nu8120762
- Cao, T. W., Geng, C. A., Jiang, F. Q., Ma, Y. B., He, K., Zhou, N. J., et al. (2013a). Chemical constituents of *Swertia yunnanensis* and their anti-Hepatitis B virus activity. *Fitoterapia* 89, 175–182. doi:10.1016/j.fitote.2013.05.023
- Cao, T. W., Geng, C. A., Ma, Y. B., He, K., Wang, H. L., Zhou, N. J., et al. (2013b). Xanthones with anti-Hepatitis B virus activity from *Swertia mussotii*. *Planta Med.* 79 (8), 697–700. doi:10.1055/s-0032-1328399
- Cao, T. W., Geng, C. A., Ma, Y. B., He, K., Zhou, N. J., Zhou, J., et al. (2015). Chemical constituents of *Swertia delavayi* and their anti-Hepatitis B virus activity. *China J. Chin. Materia Medica* 40 (5), 897–902.
- Centers for Disease Control and Prevention (2016). Hepatitis A FAQs for health professionals. Available at: <http://www.cdc.gov/hepatitis/HAV/HAVfaq.htm> (Accessed o December 1, 2016).
- Centers for Disease Control and Prevention (2019). Hepatitis A questions and answers for health professionals. Available At: <https://www.cdc.gov/hepatitis/hav/havfaq.htm> (Accessed December 19).
- Centers for Disease Control and Prevention (2020). Immunization schedules: Recommended adult immunization schedule for ages 19 and older, United States, Available At: www.hrsa.gov/vaccinecompensation.
- Chapel, C., Garcia, C., Bartosch, B., Roingeard, P., Zitzmann, N., Cosset, F. L., et al. (2007). Reduction of the infectivity of hepatitis C virus pseudoparticles by incorporation of misfolded glycoproteins induced by glucosidase inhibitors. *J. Gen. Virol.* 88, 1133–1143. doi:10.1099/vir.0.82465-0
- Chen, C., Qiu, H., Gong, J., Liu, Q., Xiao, H., Chen, X. W., et al. (2012a). (–)-Epigallocatechin-3-gallate inhibits the replication cycle of hepatitis C virus. *Arch. Virol.* 157, 1301–1312. doi:10.1007/s00705-012-1304-0
- Chen, H., Li, J., Wu, Q., Niu, X. T., Tang, M. T., Guan, X. L., et al. (2012b). Anti-HBV activities of *Streblus asper* and constituents of its roots. *Fitoterapia* 83 (4), 643–649. doi:10.1016/j.fitote.2012.01.009
- Dai, J., Carté, B. K., Sidebottom, P. J., Sek Yew, A. L., Ng, S., Huang, Y., et al. (2001). Circumdatin G, a new alkaloid from the fungus *Aspergillus ochraceus*. *J. Nat. Prod.* 64, 125–126. doi:10.1021/np000381u
- Davuluri, S., and Bansal, P. (2021). "Cryoglobulinemic vasculitis," in *StatPearls* (Treasure Island (FL): StatPearls Publishing). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK556045/>.
- Doceul, V., Bagdassarian, E., Demange, A., and Pavio, N. (2016). Zoonotic hepatitis E virus: Classification, animal reservoirs and transmission routes. *Viruses* 8, 270. doi:10.3390/v8100270
- Doutre, M. S. (2002). Ciclosporin. *Ann. Dermatol. Venereol.* 129, 392–404.
- Duan, D., Li, Z., Luo, H., Zhang, W., Chen, L., and Xu, X. (2004). Antiviral compounds from traditional Chinese medicines *Galla* Chinese as inhibitors of HCV NS3 protease. *Bioorg. Med. Chem. Lett.* 14, 6041–6044. doi:10.1016/j.bmcl.2004.09.067
- Duan, Z. H., and Chen, X. M. (2016). Research progress on the active constituents of Chinese traditional medicine for anti HBV. *J. Liaoning Univ.* 18 (11), 112–115.
- El-Fakharany, E. M., Sanchez, L., Al-Mehdar, H. A., and Redwan, E. M. (2013). Effectiveness of human, camel, bovine and sheep lactoferrin on the hepatitis C virus cellular infectivity: Comparison study. *Virol. J.* 10, 199. doi:10.1186/1743-422X-10-199
- Fisher, K., and Phillips, C. (2008). Potential antimicrobial uses of essential oils in food: Is citrus the answer? *Trends Food Sci. Technol.* 19 (3), 156–164. doi:10.1016/j.tifs.2007.11.006
- Gader, A. G., and Alhaider, A. A. (2016). The unique medicinal properties of camel products: A review of the scientific evidence. *J. Taibah Univ. Med. Sci.* 11, 98–103. doi:10.1016/j.jtumed.2015.12.007
- Garro, H. A., and Pungitore, C. R. (2015). Coumarins as potential inhibitors of DNA polymerases and reverse transcriptases. Searching new antiretroviral and antitumoral drugs. *Curr. Drug Discov. Technol.* 12, 66–79. doi:10.2174/1570163812666150716111719
- Geng, C.-A., Huang, X.-Y., Chen, X.-L., Ma, Y. B., Rong, G. Q., Zhao, Y., et al. (2015). Three new anti-HBV active constituents from the traditional Chinese herb of Yin-Chen (*Artemisia scoparia*). *J. Ethnopharmacol.* 176 (12), 109–117. doi:10.1016/j.jep.2015.10.032
- Geng, C. A., Yang, T. H., Huang, X. Y., Yang, J. I., Ma, Y. B., Li, T. Z., et al. (2018). Anti-Hepatitis B virus effects of the traditional Chinese herb *Artemisia capillaris* and its active enynes. *J. Ethnopharmacol.* 10, 283–289. doi:10.1016/j.jep.2018.06.005
- Gill, T. A., Singer, D. S., and Thompson, J. W. (2006). Purification and analysis of protamine. *Process Biochem.* 41, 1875–1882. doi:10.1016/j.procbio.2006.04.001

- Glitscher, M., Himmelsbach, K., Woytinek, K., and Johne, R. (2018). Inhibition of hepatitis E virus spread by the natural compound silvestrol. *Viruses* 10, 301. doi:10.3390/v10060301
- Goldwasser, J., Cohen, P. Y., Lin, W., Kitsberg, D., Balaguer, P., Polyak, S. J., et al. (2011). Naringenin inhibits the assembly and long-term production of infectious hepatitis C virus particles through a PPAR-mediated mechanism. *J. Hepatol.* 55 (5), 963–971. doi:10.1016/j.jhep.2011.02.011
- González, M. E., González, V. M., Montaña, M. F., Medina, G. E., Mahadevan, P., Villa, C., et al. (2017). Genome-wide association analysis of body conformation traits in Mexican Holstein cattle using a mix of sampled and imputed SNP genotypes. *Genet. Mol. Res.* 16 (2), 16. doi:10.4238/gmr16029597
- Guo, Q., Zhao, L., You, Q., Yang, Y., Gu, H., Song, G., et al. (2007). Anti-Hepatitis B virus activity of wogonin *in vitro* and *in vivo*. *Antivir. Res.* 74, 16–24. doi:10.1016/j.antiviral.2007.01.002
- Haid, S., Novodomska, A., Gentzsch, J., Grethe, C., Geuenich, S., Bankwitz, D., et al. (2012). A plant-derived flavonoid inhibits entry of all HCV genotypes into human hepatocytes. *Gastroenterology* 143 (1), 213–222. e5. doi:10.1053/j.gastro.2012.03.036
- He, W., Li, L. X., Liao, Q. J., Liu, C. L., and Chen, X. L. (2011). Epigallocatechin gallate inhibits HBV DNA synthesis in a viral replication inducible cell line. *World J. Gastroenterol.* 17, 1507–1514. doi:10.3748/wjg.v17.i11.1507
- Hegde, V. R., Pu, H., Patel, M., Das, P. R., Butkiewicz, N., Arreaza, G., et al. (2003). Two antiviral compounds from the plant *Stylogne cauliflora* as inhibitors of HCV NS3 protease. *Bioorg. Med. Chem. Lett.* 13, 2925–2928. doi:10.1016/S0960-894X(03)00584-5
- Hu, J. F., Patel, R., Li, B., Garo, E., Hough, G. W., Goering, M. G., et al. (2007). Anti-HCV bioactivity of pseudoguaianolides from *Parthenium hispidum*. *J. Nat. Prod.* 70, 604–607. doi:10.1021/np060567e
- Huang, H. L., Chang, C. G., Chen, C. F., and Chang, C. (2000). Anti-Hepatitis B virus effects of wogonin isolated from *Scutellaria baicalensis*. *Planta Med.* 66, 694–698. doi:10.1055/s-2000-9775
- Huang, Q. F., Huang, R. B., Wei, L., Chen, Y. X., Lv, S., Liang, C., et al. (2013). Antiviral activity of methyl helicterate isolated from *Helicteres angustifolia* (Sterculiaceae) against Hepatitis B virus. *Antivir. Res.* 100 (2), 373–381. doi:10.1016/j.antiviral.2013.09.007
- Huang, S.-X., Mou, J.-F., Luo, Q., Mo, Q.-H., Zhou, X.-L., Huang, X., et al. (2019). Anti-hepatitis B virus activity of esculetin from *Microsorium fortunei* *in vitro* and *in vivo*. *Molecules* 24, 3475. doi:10.3390/molecules24193475
- Huang, T. J., Liu, S. H., Kuo, Y. C., Chen, C. W., and Chou, S. H. (2014). Antiviral activity of chemical compound isolated from *Artemisia morrissonensis* against Hepatitis B virus *in vitro*. *Antivir. Res.* 10 (1), 97–104. doi:10.1016/j.antiviral.2013.11.007
- Huang, W. Y., Zhang, H. C., Liu, W. X., and Li, C. Y. (2012). Survey of antioxidant capacity and phenolic composition of blueberry, blackberry, and strawberry in Nanjing. *J. Zhejiang Univ. Sci. B* 13, 94–102. doi:10.1631/jzus.B1100137
- Jacob, J. R., Mansfield, K., You, J. E., Tennant, B. C., and Kim, Y. H. (2007). Natural iminosugar derivatives of 1-deoxynojirimycin inhibit glycosylation of hepatitis viral envelope proteins. *J. Microbiol.* 45, 431–440.
- Jenssen, H., and Hancock, R. E. (2009). Antimicrobial properties of lactoferrin. *Biochimie* 91 (1), 19–29. doi:10.1016/j.biochi.2008.05.015
- Jiang, L., Gu, Y., Ye, J., Liu, F., Zhao, Y., Wang, C., et al. (2012). Resveratrol prevents hepatic steatosis induced by hepatitis C virus core protein. *Biotechnol. Lett.* 34 (12), 2205–2212. doi:10.1007/s10529-012-1034-0
- Jiang, X., Kanda, T., Haga, Y., Sasaki, R., Nakamura, M., Wu, S., et al. (2017). Glucose-regulated protein 78 is an antiviral against hepatitis A virus replication. *Exp. Ther. Med.* 13, 3305–3308. doi:10.3892/etm.2017.4407
- Jiang, X., Ren, Q., Chen, S. P., Tong, X. B., Dong, M., Sugaya, S., et al. (2011). UVC mutagenicity is suppressed in Japanese miso-treated human R5a cells, possibly via GRP78 expression. *Biosci. Biotechnol. Biochem.* 75, 1685–1691. doi:10.1271/bbb.110175
- Jiang, Z.-Y., Liu, W.-F., Zhang, X.-M., Luo, J., Ma, Y.-B., and Chen, J. J. (2013). Anti-HBV active constituents from *Piper longum*. *Bioorg. Med. Chem. Lett.* 23 (7), 2123–2127. doi:10.1016/j.bmcl.2013.01.118
- Jie, X.-X., Geng, C.-A., Huang, X.-Y., Ma, Y. B., Zhang, X. M., Zhang, R. P., et al. (2015). Five new secoiridoid glycosides and one unusual lactonic enol ketone with anti-HBV activity from *Swertia cincta*. *Fitoterapia* 102, 96–101. doi:10.1016/j.fitote.2015.02.009
- Joshi, S. S., Howell, A. B., and D'Souza, D. H. (2016). Reduction of enteric viruses by blueberry juice and blueberry proanthocyanidins. *Food Environ. Virol.* 8 (4), 235–243. doi:10.1007/s12560-016-9247-3
- Joshi, S. S., Su, X., and D'Souza, D. H. (2015). Antiviral effects of grape seed extract against feline calicivirus, murine norovirus, and hepatitis A virus in model food systems and under gastric conditions. *Food Microbiol.* 52, 1–10. doi:10.1016/j.fm.2015.05.011
- Kang, E. H., Kwon, T. Y., Oh, G. T., Park, W. F., Park, S. I., Park, S. K., et al. (2006). The flavonoid ellagic acid from a medicinal herb inhibits host immune tolerance induced by the Hepatitis B virus-e antigen. *Antivir. Res.* 72, 100–106. doi:10.1016/j.antiviral.2006.04.006
- Kasdallah-Grissa, A., Mornagui, B., Aouani, E., Gharbi N., and Kamoun, A. (2006). Protective effect of resveratrol on ethanol-induced lipid peroxidation in rats. *Alcohol Alcohol* 41, 236–239. doi:10.1093/alcal/agh256
- Kassem, A. F., Shaheen, M. N. F., Batran, R. Z., Abbas, E. M. H., Elseginy, S. A., and Elmahdy, E. M. (2019). New 4-phenylcoumarin derivatives as potent 3C protease inhibitors: Design, synthesis, anti-HAV effect and molecular modeling. *Eur. J. Med. Chem.* 168, 447–460. doi:10.1016/j.ejmech.2019.02.048
- Kim, S., Hwang, B. Y., Su, B. N., Chai, H., Mi, Q., Kinghorn, A. D., et al. (2007). Silvestrol, a potential anticancer rocaglate derivative from *Aglaia foveolata*, induces apoptosis in LNCaP cells through the mitochondrial/apoptosome pathway without activation of executioner caspase-3 or -7. *Anticancer Res.* 27 (4B), 2175–2183.
- Kim, W. R. (2009). Epidemiology of Hepatitis B in the United States. *Hepatology* 49 (5), S28–S34. doi:10.1002/hep.22975
- Kuang, X., Lu, W., Zeng, X. P., Liang, G. Y., and Xu, B. X. (2019). Synthesis and anti-HBV activity evaluation of Matijun-Su derivatives containing veratric acid. *Chin. J. New Drugs* 28 (12), 1140–1145.
- Lahlou, M. (2013). The success of natural products in drug discovery. *Pharmacol. Pharm.* 4, 17–31. doi:10.4236/pp.2013.43a003
- Lampertico, P., Agarwal, K., and Berg, T. (2017). EASL 2017 clinical practice guidelines on the management of Hepatitis B virus infection. *J. Hepatol.* 67 (2), 370–398. doi:10.1016/j.jhep.2017.03.021
- Lan, K.-H., Wang, Y.-W., Lee, W.-P., Lan, K.-L., Tseng, S.-H., Hung, L.-R., et al. (2012). Multiple effects of Honokiol on the life cycle of hepatitis C virus. *Liver Int.* 32, 989–997. doi:10.1111/j.1478-3231.2011.02621.x
- Lee, D. Y., Chung, S. J., and Kim, K. W. (2013b). Sensory characteristics of different types of commercial soy sauce. *J. Korean Soc. Food Cult.* 28, 640–650. doi:10.7318/kjfc/2013.28.6.640
- Lee, H. Y., Yum, J. H., Rho, Y. K., Oh, S. J., Choi, H. S., Chang, H. B., et al. (2007). Inhibition of HCV replicon cell growth by 2-arylbenzofuran derivatives isolated from *Mori Cortex Radicis*. *Planta Med.* 73, 1481–1485. doi:10.1055/s-2007-990249
- Lee, M. H., Lee, B.-H., Lee, S., and Choi, C. (2013a). Reduction of hepatitis A virus on frhk-4 cells treated with Korean red ginseng extract and ginsenosides. *J. Food Sci.* 00, M1412–M1415. doi:10.1111/1750-3841.12205
- Lee, M. H., Lee, B. H., Jung, J. Y., Cheon, D. S., Kim, K. T., and Choi, C. (2011). Antiviral effect of Korean red ginseng extract and ginsenosides on murine norovirus and feline calicivirus as surrogates for human norovirus. *J. Ginseng Res.* 35 (4), 429–435. doi:10.5142/jgr.2011.35.4.429
- Li, L. Q., Li, J., Huang, Y., Wu, Q., Deng, S. P., Su, X. J., et al. (2012c). Lignans from the heartwood of *Streblus asper* and their inhibiting activities to Hepatitis B virus. *Fitoterapia* 83 (2), 303–309. doi:10.1016/j.fitote.2011.11.008
- Li, Y. T., Xu, R. A., and Cui, X. L. (2012a). Progress in anti-Hepatitis B virus natural drugs targeting different sites. *Chin. J. Pharmacol. Toxicol.* 26 (5), 702–705.
- Li, Y., Yu, S., Liu, D., Proksch, P., and Lin, W. (2012b). Inhibitory effects of polyphenols toward HCV from the mangrove plant *Excoecaria agallocha* L. *Bioorg. Med. Chem. Lett.* 22, 1099–1102. doi:10.1016/j.bmcl.2011.11.109
- Lin, L., Hsu, W., and Lin, C. (2014). Antiviral natural products and herbal medicines. *J. Tradit. Complement. Med.* 4 (1), 24–35. doi:10.4103/2225-4110.124335
- Liu, M. M., Zhou, L., He, P. L., Zhang, Y. N., Zhou, J. Y., Shen, Q., et al. (2012). Discovery of flavonoid derivatives as anti-HCV agents via pharmacophore search combining molecular docking strategy. *Eur. J. Med. Chem.* 52, 33–43. doi:10.1016/j.ejmech.2012.03.002
- Liu, Y. P., Qiu, X. Y., Liu, Y., and Ma, G. (2019). Research progress on pharmacological effect of *artemisiae scopariae* herba. *Chin. Traditional Herb. Drugs* 9, 2235–2241.
- Liu, Y. W., Wu, C., Pei, R., Song, J., and Chen, S. (2013). Dioscin's antiviral effect *in vitro*. *Virus Res.* 172 (1–2), 9–14. doi:10.1016/j.virusres.2012.12.001
- Luthra, R., Roy, A., Pandit, S., and Prasad, R. (2021). Biotechnological methods for the production of ginsenosides. *South Afr. J. Bot.* 141, 25–36. doi:10.1016/j.sajb.2021.04.026
- MacLachlan, J. H., and Cowie, B. C. (2015). Hepatitis B virus epidemiology. *Cold Spring Harb. Perspect. Med.* 5 (5), a021410. doi:10.1101/cshperspect.a021410
- Mathayan, M., Jayaraman, S., Kulanthai, L., and Suresh, A. (2019). Inhibition studies of HBV DNA polymerase using seed extracts of *Pongamia pinnata*. *Bioinformation* 15 (7), 506–512. doi:10.6026/97320630015506

- Meng, X. X., Wu, S. Z., Yang, L., Cui, C., and Cen, Z. J. (2018). Screening of marine natural active small molecules against Hepatitis B virus. *Mod. Prev. Med.* 45 (23), 4335–4340.
- Misumi, I., Li, Z., Sun, L., Das, A., Shiota, T., Cullen, J., et al. (2021). Iminosugar glucosidase inhibitors reduce hepatic inflammation in hepatitis A virus-infected *ifnar1^{-/-}* mice. *J. Virol.* 95, e0005821. doi:10.1128/JVI.00058-21
- Nelson, N. P., Link-Gelles, R., Hofmeister, M. G., Romero, J. R., Moore, K. L., Ward, J. W., et al. (2018). Update: Recommendations of the advisory committee on immunization practices for use of hepatitis A vaccine for postexposure prophylaxis and for preexposure prophylaxis for international travel. *MMWR. Morb. Mortal. Wkly. Rep.* 67 (43), 1216–1220. doi:10.15585/mmwr.mm6743a5
- Netter, H. J., Barrios, M. H., Littlejohn, M., and Yuen, L. K. W. (2021). Hepatitis delta virus (HDV) and delta-like agents: Insights into their origin. *Front. Microbiol.* 12, 652962. doi:10.3389/fmicb.2021.652962
- Newman, D. J., and Cragg, G. M. (2012). Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.* 75, 311–335. doi:10.1021/np200906s
- Parvez, M. K., Al-Dosari, M. S., Alam, P., Rehman, M. T., Alajmi, M. F., and Alqahtani, A. S. (2019). The anti-Hepatitis B virus therapeutic potential of anthraquinones derived from Aloe vera. *Phytother. Res.* 33 (11), 2960–2970. doi:10.1002/ptr.6471
- Peng, Z., Fang, G., Peng, F. H., Pan, Z. Y., Su, Z. Y., Tian, W., et al. (1970). Effects of Rubiadin isolated from *Prismatomeris connata* on anti-Hepatitis B virus activity *in vitro*. *Phytother. Res.* 31 (12), 1962–1970. doi:10.1002/ptr.5945
- Perrillo, R. (2009). Benefits and risks of interferon therapy for Hepatitis B. *Hepatology* 49 (5), S103–S111. doi:10.1002/hep.22956
- Qin, S. D., Wang, Y., Wang, W., and Zhu, W. M. (2016). Anti-H1N1-virus secondary metabolites from mangrove-derived aciduric fungus *Penicillium* sp. OUCMDZ-4736. *Chin. J. Mar. Drugs* 35, 21–28.
- Rabiei, Z., Bigdeli, M. R., and Lorigooini, Z. (2015). A review of medicinal herbs with antioxidant properties in the treatment of cerebral ischemia and reperfusion. *J. Babol Univ. Med. Sci.* 17 (12), 47–56.
- Ray, S. C., Thomas, D. L., Hepatitis, C., Bennett, J. E., Dolin, R., and Blaser, M. J. (2015). *Mandell, douglas, and bennett's principles and practice of infectious diseases*. Philadelphia, Pennsylvania: Saunders.
- Redwan, E. M., El-Fakharany, E. M., Uversky, V. N., and Linjawi, M. H. (2014). Screening the anti infectivity potentials of native N-and C-lobes derived from the camel lactoferrin against hepatitis C virus. *BMC Complement. Altern. Med.* 14, 219. doi:10.1186/1472-6882-14-219
- Redwan, E. R., and Tabll, A. (2007). Camel lactoferrin markedly inhibits hepatitis C virus genotype 4 infection of human peripheral blood leukocytes. *J. Immunoass. Immunochem.* 28, 267–277. doi:10.1080/15321810701454839
- Ringelhan, M., McKeating, J. A., and Protzer, U. (2017). Viral hepatitis and liver cancer. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 372 (1732), 20160274. doi:10.1098/rstb.2016.0274
- Rizzetto, M. (2015). Hepatitis D virus: Introduction and epidemiology. *Cold Spring Harb. Perspect. Med.* 5 (7), a021576. doi:10.1101/cshperspect.a021576
- Roy, A., and Datta, S. (2021). Medicinal plants against ischemic stroke. *Curr. Pharm. Biotechnol.* 22 (10), 1302–1314. doi:10.2174/1389201021999201209222132
- Shabanah, W., Bukhari, A., Alandijani, A., Alyasi, A., and Youssef, A.-R. (2019). Prevalence of HBV and assessment of hepatitis B vaccine response among dental health care workers in dental teaching hospital, Umm Al-qura university, Saudi arabia. *Egypt. J. Immunol.* 26, 11–17.
- Shahid, I., Alzahrani, A. R., Al-Ghamdi, S. S., Alanazi, I. M., Rehman, S., and Hassan, S. (2021). Hepatitis C diagnosis: Simplified solutions, predictive barriers, and future promises. *Diagnostics* 11 (7), 1253. doi:10.3390/diagnostics11071253
- Shin, M. S., Kang, E. H., and Lee, Y. I. (2005). A flavonoid from medicinal plants blocks Hepatitis B virus-e antigen secretion in HBV-infected hepatocytes. *Antivir. Res.* 67, 163–168. doi:10.1016/j.antiviral.2005.06.005
- Silvera, D., Formenti, S. C., and Schneider, R. J. (2010). Translational control in cancer. *Nat. Rev. Cancer* 10 (4), 254–266. doi:10.1038/nrc2824
- Sinn, D. H., Cho, E. J., Kim, J. H., Kim, D. Y., Kim, Y. J., and Choi, M. S. (2017). Current status and strategies for viral hepatitis control in Korea. *Clin. Mol. Hepatol.* 23 (3), 189–195. doi:10.3350/cmh.2017.0033
- Solati, K., Heidari-Soureshjani, S., Luther, T., and Asadi-Samani, M. (2017). Iranian medicinal plants effective on sexual disorders: A systematic review. *Int. J. Pharm. Sci. Res.* 8 (6), 2415–2420.
- Steinmann, E., Whitfield, T., Kallis, S., Dwek, R. A., Zitzmann, N., Pietschmann, T., et al. (2007). Antiviral effects of amantadine and iminosugar derivatives against hepatitis C virus. *Hepatology* 46, 330–338. doi:10.1002/hep.21686
- Steinmann, J., Buer, J., Pietschmann, T., and Steinmann, E. (2013). Anti-infective properties of epigallocatechin-3-gallate (EGCG), a component of green tea. *Br. J. Pharmacol.* 168, 1059–1073. doi:10.1111/bph.12009
- Suzuki, M., Sasaki, K., Yoshizaki, F., Oguchi, K., Fujisawa, M., and Cyong, J.-C. (2005). Anti-hepatitis C virus effect of citrus unshiu peel and its active ingredient nobletin. *Am. J. Chin. Med.* 33, 87–94. doi:10.1142/S0192415X05002680
- Takeshita, M., Ishida, Y., Akamatsu, E., Ohmori, Y., Sudoh, M., Uto, H., et al. (2009). Proanthocyanidin from blueberry leaves suppresses expression of subgenomic hepatitis C virus RNA. *J. Biol. Chem.* 284, 21165–21176. doi:10.1074/jbc.M109.004945
- Todd, D., Moeller, N., Praditya, D., Kinast, V., Friesland, M., Engelmann, M., et al. (2018). The natural compound silvestrol inhibits hepatitis E virus (HEV) replication *in vitro* and *in vivo*. *Antivir. Res.* 157, 151–158. doi:10.1016/j.antiviral.2018.07.010
- Tsukamoto, Y., Ikeda, S., Uwai, K., Taguchi, R., Chayama, K., Sakaguchi, T., et al. (2018). Rosmarinic acid is a novel inhibitor for Hepatitis B virus replication targeting viral epsilon RNA-polymerase interaction. *PLoS One* 13 (5), e0197664. doi:10.1371/journal.pone.0197664
- Ungvari, Z., Orosz, Z., Rivera, A., Labinsky, N., Xiangmin, Z., Olson, S., et al. (2007). Resveratrol increases vascular oxidative stress resistance. *Am. J. Physiol. Heart Circ. Physiol.* 292, H2417–H2424. doi:10.1152/ajpheart.01258.2006
- Wang, H. L., Geng, C. A., Ma, Y. B., Zhang, X. M., and Chen, J. J. (2013). Three new secoiridoids, swermacrolactones A-C and anti-Hepatitis B virus activity from *Swertia macrocarpa*. *Fitoterapia* 89, 183–187. doi:10.1016/j.fitote.2013.06.002
- Wang, H. N., Yin, Z. F., Yin, X., Li, H. B., and Zhao, G. Q. (2019). Chemical constituents from *Pogonatherum crinitum* and their anti-HBV activities *in vitro*. *Chin. Tradit. Pat. Med.* 41 (6), 363–368.
- Watashi, K., Ishii, N., Hijikata, M., Inoue, D., Murata, T., Miyanari, Y., et al. (2005). Cyclophilin B is a functional regulator of hepatitis C virus RNA polymerase. *Mol. Cell.* 19, 111–122. doi:10.1016/j.molcel.2005.05.014
- Watashi, K., and Shimotohno, K. (2007). Chemical genetics approach to hepatitis C virus replication: Cyclophilin as a target for anti-hepatitis C virus strategy. *Rev. Med. Virol.* 17, 245–252. doi:10.1002/rmv.534
- Wedemeyer, H., Yurdaydin, C., Dalekos, G. N., Erhardt, A., Cakaloglu, Y., Degertekin, H., et al. (2011). Peginterferon plus adefovir versus either drug alone for hepatitis delta. *N. Engl. J. Med.* 364 (4), 322–331. doi:10.1056/NEJMoa0912696
- Wedemeyer, H., Yurdaydin, C., Hardtke, S., Caruntu, F. A., Curescu, M. G., Yalcin, K., et al. (2019). Peginterferon alfa-2a plus tenofovir disoproxil fumarate for hepatitis D (HIDIT-II): A randomised, placebo controlled, phase 2 trial. *Lancet. Infect. Dis.* 19 (3), 275–286. doi:10.1016/S1473-3099(18)30663-7
- Win, N. N., Kanda, T., Nakamoto, S., Moriyama, M., Jiang, X., Suganami, A., et al. (2018). Inhibitory effect of Japanese rice-koji miso extracts on hepatitis A virus replication in association with the elevation of glucose-regulated protein 78 expression. *Int. J. Med. Sci.* 15 (11), 1153–1159. doi:10.7150/ijms.27489
- Wohlfarth, C., and Efferth, T. (2009). Natural products as promising drug candidates for the treatment of Hepatitis B and C. *Acta Pharmacol. Sin.* 30 (1), 25–30. doi:10.1038/aps.2008.5
- Wu, S. F., Lin, C. K., Chuang, Y. S., Chang, F. R., Tseng, C. K., Wu, Y. C., et al. (2012). Anti-hepatitis C virus activity of 3-hydroxy caruillignan C from *Swietenia macrophylla* stems. *J. Viral Hepat.* 19, 364–370. doi:10.1111/j.1365-2893.2011.01558.x
- Xiao, D. Y. (2018). *Experimental research of Epimedium Hyde II anti-HBV in vivo and in vitro*[D]. Zunyi Medical University.
- Xu, Z. C., Zhao, K. T., and Jiang, Y. A. (2019). Development of antiviral drugs against Hepatitis B virus. *Chin. Sci. Bull.* 64, 3123–3141. doi:10.1360/tb-2019-0038
- Yang, X. X., Cao, P. X., Huang, Z. M., and Liang, G. Y. (2014). Synthesis and anti-Hepatitis B virus activities of Matijun-Su derivatives. *Cent. South Pharm.* 12 (2), 97–102.
- Yao, X. C., Xiao, X., Huang, B. K., and Xu, Z. Y. (2019). Molecular docking and *in vitro* screening of active anti-Hepatitis B virus components from *Abrus cantoniensis*. *Chin. J. Clin. Pharmacol.* 35 (5), 439–441.
- Ye, P., Zhang, S., Zhao, L., Dong, J., Jie, S., Pang, R., et al. (2009). Tea polyphenols exerts anti Hepatitis B virus effects in a stably HBV-transfected cell line. *J. Huazhong Univ. Sci. Technol. Med. Sci.* 29, 169–172. doi:10.1007/s11596-009-0206-1
- Yi, W. S. (2012). Research progress of flavonoids biological activity. *Guangzhou Chem. Ind.* 40 (2), 47–50.
- Yun, T. K. (2001). Brief introduction of Panax ginseng C.A. Meyer. *J. Korean Med. Sci.* 16, S3–S5.
- Zarrin, A., and Akhondi, H. (2021). “Viral hepatitis,” in *StatPearls* (Treasure Island (FL): StatPearls Publishing). Available from: <https://www.ncbi.nlm.nih.gov/books/NBK556029/#>.

- Zeinab, N., and Kopaei, M. R. (2018). Effective medicinal plants in treating Hepatitis B. *Int. J. Pharm. Sci. Res.* 9 (9), 3589–3596. doi:10.13040/IJPSR.0975-8232
- Zeng, F. L., Xiang, Y. F., Liang, Z. R., Wang, X., Huang, D. E., Zhu, S. N., et al. (2013). Anti-Hepatitis B virus effects of dehydrocheilanthifoline from *Corydalis saxicola*. *Am. J. Chin. Med.* 41 (1), 119–130. doi:10.1142/S0192415X13500092
- Zhang, J. H., Liu, W. T., and Luo, H. M. (2018). Advances in activities of terpenoids in medicinal plants. *Mod. Traditional Chin. Med. Materia Medica-World Sci. Technol.* 3, 419–430.
- Zhang, Z. Q., Li, S. F., Zhang, L. W., and Chao, J. B. (2017). Chemical constituents from flowers of *Wikstroemia chamaedaphne* and their anti-Hepatitis B virus activity. *Chin. Traditional Herb. Drugs* 48 (7), 1292–1297. doi:10.7501/j.issn.0253-2670.2017.07.005
- Zhao, Y., Geng, C.-A., Chen, H., Ma, Y. B., Huang, X. Y., Cao, T. W., et al. (2015). Isolation, synthesis and anti-Hepatitis B virus evaluation of p-hydroxyacetophenone derivatives from *Artemisia capillaris*. *Bioorg. Med. Chem. Lett.* 25 (7), 1509–1514. doi:10.1016/j.bmcl.2015.02.024
- Zhao, Y., Geng, C. A., Sun, C. L., Ma, Y. B., Huang, X. Y., Cao, T. W., et al. (2014). Polyacetylenes and anti-Hepatitis B virus active constituents from *Artemisia capillaris*. *Fitoterapia* 95, 187–193. doi:10.1016/j.fitote.2014.03.017
- Zhou, W. B., Zeng, G. Z., Xu, H. M., He, W. J., and Tan, N. H. (2013). Astataricusones A-D and astataricusol A, five new anti-HBV shionane-type triterpenes from *Aster tataricus* L. f. *Molecules* 18 (12), 14585–14596. doi:10.3390/molecules181214585
- Zhou, X. L., Wen, Q. W., Lin, X., Zhang, S. J., Li, Y. X., Guo, Y. J., et al. (2014). A new phenylethanoid glycoside with antioxidant and anti-HBV activity from *Tarphochlamys affinis*. *Arch. Pharm. Res.* 37 (5), 600–605. doi:10.1007/s12272-013-0219-y
- Zhou, X., Xu, L., Wang, Y., Wang, W., Sprengers, D., Metselaar, H. J., et al. (2015). Requirement of the eukaryotic translation initiation factor 4F complex in hepatitis E virus replication. *Antivir. Res.* 124, 11–19. doi:10.1016/j.antiviral.2015.10.016
- Zoulim, F., and Durantel, D. (2015). Antiviral therapies and prospects for a cure of chronic Hepatitis B. *Cold Spring Harb. Perspect. Med.* 5 (4), a021501. doi:10.1101/cshperspect.a021501
- Zuo, G. Y., Li, Z. Q., Chen, L. R., and Xu, X. J. (2005). *In vitro* anti-HCV activities of *Saxifraga melanocentra* and its related polyphenolic compounds. *Antivir. Chem. Chemother.* 16, 393–398. doi:10.1177/095632020501600606



SAR and Optical Pixel Level Fusion Methods and Evaluations

Sanjay Singh & K. C. Tiwari

To cite this article: Sanjay Singh & K. C. Tiwari (2022): SAR and Optical Pixel Level Fusion Methods and Evaluations, Journal of Spatial Science, DOI: [10.1080/14498596.2022.2153754](https://doi.org/10.1080/14498596.2022.2153754)

To link to this article: <https://doi.org/10.1080/14498596.2022.2153754>



Published online: 14 Dec 2022.



Submit your article to this journal [↗](#)



Article views: 20



View related articles [↗](#)



View Crossmark data [↗](#)



SAR and Optical Pixel Level Fusion Methods and Evaluations

Sanjay Singh^a and K. C. Tiwari^b

^aDepartment of Electronics & Communication, Delhi Technological University, New Delhi, India;

^bDepartment of Civil Engineering, Delhi Technological University, New Delhi, India

ABSTRACT

In the past years, many vital improvements in image fusion have been recognised, particularly in the fusion of Synthetic Aperture Radar (SAR) and optical sensors. Many kinds of research have been published focusing on pixel-level fusion of SAR and optical images. This paper focuses on pixel-level SAR-optical fusion methods, along with performance assessment and applications. Five categories (component substitution, numerical method, model-based, multi-resolution, and hybrid methods) of fusion are presented. Subjective, objective, and comprehensive methods are surveyed to assess fusion. Experimental findings utilising Sentinel-1A and Sentinel-2A data corroborate the assessment. Finally, paper suggests that there is still space for investigation into SAR-optical fusion.

ARTICLE HISTORY

Received 10 October 2021

Accepted 27 November 2022

KEYWORDS

Image fusion; SAR-optical fusion; comprehensive assessment; sentinel

1. Introduction

Image fusion leads to more accurate information and is useful for various applications, ranging from object detection, urban mapping, Land-use and Land-cover (LULC), change detection, and so on. Multisensory image fusion aims to combine complementary information in a single image that serves a better understanding of the objects observed. Integrating SAR and optical data is an essential example of utilising complementary information from different sensors (Abdikan *et al.* 2015, Kumar *et al.* 2017). Whereas SAR imagery detects the physical attributes of the viewed scene and may be collected regardless of whether or not there is daylight, optical imaging measures chemical properties. It requires both daylight and a clear sky. Optical data, on the other hand, are considerably simpler for human operators to understand and generally offers more information (Herold and Haack 2002). SAR data, on the other hand, include amplitude and phase information, allowing for high-precision assessment of 3D topography and deformations (Sheoran and Haack 2014). The analysis of fused data helps in the knowledge and interpretation of the region being observed (Pohl and van Genderen 2015, Ghassemian 2016).

Pre-processing of SAR data is an essential primary step of the fusion process. Speckle reduction and image co-registration are two necessary pre-processing steps for SAR-optical fusion. SAR images are corrupted due to coherent backscattered signals, i.e. speckle noise. Speckle is generated either by the constructive or destructive interference of the coherent returns scattered by tiny reflectors (Lopes *et al.* 1990). It causes the images to seem grainy,

making it challenging to understand SAR images visually. Traditionally, two approaches have been employed for speckle reduction. The first, called multilook processing, entails the incoherent blending of multiple looks during SAR image production. The second method, which is used after the multilook SAR image has been formed, comprises adaptive spatial filtering based on evaluating the local statistics around a specific pixel (Xie *et al.* 2002). The most well-known filters are the Lee filter, Refined-Lee filter, Enhanced Frost filter, and many more (Wakabayashi and Arai 1996). Non-local filter algorithms, Gaussian denoising (Deledalle *et al.* 2017), and block matching utilising a 3D filter comprised non-local filters are examples of recent advances in speckle reduction (Deledalle *et al.* 2017). Several studies (France 2017, Wang *et al.* 2017) used a CNN technique to filter the SAR and found significant results. The technique of geometrically aligning SAR and optical images is known as co-registration. Dawn *et al.* (2010) presents a review of image registration methods used in remote sensing. SAR and optical sensor co-registration are challenging tasks (Schmidt *et al.* 2018) and another significant source of fusion errors. Though georeferenced, images captured with different sensors suffer from an issue of poor alignment since the data capture instant of two satellites is not similar, they do not all exactly trend in a similar way (Merkle *et al.* 2015). The following categories are being used to classify image registration techniques (Li *et al.* 1995): (i) feature-based approaches and (ii) intensity-based approaches. Feature-based methods rely on identifying features or items that characterise crucial landmarks, sharp edges, or forms. Images are co-registered using intensity-based approaches by comparing the intensity pixel values in two different images (Inglada and Giros 2004). Feature-based approaches are based on contour detection (Li *et al.* 1995), Canny edge detector, Scale-invariant Fourier transform-based detection, Hough transform (Palmann *et al.* 2008), and shape feature-based detection (Huang and Li 2010). Nowadays, the deep learning-based approach has emerged as the third category. Wang *et al.* (2017) offer a framework of co-registration using deep-learning techniques for satellite images, covering the co-registration of multispectral and SAR images. Figure 1 shows the georeferenced SAR and optical data used for all the experiments of the study.

Image fusion methods are categorised according to many characteristics, including the manner of data collection and the level of processing. Fusion is classified as follows (Pohl 1999):

Pixel-level fusion (PLF) is performed at the lowest processing stage at a pixel-by-pixel scale. To improve processing tasks, it produces a merged composite in which measurable physical characteristics linked with individual pixels are derived through a collection of pixels in parent images.

Feature-level fusion (FLF) involves the extraction of characteristics that are dependent on their surroundings, such as pixel values, lineaments, and shapes.

Decision-level fusion (DLF) combines decisions made by single-sensor sub-systems into a combined decision usable for end-user action selection. The obtained information is then integrated to strengthen a common interpretation (Pohl and van Genderen 2014).

Despite having a higher computational cost, PLF methods are nevertheless widely used in satellite data fusion due to their higher accuracy. There is a loss of information in FLF and DLF, while data alteration is nominal in PLF compared to other categories (Pohl and Van Genderen 1998). This makes PLF techniques better suited for SAR-optical fusion. However, numerous publications on image fusion have made important contributions in this area (Dahiya *et al.* 2013, Simone *et al.* 2002, Zhang 2010, Ghamisi *et al.* 2019) but there remains an opportunity

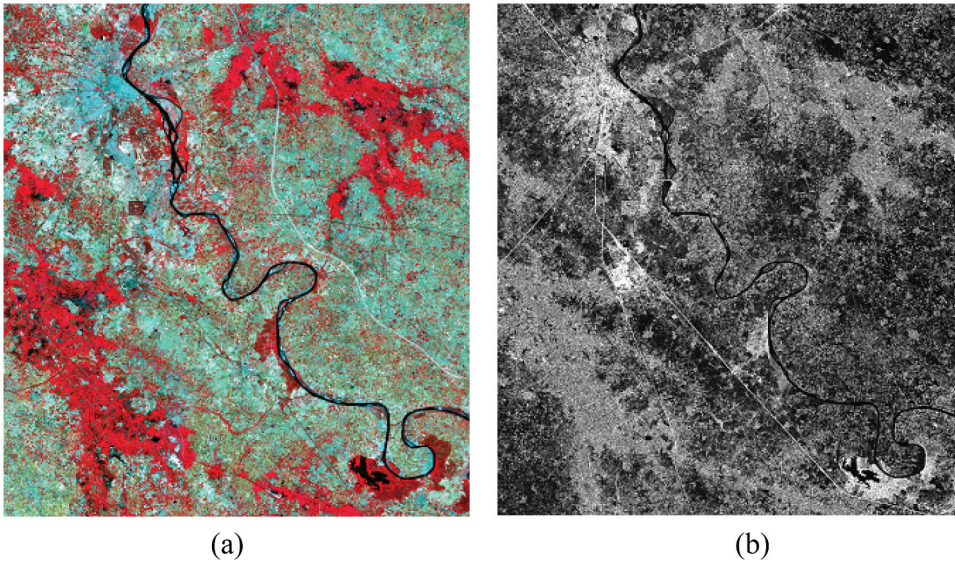


Figure 1. Mathura–Agra study region (a) Optical image (b) SAR image.

for additional critical study. Although SAR and optical fusion have been researched for several years, it has now acquired momentum, owing to two breakthroughs: (i) increasing availability of high-resolution data and, (ii) implementation of recent multinational projects (ESA’s Copernicus program). As a result, there is still significant potential for combined exploitation of SAR data combined with optical data.

The authors have conducted an extensive review of the available literature and discovered a general lack of literature that encompasses all essential elements of SAR-optical PLF methods and quality assessments i.e. subjective, objective and comprehensive, particularly the comprehensive evaluation. Therefore, the objective of the present work is as follows: (1) to conduct a pivotal study and analyse PLF methods and evaluate methodologies in all three domains (subjective, objective, and comprehensive) for SAR and optical images, and (2) to implement and evaluate the basic algorithms to include Principal Component Analysis (PCA), Brovey, Gram-Schmidt, and Colour Normalised (CN) spectral sharpening.

Section-1 has laid down the background and objectives, and [section 2](#) provides an overview of different SAR-optical image PLF techniques. [Section 3](#) outlines the many quality measures that have been utilized to evaluate the fused product. [Section 4](#) shows the fused image evaluation outcomes. [Section 5](#) provides a short overview of SAR-optical PLF applications, while [Section 6](#) summarises the conclusions.

2. Pixel-level fusion (PLF) methods

According to the literature review, PLF techniques are divided into the following categories: component substitution (CS), numerical techniques, model-based techniques, multiresolution approach (MRA), and hybrid techniques. [Table 1](#) summarises all these categories.

Table 1. SAR-optical PLF methods.

S. No.	Fusion Category	Approaches	Findings
i)	Component substitution	PCA (Herold and Haack 2002, Pal <i>et al.</i> 2007, Amarsaikhana <i>et al.</i> 2010, Abdikan and Sanli 2012), GS (Manaf <i>et al.</i> 2016), IHS (Harris <i>et al.</i> 1990, Liu <i>et al.</i> 2015, Kumar <i>et al.</i> 2017, Shao <i>et al.</i> 2020), Ehler (Ehlersa <i>et al.</i> 2010, Abdikan <i>et al.</i> 2015)	Less Computation and easier to perform. Highly dependent on correlation.
ii)	Numerical techniques	HPF (Abdikan <i>et al.</i> 2015), BT (Amarsaikhana <i>et al.</i> 2010a, Gibril <i>et al.</i> 2017) Modulation based Techniques (Facheris <i>et al.</i> 2004)	More spectral distortion. Not much suitable found for SAR-optical fusion.
iii)	Multi-Resolution Approaches	Pyramidal decomposition (Chandrakanth <i>et al.</i> 2014), Wavelet (Hong <i>et al.</i> 2009, Shah <i>et al.</i> 2019), Directional Approaches: curvelet, contourlet (Lu <i>et al.</i> 2011, El-Tawel and Helmy 2014)	Less Spectral distortion. Sometimes computationally complex
iv)	Model Based techniques	Variational model (Zhang 2010), SRBM (Huang <i>et al.</i> 2015, Zhouping 2015), CNN based (Shakya <i>et al.</i> 2020), cGAN (Grohnfeldt <i>et al.</i> 20182018, Bermudez <i>et al.</i> 2019)	Less susceptible to registration errors.
v)	Hybrid Approaches	NSCT+ IHS (Wang and Chen 2016), IHS/PCA + DWT/ AWT (Huang <i>et al.</i> 2005, Han <i>et al.</i> 2010, Kulkarni and Rege 2021), Modified BT (Chibani 2006), IHS +AWT+ EMD+AWT+IHS (Chen <i>et al.</i> 2010), weighted median filter + GS (Quan <i>et al.</i> 2020), PCA + curvelet transform (Chen <i>et al.</i> 2020)	Reduce spatial and spectral distortion.

Where AWT -Adaptive Wavelet transform; HPF- High pass Filter; CNN – convolution neural network; cGAN -conditional generative network.

2.1. Component substitution methods

The CS fusion methods are divided into three stages. First, once the multispectral (MS) bands have been registered to the SAR bands, forward transformation is enforced. Second, the higher resolution band replaces another component of the transformed data domain that is comparable to the SAR image. Third, the fused results are built using an inverse transform of the original space (Zhang 2010). Gram–Schmidt (GS), PCA, Ehlers fusion and Intensity-Hue-Saturation (IHS) are among prominent techniques in this area. IHS is a technique for obtaining spectral and spatial information from MS bands. The histogram matched SAR band replaces the intensity (I) component. Reconfigured I, Hue (H), and Saturation (S) components are translated inverted into the original domain to produce integrated MS data. Tu *et al.* (2004), who developed the generalized IHS, overcame the restriction of IHS to three bands. Researchers have suggested the adaptive IHS technique in order to continue the development of IHS-based solutions for overcoming spectral quality issues (Rahman *et al.* 2010).

To optimise the variance of the source image, PCA transforms an image by translating and rotating it into a new coordinate frame. This technique computes the PCs, then reassigns the high-resolution SAR data into the data space of the first PC and replaces it with PC1. Then, from PC1 to the original MS data, an inverse PC transform is accomplished. This technique adjusts SAR data to about the same data space as PC1 before performing the inverse PC computation (Dahiya *et al.* 2013). The first PC in Gram–Schmidt (GS) transform may be selected freely, and the other components are computed orthogonally to the first PC. The GS transform (GST) is used to convert MS and simulated SAR, and the resultant bands are utilized in the CS method. The statistically matched SAR replaces the first GS band. The fused image is the output of the inverse GS transform (Klonus and Ehlers 2007).

To enhance land cover mapping, Herold and Haack (2002) present a fusion of SAR and MS imagery utilising different layer additions and PCA techniques. After fusion with the SAR image, classification accuracy improves substantially. Harris *et al.* (1990) utilise the IHS transform to combine radar and MS images. El-Deen Taha *et al.* (2010) evaluate PLF methods based on IHS, PCA and the Brovey (BT) for a PLF of MS and SAR images to enhance the overall classification accuracy for seacoast shoreline extraction. Fusion based on IHS yields higher categorisation accuracy. To enhance the land cover categorisation accuracy, Amarsaikhan *et al.* (2011) suggest a combination of high-resolution SAR and MS bands. The multiplicative approach, IHS, PCA and BT-based PLF are used to fuse source images, and the outcomes are evaluated visually and based on overall accuracy. Manaf *et al.* (2016) evaluated IHS, BT and GS transforms to enhance classification results while extracting shorelines. It has been discovered that using integrated images increases the precision of coastline extraction. These methods provide a fused output that is good with spatial information. Due to spectral dissimilarity in the bands of heterogeneous images, there are pixel-level discrepancies between images being fused, which may develop significant spectral distortions in the integrated output (Wang *et al.* 2005, Ghassemian 2016).

2.2. Numerical methods

Mathematical combinations of various images are among the most straightforward PLF techniques. They combine the digital numbers of the images being fused on a pixel-by-pixel basis using mathematical operators such as sum, multiplication, subtraction, multiplicative, and so on. Multiplication can be a powerful fusion method and leads to perfect results for visual interpretation if optical and radar images are combined (Pohl and van Genderen 1995).

Fusion by multiplication enhances the contrast and joins MS with textural information from the SAR data. Selection of weighting and scale variables may enhance the fused image. The BT normalised MS data, and the resultant channels are multiplied by the intensity channel. It is not exactly a transform but a multiplication using a SAR band based on a normalisation of the MS bands. A smoothing filter-based intensity modulation method modulates optical data by using a ratio between the SAR and its histogram image. It utilises modulation of an intensity channel with spatial detail to make it applicable to SAR fusion (Facheris *et al.* 2004). CN Spectral Sharpening expands the BT to accommodate for more than three input channels. The method standardises the input and divides the spectral space into colour and brightness. It normalises the data by multiplying every MS channel by SAR data and dividing it by MS input images.

The High pass filter (HPF) method collects high-frequency information, which is subsequently added to every MS band of the MS band (Pohl *et al.*, 1998). For the integration of SAR with MS, Chandrakanth *et al.* (2014) use frequency domain high pass filtering. Misra *et al.* (2012) offer a PLF method for merging low-resolution (LR) optical bands with high resolution (HR) SAR data using the Colour Normalised transform, a variation of the BT.

2.3. Multi-resolution approach (MRA)

It is also called the multiscale decomposition method. The MRA relies on the insertion of spatial information that is acquired from the multi-resolution decomposition of the SAR band into the resampled multispectral images. In these techniques, the fused source images are first decomposed at various scales using an appropriate multi-resolution technique. Afterwards, fusion methods are enforced for every decomposed part, and the fused bands are inverse transformed to generate the required data (Ghassemian 2016).

Wavelet-based techniques are classified into two categories. The first method involves swapping an LR sub-band for a HR sub-band. The second method is built on inserting information from HR sub-bands into LR sub-bands (Ranchin and Wald 2000). Another common MRA technique is the non-subsampled wavelet transform. Due to its shift-invariance, this kind of multiscale transformation is ideal for multisensory data such as SAR and MS images. For integrating HR-SAR and LR optical bands, ARSIS techniques based on non-subsampled transformations are more often utilized. Abdikan *et al.* (2012) compare component replacement and wavelet-based fusion techniques. The HPF method and the wavelet-based techniques provide numerically comparable results, but the wavelet method yields aesthetically superior results.

Rahman *et al.* (2010) compare CS methods to wavelet-based fusion methods for increasing sub-surface and surface aligning accuracy. Facheris *et al.* (2004) offers an undecimated wavelet fusion technique of SAR, panchromatic (PAN), and optical bands to increase the spatial resolution of the optical band. To begin, PAN and MS bands are merged, and then texture information from SAR bands is injected into a pan-sharpened band to produce the resultant fusion output. Amarsaikhana *et al.* (2010) utilize wavelet PLF to merge optical and HR-SAR data. The results are compared to BT, PCA and Ehlers fusion for improved LULC categorisation. It has been discovered that images merged using BT have a higher spatial resolution. This study also shows that using integrated satellite data increases accuracy. Lu *et al.* (2011) combine optical and SAR data using curvelet decomposition. The results outperform IHS and the wavelet-based fusion technique.

Using a multiresolution transform adds computational complexity, but it results in improved fusion performance owing to simultaneous localization in the time and frequency domains (Ghassemian 2016). Wavelet-based techniques improve spectral resolution but introduce spatial distortion.

2.4. Hybrid methods

Hybrid techniques combine CS and MRA to make use of the benefits of both methods, resulting in improved spatial and spectral resolution. Huang *et al.* (2005) fuse SAR and MS images using the IHS and the DWT for better urban mapping. After transforming the MS bands to the IHS domain, the I-component and the SAR data are decomposed by a wavelet-based method. When this approach is compared to PCA, the suggested hybrid strategy outperforms PCA in terms of classification accuracy. Hong *et al.* (2009) proposes

a hybrid PLF technique that depends on the wavelet-based transform and IHS to merge SAR data with MS data. The PLF method in this technique depends on the neighborhood correlation between sensor images. The proposed hybrid technique is evaluated with traditional wavelet-based fusion and IHS methods.

Chibani (2006) suggests modifying the BT to include SAR characteristics in MS images. MS bands are modified by a fraction of the new I-component to the old I-component in this technique. The newer I-component is created by combining high-frequency information derived from AWT bands with SAR and inserting them into the previous component. To fuse MS and SAR images, Han *et al.* (2010) offers a hybrid method that depends on combining the adaptive wavelet transform and IHS. After converting the MS data to the IHS space, AWT is used to dissect the I-component and SAR data. Applying statistical methods, the decomposed sub-images are merged, and the merged sub-images are rehabilitated to give a merged I-component, which is then inverse transformed to the initial MS domain (Luo *et al.* 2014).

Byun *et al.* (2013) combine SAR, PAN and MS images using a hybrid method. To begin, the SAR band is divided into two regions: active and inactive. AWT is then used to combine the PAN and SAR bands, with different fusion criteria used for active and inactive regions. A hybrid pansharpening algorithm is used to combine a pan-sharpened SAR image with MS images. Zhang *et al.* (2020) examine several fusion techniques for a hybrid approach that includes IHS transformation and non-subsampled contourlet transform. The intensity picture contains spectral data, while the texture SAR image contains spatial data. To get the most spatial and spectral details in the fused data, these two images are merged using the PCA technique. It has been discovered that the fused data improves sea-ice recognition. Wang and Chen (2016) provide a hybrid method that is based on PCNN and the contourlet transform (CT). CT is acclimated to decompose the I-component of the SAR image after MS images are converted to the IHS domain. The coefficients of the SAR sub-images are then modulated using gradient transform and thresholding procedures. The output of PCNN is utilized to fuse modified coefficients of the intensity sub-images and coefficients of the SAR sub-images in the fusion stage.

Hybrid techniques combine the benefits of both MRA and CS fusion procedures. It has been discovered that hybrid fusion techniques combining MRA and CS methods result in improved SAR-optical image fusion (Abdikan *et al.* 2015).

2.5. Model-based techniques

Several model-based methods for fusing remotely sensed data have been suggested, and researchers have also implemented them for PLF of SAR and optical images. Sparse representation-based techniques (SRBM) and variational techniques are two of the most used methods.

The variational fusion technique developed for PLF of MS and PAN bands was expanded for PLF of MS and SAR bands by Zhang (2010) for improved comprehension of urban characteristics. The energy function is minimised in this fusion method, ensuring that colour details from MS data and textural details from SAR data are combined. Image fusion is approached as a restoration problem by sparse representation-based methods, which generate HR fused data from a linear integration of pixels drawn from an over-complete dictionary of HR and LR data. The sparse co-efficients of the initial data are

merged according to the PLF method, and the merged data is rebuilt adopting the aforementioned dictionary. For a better comprehension of urban characteristics, the variational fusion technique developed for the fusion of MS and PAN bands is extended to Zhang and Yu's fusion of MS and SAR bands (Chen *et al.* 2010). This fusion method minimises the energy function, ensuring that colour details from MS bands and geometric details from SAR bands are combined. Image fusion is approached as a restoration issue by sparse representation-based methods, which create HR fused data from a linear integration of pixels chosen from an over-complete dictionary of LR and HR bands. The sparse co-efficients of the original images were fused, and a fused image was created using the same language. For merging airborne SAR and optical data, Liu *et al.* (2016) propose a joint non-negative sparse technique using IHS. To begin, an I-component is retrieved by applying the IHS transformation to MS images. The sparse coefficients of the I-component are combined with the sparse coefficients of the SAR band. Fused data is produced by inverting the changed I-component. In certain cases, sparse representation methods outperform other techniques; nevertheless, dictionary creation and sparse coding are the main problems (Ghahremani and Ghassemian 2016).

It has been noted that hybrid fusion approaches are more attractive owing to their capability to manage the issue of spatial and spectral distortion and their lower complication when compared to sparse representation techniques.

3. Quality evaluation measures of SAR-optical PLF

Image fusion has several applications, and no image fusion technique performs equally well in every circumstance, particularly in SAR-optical PLF. To verify the fusion technique, the fused images were evaluated using three methods: objective, subjective, and comprehensive assessments (Meijie *et al.* 2015). Objective assessment has been done using quality measures (based on the presence or absence of reference data). The reduced resolution assessment (Wald protocol) is used when reference images may not be available.

3.1. Objective evaluation (quality metrics)

Jagalingam and Hegde (2015) provide a comprehensive analysis of quantitative measures used to assess fused images. A few assessment measures are described in further detail below.

3.1.1. Metrics (when the reference data is available)

In the presence of a reference image, quantitative measures such as correlation coefficient (CC), mean bias (MB), signal-to-noise ratio, mutual information (MI), ERGAS, structural similarity index measure (SSIM), root mean square error (RMSE), spectral angular mapper (SAM), universal quality index (UIQI/Q) etc. are calculated for SAR-optical PLF (Table 2).

3.1.2. Metrics that do not need reference data

In the absence of a reference image, quantitative measures such as entropy, cross entropy, standard deviation (SD), fusion quality index (FQI), mutual information, spatial frequency (SF), etc. have been used (Table 3).

Table 2. Metrics (when the reference data is available).

SN	Measures	References	Main points/Formula
1.	CC	Abdikan <i>et al.</i> (2012), Chen <i>et al.</i> (2010), Ehlers <i>et al.</i> (2010), El-Tawel and Helmy (2014), Hong <i>et al.</i> (2009), Kulkarni and Rege (2019)2019, Liu <i>et al.</i> (2016), Parcharidis and Kazi-Tani (2000), Wald <i>et al.</i> (1997), and Zhang and Yu (2010)	<p>Calculate the spectral feature similarity between the integrated and reference.</p> $CC = \frac{\sum_{j=1}^m \sum_{i=1}^n (R_{ij} - \bar{R})(F_{ij} - \bar{F})}{\sqrt{\sum_{j=1}^m \sum_{i=1}^n ((R_{ij} - \bar{R})^2 (F_{ij} - \bar{F})^2)}$ <p>R and F are a reference, and fusion image, respectively and \bar{R} and \bar{F} are corresponding mean images.</p>
2.	SAM	Byun <i>et al.</i> (2013), Kulkarni and Rege (2020)2020, and Liu <i>et al.</i> (2016)	<p>Calculate angle between integrated and reference images.</p> $SAM = \arccos \left(\frac{\sum_{j=1}^m \sum_{i=1}^n F_i R_i}{(\sum_{j=1}^m F_i^2)^{1/2} (\sum_{i=1}^n R_i^2)^{1/2}} \right)$ <p>B represents the number of bands.</p>
3.	UIQI/Q4 index	Abdikan <i>et al.</i> (2012), Byun <i>et al.</i> (2013), Chen <i>et al.</i> (2010), and El-Tawel and Helmy (2014)	<p>Calculates the amount of relevant information transform from reference to integrated images.</p> $UIQI = \frac{4\sigma_R \sigma_F}{(\sigma_F^2 + \sigma_R^2) (\bar{F}^2 + \bar{R}^2)}$ <p>σ^2 is variance.</p>
4.	RMSE	Chen <i>et al.</i> (2010), Chikr El-Mezouar <i>et al.</i> (2011)2011, Hong <i>et al.</i> (2009), Luo <i>et al.</i> (2014), and Parcharidis and Kazi-Tani (2000)	<p>Evaluate the variance in pixels to get the difference between the reference and integrated images.</p> $RMSE = \sqrt{\frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N (R_{ij} - F_{ij})^2}$ <p>M and N are rows and columns.</p>
5.	ERGAS	Byun <i>et al.</i> (2013), Kulkarni and Rege (2019), and Liu <i>et al.</i> (2016)	<p>Calculates the attribute of the integrated data in perspective of the normalised error of individual processed image. Higher value, more distortion in the merged image.</p> $ERGAS = 100 \frac{s}{m} \left(\frac{1}{B} \sum_{i=1}^B \left(\frac{RMSE_i^2}{mean^2} \right) \right)^{\frac{1}{2}}$ <p>Where s and m are the SAR and multispectral image resolution, respectively.</p>
6.	Mean Bias	Abdikan <i>et al.</i> (2012), Chibani (2006), Parcharidis and Kazi-Tani (2000), and Wald <i>et al.</i> (1997)	<p>difference between mean of fused and reference image.</p> $Mean\ Bias = \frac{MS_{mean} - Fused_{mean}}{MS_{mean}}$
7.	SSIM	Ehlersa <i>et al.</i> (2010), and Khosravi <i>et al.</i> (2017)2017)	<p>Compute the pixel patterns between the integrated and reference images locally.</p> $SSIM = \frac{(2\mu_R \mu_F + C_1)(2\sigma_{RF} + C_2)}{(\mu_R^2 + \mu_F^2 + C_1)(\sigma_R^2 + \sigma_F^2 + C_2)}$ <p>The variables C1 and C2 are utilised to stabilise the denominator where μ is average.</p>
8.	Relative SDD	Abdikan <i>et al.</i> (2012), and Chikr El-Mezouar <i>et al.</i> (2011)	<p>Compute SD difference between fused and reference data and divided by the mean of the reference images.</p> $Rel.\ SDD = \frac{Standard\ deviation(Reference - Fused\ band)}{Mean\ MS\ band}$

Table 3. Objective evaluation measures (reference data not required).

SN	measures	Reference	Points
1.	Entropy	Chandrakanth <i>et al.</i> (2011), Chibani (2006), El-Tawel and Helmy (2014), Kulkarni and Rege (2019), and Liu <i>et al.</i> (2015)	Measure the information content of resultant. $\text{Entropy} = - \sum_{i=0}^{L-1} p_i \log_2 p_i$ L and p_i are the total number of grey levels and corresponding normalised histograms, respectively.
2.	SD	Abdikan Fusun Balik Sanli (2012), Chen <i>et al.</i> (2010), Kulkarni and Rege (2019), Meijie <i>et al.</i> (n.d.), and Parihar <i>et al.</i> (2017)	measure the contrast of resultant $\text{SD} = \sqrt{\sum_{i=1}^M \sum_{j=1}^N (F - \bar{F})^2}$
3.	MI	Li <i>et al.</i> (2008), Qu <i>et al.</i> (2002)4, and You <i>et al.</i> (2014)	compare the resemblance of reference and fused images $\text{MI} = \sum_{R,F} H_{R,F}(r, f) \log \left[\frac{H_{R,F}(r, f)}{H_R(r)H_F(f)} \right]$ Where $H_R(r)$ and $H_F(f)$ represent the histogram of MS image R and the fused image F. $H_{R,F}(r, f)$ denote joint histogram of MS and Fused image.
4.	SF	Kulkarni and Rege (2019), and Zheng <i>et al.</i> (2007)	measures the overall activity level $\text{SF} = \sqrt{\sum_{i=1}^M \sum_{j=1}^N \left[\{F(i, j) - F(i, j - 1)\}^2 - \{F(i, j) - F(i - 1, j)\}^2 \right]}$

Table 4. Evaluation of fusion image with classification accuracy.

S. No.	Classification Method	References
1.	MLC	Abdikan <i>et al.</i> (2015), Ali <i>et al.</i> (2018), El-Deen Taha and Elbeih (2010), Facheris <i>et al.</i> (2004), Huang <i>et al.</i> (2005), Kulkarni and Rege (2019), Manaf <i>et al.</i> (2016), Neetu and Ray (2020), Rusmini <i>et al.</i> (2012), and Sheoran and Haack (2014)
2.	SVM	Abdikan <i>et al.</i> (2015), Clerici <i>et al.</i> (2017), Gibril <i>et al.</i> (2017), and Manaf <i>et al.</i> (2016)
3.	RF	Abdikan <i>et al.</i> (2015), Clerici <i>et al.</i> (2017), Manakos <i>et al.</i> (2020), and Veerabhadraswamy <i>et al.</i> (2021)
4.	Fuzzy classification	El-Deen Taha and Elbeih (2010), and Riedel <i>et al.</i> (2007)
5.	K-NN	Abdikan <i>et al.</i> (2015), and Clerici <i>et al.</i> (2017)
6.	Neural Network	Manaf <i>et al.</i> (2016), and Meraner <i>et al.</i> (2020)

3.2. Comprehensive assessment

It is proposed that quality measures should not be the sole means of interpreting fused images. In a few applications, objective metric analysis of an image with a low value may provide superior class accuracy. To enhance the assessment of the fusion outcome, the analysis may be built from the application perspective (Huang *et al.* 2005). The fusion image may be assessed by comparing it to the various classification findings. Abdikan *et al.* (2015) investigate the impact of fusion techniques on classification accuracies by comparing the Maximum Likelihood Classifier (MLC), K-nearest neighbours (K-NN), support vector machine (SVM) and Random Forest (RF) as statistical models. Neetu and Ray (2020) compared different fusion techniques and showed results visually, statistically, and through image classification for crop classification (Table 4).

4. Experimental result

This section provides a subjective, objective and comprehensive assessment of fusion techniques, including PCA, CN spectral sharpening, BT, and Gram–Schmidt methods. Here,

Table 5. Details of used satellite images.

C- Band SAR sentinel-1A data				
Acquisition Date	Mode	Orbit	Resolution (m)	Polarization
October 3 rd , 2017	Interferometry Mode	Ascending	5 X 20 m	VV and VH
Optical Sentinel-2A data				
Acquisition Date	Orbit	Resolution (m)	Bands selected	Cloud Cover (%)
Oct. 07 th , 2017	Ascending	10 m	Red, Green, Blue, NIR	0%

three bands (NIR, Red, and Green) of Sentinel-2 A multispectral images were fused with a VH polarized Sentinel-1A SAR image. Table 5 shows the details of data that have been utilised for experiments. The Agra-Mathura region of India has been taken as a study area. The climate of this region is dry. We performed calibration, speckle reduction, and terrain correction as SAR preprocessing steps. The optical image has been preprocessed for atmospheric correction. Later on, SAR and MS images were reprojected to the WGS-84 band.

4.1. Subjective assessment

The subjective (visual) findings of PCA, CN spectral sharpening, BT, and GS methods show that of all these methods tested, PCA fusion produces the best fused output (Figure 2).

The use of BT and GS fusion methods has resulted in significant spectrum distortion, as shown in the visual findings. The findings also show that some texture information accessible in the SAR data is fed into the fused output. The visual findings of CN spectral sharpening fusion show that the outcomes are visually superior to the GS and BT techniques. On visual inspection, some details of the CN technique are found to be sharpened than that in the PCA, GS and BT techniques. However, simple GS and CN-based methods have enough room for enhancement and result in spectral mismatch between SAR and optical image distortion.

4.2. Objective assessment

Tables 6 and 7 present the evaluation of some widely used objective metrics. All fused images are evaluated using metrics: CC, UIQI, SSIM, and Mean Bias require a reference image (Table 6), whereas Entropy, SD, and MI do not require a reference image (Table 7) as evaluated PCA and GS methods show better results. Among all considered methods, the performance of BT was significantly poor. The CC value of all fusions shows little difference. PCA and GS fusion methods result in better values of metrics. In the absence of high-resolution reference image, metrics are evaluated using the Wald protocol at reduced resolution (Wald *et al.* 1997).

4.3. Comprehensive assessment

Figure 3 shows the comprehensive assessment of fusion methods using SVM classification and the outcome of assessment in terms of kappa coefficient and accuracy is shown in

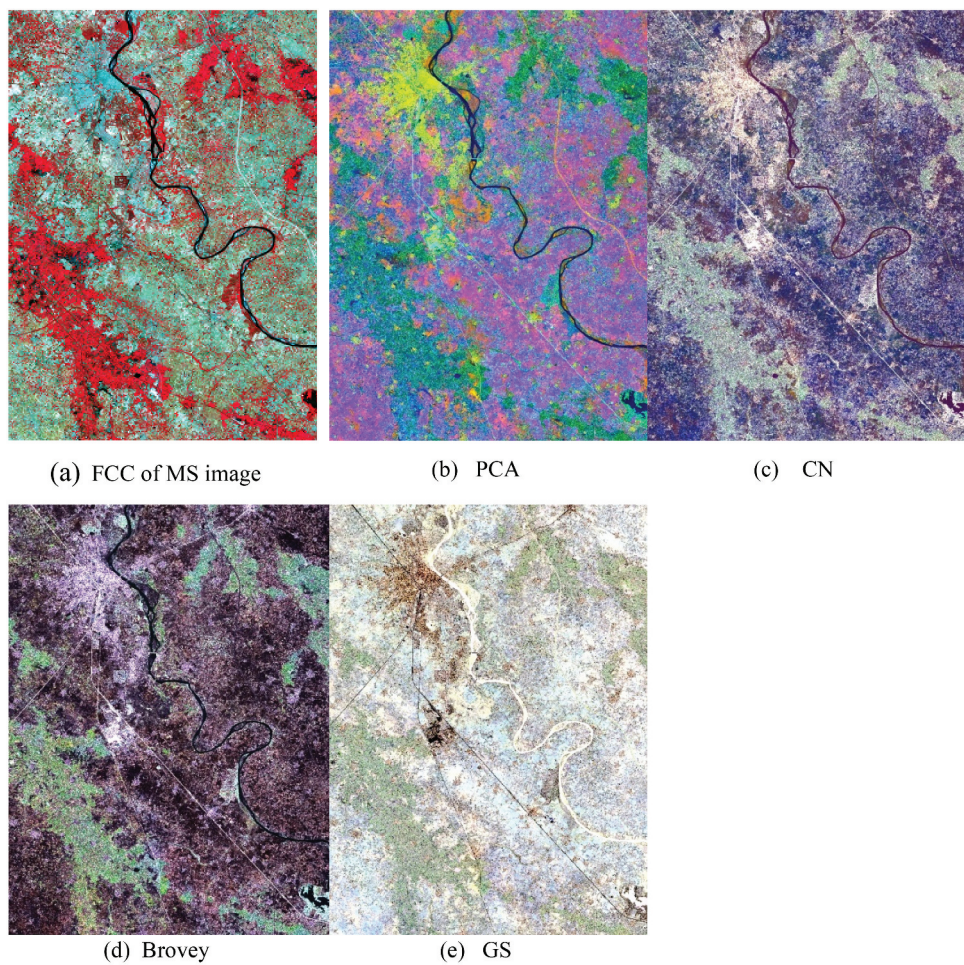


Figure 2. SAR-optical fusion of S1A and S2A.

Table 6. Experimental results of objective measures (reference data required).

Parameter	Fusion method	Fused images			Mean Value	Ideal value
		Green	Red	NIR		
CC	PCA	0.665	0.676	0.710	0.683	1
	GS	0.555	0.555	0.610	0.573	1
	BT	0.056	0.076	0.272	0.134	1
	CN	0.445	0.445	0.552	0.481	1
UIQI	PCA	−0.045	0.082	0.025	0.021	−1 to 1
	GS	−0.069	−0.055	0.210	0.029	−1 to 1
	BT	−0.049	−0.022	−0.028	−0.033	−1 to 1
	CN	−0.015	−0.045	0.030	−0.01	−1 to 1
SSIM	PCA	0.570	0.550	0.258	0.459	−1 to 1
	GS	0.462	0.380	0.376	0.406	−1 to 1
	BT	0.250	0.183	0.190	0.208	−1 to 1
	CN	0.444	0.418	0.276	0.379	−1 to 1
Mean Bias	PCA	0.260	0.220	0.240	0.240	0
	GS	0.000	0.001	0.002	0.001	0
	BT	0.816	0.803	0.595	0.738	0
	CN	0.304	0.266	0.188	0.249	0

Table 7. Experimental results of objective measures (reference data not required).

Parameter	Band	Initial Value	Fusion method			
			PCA	GS	BT	CN
Entropy	Green	4.520	4.750	5.455	4.002	4.650
	Red	4.725	4.730	4.995	3.990	4.854
	NIR	5.445	5.556	5.766	4.206	5.485
SD	Green	124.540	85.224	110.56	17.820	113.35
	Red	140.578	85.442	106.34	24.110	142.54
	NIR	154.660	162.57	168.577	72.340	168.54
MI	Green	9.486	9.210	8.824	4.324	9.256
	Red	9.540	9.846	9.276	4.540	9.664
	NIR	10.048	11.324	10.667	6.548	10.424

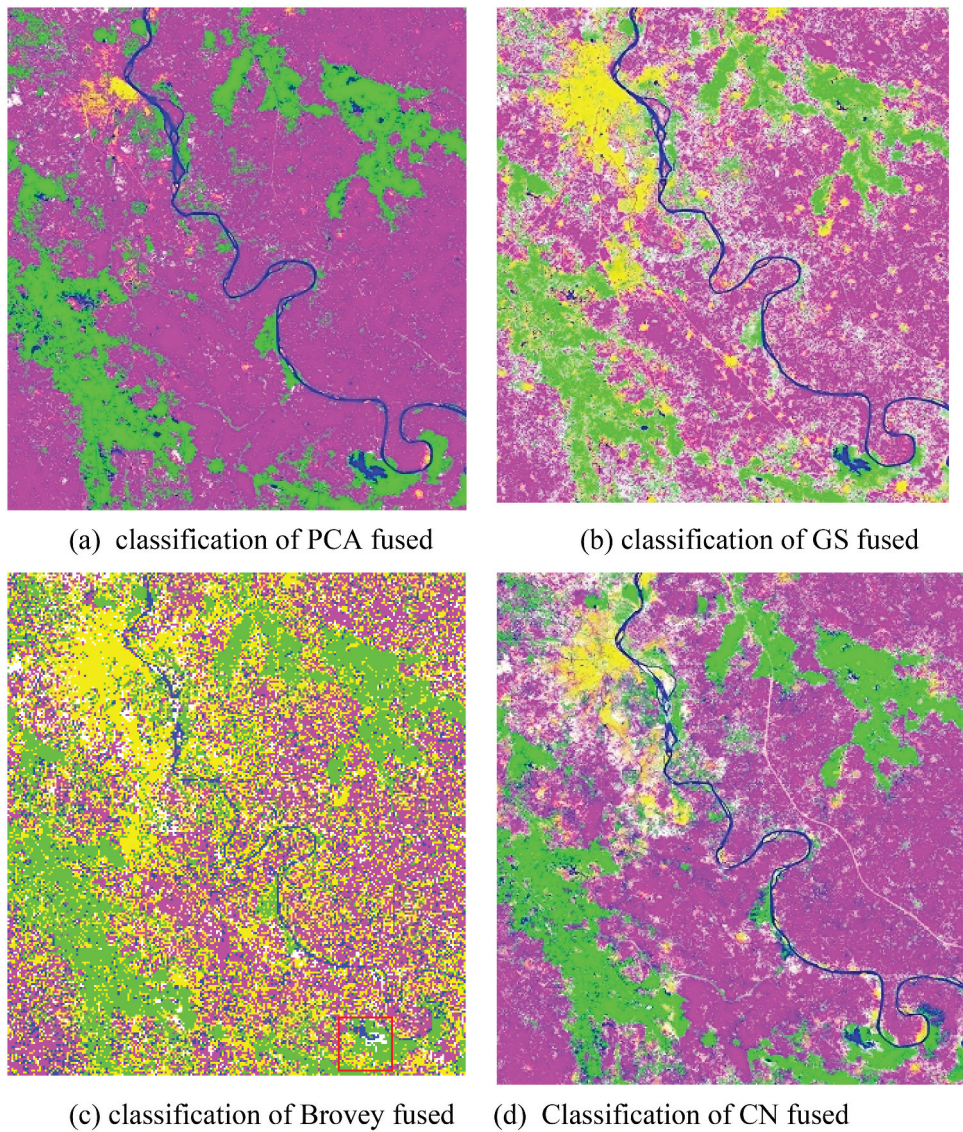


Figure 3. Comprehensive assessment of SAR-optical fusion with SVM classification.

Table 8. Details of classes.

Class	map colour	Training Pixel	Testing Pixel
Man-made Structure		963	780
Water		1025	916
Forest		1011	912
Barren land		1215	1065
Veg. Agriculture		1033	845

Table 9. Classification outcome evaluation of SAR-optical image fusion method.

		Class (percentage)					Overall Accuracy	Kappa Coeff.
	Accuracy	Settlement	Barrenland	Water	Forest	Veg. agriculture		
SAR only	Prod Acc.	70.21	23.72	65.62	38.55	67.32	39.7121	0.2666
	User Acc.	73.62	52.33	78.06	79.42	45.73		
MS only	Prod Acc.	93.45	87.19	95.34	74.48	66.52	78.1531	0.6835
	User Acc.	75.66	81.24	92.89	81.70	72.80		
PCA	Prod Acc.	71.46	99.81	100	46.27	87.46	82.6276	0.7817
	User Acc.	86.90	96.50	96.76	99.74	49.21		
GS	Prod Acc.	96.66	97.49	65.84	50.48	86.52	79.1190	0.7175
	User Acc.	73.73	83.34	90.89	91.70	62.80		
BT	Prod Acc.	90.63	3.72	5.62	44.35	89.50	39.4871	0.2778
	User Acc.	77.82	43.01	53.06	89.56	20.70		
CN	Prod Acc.	96.67	92.01	100	39.66	89.50	83.3521	0.7926
	User Acc.	81.53	84.76	96.86	77.10	69.04		

Table 10. CPU processing time for fusion methods (Intel i5, 8GB RAM).

	PCA	GS	BT	CN
Processing Time	30 min 50 sec	29 min 05 sec	02 min 30 sec	30 min 15 sec

Table 9. LULC classes (man-made structure, water, forest, barren land and veg. agriculture) have been included in the experiment (**Table 8**). The CN fusion provides a maximum accuracy of 83.3521% and BT gives the worst accuracy (**Table 9**).

4.4. Computational cost

Computational cost is nowadays a crucial part of analysing the performance of fusion methods. Here, we compute the processing time of each fusion method. Even though it is vital to compare in terms of computational costs (processing time, storage), it is difficult as the approaches depended on processing parameters. In this study, the same data set and pre-processing steps have been used for all algorithms. In the pixel-level approach, the three bands of optical images have been fused with a VH polarised SAR image. **Table 10** shows the processing time for the examined fusion methods. Therefore, based on computational costs, the BT fusion had the lowest processing time, followed by GS. The results show a trade-off between accuracy and computational cost.

5. Applications

Various applications have been summarised in **Table 11**.

Table 11. Application of SAR-optical fusion.

SN	Application	Details with Reference
1.	Urban	Fusion of SAR images (sensitive to textural variations) with the optical image is beneficial for taking out urban characteristics. Fusion enhanced the urban features and classification of urban land classes (Lopes <i>et al.</i> 1990, Tupin and Roux 2003, Huang <i>et al.</i> 2005, Amarsaikhana <i>et al.</i> 2010, Werner <i>et al.</i> 2014, Salentinig and Gamba (2015). Facheris <i>et al.</i> (2004) demonstrate proper spectral preservation on agriculture areas, bare-soil, and textured regions (road networks and buildings). Garzelli <i>et al.</i> (2002) exhibit the fusion capability to inject 'urban' information without affecting the spatial resolution of the optical data.
2.	LULC	Fusion aids in distinguishing various kinds of LULC classes that are indistinguishable in optical image owing to comparable spectral features of landcover objects (Dupas 2000, Herold and Haack 2002, Abdikan <i>et al.</i> 2015, Gibril <i>et al.</i> 2017). Parihar <i>et al.</i> (2017) evaluate and find that fusion of cross-polarised SAR image with optical image significantly enhance LULC class accuracy. Gaetano <i>et al.</i> (2017) exhibit that optical-driven speckled SAR image vastly improves class accuracy w.r.t the reference image and even multitemporal image.
3.	Change Detection	Multitemporal and multisensor images recorded for a certain region of the earth offer complementary object details (Gibril <i>et al.</i> 2017). Zhang <i>et al.</i> (2010) present the outcomes of multitemporal SAR and optical data fusion algorithms for land-cover change detection and hard- and soft-decision-based change detection. Pal <i>et al.</i> (2007) found that change detection helps locate the zones of likely changes between the lithological interpretations and the published geological map.
4.	Forest Mapping	Fusion is a useful technique in forest mapping because of the complementary information given by SAR and optical sensors (Moghaddam <i>et al.</i> 2002, Cutler <i>et al.</i> 20122012).Veerabhadraswamy <i>et al.</i> (2021) demonstrate that RF classification of fuse images improves the forest and non-forest mapping by 9.15% and 13.50%, compared to optical and SAR.
5.	Flood mapping	Because of its sensitivity to the dielectric constant of soil, a SAR signal may detect soil dampness. Optical data may reveal the soil's pre-flood condition. As a result, combining SAR and MS images is beneficial for flood mapping. Quang <i>et al.</i> (2019) show that fusing optical and SAR images reduces the influences of clouds and cloud shadows in optical data. It adopts the method of water extraction model for rapid flood mapping. Tong <i>et al.</i> (2018) used a map difference method for a large-scale flood inundation mapping and proposed an approach based on SVM and the active contour technique.
6.	Topographic map	A sensor may cover a region that is uncovered by another sensor. Due to the availability of SAR data under all weather conditions, information may be accessible with SAR data. The combination of SAR and optical images is beneficial for revamping topographic mapping (Zhang 2010). Hammam <i>et al.</i> (2020) exhibited the superior performance of SAR image for surface and near-surface structural delineation in semi-flat terrains of arid environments using its penetration capability.

6. Conclusions

The recent research and launch of SAR sensors (specially ESA's Sentinel-1 and Sentinel-2) opens several new trends in SAR-optical fusion in many dimensions. Since SAR and optical images provide complementary details, this kind of fusion is the way forward for a variety of applications, despite the challenges (spatial/spectral distortions, computational complexity and misregistration).

The following findings regarding the five major PLF methods are drawn from the provided literature review. This paper examines fusion techniques based on PCA, BT, GS and CN spectral sharpening (CN) PLF. Fusion techniques based on CS are less computationally demanding and easier to implement. Numerical-based methods are less complex

but offer serious spectral and spatial distortions. MRA methods reduce spatial and spectral distortions. Model-based methods have greater computational complexity but outperform them in terms of performance. Hybrid techniques combine the benefits of CS and MRA methodologies, resulting in a superior fused outcome than separate methods. Along with various fusion techniques, three major types of fusion assessment are described here for evaluating fusion performance. The visual effect is subjective in nature and varies from researcher to researcher. Comprehensive assessment shows promising results here and this review paper shows the experimental outcomes of PCA, CN, GS and BT fused images. With the launch of new SAR satellites with higher resolution, SAR and optical fusion remains an active research field that will be useful for a wide range of remote sensing applications. Recent findings indicate that SAR-optical PLF development is moving towards deep learning, big data, and cloud computing.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- Abdikan, S., *et al.*, 2015. Enhancing land use classification with fusing dual-polarized TerraSAR-X and multispectral RapidEye data. *Journal of Applied Remote Sensing*, 9 (1), 096054. doi:[10.1117/1.jrs.9.096054](https://doi.org/10.1117/1.jrs.9.096054).
- Abdikan, S., and Sanli, F.B., 2012. Comparison of different fusion algorithms in urban and agricultural areas using sar (palsar and radarsat) and optical (spot) images. *Bol. Ciênc. Geod.*, 18 (4), 509–531. doi:[10.1590/S1982-21702012000400001](https://doi.org/10.1590/S1982-21702012000400001).
- Ali, M.Z., Qazi, W., and Aslam, N., 2018. A comparative study of ALOS-2 PALSAR and landsat-8 imagery for land cover classification using maximum likelihood classifier. *The Egyptian Journal of Remote Sensing and Space Science*, 21 S29–S35. doi:[10.1016/j.ejrs.2018.03.003](https://doi.org/10.1016/j.ejrs.2018.03.003).
- Amarsaikhan, D., *et al.*, 2011. Applications of GIS and very high-resolution RS data for urban land use change studies in Mongolia. *International Journal of Navigation and Observation*. doi:[10.1155/2011/314507](https://doi.org/10.1155/2011/314507).
- Amarsaikhana, D., *et al.*, 2010. Fusing high-resolution SAR and optical imagery for improved urban land cover study and classification. *International Journal of Image and Data Fusion*, 1 (1), 83–97. doi:[10.1080/19479830903562041](https://doi.org/10.1080/19479830903562041).
- Bermudez, J.D., *et al.*, 2019. Synthesis of Multispectral Optical Images from SAR/Optical Multitemporal Data Using Conditional Generative Adversarial Networks. *IEEE Geosci. Remote Sensing Lett.*, 16(8), 1220–1224. doi:[10.1109/LGRS.2019.2894734](https://doi.org/10.1109/LGRS.2019.2894734).
- Byun, Y., Choi, J., and Han, Y., 2013. An area-based image fusion scheme for the integration of SAR and optical satellite imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6 (5), 2212–2220. doi:[10.1109/JSTARS.2013.2272773](https://doi.org/10.1109/JSTARS.2013.2272773).
- Chandrakanth, R., *et al.*, 2014. A novel image fusion system for multisensor and multiband remote sensing data. *IETE Journal of Research*, 60 (2), 168–182. doi:[10.1080/03772063.2014.914697](https://doi.org/10.1080/03772063.2014.914697).
- Chandrakanth, R., *et al.*, 2011. Feasibility of high resolution SAR and multispectral data fusion. *Int. Geosci. Remote Sens. Symp.*, 356–359. doi:[10.1109/IGARSS.2011.6048972](https://doi.org/10.1109/IGARSS.2011.6048972).
- Chen, S., *et al.*, 2010. SAR and multispectral image fusion using generalized IHS transform based on à Trouis Wavelet and EMD decompositions. *IEEE Sensors Journal*, 10 (3), 737–745. doi:[10.1109/JSEN.2009.2038661](https://doi.org/10.1109/JSEN.2009.2038661).
- Chen, C., *et al.*, 2020. A pixel-level fusion method for multi-source optical remote sensing image combining the principal component analysis and curvelet transform. *Earth Sci Inform*, 13 (4), 1005–1013. doi:[10.1007/s12145-020-00472-7](https://doi.org/10.1007/s12145-020-00472-7).

- Chibani, Y., 2006. Additive integration of SAR features into multispectral SPOT images by means of the à trous wavelet decomposition. *ISPRS Journal of Photogrammetry and Remote Sensing*, 60 (5), 306–314. doi:[10.1016/j.isprsjprs.2006.05.001](https://doi.org/10.1016/j.isprsjprs.2006.05.001)
- Chikr El-Mezouar, M., et al., 2011. An IHS-Based Fusion for Color Distortion Reduction and Vegetation Enhancement in IKONOS Imagery. *IEEE Trans. Geosci. Remote Sensing*, 49 (5), 1590–1602. doi:[10.1109/TGRS.2010.2087029](https://doi.org/10.1109/TGRS.2010.2087029)
- Clerici, N., Valbuena Calderón, C.A., and Posada, J.M., 2017. Fusion of Sentinel-1A and Sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, Colombia. *Journal of Maps*, 13 (2), 718–726. doi:[10.1080/17445647.2017.1372316](https://doi.org/10.1080/17445647.2017.1372316).
- Cutler, M., et al. 2012. Estimating tropical forest biomass with a combination of SAR image texture and Landsat TM data: An assessment of predictions between regions. *Isprs Journal of Photogrammetry and Remote Sensing*, 70, 66–77. doi:[10.1016/j.isprsjprs.2012.03.011](https://doi.org/10.1016/j.isprsjprs.2012.03.011).
- Dahiya, S., Garg, P.K., and Jat, M.K., 2013. A comparative study of various pixel-based image fusion techniques as applied to an urban environment. *International Journal of Image and Data Fusion*, 4 (3), 197–213. doi:[10.1080/19479832.2013.778335](https://doi.org/10.1080/19479832.2013.778335)
- Dawn, S., Saxena, V., and Sharma, B., 2010. Remote Sensing Image Registration Techniques: A Survey. *Image and Signal Processing. Springer Berlin Heidelberg*, 6134 (c), 103–112.
- Deledalle, C.A., et al., 2017. MuLoG, or how to apply gaussian denoisers to multi-channel SAR speckle reduction? *IEEE Transactions on Image Processing*, 26, 4389–4403. doi:[10.1109/TIP.2017.2713946](https://doi.org/10.1109/TIP.2017.2713946)
- Ehlers, M., et al. 2010. Multi-sensor image fusion for pansharpening in remote sensing. *International Journal of Image and Data Fusion*, 1 (1), 25–45. doi:[10.1080/19479830903561985](https://doi.org/10.1080/19479830903561985).
- El-Deen Taha, L.G. and Elbeih, S.F., 2010. Investigation of fusion of SAR and Landsat data for shoreline super resolution mapping: the northeastern Mediterranean sea coast in Egypt. *Applied Geomatics*, 2, 177–186. doi:[10.1007/s12518-010-0033-x](https://doi.org/10.1007/s12518-010-0033-x)
- El-Tawel, G.S., and Helmy, A.K., 2014. Fusion of Multispectral and Full Polarimetric SAR Images in NSST Domain. *International Journal of Image Processing*, 8 (6), 497–513.
- Facheris, L., et al., 2004. Fusion of multispectral and SAR images by intensity modulation nefocast view project ANISAP view project fusion of multispectral and SAR images by intensity modulation.
- France, F.-N., 2017. Sar image despeckling through convolutional neural networks G . Chierchia Universit ´e Paris Est LIGM UMR 8049, CNRS, ESIEE Paris D . Cozzolino, G . Poggi, L . Verdoliva DIETI 5438–5441.
- Gaetano, R., et al., 2017. Fusion of sar-optical data for land cover monitoring. *Int. Geosci. Remote Sens. Symp*, 5470–5473. doi: [10.1109/IGARSS.2017.8128242](https://doi.org/10.1109/IGARSS.2017.8128242).
- Garzelli, A., 2002. Wavelet-Based Fusion of Optical and Sar Image Data Over Urban Area. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* Vol. 34, Issue 3/B, *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* CANADA 59–62.
- Ghahremani, M. and Ghassemian, H., 2016. A compressed-sensing-based pan-sharpening method for spectral distortion reduction. *IEEE Transactions on Geoscience and Remote Sensing*, 54 (4), 2194–2206. doi:[10.1109/TGRS.2015.2497309](https://doi.org/10.1109/TGRS.2015.2497309)
- Ghamisi, P., et al., 2019. Multisource and multitemporal data fusion in remote sensing: a comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7, 6–39. doi:[10.1109/MGRS.2018.2890023](https://doi.org/10.1109/MGRS.2018.2890023)
- Ghassemian, H., 2016. A review of remote sensing image fusion methods. *Information Fusion*, 32, 75–89. doi:[10.1016/j.inffus.2016.03.003](https://doi.org/10.1016/j.inffus.2016.03.003)
- Gibril, M.B., et al., 2017. Fusion of RADARSAT-2 and multispectral optical remote sensing data for LULC extraction in a tropical agricultural area. *Geocarto international*, 32 (7), 735–748. doi:[10.1080/10106049.2016.1170893](https://doi.org/10.1080/10106049.2016.1170893).
- Grohnfeldt, C., Schmitt, M., and Zhu, X., 2018. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. *International Geoscience and Remote Sensing Symposium (IGARSS)*, july-2018, 1726–1729.

- Han, N., Hu, J., and Zhang, W., 2010. Multi-spectral and SAR images fusion via Mallat and à trous wavelet transform. In: *2010 18th International Conference on Geoinformatics*. doi:[10.1109/GEOINFORMATICS.2010.5567653](https://doi.org/10.1109/GEOINFORMATICS.2010.5567653).
- Harris, J.R., Murray, R., and Hirose, T., 1990. IHS transform for the integration of radar imagery with other remotely sensed data. *Photogrammetric Engineering & Remote Sensing*, 56, 1631–1641.
- Herold, N.D. and Haack, B.N., 2002. Fusion of radar and optical data for land cover mapping. *Geocarto International*, 17 (2), 21–30. doi:[10.1080/10106040208542232](https://doi.org/10.1080/10106040208542232)
- Hong, G., Zhang, Y., and Mercer, B., 2009. A wavelet and IHS integration method to fuse high resolution SAR with moderate resolution multispectral images. *Photogrammetric Engineering & Remote Sensing*, 75 (10), 1213–1223. doi:[10.14358/PERS.75.10.1213](https://doi.org/10.14358/PERS.75.10.1213)
- Huang, Y., et al., 2005. The fusion of multispectral and SAR images based wavelet transformation over urban area. *International Geoscience and Remote Sensing Symposium*, 6, 3942–3944. doi:[10.1109/IGARSS.2005.1525774](https://doi.org/10.1109/IGARSS.2005.1525774)
- Huang, B., et al., 2015. Cloud Removal from Optical Satellite Imagery with SAR Imagery Using Sparse Representation. *IEEE Geosci. Remote Sensing Lett.*, 12(5), 1046–1050. doi:[10.1109/LGRS.2014.2377476](https://doi.org/10.1109/LGRS.2014.2377476).
- Huang, L. and Li, Z., 2010. Feature-based image registration using the shape context. *International Journal of Remote Sensing*, 31 (8), 2169–2177. doi:[10.1080/01431161003621585](https://doi.org/10.1080/01431161003621585)
- Inglada, J. and Giros, A., 2004. On the possibility of automatic multisensor image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 42 (10), 2104–2120. doi:[10.1109/TGRS.2004.835294](https://doi.org/10.1109/TGRS.2004.835294)
- Jagalingam, P. and Hegde, A.V., 2015. A review of quality metrics for fused image. *Aquatic Procedia*, 4, 133–142. doi:[10.1016/j.aqpro.2015.02.019](https://doi.org/10.1016/j.aqpro.2015.02.019)
- Khosravi, M.R., et al., 2017. MRF-Based Multispectral Image Fusion Using an Adaptive Approach Based on Edge-Guided Interpolation. *JGIS*, 9 (2), 114–125. doi:[10.4236/jgis.2017.92008](https://doi.org/10.4236/jgis.2017.92008).
- Klonus, S. and Ehlers, M., 2007. Image fusion using the Ehlers spectral characteristics preservation algorithm. *GIScience & Remote Sensing*, 44 (2), 93–116. doi:[10.2747/1548-1603.44.2.93](https://doi.org/10.2747/1548-1603.44.2.93)
- Kulkarni, S.C. and Rege, P.P., 2019. Fusion of RISAT-1 SAR Image and Resourcesat-2 Multispectral Images Using Wavelet Transform. *6th Int. Conf. Signal Process. Integr. Networks, SPIN 2019*, 45–52. doi:[10.1109/SPIN.2019.8711589](https://doi.org/10.1109/SPIN.2019.8711589).
- Kulkarni, S.C. and Rege, P.P., 2020. Pixel level fusion techniques for SAR and optical images: A review. *Information Fusion*, 59, 13–29. doi:[10.1016/j.inffus.2020.01.003](https://doi.org/10.1016/j.inffus.2020.01.003).
- Kulkarni, S.C. and Rege, P.P., 2021. Application of Taguchi method to improve land use land cover classification using PCA-DWT-based SAR-multispectral image fusion. *J. Appl. Rem. Sens.*, 15 (1). doi:[10.1117/1.JRS.15.014509](https://doi.org/10.1117/1.JRS.15.014509).
- Kumar, M., et al., 2017. Study of mangrove communities in Marine National Park and Sanctuary, Jamnagar, Gujarat, India, by fusing RISAT-1 SAR and resourcesat-2 LISS-IV images. *International Journal of Image and Data Fusion*, 8 (1), 73–91. doi:[10.1080/19479832.2016.1232755](https://doi.org/10.1080/19479832.2016.1232755)
- Li, S., Hong, R., and Wu, X., 2008. A novel similarity based quality metric for image fusion. *ICALIP 2008 - 2008 Int. Conf. Audio, Lang. Image Process. Proc.*, 167–172. doi:[10.1109/ICALIP.2008.4589989](https://doi.org/10.1109/ICALIP.2008.4589989).
- Li, H., Manjunath, B.S., and Mitra, S.K., 1995. A contour-based approach to multisensor image registration. *IEEE Transactions on Image Processing*, 4 (3), 320–334. doi:[10.1109/83.366480](https://doi.org/10.1109/83.366480)
- Liu, J., et al., 2015. Human visual system consistent quality assessment for remote sensing image fusion. *Isprs Journal of Photogrammetry and Remote Sensing*, 105 79–90. doi:[10.1016/j.isprsjprs.2014.12.018](https://doi.org/10.1016/j.isprsjprs.2014.12.018).
- Liu, C., Qi, Y., and Ding, W., 2016. Airborne SAR and optical image fusion based on IHS transform and joint non-negative sparse representation. In: *International Geoscience and Remote Sensing Symposium*, 7196–7199. November. doi:[10.1109/IGARSS.2016.7730877](https://doi.org/10.1109/IGARSS.2016.7730877).
- Lopes, A., Touzi, R., and Nezry, E., 1990. Adaptive speckle filters and scene heterogeneity. *IEEE Transactions on Geoscience and Remote Sensing*, 28 (6), 992–1000. doi:[10.1109/36.62623](https://doi.org/10.1109/36.62623)

- Lu, Y., et al., 2011. SAR and MS image fusion based on curvelet transform and activity measure. In: *2011 International Conference on Electric Information and Control Engineering (ICEICE 2011)-Proceedings*. 1680–1683. doi:10.1109/ICEICE.2011.5777893.
- Luo, D., et al., 2014. Fusion of high spatial resolution optical and polarimetric SAR images for urban land cover classification. In: *3rd International Workshop on Earth Observation and Remote Sensing Applications (EORSA 2014) – Proceedings*, 362–365. doi:10.1109/EORSA.2014.6927913.
- Manaf, S.A., et al., 2016. Comparison of classification techniques on fused optical and SAR images for shoreline extraction: a case study at Northeast Coast of Peninsular Malaysia. *Journal of Computer Science*, 12, 399–411. doi:10.3844/jcssp.2016.399.411
- Manakos, I., Kordelas, G.A., and Marini, K., 2020. Fusion of Sentinel-1 data with Sentinel-2 products to overcome non-favourable atmospheric conditions for the delineation of inundation maps. *European Journal of Remote Sensing*, 53 (sup2), 53–66. doi:10.1080/22797254.2019.1596757.
- Meijie, L., et al., 2015. PCA-based sea-ice image fusion of optical data by HIS transform and SAR data by wavelet transform. *Acta Oceanologica Sinica*, 34 (3), 59–67. doi:10.1007/s13131-015.
- Meraner, A., et al., 2020. Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. *Isprs Journal of Photogrammetry and Remote Sensing*, 166, 333–346. doi:10.1016/j.isprsjprs.2020.05.013.
- Merkle, N., et al., 2015. A new approach for optical and sar satellite image registration. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Copernicus GmbH, 119–126. doi:10.5194/isprsannals-II-3-W4-119-2015.
- Misra, I., et al., 2012. An efficient algorithm for automatic fusion of RISAT-1 SAR data and Resourcesat-2 optical images. In: *4th International Conference on Intelligent Human Computer Interaction: Advancing Technology for Humanity, IHCI 2012*. doi:10.1109/IHCI.2012.6481838.
- Moghaddam, M., Dungan, J., and Acker, S., 2002. Forest variable estimation from fusion of SAR and multispectral optical data. *IEEE Trans. Geosci. Remote Sensing*, 40 (10), 2176–2187. doi:10.1109/TGRS.2002.804725.
- Neetu, N. and Ray, S.S., 2020. Evaluation of different approaches to the fusion of Sentinel –1 SAR data and Resourcesat 2 LISS III optical data for use in crop classification. *Remote Sensing Letters*, 11 (12), 1157–1166. doi:10.1080/2150704X.2020.1832278
- Pal, S., Majumdar, T., and Bhattacharya, A.K., 2007. ERS-2 SAR and IRS-1C LISS III data fusion: A PCA approach to improve remote sensing based geological interpretation. *Isprs Journal of Photogrammetry and Remote Sensing*, 61 (5), 281–297. doi:10.1016/j.isprsjprs.2006.10.001.
- Palmann, C., Mavromatis, S., and Sequeira, J., 2008. Sar image registration using a new approach based on the generalized Hough transform. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37, 145–152.
- Parcharidis, I., Kazi-Tani, L.M., 2000. Landsat TM and ERS data fusion: A statistical approach evaluation for four different methods. *Int. Geosci. Remote Sens. Symp.* 5, 2120–2122. <https://doi.org/10.1109/igarss.2000.858315>
- Parihar, N., Rathore, V., and Mohan, S., 2017. Combining ALOS PALSAR and AVNIR-2 data for effective land use/land cover classification in Jharia coalfields region. *International Journal of Image and Data Fusion*, 8(2), 130–147. doi:10.1080/19479832.2016.1273258.
- Pohl, C., 1999. Tools and methods for fusion of images of different spatial resolution. *Archives*, 3–4.
- Pohl, C. and van Genderen, J.L., 1995. Image fusion of microwave and optical remote sensing data for topographic map updating in the tropics. *Image and Signal Processing for Remote Sensing II*, 2579, 2–10. doi:10.1117/12.226824
- Pohl, C. and Van Genderen, J.L., 1998. Review article Multisensor image fusion in remote sensing: concepts, methods and applications. *International Journal of Remote Sensing*, 19 (5), 823–854. doi:10.1080/014311698215748
- Pohl, C. and van Genderen, J., 2014. Remote sensing image fusion: an update in the context of Digital Earth. *International Journal of Digital Earth*, 7 (2), 158–172. doi:10.1080/17538947.2013.869266
- Pohl, C. and van Genderen, J., 2015. Structuring contemporary remote sensing image fusion. *International Journal of Image and Data Fusion*, 6 (1), 3–21. doi:10.1080/19479832.2014.998727

- Quan, Y., *et al.*, 2020. A Novel Image Fusion Method of Multi-Spectral and SAR Images for Land Cover Classification. *Remote Sensing*, 12(22), 3801. doi:[10.3390/rs12223801](https://doi.org/10.3390/rs12223801).
- Quang, N.H., *et al.*, 2019. Synthetic aperture radar and optical remote sensing image fusion for flood monitoring in the Vietnam lower Mekong basin: a prototype application for the Vietnam Open Data Cube. *European Journal of Remote Sensing*, 52(1), 599–612. doi:[10.1080/22797254.2019.1698319](https://doi.org/10.1080/22797254.2019.1698319).
- Qu, G., Zhang, D., and Yan, P., 2002. Information measure for performance of image fusion. *Electronics letters*, 38, (7), 313. doi:[10.1049/el:20020212](https://doi.org/10.1049/el:20020212).
- Rahman, M.M., Sumantyo, J.T.S., and Sadek, M.F., 2010. Microwave and optical image fusion for surface and sub-surface feature mapping in eastern Sahara. *International Journal of Remote Sensing*, 31 (20), 5465–5480. doi:[10.1080/01431160903302999](https://doi.org/10.1080/01431160903302999)
- Ranchin, T. and Wald, L., 2000. Fusion of high spatial and spectral resolution images: the ARSIS concept and its implementation. *Photogrammetric Engineering & Remote Sensing*, 66, 49–61.
- Riedel, T., Thiel, C., and Schmullius, C., 2007. Fusion of optical and SAR satellite data for improved land cover mapping in agricultural areas. *Eur. Sp. Agency*, (Special Publ. ESA SP).
- Rusmini, M., *et al.*, 2012. High-resolution SAR and high-resolution optical data integration for sub-urban land-cover classification. *Int. Geosci. Remote Sens. Symp.*, 4986–4989. doi:[10.1109/IGARSS.2012.6352492](https://doi.org/10.1109/IGARSS.2012.6352492).
- Salentinig, A., and Gamba, P., 2015. Combining SAR-Based and Multispectral-Based Extractions to Map Urban Areas at Multiple Spatial Resolutions. *IEEE Geosci. Remote Sens. Mag.*, 3 (3), 100–112. doi:[10.1109/MGRS.2015.2430874](https://doi.org/10.1109/MGRS.2015.2430874)
- Schmidt, J., *et al.*, 2018. Synergetic use of Sentinel-1 and Sentinel-2 for assessments of heathland conservation status. *Remote Sensing in Ecology and Conservation*, 4 (3), 225–239. doi:[10.1002/rse2.68](https://doi.org/10.1002/rse2.68)
- Shah, E., Jayaprasad, P., and James, M.E., 2019. Image Fusion of SAR and Optical Images for Identifying Antarctic Ice Features. *J Indian Soc Remote Sens*, 47(12), 2113–2127. doi:[10.1007/s12524-019-01040-3](https://doi.org/10.1007/s12524-019-01040-3).
- Shakya, A., Biswas, M., and Pal, M., 2020. CNN-based fusion and classification of SAR and Optical data. *International Journal of Remote Sensing*, 41(22), 8839–8861. doi:[10.1080/01431161.2020.1783713](https://doi.org/10.1080/01431161.2020.1783713).
- Shao, Z., Wu, W., and Guo, S., 2020. IHS-GTF: A Fusion Method for Optical and Synthetic Aperture Radar Data. *Remote Sensing*, 12(17), 2796. doi:[10.3390/rs12172796](https://doi.org/10.3390/rs12172796).
- Sheoran, A. and Haack, B., 2014. Optical and radar data comparison and integration: Kenya example. *Geocarto International*, 29 (4), 370–382. doi:[10.1080/10106049.2013.769027](https://doi.org/10.1080/10106049.2013.769027)
- Simone, G., *et al.*, 2002. Image fusion techniques for remote sensing applications. *Information Fusion*, 3 (1), 3–15. doi:[10.1016/S1566-2535\(01\)00056-2](https://doi.org/10.1016/S1566-2535(01)00056-2)
- Tong, X., *et al.*, 2018. An approach for flood monitoring by the combined use of Landsat 8 optical imagery and COSMO-SkyMed radar imagery. *Isprs Journal of Photogrammetry and Remote Sensing*, 136, 144–153. doi:[10.1016/j.isprsjprs.2017.11.006](https://doi.org/10.1016/j.isprsjprs.2017.11.006).
- Tu, T.M., *et al.*, 2004. A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Geoscience and Remote Sensing Letters*, 1, 309–312. doi:[10.1109/LGRS.2004.834804](https://doi.org/10.1109/LGRS.2004.834804)
- Tupin, F. and Roux, M., 2003. Detection of building outlines based on the fusion of SAR and optical features. *Isprs Journal of Photogrammetry and Remote Sensing*, 58 (1–2), 71–82. doi:[10.1016/S0924-2716\(03\)00018-2](https://doi.org/10.1016/S0924-2716(03)00018-2).
- Veerabhadraswamy, N., Devagiri, G.M., and Khaple, A.K., 2021. Fusion of Complementary Information of SAR and Optical Data for Forest Cover Mapping using Random Forest Algorithm. *Current science*, 120(1), 193. doi:[10.18520/cs/v120/i1/193-199](https://doi.org/10.18520/cs/v120/i1/193-199).
- Wakabayashi, H. and Arai, K., 1996. A method of speckle noise reduction for SAR data. *International Journal of Remote Sensing*, 17 (10), 1837–1849. doi:[10.1080/01431169608948742](https://doi.org/10.1080/01431169608948742)
- Wald, L., Ranchin, T., and Mangolini, M., 1997. Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images. *Photogrammetric Engineering & Remote Sensing*, 63, 691–699.
- Wang, Z., *et al.*, 2005. A comparative analysis of image fusion methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43 (6), 1391–1402. doi:[10.1109/TGRS.2005.846874](https://doi.org/10.1109/TGRS.2005.846874)

- Wang, X.L. and Chen, C.X., 2016. Image fusion for synthetic aperture radar and multispectral images based on sub-band modulated non-subsampled contourlet transform and pulse coupled neural network methods. *The Imaging Science Journal*, 64, 87–93. doi:[10.1080/13682199.2015.1136101](https://doi.org/10.1080/13682199.2015.1136101)
- Wang, P., Zhang, H., and Patel, V.M., 2017. SAR image despeckling using a convolutional neural network. *IEEE Signal Processing Letters*, 24 (12), 1763–1767. doi:[10.1109/LSP.2017.2758203](https://doi.org/10.1109/LSP.2017.2758203)
- Werner, A., Storie, C.D., and Storie, J., 2014. Evaluating SAR-Optical Image Fusions for Urban LULC Classification in Vancouver Canada. *Canadian Journal of Remote Sensing*, 40(4), 278–290. doi:[10.1080/07038992.2014.976700](https://doi.org/10.1080/07038992.2014.976700).
- Xie, H., Pierce, L.E., and Ulaby, F.T., 2002. SAR speckle reduction using wavelet denoising and Markov random field modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 40 (10), 2196–2212. doi:[10.1109/TGRS.2002.802473](https://doi.org/10.1109/TGRS.2002.802473)
- You, C., et al., 2014. An Objective Quality Metric for Image Fusion based on Mutual Information and Multi-scale Structural Similarity. *JSW*, 9 (4). doi:[10.4304/jsw.9.4.1050-1054](https://doi.org/10.4304/jsw.9.4.1050-1054).
- Zhang, J., 2010. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1 (1), 5–24. doi:[10.1080/19479830903561035](https://doi.org/10.1080/19479830903561035)
- Zhang, R., et al., 2020. A novel feature-level fusion framework using optical and SAR remote sensing images for land use/land cover (LULC) classification in cloudy mountainous area. *Applied Sciences*, 10. doi:[10.3390/APP10082928](https://doi.org/10.3390/APP10082928)
- Zhang, W. and Yu, L. 2010. SAR and Landsat ETM+ image fusion using variational model. *International Conference on Computer and Communication Technologies in Agriculture Engineering*, 3 (205–207).
- Zheng, Y., et al., 2007. A new metric based on extended spatial frequency and its application to DWT based fusion algorithms. *Information Fusion*, 8 (2), 177–192. doi:[10.1016/j.inffus.2005.04.003](https://doi.org/10.1016/j.inffus.2005.04.003).
- Zhouping, Y., 2015. Fusion Algorithm of Optical Images and SAR with SVT and Sparse Representation. *International Journal on Smart Sensing and Intelligent Systems*, 8(2), 1123–1141. doi:[10.21307/ijssis-2017-799](https://doi.org/10.21307/ijssis-2017-799).

Seismic Analysis of Sagging Elasto-flexible Cable using Placement Model

Pankaj Kumar ^{1,*}, Sanjay Tiwari ², S.K. Jain ², Ritu Raj ³

¹Department of Civil Engineering, Assistant Professor, Madhav Institute of Technology and Science Gwalior 474005, India

²Department of Civil Engineering, Professor, Madhav Institute of Technology and Science Gwalior 474005, India

³Department of Civil Engineering, Assistant Professor, Delhi Technological University, Delhi 110042, India

Paper ID - 060401

Abstract

Weightless sagging elasto-flexible cables lack unique natural state. However, heavy cables are assumed their natural state under their self-weight for predicting their static and dynamic response. In most of the existing literatures the nodal coordinates or placements are generally defined in reference to the chosen Cartesian coordinate system for the discrete formulation called here Placement Model. In this paper, Placement Model is applied to predict the seismic response of weightless sagging cables. The cable is fixed at their both the ends and two masses are attached in its intermediate points leading to a 2-node 4-DOF system. The dynamic response of the cable node is investigated by applying two different seismic excitations in three different conditions as only horizontal excitation, only vertical excitation and both horizontal and vertical excitation act simultaneously. El-Centro and Loma Preta Seismic excitation is applied to predict the seismic response of the cables.

Keywords: Sagging elastic cable, seismic response, earthquake, placement model, symmetric vibration, anti-symmetric vibration

1. Introduction

The cables are used in diverse areas of structural engineering. It is well-known that elasto-flexible sagging cables lack unique natural reference configuration in their passive state. However, it is assumed that the equilibrium configuration under its self-weight as the reference configuration. Now these days quite sophisticated analytical and computational techniques have been used for obtaining the static and dynamic response of cables. The behaviour of cables shows inherent nonlinearity. In addition to this, the presence of self-weight renders the response of the cable to another type of nonlinearity. Linear modal frequencies of single sagging cables depend upon their elasto-geometrical parameters [1, 2]. Most of the investigations on cables are considering its self-weight and under points load. Under self-weight the shape of the cable is found catenary [3, 4]. A spatial two- node catenary cable element with derived tangent stiffness matrix is proposed for conducting nonlinear seismic analysis of cable structures under self-weight and concentrated loads [5]. Such a popular approach using elastic and geometric stiffness matrices for dynamic analysis for nonlinear cables has been criticized [6]. Phenomenon of internal resonances and subharmonic resonances has been predicted for these nonlinear structures [7].

The main point of departure in the approach by the authors is the assumed weightlessness of the cables and the lack of their unique natural configuration. Static and dynamic response of a single weightless elasto-flexible sagging planer cable carrying lumped nodal masses and

applied nodal loads leading to a 2-node 4-DOF cable system has earlier been investigated. Rate-type constitutive equations and third order differential equations of motion of have been derived. The dynamic response of such structures to harmonic nodal force is determined for different sustained nodal forces, axial elastic stiffness and sag/span ratios. Sub harmonic resonances, jump and beat phenomenon are predicted for elastic and inextensible cables [8]. In another approach the incremental equation of motion involving tangent stiffness matrix employed is equivalent to the third order equation of motion proposed [5].

In the discrete formulation, the nodal coordinates or placements are generally defined in reference to the chosen Cartesian coordinate system and this approach is called Placement Model. In Placement Models, the nodal deformed placements constitute the primary kinematic variables in the equation of motion.

In this paper we have adopted the Placement Model to develop the stiffness matrix of the 2-node 4-DOF cable system and further dynamic analysis. The constitutive equations and equations of motion earlier developed [8] is applied here for seismic analysis of a particular sagging planer cable structure. Also, the Placement Model developed is compared with the earlier developed Force Model [9].

As required by the equation of motion, the time-derivative of the two well-known seismic excitations is obtained. Dynamic response is investigated under horizontal and vertical seismic excitations applied separately as well as

*Corresponding author. Tel: +919968270408; E-mail address: pankaj437@gmail.com

simultaneously. In this paper, it is assumed that the vertical component of the ground acceleration is two-third of the horizontal component of the ground acceleration.

For seismic analysis of Indiano Bridge, finite element analysis is used and the material is assumed as linear elastic. High amplitude longitudinal vibrations is observed [10, 11]. In this paper the material is assumed as linear elastic and the damping is assumed as instantaneously classically damped i.e. the damping matrix is a function of stiffness and mass matrices.

Seismic analysis of transmission line system under El Centro ground excitation on are investigated with different boundary conditions. The vertical displacement of middle span of the transmission line is obtained maximum due to anti-symmetric modes of the transmission lines (Tian L. and Xia G. 2016). In this paper the assumed cable system is investigated under both El Centro and Loma Prieta ground excitations.

2. Equation of Motion and constitutive relation

The placement model is derived here for the 2-node 4-DOF planar sagging cable tied at both the ends (Fig. 1). Undeformed segment lengths of the cable are known in addition to the coordinates of the tied end D in reference to the coordinate system with origin at the end A.

Assuming the nodal placements $\{y_i\}$ in the deformed state of the cable as y_i , the axial extensions as Δ_i and tensile forces in its segments as T_i . The segment tensile forces are obtained as

$$T_1 = \frac{EA\Delta_1}{s_1} \quad T_2 = \frac{EA\Delta_2}{s_2} \quad T_3 = \frac{EA\Delta_3}{s_3} \quad (1)$$

Also, the axial elastic extension in the cable segments are expressed as

$$\begin{aligned} \Delta_1 &= \sqrt{(y_1^2 + y_2^2)} - s_1; \\ \Delta_2 &= \sqrt{[(y_3 - y_1)^2 - (y_4 - y_2)^2]} - s_2; \\ \Delta_3 &= \sqrt{[(L - y_3)^2 - (H - y_4)^2]} - s_3 \end{aligned} \quad (2)$$

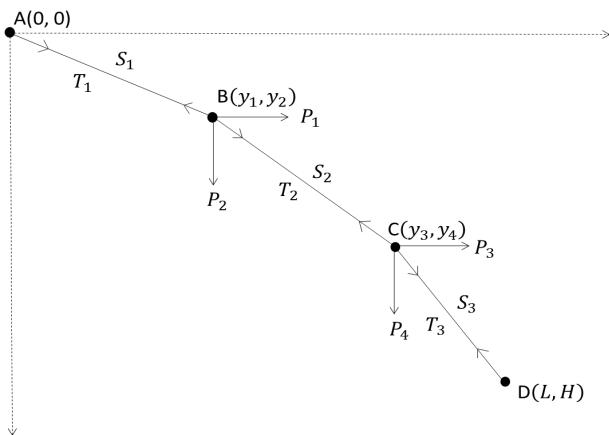


Fig. 1 Two-node four-DOF cable tied at both ends

The nodal equilibrium conditions are used to obtained the nodal loads $\{P_i\}$ in terms of the deformed state nodal placements $\{y_i\}$ as

$$\begin{aligned} P_1 &= T_1 \left(\frac{y_1}{L_1} \right) - T_2 \left(\frac{y_2 - y_1}{L_2} \right); \\ P_2 &= T_1 \left(\frac{y_2}{L_1} \right) - T_2 \left(\frac{y_4 - y_2}{L_2} \right); \\ P_3 &= T_2 \left(\frac{y_3 - y_1}{L_2} \right) - T_3 \left(\frac{L - y_3}{L_3} \right); \\ P_4 &= T_2 \left(\frac{y_4 - y_2}{L_2} \right) - T_3 \left(\frac{H - y_4}{L_3} \right); \end{aligned} \quad (3)$$

The nodal elastic displacements $\{u_i\}$ can also be represented as the components of the axial elastic extensions $\{\Delta_i\}$ along the coordinate axes and are expressed as

$$\begin{aligned} u_1 &= \Delta_1 \frac{y_1}{L_1}; & u_2 &= \Delta_1 \frac{y_2}{L_1}; \\ u_3 &= -\Delta_2 \frac{(L - y_3)}{L_2}; & u_4 &= -\Delta_2 \frac{(H - y_4)}{L_2}; \end{aligned} \quad (4)$$

Substituting of expressions for tensile forces in above equations in terms of nodal placements, the following nodal load-placement relations $\{P = P(y)\}$ are obtained as

$$\begin{aligned} P_1 &= EA \left[\left\{ \left(\frac{y_2 - y_1}{L_2} \right) - \left(\frac{y_1}{L_1} \right) \right\} \right] \\ P_2 &= EA \left[\left\{ \left(\frac{y_4 - y_2}{L_2} \right) - \left(\frac{y_2}{L_1} \right) \right\} \right] \\ P_3 &= EA \left[\left\{ \left(\frac{L - y_3}{L_3} \right) - \left(\frac{y_3 - y_1}{L_2} \right) \right\} \right] \\ P_4 &= EA \left[\left\{ \left(\frac{H - y_4}{L_3} \right) - \left(\frac{y_4 - y_2}{L_2} \right) \right\} \right] \end{aligned} \quad (5)$$

The incremental constitutive equation is established as $dP_i = K_{ij} dy_j$, where the tangent stiffness matrix coefficients as $K_{ij} = \partial P_i / \partial y_j$. The rate-type constitutive equation relating these internal nodal forces P_i and the nodal coordinates y_i is stated as

$$\dot{P}_i = K_{ij} \dot{y}_j \quad (6)$$

Consider a two-node four-DOF weightless planar sagging elasto-flexible cable with attached masses as shown in Fig. 2. Let y_i denote the deformed nodal coordinates of the cable carrying sustained nodal loads.

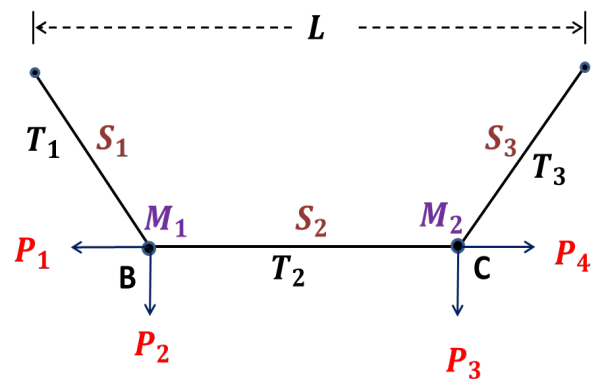


Fig. 2 Two-node four-DOF symmetrical system

The following third order coupled nonlinear differential equations of motion are derived for such cables carrying lumped masses M_{ij} .

$$M_{ij} \ddot{y}_i + C_{ij} \dot{y}_i + K_{ij} y_i = \hat{F}_i(t) \quad (7)$$

The internal resistive nodal forces \mathbf{R}_i are obtained as

$$\mathbf{R}_i = \mathbf{F}_i(\mathbf{t}) - \mathbf{M}_{ij}\ddot{\mathbf{y}}_i - \mathbf{C}_{ij}\dot{\mathbf{y}}_i \quad (8)$$

Here, the nonlinear cable structure is assumed to be instantaneously classically damped with the instantaneous damping matrix \mathbf{C}_{ij} being determined as

$$\mathbf{C}_{ij} = \alpha_0 \mathbf{M}_{ij} + \alpha_1 \mathbf{K}_{ij} \quad (9)$$

3. Structural and loading details

The seismic analysis of a planar cable net has been investigated earlier by Thai and Kim (2011). For investigation in this paper a particular single planar sagging cable is chosen from that cable net by idealization as shown in Fig. 2. The structural details are given below:

$$\mathbf{E} = 82737 \text{ MPa}; \mathbf{L} = 91.44 \text{ m}; \mathbf{A} = 146.45 \text{ mm}^2; \mathbf{H} = 0 \\ \mathbf{F}_0 = (0, 17.793, 0, 17.793) \text{ kN}$$

Two equal masses ($M = 4380 \text{ kg}$) are attached at the two nodes. Self-weight of the cable is ignored. Using Newton-Raphson Method following cable configuration and segment tension are obtained:

$$\mathbf{y}_1 = 30.41; \mathbf{y}_2 = 9.87 \text{ m}; \mathbf{y}_3 = 61.02 \text{ m}; \mathbf{y}_4 = 9.87 \text{ m} \\ \mathbf{T}_1 = 57.62 \text{ kN}; \mathbf{T}_2 = 54.80 \text{ kN}; \mathbf{T}_3 = 57.62 \text{ kN}$$

The linear modal frequencies given by eigenvalues of the matrix $\mathbf{M}_{ij}^{-1}\mathbf{K}_{ij}$ are obtained as $\omega_{n1} = 1.07 \text{ rad/s}$; $\omega_{n2} = 2.47 \text{ rad/s}$; $\omega_{n3} = 9.32 \text{ rad/s}$; $\omega_{n4} = 16.20 \text{ rad/s}$.

In the presence of seismic excitation, the nodal force vector is obtained as

$$\mathbf{F}_i(\mathbf{t}) = \mathbf{F}_{0i} + \mathbf{E}_i(\mathbf{t}) \quad (10)$$

Where $\mathbf{E}_i(\mathbf{t})$ represents the applied time-dependent seismic forces. It is assumed that the dynamic forces introduced by the earthquakes and acting on the nodal masses are lying in the plane of the cable. Also, these forces depend upon the horizontal and vertical components of the seismic acceleration. In the case when only horizontal seismic forces are considered

$$\mathbf{E}_h(\mathbf{t}) = (-M_1\ddot{y}_h, 0, -M_2\ddot{y}_h, 0) \quad (11)$$

Similarly, when only vertical seismic forces are acting

$$\mathbf{E}_v(\mathbf{t}) = (0, -M_1\ddot{y}_v, 0, -M_2\ddot{y}_v) \quad (12)$$

In fact, the horizontal and the vertical components of the seismic forces act simultaneously. In this paper, it is assumed that the vertical component of the ground acceleration is two- third of the horizontal component of the ground acceleration with same frequency content. As the equation of motion is third order differential equation, rate of loading vector is also needed. In the case of horizontal and vertical seismic accelerations, the loading rate vectors are represented as

$$\begin{aligned} \dot{\mathbf{E}}_h(\mathbf{t}) &= (-M_1\ddot{y}_h, 0, -M_2\ddot{y}_h, 0) \\ \dot{\mathbf{E}}_v(\mathbf{t}) &= (0, -M_1\ddot{y}_v, 0, -M_2\ddot{y}_v) \end{aligned} \quad (13)$$

Here, \ddot{y}_h and \ddot{y}_v are the rates of change of horizontal and vertical components of ground acceleration respectively. Here, \ddot{y}_h and \ddot{y}_v are determined from the available ground acceleration components \ddot{y}_h and \ddot{y}_v respectively as

$$\begin{aligned} \ddot{y}_h(t_i + \frac{\Delta t}{2}) &= \frac{\ddot{y}_h(t_i + \Delta t) - \ddot{y}_h(t_i - \Delta t)}{2\Delta t} \\ \ddot{y}_v(t_i + \frac{\Delta t}{2}) &= \frac{\ddot{y}_v(t_i + \Delta t) - \ddot{y}_v(t_i - \Delta t)}{2\Delta t} \end{aligned} \quad (14)$$

From the available records for the El Centro Southern California, US earthquake (1940), the horizontal ground acceleration components and the deduced rate of acceleration normalized with respect to acceleration due to gravity is plotted in Fig. 3(a) and Fig. 3(b) respectively. As the seismic response of structures also depend upon their modal frequencies as well as the frequency content of the applied excitation. For better understanding and easy interpretation of dynamic response, Fast Fourier Transform (FFT) of the same ground acceleration as well as the rate of ground acceleration is presented in Fig. 3(c) and 3(d).

Similar characteristics of the Loma Prieta, California's Central Coas, US earthquake (1989) are presented in Fig. 4(a), 4(b), 4(c) and 4(d) respectively. It is observed that the peak ground acceleration (PGA) of the El Centro and the Loma Prieta earthquakes are 0.3163g and 0.5235g respectively with 'g' as acceleration due to gravity. It is also observed that the dominant frequency ranges for the El Centro and the Loma Prieta seismic accelerations lies between 5 to 45 rad/s and 4 to 25 rad/s respectively. The corresponding frequency ranges for of the dominant rate of seismic accelerations are 13 to 42 rad/s² and 17 to 52 rad/s².

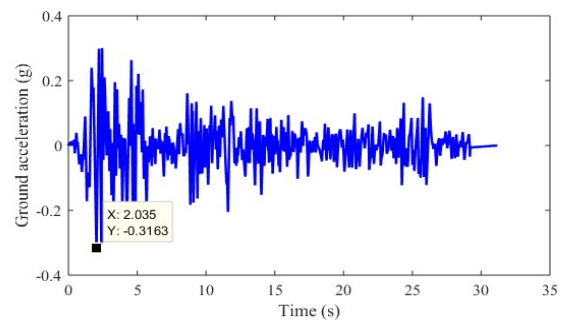


Fig. 3(a) Ground acceleration (El Centro)

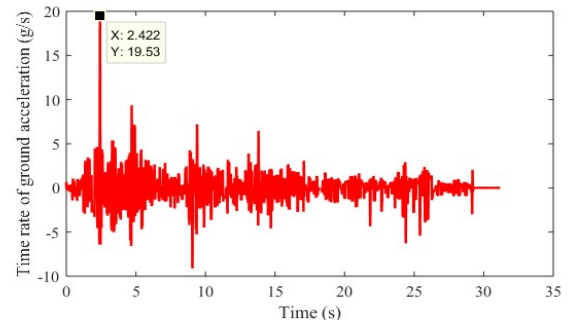


Fig. 3(b) Time rate of ground acceleration (El Centro)

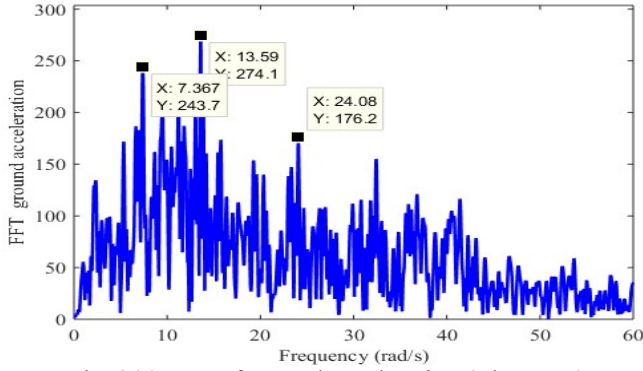


Fig. 3(c) FFT of ground acceleration (El Centro)

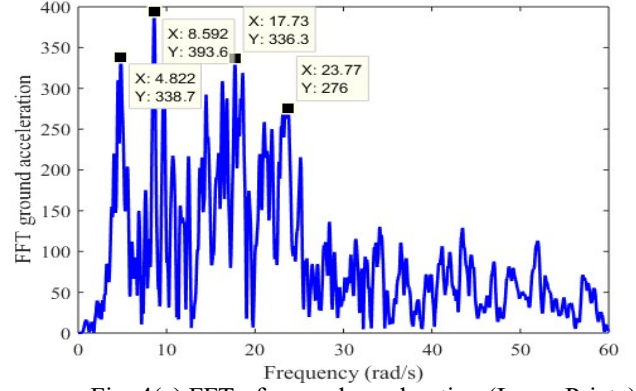


Fig. 4(c) FFT of ground acceleration (Loma Prieta)

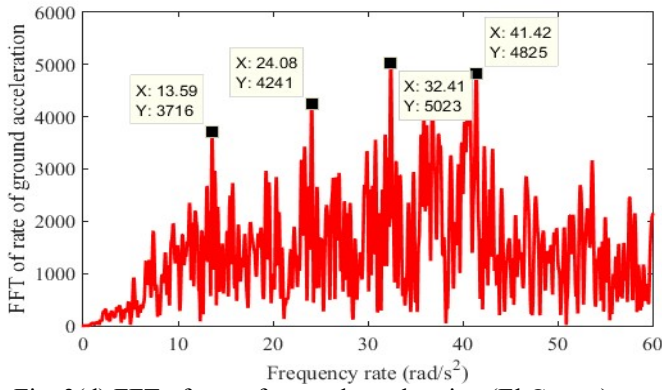


Fig. 3(d) FFT of rate of ground acceleration (El Centro)

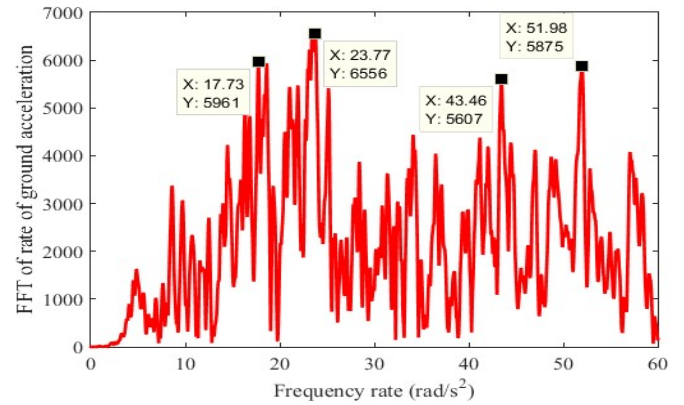


Fig. 4(d) FFT of rate of ground acceleration (Loma Prieta)

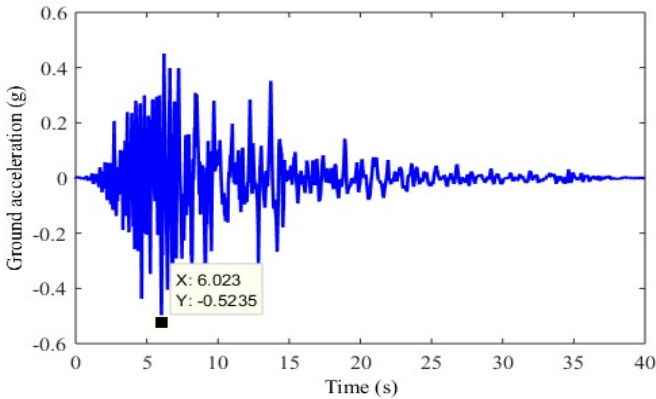


Fig. 4(a) Ground acceleration (Loma Prieta)

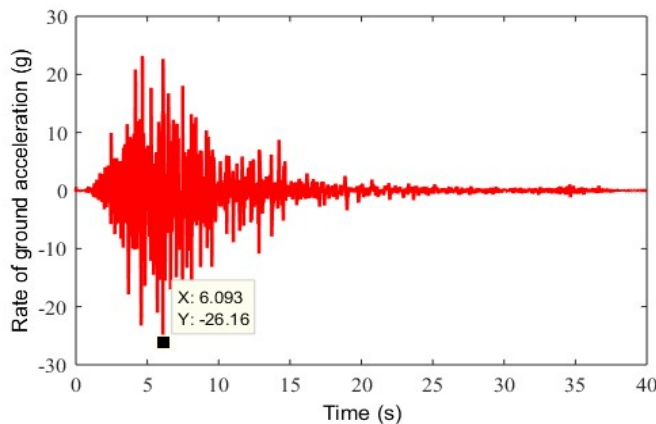


Fig. 4(b) Time rate of ground acceleration (Loma Prieta)

4. Predicted seismic response: El Centro Earthquake

The dynamic response of the chosen cable structure is investigated with following three cases of ground motion as (i) Horizontal excitation (ii) Vertical excitation and (iii) Combination of both horizontal and vertical excitation.

4.1 Only horizontal excitation is applied

When only horizontal seismic force is applied to the cable structure, then peak vertical nodal response is about twice than the peak horizontal nodal response as shown in the Fig. 5(a) and 5(b). Also, the vertical nodal responses of both nodes are in opposite phase but same magnitude, whereas the horizontal response of both the nodes are in same phase and same magnitude.

Tension increment is plotted for all three cable segments. It is very interesting to note that, the predicted tension increment in the inclined cables are much large than that of horizontal cable segment as shown in Fig. 6(a) and Fig. 6(b) respectively. Also, the tension increment waveform for inclined cable shows opposite phase with same magnitude. Here negative sign does not mean compressive axial force, but the decrease in the axial tensile force.

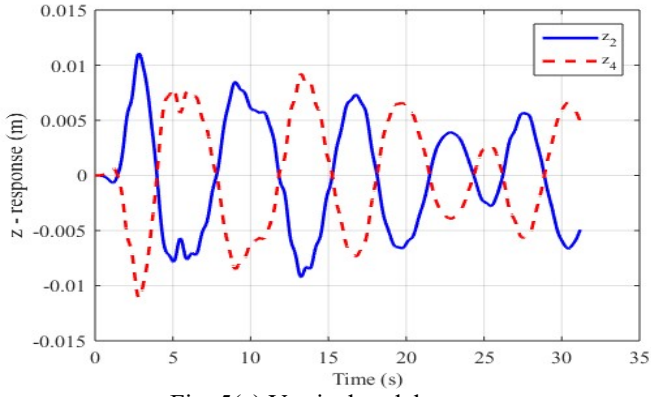


Fig. 5(a) Vertical nodal response

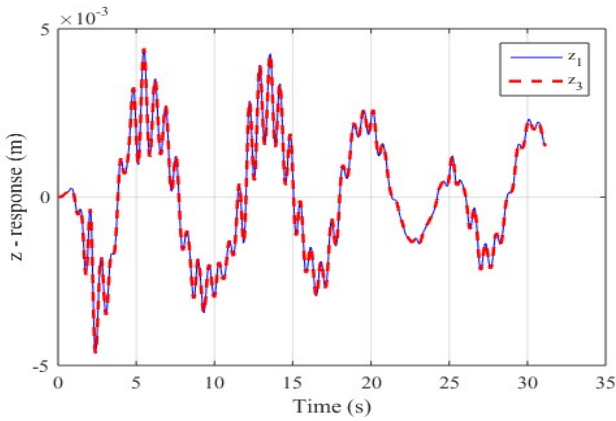


Fig. 5(b) Horizontal nodal response

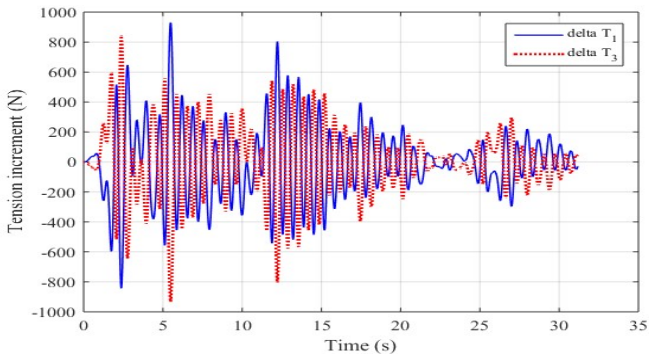


Fig. 6(a) Tension increment of inclined cable segments

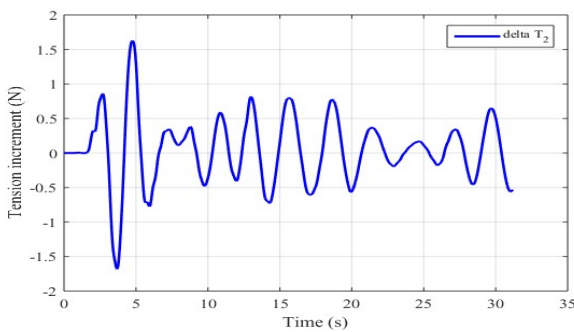


Fig. 6(b) Tension increment of horizontal cable segments

4.2 Only vertical excitation is applied

In case when only vertical seismic excitation is applied to the system, the peak vertical nodal response of node 2 is about ten times the peak horizontal nodal response as shown

in the Fig. 7(a) and 7(b). Also, the vertical nodal response of node 1 is interestingly much lower than the vertical nodal response of node 2. The horizontal response of both the nodes are same in magnitude but with opposite phase.

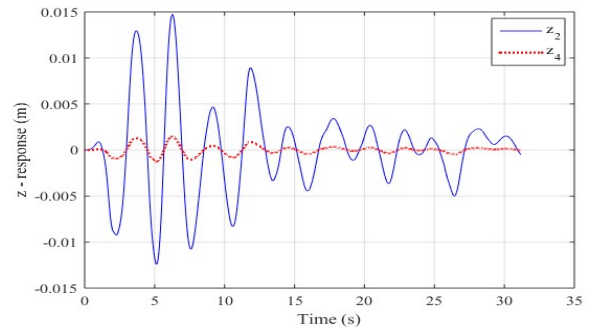


Fig. 7(a) Vertical nodal response

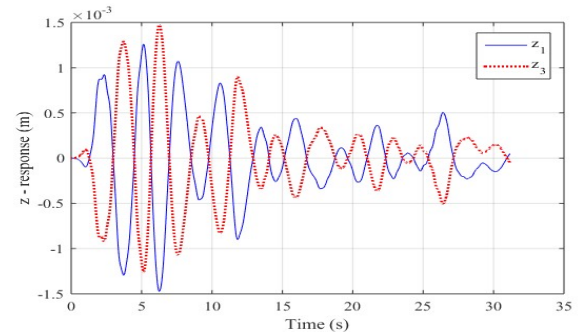


Fig. 7(b) Horizontal nodal response

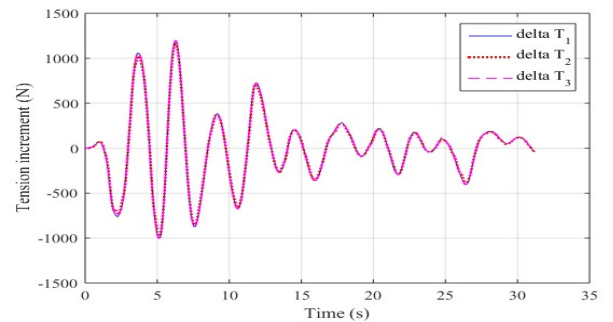


Fig. 8 Tension increment in all cable segments

Tension increment is plotted for each three cable segments in Fig. 8. The predicted tension increments in all three cable segments for vertical seismic excitation are same.

4.3 Both horizontal and vertical excitation is applied

When both vertical and horizontal seismic excitation is applied simultaneous to the cable structures, both the peak horizontal and vertical nodal response is observed to be higher compared to earlier cases when only horizontal or only vertical excitation is applied respectively. Also the nature of predicted waveform is different in this case as shown in in Fig. 9(a) and 9(b).

The predicted tension increment in all three cable segments of each three cable segments are shown in the Fig. 10. The predicted tension response shows that, the increment in tension of inclined segment is slightly more than the horizontal segment. It is interesting to note that, the pattern of tension increment is similar for all three segments.

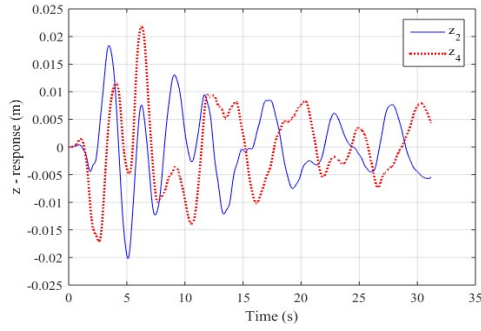


Fig. 9(a) Vertical nodal response

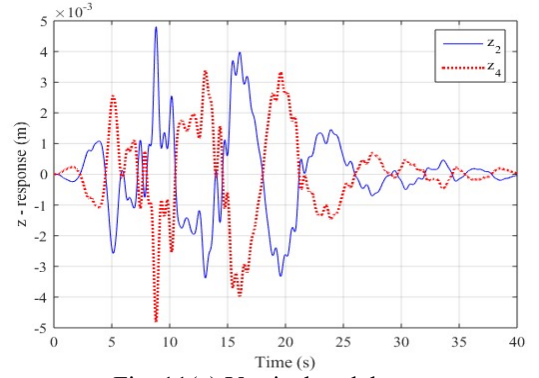


Fig. 11(a) Vertical nodal response

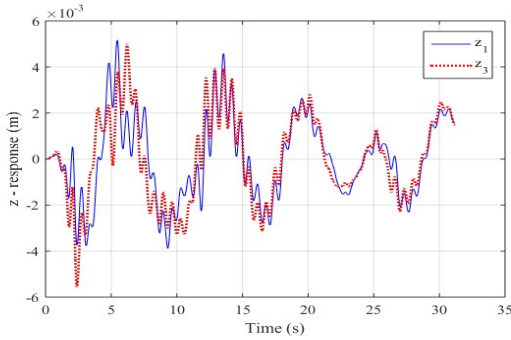


Fig. 9(b) Horizontal nodal response

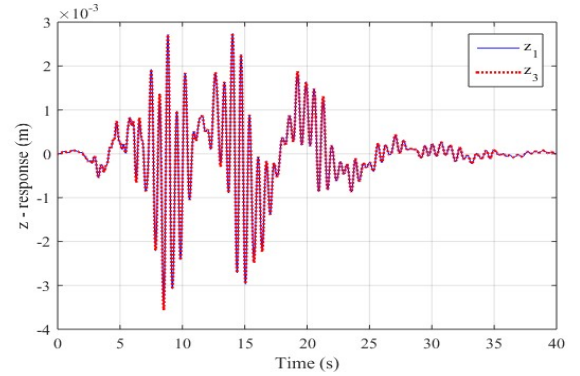


Fig. 11(b) Horizontal nodal response

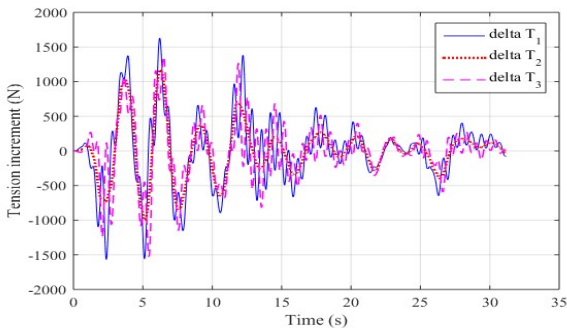


Fig. 10 Tension increment in all cable segments

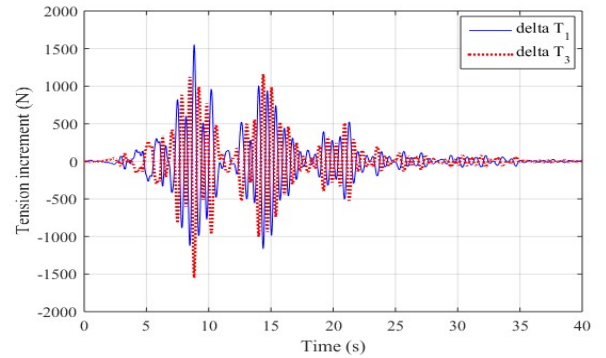


Fig. 12(a) Tension increment of inclined cable segments

5. Predicted seismic response: Loma Preita Earthquake

5.1. Only horizontal excitation is applied

In this case the Loma Preita seismic excitation is applied in all three cases. When only horizontal seismic force is applied to the cable structure, the vertical nodal response of both the nodes shows similar response but out of phase as shown in the Fig. 11(a). Whereas the horizontal nodal response of both the nodes are exactly similar in nature as shown in Fig. 11(b).

The predicted tension response of inclined cable segments shows much higher increment than horizontal cable segments. The tension increments of both the inclined segments are similar in magnitudes but opposite in phase as shown in Fig. 12(a). Almost no change in the tension of horizontal segment is occurred as shown in Fig. 12(b).

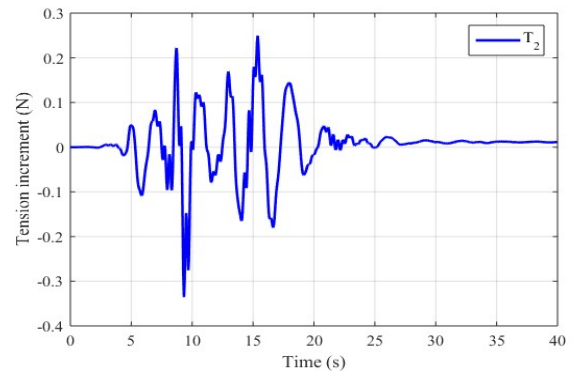


Fig. 12(b) Tension increment of horizontal cable segments

5.2. Only vertical excitation is applied

When only vertical Loma Preita seismic excitation is applied to the cable structure, the vertical nodal response of both the nodes as shown in the Fig. 13(a) have similar

response with same phase. In other words the vertical response of both the nodes coincides with each other as shown in the figure. However the horizontal nodal responses of both the nodes are similar in magnitudes but having opposite phase as shown in Fig. 13(b).

The predicted tension of all the three cable segments shows almost similar response as shown in Fig. 14. However at few time instant the tension response of horizontal cable segment differ slightly with inclined cable tension response. Also the magnitude of tension increment in this case is about half of that of previous case when only horizontal seismic force is applied.

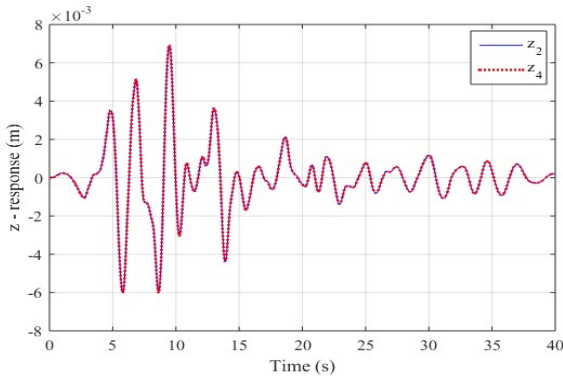


Fig. 13(a) Vertical nodal response

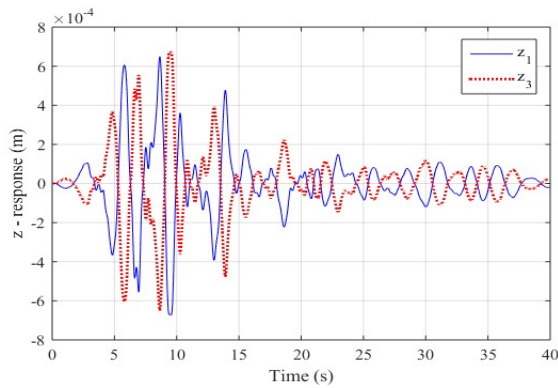


Fig. 13(b) Horizontal nodal response

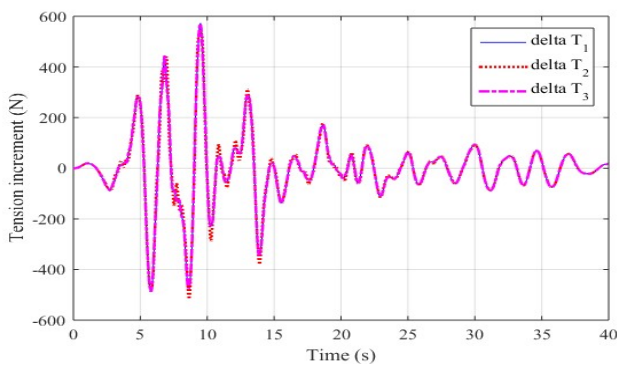


Fig. 14 Tension increment in all cable segments

5.3. Both horizontal and vertical excitation is applied

When both vertical and horizontal seismic excitation is applied simultaneous to the cable structures, the vertical nodal response shows similar phase in first last few seconds

while opposite phase in middle of the excitation. However the magnitude of vertical response of both the nodes changes relatively with time passes as shown in in Fig. 15(a). The horizontal nodal response shows about half the magnitude of corresponding vertical nodal response. Also, both nodes show relatively similar horizontal response as shown in Fig. 15(b).

The predicted tension increment in inclined cable segments are about double that of horizontal cable segments are shown in the Fig. 16(a) and Fig. 16(b) respectively.

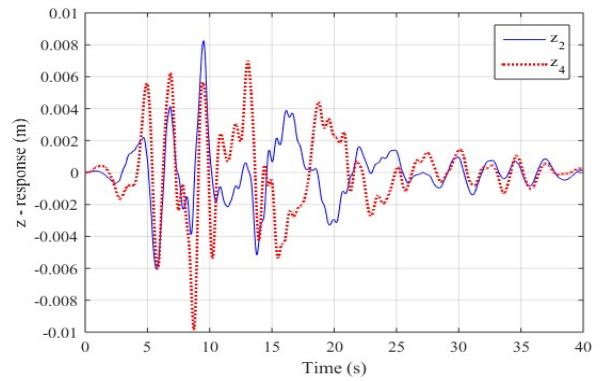


Fig. 15(a) Vertical nodal response

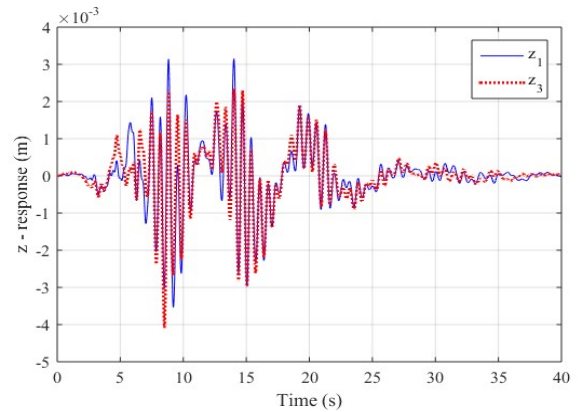


Fig. 15(b) Horizontal nodal response

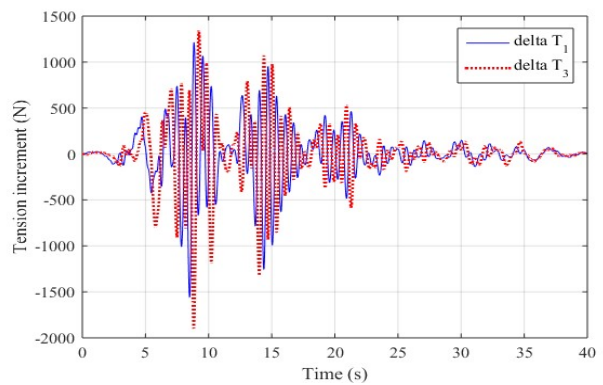


Fig. 16(a) Tension increment of inclined cable segments

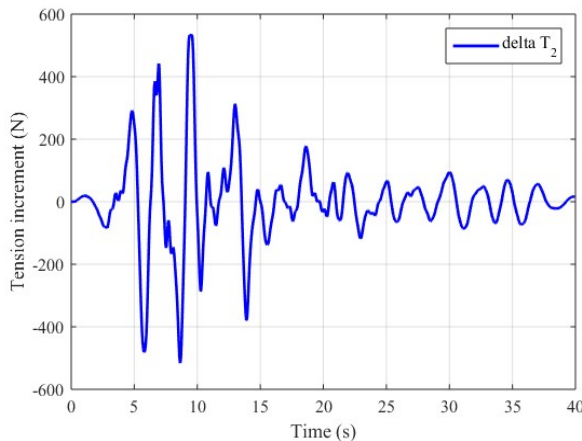


Fig. 16(b) Tension increment of horizontal cable segments

6. Discussion

In this paper the seismic response of sagging elasto-flexible cable is investigated with two distinct seismic excitation available named El Centro (1940) and Loma Prieta (1989). The author earlier investigated the same problem by adopting Force Model and find interesting results of sagging cable vibration [13]. The earthquake excitation is applied to the cable structure in three different ways. In first case, only horizontal seismic force is applied while in the second case only vertical seismic forces is applied. In third case both horizontal as well as vertical seismic force is applied simultaneously. It is important to note that, the vertical seismic force vector remains parallel to the sustained load vector. As the relative magnitudes of the sustained nodal forces remain same for the duration of the earthquakes, the vertical seismic force vector represents the proportionate load vector. On the other hand, the horizontal seismic force vector observed to remain orthogonal to the sustained load vector.

It is predicted that the tension increment response in the inclined cable segments are of similar order irrespective of the case of application of the seismic excitation. In other words the tension increments in the inclined cable segments due to seismic force are more or less same in all the cases of seismic force. This behaviour is shown by both El Centro and Loma Prieta ground excitation. This is due to the facts that the part or full inertia force is always taken by inclined cable segments.

The tension increment in the horizontal cable segment is negligible when only horizontal seismic force is applied whereas the tension increment in all the cable segments are same when only vertical seismic force is applied. This can be due to horizontal seismic excitation, relative change in horizontal inertial force is zero due to similar mass at the both node. This confirms the facts that due to horizontal seismic force horizontal nodal vibration are symmetric and hence no tension increments should be there.

It is very interesting to note that, when only vertical seismic forces is applied the maximum horizontal nodal response amplitude is negligible although the tension increment in the horizontal cable segment is of the order 1.2 kN. This is due to anti-symmetric vibration of the both node in horizontal direction.

7. Conclusion

The predicted seismic responses confirm the facts that the vertical vibration of the cable nodes is symmetric under vertical seismic force and anti-symmetric under horizontal seismic forces. Similarly the horizontal vibration of the cable nodes is symmetric under horizontal seismic forces and anti-symmetric under vertical seismic forces. The change in the tension in the inclined cable shows maximum variation in all the condition of application of the seismic excitation. The horizontal cable segment shows maximum tension variation in case of anti-symmetric nodal vibration in horizontal direction, whereas minimum or negligible variation in case of symmetric horizontal nodal vibration. The present investigation shows interesting facts about the nonlinear characteristics of the cable vibration with simple approach. The investigation may help to understand the basic dynamics of the sagging cable.

Acknowledgement

The authors express their sincere thanks towards the National Project Implementation Unit (NPIU) MHRD, New Delhi, India for providing research grant for this work.

References

1. Irvine, H.M. and Caughey, T.K. The linear theory of free vibrations of a suspended cable. *Mathematical and Physical Sciences*, 1974; 341(1626): 299-315.
2. Lacarbonara, W., Paolone, A. and Vestroni, F. Non-linear modal properties of non-shallow cables. *Int. J. of Non-Linear Mechanics*, 2007; 42: 542-554.
3. Fried, I. Large deformation static and dynamic finite element analysis of extensible cables, *Computers and Structures*, 1982; 15(3): 315-319.
4. Koh, C.G, Zhang, Y. and Quek, S.T. Low-tensioned cable dynamics: numerical and experimental Studies. 1999; *J. of Eng. Mechanics*, 125(3): 347-354.
5. Thai, H.T. and Kim, S.E. Nonlinear static and dynamic analysis of cable structures, *Finite Elements in Analysis and Design*, 2011; 47: 237-246.
6. Volokh, K.Y., Vilnay, O. and Averbuh, I. Dynamics of cable structures. *Journal of Engineering Mechanics*, 2003; 129(2): 175-180.
7. Kamel, M.M. and Hamed, Y.S. Nonlinear analysis of an elastic cable under harmonic excitation." *Acta Mech. Springer-Verlag*, 2010; 214: 135-325.
8. Kumar P., Ganguli A. and Benipal G.S. Theory of weightless sagging elasto-flexible cables. *Latin American Journal of Solids and Structures*, 2016; 13(1): 155-174.
9. Kumar P., Ganguli A. and Benipal G.S. Comparative assessment of the contending force and placement methods for weightless sagging cables. *Asian Journal of Civil Engineering*, 2019; 20(7): 1049-1062.

10. Clemente, P., Celebi, M., Bongiovanni, G., and Rinaldis, D. Seismic analysis of the Indiano cable-stayed bridge. In 13th World Conference on Earthquake Engineering, Vancouver, BC, Canadá, 2004. paper (No. 3296).
11. Gong, Jun, Xudong Zhi, Feng Fan and Shizhao Shen. Static and dynamic stiffness in the modeling of inclined suspended cables. Journal of Constructional Steel Research 172 (2020), pp. 106210.
12. Tian, Li, and Xia Gai. Nonlinear seismic behavior of different boundary conditions of transmission line systems under earthquake loading. Shock and Vibration (2016).
13. Kumar, P., Ganguli, A., and Benipal, G.S. Seismic analysis of weightless sagging elasto-flexible cables. In Advances in Structural Engineering, Springer, New Delhi, 2015; 1543-1562.



Sensitivity Investigation of Junctionless Gate-all-around Silicon Nanowire Field-Effect Transistor-Based Hydrogen Gas Sensor

Rishu Chaujar¹ · Mekonnen Getnet Yirak^{1,2}

Received: 16 May 2022 / Accepted: 19 November 2022
© Springer Nature B.V. 2022

Abstract

In this work, a junctionless (JL) gate all around (GAA) silicon nanowire field-effect transistor sensor for the detection of hydrogen (H_2) has been carried out. The sensors are designed to specify hydrogen gas (H_2) existence. Unsafe conditions can result if hydrogen escapes and accumulates in an enclosed space throughout the purifying process; this is why we try to investigate technologically ultra-small-scale hydrogen gas sensor devices. The sensor also showed satisfactory characteristics for ensuring safety when handling hydrogen and remarkable selectivity for monitoring H_2 among other gases, such as LPG, NH_3 , and CO. The temperature and palladium (Pd) gate work function variations in the translation processes are well-thought-out throughout a change in palladium (Pd) gate work function following exposure to the hydrogen gas molecule (H_2). Due to its sensitivity to H_2 gas, palladium (Pd) is employed as a gate electrode in H_2 gas detection. Shift in threshold voltage (V_{th}), Ion and Ioff as a result of the metal work function at the gate changing when gas is present; these changes can be regarded as sensitivity parameters for sensing hydrogen gas molecules. ATLAS-3D device simulator has been conducted at low drain bias voltage (0.05V). This study focuses on temperature variation (300K to 500K) and palladium (Pd) metal gate work function variations (5.20eV to 5.40eV) to examine the existence of hydrogen molecule (H_2) and its effect on the performance of junctionless SiNW-GAA field-effect transistor gas sensors. When the sensitivity ($S_{I_{OFF}}$), of proposed JL-GAA-SiNWFET is compared with GAA-MOSFET and bulk MOSFET, JL-GAA-SiNWFET shows improved sensitivity. The results show that as 150mV Pd work function shifts at the gate, the sensitivity improvement with JL-GAA-SiNWFET-based hydrogen gas sensors are 51.65% and 124.51% compared with GAA-MOSFET and MOSFET, respectively. High dielectric oxide (HfO_2) and interface oxide (SiO_2) is also employed at the gate to overcome electron tunneling. The study of this work proves that a silicon nanowire field-effect transistor with a junctionless gate all around catalytic palladium (Pd) metal gate is the best candidate for sensing hydrogen gas molecules than a bulk metal oxide semiconductor field-effect transistor (MOSFET).

Keywords Hydrogen gas-sensor · Junctionless (JL) · Silicon nanowire FET · Gate-all-around (GAA)

1 Introduction

Hydrogen is recognized as one of the most significant clean energy carriers and the ultimate fossil fuel candidate and renewable energy source [1] because of its high

heat of combustion, low minimum ignition energy, and high combustion velocity. Due to its robust reducing characteristics, hydrogen is also employed in metal smelting, petroleum extraction, semiconductor processing, glass-making, and the daily chemical industry, among other things [2]. It is owing to the growing demand for gas sensing sensors for seismic monitoring applications, environmental monitoring, medical and automotive industries, in addition to domestic usages, such as detecting pollutants, fueling stations, petroleum refineries, and detecting certain types of bacterial infection, which are continuously at high perilous of gas leakage [1–6]. Designing a hydrogen gas sensor based on a GAA-JL-SiNWFET device is an exciting option for gas sensors. It offers low power consumption, high sensitivity, low cost, portability,

✉ Rishu Chaujar
chaujar.rishu@dtu.ac.in
Mekonnen Getnet Yirak
mekonnengetnet01@gmail.com

¹ Applied Physics Department, Delhi Technological University, Delhi, India

² Physics Department, Debre Tabor University, Debre Tabor, Ethiopia

technology compatibility, on-chip integration, small size, and CMOS compatibility [7, 8]. Humans cannot smell hydrogen gas since it is colorless and tasteless [9, 10]. It is easily flammable and explosive due to its low explosion energy and extensive flammable range. As a result, an effective and reliable hydrogen sensing device is required for hydrogen manufacturing and consumption, monitoring and managing hydrogen concentrations in nuclear reactors and coal mines, and detecting and alarming H_2 leakage during storage, transportation, production and usage [1, 3, 6, 7]. As a result, such sensors seem to be among the most straightforward, inexpensive, and efficient tools for real-time measurement or gas leak detection [10]. Due to various reasons, different types of SiNWFET-based hydrogen gas detecting devices have been designed in recent years to identify gas molecules by analyzing the induced change in work function at the surface of an attractive film [6, 9]. Numerous types of gas detectors are available, but FET-based gas detectors have received much attention [11]. Device engineering is being used in this area of research and development, including modeling and evaluating the field-effect device to improve sensitivity [5]. Floating gate MOSFETs [12], SOI MOSFETs, dual-gate MOSFETs [13], and now nanowire MOSFETs [14] have all been considered in device engineering. A high surface area to volume ratio is necessary to boost sensitivity by raising the potential for surface interactions [8]. Gas sensors in this device depend on the interaction of a thin Pd layer with hydrogen gas [9, 15, 16]. A junctionless nanowire transistor is a gated resistor with the same doping type on the source, channel, and drain without junctions [17]. For example, leakages are always a hazard at gas stations and refineries, and early recognition is thoughtful to minimize dangers and accidents [1, 2]. A silicon nanowire Field Effect Transistor (SiNWFET) sensors are an enticing proposition for gas sensing [9] due to technology compatibility [18] for on-chip integration, portability, low power consumption, and the ability to detect both weakly bound strongly bonded and chemical bonding species at room temperature [19, 20]. The detecting mechanism is the interfacial adsorption of disassociated hydrogen molecules into the palladium gate results in the formation of a dipole layer, which alters the gate's work function and causes a significant shift in threshold voltage (ΔV_{th}) [3, 9, 19]. For example, different catalytic gate metals have been utilized to realize the hydrogen gas sensor, such as Palladium [15, 21], Platinum [1, 7], and poly-methylmethacrylate-platinum [8]. Semiconductors such as silicon nanowires (SiNWs) and thin films have been utilized as sensing materials for the development of high hydrogen gas sensors in recent years [1] [9] due to their huge specific surface area and unique electron transportation characteristics.

Numerous nanoelectronics devices with multiple gate materials, for instance, floating gate MOSFETs [20], Palladium (Pd) gate MOSFETs [22], and Tunnel-FET (TFET) [23], etc., have been designed to boost the sensitivity of SiNWFET based sensors [16]. The planer MOSFET is the most ideal among them because of its production ease, although it has a number of drawbacks in the ultra-small scale dimension, such as short channel effects (SCEs), Subthreshold Swing [24–26]. Under ambient conditions, a junctionless catalytic metal gate-all-around silicon nanowire FET [27–31] device is best for overcoming these issues and providing better performance in terms of sensitivity and response time. Because of a more active surface interaction of hydrogen gas molecules with palladium [19], nanostructure palladium gate materials have a high attraction for hydrogen gas and provide superior sensing capability [22, 32] than bulk palladium materials. Due to these and other physical significance, we employed Nanowire palladium materials as gate electrodes for our proposed device.

In this work, high-k materials for gate oxide, like hafnium oxide (HfO_2) and interface oxide (SiO_2), were chosen to develop new and high-performance electrical devices at the nanoscale that has relied heavily on gate dielectric materials [33–35]. Since hafnium oxide (HfO_2) provides the most excellent and powerful dielectric materials, enhancing the sensing performance compared to other simulated dielectric materials [33, 36] in controlling tunneling/leakage current. The significance of utilizing Pd metal gate as a catalyst to improve SiNW thin film H_2 -sensing performance is the straightforward synthesis, which allows for precise control of the thicknesses of SiNW and Pd metal gate to produce detectors with the maximum possible sensitivity [37]. Because palladium electrode selectively absorbs hydrogen gas and produces palladium hydride, as it is employed in various industries [22].

Therefore, we have designed a p-type substrate junctionless gate-all-around SiNWFET-based sensor to investigate hydrogen gas using Atlas-3D-TCAD device simulator tool.

1.1 Device structure

The designed structure for our proposed device is illustrated in Fig. 1) here, L (20nm) is the length of dielectric material (HfO_2), channel, (SiO_2) is an interface oxide layer and T_1 , T_2 , and T_3 are the thickness of the metal Gate, hafnium oxide, and interface (SiO_2) oxides, respectively, and $2R$ is channel diameter (silicon film thickness), as shown in Table 1. The device gate length is 20nm for all simulations, and source and drain lengths are each 10nm, as shown in Table 1. To achieve a tolerable threshold voltage, a high doping concentration ($1 \times 10^{19} \text{cm}^{-3}$) was applied uniformly through the channel

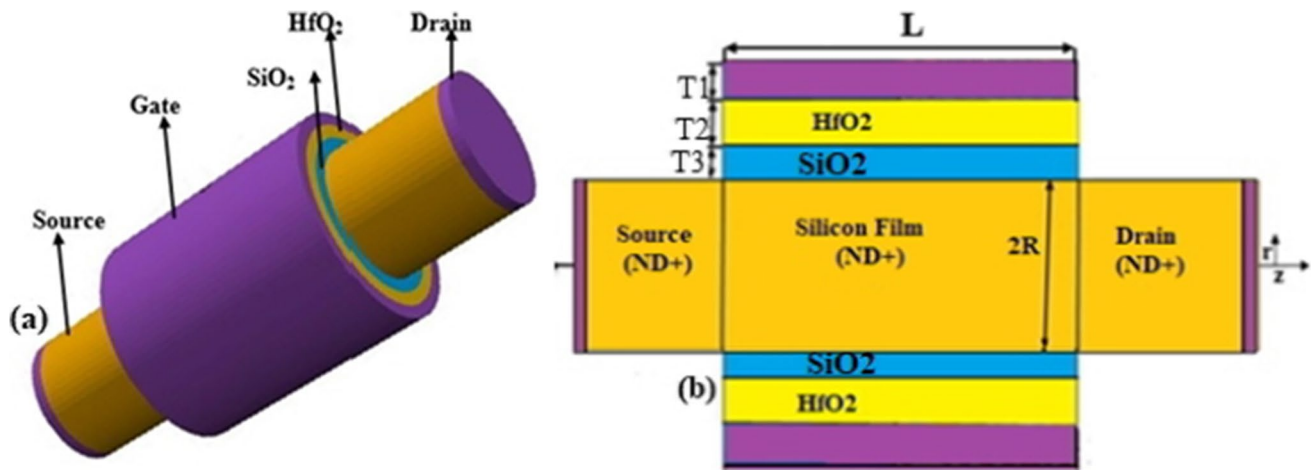


Fig. 1 illustrates (a) 3D representation diagram and (b) 2D cross-sectional view for p-type substrate cylindrical JL-GAA-SiNWFET-based hydrogen sensor

Table 1 Technology parameters

Device Parameters	GAA-JL-SiNWFET
Channel length (nm)	20.00
Thickness of oxide HfO ₂ & SiO ₂ respectively, (nm)	1.50 & 0.30
Interface Oxide (SiO ₂) thickness, (nm)	1.00
Oxide (SiO ₂) length, (nm)	20.00
Source and Drain length/thickness (nm)	10.00
Hafnium Oxide (HfO ₂) length, (nm)	20.00
Radius of silicon film (nm)	10.00
Drain, Source & Channel Doping (N _D ⁺)	10 ¹⁹ cm ⁻³
Oxide dielectric constant, HfO ₂ & SiO ₂	25.00 & 3.90
Reference gate work function (Palladium), (eV)	5.20

from the source to drain for our designed p-type substrate devices. The supply gate-source voltage (0.6V) with a consistent drain-source voltage (0.05V) have been employed for all simulations.

1.2 Simulation methodology

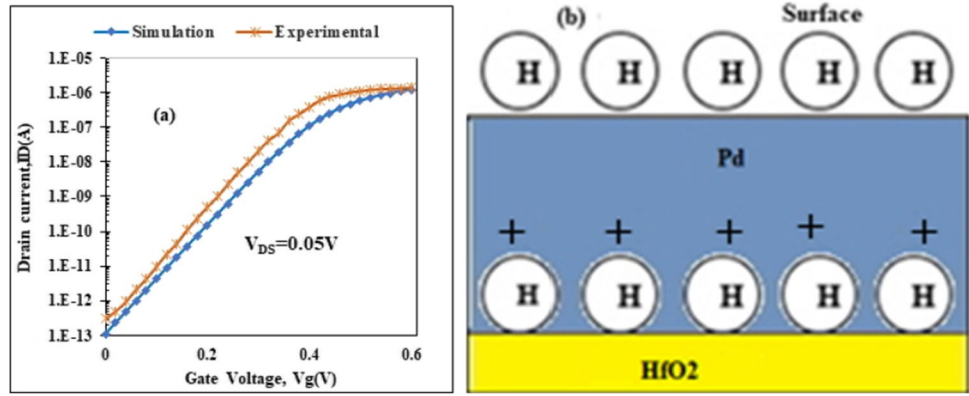
Silvaco TCAD-ATLAS tool is used for all simulations in this work. Concentration-dependent mobility, drift-diffusion, and field-dependent mobility models are activated to incorporate electron mobility models [38]. To account for the recombination of majority and minority charge carriers, the Shockley-Read-Hall (SRH) recombination model is also activated [39]. The drift-diffusion model also accounts for the driving current caused by the charge carrier following. Due to heavy source and drain doping, Fermi Dirac Statistics have been introduced.

The Boltzmann transport statistics and concentration, voltage, and temperature (CVT) [39] Lombardi mobility model [16] have accounted for parallel and perpendicular field mobility [39]. Palladium (Pd) has a high affinity for hydrogen, making it an excellent material for detecting hydrogen storage (reversibly introduced) [10, 40, 41]; when interacting with the palladium surface, Van der Waals forces interact between hydrogen gas molecules and palladium atoms [22].

Figure 1 demonstrates (a) the three-dimensional structures and (b) the two-dimensional cross-sectional view of the p-type palladium metal gate Junctionless (JL) GAA SiNWFET-based hydrogen gas sensor. The GAA SiNWFET structure includes a 40 nm long p-type doped channel and applied $1 \times 10^{19} \text{ cm}^{-3}$ doping concentration from the source to drain through the channel uniformly. Palladium metals were applied as the gate material because hydrogen molecules at the palladium surface breakdown when H₂ gas is exposed to a palladium metal gate, which causes dissociated molecules to diffuse into the gate [10, 22].

In this study, we have considered the catalytic metal gate method to describe the behavior of JL-GAA-SiNWFET based hydrogen gas sensor. The Pd work function must be a critical factor in altering the electrical field properties of the device as it changes [40]. The H₂ gas molecules break down at the metal surface of the metal gate (Pd) after exposure to the gas [40], and the disassociated molecules subsequently diffuse within the metal gate, as shown in Fig. 2b. Consequently, some hydrogen atoms diffuse through the gate metal, eventually producing the dipole at and within the interface by changing the metal-work function. As a result, we have examined the $I_{\text{ON}}/I_{\text{OFF}}$

Fig. 2 illustrates (a) Calibration with simulation results at $V_{DS}=0.05V$ with the experimental result [42] and (b) 2D Electrical dipole generation at the Pd/HfO₂ interface



ratio, drain-off sensitivity (S_{Ioff}), and the proposed device's performance shift in threshold voltage.

1.3 Analytical modeling

Using boundary conditions, surface potential in the radial direction is obtained;

$$\phi_s(z) = Ae^{kz} + Be^{-kz} + \Phi \quad (1)$$

Here k is described by

$$k = \sqrt{\frac{2\epsilon_{OX}}{\epsilon_{Si}R^2 \ln\left(1 + \frac{t_{OX}}{R}\right)}} \quad (2)$$

And Φ is given by

$$\Phi = V_{gs} - V_{fb} - qN_{Si}/\epsilon_{Si}k^2 \quad (3)$$

I) As a function of z , the surface potential is given by:

$$\phi(r=0, z) = \phi_s(z) \quad (4)$$

II) At the center of silicon substrate's electric field is zero and expressed as:

$$\left. \frac{\partial \phi(r, z)}{\partial r} \right|_{r=0} = 0 \quad (5)$$

At the boundary of silicon oxide, the electric field is computed as follows:

$$\left. \frac{\partial \phi(r, z)}{\partial r} \right|_{r=\frac{t_{Si}}{2}} = \frac{C_{OX}}{\epsilon_{Si}} \left(V_{gs} - V_{fb} - \phi\left(r = \frac{t_{Si}}{2}, z\right) \right) \quad (6)$$

Oxide capacitance per unit area (C_{OX}) is obtained;

$$C_{OX} = \frac{\epsilon_{OX}}{(R/2) \ln\left(1 + \frac{t_{OX}}{R}\right)} \quad (7)$$

Here, t_{Si} is silicon thickness, R is silicon (channel) radius, ϵ_{Si} is permittivity of silicon, and ϵ_{OX} is oxide layer permittivity. The variation in the catalytic metal work function at the metal surface by the reactivity of gas molecules is denoted by $\Delta\Phi_m$, and the flat-band voltage is V_{fb} , and V_{fb} is described by [5, 9];

$$V_{fb} = \phi_m - \phi_{Si} \pm \Delta\Phi_m \quad (8)$$

where ϕ_{Si} represents for a silicon work function and is obtained by;

$$\phi_s = \frac{E_g}{2} + \chi + q\phi_{fp} \quad (9)$$

The value of $\Delta\Phi_m$ is expressed using Eq. (10):

$$\Delta\phi_M = \text{cont.}(\Phi_m) - \left(\frac{RT}{4F}\right) \ln P \quad (10)$$

Where T is for absolute temperature, P is for partial hydrogen gas pressure, R is for hydrogen gas constant, and F is for Faraday's constant. A and B are coefficients obtained using source and drain boundary conditions and determined using the formula;

$$A = \frac{(V_{bi} + \phi)(1 - e^{-kL}) + V_{ds}}{2\sinh(kL)} \quad (11)$$

$$B = \frac{(V_{bi} + \phi)(e^{kL} - 1) - V_{ds}}{2\sinh(kL)} \quad (12)$$

As illustrated below, the quasi-fermi-potential changes along the channel direction are used to calculate the drain current from the source to the drain.

$$\phi(r, z) = \phi_s(z) + \frac{C_{OX}}{2\epsilon_{Si}R} (V_{gs} - V_{fb} - \phi_s(z))(r^2 - R^2) \quad (13)$$

As seen below, the subthreshold current is determined using a 2-D potential relation.

$$I_{sub} = 2\pi R \mu q n_i \frac{\int_{V_s}^{V_d} e^{-qV(z)/KT} dV(z)}{\int_0^L \frac{dz}{\int_0^R e^{q\phi(r,z)/KT} dr}} \quad (14)$$

2 Threshold Voltage (V_{th}) Modeling

For a p-channel MOSFET device, threshold voltage (V_{th}) in an enhancement mode can be obtained [43] using Eq. (15).

$$V_{th} = V_{(T,0)} + \gamma \left(\sqrt{|V_{SB} + 2\phi_F|} - \sqrt{|2\phi_F|} \right) \quad (15)$$

where V_{th} is the threshold voltage, $2\phi_F$ is the surface potential, V_{SB} is the source-to-body substrate bias, $V_{(T,0)}$ is the zero substrate bias threshold voltage, and (γ) is a constant body effect parameter given by;

$$\gamma = (t_{ox}/\epsilon_{ox}) \sqrt{2q\epsilon_{Si}N_A} \quad (16)$$

Here, t_{ox} is oxide thickness, ϵ_{ox} is the relative permittivity of oxide, N_A is the doping concentration, q is the charge of an electron, and ϵ_{Si} is the relative permittivity of silicon semiconductors.

Temperature affects the threshold voltage of a CMOS device, in addition to how oxide thickness affects threshold voltage as shown in Eq. (2);

$$\phi_F = \frac{KT}{q} \ln \left(\frac{N_A}{n_i} \right) \quad (17)$$

Where n_i is the silicon intrinsic doping Parameter, k is Boltzmann's constant, ϕ_F is the contact potential, and T is Temperature [12].

$$n_i = 5.2 \times 10^{15} x T^{3/2} x \exp \left(\frac{-E_g}{2KT} \right) \quad (18)$$

Here E_g is the bandgap energy of the silicon channel material. For GAA-JL-SiNWFET, the equation for the threshold voltage depending upon the device's radius is given by Eq. (19) [44].

$$V_{th} = \Delta\phi + \frac{KT}{q} \ln \left(\frac{8KT\epsilon_{Si}}{q^2 n_i} \right) - \frac{2KT}{q} \ln \left[R \left(\frac{1+t_{ox}}{R} \right)^{\frac{2\epsilon_{Si}}{\epsilon_{ox}}} \right] \quad (19)$$

Here R is the device's radius, and the work function difference is $\Delta\phi$.

Summary of fabrication flowchart for the proposed device [24] Fig. 3.

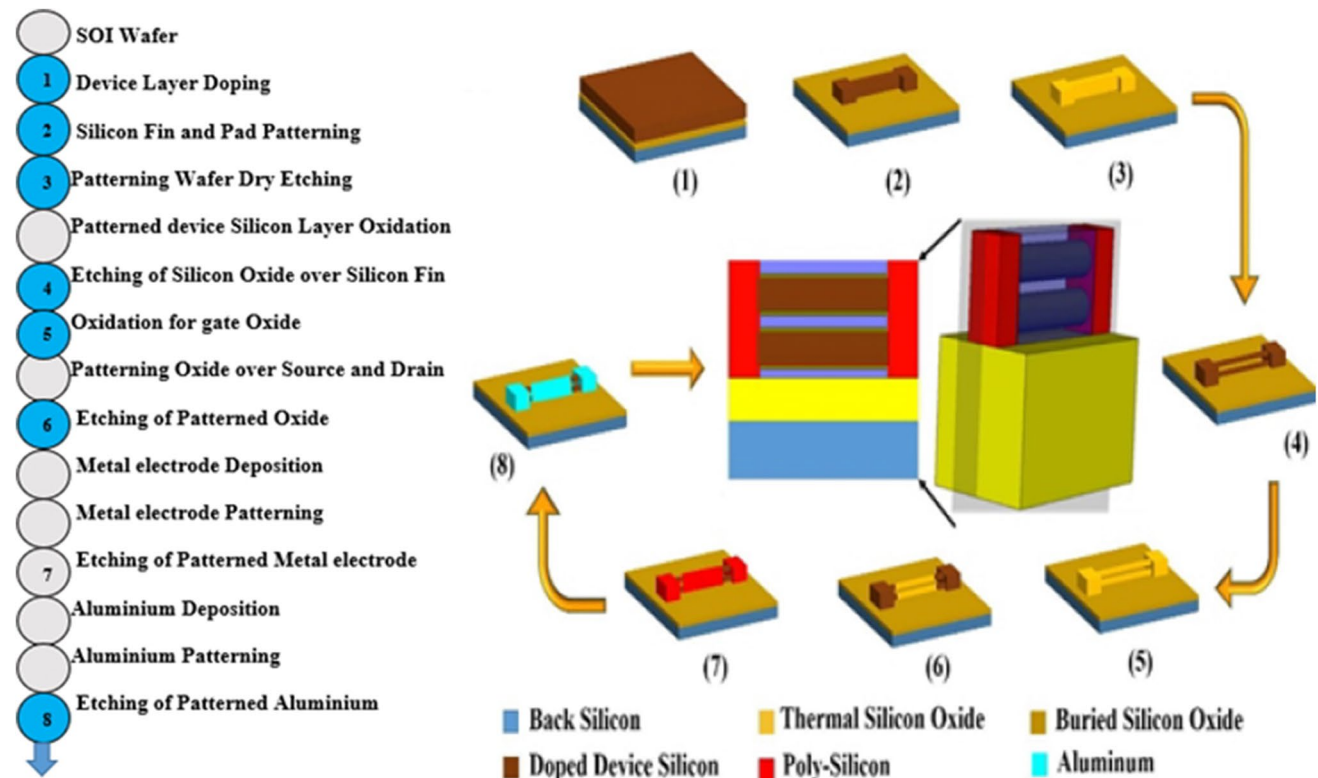
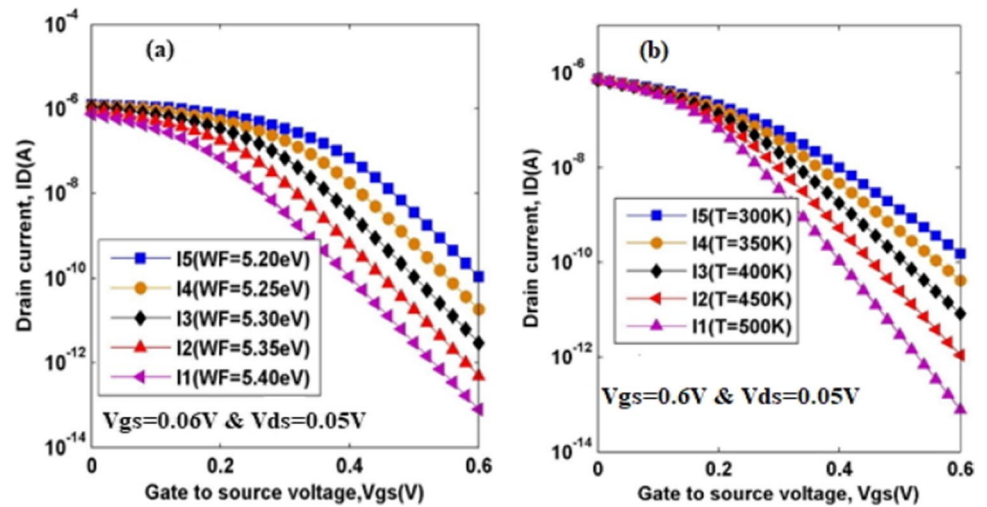


Fig. 3 Describes the fabrication process flowchart and schematic view for the JL-GAA-SiNWFET fabrication process, adapted [24]

Fig. 4 illustrates the impact of (a) palladium (Pd) work function and (b) Temperature on I_D - V_G for p-type substrate cylindrical JL-GAA-SiNWFET



3 Results and Discussion

3.1 Change in drain current

In this work, a change in drain current can also be considered a critical characteristic for identifying hydrogen gas molecules. The change in drain current for a p-type gate-all-around junctionless SiNWFET sensor with a palladium metal gate work function and temperature variation is depicted in Fig. 4. The work function of the catalytic metal gate is controlled by the chemical reaction of hydrogen gas molecules on its surface [22]. In this case, the device's hydrogen gas sensitivity is expressed as a change in the threshold voltage and drain current [20]. For instance, shifting in drain current for the proposed device when the work function changes from 5.20eV to 5.40eV is 1.08×10^{-10} A and when the temperature varies from 300K to 500K is 1.0×10^{-8} A, as illustrated in Fig. 4a and b, respectively. In both cases, OFF-current changes rapidly in puny inversion region and is inversely proportional to hydrogen gas

concentration due to the impact of metalwork function and temperature variation. As a result, the subthreshold zone provides substantially higher sensitivity while operating at low power, resulting in a low-cost hydrogen gas sensor device. This enhanced sensitivity in the subthreshold region is attributable to different band bending in the nonappearance of Fermi level restraining caused by a shift in palladium metal gate work function following hydrogen gas molecule surface reactivity [9]. We conclude that the proposed device will be desirable for detecting hydrogen gas molecule leaks that could have severe impacts, like an explosion, and the device's sensitivity is obtained Eq. (20).

a) Change in surface potential

Figure 5a shows the change in surface potential induced by a shift in the palladium metal work function (5.20 and 5.40 eV). The work function of the catalytic metal gate is altered by the reactivity of hydrogen gas

Fig. 5 impact of (a) palladium (Pd) work function and (b) Temperature on center potential (V) for p-type substrate cylindrical JL-GAA-SiNWFET

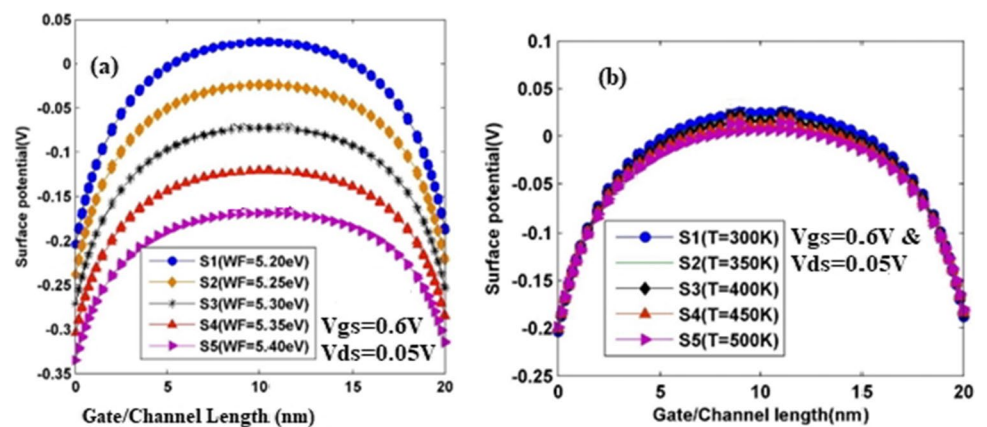
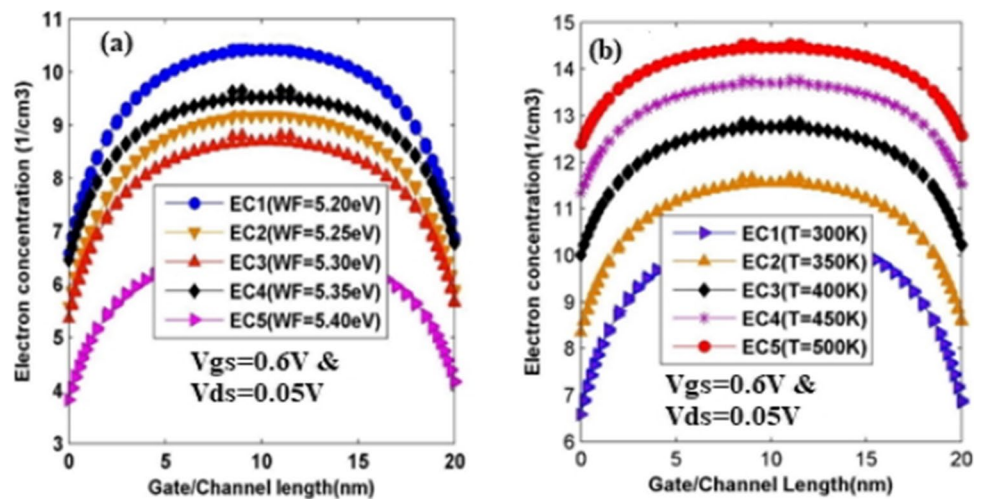


Fig. 6 influence of (a) palladium (Pd) gate work- function and (b) Temperature variation on electron concentration ($1/\text{cm}^3$) for p-type substrate cylindrical JL-GAA-SiNWFET



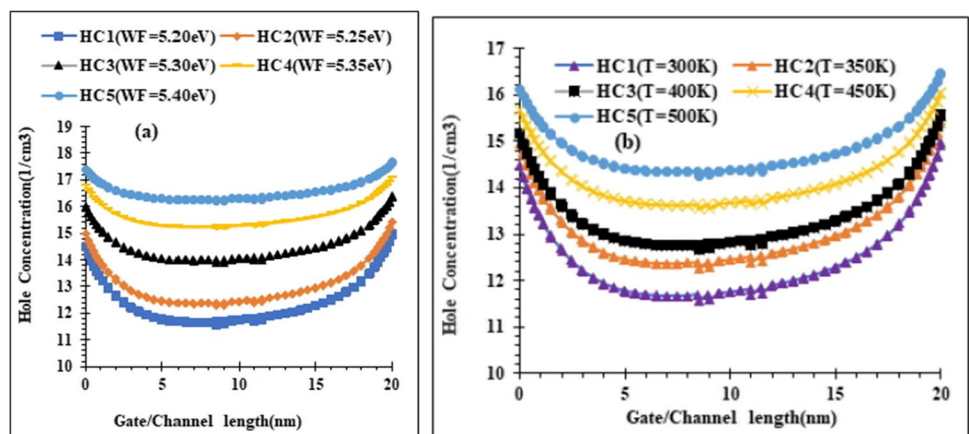
molecules at the gate surface, resulting in further band bending and a change in flat-band voltage, as indicated by Eq. (8) [3, 9].

Figure 5a shows that adjusting the work function impacts the surface potential of p-channel junctionless GAA-SiNWFETs with palladium (Pd) metal gates. The work function of the catalytic metal gate is altered by the reactivity of hydrogen gas molecules at the gate surface, resulting in considerable band bending and a shift in flat-band voltage [3, 5], which causes electrical outputs. Such as drain current, surface potential, and threshold voltage (V_{th}) shift when the flat-band voltage varies [45]. Using a palladium catalytic metal gate, it is feasible to sense the existence of hydrogen gas molecules by measuring the shifting of I_{OFF} , switching ratio, and V_{th} , as clearly described in Fig. 6. Variation of temperature also impacts surface potential, as depicted in Fig. 5b) and significantly represents the proposed device sensing capability.

b) Change in electron mobility

The electron mobility throughout the channel was also extracted, as shown in Fig. 6. The change in electron concentration due to the shift in palladium metal work function (5.20 and 5.40 eV) is examined in Fig. 6a). The work function of the catalytic metal gate is altered by the reactivity of hydrogen gas molecules at the gate surface, leading to different band bending and a shift in flat-band voltage [3, 9], causing mobility of electrons across the channel. As seen in Fig. 6), the shift in electron concentration in the channel region is substantially more significant than in the source and drain regions. Because the electric field in the channel is affected by electron concentration and the flow of charges in the channel [24]. Figure 6b) shows the impact of temperature on electron concentration for the proposed device.

Fig. 7 effect of (a) palladium (Pd) work function and (b) Temperature variation on hole concentration ($1/\text{cm}^3$) for p-type substrate cylindrical JL-GAA-SiNWFET



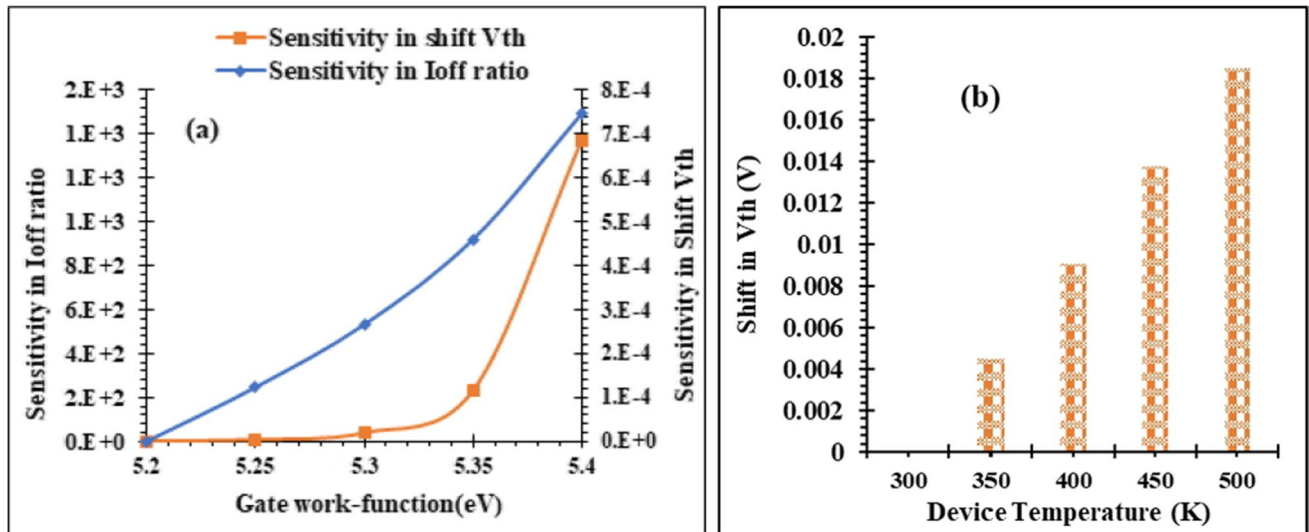


Fig. 8 effect of (a) palladium (Pd) metal gate work and (b) Temperature on I_{OFF} -current ratio for p-type substrate cylindrical JL-GAA-SiNWFET

c) Change in hole mobility

The hole mobility throughout the channel was also extracted, as shown in Fig. 7. The change in hole concentration due to the shift in palladium metal work function (5.20 and 5.40 eV) is examined in Fig. 7a). The reactivity of hydrogen gas molecules at the catalytic metal gate surface alters the gate metal's work function, resulting in further band bending and a change in flat-band voltage [3, 9]. As shown in Fig. 7), the channel's hole concentration difference is considerably less than in the source and drain regions. Because the electric field in the channel is affected by hole concentration, the flow of charges in the channel is also influenced, resulting in drain current and, eventually, device sensitivity. The effect of temperature on hole concentration is depicted in Fig. 7b), and device performance should be significant at room temperature. We examine that the proposed technology has shown to be promising for hydrogen gas detection applications.

d) Shifting in drain current and threshold voltage

The impact of palladium (Pd) work function and temperature variations on device sensitivity are investigated to assess device performance and stability shown in Fig. 8. Figure 8a reflects the sensitivity of gate all around junctionless SiNWFETs as a palladium metal work function in terms of I_{off} ratio and shift in threshold voltage (ΔV_{th}). It can be shown that gate all around junctionless SiNWFETs has better sensitivity at higher palladium (Pd) metalwork functions [41]. Since the flat-band voltage changes as the palladium gate's metal work function rise due to higher band bending. Due to variations in the palladium metal gate work function, a

change in flat-band voltage induces a shift in drain current, threshold voltage (V_{th}), and [9]. It is thus feasible to identify the existence of hydrogen gas molecules by monitoring changes in I_{ON} , ΔV_{th} , and I_{OFF} .

Consequently, some hydrogen atoms diffuse through the gate metal, eventually producing the dipole at and within the interface by changing the metalworking function. In this regard, we have examined the I_{ON}/I_{OFF} ratio, drain-off sensitivity ($S_{I_{off}}$), and shift in threshold voltage of the proposed devices extracting those output results during the simulation, and factors can be regarded as sensitivity variables. We have also shown electron and hole mobility and potential surface distribution along the channel, and carrier transport mechanism has been obtained through NEGF model simulations to obtain the drain current, surface potential, electron and hole mobility and subsequently threshold voltage concerning variation Pd work function and temperature.

When the work function is increased, sensitivity changes exponentially, as seen in Fig. 8a), and it may be estimated using Eq. (20).

$$S_{I_{OFF}} = \frac{I_{OFF(after\ gas\ reaction)} - I_{OFF(before\ gas\ reaction)}}{I_{OFF(before\ gas\ reaction)}} \quad (20)$$

Another essential parameter employed in detecting gas molecules is shifting threshold voltage (ΔV_{th}) and defined as the difference between the threshold voltage with and without hydrogen gas adsorption is defined as (ΔV_{th}) is depicted in Fig. 8) as a function of palladium metal gate work function and temperature. Higher (ΔV_{th}) and $S_{I_{OFF}}$ (shown in Fig. 8) reflects higher palladium metal gate and temperature values, indicating that JL-SiNWFET is well suited for hydrogen gas sensing. As the palladium metal work function

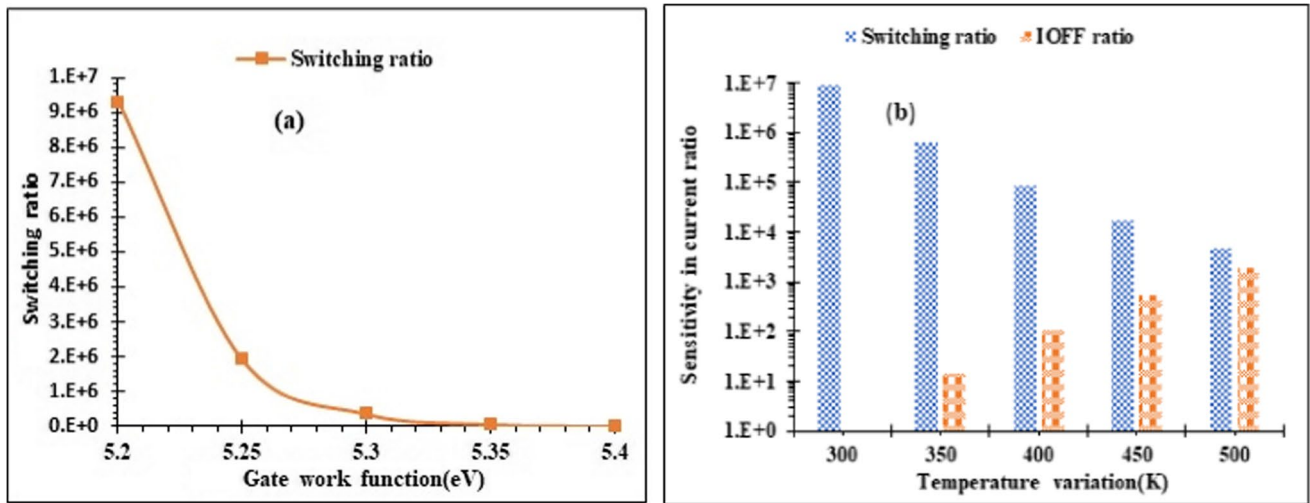


Fig. 9 effect of (a) palladium (Pd) metal gate work function on switching ratio and (b) Temperature on shifting threshold voltage for p-type substrate cylindrical JL-GAA-SiNWFET

and temperature increase, a shift in threshold voltage (V_{th}) arise, resulting in increased hydrogen gas molecule concentration, as seen in Fig. 8), which can be calculated using Eq. (21).

$$\Delta V_{th} = \left| V_{th(after\ gas\ reaction)} - V_{th(before\ gas\ reaction)} \right| \quad (21)$$

The impact of palladium (Pd) work function and temperature variations on device switching ratio using different Pd work functions and temperatures to assess device performance and stability as illustrated in Fig. 9. Figure 9a reflects the impact of varying palladium metal

work functions on the switching ratio for junctionless gate all around SiNWFETs device. Our proposed device has a lower switching ratio at higher palladium (Pd) metal work functions. Sensitivity in the switching ratio is lowered as temperature rises (as illustrated in Fig. 9b); sensitivity in terms of I_{OFF} ratio increases as temperature increases (Fig. 9b).

Figure 10 effect of palladium (Pd) work function variation on (a) switching ratio and (b) leakage current reflects the impact of changing the work function on the switching ratio and leakage current characteristics on our suggested device. Figure 10) illustrates the analytical model validation with

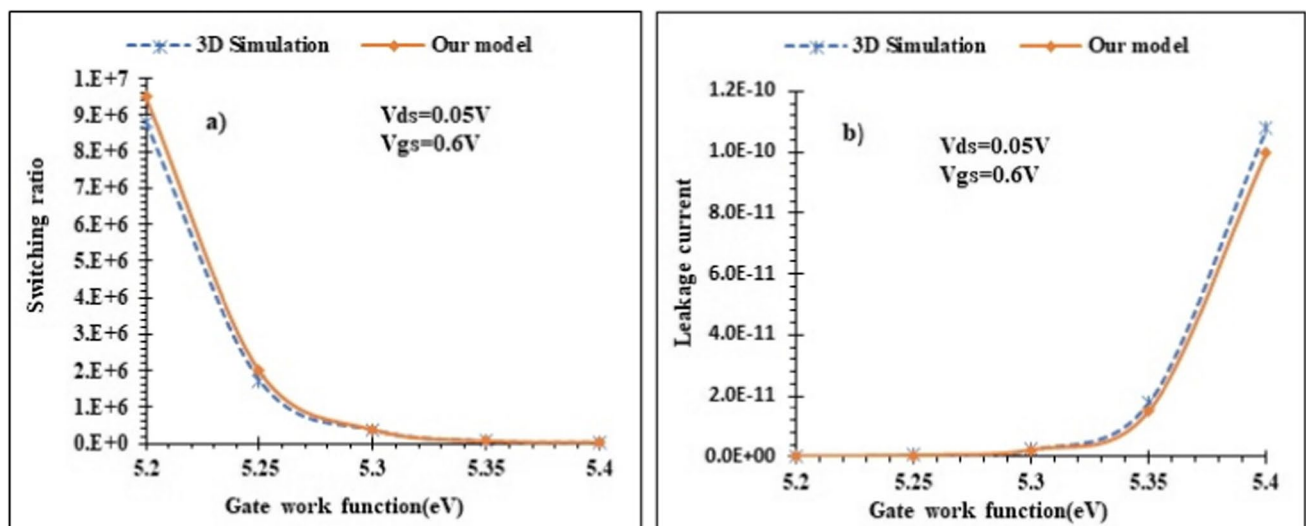


Fig. 10 effect of palladium (Pd) work function variation on (a) switching ratio and (b) leakage current for p-type substrate cylindrical JL-GAA-SiNWFET

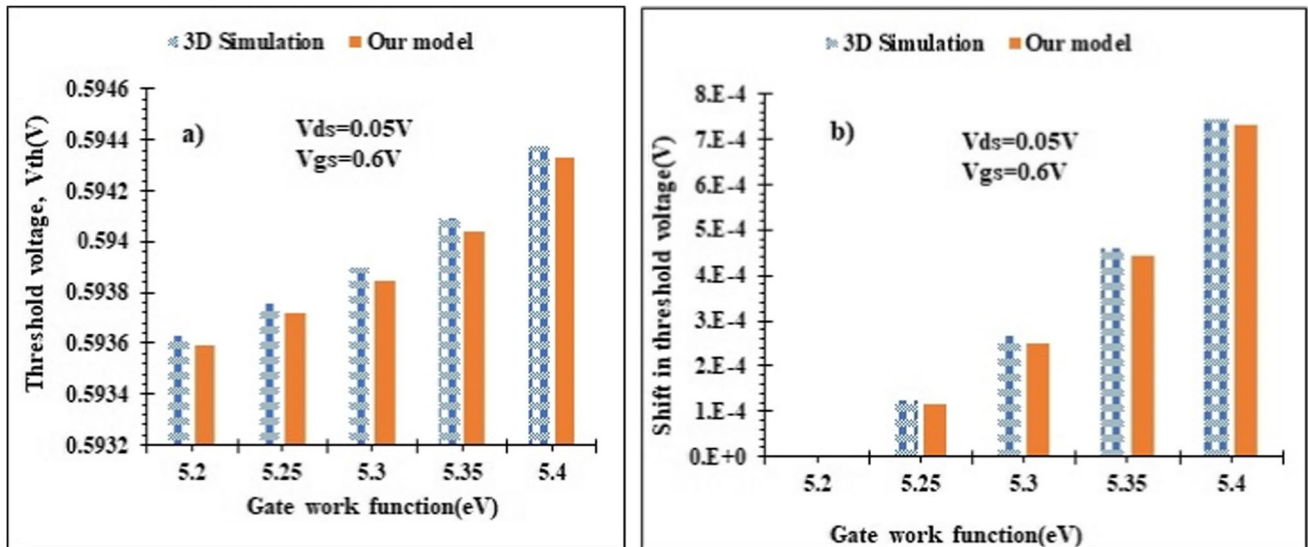


Fig. 11 impact of palladium (Pd) work function variation on (a) threshold voltage (V_{th}) and (b) shifting on threshold voltage (ΔV_{th}) for p-type substrate cylindrical JL-GAA-SiNWFET

the simulation results, which is validated more by the close proximity of our proposed devices.

Figure 11) illustrates the influence of changing the work function on the threshold voltage and shifting threshold voltage characteristics on our suggested device. Figure 11) clearly demonstrates that the analytical model is validated more by the proximity of our proposed device's analytical and simulated results.

Table 2 Examines sensitivity comparison of bulk MOSFET, GAA MOSFET, and JL-GAA-SiNWFET device concerning Off-state current for hydrogen gas sensor after and before gas reaction generated by gas molecules when the threshold voltage of all devices was adjusted at the same

Table 2 Pd gate sensitivity comparison shows the p-type substrate of bulk MOSFET, GAA-MOSFET, and JL-GAA-SiNWFET

$$S_{I_{OFF}} = \frac{I_{OFF(after\ gas\ reaction)} - I_{OFF(before\ gas\ reaction)}}{I_{OFF(before\ gas\ reaction)}}$$

	Previously designed device [5]		Proposed device
Shifting in Pd work function	Bulk-MOSFET $t_{Si}=20nm$	GAA MOSFET $t_{Si}=20nm$ R=10nm	JL-GAA-SiNWFET $t_{Si}=10nm$ R=5nm
$\Delta\Phi_m=50mV$	5.08	5.96	6.17
$\Delta\Phi_m=100mV$	4.56	33.10	37.80
$\Delta\Phi_m=150mV$	102	151	229

Device parameters: Drain, Source, and Channel doping (N_{Si})= $10^{19}cm^{-3}$, Oxide thickness is 1.5&0.3nm, oxide dielectric constants (HfO_2 & SiO_2 are 25.0 & 3.90, respectively), channel length(L)=40nm, drain to source voltage (V_{DS})=0.05V, gate to source voltage (V_{GS})=0.6V, and radius(R)=5nm.

values. When the sensitivity of JL-GAA-SiNWFET was compared to the sensitivity of bulk MOSFET and GAA MOSFET, the sensitivity was found to be more in JL-GAA-SiNWFET because the sensitivity ($S_{I_{off}}$) equation tells us that the hole mobility is related to the subthreshold leakage current. This provides that the subthreshold current in bulk MOSFET and GAA MOSFET devices is higher than JL-GAA-SiNWFET. Since the JL-GAA-SiNWFET structure experience, a higher surface-to-volume ratio and its channel exposed to more effective gate control than others at gate-source voltage are zero, resulting in a more significant variation in subthreshold current when the work function of the gate metal was altered as the gas molecules react with the catalytic metal gate. For instance, the sensitivity ($S_{I_{OFF}}$), of proposed JL-GAA-SiNWFET compared with GAA-MOSFET and bulk MOSFET, JL-GAA-SiNWFET shows improved sensitivity. The results show that as 150mV work function shift of Pd at the gate, the sensitivity improvement with JL-GAA-SiNWFET based hydrogen gas sensors is 51.65% and 124.51% compared with GAA-MOSFET and MOSFET, respectively. Due to high dielectric oxide(HfO_2) and interface oxide(SiO_2) suppressing electron tunneling and hole mobility at gate-source voltage vanishes.

Finally, we have summarized the results here; as we have studied different articles, hydrogen is one of the essential future clean energy sources on the road to a more sustainable world and replacing fossil fuels [45, 46]. For instance, the availability of hydrogen may serve as one of the primary drivers of the energy shift and decarbonization [37]. In order to handle hydrogen safely, robust sensors are highly desired. Particularly, it has been shown that the active materials

(Palladium) employed in these sensors exhibit the high sensitivity to H_2 required for practical applications and that nanostructuring of these materials enables a reduction in response time of the sensors and a close approximation to the industry standard [22]. Due to these and other applications, we studied and designed a Palladium gate modulated JL-GAA-SiNWFET based hydrogen sensor, and it is crucial in applications where health is of particular significance due to their unique qualities, particularly their innately low fire risk, making them the technology of choice; for instance, in mass transit hydrogen-powered vehicles and H_2 accidental leakage [5]. Integrating the palladium electrode in the proposed device enhances device sensitivity performance, lifetime, and reliability. Therefore palladium electrode material is a very sensitive and selective material for H_2 and does not require oxygen to carry out [45]. For various applications, Palladium JL-GAA-SiNWFET based hydrogen sensor has been examined; since hydrogen is odorless compared to gasoline fuel, it is used for the detection of H_2 leakage in the area of hydrogen fueling stations [37], hydrogen pipelines distribution and transmissions [15], cryogenic hydrogen storage tanks (since such storage tanks constantly release H_2 , leads to change in partial pressure), hydrogen fuel cells (utilized in automobiles and can serve as a backup for generators and small power plants) [41], its water resistance (since, the majority of fuel cells operate with surplus liquids, including water) [20], hydrogen safety and control (increased use of hydrogen fuel leads to more H_2 infrastructure incidents) [15], and widely used in industrial settings due to their dependability and excellent sensitivity of H_2 . Because of the increasing need for hydrogen fuel, efficient hydrogen detection is crucial in many industries for everyday safety and process development. Not only these, Junctionless Gate-All-Around SiNWFET-based hydrogen gas sensor is excellent electrostatic control of short channel effects (SCEs). Due to these and other physical significance, we have studied Palladium integrated junctionless gate all around SiNWFET-based hydrogen gas sensor.

4 Conclusion

Through Silvaco-TCAD simulations and analytical model development, this work verified a Junctionless GAA silicon Nanowire transistor with a palladium (Pd) metal gate as a viable sensor for detecting hydrogen gas based on an electrical detecting approach. The resulting analytical model's shifting threshold voltage (ΔV_{th}) and shifting subthreshold current (S_{IOFF}) sensitivities are consistent with simulation data. These results indicate higher sensitivity values at higher palladium metal work function and temperature variations due to increased gas surface covering over the Pd metal gate $|\Delta\phi_m|$. We have examined that the catalytic palladium metal gate JL-GAA-SiNWFET sensor has a higher hydrogen gas molecule sensitivity than GAA-MOSFET and conventional bulk MOSFET due to its larger surface-to-volume ratio in addition to improved performance.

As is confirmed in Table 2, the percentage improvement in the subthreshold drain current ratio's sensitivity (S_{IOFF}) are 124.51% and 51.65% when JL-GAA-SiNWFET compared with bulk MOSFET and GAA-MOSFET, respectively. So, the sensitivity parameter for hydrogen gas sensing in this investigation involves a change in subthreshold current, and it is a critical concern in addition to shifting threshold voltage (ΔV_{th}). This finding provides novel promises for using Pd island gate junctionless gates all around SiNW field-effect transistor sensors to detect hydrogen gas and is applicable for industries such as petrochemical plants, nuclear reactors, hydrogen manufacturing facilities, petroleum refineries, space launching, leak detection, fuel cells, medical diagnostics, and nuclear power plants.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s12633-022-02242-0>.

Acknowledgment The authors thank Ethiopia's Ministry of Higher Education and Microelectronics Research Laboratory, Delhi Technological University, for supporting the work.

Data Availability The authors mentioned above have all relevant data related to this study effort and will be willing to disclose it if asked to do so in the future.

Authors' Contribution

All of the authors contributed to the study's inception and design.

Compliance with Ethical Standards The authors have reviewed all of the Ethical Standards and are expected to adhere to them in the future.

Conflict of Interests The authors declare they have no competing interests.

Consent to Participate & for the Publication Since the study report in question is for a 'non-life science journal,' So 'Not Applicable' in this case.

References

- Gu H, Wang Z, Hu Y (2012) Hydrogen gas sensors based on semiconductor oxide nanostructures. *Sensors (Switzerland)* 12:5517. <https://doi.org/10.3390/s120505517>
- Kim BJ, Kim JS (2013) Dual MOSFET hydrogen sensors with thermal island structure. *Key Eng Mater* 543:93. <https://doi.org/10.4028/www.scientific.net/KEM.543.93>
- Pratap Y, Kumar M, Gupta M, Haldar S, Gupta RS, Deswal SS (2016) Sensitivity investigation of gate-all-around Junctionless transistor for hydrogen gas detection, 2016 IEEE Int. Nanoelectron Conf 1:1. <https://doi.org/10.1109/inec.2016.7589308>
- Cao A, Sudhölter EJR, de Smet LCPM (2013) Silicon nanowire-based devices for gas-phase sensing. *Sensors (Switzerland)* 14:245. <https://doi.org/10.3390/s140100245>
- Gautam R, Saxena M, Gupta RS, Gupta M (2013) Gate-all-around nanowire MOSFET with catalytic metal gate for gas sensing applications. *IEEE Trans Nanotechnol* 12:939. <https://doi.org/10.1109/TNANO.2013.2276394>
- Madan J, Pandey R, Chaujar R (2020) Conducting polymer based gas sensor using PNIN- gate all around - tunnel FET. *Silicon*. <https://doi.org/10.1007/s12633-020-00394-5>

7. Kim JS, Kim BJ (2013) Highly sensitive and stable Mosfet-type hydrogen sensor with dual Pt-Fets. *Nanosci Nanotechnol Lett* 8:43–47. <https://doi.org/10.1166/nnl.2016.2097>
8. Wang Z, Lee S, Koo K, Kim K (2016) Nanowire-based sensors for biological and medical applications. *IEEE Trans Nanobioscience* 15:186. <https://doi.org/10.1109/TNB.2016.2528258>
9. Mokkapati S, Jaiswal N, Gupta M, Kranti A (2019) Gate-all-around nanowire Junctionless transistor-based hydrogen gas sensor. *IEEE Sensors J* 19:4758. <https://doi.org/10.1109/JSEN.2019.2903216>
10. Van Toan N, Viet Chien N, Van Duy N, Si Hong H, Nguyen H, Duc Hoa N, Van Hieu N (2016) Fabrication of highly sensitive and selective H₂ gas sensor based on SnO₂ thin film sensitized with micro-sized Pd Islands. *J Hazard Mater* 301:433. <https://doi.org/10.1016/j.jhazmat.2015.09.013>
11. Sharma B, Kim JS (2018) MEMS based highly sensitive dual FET gas sensor using graphene decorated Pd-ag alloy nanoparticles for H₂ detection. *Sci Rep* 8:1. <https://doi.org/10.1038/s41598-018-24324-z>
12. Park KY, Kim MS, Choi SY (2015) Fabrication and characteristics of MOSFET protein chip for detection of ribosomal protein. *Biosens Bioelectron* 20:2111. <https://doi.org/10.1016/j.bios.2004.08.037>
13. Sze (2014) Performance investigation of dual material gate stack Schottky-barrier source/drain GAA MOSFET for analog applications. *Phys Semicond Devices, Environ Sci Eng* 10:739. <https://doi.org/10.1007/978-3-319-03002-9>
14. Generalov VM, Naumova OV, Fomin BI (2019) Detection of Ebola virus VP40 protein using a nanowire SOI biosensor. *Optoelectron Instrum Data Process* 55:618. <https://doi.org/10.3103/S875669901906013X>
15. Najjar YS (2019) Hydrogen leakage sensing and control: (review). *Biomed J Sci Tech Res* 21(16228). <https://doi.org/10.26717/bjstr.2019.21.003670>
16. Gupta N, Chaujar R (2016) Influence of gate metal engineering on small-signal and noise behaviour of silicon nanowire MOSFET for low-noise amplifiers. *Appl Phys A Mater Sci Process* 122:1. <https://doi.org/10.1007/s00339-016-0239-9>
17. Kumar A, Gupta N, Tripathi MM, Chaujar R (2020) Analysis of structural parameters on sensitivity of black phosphorus Junctionless Recessed Channel MOSFET for biosensing application. *Microsyst Technol* 26:2227. <https://doi.org/10.1007/s00542-019-04545-6>
18. Gupta N, Chaujar R (2016) Investigation of temperature variations on analog/RF and linearity Performance of stacked gate GEWE-SiNW MOSFET for improved device reliability. *Microelectron Reliab* 64:235. <https://doi.org/10.1016/j.microrel.2016.07.095>
19. Madan J, Chaujar R (2016) Palladium gate all around - hetero dielectric -tunnel FET based highly sensitive hydrogen gas sensor. *Superlattice Microst* 1. <https://doi.org/10.1016/j.spmi.2016.09.050>
20. Kim CH, Cho IT, Shin JM, Choi KB, Lee JK, Lee JH (2014) A new gas sensor based on MOSFET having a horizontal floating-gate. *IEEE Electron Device Lett* 35:265. <https://doi.org/10.1109/LED.2013.2294722>
21. Sundgren H, Lundström I, Winquist F, Lukkari I, Carlsson R, Wold S (1990) Evaluation of a multiple gas mixture with a simple MOSFET gas sensor Array and pattern recognition. *Sensors Actuators B Chem* 2:115. [https://doi.org/10.1016/0925-4005\(90\)80020-Z](https://doi.org/10.1016/0925-4005(90)80020-Z)
22. Kumar A (2020) Palladium-based trench gate MOSFET for highly sensitive hydrogen gas sensor. *Mater Sci Semicond Process* 120:105274. <https://doi.org/10.1016/j.mssp.2020.105274>
23. Raad BR, Tirkey S, Sharma D, Kondekar P (2017) A new design approach of Dopingless tunnel FET for enhancement of device characteristics. *IEEE Trans. Electron Devices* 64:1830. <https://doi.org/10.1109/TED.2017.2672640>
24. Peesa RB, Panda DK (2021) Rapid detection of biomolecules in a junction less tunnel field-effect transistor (JL-TFET) biosensor. *Silicon* 4. <https://doi.org/10.1007/s12633-021-00981-0>
25. Veera Boopathy E, Raghu G, Karthick K, Power L, High-Performance MOSFET (2015) International conference on VLSI systems, architecture, technology and applications, VLSI-SATA 2015. Vol. 2(2015). <https://doi.org/10.1109/VLSI-SATA.2015.7050455>
26. Guerfi Y, Larrieu G (2016) Vertical silicon nanowire field effect transistors with nanoscale gate-all-around. *Nanoscale Res Lett* 11:1. <https://doi.org/10.1186/s11671-016-1396-7>
27. Hossain NMM, Quader S, Siddik AB, Chowdhury MIB (2017) TCAD based Performance analysis of Junctionless cylindrical double gate all around FET up to 5nm technology node, 20th Int. Conf Comput Inf Technol ICCIT 2017:1. <https://doi.org/10.1109/ICCITECHN.2017.8281858>
28. Chowdhury MIB, Islam MJ, Islam MJ, Hasan MM, Farwah SU (2016) Silvaco TCAD based analysis of cylindrical gate -all-around FET having indium arsenide as channel and Aluminium oxide as gate dielectrics. *J Nanotechnol Its Appl Eng* 1:1
29. Pratap Y, Kumar M, Kabra S, Haldar S, Gupta RS, Gupta M (2018) Analytical modeling of gate-all-around Junctionless transistor based biosensors for detection of neutral biomolecule species. *J Comput Electron* 17:288. <https://doi.org/10.1007/s10825-017-1041-4>
30. Getnet M, Chaujar R (2022) Sensitivity analysis of biomolecule Nanocavity immobilization in a dielectric modulated triple - hybrid metal gate - all-around Junctionless NWFET biosensor for detecting various diseases. *J Electron Mater*. <https://doi.org/10.1007/s11664-022-09466-1>
31. Chong C, Liu H, Wang S, Chen S, Xie H (2021) Sensitivity analysis of biosensors based on a dielectric-modulated I-shaped gate field-effect transistor. *Micromachines* 12:1. <https://doi.org/10.3390/mi12010019>
32. Kumar P, Sharma SK, Balwinder R (2021) Comparative analysis of nanowire tunnel field effect transistor for biosensor applications. *Silicon* 13:4067. <https://doi.org/10.1007/s12633-020-00718-5>
33. Kosmani NF, Hamid FA, Razali MA (2020) Effects of High-k dielectric materials on electrical Performance of double gate and gate-all-around MOSFET. *Int J Integr Eng* 12(81). <https://doi.org/10.30880/ijie.2020.12.02.010>
34. Dhiman G (2019) Investigation of junction - less double gate MOSFET with High - k gate - oxide and metal gate layers. *Int J Innov Res Sci Eng Technol* 8:289
35. Li Q, Zhu H, Yuan H, Kirillov O, Ioannou D, Suehle J, Richter CA (2012) A study of metal gates on HfO₂ using Si nanowire field effect transistors as platform, ECS. Meet Abstr MA2012-02:2614. <https://doi.org/10.1149/ma2012-02/31/2614>
36. Sarangadharan I, Pulikkathodi AK, Chu C-H, Chen Y-W, Regmi A, Chen P-C, Hsu C-P, Wang Y-L (2018) Review—High field modulated FET biosensors for biomedical applications. *ECS J Solid State Sci Technol* 7:Q3032. <https://doi.org/10.1149/2.0061807jss>
37. Koo WT, Cho HJ, Kim DH, Kim YH, Shin H, Penner RM, Kim ID (2020) Chemiresistive hydrogen sensors: fundamentals. Recent Advances, and Challenges, *ACS Nano* 14:14284. <https://doi.org/10.1021/acsnano.0c05307>
38. Siddik AB, Hossain NMM, Quader S, Chowdhury MIB (2018) Silicon on metal technology merged with cylindrical gate all around fet for enhanced performance. In: 3rd International Conference on Electrical Information and Communication Technology (EICT). <https://doi.org/10.1109/EICT.2017.8275181>
39. Ha MW, Seok O, Lee H, Lee HH (2020) Mobility models based on forward current-voltage characteristics of P-Type Pseudo-Vertical diamond schottky barrier diodes. *Micromachines* 11:598. <https://doi.org/10.3390/M11060598>
40. Hung CW, Lin KW, Liu RC, Tsai YY, Lai PH, Fu SI, Chen TP, Chen HI, Liu WC (2007) On the hydrogen sensing properties

- of a Pd/GaAs transistor-type gas sensor in a nitrogen ambience, sensors actuators. B Chem 125:22. <https://doi.org/10.1016/j.snb.2007.01.027>
41. Behzadi pour G, Fekri aval L (2017) Highly sensitive work function hydrogen gas sensor based on PdNPs/SiO₂/Si structure at room temperature. Results Phys 7:1993. <https://doi.org/10.1016/j.rinp.2017.06.026>
 42. Chen ZX, Yu HY, Singh N, Shen NS, Sayanthan RD, Lo GQ, Kwong D (2009) Demonstration of tunneling FETs based on highly scalable vertical silicon nanowires. IEEE Electron Device Lett 30:754
 43. Arefin A (2015). Impact of Temperature on Threshold Voltage of Gate-All-Around Junctionless Nanowire Field-Effect Transistor 6:14
 44. Sgmosfets SM (2008) Continuous Analytic Current-Voltage ($I-V$) Model for Long-Channel Doped J J, 315
 45. Choi JH, Jo MG, Han SW, Kim H, Kim SH, Jang S, Kim JS, Cha HY (2017) Hydrogen gas sensor of Pd-functionalised AlGaIn/GaN Heterostructure with High sensitivity and low-Power consumption. Electron Lett 53:1200. <https://doi.org/10.1049/el.2017.2107>
 46. Gu H, Wang Z, Hu Y (2012) Hydrogen Gas Sensors Based on Semiconductor Oxide Nanostructures 12. <https://doi.org/10.3390/s120505517>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

SHARP THIRD HANKEL DETERMINANT BOUND FOR $\mathcal{S}^*(\alpha)$

NEHA VERMA AND S. SIVAPRASAD KUMAR

ABSTRACT. For $0 \leq \alpha < 1$, a normalized analytic function $f(z)$ is said to be starlike of order α , denoted by $\mathcal{S}^*(\alpha)$, if and only if $\operatorname{Re}\{zf'(z)/f(z)\} > \alpha$ for $z \in \mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$. This paper provides the sharp bound of $H_{3,1}(f)$, third Hankel determinant of functions belonging to the class $\mathcal{S}^*(\alpha)$, given by

$$|H_{3,1}(f)| \leq \frac{4(1-\alpha)^2}{9}$$

with some restrictions on α , which solves a longstanding coefficient problem.

1. INTRODUCTION

Let \mathcal{A} be the class of normalized analytic functions classified on the open unit disk $\mathbb{D} := \{z \in \mathbb{C} : |z| < 1\}$ such that

$$f(z) = z + \sum_{n=2}^{\infty} a_n z^n. \quad (1.1)$$

Let $\mathcal{S} \subset \mathcal{A}$ be the class of univalent functions and \mathcal{P} be the collection of analytic functions defined on \mathbb{D} having positive real part, given by $p(z) = 1 + \sum_{n=1}^{\infty} p_n z^n$. For any two analytic functions, h and g , we say h is subordinate to g or $h \prec g$, if there exists a Schwarz function w with $w(0) = 0$ and $|w(z)| \leq |z|$ such that $h(z) = g(w(z))$.

Bieberbach conjecture [5, Page no. 17] has immensely inspired the growth of univalent function theory and the creation of new coefficient problems since 1916. Consequently, in 1966, Pommerenke [14] introduced the q^{th} Hankel determinants $H_{q,n}(f)$ of analytic functions f in (1.1) for $n, q \in \mathbb{N}$, defined as follows:

$$H_{q,n}(f) = \begin{vmatrix} a_n & a_{n+1} & \cdots & a_{n+q-1} \\ a_{n+1} & a_{n+2} & \cdots & a_{n+q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n+q-1} & a_{n+q} & \cdots & a_{n+2q-2} \end{vmatrix}. \quad (1.2)$$

By appropriately choosing n and q , we come to its particular cases. The expression of second order Hankel determinant is derived by considering $q = 2$. Several studies have established the sharp bound of second-order Hankel determinants for various subclasses of \mathcal{S} , see [4, 6, 8, 9]. If we take $q = 3$ and $n = 1$ in (1.2), the expression of Hankel determinant of order three, is obtained assuming $a_1 := 1$. It is as follows:

$$H_{3,1}(f) := \begin{vmatrix} 1 & a_2 & a_3 \\ a_2 & a_3 & a_4 \\ a_3 & a_4 & a_5 \end{vmatrix} = 2a_2a_3a_4 - a_3^3 - a_4^2 - a_2^2a_5 + a_3a_5. \quad (1.3)$$

In 1936, Robertson [16] introduced and investigated the class of starlike functions of order α , defined as:

2010 *Mathematics Subject Classification.* 30C45, 30C50.

Key words and phrases. Starlike, Sharp, Hankel determinant, Order alpha .

Definition 1.1. [16] For $0 \leq \alpha < 1$, we say that a function $f \in \mathcal{A}$ is starlike of order α if and only if

$$\operatorname{Re} \left(\frac{zf'(z)}{f(z)} \right) > \alpha, \quad z \in \mathbb{D}.$$

The class of all such functions is represented by $\mathcal{S}^*(\alpha)$.

In 1992, Ma and Minda [13] introduced the following class of starlike functions,

$$\mathcal{S}^*(\varphi) = \left\{ f \in \mathcal{A} : \frac{zf'(z)}{f(z)} \prec \varphi(z) \right\},$$

where φ is an analytic univalent function such that $\operatorname{Re} \varphi(z) > 0$, $\varphi(\mathbb{D})$ is symmetric about the real axis and starlike with respect to $\varphi(0) = 1$ with $\varphi'(0) > 0$.

Using subordination, the class $\mathcal{S}^*(\alpha)$ can be characterized as:

$$\mathcal{S}^*(\alpha) = \left\{ f \in \mathcal{A} : \frac{zf'(z)}{f(z)} \prec \frac{1 + (1 - 2\alpha)z}{1 - z} \right\} \quad \text{for } \alpha \in [0, 1].$$

Note that $\mathcal{S}^*(0) = \mathcal{S}^*$ and $\mathcal{S}^*(\varphi) \subset \mathcal{S}^*(\alpha)$ for some α depending upon the choice of φ . The estimation of third order Hankel determinant is more difficult than second order Hankel determinant, see [10, 22]. Some of the sharp estimates of third order Hankel determinant of functions in $\mathcal{S}^*(\varphi)$ for different choices of $\varphi(z)$, are enlisted below:

TABLE 1. List of third order Hankel determinants

Class	Sharp bound	Reference
$\mathcal{S}^* := \mathcal{S}^*(0)$	4/9	[7]
$\mathcal{S}^*(1/2)$	1/9	[15]
$\mathcal{S}_\rho^* := \mathcal{S}^*(1 + ze^z)$	1/9	[19], [10, Conjecture on Page no. 33]
$\mathcal{SL}^* := \mathcal{S}^*(\sqrt{1+z})$	1/36	[1]
$\mathcal{S}_e^* := \mathcal{S}^*(e^z)$	1/9	[18]
$\mathcal{S}_\rho^* := \mathcal{S}^*(1 + \sinh^{-1}(z))$	1/9	[17]
$\mathcal{S}_{Ne}^* := \mathcal{S}^*(1 + z - z^3/3)$	—	—

However, the sharp estimate of $H_{3,1}(f)$ for \mathcal{S}_{Ne}^* is yet to be estimated.

For the class $\mathcal{S}^*(\alpha)$, Krishna and Ramreddy [8] computed the bound of the second order Hankel determinant, $|a_2a_4 - a_3^2| \leq (1 - \alpha)^2$, $\alpha \in [0, 1/2]$ while Xu and Fang [21] calculated the sharp bounds of the Fekete and Szegő functional $|a_3 - \lambda a_2^2| \leq (1 - \alpha) \max\{1, |3 - 2\alpha - 4\lambda(1 - \alpha)|\}$, $\lambda \in \mathbb{C}$ and $\alpha \in [0, 1]$. We refer [2] for further information on Hankel determinants associated with the class $\mathcal{S}^*(\alpha)$.

The purpose of this study is to establish the sharp bound of third order Hankel determinant for functions belonging to the class, $\mathcal{S}^*(\alpha)$. At the end of this article, we point out the validation of our main result by considering the class $\mathcal{S}^*(\alpha)$ for $\alpha = 0$, clubbed with some applications.

2. PRELIMINARIES

Let $f \in \mathcal{S}^*(\alpha)$, then a Schwarz function $w(z)$ exists such that

$$\frac{zf'(z)}{f(z)} = \frac{1 + (1 - 2\alpha)w(z)}{1 - w(z)}. \quad (2.1)$$

Let $p(z) = 1 + \sum_{n=2}^{\infty} p_n z^n \in \mathcal{P}$ and $w(z) = (p(z) - 1)/(p(z) + 1)$. The expressions of a_i ($i = 2, 3, 4, 5$) are obtained in terms of p_j ($j = 1, 2, 3, 4$) by substituting $w(z)$, $p(z)$, and $f(z)$ in equation (2.1) with suitable comparison of coefficients so that

$$a_2 = p_1(1 - \alpha), \quad (2.2)$$

$$a_3 = \frac{(1 - \alpha)}{2} \left(p_2 + p_1^2(1 - \alpha) \right), \quad (2.3)$$

$$a_4 = \frac{(1 - \alpha)}{6} \left(2p_3 + 3p_1p_2(1 - \alpha) + p_1^3(1 - \alpha)^2 \right), \quad (2.4)$$

and

$$a_5 = \frac{(1 - \alpha)}{24} \left\{ 6p_4 + (1 - \alpha) \left(3p_2^2 + 8p_1p_3 \right) + (1 - \alpha)^2 \left(6p_1^2p_2 + p_1^4(1 - \alpha) \right) \right\}. \quad (2.5)$$

The formula for p_j ($j = 2, 3, 4$), which plays a significant role in finding the sharp bound of the Hankel determinant and has been prominently exploited in the main theorem, is contained in the Lemma 2.1 below. For further details on the class \mathcal{P} coefficients, refer to the survey article, [3].

Lemma 2.1. [11, 12] *Let $p \in \mathcal{P}$ has the form $1 + \sum_{n=1}^{\infty} p_n z^n$. Then*

$$2p_2 = p_1^2 + \gamma(4 - p_1^2),$$

$$4p_3 = p_1^3 + 2p_1(4 - p_1^2)\gamma - p_1(4 - p_1^2)\gamma^2 + 2(4 - p_1^2)(1 - |\gamma|^2)\eta,$$

and

$$8p_4 = p_1^4 + (4 - p_1^2)\gamma(p_1^2(\gamma^2 - 3\gamma + 3) + 4\gamma) - 4(4 - p_1^2)(1 - |\gamma|^2)(p_1(\gamma - 1)\eta + \bar{\gamma}\eta^2 - (1 - |\eta|^2)\rho),$$

for some γ , η and ρ such that $|\gamma| \leq 1$, $|\eta| \leq 1$ and $|\rho| \leq 1$.

3. SHARP $H_{3,1}$ FOR $\mathcal{S}^*(\alpha)$

Recently, Kowalczyk [7] and Rath et al. [15] determined the sharp bound of Hankel determinant of order three for the functions in the class $\mathcal{S}^*(\alpha)$ when $\alpha = 0$ and $\alpha = 1/2$, respectively. In this section, we calculate the sharp bound of $H_{3,1}(f)$ for the functions belonging to the class $\mathcal{S}^*(\alpha)$, for some more values of α . The graphs provided in this article are all prepared using MATLAB.

Theorem 3.1. *Let $f \in \mathcal{S}^*(\alpha)$. Then*

$$|H_{3,1}(f)| \leq \frac{4(1 - \alpha)^2}{9}, \quad \alpha \in [0, \alpha_0) \cup (\alpha_0, \alpha_1), \quad (3.1)$$

is the sharp bound. Here, $\alpha_0 \in [0, 1)$ is the smallest positive root of $1 - 4\alpha + 6\alpha^2 - 16\alpha^3 + 4\alpha^4 = 0$ and $\alpha_1 \approx 0.370803$.

Proof. Suppose that $p_1 =: p \in [0, 2]$ due to the invariant property of class \mathcal{P} , under rotation. The expressions of a_i ($i = 2, 3, 4, 5$) from equations (2.2)-(2.5) are substituted in equation (1.3). We get

$$H_{3,1}(f) = \frac{(1 - \alpha)^2}{144} \left(- (1 - \alpha)^4 p^6 + 3(1 - \alpha)^3 p^4 p_2 + 8(1 - \alpha)^2 p^3 p_3 + 24(1 - \alpha) p p_2 p_3 - 9(1 - \alpha) p_2^3 \right. \\ \left. - 18(1 - \alpha) p^2 p_4 - 9(1 - \alpha)^2 p^2 p_2^2 - 16p_3^2 + 18p_2 p_4 \right).$$

After simplifying the calculations through Lemma 2.1, we obtain

$$H_{3,1}(f) = \frac{1}{1152} \left(\Delta_1(p, \gamma) + \Delta_2(p, \gamma)\eta + \Delta_3(p, \gamma)\eta^2 + \Phi(p, \gamma, \eta)\rho \right), \quad \text{for } \gamma, \eta, \rho \in \mathbb{D}.$$

Here

$$\begin{aligned}
\Delta_1(p, \gamma) &:= (1 - \alpha)^2 \left(\alpha(1 - 2\alpha)^2(3 - 2\alpha)p^6 - (2 - 15\alpha + 18\alpha^2)p^2\gamma^2(4 - p^2)^2 + p^2\gamma^4(4 - p^2)^2 \right. \\
&\quad - (10 - 15\alpha)p^2\gamma^3(4 - p^2)^2 + 36\alpha\gamma^3(4 - p^2)^2 + (3 - 12\alpha^3 + 32\alpha^2 - 19\alpha)p^4\gamma(4 - p^2) \\
&\quad \left. + (3 - 16\alpha^2 + 2\alpha)p^4\gamma^2(4 - p^2) - 9(1 - 2\alpha)p^4\gamma^3(4 - p^2) - 36(1 - 2\alpha)p^2\gamma^2(4 - p^2) \right), \\
\Delta_2(p, \gamma) &:= 4(1 - |\gamma|^2)(4 - p^2)(1 - \alpha)^2 \left((8\alpha^2 - 10\alpha + 3)p^3 + 9(1 - 2\alpha)p^3\gamma + (5 - 12\alpha)p\gamma(4 - p^2) \right. \\
&\quad \left. - p\gamma^2(4 - p^2) \right), \\
\Delta_3(p, \gamma) &:= 4(1 - |\gamma|^2)(4 - p^2)(1 - \alpha)^2 \left(-8(4 - p^2) - |\gamma|^2(4 - p^2) + 9(1 - 2\alpha)p^2\bar{\gamma} \right), \\
\Phi(p, \gamma, \eta) &:= 36(1 - |\gamma|^2)(4 - p^2)(1 - |\eta|^2)(1 - \alpha)^2 \left((4 - p^2)\gamma - (1 - 2\alpha)p^2 \right).
\end{aligned}$$

Assume $x := |\gamma|$, $y := |\eta|$ and since $|\rho| \leq 1$, the above expression reduces to

$$|H_{3,1}(f)| \leq \frac{1}{1152} \left(|\Delta_1(p, \gamma)| + |\Delta_2(p, \gamma)|y + |\Delta_3(p, \gamma)|y^2 + |\Phi(p, \gamma, \eta)| \right) \leq Z(p, x, y),$$

where

$$Z(p, x, y) = \frac{1}{1152} \left(z_1(p, x) + z_2(p, x)y + z_3(p, x)y^2 + z_4(p, x)(1 - y^2) \right) \quad (3.2)$$

with

$$\begin{aligned}
z_1(p, x) &:= (1 - \alpha)^2 \left(\alpha(1 - 2\alpha)^2(3 - 2\alpha)p^6 + (2 - 15\alpha + 18\alpha^2)p^2x^2(4 - p^2)^2 + p^2x^4(4 - p^2)^2 \right. \\
&\quad + (10 - 15\alpha)p^2x^3(4 - p^2)^2 + 36\alpha x^3(4 - p^2)^2 + (3 - 12\alpha^3 + 32\alpha^2 - 19\alpha)p^4x(4 - p^2) \\
&\quad \left. + (3 - 16\alpha^2 + 2\alpha)p^4x^2(4 - p^2) + 9(1 - 2\alpha)p^4x^3(4 - p^2) + 36(1 - 2\alpha)p^2x^2(4 - p^2) \right), \\
z_2(p, x) &:= 4(1 - x^2)(4 - p^2)(1 - \alpha)^2 \left((8\alpha^2 - 10\alpha + 3)p^3 + 9(1 - 2\alpha)p^3x + (5 - 12\alpha)px(4 - p^2) \right. \\
&\quad \left. + px^2(4 - p^2) \right), \\
z_3(p, x) &:= 4(1 - x^2)(4 - p^2)(1 - \alpha)^2 \left(8(4 - p^2) + x^2(4 - p^2) + 9(1 - 2\alpha)p^2x \right), \\
z_4(p, x) &:= 36(1 - x^2)(4 - p^2)(1 - \alpha)^2 \left((4 - p^2)x + (1 - 2\alpha)p^2 \right).
\end{aligned}$$

We maximise $Z(p, x, y)$ within the closed cuboid $Y : [0, 2] \times [0, 1] \times [0, 1]$, by finding the maximum values in the interior of Y , in the interior of the six faces and on the twelve edges.

Case I:

We begin with every interior point of Y assuming $(p, x, y) \in (0, 2) \times (0, 1) \times (0, 1)$. We determine

$\partial Z/\partial y$ to examine the points of maxima in the interior of Y . Thus

$$\begin{aligned} \frac{\partial Z}{\partial y} = \frac{(4-p^2)(1-x^2)(1-\alpha^2)}{288} & \left(8(8-9x+x^2)y - 2p^2(1-x)y(17-x-18\alpha) \right. \\ & + p^3(3-x^2-10\alpha+8\alpha^2+x(4-6\alpha)) \\ & \left. + 4xp(5+x-12\alpha) \right). \end{aligned}$$

Now, $\frac{\partial Z}{\partial y} = 0$ gives

$$y = y_0 := \frac{4xp(5+x-12\alpha) + p^3(3-x^2-10\alpha+8\alpha^2+x(4-6\alpha))}{2(1-x)(-4(8-x) + p^2(17-x-18\alpha))}.$$

The existence of critical points require that $y_0 \in (0, 1)$ and can only exist when

$$\begin{aligned} 2p^2(1-x)(17-x-18\alpha) & > -p^3(-3+x^2+10\alpha-8\alpha^2-x(4-6\alpha)) \\ & + 4px(5+x-12\alpha) + 8(1-x)(8-x). \end{aligned} \quad (3.3)$$

We try finding the solution satisfying the inequality (3.3) for finding critical points. If we assume $p \rightarrow 0$, then no $x \in (0, 1)$ exists, satisfying the equation (3.3). But, when $p \rightarrow 2$, the equation (3.3) holds for all $x < (3+2\alpha)/9$ and $\alpha \in [0, 1/2)$. Similarly, if we assume $x \rightarrow 0$ and 1, then no such $p \in [0, 2]$ exists so that the equation (3.3) holds. So, we observe that the function Z has no critical point in the desired domain.

Case II:

The interior of six faces of the cuboid Y , is now under consideration, for the further calculations. On $p = 0$, $Z(p, x, y)$ turns into

$$s_1(x, y) := \frac{(1-\alpha)^2((1-x^2)((8+x^2)y^2 + 9x(1-y^2)) + 9x^3\alpha)}{18}, \quad x, y \in (0, 1). \quad (3.4)$$

Since

$$\frac{\partial s_1}{\partial y} = \frac{(1-x^2)(x+1)(8-x)(1-\alpha)^2y}{9} \neq 0, \quad x, y \in (0, 1).$$

Thus, s_1 has no critical point in $(0, 1) \times (0, 1)$.

On $p = 2$, $Z(p, x, y)$ reduces to

$$Z(2, x, y) := \frac{\alpha(1-\alpha)^2(1-2\alpha)^2(3-2\alpha)}{18}, \quad x, y \in (0, 1). \quad (3.5)$$

On $x = 0$, $Z(p, x, y)$ becomes

$$\begin{aligned} s_2(p, y) := \frac{(1-\alpha)^2}{1152} & \left(\alpha(3-2\alpha)(1-2\alpha)^2p^6 + 36(1-2\alpha)p^2(4-p^2)(1-y^2) \right. \\ & \left. + 32(4-p^2)^2y^2 + 4(3-10\alpha+8\alpha^2)p^3y(4-p^2) \right) \end{aligned} \quad (3.6)$$

with $p \in (0, 2)$ and $y \in (0, 1)$. On solving $\partial s_2/\partial p$ and $\partial s_2/\partial y$, to find the points of maxima. After resolving $\partial s_2/\partial y = 0$, we get

$$y = \frac{p^3(3-10\alpha+8\alpha^2)}{2(17p^2-32-18p^2\alpha)} (=: y_0). \quad (3.7)$$

Upon calculations, we observe that to have $y_0 \in (0, 1)$ for the given range of y , $p =: p_0 > \approx A(\alpha)$ is needed with $\alpha \in [0, \beta_0)$. This $\beta_0 \in [0, 1)$ is the smallest positive root of $-3 + 10\alpha - 8\alpha^2 = 0$ and no such $p \in (0, 2)$ exists when $\alpha \in (\beta_0, 1)$. Here,

$$A(\alpha) := \frac{1}{-3 + 10\alpha - 8\alpha^2} \left(-11.3333 + 12\alpha - \frac{(272.143 - 471.365i)(0.944444 - \alpha)^2}{B} - (0.132283 + 0.229122i)B \right),$$

with

$$B := \left(C + \sqrt{-8.70713 \times 10^9 (0.944444 - \alpha)^6 + C^2} \right)^{1/3}$$

and

$$C := -63056 + 146016\alpha - 8640\alpha^2 - 183168\alpha^3 + 110592\alpha^4.$$

Based on computations, $\partial s_2 / \partial p = 0$ gives

$$\begin{aligned} 0 = & 16p(9 - 18\alpha - y^2(25 - 18\alpha)) - 2p^2y(3 - 10\alpha + 8\alpha^2)(5p^2 - 12) \\ & + 3\alpha(1 - 2\alpha)^2(3 - 2\alpha)p^5 - 8p^3(9 - 18\alpha - y^2(17 - 18\alpha)). \end{aligned} \quad (3.8)$$

After substituting equation (3.7) into equation (3.8), we have

$$\begin{aligned} 0 = & p \left(49152(1 - 2\alpha) - 3072p^2(25 - 68\alpha + 36\alpha^2) - p^8(1 - 2\alpha)^2(153 - 1437\alpha + 3118\alpha^2 \right. \\ & - 2484\alpha^3 + 648\alpha^4) + 16p^4(2427 - 7890\alpha + 6020\alpha^2 + 616\alpha^3 - 1024\alpha^4) \\ & \left. - 128p^6(48 - 153\alpha + 20\alpha^2 + 340\alpha^3 - 352\alpha^4 + 96\alpha^5) \right). \end{aligned} \quad (3.9)$$

A numerical calculation suggests that the solution of (3.9) in the interval $(0, 2)$ is $p \approx B(\alpha)$ whenever $\alpha \in [0, \alpha_2)$, where $\alpha_2 \in [0, 1)$ is the smallest positive root of $153 - 1437\alpha + 3118\alpha^2 - 2484\alpha^3 + 648\alpha^4 = 0$, otherwise no such $p \in (0, 2)$ exists, see Fig. 1. Thus, s_2 does not have any critical point in $(0, 2) \times (0, 1)$.

Here

$$\begin{aligned} B(\alpha) := & \frac{1}{\sqrt{2}} \left[\left\{ \frac{1}{F} \left(-3072 + 3648\alpha + 6016\alpha^2 - 9728\alpha^3 + 3072\alpha^4 \right. \right. \right. \\ & - \frac{4F}{\sqrt{3}} \left\{ \frac{1}{E^2} \left(\frac{768J^2}{(1-2\alpha)^2} + \frac{2GE}{(1-2\alpha)} + \frac{HE}{I} + \frac{IE}{(1-2\alpha)^2} \right) \right\}^{1/2} \\ & + \frac{4F}{\sqrt{3}} \left\{ \frac{-1}{E^3} \left(\frac{-1536J^2E}{(1-2\alpha)^2} - \frac{4GE^2}{(1-2\alpha)} + \frac{HE^2}{I} + \frac{IE^2}{(1-2\alpha)^2} \right. \right. \\ & \left. \left. - \frac{K}{(1-2\alpha)^3 \left\{ \frac{1}{E^2} \left(\frac{768J^2}{(1-2\alpha)^2} + \frac{2GE}{(1-2\alpha)} + \frac{HE}{I} + \frac{IE}{(1-2\alpha)^2} \right) \right\}^{1/2}} \right) \right\}^{1/2} \right\}^{1/2} \right], \end{aligned}$$

with

$$E := 153 - 1437\alpha + 3118\alpha^2 - 2484\alpha^3 + 648\alpha^4;$$

$$F := (1 - 2\alpha)E;$$

$$G := 2427 - 3036\alpha - 52\alpha^2 + 512\alpha^3;$$

$$H := 8217 - 173160\alpha + 1260312\alpha^2 - 2415264\alpha^3 + 2091664\alpha^4 - 1048576\alpha^5 + 262144\alpha^6;$$

$$\begin{aligned}
I := & \left\{ 127065213 - 1886889978\alpha + 12579196752\alpha^2 - 49871499552\alpha^3 + 132494582880\alpha^4 \right. \\
& - 253944918720\alpha^5 + 368411062528\alpha^6 - 410152327680\alpha^7 + 340236674304\alpha^8 \\
& - 196757891584\alpha^9 + 71861010432\alpha^{10} - 14168358912\alpha^{11} + 1073741824\alpha^{12} \\
& + 288\sqrt{6} \left((1-2\alpha)^8(3-4\alpha)^2(3604621581 - 39763739565\alpha + 202125486510\alpha^2 \right. \\
& - 633657349224\alpha^3 + 1436021769744\alpha^4 - 2516421142080\alpha^5 + 34829931648\alpha^6 \\
& - 3872882513280\alpha^7 + 3466438619648\alpha^8 - 2402201403136\alpha^9 + 1198713174528\alpha^{10} \\
& \left. \left. - 394099818496\alpha^{11} + 75581358080\alpha^{12} - 6442450944\alpha^{13} \right) \right)^{1/2} \Big\}^{1/3}; \\
J := & -48 + 57\alpha + 94\alpha^2 - 152\alpha^3 + 48\alpha^4;
\end{aligned}$$

and

$$\begin{aligned}
K := & 288\sqrt{3} \left(15963705 - 129546873\alpha + 510658314\alpha^2 - 1308834456\alpha^3 + 2415583204\alpha^4 \right. \\
& - 3321041560\alpha^5 + 3420107120\alpha^6 - 2619528992\alpha^7 + 1464766656\alpha^8 \\
& \left. - 575732096\alpha^9 + 147709696\alpha^{10} - 21284352\alpha^{11} + 1179648\alpha^{12} \right).
\end{aligned}$$

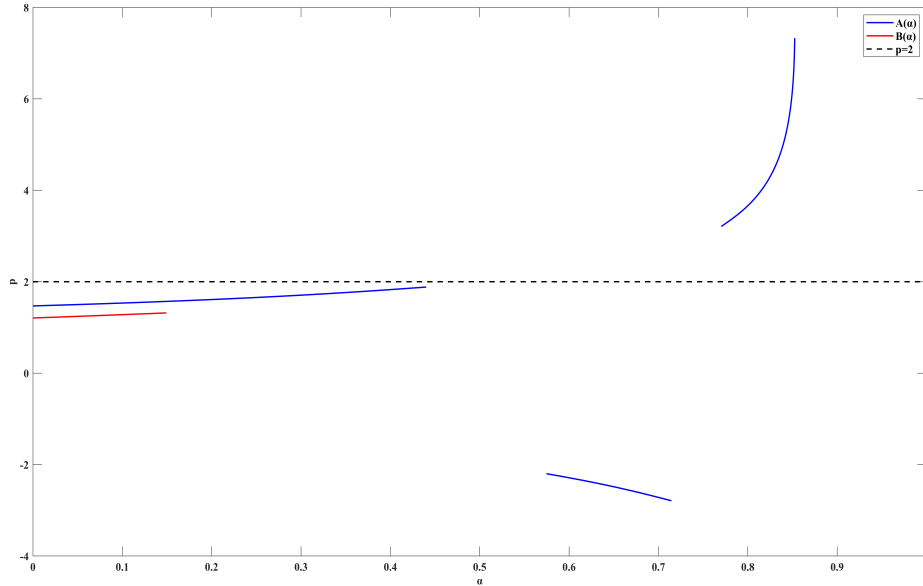


FIGURE 1. Graphical representation of p versus α . Here, $B(\alpha)$ (solid Red) and $A(\alpha)$ (solid blue) do not intersect for any choice of α . Dashed black line represents $p = 2$.

On $x = 1$, $Z(p, x, y)$ reduces into

$$s_3(p, y) := \frac{(1 - \alpha)^2}{576} \left(288\alpha + 16p^2(11 - 33\alpha + 9\alpha^2) - 8p^4(5 - 13\alpha + 5\alpha^2 + 3\alpha^3) - p^6(1 - 4\alpha + 6\alpha^2 - 16\alpha^3 + 4\alpha^4) \right), \quad p \in (0, 2). \quad (3.10)$$

While computing $\partial s_3 / \partial p = 0$, $p =: p_0 \approx 2L(\alpha)$ for $\alpha \in [0, \alpha_0) \cup (\alpha_0, \alpha_1)$, comes out to be the critical point, where $\alpha_0 \in [0, 1)$ is the smallest positive root of $1 - 4\alpha + 6\alpha^2 - 16\alpha^3 + 4\alpha^4 = 0$ and $\alpha_1 (\approx 0.370803927) \in [0, 1)$ (see Fig. 2) is the largest value so that $p \in (0, 2)$ otherwise no such real $p \in (0, 2)$ exists beyond this α_1 . Here

$$\left. \begin{aligned} L(\alpha) &:= \sqrt{\frac{-10 + 26\alpha - 10\alpha^2 - 6\alpha^3 + M}{3N}}; \\ M &:= \sqrt{133 - 751\alpha + 1497\alpha^2 - 1630\alpha^3 + 1666\alpha^4 - 708\alpha^5 + 144\alpha^6}; \\ N &:= 1 - 4\alpha + 6\alpha^2 - 16\alpha^3 + 4\alpha^4. \end{aligned} \right\} \quad (3.11)$$

Undergoing simple calculations, s_3 achieves its maximum value, approximately equals $P(\alpha)$, $\alpha \in [0, \alpha_0) \cup (\alpha_0, \alpha_1)$ at p_0 . Here

$$P(\alpha) := \frac{(1 - \alpha)^2}{486} \left(243\alpha - \frac{18(11 - 33\alpha + 9\alpha^2)(10 - 26\alpha + 10\alpha^2 + 6\alpha^3 - M)}{N} - \frac{12(5 - 13\alpha + 5\alpha^2 + 3\alpha^3)(-10 + 26\alpha - 10\alpha^2 - 6\alpha^3 + M)^2}{N^2} - \frac{2(-10 + 26\alpha - 10\alpha^2 - 6\alpha^3 + M)^3}{N^2} \right). \quad (3.12)$$

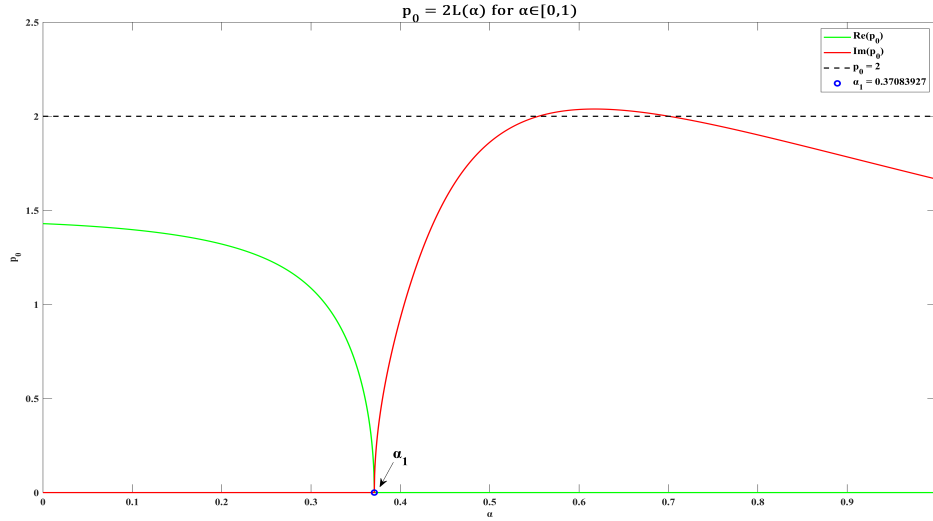


FIGURE 2. Graphical representation of p_0 versus α . Here, $\text{Re}(p_0)$ (solid green) and $\text{Im}(p_0)$ (solid red) represent the value of p_0 at different α , where α_1 (blue circle) is the point at which p_0 transforms from completely real to imaginary. Dashed black line represents $p_0 = 2$.

On $y = 0$, $Z(p, x, y)$ can be seen as

$$\begin{aligned} s_4(p, x) := \frac{(1-\alpha)^2}{1152} & \left\{ 576 \left(x - x^3(1-\alpha) \right) + 16p^2 \left(9 - 18x + x^4 + x^3(28 - 33\alpha) - 18\alpha \right. \right. \\ & \left. \left. + x^2(2 - 15\alpha + 18\alpha^2) \right) - 4p^4 \left(9 + 2x^4 + x^3(20 - 21\alpha) - 18\alpha \right. \right. \\ & \left. \left. + x^2(1 - 32\alpha + 52\alpha^2) + x(-12 + 19\alpha - 32\alpha^2 + 12\alpha^3) \right) \right. \\ & \left. + p^6 \left(x^4 + \alpha(1 - 2\alpha)^2(3 - 2\alpha) + x^3(1 + 3\alpha) - x^2(1 + 17\alpha \right. \right. \\ & \left. \left. - 34\alpha^2) + x(-3 + 19\alpha - 32\alpha^2 + 12\alpha^3) \right) \right\}. \end{aligned}$$

Furthermore, through some calculations, such as

$$\begin{aligned} \frac{\partial s_4}{\partial x} = \frac{(1-\alpha)^2}{1152} & \left\{ 576 \left(1 - 3x^2(1-\alpha) \right) - 16p^2 \left(18 - 4x^3 - 3x^2(28 - 33\alpha) - 2x(2 - 15\alpha \right. \right. \\ & \left. \left. + 18\alpha^2) \right) + p^6 \left(4x^3 + 3x^2(1 + 3\alpha) - 2x(1 + 17\alpha - 34\alpha^2) - 3 + 19\alpha \right. \right. \\ & \left. \left. - 32\alpha^2 + 12\alpha^3 \right) - 4p^4 \left(8x^3 + 3x^2(20 - 21\alpha) + 2x(1 - 32\alpha + 52\alpha^2) \right. \right. \\ & \left. \left. - 12 + 19\alpha - 32\alpha^2 + 12\alpha^3 \right) \right\} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial s_4}{\partial p} = \frac{(1-\alpha)^2}{1152} & \left\{ 32p \left(9 - 18x + x^4 + x^3(28 - 33\alpha) - 18\alpha + x^2(2 - 15\alpha + 18\alpha^2) \right) \right. \\ & \left. - 16p^3 \left(9 + 2x^4 + x^3(20 - 21\alpha) - 18\alpha + x^2(1 - 32\alpha + 52\alpha^2) \right. \right. \\ & \left. \left. + x(-12 + 19\alpha - 32\alpha^2 + 12\alpha^3) \right) + 6p^5 \left(x^4 + \alpha(1 - 2\alpha)^2(3 - 2\alpha) \right. \right. \\ & \left. \left. + x^3(1 + 3\alpha) - x^2(1 + 17\alpha - 34\alpha^2) + x(-3 + 19\alpha - 32\alpha^2 + 12\alpha^3) \right) \right\}, \end{aligned}$$

indicates that there does not exist any common solution for the system of equations $\partial s_4 / \partial x = 0$ and $\partial s_4 / \partial p = 0$, thus, s_4 has no critical points in $(0, 2) \times (0, 1)$.

On $y = 1$, $Z(p, x, y)$ reduces to

$$\begin{aligned} s_5(p, x) := \frac{(1-\alpha)^2}{1152} & \left\{ 64px(1 - x^2)(5 + x - 12\alpha) + 64(8 - 7x^2 - x^4 + 9\alpha) + 16p^3(1 - x^2)(3 - x \right. \\ & \left. - 2x^2 - 10\alpha + 6x\alpha + 8\alpha^2) - 2p^6(1 - 4\alpha + 6\alpha^2 - 16\alpha^3 + 4\alpha^4) \right. \\ & \left. + 16p^2 \left(6 + 14x^2 + 2x^4 + x(9 - 18\alpha) - 66\alpha + 18\alpha^2 - 9x^3(1 - 2\alpha) \right) \right. \\ & \left. + 4p^5(1 - x^2) \left(x^2 - 3 + 10\alpha - 8\alpha^2 - x(4 - 6\alpha) \right) - 4p^4 \left(7x^2 + x^4 + x(9 - 18\alpha) \right. \right. \\ & \left. \left. - 9x^3(1 - 2\alpha) + 4(3 - 13\alpha + 5\alpha^2 + 3\alpha^3) \right) \right\}. \end{aligned}$$

We note that the equations $\partial s_5/\partial x = 0$ and $\partial s_5/\partial p = 0$ possess no common solution in $(0, 2) \times (0, 1)$.

Case III:

Now, we determine the maximum values that $Z(p, x, y)$ may obtain on the cuboid Y 's edges. From equation (3.6), we have

$$Z(p, 0, 0) = r_1(p) := \frac{p^2(1-\alpha)^2(1-2\alpha)(144-36p^2+p^4\alpha(3-8\alpha+4\alpha^2))}{1152}.$$

Here, we consider the following three subcases for different choices of α .

- (1) For $\alpha = 0$, $r_1(p)$ reduces to $p^2(4-p^2)/32$ and $r'_1(p) = 0$ for $p = 0$, the point of minima and $p = \sqrt{2}$, the point of maxima. Therefore

$$Z(p, 0, 0) \leq \frac{1}{8}, \quad p \in [0, 2].$$

- (2) For $\alpha = 1/2$, $r_1(p) = 0$.

- (3) For $\alpha = (0, 1/2) \cup (1/2, 1)$, $r'_1(p) = p(1-\alpha)^2(1-2\alpha)(48-24p^2+p^4\alpha(3-8\alpha+4\alpha^2)) = 0$ for $p = 0$ and $p = 2((3-R(\alpha))/(3\alpha-8\alpha^2+4\alpha^3))^{1/2}$ as the points of minima and maxima respectively. So,

$$Z(p, 0, 0) \leq \frac{(1-\alpha)^2(3-R(\alpha))(-3+6\alpha-16\alpha^2+8\alpha^3+R(\alpha))}{6(3-2\alpha)^2\alpha^2(1-2\alpha)},$$

$$\text{with } R(\alpha) := \sqrt{3(3-3\alpha+8\alpha^2-4\alpha^3)}.$$

Now, equation (3.6) at $y = 1$, implies that $Z(p, 0, 1) = r_2(p) := (1-\alpha)^2(32(4-p^2)^2 + \alpha(1-2\alpha)^2(3-2\alpha)p^6 + 4p^3(4-p^2)(3-10\alpha+8\alpha^2))/1152$. Note that $r'_2(p)$ is a decreasing function in $[0, 2]$ and hence $p = 0$ becomes the point of maxima. Thus

$$Z(p, 0, 1) \leq \frac{4(1-\alpha)^2}{9}, \quad p \in [0, 2].$$

Through calculations, equation (3.6) shows that $Z(0, 0, y)$ attains its maximum value at $y = 1$, which implies that

$$Z(0, 0, y) \leq \frac{4(1-\alpha)^2}{9}, \quad y \in [0, 1].$$

Since, the equation (3.10) is free from y , we have

$$Z(p, 1, 1) = Z(p, 1, 0) = r_3(p) := \frac{(1-\alpha)^2}{576} \left(288\alpha + 16p^2(11-33\alpha+9\alpha^2) - 8p^4(5-13\alpha+5\alpha^2 + 3\alpha^3) - p^6(1-4\alpha+6\alpha^2-16\alpha^3+4\alpha^4) \right).$$

Now, $r'_3(p) = 32p(11-33\alpha+9\alpha^2) - 32p^3(5-13\alpha+5\alpha^2+3\alpha^3) - 6p^5(1-4\alpha+6\alpha^2-16\alpha^3+4\alpha^4) = 0$ when $p = \delta_1 := 0$ and $p = \delta_2 := 2L(\alpha)$ for $\alpha \in [0, \alpha_0) \cup (\alpha_0, \alpha_1)$, as the points of minima and maxima respectively, in the interval $[0, 2]$. The justification of $P(\alpha)$, α_0 and α_1 are provided above through equation (3.11) and (3.12). Thus, from equation (3.10),

$$Z(p, 1, 1) = Z(p, 1, 0) \leq P(\alpha), \quad p \in [0, 2] \quad \text{and} \quad \alpha \in [0, \alpha_0) \cup (\alpha_0, \alpha_1).$$

Consider equation (3.10) at $p = 0$, we get

$$Z(0, 1, y) = \frac{\alpha(1-\alpha)^2}{2}.$$

Equation (3.5) indicates that

$$Z(2, 1, y) = Z(2, 0, y) = Z(2, x, 0) = Z(2, x, 1) = \frac{\alpha(1-2\alpha)^2(1-\alpha)^2(3-2\alpha)}{18}.$$

Using equation (3.4), $Z(0, x, 1) = r_4(x) := (1-\alpha)^2(8-7x^2-x^4+9x^3\alpha)/18$. Upon calculations, we see that r_4 is a decreasing function of x in $[0, 1]$ and therefore $x = 0$ is the point of maxima. Hence

$$Z(0, x, 1) \leq \frac{4(1-\alpha)^2}{9}, \quad x \in [0, 1].$$

On again using equation (3.4), $Z(0, x, 0) = r_5(x) := x(1-(1-\alpha)x^2)(1-\alpha)^2/2$. Moreover, $r_5'(x) = 0$ when $x = \delta_3 := 1/\sqrt{3(1-\alpha)}$. Observe that $r_5(x)$ increases in $[0, \delta_3)$ and decreases in $(\delta_3, 1]$. Hence,

$$Z(0, x, 0) \leq \frac{(1-\alpha)^2}{3\sqrt{3(1-\alpha)}}, \quad x \in [0, 1].$$

Now, we provide a graphical representation of five upper-bounds (u.b) of $H_{3,1}(f)$ as follows:

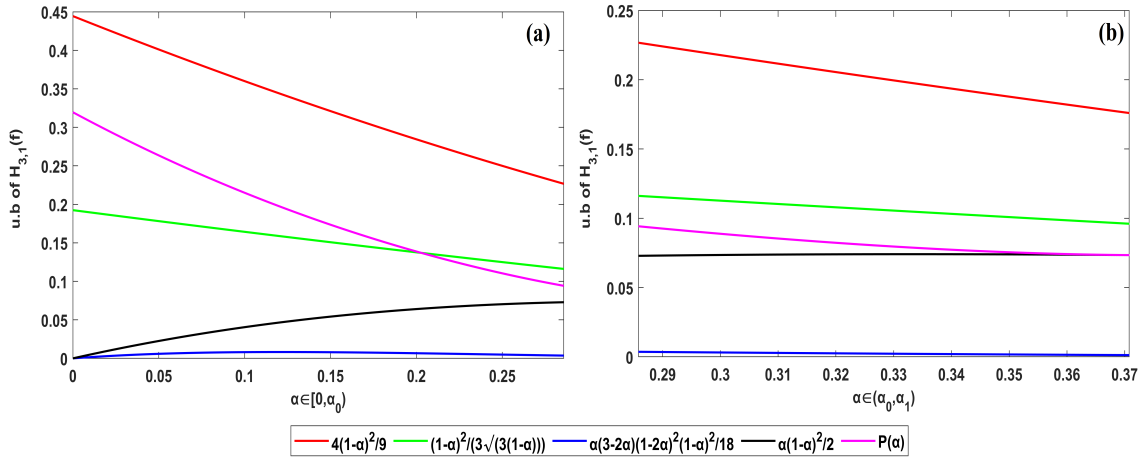


FIGURE 3. Graph of five upper-bounds (u.b) versus α . U.b's of $H_{3,1}(f)$ are $4(1-\alpha)^2/9$ (solid red), $(1-\alpha)^2/(3\sqrt{3(1-\alpha)})$ (solid green), $\alpha(1-2\alpha)^2(1-\alpha)^2(3-2\alpha)/18$ (solid blue), $\alpha(1-2\alpha)^2/2$ (solid black) and $P(\alpha)$ (solid pink) for $\alpha \in [0, \alpha_1]$ in (a) and $\alpha \in (\alpha_0, \alpha_1)$ in (b).

Given all the cases (see Fig. 3), the sharp inequality $|H_{3,1}(f)| \leq 4(1-\alpha)^2/9$, holds for every $\alpha \in [0, \alpha_0] \cup (\alpha_0, \alpha_1)$.

Let the function $f_0 \in \mathcal{S}^*(\alpha) : \mathbb{D} \rightarrow \mathbb{C}$, be defined as

$$f_0(z) = z \exp \left(\int_0^z \frac{\frac{1+(1-2\alpha)s^3}{1-s^3} - 1}{s} ds \right) = z + \frac{2(1-\alpha)}{3} z^4 + \frac{(1-\alpha)(5-2\alpha)}{9} z^7 + \dots,$$

with $f_0(0) = 0$ and $f_0'(0) = 1$, plays the role of an extremal function for the inequality presented in equation (3.1) with $a_2 = a_3 = a_5 = 0$ and $a_4 = 2(1-\alpha)/3$. ■

Now, we provide a corollary which incorporates the bound of $|H_{3,1}(f)|$ for the class \mathcal{S}^* , a subclass of $\mathcal{S}^*(\alpha)$, given as follows:

Corollary 3.2. *On substituting $\alpha = 0$ in Theorem 3.1, $\mathcal{S}^*(0) = \mathcal{S}^*$ and from equation (3.1), we get $|H_{3,1}(f)| \leq 4/9$. This bound is sharp and coincides with that of Kowalczyk et al. [7].*

For some already known sharp bounds of $H_{3,1}(f)$, regarding various choices of $\varphi(z)$, See Table 1. We note that the same bound is not available for $\varphi(z) := 1 + z - z^3/3$. Hence, as an application of Theorem 3.1, we provide a better bound of $|H_{3,1}(f)|$ for functions belonging to the class, $\mathcal{S}_{Ne}^* := \mathcal{S}^*(1 + z - z^3/3)$.

Corollary 3.3. *If $f \in \mathcal{S}_{Ne}^*$. Then $|H_{3,1}(f)| \leq 32/81 \approx 0.395062$.*

Proof. From [20], we have

$$\min_{|z|=r} \operatorname{Re}(\varphi(z)) = \begin{cases} 1 - r + \frac{1}{3}r^3, & r \leq 1/\sqrt{3} \\ 1 - \frac{1}{3}(1 + r^2)^{3/2}, & r \geq 1/\sqrt{3}. \end{cases}$$

We note that $\alpha = \min_{|z|=r} \operatorname{Re}(\varphi(z)) = 1 - \frac{2\sqrt{2}}{3}$ as r tends to 1. Now, substitution of $\alpha = 1 - 2\sqrt{2}/3 \approx 0.057191 \in [0, \alpha_0) \cup (\alpha_0, \alpha_1)$ in equation (3.1) implies that $|H_{3,1}(f)| \leq 32/81 \approx 0.395062$. ■

Remark 3.4. We have attempted to provide the sharp bound of $H_{3,1}(f)$ of functions, $f \in \mathcal{S}^*(\alpha)$ for $\alpha \in [0, \alpha_0) \cup (\alpha_0, \alpha_1)$ in Theorem 3.1. As a future scope, this result is still open for the remaining choices of α i.e., $\alpha \in (\alpha_1, 1)$.

Acknowledgment. Neha is thankful for the Research Grant from Delhi Technological University, New Delhi-110042.

REFERENCES

- [1] S. Banga and S. S. Kumar, The sharp bounds of the second and third Hankel determinants for the class \mathcal{SL}^* , Math. Slovaca **70** (2020), no. 4, 849–862
- [2] N. E. Cho et al., Some coefficient inequalities related to the Hankel determinant for strongly starlike functions of order alpha, J. Math. Inequal. **11** (2017), no. 2, 429–439.
- [3] N. E. Cho, V. Kumar and V. Ravichandran, A survey on coefficient estimates for Carathéodory functions, Appl. Math. E-Notes **19** (2019), 370–396
- [4] S. Fadaei, A. Ebadian and E. A. Adegani, The second Hankel determinant problem for a certain subclass of bi-univalent functions, Afr. Mat. **33** (2022), no. 2, Paper No. 48
- [5] A. W. Goodman, Univalent functions. Vol. I, Mariner Publishing Co., Inc., Tampa, FL, 1983
- [6] A. Janteng, S. A. Halim and M. Darus, Hankel determinant for starlike and convex functions, Int. J. Math. Anal. (Ruse) **1** (2007), no. 13-16, 619–625.
- [7] B. Kowalczyk, A. Lecko, and D. K. Thomas, The sharp bound of the third Hankel determinant for starlike functions, Forum Mathematicum. De Gruyter, (2022)
- [8] D. V. Krishna and T. Ramreddy, Hankel determinant for starlike and convex functions of order alpha, Tbil. Math. J. **5** (2012), 65–76.
- [9] D. V. Krishna and T. RamReddy, Second Hankel determinant for the class of Bazilevic functions, Stud. Univ. Babeş-Bolyai Math. **60** (2015), no. 3, 413–420
- [10] S. S. Kumar and G. Kamaljeet, A cardioid domain and starlike functions, Anal. Math. Phys. **11** (2021), no. 2, Paper No. 54, 34 pp
- [11] O. S. Kwon, A. Lecko and Y. J. Sim, On the fourth coefficient of functions in the Carathéodory class, Comput. Methods Funct. Theory **18** (2018), no. 2, 307–314

- [12] R. J. Libera and E. J. Zlotkiewicz, Early coefficients of the inverse of a regular convex function, *Proc. Amer. Math. Soc.* **85** (1982), no. 2, 225–230
- [13] W. C. Ma and D. Minda, A unified treatment of some special classes of univalent functions, in *Proceedings of the Conference on Complex Analysis (Tianjin, 1992)*, 157–169, Conf. Proc. Lecture Notes Anal., I, Int. Press, Cambridge.
- [14] C. Pommerenke, On the coefficients and Hankel determinants of univalent functions, *J. London Math. Soc.*, **41** (1966), 111–122.
- [15] B. Rath, K. S. Kumar, D. V. Krishna and A. Lecko, The sharp bound of the third Hankel determinant for starlike functions of order $1/2$, *Complex Anal. Oper. Theory* **16** (2022), no. 5, Paper No. 65, 8 pp.
- [16] M. I. S. Robertson, On the theory of univalent functions, *Ann. of Math. (2)* **37** (1936), no. 2, 374–408.
- [17] S. Sivaprasad Kumar and N. Verma, Certain Coefficient Problems of \mathcal{S}_e^* and \mathcal{C}_e , arXiv e-prints, pp.arXiv:2208.14644
- [18] S. Sivaprasad Kumar and N. Verma, Coefficient problems for starlike functions associated with a petal shaped domain, arXiv e-prints, pp.arXiv:2210.01435
- [19] N. Verma and S. Sivaprasad Kumar, A Conjecture on $H_3(1)$ for certain Starlike Functions, arXiv e-prints, pp.arXiv:2208.02975(2-22).
- [20] L. A. Wani and A. Swaminathan, Starlike and convex functions associated with a nephroid domain, *Bull. Malays. Math. Sci. Soc.* **44** (2021), no. 1, 79–104.
- [21] Q. H. Xu, F. Fang and T. S. Liu, On the Fekete and Szegő problem for starlike mappings of order α , *Acta Math. Sin. (Engl. Ser.)* **33** (2017), no. 4, 554–564.
- [22] P. Zaprawa, Third Hankel determinants for subclasses of univalent functions, *Mediterr. J. Math.* **14** (2017), no. 1, Paper No. 19, 10 pp.

DEPARTMENT OF APPLIED MATHEMATICS, DELHI TECHNOLOGICAL UNIVERSITY, DELHI-110042, INDIA
Email address: nehaverma1480@gmail.com

DEPARTMENT OF APPLIED MATHEMATICS, DELHI TECHNOLOGICAL UNIVERSITY, DELHI-110042, INDIA
Email address: spkumar@dce.ac.in

SMOTE-LASSO-DeepNet Framework for Cancer Subtyping from Gene Expression Data

Yashpal Singh and Seba Susan
Department of Information Technology
Delhi Technological University, Delhi, India
seba_406@yahoo.in

Abstract—Cancer subtyping from gene expression data is trending research in the field of bioinformatics. Classification of gene expression data is a challenging task due to the small number of samples and large number of features involved. The problem is further complicated due to the strong class imbalance issue prevalent in gene expression datasets. The challenge here is to find an end-to-end machine learning solution to classify cancer subtypes from small sample, high-dimensional, imbalanced gene expression datasets. In this study, we propose a SMOTE-LASSO-DeepNet framework for the identification of cancer subtypes from gene expression data. The proposed framework balances the training set using SMOTE, and then finds the most informative genes using LASSO. The balanced and pruned training set is then applied as input to a deep neural network (DeepNet) with four hidden layers having 512, 256, 128 and 64 neurons respectively. We tested our framework on four different cancer gene expression datasets: Leukemia, Lung cancer, Brain cancer and Breast cancer. It is observed from the results that our proposed SMOTE-LASSO-DeepNet framework performs consistently best as compared to the existing methods.

Keywords—SMOTE; LASSO; Deep learning; Gene Expression

I. INTRODUCTION

It is well-known that cancer is one of the leading causes of death globally. It is caused by the abnormal rapid production of cells, producing tumors with different behaviors [1]. On an average, worldwide, one out of six deaths are because of cancer [2]. These facts give rise to the need for an early and accurate diagnosis, which also reduces the side effects of treatment. Gene expression profiling provides valuable and early information about differentially expressed genes associated with different cancer types [3]. The raw microarray gene expression data is in the form of two-dimensional data where the columns represent the genes or features and the rows represent the samples. One problem associated with cancer gene expression datasets is the class imbalance issue in which the population of one class (majority class) far exceeds the population of the other classes (minority classes) [4]. In the current work, we use the most convenient and popular balancing technique suitable for multi-class datasets: - Synthetic Minority Over-Sampling Technique (SMOTE) [5], to oversample the minority classes and balance the class distribution. Mining the gene expression data is a challenging task because of the thousands of genes involved with very few samples available. And not all the genes in the dataset make an impact on the final classification results; only a few among

the thousands are significant for the model training [30]. In our work, we use Least Absolute Shrinkage and Selection Operator (LASSO) [6] for selecting the most informative genes from the gene expression data prior to the classification phase. LASSO has proved to outperform the state of the art in feature ranking and selection in the past [7].

In our proposed end-to-end classification framework for identification of cancer subtypes, we use both SMOTE and LASSO along with a deep neural network (DeepNet) to address, simultaneously, the class imbalance issue and high-dimensionality problem associated with cancer gene expression datasets. The deep neural network (DeepNet) is one of the best classifiers available today and is known to achieve high accuracies [8]. For the small number of samples found in cancer gene expression datasets, it may suffer from the overfitting problem, which we overcome by using SMOTE for minority class data augmentation. DeepNets achieve good results because of the many hidden layers that facilitate feature transformation and feature extraction, that trains the model much better as compared to other machine learning algorithms [29]. In this study, we use gene expression data belonging to four different types of cancers (Lung cancer, Breast cancer, Brain cancer, and Leukemia). The remaining part of the paper is organized as follows. In section II we review some works from literature on the classification of gene expression data. In section III we present the SMOTE-LASSO-DeepNet framework and the process flow. In section IV we discuss the experimental setup, and analyze the accuracies and F1 scores obtained. Section V concludes the paper.

II. RELATED WORK

Till today many researchers have worked on microarray data with different aims and techniques. Different soft computing techniques have been put in use in the field of bioinformatics [32]. The previous work of the authors tested the combination of fuzzy min-max classifier with LASSO based gene selection (FMM-LASSO) for identifying the lung cancer subtypes [9]. Chen *et al.* [10] presented a deep learning method called D-GEX that used an omnibus dataset having 111K gene expression profiles. D-GEX revealed complex patterns of gene expression. It was proved that deep learning-based D-GEX outperformed Linear Regression for gene expression inference on GEO microarray data. Tabares-Soto *et al.* [11] compared different machine learning algorithms on the 11_tumors dataset which has eleven types of tumors, and achieved high accuracies using the Convolution Neural Network (CNN). Lyu *et al.* [12], Gullien *et al.* [13] and Mohammed *et al.* [14] also implemented classification models

based on deep learning architectures for cancer gene expression data classification. Mostavi *et al.* [15] tested different Convolutional Neural Network (CNN) models on 33 cancer types; they implemented 1D-CNN, 2D-CNN, and 2D-Hybrid-CNN. The authors achieved the highest accuracies (> 95%) for ID-CNN and 2D-Hybrid-CNN.

One of the works most related to our approach is that of Urda *et al.* [16] who proposed a deep neural network with 2 to 4 hidden layers, each having neurons in the range 10 to 200, in combination with LASSO feature selection, for effective gene expression classification. This method however, restricts the maximum number of hidden units to 200; it does not recommend any optimal customized architecture suitable for classification of cancer gene expression data. Also, no solution for class imbalance is provided. There are some works available in literature that have implemented SMOTE for balancing the gene expression datasets [17]. A few have also combined SMOTE or variants of SMOTE with feature selection techniques such as LASSO and Information Gain [18, 19]. Several researchers have tested feature selection and classifier combinations for gene expression data classification, such as Principal Component Analysis (PCA) with Support Vector Machine (SVM) [35], and T-Test with Fuzzy Neural Network and SVM [31]. However, to the best of our knowledge, there is no prior work that has combined SMOTE and LASSO with DeepNets for Cancer subtyping from gene expression data, which is the idea presented in this paper. In our proposed framework, we seek to simultaneously address the different issues affecting gene expression profile datasets for accurate cancer diagnosis.

III. MATERIALS AND METHODS

A. Overall process flow

Fig. 1 presents the overall process flow for all our experiments, including the methods used for comparison. The LASSO feature selector is used with all models to induce fair comparison. As noted from Fig. 1, we start by loading the cancer gene expression data. The first step is preprocessing the dataset in which the values of the features are scaled in the range of 0 to 1 using the *MinMaxScaler* function. Then we divide the dataset into two equal sets: - validation (*V*) and cross-validation (*CV*) by selecting alternate samples for training and testing. For the *V* part, we use the original training and test sets for training and testing, respectively. For the *CV* part, the original testing set is the new training set, and the original training set is used for testing. Before training the model, we extract the important genes from the training set using the LASSO feature selector. For performance evaluation, we use six different classifier models apart from the proposed DeepNet model. For all the models, the process flow is the same. However, in our proposed method we use SMOTE for minority class data augmentation prior to the feature selection stage. The last step is the prediction on the test set, and analyzing the performance of different models. We compare the validation (*V*) and cross-validation (*CV*) accuracies of all classifiers and also calculate the F1-Scores.

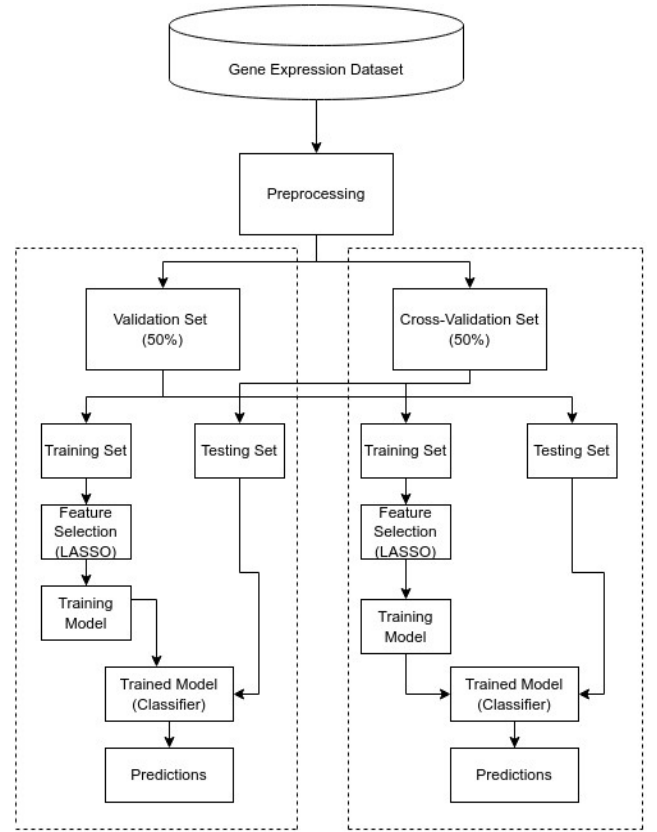


Fig 1. Overall process flow

B. Feature selection using LASSO

While analyzing the complexities of handling the gene expression data, the main problem found is that these datasets are high-dimensional, containing thousands of genes, and the majority of the features are irrelevant for cancer subtyping, and have minimal impact on the final classification results. These unwanted genes make the whole process slow and also lower the performance of the model. To overcome this problem, we used LASSO feature selection [6] whose main purpose is to select only those genes that are important for the learning process, and remove those that are unwanted. LASSO uses the following cost function to minimize the differences between the real and predicted values.

$$l_1 = \frac{1}{2X_{training}} \sum_{i=1}^{X_{training}} (Y_{real}^{(i)} - Y_{pred}^{(i)})^2 + \alpha \sum_{j=1}^n |\vartheta_j| \quad (1)$$

In (1), ϑ_j is the coefficient of the j^{th} feature and α is the hyper parameter that sets the penalty term; we set it to 0.001.

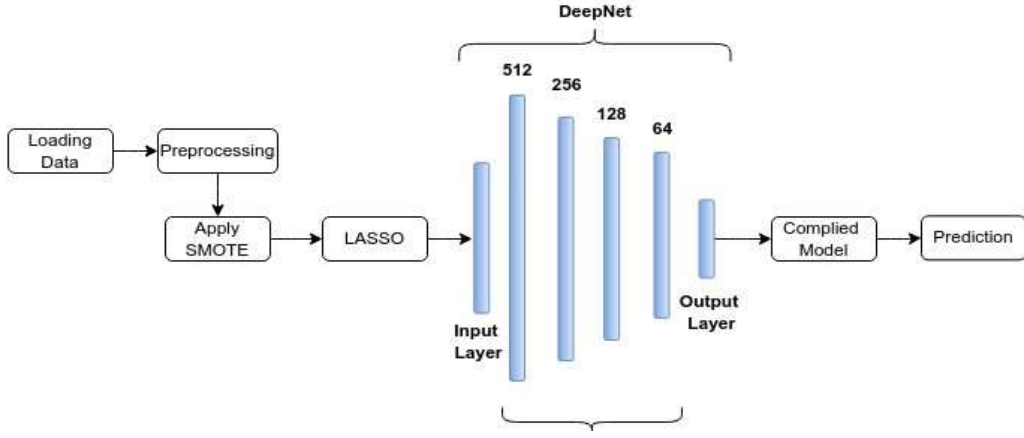


Fig 2. Architecture of proposed SMOTE-LASSO-DeepNet framework for cancer subtype classification

The aim is to optimize the cost function by reducing the absolute values of the coefficient. It selects those features which are useful and discards those which are unwanted by making their coefficient value zero.

C. Proposed SMOTE-LASSO-DeepNet framework

The proposed framework uses DeepNet to identify the cancer subtypes from gene expression datasets. A deep neural network has an architecture resembling that of a multi-layer perceptron [23], but with deeper layers. The proposed DeepNet architecture has four hidden layers with [512, 256, 128, 64] units respectively, as shown in Fig. 2. The proposed SMOTE-LASSO-DeepNet framework is an end-to-end solution for the classification of cancer gene expression data. This method has three phases, and each phase overcomes specific challenges associated with the gene expression data. In Phase 1, we apply SMOTE on the minority classes of the training set to balance the class distribution. SMOTE is the most preferable method to deal with imbalanced data in multi-class classification problems, due to its ease of application. It creates synthetic data points which are slightly different from the original data points by the process of interpolation.

In Phase 2, we apply LASSO feature selection on the balanced training data; it selects the important features and also speeds up the whole process of classification. We use the final selected features for the model training; the scores are computed separately for the V and CV cross-validation process as illustrated in the process flow in Fig. 1. In Phase 3, we use a DeepNet of architecture $X - 512 - 256 - 128 - 64 - Y$, where, X is the number of input features selected by LASSO, and Y is the number of target classes or cancer subtypes. For the hidden layers, we used 'ReLU' as an activation function, and for the output layer 'softmax' function is used because this model is used for multi-class classification. For all the datasets, we used a batch size of 30, with 100 epochs. We have compared the performance of the proposed deep learning framework with popular machine learning algorithms using the performance metrics of accuracy and F1-score. The performance of the proposed framework is compared both with and without SMOTE.

IV. EXPERIMENTAL SETUP AND RESULTS

In this section, we describe the dataset and experimental setup; and later on, we analyze the results.

A. Datasets

We used four benchmark gene expression datasets for cancer subtype identification: - Lung cancer, Brain cancer, Breast cancer and Leukemia (Blood cancer) [20, 21, 22]. In Table I, we show the description of these datasets.

TABLE I. DETAILS OF CANCER GENE EXPRESSION DATASETS

Cancer type	Samples	Genes	Classes
Leukemia	64	22,284	5
Lung Cancer	203	12,600	5
Brain Cancer	130	54,676	5
Breast Cancer	151	54,676	6

The datasets have good sample quality and they are manually curated for research purposes. All four datasets have large numbers of genes with limited samples, and an imbalanced class distribution. In the Brain and Breast cancer dataset, there are 54,676 genes which is the highest among all datasets.

B. Experimental setup

Data preprocessing is the first step to perform before any machine learning task because it helps to improve the overall performance. In this step, we normalize all the features in the dataset so that they lie in the range of 0 to 1. For model construction and training, Python *Keras* library is used which is run on TensorFlow's framework on a 2.00GHz Intel core™ i3 PC. *Imblearn* library is used for SMOTE implementation, which is present in the package of *over_sampling*. *Sklearn* is used for the LASSO implementation. In Table II, hyper parameter settings are shown for all the learning algorithms in our experiments. Other than the proposed SMOTE-LASSO-DeepNet framework, we also find the accuracies of the combinations LASSO-SMOTE-DeepNet (proposed method with LASSO feature selection performed before SMOTE), and LASSO-DeepNet (proposed method without SMOTE), to verify the effectiveness of the proposed combination. For proving the effectiveness of the proposed SMOTE-LASSO-DeepNet framework, we compare its

performance with that of six machine learning algorithms [9, 24-28] used in combination with LASSO feature selection; these methods along with their hyperparameters are listed in Table II.

TABLE II. HYPER-PARAMETER SETTINGS

Methods	Hyper parameters	values
SMOTE-LASSO-DeepNet (proposed)	Activation Function	ReLU (for i/p and hidden layer), softmax (o/p layer)
	Hidden Layers	4
	Number of units per layer	[512, 256, 128, 64]
	Optimizer	adam
	Loss Function	sparse_categorical_crossentropy
LASSO-SMOTE-DeepNet	Activation Function	ReLU (for i/p and hidden layer), softmax (o/p layer)
	Hidden Layer	4
	Number of Units per layer	[512, 256, 128, 64]
	Optimizer	adam
	Loss Function	sparse_categorical_crossentropy
LASSO-DeepNet	Activation Function	ReLU (for i/p and hidden layer), softmax (o/p layer)
	Hidden Layers	4
	Number of units per layer	[512, 256, 128, 64]
	Optimizer	adam
	Loss Function	sparse_categorical_crossentropy
FMM-LASSO [9]	Hyperbox Coefficient	0.7
	Sensitivity	1
Support Vector Machine [24] (with LASSO)	Regularization parameter c	0.0018
	gamma	0.126
Random Forest [25] (with LASSO)	N_estimators	100
	Max_depth	3
Logistic Regression [26] (with LASSO)	C	1
	Penalty	12
	Solver	lbfgs
K-Nearest Neighbor [27] (with LASSO)	No. of neighbors	5
Naïve Bayes [28] (with LASSO)	Var_Smoothing	$1e^{-9}$

C. Result Analysis

All the results were calculated in a setting where the dataset is divided into two sets, one for odd-numbered points and the other for even-numbered points, as illustrated in the process flow in Fig. 1. In the validation (*V*) stage, the training and test sets were used as it is, while in the cross-validation (*CV*) stage, the training and test sets were interchanged. For both parts, LASSO was used for feature selection. For the performance analysis, we calculated the accuracy and F1-score of all

methods. The classification results of all four datasets are given in Tables III to VI for the Leukemia, Lung cancer, Brain cancer and Breast cancer datasets, respectively.

TABLE III. RESULT FOR LEUKEMIA*

Methods	Accuracy (in%)		F1-Score	
	Validation	Cross-validation	Validation	Cross-Validation
SMOTE-LASSO-DeepNet	98.13	100	0.98	1.0
LASSO-SMOTE-DeepNet	96.25	100	0.95	1.0
LASSO-DeepNet	96.25	98.125	0.94	0.98
FMM-LASSO	90.62	93.75	0.92	0.94
SVM-LASSO	96.87	96.87	0.95	0.96
RF-LASSO	96.25	96.25	0.96	0.95
KNN-LASSO	96.87	100	0.95	1.0
NB-LASSO	81.25	84.37	0.75	0.82
LR-LASSO	96.87	100	0.95	1.0

* Highest values shown in bold

TABLE IV. RESULT FOR LUNG CANCER*

Methods	Accuracy (in %)		F1-Score	
	Validation	Cross-validation	Validation	Cross-Validation
SMOTE-LASSO-DeepNet	95.68	94.06	0.91	0.85
LASSO-SMOTE-DeepNet	95.49	91.28	0.91	0.78
LASSO-DeepNet	94.31	89.9	0.87	0.76
FMM-LASSO	90.19	93	0.86	0.85
SVM-LASSO	93.13	92.07	0.74	0.72
RF-LASSO	91.76	85.34	0.7	0.59
KNN-LASSO	94.11	89.1	0.83	0.68
NB-LASSO	94.11	84.15	0.75	0.6
LR-LASSO	95	89.1	0.85	0.74

* Highest values shown in bold

In Table III, as we observe in the case of the Leukemia dataset, SMOTE-LASSO-DeepNet performs best among all the classifiers. KNN-LASSO, LR-LASSO, and LASSO-SMOTE-DeepNet also achieve 100% accuracy in cross-validation, but in the validation case, the highest accuracy is 98.13% with an F1-score of 0.98 for the proposed method. In the Lung cancer results shown in Table IV, SMOTE-LASSO-DeepNet achieves the highest accuracies of 95.68% and 94.06% (for *V* and *CV* respectively). However, in case of the Brain cancer results shown in Table V, the SVM-LASSO and LR-LASSO perform slightly better in the *V* and *CV* cases, respectively. SMOTE-LASSO-DeepNet achieves a consistently good

performance for both V and CV cases, unlike some other classifiers in Table V for which a dip in performance was noted when cross-validating the results. In case of the Breast cancer results shown in Table VI, LR-LASSO achieves the highest accuracy of 86.84% (F1-score=0.86) for the validation case, but in cross-validation, SMOTE-LASSO-DeepNet achieves the highest accuracy of 91.46% (F1-score=0.92).

TABLE V. RESULT FOR BRAIN CANCER*

Methods	Accuracy (in %)		F1-Score	
	Validation	Cross-validation	Validation	Cross-Validation
SMOTE-LASSO-DeepNet	91.3	83.3	0.89	0.83
LASSO-SMOTE-DeepNet	90.7	84	0.89	0.83
LASSO-DeepNet	90.7	83.6	0.89	0.82
FMM-LASSO	89.23	84.6	0.86	0.82
SVM-LASSO	92.3	84.6	0.9	0.82
RF-LASSO	90.15	86.15	0.87	0.84
KNN-LASSO	89.23	78.46	0.86	0.79
NB-LASSO	87.69	80	0.85	0.77
LR-LASSO	90.76	87.69	0.88	0.88

* Highest values shown in bold

TABLE VI. RESULT FOR BREAST CANCER*

Methods	Accuracy (in %)		F1-Score	
	Validation	Cross-validation	Validation	Cross-Validation
SMOTE-LASSO-DeepNet	84.47	91.46	0.84	0.92
LASSO-SMOTE-DeepNet	81.31	91.46	0.83	0.91
LASSO-DeepNet	81.31	90.93	0.82	0.91
FMM-LASSO	72.36	81.33	0.76	0.81
SVM-LASSO	85.52	84	0.83	0.77
RF-LASSO	78.15	91.46	0.77	0.79
KNN-LASSO	80.26	81.33	0.81	0.77
NB-LASSO	84.21	89.33	0.84	0.77
LR-LASSO	86.84	88	0.86	0.87

* Highest values shown in bold

As an overall observation, we can say that classifiers other than the proposed method, especially SVM-LASSO and LR-LASSO have also performed well, but their performance was not consistent across all four datasets. Our proposed method SMOTE-LASSO-DeepNet performs consistently best in case of all four datasets for both V and CV cases. Apart from the proposed method, we also study two DeepNet variations of our framework (same architecture of DeepNet is maintained):

LASSO-SMOTE-DeepNet and LASSO-DeepNet; we find from Tables III to VI that their performances were not at par with that of the proposed SMOTE-LASSO-DeepNet framework.

V. CONCLUSION

In this paper we present a deep learning framework called SMOTE-LASSO-DeepNet for identifying cancer subtypes from microarray gene expression data. In our methodology, we first use SMOTE on the training set for minority class augmentation. In the next phase, we apply LASSO for feature selection to select the relevant genes for the classification. Finally, we train the data using a DeepNet model with four hidden layers containing 512, 256, 128 and 64 hidden units, respectively. After analyzing the results, we establish that our method gives consistently the best performance among all methods for the four benchmark cancer datasets. We will be extending this work in future by exploring deep ensemble frameworks for classifying gene expression data.

REFERENCES

- [1] Varadhachary, Gauri, and James L. Abbruzzese. "Carcinoma of unknown primary." In *Abeloff's Clinical oncology*, pp. 1694-1702. Elsevier, 2020.
- [2] Torre, Lindsey A., Britton Trabert, Carol E. DeSantis, Kimberly D. Miller, Goli Samimi, Carolyn D. Runowicz, Mia M. Gaudet, Ahmedin Jemal, and Rebecca L. Siegel. "Ovarian cancer statistics, 2018." *CA: a cancer journal for clinicians* 68, no. 4 (2018): 284-296.
- [3] Lu, Ying, and Jiawei Han. "Cancer classification using gene expression data." *Information Systems* 28, no. 4 (2003): 243-268.
- [4] Susan, Seba, and Amitesh Kumar. "The balancing trick: Optimized sampling of imbalanced datasets—A brief survey of the recent State of the Art." *Engineering Reports* 3, no. 4 (2021): e12298.
- [5] Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. "SMOTE: synthetic minority over-sampling technique." *Journal of artificial intelligence research* 16 (2002): 321-357.
- [6] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58, no. 1 (1996): 267-288.
- [7] Susan, Seba, and Madasu Hanmandlu. "Smaller feature subset selection for real-world datasets using a new mutual information with Gaussian gain." *Multidimensional Systems and Signal Processing* 30, no. 3 (2019): 1469-1488.
- [8] Jain, Anmol, Aishwary Kumar, and Seba Susan. "Evaluating Deep Neural Network Ensembles by Majority Voting Cum Meta-Learning Scheme." In *Soft Computing and Signal Processing*, pp. 29-37. Springer, Singapore, 2022.
- [9] Singh, Yashpal, and Seba Susan. "Optimal Gene Selection and Classification of Microarray Data Using Fuzzy Min-Max Neural Network with LASSO." In *International Conference on Intelligent and Fuzzy Systems*, pp. 777-784. Springer, Cham, 2022.
- [10] Chen, Yifei, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. "Gene expression inference with deep learning." *Bioinformatics* 32, no. 12 (2016): 1832-1839.
- [11] Tabares-Soto, Reinel, Simon Orozco-Arias, Victor Romero-Cano, Vanesa Segovia Bucheli, José Luis Rodríguez-Sotelo, and Cristian Felipe Jiménez-Varón. "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data." *PeerJ Computer Science* 6 (2020): e270.
- [12] Lyu, Boyu, and Anamul Haque. "Deep learning based tumor type classification using gene expression data." In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pp. 89-96. 2018.
- [13] Guillen, Pablo, and Jerry Ebalunode. "Cancer classification based on microarray gene expression data using deep learning." In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 1403-1405. IEEE, 2016.

- [14] Mohammed, Mohanad, Henry Mwambi, Innocent B. Mboya, Murtada K. Elbashir, and Bernard Omolo. "A stacking ensemble deep learning approach to cancer type classification based on TCGA data." *Scientific reports* 11, no. 1 (2021): 1-22.
- [15] Mostavi, Milad, Yu-Chiao Chiu, Yufei Huang, and Yidong Chen. "Convolutional neural network models for cancer type prediction based on gene expression." *BMC medical genomics* 13, no. 5 (2020): 1-13.
- [16] Urda, Daniel, Julio Montes-Torres, Fernando Moreno, Leonardo Franco, and José M. Jerez. "Deep learning to analyze RNA-seq gene expression data." In *International work-conference on artificial neural networks*, pp. 50-59. Springer, Cham, 2017.
- [17] Blagus, Rok, and Lara Lusa. "SMOTE for high-dimensional class-imbalanced data." *BMC bioinformatics* 14, no. 1 (2013): 1-16.
- [18] Hamzeh, Osama, Abedalrhman Alkhateeb, Julia Zheng, Srinath Kandalam, and Luis Rueda. "Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data." *BMC bioinformatics* 21, no. 2 (2020): 1-10.
- [19] Roy, Shikha, Rakesh Kumar, Vaibhav Mittal, and Dinesh Gupta. "Classification models for Invasive Ductal Carcinoma Progression, based on gene expression data-trained supervised machine learning." *Scientific reports* 10, no. 1 (2020): 1-15.
- [20] Grisci, Bruno Iochins, Bruno César Feltes, and Marcio Dorn. "Neuroevolution as a tool for microarray gene expression pattern identification in cancer research." *Journal of biomedical informatics* 89 (2019): 122-133.
- [21] Feltes, Bruno Cesar, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Marcio Dorn. "Cumida: an extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research." *Journal of Computational Biology* 26, no. 4 (2019): 376-386.
- [22] Bhattacharjee, Arindam, William G. Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd et al. "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses." *Proceedings of the National Academy of Sciences* 98, no. 24 (2001): 13790-13795.
- [23] Susan, Seba, and Jatin Malhotra. "Learning interpretable hidden state structures for handwritten numeral recognition." In *2020 4th International Conference on Computational Intelligence and Networks (CINE)*, pp. 1-6. IEEE, 2020.
- [24] Vapnik, Vladimir. *The nature of statistical learning theory*. Springer science & business media, 1999.
- [25] Breiman, Leo. "Random forests." *Machine learning* 45, no. 1 (2001): 5-32.
- [26] Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. "An introduction to logistic regression analysis and reporting." *The journal of educational research* 96, no. 1 (2002): 3- 14.
- [27] Guo, Gongde, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. "KNN model-based approach in classification." In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"*, pp. 986-996. Springer, Berlin, Heidelberg, 2003.
- [28] Wickramasinghe, Indika, and Harsha Kalutarage. "Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation." *Soft Computing* 25, no. 3 (2021): 2277-2293.
- [29] Susan, Seba, and Jatin Malhotra. "Recognising devanagari script by deep structure learning of image quadrants." *DESIDOC Journal of Library & Information Technology* 40, no. 5 (2020): 268-271.
- [30] Wang, Lipo, Yaoli Wang, and Qing Chang. "Feature selection methods for big data bioinformatics: a survey from the search perspective." *Methods* 111 (2016): 21-31.
- [31] Wang, Lipo, Feng Chu, and Wei Xie. "Accurate cancer classification using expressions of very few genes." *IEEE/ACM Transactions on computational biology and bioinformatics* 4, no. 1 (2007): 40-53.
- [32] S. Mitra and Y. Hayashi, "Bioinformatics with soft computing," *IEEE Trans. Systems, Man and Cybernetics, Part C*, vol.36, pp.616 -635, 2006.
- [33] Chu, Feng, and Lipo Wang. "Applications of support vector machines to cancer classification with microarray data." *International journal of neural systems* 15, no. 06 (2005): 475-484.

Article

Spatiotemporal Activity Mapping for Enhanced Multi-Object Detection with Reduced Resource Utilization

Shashank * and Indu Sreedevi

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

* Correspondence: shashank.ece.dtu@gmail.com or shashank_2k17phdec05@dtu.ac.in

Abstract: The accuracy of data captured by sensors highly impacts the performance of a computer vision system. To derive highly accurate data, the computer vision system must be capable of identifying critical objects and activities in the field of sensors and reconfiguring the configuration space of the sensors in real time. The majority of modern reconfiguration systems rely on complex computations and thus consume lots of resources. This may not be a problem for systems with a continuous power supply, but it can be a major set-back for computer vision systems employing sensors with limited resources. Further, to develop an appropriate understanding of the scene, the computer vision system must correlate past and present events of the scene captured in the sensor's field of view (FOV). To address the abovementioned problems, this article provides a simple yet efficient framework for a sensor's reconfiguration. The framework performs a spatiotemporal evaluation of the scene to generate adaptive activity maps, based on which the sensors are reconfigured. The activity maps contain normalized values assigned to each pixel in the sensor's FOV, called normalized pixel sensitivity, which represents the impact of activities or events on each pixel in the sensor's FOV. The temporal relationship between the past and present events is developed by utilizing standard half-width Gaussian distribution. The framework further proposes a federated optical-flow-based filter to determine critical activities in the FOV. Based on the activity maps, the sensors are re-configured to align the center of the sensors to the most sensitive area (i.e., region of importance) of the field. The proposed framework is tested on multiple surveillance and sports datasets and outperforms the contemporary reconfiguration systems in terms of multi-object tracking accuracy (MOTA).



Citation: Shashank; Sreedevi, I. Spatiotemporal Activity Mapping for Enhanced Multi-Object Detection with Reduced Resource Utilization.

Electronics **2023**, *12*, 37. <https://doi.org/10.3390/electronics12010037>

Academic Editor: Jungong Han

Received: 11 November 2022

Revised: 6 December 2022

Accepted: 15 December 2022

Published: 22 December 2022



Copyright: © 2022 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/>).

Keywords: computer vision; activity mapping; reconfiguration; multi-object tracking; spatiotemporal analysis; federated optical flow

1. Introduction

Computer vision applications have experienced tremendous growth over the past two decades. Most commonly, the computer vision applications cover the automotive, sports, entertainment, healthcare, robotics, security and surveillance areas [1]. Computer vision systems perform data processing to obtain an understanding of a scene. Systems employing robotic vision [2] take it to the next level by actuating actions based on the understanding developed by computer vision. Most of the modern computer vision applications rely on one or more sensors co-operatively sensing their environment to detect and track objects of interest and obtain information of activities or events in the surroundings.

The performance of a computer vision application depends on the accuracy of the data captured by the sensors used in terms of application-specific information. An accurate computer vision system is capable of distinguishing object(s) of interest prior to capturing the image of the scene, so the data bearing high information of the relevant object(s) of interest can be made available for processing. Further, if the object(s) of interest are captured in the center of the sensor's field of view (FOV), the information and understanding of the event obtained by processing the data are optimized. The performance of the computer

vision system relies on the processing of the data bearing information about the object(s) of interest that are captured by the sensors, whereas the sensors can be made to capture the data with high application-specific accuracy by way of calibration. Therefore, since the sensor calibration and data processing of computer vision systems are inter-related [3], the re-configuration of sensors in real time is a major challenge, especially for systems with limited resources, which cannot afford a high computational complexity. Therefore, there is a need for a framework for sensor reconfiguration that can efficiently address the challenge of limited resources.

Applications such as driverless automobiles [4], sports analytics [5] and the surveillance of large areas using UAVs [6,7] employ mobile sensors to capture the environment. The sensors of such applications often have limited resources; therefore, optimized resource utilization becomes essential. In some real-time computer vision applications, due to high computational complexity requirements, the resources are exhausted at a higher pace; thus, resources become critical.

Active computer vision systems (also known as active vision systems) are capable of calibrating (reconfiguring) the internal and external parameters of their sensors according to the needs of the system. Most of the active vision systems, such as the systems proposed in [8–10], rely on a pre-defined prioritization of the area under observation, based on assumptions regarding the sensor's field of view and surveillance area. As the activities are highly dynamic in occurrence, the pre-defined placement and orientation of the sensors result in an inappropriate sensor's pose, which hardly enables the object(s) of interest to be captured in the center of the sensor's FOV. Such a setting of the sensors in the network results in extraction of data with insufficient information of an activity or event.

It should be noted that the pattern of activities also affects the understanding of an event. A computer vision system must therefore take into consideration the events and activities of the past as well as the present activities when developing an understanding of a scene. For example, in a surveillance system, if an event occurs repetitively, the computer vision system must identify the area or site of the event and must prioritize that area or site for inspection. However, an active vision system requires highly complex computations to obtain such an understanding, utilizing a lot of resources.

Based on the abovementioned challenges, a computer vision system capable of reconfiguring the sensors based on spatiotemporal activity analysis, utilizing low resources, is desirable. This article provides a simple yet effective framework for the reconfiguration of sensors participating in an active vision system, thus yielding a better activity analysis and scene understanding with a very low computation complexity.

The framework performs a spatiotemporal evaluation of the scene to generate adaptive activity maps, based on which the sensors are reconfigured. To minimize the resource utilization, the proposed framework proposes a model-based approach (i.e., a non-learning-based approach) for the spatiotemporal evaluation of the scene by utilizing simpler concepts of image processing in combination. Based on the spatiotemporal activity analysis of the scene, the framework assigns a normalized sensitivity value to each pixel of the frame, which represents the impact of present and past activities on the corresponding pixel. Standard half-width Gaussian distribution is used to develop the temporal relationship between the past and present events. The framework further provides a federated optical-flow-based filter to identify and distinguish critical activities or events captured within the field of view of the sensors. Based on the normalized pixel sensitivity values of the adaptive activity maps, the framework reconfigures the sensors to align the center of the FOV of the sensors to the most sensitive portion of the scene, thus capturing the data with the best information in each frame.

The proposed framework can be utilized in a single-camera setting but is more suitable for a camera-pair configuration, with each node having two sensors (i.e., a primary camera and a secondary camera cooperatively operating together). In the camera-pair configuration, the primary camera includes basic image-processing capabilities and is used to generate activity maps for a scene. Based on the activity maps, the secondary

camera can be calibrated (or reconfigured) to obtain optimized data. The performance of the framework is determined in terms of multi-object tracking accuracy (MOTA). In the case of the single-camera setting, the framework is capable of determining object(s) of interest in the frame initially; however, the calibration of the camera may result in the loss of a new object of interest being captured by the sensor, as the field of view (FOV) of the sensor is adjusted in accordance with the knowledge of the object(s) of interest previously present in the frame. It must be noted that the objective of the proposed framework is not to obtain a high-performance activity tracking system; rather, the proposed framework intends to provide a balance between resource utilization and tracking accuracy and enables spatiotemporal activity analysis using simple image-processing methods.

2. Background

To capture accurate data for the optimized performance of the computer vision system, the system must identify the region of importance (ROI) in the FOV of the sensors and reconfigure the sensors to match the center of the ROI with the center of the FOV. There are several methods for determining the ROI, activity mapping [11] being one of the most efficient ones. Most systems proposed for ROI determination focus on the activities detected in the present frame for activity mapping and neglect the impact of past events (i.e., those events that are captured by a sensor in the past while capturing a frame for a timeframe) on the activity map. Such systems usually focus on reducing noise in the frame rather than establishing a spatiotemporal relationship between present and past events. The expression “past events” must be interpreted as the events captured by a sensor and detected by processing a frame that is in the past in relation to a current frame of the sensor. The expression “present events” must be interpreted as the events derived by the processing of the current frame captured by the sensor. For example, the past events can be events captured by the sensor through the analysis of the most recent frame captured by the sensor, whereas all the events derived by processing the frame prior to the current frame can be referred to as “past events”. Thus, the activity map generated by such systems is only a filtered set of frames accumulated together.

It must also be noted that the system must be capable of identifying critical activities or events and distinguishing them from the undesirable activities captured in the scene. The ROI must only be specific to the desired activities dedicated to the overall functionality of the system. For example, a traffic monitoring system should avoid the falling of leaves from trees in the scene and must treat such an activity as undesirable. On the other hand, a computer vision system designed to estimate the effects of climatic change on trees must consider the falling of leaves as a critical activity and must treat other activities in the scene as undesirable.

Pan et al. [12] and Mehboob et al. [13] proposed region of interest (ROI) estimation for vehicle flow detection using morphological operations to filter noise. Pan et al. [12] proposed a self-adaptive window-based traffic estimation using background subtraction, edge detection for object detection and morphological features for noise reduction. Mehboob et al. [13] proposed centroid detection using morphological close and erode operation for noise reduction and motion vectors for traffic flow estimation. The morphological operations used in [12,13] attempted to reduce the noise but failed to prevent the detection of undesirable activities in the scene (i.e., failed to determine critical activities and prevent the detection of undesirable activities from the scene). Both systems further developed an understanding of the ROI based only on the activities in the present frame and neglected the impact of past events on the ROI. To address this gap, the approach proposed in [14] considered spatiotemporal evaluation for activity mapping. The approach proposed a spatiotemporal relationship between the present and past activities in the frame; however, it did not provide a filter to remove noise or undesirable object detection from the scene.

Contemporary approaches for activity mapping either employ highly complex computation models or artificial intelligence approaches for ROI detection. Marvin and Moritz [15] presented a non-parametric model for spatiotemporal activity based on Gaussian process

regression (GPR). Sattar et al. [16] proposed a convolution neural network (CNN)-based spatiotemporal activity mapping method for group activity detection. Zhao and Gao [17] proposed an online feature learning model for spatiotemporal event forecasting. Liu and Jing [18] proposed an artificial intelligence (AI)-based activity mapping method for sports analytics using spatiotemporal activity patterns. Yan et al. [19] proposed an end-to-end position-aware spatiotemporal activity analysis using a long short-term memory (LSTM) approach.

The approaches in [12–14] provide simple models for activity mapping and ROI detection; however, they lack the accuracy of ROI prediction. The methods proposed in [15–19] provide efficient spatiotemporal activity mapping for ROI detection, but at the cost of high computational complexity, and are thus not suitable for systems with low or limited resources. Artificial intelligence (AI)-based systems [16–19] require high computation and storage capabilities to train the model and further require iteratively changing the training model in unforeseen conditions, illumination changes, etc. AI-based multi-object tracking systems are also susceptible to adversarial attacks, which makes the timely training of the system critically necessary. The AI-based systems, due to re-iterative training to deal with unforeseen conditions and adversarial attacks, require higher computational resources. There is a trade-off between the accuracy and computational complexity of the activity mapping approaches. The proposed framework strikes a balance between resource utilization and activity tracking accuracy, enabling spatiotemporal activity analysis using simple image-processing techniques for accurate ROI detection.

3. Spatiotemporal Activity Mapping Model

Considering the requirement of spatiotemporal activity mapping for accurate ROI detection and better tracking accuracy, this article provides an efficient framework with a low resource utilization for computer vision systems with limited resources. The framework strikes a balance between resource utilization and activity tracking accuracy, allowing for spatiotemporal activity analysis using simple image-processing techniques. The framework (as shown in Figure 1) generates adaptive spatiotemporal activity maps based on the sensed data obtained by the primary camera in consecutive temporal frames, which can be processed to derive information to calibrate the secondary camera of the computer vision system. Each adaptive spatiotemporal activity map is derived by the past frames and the present frame for each timeframe. The importance of the past frames and the present frame is derived through a normalized half-width Gaussian distribution function such that events in the present frame are assigned the highest importance, whereas the events in the past frames are assigned importance across the normalized half-width of the Gaussian distribution function according to their relative timeframes. The adaptive activity map is further used to determine the dynamic ROI of the scene. The framework further includes a re-configurator that alters the configuration space of the sensor such that the center of the ROI matches with the center of the sensor's FOV. The revised reconfiguration results in the improved accuracy of activity tracking and thus yields better scene understanding.

Notations Used:

- k : Number of past frames captured by the primary camera;
- $S(t)$: Data sensed by the primary camera in the present frame;
- $S(t - i)$: Data sensed by the primary camera in “i” frames prior to the present frame;
- S_{1i} : Data obtained by object detection on $S(t - i)$;
- S_{2i} : Data obtained by applying adaptive thresholding on S_{1i} ;
- S_{3i} : Data obtained by the binarization of S_{2i} ;
- S_{4i} : Data obtained by filtering S_{3i} using federated optical flow;
- $H(t - i)$: Temporal function for “i” frames prior to the present frame;
- S_{5i} : Temporal component of S_{4i} obtained as a product of S_{4i} and $H(t - i)$;
- X : Cumulative spatiotemporal activity map for the present frame;
- N : Normalized spatiotemporal activity map for the present frame;
- R : Reconfiguration parameters for the secondary camera;

C: Data from the calibrated secondary camera; and
Y: Activity analysis after processing C.

The following subsections of Section 3 intend to discuss the functionality of the proposed framework for spatiotemporal activity mapping. Through Section 3.1, we discuss some basic image-processing tools/methods used by the proposed framework for generating adaptive spatiotemporal activity maps, whereas the steps are discussed in Section 3.2.

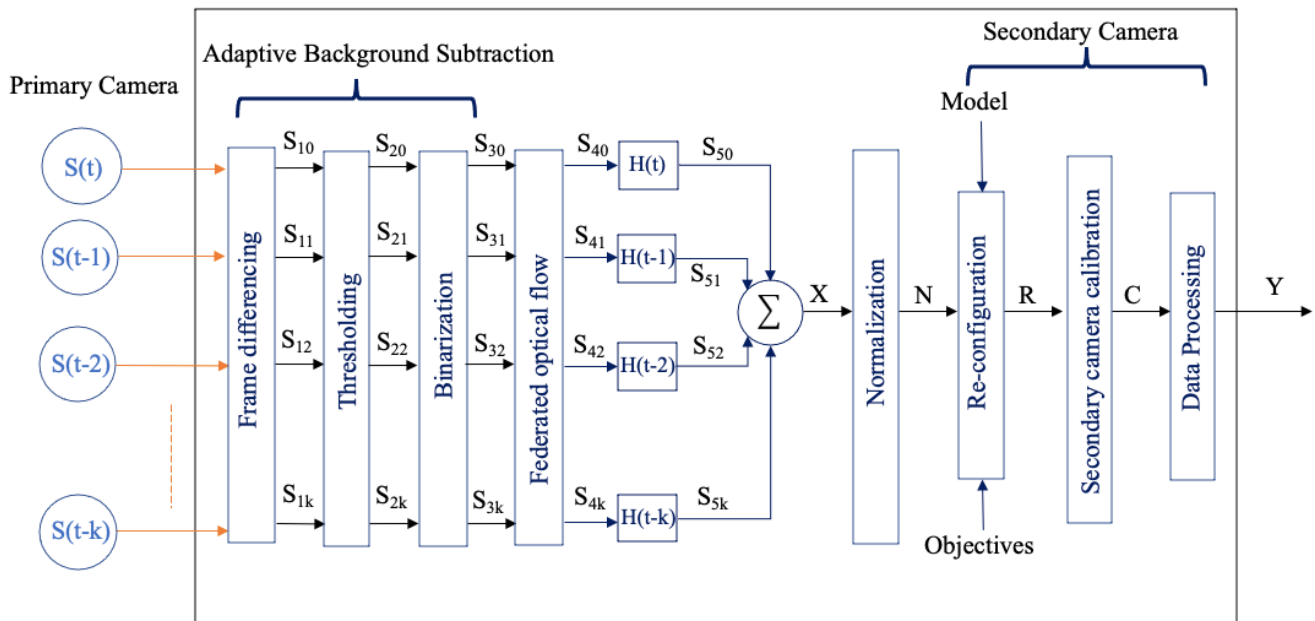


Figure 1. Adaptive spatiotemporal activity mapping framework.

3.1. Methods

• Adaptive background subtraction

To detect the moving pixels in the FOV from the data captured by the sensor, the framework utilizes frame differentiation-inspired adaptive background subtraction, which enables activity region detection and background allocation for each frame. The pixel information from each preceding frame is utilized to obtain background and foreground pixels in the next frame. For example, if $S(t)$ is the sensed data in the frame at time “t” and $S(t - 1)$ is the sensed data in the frame at time “t – 1”, then the foreground is initially extracted using frame differentiation such that each frame has its own background and foreground with reference to a relative previous frame, which can be utilized for the detection of objects in the camera FOV. The foreground information at time t (i.e., $F(t)$) is obtained by using Equation (1):

$$F(t) = S(t) - S(t - 1); \quad (1)$$

The foreground is further filtered (pre-processed) by adaptive thresholding followed by binarization to reduce noise from the foreground. A combination of frame differencing with adaptive thresholding and binarization results in adaptive background subtraction.

• Normalized Half-width Gaussian Distribution

The proposed framework uses the half-width of the normalized Gaussian distribution function to allocate temporal importance to past frames captured by the sensors. The half-width Gaussian distribution provides an accurate prioritization and importance to the past events captured by the sensor. Due to a continuous curve, the half-width Gaussian distribution provides flexibility such that it can be fragmented into any number of segments and thus is the most suitable function for the temporal relationship between the present and past frames captured by the sensor. Further, as the Gaussian distribution finds vast

applications in building temporal relationships [20,21], thus, to relate the past events to the present activities, the half-width of the normalized Gaussian distribution has been used. For the temporal relationship between the past and present frames, we used the first half of the standard normalized Gaussian distribution curve (as shown in Figure 2).

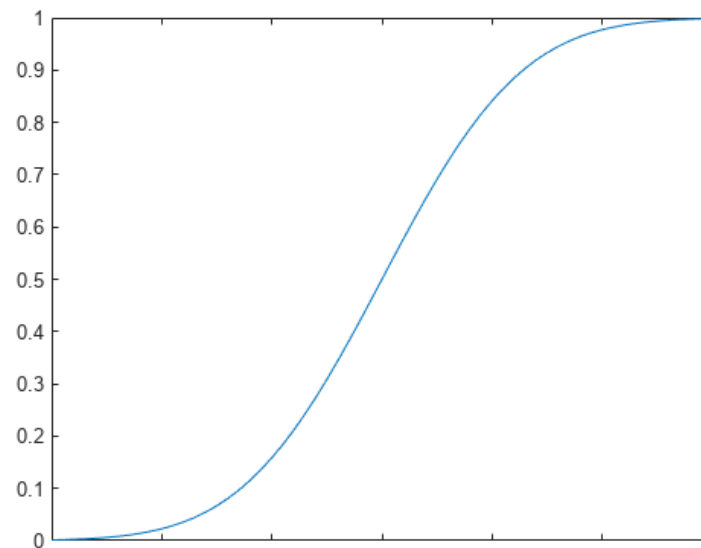


Figure 2. Half-width Gaussian Distribution function.

The normalized half Gaussian distribution function is used to assign importance to events in the present and past frames in order to assign a spatiotemporal relationship between consecutive frames captured by the primary camera. Specifically, the present events are assigned a normalized importance “1”, whereas the past events are assigned a normalized importance value according to the normalized half-width Gaussian curve and the frames relative to the time of the present frame. The half-width Gaussian curve is a continuous curve, can be segregated into any number of values temporally and is thus ideal for the spatiotemporal relationship of events.

- *Federated Optical flow*

Federated systems [22,23] combine information from various sources to reach a consensus decision. The adaptive behavior of federated systems enhances the system’s overall performance. Optical flow methods [24–26] have been extensively used to track the motion sequence of a group of pixels in consecutive frames. The two assumptions for obtaining optical flow are: (i) the group of pixels moves simultaneously in the consecutive frames and (ii) the illumination does not change in consecutive frames. In [27], Iqbal et al. presented the use of optical flow and Lukas–Kanade approaches in computer vision applications and activity mapping. The movement of pixels is obtained by determining the change in position of a pixel or cluster of pixels relative to the neighbor pixels. In the proposed framework, federated optical flow is used to design a filter to avoid the impact of undesirable activities on the ROI. The proposed framework determines the optical flow in each consecutive temporal frame captured by the sensor. Further, to filter out the unwanted objects or noise from the pre-processed foreground of each frame, the pixels from each temporal frame obtained after adaptive background subtraction are segregated into clusters by determining clusters of pixels with a consistent optical flow in consecutive temporal frames. Adaptive thresholding is used on the obtained optical flow to enhance the accuracy of the segmentation of the ROI, thus resulting in a federated optical flow.

3.2. Process

The proposed framework is designed to detect important activities captured over a period of time in the sensor’s FOV and generate an adaptive activity map for the scene.

Each pixel in the adaptive activity map is assigned dynamic pixel-sensitivity values to determine regions of importance in the scene. The adaptive activity map is further used to reconfigure the configuration space of the sensor such that the orientation of the sensor is altered to capture the ROI in the center of the sensor's FOV. A process flow of the framework is shown in Figure 3.

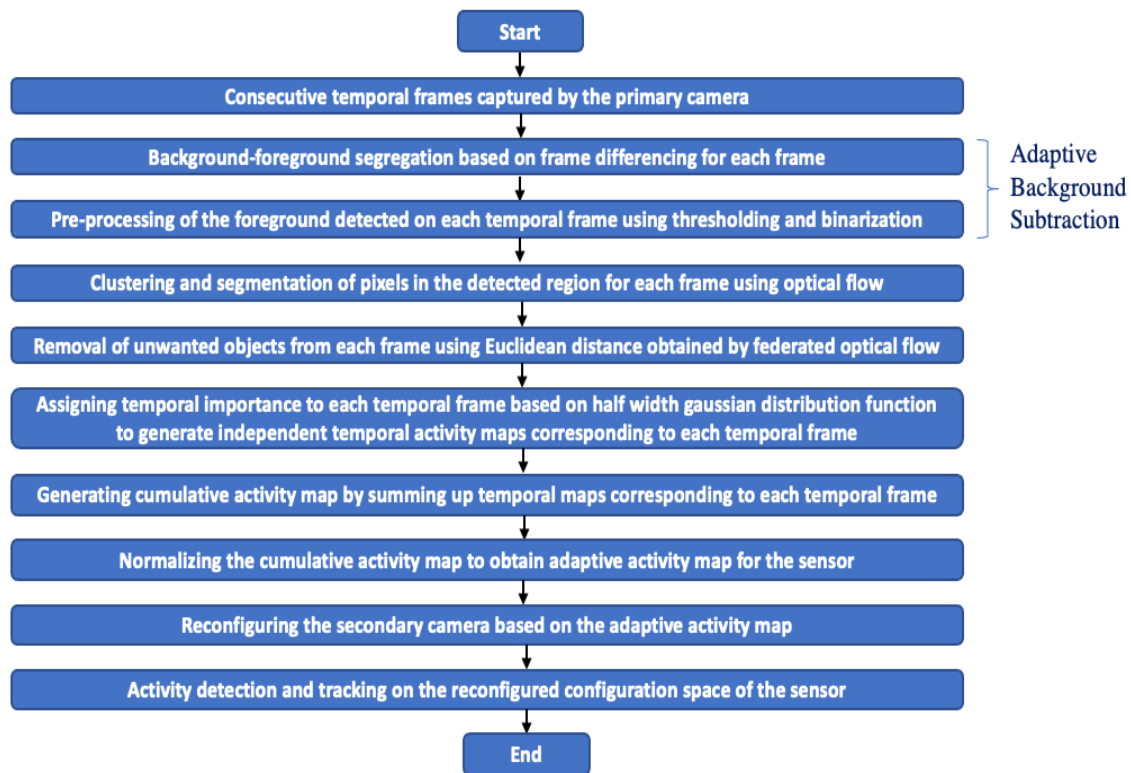


Figure 3. Process flow for adaptive spatiotemporal activity mapping.

Initially, the primary camera is configured to obtain data over time in the form of temporal frames. Each temporal frame is assigned a unique reference number based on the timestamp by the primary camera for relative reference. The data acquired by the primary camera in each temporal frame are processed to detect the foreground and background using adaptive background subtraction. The proposed framework, by way of frame differencing, generates one or more initial or unfiltered regions of importance (ROI) for each temporal frame from the determined foreground. The initial ROI for each temporal frame is pre-processed through adaptive thresholding, followed by binarization to filter the initial ROI in order to reduce noise from the foreground. A combination of frame differencing with adaptive thresholding and binarization results in adaptive background subtraction.

After pre-processing, the initial ROI of each temporal frame is clustered and segmented using optical flow. Undesirable objects are removed from each temporal frame using thresholding on a Euclidean distance obtained through federated optical flow. The filtered regions of importance of each temporal frame are assigned a temporal importance based on their relative temporal positions and half-width Gaussian distribution function, generating independent temporal maps corresponding to each temporal frame. The framework further proposes cumulating the independent temporal maps and normalizing the cumulative temporal map to generate the adaptive activity map for the scene. The regions of importance in the adaptive activity map are determined on the basis of normalized pixel sensitivity values (which are assigned to each pixel in the activity map), thus providing a pixel-wide, accurate spatiotemporal activity mapping of the scene.

The framework further utilizes a re-configurator for the reconfiguration of the secondary camera's configuration space such that the center of the ROI can be captured in the center of the FOV of the secondary camera. The re-configurator is provided with an objective specific reconfiguration model. The objectives further decide the overall functionality of the system and thus decide which activities are critical or undesirable. The adaptive activity maps are used by the re-configurator system to alter the orientation of the secondary camera and capture data with higher accuracy, thus providing better activity detection, tracking accuracy and scene understanding. In a single-camera setting, the primary camera acts as the secondary camera, and upon the generation of the adaptive spatiotemporal activity maps, it is configured to be calibrated based on the spatiotemporal activity map corresponding to the current (or latest) temporal frame.

Our contribution through this framework is twofold. Firstly, the present framework provides a spatiotemporal relationship between past and present events (or frames) captured by a sensor (i.e., the primary camera) to generate adaptive spatiotemporal activity maps for each frame, with the impact of past and present events on each activity map. The impact can be quantified in terms of a normalized sensitivity value assigned to each pixel of the activity map. Secondly, to remove the noise and improve the accuracy of the detection of objects of interest in each frame, the framework proposes federated optical-flow-based filtering prior to generating the adaptive spatiotemporal activity maps such that the accuracy of the adaptive spatiotemporal activity maps is enhanced. The adaptive spatiotemporal activity maps generated by the processing data sensed by the primary camera are further utilized for the calibration of the secondary camera to capture objects of interest with a high resolution, thus resulting in data with high information. The data captured by the secondary camera further result in better object detection and tracking and thus enhanced image understanding.

The performance of the proposed framework is evaluated in terms of multi-object tracking accuracy (MOTA) by the secondary camera in terms of truly (or accurately) detected positive and negative values of pixels in each frame and falsely (or inaccurately) detected pixels in each frame. The MOTA (%) increases if the truly detected pixel count increases and the falsely detected pixel count decreases. The performance parameters are elaborated on in Section 4.

4. Simulations and Results

The proposed framework is tested on several datasets to evaluate the efficacy on spatiotemporal activity mapping, activity detection and tracking and scene understanding. For illustration, we tested the framework using a single-camera setting on several surveillance and sports video datasets (10 seconds @ 30 fps each, i.e., 300 consecutive frames of 360×640 resolution for each video dataset) to obtain a spatiotemporal activity map after ten seconds for each video dataset for multi-object detection and tracking. The framework can be used for the calibration of reconfiguration parameters based on the generated activity map. However, the demonstration of the same requires primary and secondary sensors performing in real time. For illustration, the performance of the proposed framework is measured in terms of MOTA (%). Further, to demonstrate the results, the half-width of the half-maxima Gaussian distribution has been used, with the normalized values of the mean (μ) and standard deviation (α) being 1 and 0.5, respectively, for the temporal relation between the consecutive temporal frames. It must be noted that the half-width Gaussian distribution can be utilized for longer durations of activity analysis and thus should not be considered as a limitation to the framework. The half-width half-maxima (HWHM) Gaussian is represented in terms of the standard deviation (α) by Equation (2):

$$\text{HWHM} = \sqrt{2 \ln(2)} \alpha = 1.1799\alpha; \quad (2)$$

4.1. Performance Parameters

We have utilized markers on the consecutive temporal frames to obtain the true data of each temporal frame in the video dataset for the evaluation of the performance. The performance of the framework is analyzed in terms of multi-object tracking accuracy (MOTA). To determine the MOTA, the true positive pixel count (TPC), true positive pixel detection rate (TPR), false positive pixel count (FPC), false positive pixel detection rate (FPR), true negative pixel count (TNC), true negative pixel detection rate (TNR), false negative pixel count (FNC) and false negative pixel detection rate (FNR) are used as primary performance parameters. The expression for MOTA is formulated by Equation (3):

$$\text{MOTA (\%)} = \{(P_t - P_f)/P_t\} * 100; \quad (3)$$

where P_t represents the total pixel count in the activity map, and P_f represents the count of falsely detected or non-detected pixels in the activity map.

The total pixel count (P_t) in the activity map is represented by Equation (4):

$$P_t = \text{TPC} + \text{FPC} + \text{TNC} + \text{FNC}; \quad (4)$$

The count of falsely detected or non-detected pixels in the activity map (P_f) is represented by Equation (5):

$$P_f = \text{FPC} + \text{FNC}; \quad (5)$$

The pre-processed activity data (A_{ij}) provide the pixel activity value of each temporal frame. The pixel sensitivity is obtained by the cumulative weighted sum of the activity values obtained from all past and present frames. A weight (W_k) is assigned to each temporal frame in accordance with the HWHM Gaussian distribution. The pixel sensitivity value of the activity map is formulated by Equation (6):

$$S_{ij} = \sum_{(k)} W_k * A_{ijk} \quad (6)$$

where k represents the number of frames captured by the sensor, W_k represents the weight assigned to each temporal frame and A_{ijk} represents the pre-processed activity data of the k th temporal frame.

Further, the normalized pixel sensitivity of each pixel in the adaptive activity map can be formulated by Equation (7):

$$S_{ij} \text{ (normalized)} = S_{ij} / \text{Max}(S_{ij}); \quad (7)$$

where $\text{Max}(S_{ij})$ is the maximum pixel sensitivity value in the activity map.

Spatiotemporal activity mapping is carried out by assigning a normalized pixel sensitivity value to each pixel in the scene. ROIs can be depicted as the clusters of pixels with a high value of normalized pixel sensitivity.

4.2. Simulations

We have compared our approach with the approaches proposed in [12–14] for multi-object tracking and traffic flow estimation. The simulations and results are derived using the MatLab Image Processing Toolbox on a work station (GPU) with 128 GB of random-access memory and an Intel(R) Xeon(R) Silver 4214 CPU @ 2.19–2.20 GHz. The results on randomly selected frames from the video dataset 1 after frame extraction, binarized adapted background subtraction and Lukas–Kanade optical flow estimation are shown in Figure 4. A comparison of the activity map and normalized pixel sensitivity derived from different approaches for video dataset 1 is shown in Figure 5. The results on randomly selected frames from the video dataset 2 after frame extraction, binarized adapted background subtraction and Lukas–Kanade optical flow estimation are shown in Figure 6. A comparison of the activity map and normalized pixel sensitivity derived from different approaches for video dataset 2 is shown in Figure 7. The results on randomly selected frames from

the video dataset 3 after frame extraction, binarized adapted background subtraction and Lukas–Kanade optical flow estimation are shown in Figure 8. A comparison of the activity map and normalized pixel sensitivity derived from different approaches for video dataset 3 is shown in Figure 9. The results on randomly selected frames from the video dataset 4 after frame extraction, binarized adapted background subtraction and Lukas–Kanade optical flow estimation are shown in Figure 10. A comparison of the activity map and normalized pixel sensitivity derived from different approaches for video dataset 4 is shown in Figure 11. The results on randomly selected frames from the video dataset 5 after frame extraction, binarized adapted background subtraction and Lukas–Kanade optical flow estimation are shown in Figure 12. A comparison of the activity map and normalized pixel sensitivity derived from different approaches for video dataset 5 is shown in Figure 13. The results on randomly selected frames from the video dataset 6 after frame extraction, binarized adapted background subtraction and Lukas–Kanade optical flow estimation are shown in Figure 14. A comparison of the activity map and normalized pixel sensitivity derived from different approaches for video dataset 6 is shown in Figure 15.

A. Video dataset 1 (Traffic surveillance):

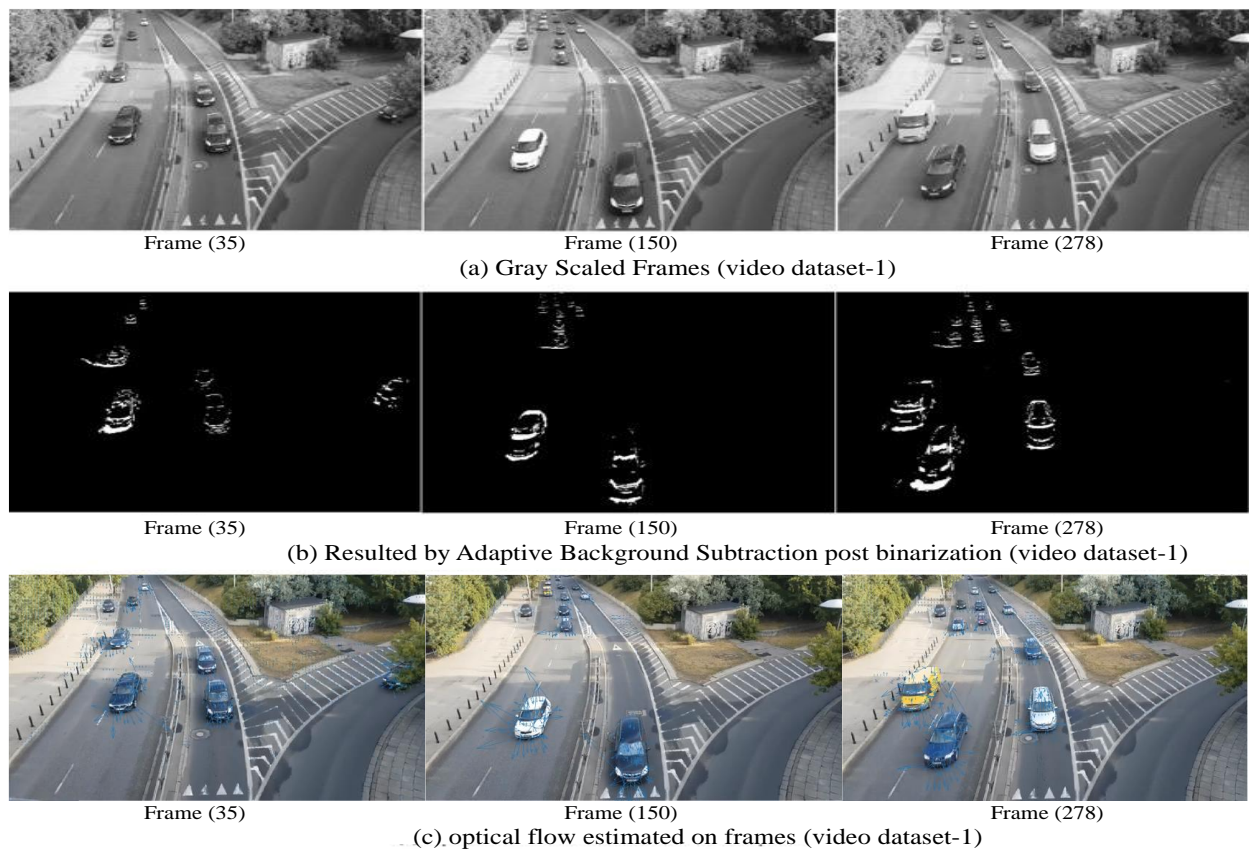


Figure 4. Simulation from video dataset 1 by the proposed framework.

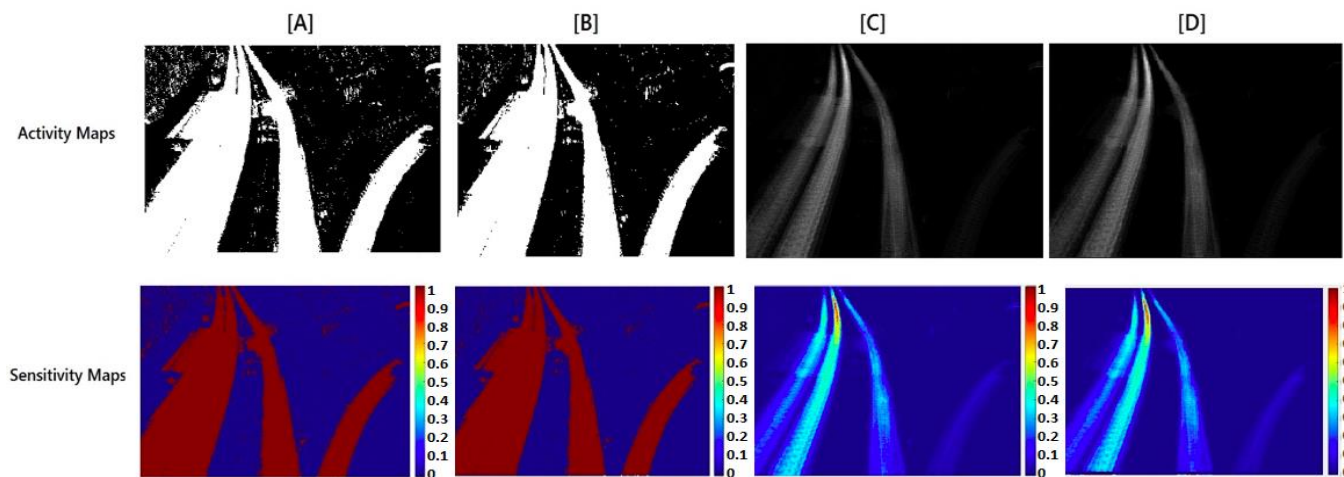


Figure 5. Activity map and pixel sensitivity maps of video dataset 1 by different approaches: (A) by Pan et al. in [12]; (B) by Mehboob et al. in [13]; (C) by Indu, S. in [14]; (D) our proposed framework.

B. Video dataset 2 (Traffic surveillance):

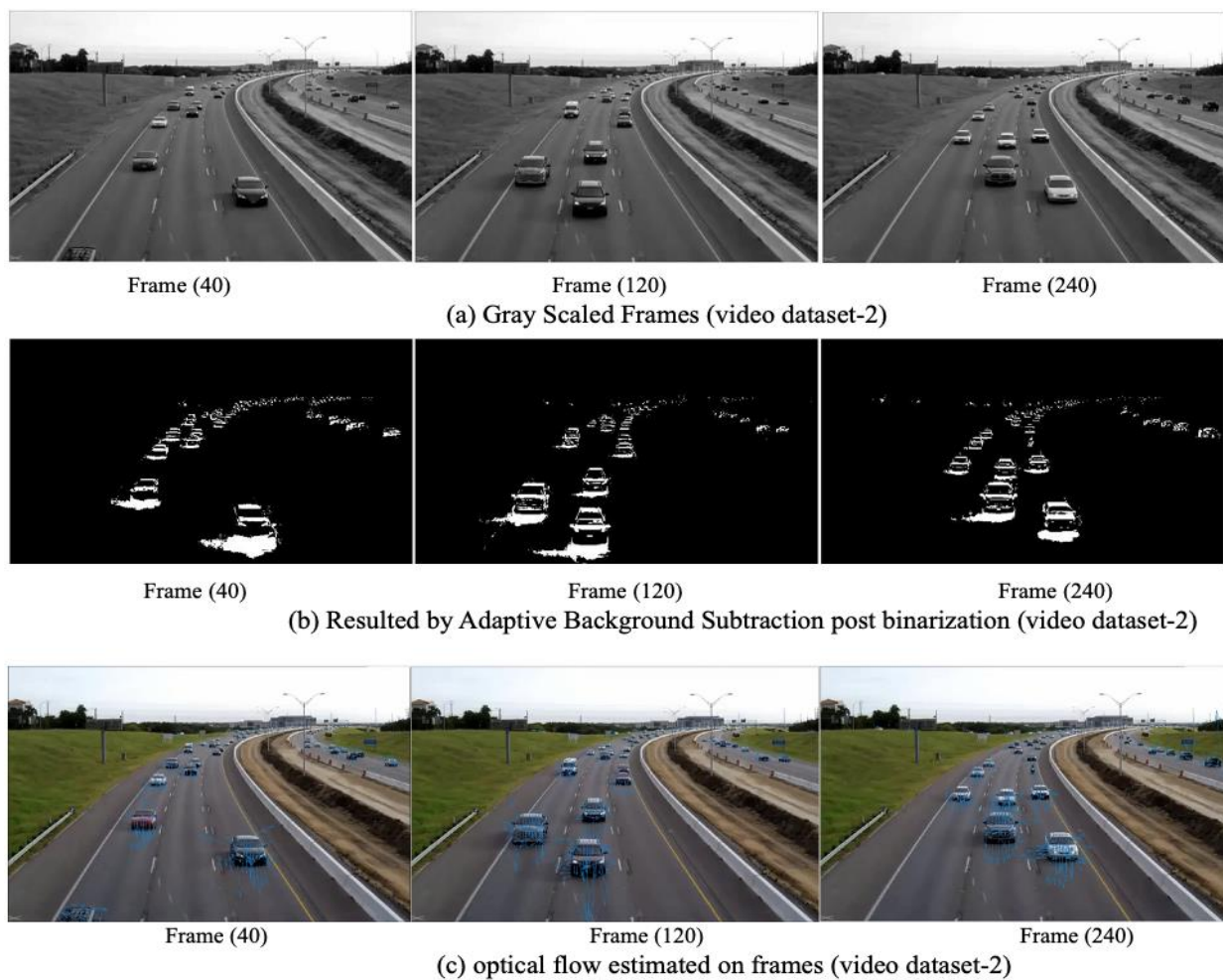


Figure 6. Simulation from video dataset 2 by the proposed framework.

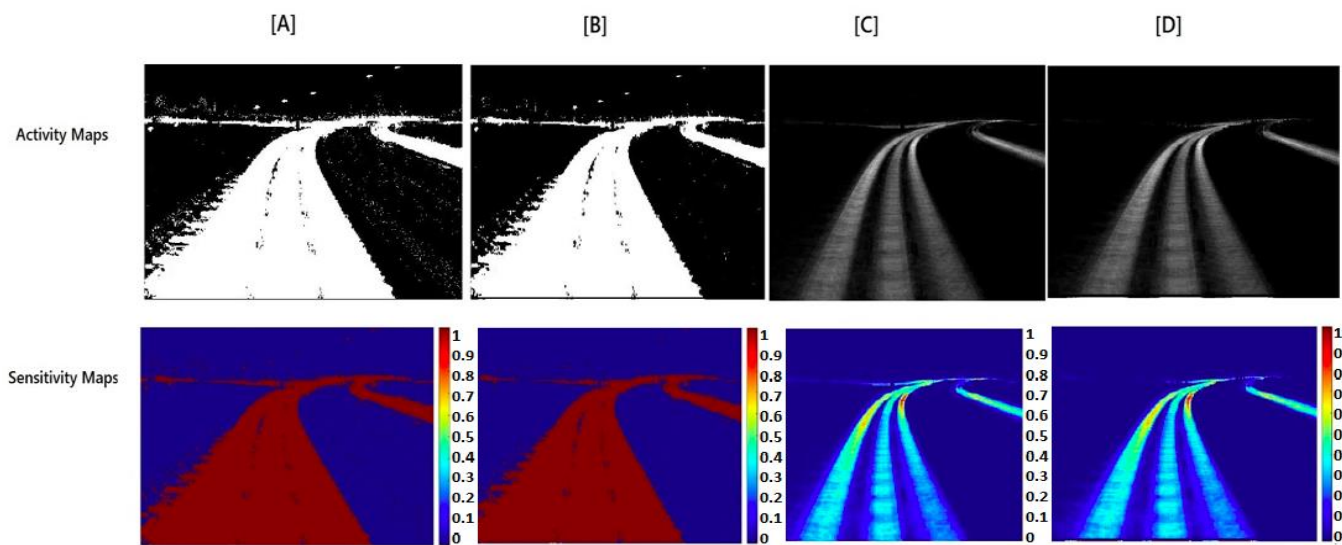


Figure 7. Activity map and pixel sensitivity maps of video dataset 2 by different approaches: (A) by Pan et al. in [12]; (B) by Mehboob et al. in [13]; (C) by Indu, S. in [14]; (D) our proposed framework.

C. Video dataset 3 (Traffic surveillance):



Figure 8. Simulation results from video dataset 3 by the proposed framework.

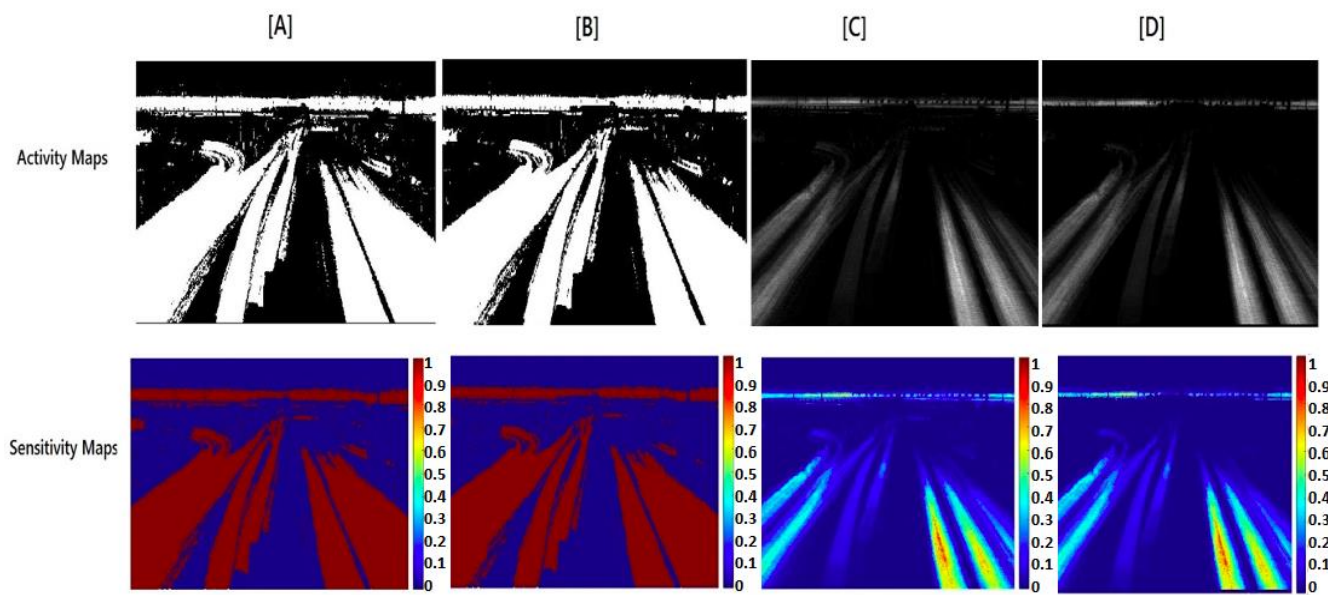


Figure 9. Activity map and pixel sensitivity maps of video dataset 3 by different approaches: (A) by Pan et al. in [12]; (B) by Mehboob et al. in [13]; (C) by Indu, S. in [14]; (D) our proposed framework.

D. Video dataset 4 (Sports—Badminton):

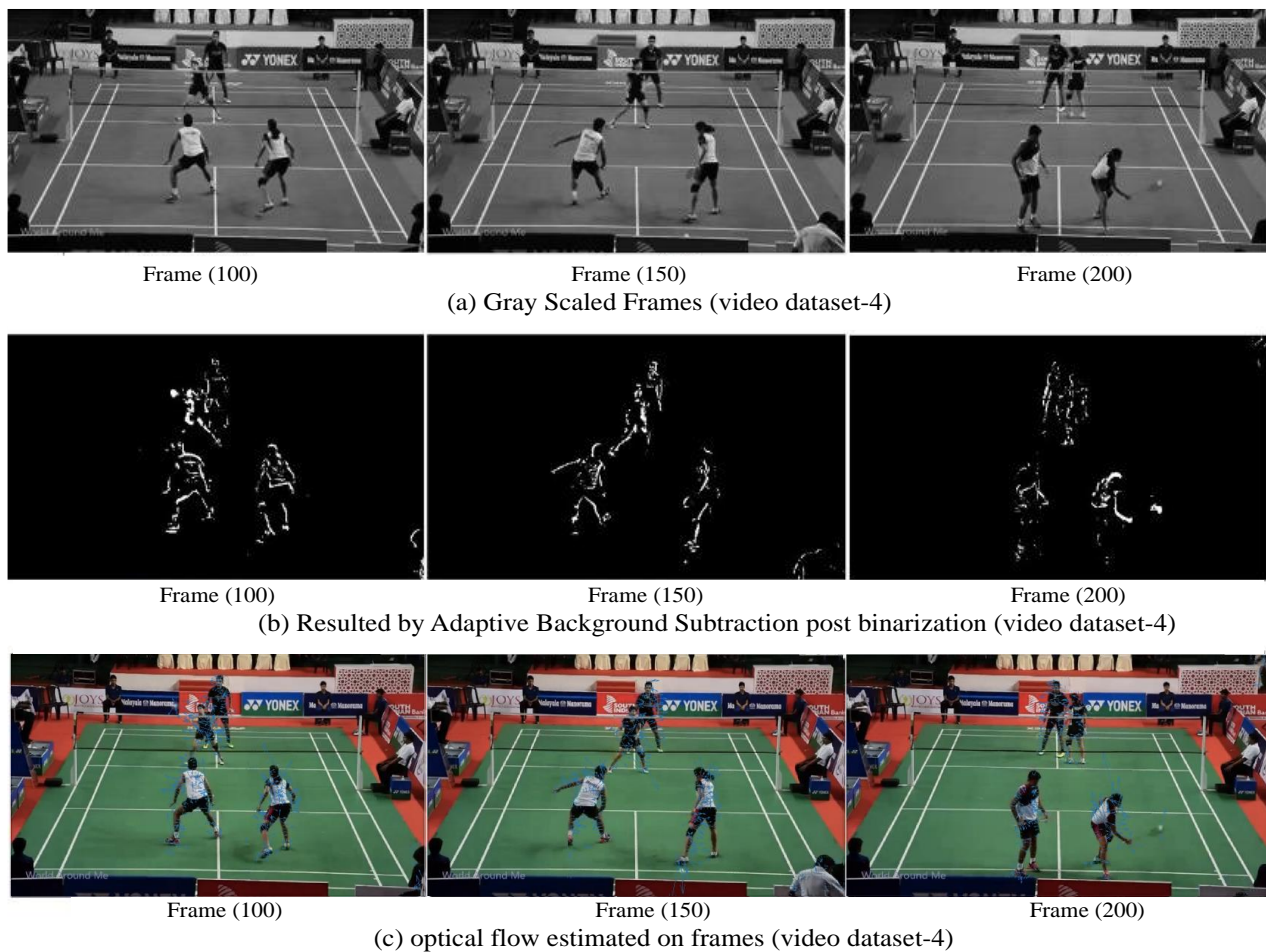


Figure 10. Simulation results from video dataset 4 by the proposed framework.

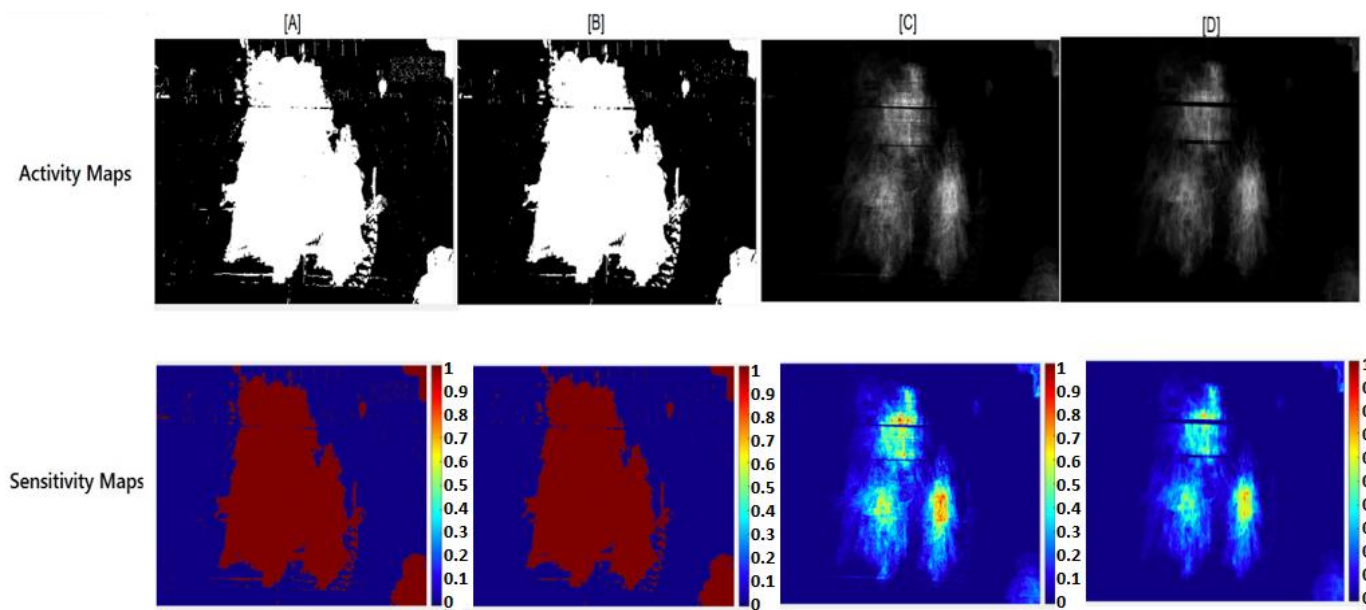


Figure 11. Activity map and pixel sensitivity maps of video dataset 4 by different approaches: (A) by Pan et al. in [12]; (B) by Mehboob et al. in [13]; (C) by Indu, S. in [14]; (D) our proposed framework.

E. Video dataset 5 (Sports—Sword fight):

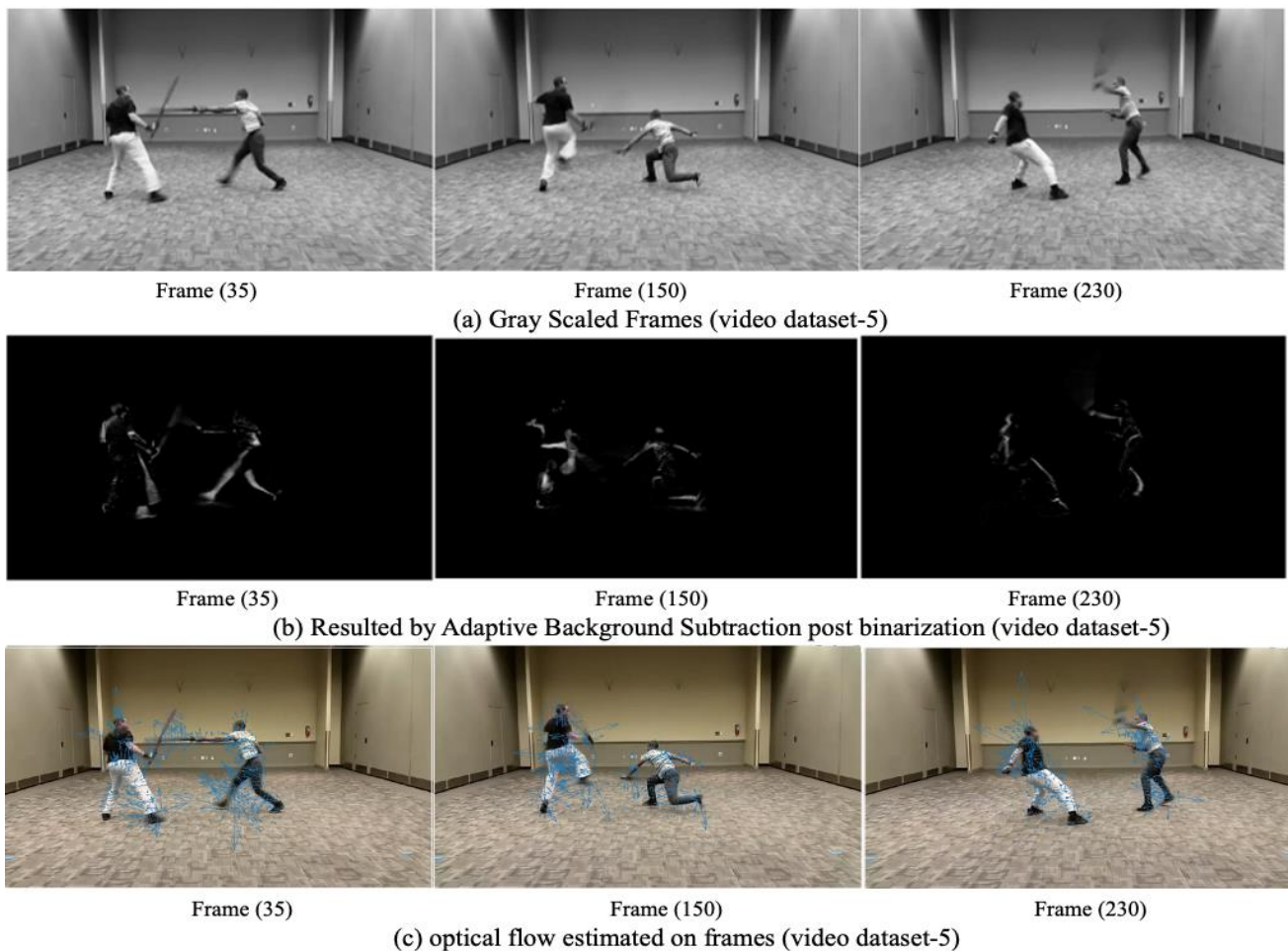


Figure 12. Simulation results from video dataset 5 by the proposed framework.

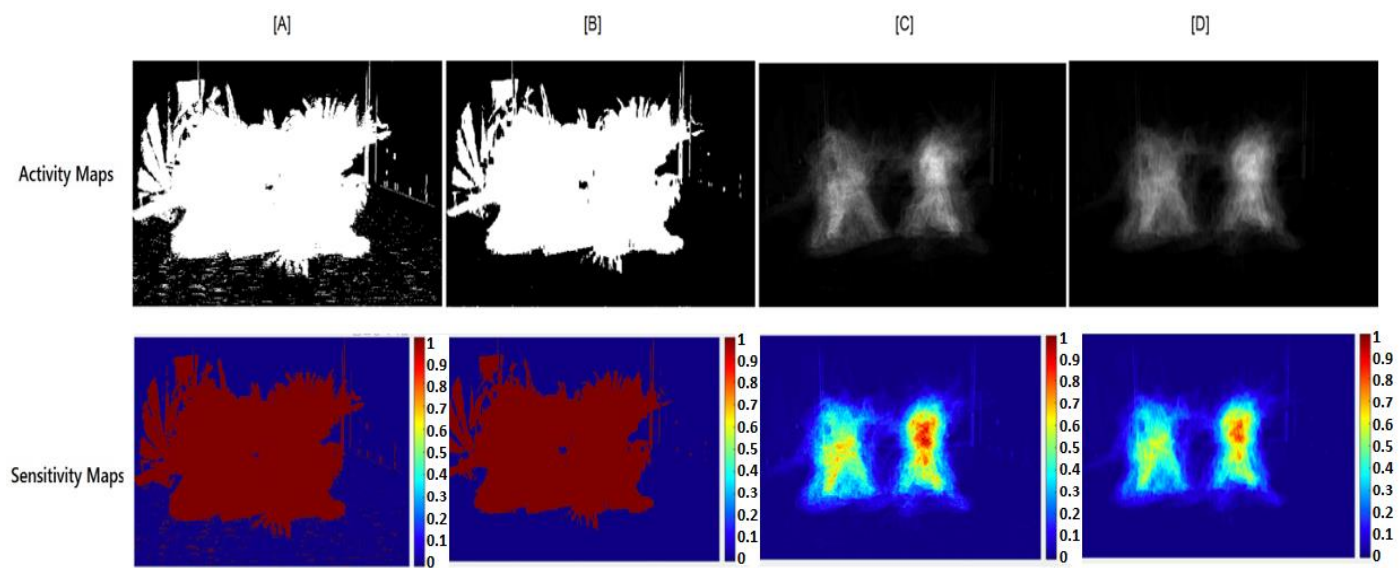


Figure 13. Activity map and pixel sensitivity maps of video dataset 5 by different approaches: (A) by Pan et al. in [12]; (B) by Mehboob et al. in [13]; (C) by Indu, S. in [14]; (D) our proposed framework.

F. Video dataset 6 (Sports—Tennis):



Figure 14. Simulation results from video dataset 6 by the proposed framework.

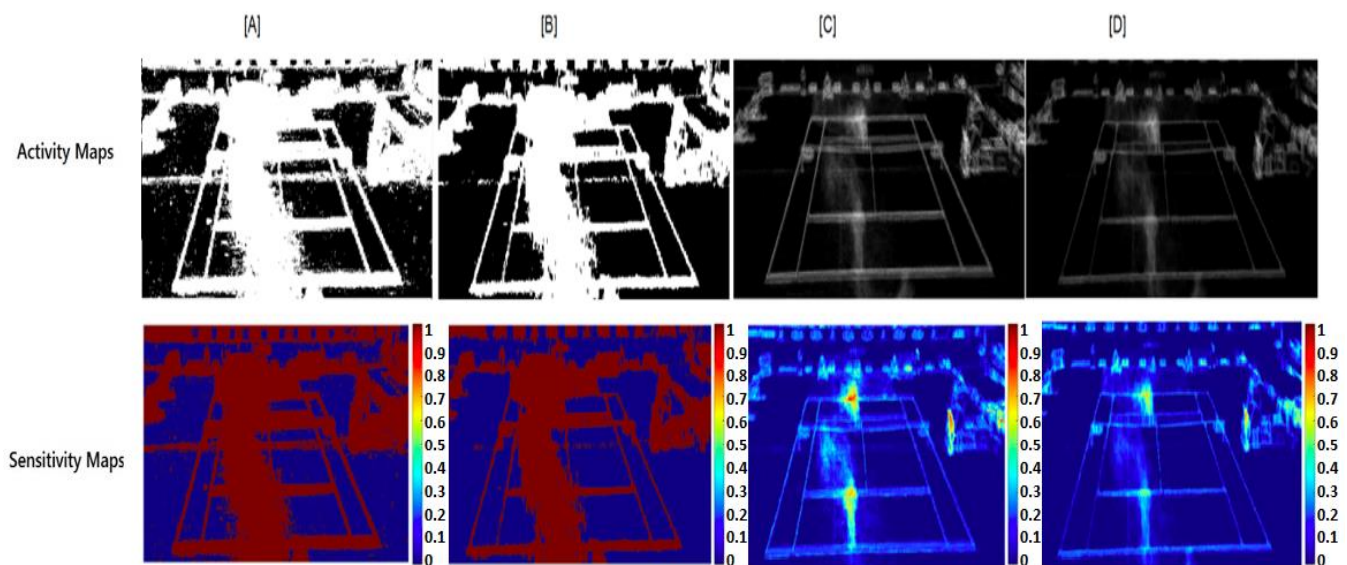


Figure 15. Activity map and pixel sensitivity maps of video dataset 6 by different approaches: (A) by Pan et al. in [12]; (B) by Mehboob et al. in [13]; (C) by Indu, S. in [14]; (D) our proposed framework.

4.3. Discussion on Simulations

The grayscale images, detected and filtered foreground and optical flow of randomly selected frames are shown in Figure 4 for video dataset 1, Figure 6 for video dataset 2, Figure 8 for video dataset 3, Figure 10 for video dataset 4, Figure 12 for video dataset 5 and Figure 14 for video dataset 6, respectively. The filtered foreground images are obtained by adaptive background subtraction by frame differencing, followed by adaptive thresholding and binarization.

Federated optical flow is used to further remove noise from the foreground generated by adaptive background subtraction. Each frame is then assigned a normalized weight based on its relative temporal position using HWHM Gaussian distribution.

Comparisons of activity maps and normalized pixel sensitivity maps derived by Pan et al. in [12], as [A], Mehboob et al. in [13], as [B], Indu, S. in [14], as [C], and our proposed framework, as [D], are shown in Figure 5 for video dataset 1, Figure 7 for video dataset 2, Figure 9 for video dataset 3, Figure 11 for video dataset 4, Figure 13 for video dataset 5 and Figure 15 for video dataset 6, respectively. The approaches presented in [12,13] do not include any temporal relation between past and present frames. Pan et al. in [12] do not present any filter for the pre-processing of the raw images obtained by the sensor. Mehboob et al., in [13], proposed the filtering of the foreground through morphological operations (close and erode), and their approach thus performed better than the approach presented by Pan et al. in [12]. However, both Pan et al. in [12] and Mehboob et al. in [13] failed to showcase the temporal effect of past frames on the present frame, which can be seen in Figure 5 for video dataset 1, Figure 7 for video dataset 2, Figure 9 for video dataset 3, Figure 11 for video dataset 4, Figure 13 for video dataset 5 and Figure 15 for video dataset 6, respectively.

Indu, S., in [14], proposed a spatiotemporal relation between past and present frames; however, the proposed method in [14] failed to provide efficient filtering of the foreground and thus resulted in inaccurate spatiotemporal activity and sensitivity maps due to the noise and undesirable regions detected in the activity map. The proposed framework, in comparison to that of Pan et al. in [12] and that of Mehboob et al. in [13], presents better spatiotemporal activity tracking in terms of normalized pixel sensitivity. Further, the proposed framework outperforms the spatiotemporal activity mapping proposed by Indu, S. in [14] by filtering out the unwanted objects and noise from the detected ROI. The

impact of falsely detected or undesirable objects in the activity map can be derived through multi-object tracking accuracy in the following Section 4.4.

4.4. Results

The comparative performance analysis of different approaches in terms of performance parameters is shown in Table 1 for video dataset 1, Table 2 for video dataset 2, Table 3 for video dataset 3, Table 4 for video dataset 4, Table 5 for video dataset 5 and Table 6 for video dataset 6, respectively.

Table 1. Comparison of performance parameters by different approaches tested on video dataset 1: [A] by Pan et al. in [12]; [B] by Mehboob et al. in [13]; [C] by Indu, S. in [14]; [D] our proposed approach; [E] the true data obtained using markers.

Ref.	TPC	TPR (%)	FPC	FPR (%)	TNC	TNR (%)	FNC	FNR (%)	MOTA (%)
Video Dataset 1 (Traffic Surveillance):									
[A]	67,830	92	39,114	53.09	117,558	75.03	5898	3.76	43.15
[B]	66,245	89.85	32,761	44.43	123,911	79.09	7483	4.77	50.80
[C]	65,924	89.41	6273	8.51	150,399	95.99	7804	4.98	86.51
[D]	65,138	88.34	1071	0.14	155,591	99.31	8590	5.48	94.38
[E]	73,728	100	0	0	1,56,672	100	0	0	100

Table 2. Comparison of performance parameters by different approaches tested on video dataset 2: [A] by Pan et al. in [12]; [B] by Mehboob et al. in [13]; [C] by Indu, S. in [14]; [D] our proposed approach; [E] the true data obtained using markers.

Ref.	TPC	TPR (%)	FPC	FPR (%)	TNC	TNR (%)	FNC	FNR (%)	MOTA (%)
Video Dataset 2 (Traffic Surveillance):									
[A]	83,262	94.50	43,149	48.97	99,146	69.67	4843	3.40	47.63
[B]	81,989	93.05	38,102	43.24	104,193	73.22	6116	4.29	52.46
[C]	80,594	91.47	8122	9.21	134,173	94.29	7511	5.27	85.52
[D]	79,813	90.59	1622	0.18	140,673	98.86	8292	5.82	94.00
[E]	88,105	100	0	0	142,295	100	0	0	100

Table 3. Comparison of performance parameters by different approaches tested on video dataset 3: [A] by Pan et al. in [12]; [B] by Mehboob et al. in [13]; [C] by Indu, S. in [14]; [D] our proposed approach; [E] the true data obtained using markers.

Ref.	TPC	TPR (%)	FPC	FPR (%)	TNC	TNR (%)	FNC	FNR (%)	MOTA (%)
Video Dataset 3 (Traffic Surveillance):									
[A]	106,274	96.15	41,827	37.84	78,051	65.11	4248	3.54	57.11
[B]	104,483	94.53	34,131	30.88	85,747	71.54	6039	5.03	64.09
[C]	102,173	92.44	9138	8.27	110,740	92.37	8349	6.96	84.77
[D]	100,628	91.04	1122	0.10	118,756	99.06	9894	8.25	91.65
[E]	110,522	100	0	0	119,878	100	0	0	100

Table 4. Comparison of performance parameters by different approaches tested on video dataset 4: [A] by Pan et al. in [12]; [B] by Mehboob et al. in [13]; [C] by Indu, S. in [14]; [D] our proposed approach; [E] the true data obtained using markers.

Ref.	TPC	TPR (%)	FPC	FPR (%)	TNC	TNR (%)	FNC	FNR (%)	MOTA (%)
Video Dataset 4 (Sports—Badminton):									
[A]	82,107	95.35	26,187	30.41	118,105	81.85	4001	2.77	66.82
[B]	80,926	93.98	19,223	22.32	125,069	86.68	5182	3.59	74.09
[C]	78,446	91.10	5982	6.94	138,310	95.8	7662	5.31	87.75
[D]	77,102	89.54	2321	2.69	141,971	98.39	9006	6.24	91.07
[E]	86,108	100	0	0	144,292	100	0	0	100

Table 5. Comparison of performance parameters by different approaches tested on video dataset 5: [A] by Pan et al. in [12]; [B] by Mehboob et al. in [13]; [C] by Indu, S. in [14]; [D] our proposed approach; [E] the true data obtained using markers.

Ref.	TPC	TPR (%)	FPC	FPR (%)	TNC	TNR (%)	FNC	FNR (%)	MOTA (%)
Video Dataset 5 (Sports—Sword Fight):									
[A]	66,121	91.87	23,877	33.17	128,547	81.14	5885	3.71	63.12
[B]	63,964	88.86	21,232	29.49	137,192	86.59	8012	5.05	65.46
[C]	58,372	81.10	12,121	16.84	146,303	92.35	13,604	8.58	74.58
[D]	54,232	75.34	8962	12.45	149,462	94.32	17,744	11.2	76.35
[E]	71,976	100	0	0	158,424	100	0	0	100

Table 6. Comparison of performance parameters by different approaches tested on video dataset 6: [A] by Pan et al. in [12]; [B] by Mehboob et al. in [13]; [C] by Indu, S. in [14]; [D] our proposed approach; [E] the true data obtained using markers.

Ref.	TPC	TPR (%)	FPC	FPR (%)	TNC	TNR (%)	FNC	FNR (%)	MOTA (%)
Video Dataset 6 (Sports—Tennis):									
[A]	46,185	94.34	20,863	42.61	160,602	88.50	2768	1.52	55.87
[B]	43,266	88.38	18,286	37.35	163,179	89.92	5687	3.13	59.52
[C]	38,128	77.89	12,112	24.74	169,353	93.32	10,825	5.97	69.29
[D]	37,109	75.81	8934	18.25	172,531	95.08	11,844	6.52	75.23
[E]	48,953	100	0	0	181,465	100	0	0	100

The average performance of the proposed framework for traffic surveillance has been obtained and compared with contemporary traffic surveillance multi-object tracking systems [28,29] (MOTA average (%) calculated for the first ten video sequences of [28,29]). Further, the average performance of the proposed framework for sports analytics has been obtained and compared with a contemporary sports activity tracking system [30] (MOTA average (%) obtained from the RGB sequence of [30]). The comparison of the performance of the proposed framework with that of [28–30] in terms of average MOTA (%) is shown in Figure 16.

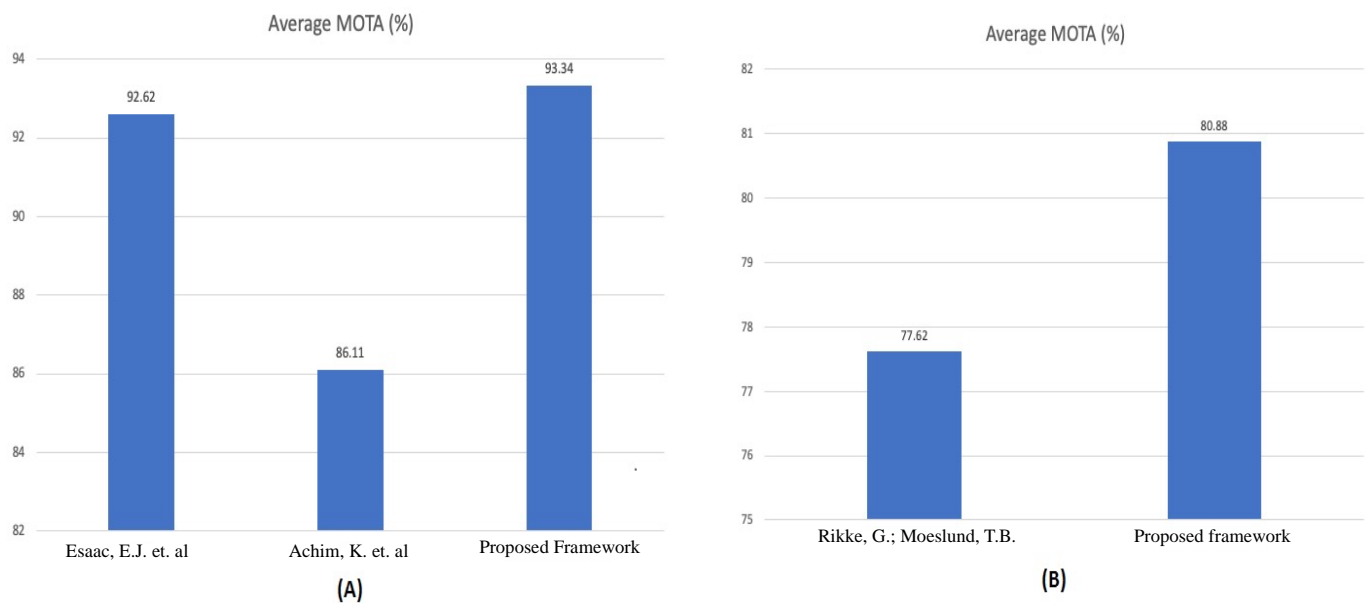


Figure 16. Comparison of the performance in terms of average MOTA %: (A): Comparison of the average MOTA % of the proposed framework with that of Isaac, E.J. et. al in [28] and Achim, K. et.al in [29]; (B): Comparison of the average MOTA % of the proposed framework with that of Rikke, G.; Moeslund, T.B. in [30].

The average MOTA (%) of the proposed framework for traffic surveillance is obtained as 93.34%, as calculated by Equation (8):

$$\{\text{MOTA (dataset-1)} + \text{MOTA (dataset-2)} + \text{MOTA (dataset-3)}\} / 3; \quad (8)$$

The average MOTA (%) of the proposed framework for sports analytics is obtained as 80.88%, as calculated by Equation (9):

$$\{\text{MOTA (dataset-4)} + \text{MOTA (dataset-5)} + \text{MOTA (dataset-6)}\} / 3; \quad (9)$$

5. Conclusions

The accuracy of the data collected by sensors and the spatiotemporal understanding of the scene are two key components of a successful computer vision system. Most of the advanced computer vision systems address both of the abovementioned components using a highly complex computation model, which may be a problem for most of the systems with limited resources. Through this article, a framework for spatiotemporal activity mapping capable of handling the trade-off between resource limitations and the performance of the computer vision system is proposed.

The framework evaluates the scene spatiotemporally and produces adaptive activity maps for the re-configuration of the sensor such that the region(s) of importance can be captured in the center of the sensor's field of view. The framework utilizes simple image-processing tools such as adaptive background subtraction, binarization, thresholding and federated optical flow for pre-processing the sensor data. Half-width Gaussian distribution is used for the temporal relationship between the present and past frames. The simple model of the proposed framework results in a low computational complexity and thus low resource utilization.

The performance of the framework is compared in terms of multi-object tracking accuracy (MOTA) and has been tested on multiple traffic surveillance and sports datasets. The framework outperforms the contemporary systems presented in [12–14,28–30]. The framework showcased a 0.71% better average MOTA compared to [28] and an 8.39% better average MOTA compared to [29] when tested on traffic surveillance datasets (i.e., datasets 1, 2 and 3). The framework further showcased a 4.21% better average MOTA compared

to [30] when tested on sports datasets (i.e., datasets 4, 5 and 6). Artificial intelligence (AI)-based systems [16–19] require high computation and storage capabilities to train the model and further require iteratively changing the training model in unforeseen conditions, illumination changes, etc. AI-based multi-object tracking systems are also susceptible to adversarial attacks, which makes the timely training of the system critically necessary. The proposed framework performs similarly to the AI-based multi-object detection systems proposed in [16–19] without the high computation or storage requirements needed to handle unforeseen conditions. Thus, the proposed framework showcases a balance in terms of resource utilization and performance.

Author Contributions: S.: Lead author of the manuscript (corresponding author), conceptualization and methodology, writing—original draft preparation, investigation and editing; I.S.: Second author, research design, guidance and reviewing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The proposed framework has been tested on three surveillance datasets (i.e., video dataset 1, video dataset 2 and video dataset 3) and three sports datasets (video dataset 4, video dataset 5 and video dataset 6). The video datasets and codes for the simulation models for the proposed framework can be accessed by clicking on this <https://drive.google.com/drive/folders/1a3SG5qEOL25e7pFIPOb0GCvGKL6X7ghN> (last accessed on 12 December 2022).

Acknowledgments: This work was carried out under the supervision of Indu Sreedevi at the Department of ECE, Delhi Technological University, New Delhi, India, and Shashank expresses immense gratitude to his guide and to UGC for enlightening him throughout the process.

Conflicts of Interest: The authors declare that they have no conflict of interest. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

References

1. AI in Computer Vision Market Research Report by Component (Hardware, Software), Vertical (Healthcare, Security, Automotive, Agriculture, Sports & Entertainment, and Others), and Region—Global Forecast to 2027. Available online: <https://www.expertmarketresearch.com/reports/ai-in-computer-vision-market> (accessed on 22 August 2022).
2. Tadic, V.; Toth, A.; Vizvari, Z.; Klincsik, M.; Sari, Z.; Sarcevic, P.; Sarosi, J.; Biro, I. Perspectives of RealSense and ZED Depth Sensors for Robotic Vision Applications. *Machines* **2022**, *10*, 183. [\[CrossRef\]](#)
3. Shashank; Sreedevi, I. Distributed Network of Adaptive and Self-Reconfigurable Active Vision Systems. *Symmetry* **2022**, *14*, 2281. [\[CrossRef\]](#)
4. Li, S.; Huang, M.; Guo, M.; Yu, M. Evaluation model of autonomous vehicles' speed suitability based on overtaking frequency. *Sensors* **2021**, *21*, 371. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Saini, N.; Bonetto, E.; Price, E.; Ahmad, A.; Black, M.J. AirPose: Multi-View Fusion Network for Aerial 3D Human Pose and Shape Estimation. *IEEE Robot. Autom.* **2022**, *7*, 4805–4812. [\[CrossRef\]](#)
6. Lo, L.Y.; Yiu, C.H.; Tang, Y.; Yang, A.S.; Li, B.; Wen, C.Y. Dynamic Object Tracking on Autonomous UAV System for Surveillance Applications. *Sensors* **2021**, *21*, 7888. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Xu, C.; Zhang, K.; Jiang, Y.; Niu, S.; Yang, T.; Song, H. Communication aware UAV swarm surveillance based on hierarchical architecture. *Drones* **2021**, *5*, 33. [\[CrossRef\]](#)
8. Indu, S.; Chaudhury, S.; Mittal, N.R.; Bhattacharyya, A. Optimal sensor placement for surveillance of large spaces. In Proceedings of the Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), Como, Italy, 30 August–2 September 2009; pp. 1–8.
9. Zhang, G.; Dong, B.; Zheng, J. Visual Sensor Placement and Orientation Optimization for Surveillance Systems. In Proceedings of the 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA), Krakow, Poland, 4–6 November 2015; pp. 1–5.
10. Silva, L.S.B.D.; Bernardo, R.M.; Oliveira, H.A.; Rosa, P.F.F. Multi-UAV agent-based coordination for persistent surveillance with dynamic priorities. In Proceedings of the International Conference on Military Technologies (ICMT), Brno, Czech Republic, 31 May–2 June 2017; pp. 765–771.
11. Ahad, M.A.R. Action Representations. In *Motion History Images for Action Recognition and Understanding*; Book Chapter; Springer: Berlin, Germany, 2013; pp. 19–29.
12. Pan, X.; Guo, Y.; Men, A. Traffic Surveillance System for Vehicle Flow Detection. In Proceedings of the Second International Conference on Computer Modeling and Simulation, Sanya, China, 22–24 January 2010; pp. 314–318.

13. Mehboob, F.; Abbas, M.; Almotaeryi, R.; Jiang, R.; Maadeed, S.A.; Bouridane, A. Traffic Flow Estimation from Road Surveillance. In Proceedings of the IEEE International Symposium on Multimedia (ISM), Miami, FL, USA, 14–16 December 2015; pp. 605–608.
14. Shashank; Indu, S. Sensitivity-Based Adaptive Activity Mapping for Optimal Camera Calibration. In Proceedings of the International Conference on Intelligent Computing and Smart Communication, Tehri, India, 20–21 April 2019; Springer: Berlin, Germany, 2019; pp. 1211–1218.
15. Stuede, M.; Schappler, M. Non-Parametric Modeling of Spatio-Temporal Human Activity Based on Mobile Robot Observations. *arXiv* **2022**, arXiv:2203.06911.
16. Sattar, S.; Sattar, Y.; Shahzad, M.; Fraz, M.M. Group Activity Recognition in Visual Data: A Retrospective Analysis of Recent Advancements. In Proceedings of the International Conference on Digital Futures and Transformative Technologies (ICoDT2), Islamabad, Pakistan, 20–21 May 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
17. Zhao, L.; Gao, Y.; Ye, J.; Chen, F.; Ye, Y.; Lu, C.T.; Ramakrishnan, N. *Online Dynamic Multi-Source Feature Learning and Its Application to Spatio-Temporal Event Forecasting*; ACM Transactions on Knowledge Discovery from Data. 2021, Volume 1. Available online: http://cs.emory.edu/~jlzhao41/materials/papers/TKDD2020_preprinted.pdf (accessed on 14 December 2022).
18. Yuanqiang, L.; Jing, H. A Sports Video Behavior Recognition Using Local Spatiotemporal Patterns. *Mob. Inf. Syst.* **2022**, 2022. [CrossRef]
19. Yan, R.; Shu, X.; Yuan, C.; Tian, Q.; Tang, J. Position-aware participation-contributed temporal dynamic model for group activity recognition. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 33, 7574–7588. [CrossRef] [PubMed]
20. Piergiovanni, A.J.; Michael, R. Temporal Gaussian mixture layer for videos. In Proceedings of the International Conference on Machine learning, Long Beach, CA, USA, 9–15 June 2019; pp. 5152–5161.
21. Wang, B.; Kuo, J.; Bae, S.C.; Granick, S. When Brownian diffusion is not Gaussian. *Nat. Mater.* **2012**, 11, 481–485. [CrossRef] [PubMed]
22. Jakub, K.; McMahan, H.B.; Yu, X.F.; Richtárik, P.; Suresh, A.T.; Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv* **2016**, arXiv:1610.05492.
23. Keith, B.; Eichner, H.; Grieskamp, W.; Huba, D.; Ingerman, A.; Ivanov, V.; Kiddon, C. Towards federated learning at scale: System design. *arXiv* **2019**, arXiv:1902.01046.
24. Jeongho, S.; Kim, S.; Kang, S.; Lee, S.W.; Paik, J.; Abidi, B.; Abidi, M. Optical flow-based real-time object tracking using non-prior training active feature model. In Proceedings of the Advances in Multimedia Information Processing—5th Pacific Rim Conference on Multimedia, Tokyo, Japan, 30 November–5 December 2004; pp. 69–78.
25. Baker, S.; Matthews, I. Lucas-kanade 20 years on: A unifying framework. *Int. J. Comput. Vis.* **2004**, 56, 221–255. [CrossRef]
26. Nusrat, S.; Brad, R. Optimal Filter Estimation for Lucas-Kanade Optical Flow. *Sensors* **2012**, 12, 12694–12709. [CrossRef]
27. Umair, I.; Perez, P.; Li, W.; Barthelmy, J. How computer vision can facilitate flood management: A systematic review. *Int. J. Disaster Risk Reduct.* **2021**, 53, 102030.
28. Isaac, E.J.; Martin, J.; Barco, R. A low-complexity vision-based system for real-time traffic monitoring. *IEEE Trans. Intell. Transp. Syst.* **2016**, 18, 1279–1288.
29. Achim, K.; Sefati, M.; Arya, S.; Rachman, A.; Kreisköther, K.; Campoy, P. Towards Multi-Object Detection and Tracking in Urban Scenario under Uncertainties. In Proceedings of the 4th International Conference on Vehicle Technology and Intelligent Transport Systems, VEHITS, Funchal, Madeira, Portugal, 16–18 March 2018; pp. 156–167.
30. Rikke, G.; Moeslund, T.B. Constrained multi-target tracking for team sports activities. *IPSJ Trans. Comput. Vis. Appl.* **2018**, 10, 2.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Study of third harmonic generation in $\text{In}_x\text{Ga}_{1-x}\text{As}$ semi-parabolic 2-D quantum dot under the influence of Rashba spin-orbit interactions (SOI): Role of magnetic field, confining potential, temperature & hydrostatic pressure

Suman Dahiya^a, Siddhartha Lahon^{b,*}, Rinku Sharma^{a,**}

^a Department of Applied Physics, Delhi Technological University, Delhi, 110042, India

^b Physics Department, KMC, University of Delhi, Delhi, 110007, India

ARTICLE INFO

Handling editor: J. Nitta

Keywords:

Quantum dot
Spin-orbit interaction
Magnetic field
Third harmonic generation
Temperature

ABSTRACT

In the present case, nonlinear optical susceptibility for $\text{In}_x\text{Ga}_{1-x}\text{As}$ 2-D semi-parabolic Quantum Dot with the key prominence given to the magnetic field, Hydrostatic pressure, confining potential and Temperature on THG in the presence of Rashba SOI is investigated. The main expression of THG is attained using a formalism of compact density matrix. Our results are showing that rise/diminution in the Rashba SOI coefficient strongly affects the THG peaks. Also, a blue/redshift is observed with an increase/decrease in external factors such as Hydrostatic Pressure, magnetic field, Temperature, confining potential & Rashba SOI with a corresponding increase/decrease in peak height. According to the observation, two-photon resonance peaks are found to be stouter than one & three-photon resonance peaks as with a significant increase in the coupling, the strength of the dipole matrix element also increases in correspondence to the peak height. The conclusions are demonstrating that for the comprehensive engineering of optical devices based on the QDs, SOI must be taken into consideration, and hence by tuning the strength parameter, the optical properties of the optoelectronic devices can be controlled.

1. Introduction

As quantum confinement has become apparent in all spatial directions, ultra-small semiconductor heterostructures have gained the immense focus of researchers around the world. These heterostructures such as quantum heterostructure including quantum dot, quantum wire, etc. Have many unique potential applications due to the confinement as the movement of charge carrier is restricted and hence leads to the development of a set of discrete energy levels where the carriers may exist. There are many microfabrication techniques including MBE, lithography, vapor deposition technique, etc. to fabricate these quantum heterostructures of different shapes and sizes. Explicitly, optical properties such as linear and nonlinear optical properties & susceptibilities have more scientific interest as they offer massive efficacy in understanding numerous semiconductor optoelectronic devices such as quantum LED's, solar cells, quantum dot lasers, single-electron transistors & quantum computing computers, etc. [1–7].

The role of externally applied fields, SOI, temperature & hydrostatic pressure, etc. is very significant in altering the properties of the quantum heterostructures as any prominent change in the property can result in momentous changes in the working of the nano-scale device [8–11]. This dependence has often been used to externally alter the properties of these nano-scale devices to the operator's own will and thus has been one of the most widely researched areas in recent times. On application of a perpendicular magnetic field to the plane of QD, energy levels are supplied with a supplementary structure as well as interacting electrons confined in QD results in correlation effect. The study of the electronic, optical and thermodynamic properties has been done by many different techniques. Theoretically, the 2-electron QD Hamiltonian with the inclusion of the effect of the magnetic field has been solved by many authors for obtaining the respective eigen energies and eigenstates of the QD-system [12–20] but to our sincere belief, the effect of external factors & SOI with doping is relatively a less discovered area.

Second and third-order nonlinear optical interaction of 2- incident

* Corresponding author.

** Corresponding author.

E-mail addresses: dahiyasuman90@gmail.com (S. Dahiya), sid.lahon@gmail.com (S. Lahon), rinkusharma@dtu.ac.in (R. Sharma).

<https://doi.org/10.1016/j.physe.2022.115620>

Received 20 September 2022; Received in revised form 14 November 2022; Accepted 16 December 2022

Available online 21 December 2022

1386-9477/© 2022 Elsevier B.V. All rights reserved.

fields with optical media results in the generation of SHG & THG and ORC. Xie and Bass et al., in 1962 performed initial work on OR [21], whereas Franken et al. were the first to report the experimental observation of SHG [22,23]. Baskoutas et al. investigated the impact of exciton in optical susceptibility for a semi-parabolic Quantum dot for a semi-parabolic potential [24] whereas F. Ungan et al. focused on the impact of the Electric field on TAC and RICC of an asymmetric Quantum dot [25]. Xuechao Li et al. [26], focused on the outcome of the Magnetic field on TAC of an asymmetric Quantum dot and studied the variations with parameters such as radius and magnetic field intensity.

Many important factors such that such as temperature, hydrostatic pressure, applied electric and magnetic fields and intense laser fields have a significant effect on the linear and nonlinear optical properties of Q D's. On the application of these external field factors, the band structure, and the optical nonlinearity of the QD system is controlled and altered. Hence, concluding that the geometry of systems, as well as external perturbations, are equally contributing significantly influencing the nonlinear optical properties of semiconductor structures. Uni-dimensional quantum dots have been studied immensely in recent times, but studies on the 2D semi-parabolic quantum dot with doping focusing on the dependence of its properties on external factors are scarce. In this study, we explored the variation of optical properties, such as THG on applied hydrostatic pressure, temperature, magnetic field, Rashba spin for $\text{In}_x\text{Ga}_{1-x}\text{As}$ Quantum dot in the presence of SOI. The first section contains a crisp introduction of the topic, whereas the theory and the formulas have been mentioned in the second section. The third section consists of the obtained results and the graphs and is followed by the fourth section which contains well-drawn solutions.

1.1. Theoretical model

This section describes the detailed theory of 1-electron QD consisting of two parts given as (1) QD Hamiltonian (2) exact diagonalization method for the $\text{In}_x\text{Ga}_{1-x}\text{As}$ QD.

Considering a 2-D $\text{In}_x\text{Ga}_{1-x}\text{As}/\text{In}_y\text{Al}_{1-y}\text{As}$ QD (see Fig. 1) with a semi-parabolic confining potential as given in equation (3), with a vertical magnetic field given as $\vec{B} = B\hat{k}$ having a symmetric gauge as: $A = B(-y, x, 0)/2$, the Hamiltonian for a one-electron system within effective mass taking spin into account is given as [27–31]:

$$H_{ST} = H_{ws} + H_{so}^R + \frac{1}{2}g^*\mu_B B\sigma \quad (1)$$

where

$$H_{ws} = \frac{1}{2m^*} \left[P - \frac{e}{c} A(r) \right]^2 + V_r \quad (2)$$

And the potential is given as

$$V_r = \frac{1}{2}m^*(P, T)\omega_0^2(x^2 + y^2) \quad (3)$$

In eq. (1), g^* represents the effective Lande factor for the semiconductor and electron spin z projection is given by $\frac{1}{2}\sigma$ having $\sigma = \pm$. Here up spin and down spin is represented by $\sigma = +1$ and $\sigma = -1$ respectively. SOI term H_{SO} contains two parts (i) Dresselhaus SOI term, H_{SO}^D (ii) Rashba SOI term, H_{SO}^R . As H_{SO}^R dominates over H_{SO}^D for the narrow gap, hence, neglecting H_{SO}^D , we have H_{SO}^R as:

$$H_{SO}^R = \frac{\alpha}{\hbar} \left[\vec{\sigma} \times \left(\vec{p} - \frac{e}{c} \vec{A} \right) \right]_z \quad (4)$$

The term $\frac{\alpha}{\hbar} \left[\vec{\sigma} \times \left(\vec{p} - \frac{e}{c} \vec{A} \right) \right]_z$ represents the spin-orbit coupling due to the inhomogeneous potential confining the electrons to the 2D plane and possible external gate voltages applied on the top of the dot. The strength of this coupling is determined by these parameters with a variation in magnitude when external gate voltages are applied. Here, the Rashba coupling coefficient is given by α and is controllable by varying the applied gate voltage in z direction, spinors and canonical momentum are represented by σ_x & σ_y and p_x & p_y respectively.

In effect, mass approx., total Hamiltonian H_{ST} for the combination of H_{ws} and H_{SO}^R is given as:

$$H_{ST} = \frac{p^2}{2m^*} + \frac{e}{m^*} A \cdot p + \frac{e^2 A^2}{2m^*} + V_r + \frac{1}{2}g^*\mu_B B\sigma + \frac{\alpha}{\hbar} \left[\vec{\sigma} \times \left(\vec{p} - \frac{e}{c} \vec{A} \right) \right]_z \quad (5a)$$

$$H_{ST} = \frac{p^2}{2m^*} + \frac{e}{m^*} A \cdot p + \frac{e^2 A^2}{2m^*} + V_r + \frac{1}{2}g^*\mu_B B\sigma + \frac{\alpha}{\hbar} [\sigma \times p]_z + \frac{e\alpha}{\hbar} [\sigma \times A]_z \quad (5b)$$

Here:

m^* = represents effect. mass of charge carrier.

$\hbar\omega_0$ = confining potential strength corresponding to the size of the QD.

The Temperature and hydrostatic dependent Effect. mass of the electron for GaAs is given as [32]:

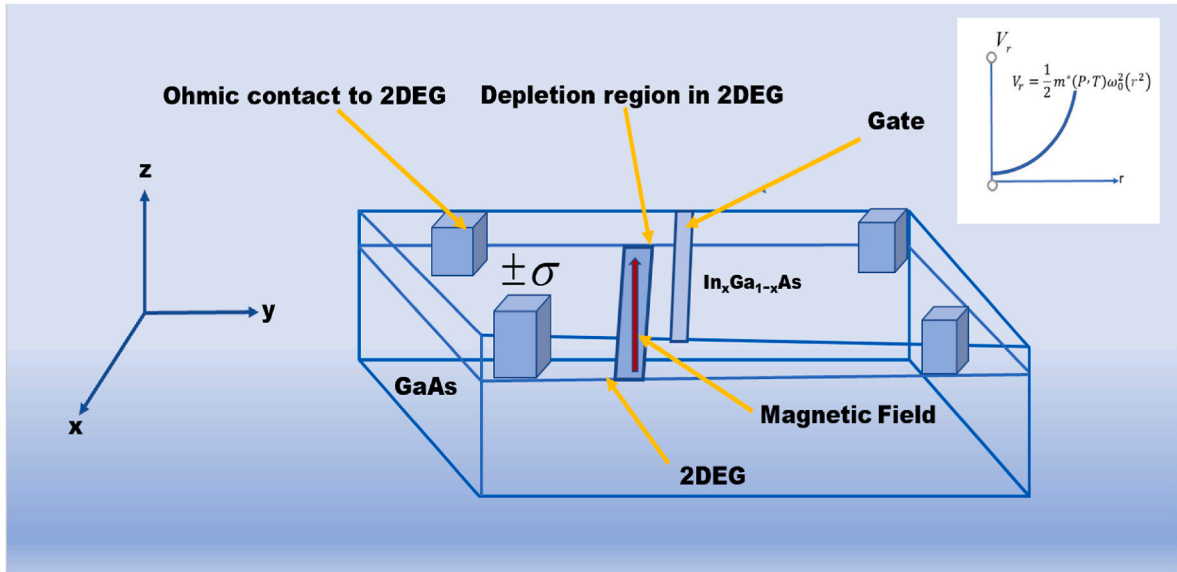


Fig. 1. Schematic diagram of an $\text{In}_x\text{Ga}_{1-x}\text{As}$ Quantum Dot.

$$m_e^*(P, T) = m_o \left[1 + \frac{7510}{E_g(P, T) + 341} + \frac{15020}{E_g(P, T)} \right]^{-1} \quad (6a)$$

$$\text{With } E_g(P, T) = \left[1519 - \frac{0.5405T^2}{T + 204} + 10.7P \right] \quad (6b)$$

Here Temperature and hydrostatic pressure-dependent energy gap for GaAs, E_g is in meV, P is in “kbar” and T is in “Kelvin.” The pressure-dependent oscillator frequency is expressed as

$$\omega(P) = \omega_0 / [1 - 2P(1.16 \times 10^{-3} - 7.4 \times 10^{-4})] \quad (6c)$$

Eigenfunction of Hamiltonian in eq (1) is represented by Fock-Darwin states $|n, l\rangle$ and is given as:

$$\psi_{nl\sigma}(r) = \frac{1}{\sqrt{2\pi}} R_{nl}(r) e^{i n \phi} \chi_\sigma \quad (7a)$$

with

$$R_{nl}(r) = \sqrt{\frac{2}{c_l}} \sqrt{\frac{(2n+1)!}{(2n+1+|l|)!}} \exp\left(\frac{-r^2}{2a^2}\right) + \left(\frac{r^2}{a^2}\right)^{|l|/2} L_{2n+1}^{|l|} \left(\frac{r^2}{a^2}\right) \chi_\sigma \quad (7b)$$

Where $a = \left(\frac{\hbar}{m^* \Omega}\right)^{1/2}$, $\Omega^2 = \omega_o^2 + \frac{\omega_c^2}{4}$, χ_σ = spinor function and $\omega_c = \frac{eB}{m^*}$ (cyclotron frequency).

Eigen energies for eq. (1) is given as: $E_{nl} = (2n + |l| + 3/2)\hbar\Omega + \frac{\hbar}{2}\omega_c + \sigma(\frac{1}{2}g^*\mu_B B + lam^*\omega_o^2)$, where

$$m = 0, 1, 2, 3, \dots, l = 0, \pm 1, \pm 2, \dots \text{ and } \Omega_o^2 = \omega_o^2 + \frac{\omega_c^2}{4} + \frac{\sigma lam^* \omega_o^2 \omega_c}{\hbar} \quad (8)$$

Also, Ω_o is influenced by Rashba SOI, resulting in an upsurge in the up-spin energy gap with diminution in spin-down energy. SOI also affects the energy term in such a way that at $B = 0$ energy term becomes independent of the magnetic field and helps in uplifting spin degeneracy states.

An analytic expression for THG related to an optical inter-subband transition can be obtained using the density matrix formalism and an iterative procedure [33–35].

The wavefunction $\psi_{nl}(r)$ of the quantum dot with Rashba can be expanded in terms of an orthogonal and complete set of eigenvectors of H_0 . The expansion takes the form [36]:

Hence the Schrödinger equation $H\psi = E\psi$ becomes:

$$(E_{n,l,\sigma}^0 - E)C_{n,l}^\sigma + \sum_{n',l',\sigma'} (H_R)_{nn',ll'}^{\sigma,\sigma'} C_{n',l'}^{\sigma'} = 0 \quad (9a)$$

$$\text{With } \psi_{nl\sigma} = \sum_{n,l,\sigma} C_{nl}^\sigma \varphi_{nl\sigma} \quad (9b)$$

The matrix elements are given by:

$$(H_R)_{nn',ll'}^{\sigma,\sigma'} = \langle \varphi_{n,l}^\sigma | H_R | \varphi_{n',l'}^{\sigma'} \rangle = a \delta_{l,l+1} \sum \left(C_n^{\sigma'} C_n^\sigma \sqrt{n+l+1} - C_{n-1}^{\sigma'} C_n^\sigma \sqrt{n} + C_n^{\sigma'} C_n^\sigma \sqrt{n+l+2} - C_{n-1}^{\sigma'} C_n^\sigma \sqrt{n} \right) \quad (10)$$

The polarized electromagnetic field for an exciting system with frequency ω is given as:

$$E(t) = E e^{i\omega t} + E^* e^{-i\omega t} \quad (11)$$

Also, relationship between the electronic polarization $P(t)$ & polarized electromagnetic field are expressed by:

$$P(t) = \varepsilon_o \chi(\omega) E e^{-i\omega t} + \varepsilon_o \chi(-\omega) E^* e^{-i\omega t} = \frac{1}{V} \text{Tr}(\rho M) \quad (12)$$

Where $\chi_{\omega}^{(1)}$, $\chi_{2\omega}^{(2)}$, $\chi_{\omega}^{(2)}$, and $\chi_{3\omega}^{(3)}$ are representing the susceptibilities such as linear, SHG, OR and THG, respectively. THG relation is given as [17, 37–40]:

$$\chi_{3\omega}^{(3)} = \frac{n_o e^4}{\hbar^3} \frac{M_{01} M_{12} M_{23} M_{30}}{(\omega - \omega_{10} + i\Gamma_o)(2\omega - \omega_{20} + i\Gamma_o)(3\omega - \omega_{30} + i\Gamma_o)} \quad (13)$$

Here, the system's electron density is given as n_o , e is representing electronic charge, and relaxation time is represented by Γ_o . $\omega_{ij} = (E_i - E_j)/\hbar$ shows the frequency for the transition and matrix elements are given by $M_{ij} = |\langle \psi_i | e r | \psi_j \rangle|$ ($i, j = 0, 1, 2, 3$) such as $M_{01} = |\langle \psi_{nl\sigma} | e r | \psi_{n'l'\sigma'} \rangle|$ where $0 = nl\sigma$ and $1 = n'l'\sigma'$.

2. Result and discussion

In this section, the simultaneous effect of the magnetic field, Hydrostatic Pressure, Temperature, confining potential in the presence of SOI on THZ in $\text{In}_x\text{Ga}_{1-x}\text{As}$ semi-parabolic 2-D quantum dot is calculated. For this purpose, the physical parameters that have been used are given as follows: $\Gamma_o = 0.66\text{ps}$ and $m_e^* = 0.041m_o$, where mass of a free electron m_o . σ_s is taken as $5 \times 10^{22} \text{ m}^{-3}$, $\varepsilon_r = 12.53$, and $g = -15$ [41,42].

Considering a QD system in semi parabolic potential having (0 0–1) as ground state and (1 –1 –1) and (1 1–1) as excited states having degenerate intermediate states, when both the magnetic field and spin are zero. This degeneracy can be wrecked by applying a magnetic field ‘B’ and introducing an “ α ”, and; hence, by using these two parameters some manipulation in the energy can be done. Fig. 2 represents a Schematic conduction band energy level diagram (n 1 –1) for InAlAs/InGaAs semi-parabolic quantum dot having four possible routes.

For THG, four possible routes having (0 0–1) ($nl\sigma$) as ground state and their corresponding transition energies with potential confinement at $\hbar \omega = 10 \text{ meV}$, Rashba factor $\alpha = 10 \text{ meV nm}$, Pressure 10 kbar, temperature = 10 K and magnetic field $B = 1 \text{ T}$ is given in Fig. 2. Fig. 3 represents four individual possible paths for transition for the THG coefficients vs incident photon energy with confining potential keeping at $\omega = 10 \text{ meV}$, fixing the Rashba SOI parameter at $\alpha = 10 \text{ meV nm}$, Hydrostatic Pressure at 10 kbar, Temperature at 10 K and applying an external magnetic field, $B = 1 \text{ T}$ to the QD. As it can be observed from Fig. 3 that one, two, three-photon resonances are occurring at different photon energies due to the intermediate ladder states. Alteration in peaks as well as in peak heights can also be observed for different transition energies in each dissimilar path. According to the observation, two-photon resonance peaks are found to be stouter than one & three-photon resonance peaks as peak height corresponding to the strength of the dipole matrix element also increases with a significant increase in the coupling.

The magnetic field is having a significant effect in shifting the peaks as well as changing the magnitude of the peaks which is evident in Fig. 4 where THG coefficient vs incident photon energy has been plotted for

four different values of the magnetic field. As the energy levels are getting affected by the magnetic field hence in Fig. 4a–d, a 2-way shifting of different positions of resonance can be observed. Here, the cyclotron frequency term supports dropping the energy level as well as (–) 1 whereas the Zeeman term helps in boosting the energy. As B is independent of the sign of l i.e., whether l is + or –, it will always help in enhancing $\Omega\sigma$. Same can be observed from the expression of eigen energy given in equation (8). Further, it is also observed from the figure that for 1st two paths, the three-photon resonance is having some

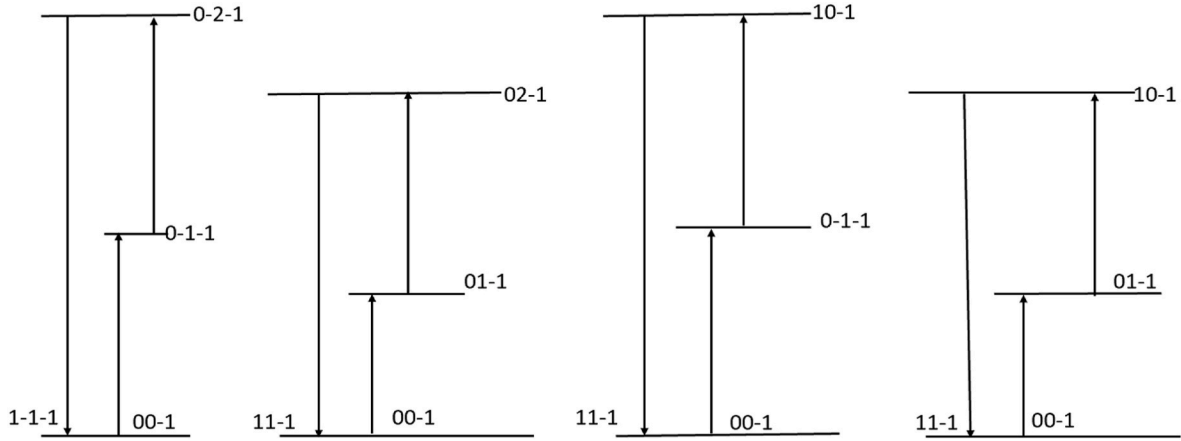


Fig. 2. Schematic conduction band energy level diagram ($n l - \sigma$) for InAlAs/InGaAs quantum dot having four possible routes.

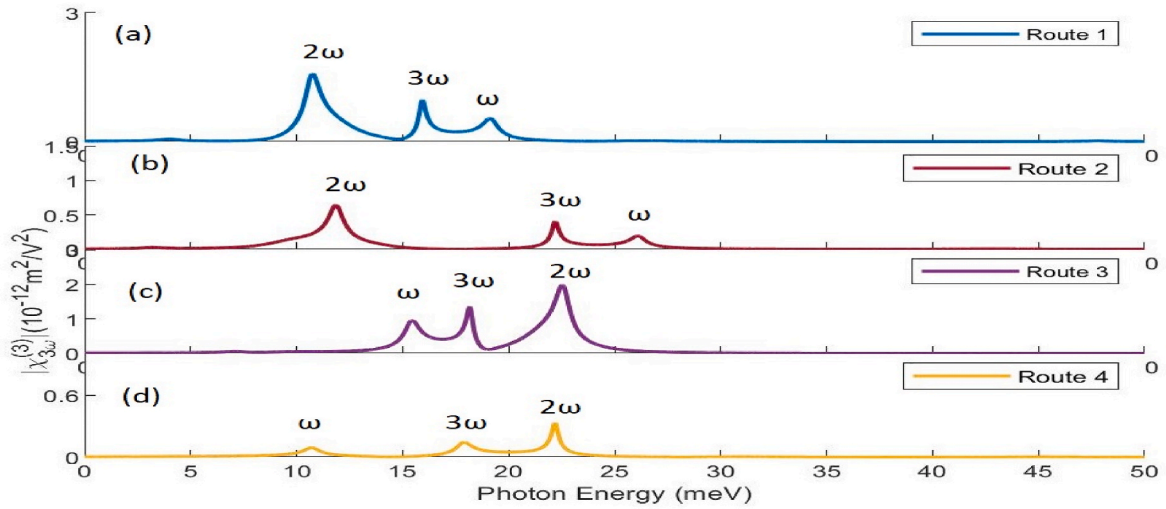


Fig. 3. Coefficient of THG vs photon energy for four possible routes (a) 1st route; (b) 2nd route; (c) 3rd route; (d) 4th route having $\hbar \omega = 10$ meV, $\alpha = 10$ meV nm, $P = 10$ kbar, $T = 10$ K and $B = 1$ T.

significance on the right side of the highest peak while for the 3rd and 4th path, peaks are enhancing but three-photon resonances are not having much significant value due to involvement of -1 states in the resonance position which are independent of three photons resonances and hence shifting towards the lower energy side, whereas the involvement of $+1$ states in the resonance positions helps in shifting the excited states to a state of higher photon energy.

In Fig. 5, THG Coefficient and photon energy for dissimilar values of the QD confinement potential (a) $\hbar \omega = 10$ meV, (b) $\hbar \omega = 15$ meV, (c) $\hbar \omega = 20$ meV, (d) $\hbar \omega = 25$ meV keeping $B = 1$ T, $P = 10$ kbar, $T = 10$ K and $\alpha = 10$ meV nm are plotted. From this figure it is concluded that both the blue/redshifts are observed for the resonant peaks of the THG i.e., some resonant peaks are moving toward lower photon energies exhibiting redshift and some are moving toward higher photon energies displaying blue shift with an increase in the confining potential. Due to the quantum confinement effect, an increased confinement potential roots towards the lesser radius of charge carriers in a QD. Due to weak confinement, the energy separation between the states tends to decrease and hence, exhibits a blue shift in the peaks. It is also observed that the peaks are having unequal spacing with a difference in the number of photon resonance as they are belonging to different energy levels.

Fig. 6 represents a plot between THG coefficient and incident photon energy for different values of α at a confining potential of $\omega = 10$ meV and $B = 1$ T, where α is increasing at a step of 5 meV nm. With an

increase in α from a value of 10 meV nm to 25 meV nm, a slight redshift can be observed in one and two-photon resonance peaks, whereas a shifting towards the higher energy end can be observed in three-photon resonance positions resulting in a blue shift. As α is playing a two-way role while handling the values of energy levels hence red/blue shifts are observed. Further, a decrease in peak height can also be observed with an increase in the value of the Rashba parameter.

Fig. 7(a) is showing the coefficient of THG vs incident photon energy for four diverse values of pressure fixing confining potential $\hbar \omega = 10$ meV, $T = 10$ K, $\alpha = 10$ meV nm and $B = 1$ T. With an increase in Hydrostatic Pressure, the magnitude of THG resonant peaks increases with a slight decrease in peak height as the peaks are shifting towards lower energy. It is observed that the change in the magnitude of the THG resonant peaks is directly correlated to the dipole matrix element term $M_{01}M_{12}M_{23}M_{30}$ in the numerator as well as to the energy interval ω_{10} , ω_{20} and ω_{30} in the denominator. Additionally, as dipole matrix elements are decreasing with an increase in pressure hence the red shift in resonant peaks is observed with an increase in hydrostatic pressure, as shown in Fig. 7(b). This is explained by the fact that the quantum confinement becomes weak with the decrease in the energy interval with a rise in hydrostatic pressure.

Fig. 8(a) is showing a plot of THG and incident photon energy for four diverse values of temperature fixing confining potential $\hbar \omega = 10$ meV, $P = 10$ kbar, $\alpha = 10$ meV nm and $B = 1$ T. Blueshift can be observed

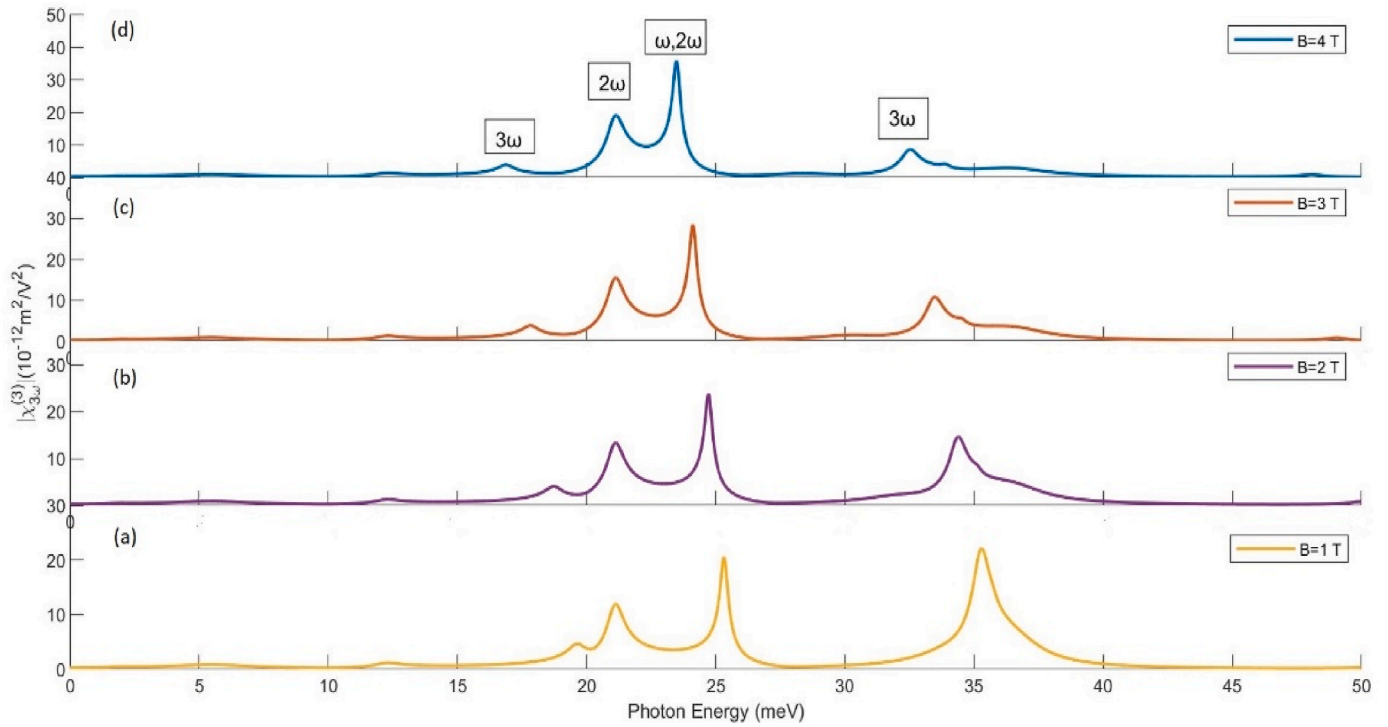


Fig. 4. Coefficient of THG vs incident photon energy at diverse Magnetic field values keeping $\alpha = 10$ meV nm, $P = 10$ kbar, $T = 10$ K and confining potential $\hbar \omega = 10$ meV.

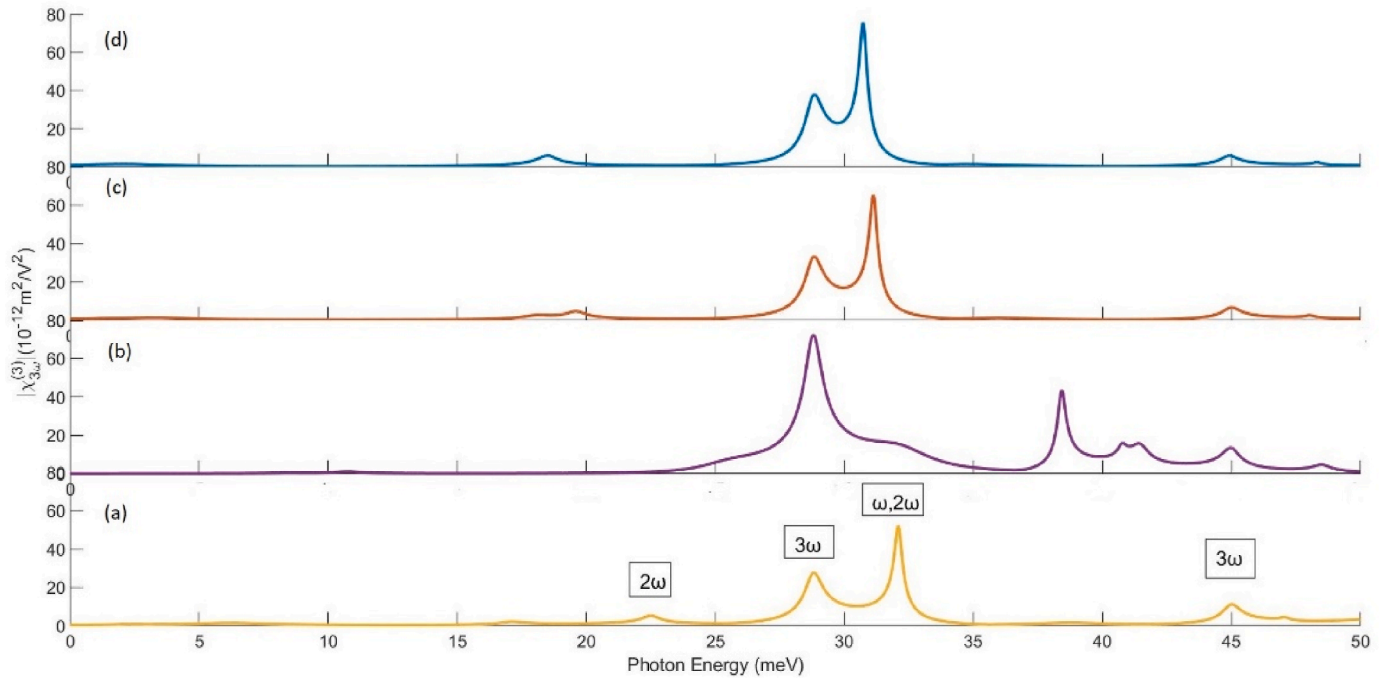


Fig. 5. Coefficient of THG vs incident photon energy for the diverse value of confining potential keeping $B = 1$ T, $P = 10$ kbar, $T = 10$ K and $\alpha = 10$ meV nm.

with an increase in the temperature as the resonant peaks are moving towards higher energy region with a significant increase in the peak height. As one can see from Fig. 8(b) that the dipole matrix element M01 gets enhanced with increase in temperature, consequently the peak height corresponding to $(\omega, 2\omega)$ gets enhance along with its blueshift. The dipole matrix element M01's value increases with temperature as the effective radius changes with an increment in temperature.

However, the peak corresponding to 3ω at higher energy gets suppressed at higher temperature. This can be attributed to the fact that as temperature increases, the sharp resonances get fuzzy due to the thermal energy of the electrons.

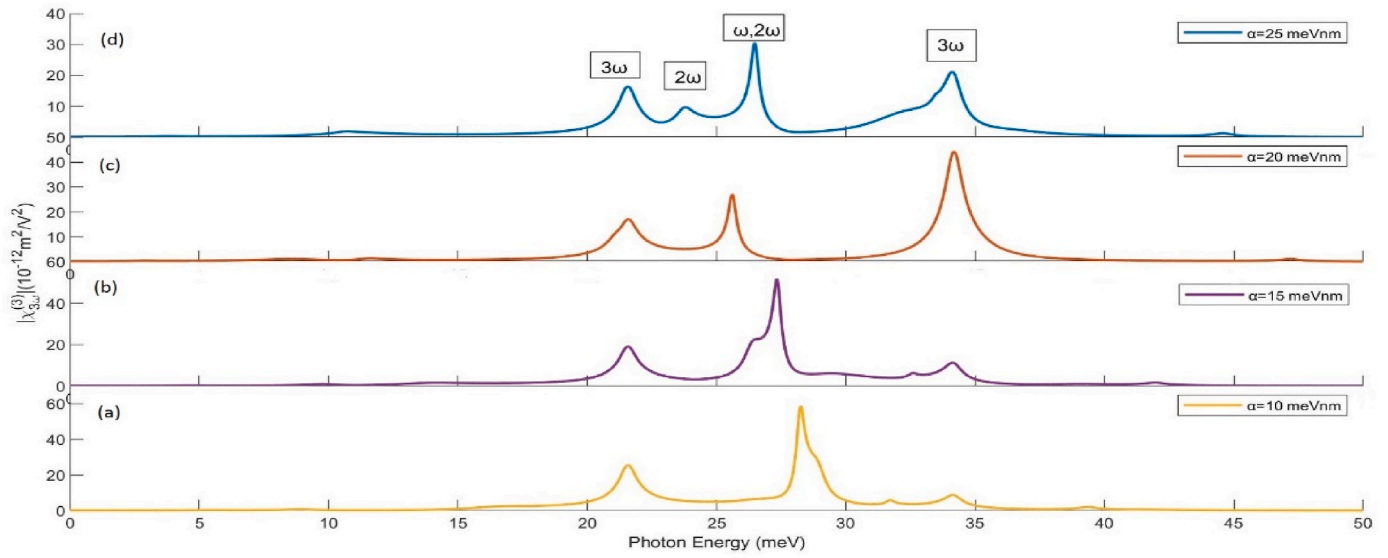
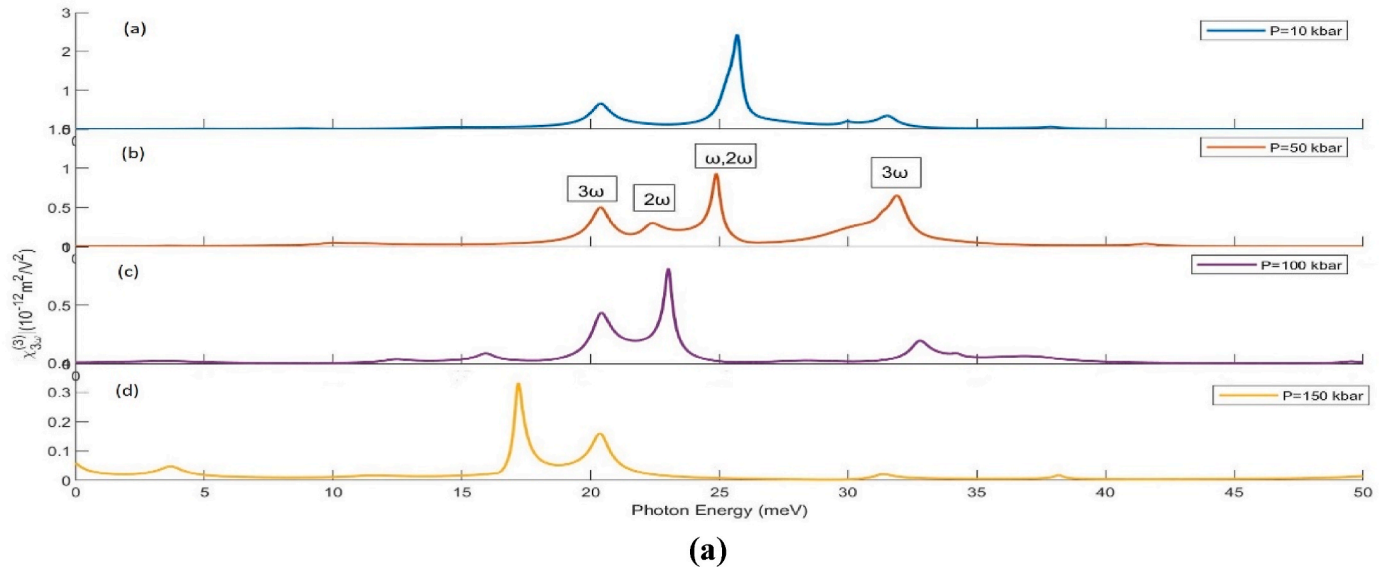
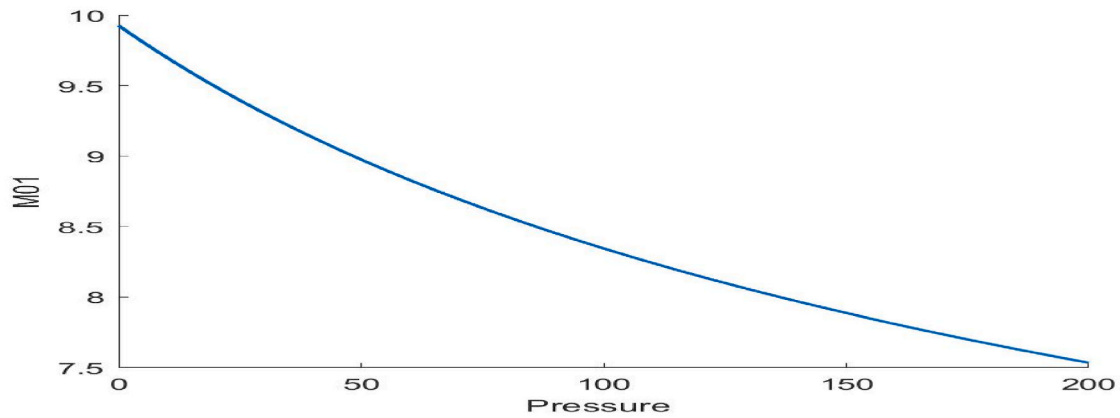


Fig. 6. Coefficient of THG vs incident photon energy for the diverse value of Rashba SOI coupling factor keeping confining potential $\hbar \omega = 10$ meV, $P = 10$ kbar, $T = 10$ K and $B = 1$ T.



(a)



(b)

Fig. 7(a). Coefficient of THG vs incident photon energy for diverse value of pressure fixing confining potential $\hbar \omega = 10$ meV, $T = 10$ K, $\alpha = 10$ meV nm and $B = 1$ T.

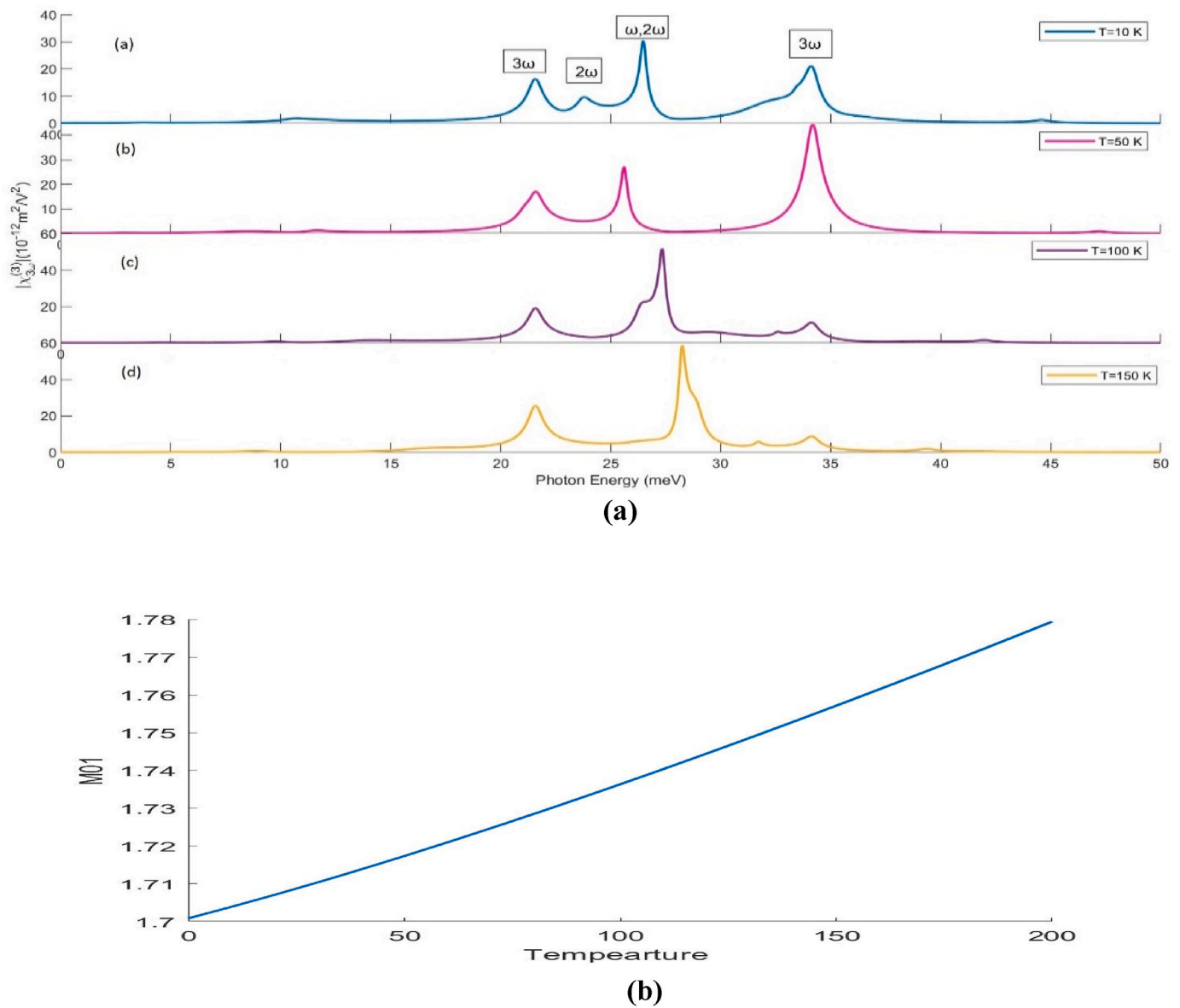


Fig. 8(a). Coefficient of THG vs incident photon energy for diverse value of Temperature fixing confining potential $\hbar\omega = 10 \text{ meV}$, $P = 10 \text{ kbar}$, $\alpha = 10 \text{ meV nm}$ and $B = 1 \text{ T}$.

3. Conclusion

A detailed investigation for the THG coefficients for an $\text{In}_x\text{Ga}_{1-x}\text{As}$ QD in THz laser field with Rashba SOI with the impact of the magnetic field in the vertical direction is carried out in the present study. To carry out the detailed investigation, energy levels with respective wave functions within the effect. mass approx. is being determined using the variational technique. The variation of THG coefficients vs incident photon energy is explored for various external parameters such as temperature, hydrostatic pressure, confining potential, the magnetic field in the presence of Rashba SOI strength. Results are signifying that with an upsurge in the Rashba SOI coefficient, a strong effect on the THG peak positions is observed. It can also be observed that the two-photon resonance peaks are stronger than three-photon resonance peaks due to an increase in the coupling of levels as the peak height corresponds to the strength of the dipole matrix element. The outcomes are displaying that for the detailed engineering of optical devices based on the QD's, consideration of SOI is a must and optical properties of the optoelectronic devices are controllable using the tunable strength parameter.

Declaration of competing interest

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

Data availability

No data was used for the research described in the article.

Acknowledgement

We acknowledge sincere gratitude to Delhi Technological University to appreciate their help and enhance our services and facilities.

References

- [1] S.J. Liang, W. Xie, *Eur. Phys. J. B* 81 (2011) 79.
- [2] H. Hartmann & R. Schuck, *Int. J. Quant. Chem.* 18 (1980) 125.
- [3] I. Zutic, J. Fabian, S. Das Sarma, *Rev. Mod. Phys.* 76 (2004) 323.

- [4] J. Jayabalan, M.P. Singh, K.C. Rustagi, *Phys. Rev. B* 68 (2003), 075319.
- [5] R. Khordad, *Opt. Quant. Electron.* 46 (2014) 283.
- [6] R. Khordad, H. Bahramiyan, *Pramana - J. Phys.* 88 (2017) 50.
- [7] S.A. Wolf, et al., *Science* 294 (2001) 148.
- [8] W. Xie, S. Liang, *Physica B* 406 (2011) 4657.
- [9] R. Khordad, S.K. Khaneghah, M. Masoumi, *Superlattice. Microst.* 47 (2010) 538.
- [10] R. Khordad, *Superlattice. Microst.* 47 (2010) 422.
- [11] G. Rezaei, S. S. Superlattice. *Microst.* 53 (2013) 99.
- [12] B. Gil, P. Lefebvre, P. Boring, *Phys. Rev. B* 44 (1991) 1942.
- [13] R. Khordad, *J. Lumin.* 134 (2013) 201.
- [14] R. Khordad, *Int. J. Mod. Phys. B* 3 (2017), 1750055.
- [15] S. Liang, W. Xie, X. Li, et al., *Superlattice. Microst.* 49 (2011) 623.
- [16] P. Lefebvre, B. Gil, H. Mathieu, *Phys. Rev. B* 35 (1987) 5630.
- [17] R. Khordad, *Opt Commun.* 391 (2017) 121.
- [18] İ. Karabulut, H. Şafak, *Physica B* 82 (2005) 368.
- [19] C.M. Duque, M.E. Mora-Ramos, C.A. Duque, *J Nano part Res* 13 (2011) 6103.
- [20] S. Baskoutas, E. Paspalakis, A.F. Terzis, *Phys. Rev. B* 74 (2006), 153306.
- [21] W. Xie, *J. Lumin.* 131 (5) (2011) 943.
- [22] M. Bass, P.A. Franken, J.F. Ward, G. Weinreich, *Phys. Rev. Lett.* 9 (11) (1962) 446.
- [23] P.A. Franken, A.E. Hill, C.W. Peters, G. Weinreich, *Phys. Rev. Lett.* 7 (4) (1961) 118.
- [24] S. Baskoutas, E. Paspalakis, A.F. Terzis, *Phys. Rev. B* 74 (2006) 15.
- [25] F. Ungan, M.K. Bahar, J.C. Martinez-Orozco, M.E. Mora-Ramos, *Photonics and Nanostructures - Fundamentals and Applications* 41 (2020), 100833.
- [26] Xuechao Li, Chaojin Zhang, *Superlattice. Microst.* 60 (2013) 40.
- [27] Manoj Kumar, Sukirti Gumber, Siddhartha Lahon, Pradip Kumar Jha, Man Mohan, *Eur. Phys. J. B* 87 (2014) 71.
- [28] L. Jacak, P. Hawrylak, A. Wojs, *Quantum Dots*, Springer, Berlin, 1997.
- [29] R. Khordad, B. Vaseghi, *Chin. J. Phys.* 59 (2019) 473.
- [30] R. Khordad, H. Bahramiyan, *Commun. Theor. Phys.* 62 (2014) 283.
- [31] Y. Khoshbakht, R. Khordad, Rastegar Sedehi, *J. Low Temp. Phys.* 202 (2021) 59.
- [32] S. Dahiya, S. Lahon, R. Sharma, *Physica E* 118 (2020), 113918.
- [33] R. Khordad, *Opt. Quant. Electron.* 46 (2014) 283.
- [34] S. Dahiya, M. Verma, S. Lahon, R. Sharma, *Journal of Atomic, Molecular, Condensed Matter and Nano Physics* 5 (1) (2018) 41–53.
- [35] R.F. Kopf, M.H. Herman, Lamont Schnoes, M. Perley, A.P. Livescu, G. Ohring, *J. Appl. Phys.* 71 (1992) 5004–5011.
- [36] R. Chaurasiya, S. Dahiya, R. Sharma, *IEEE International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology (5NANO)*, 2022, pp. 1–3.
- [37] Zhi-Hai Zhang, Kang-Xian Guo, Bin Chen, Rui-Zhen Wang, Min-Wu Kang, *Superlattice. Microst.* 46 (2009) 672.
- [38] L. Zhang, H.J. Xie, *Phys. E* 22 (2004) 791.
- [39] G.H. Wang, *Phys. Rev. B* 72 (2005), 155329.
- [40] Shuai Shao, Kang Xian Guo, Zhi Hai Zhang, Li Ning, Chao Pen, *Solid State Commun.* 151 (2011) 589.
- [41] Junsaku Nitta, Tatsushi Akazaki, Hideaki Takayanagi, Takatomo Enoki, *Phys. Rev. Lett.* 78 (1997) 1335.
- [42] M. Solaimani, L. Lavaei, S.M.A. Aleomraninejad, 1989, *J. Opt. Soc. Am. B* 9 (2017) 34.

Toeplitz determinants on bounded starlike circular domain in \mathbb{C}^n

Surya Giri¹ and S. Sivaprasad Kumar*

Abstract

In this paper, we derive the sharp bounds of Toeplitz determinants for a class of holomorphic mappings on the bounded starlike circular domain Ω in \mathbb{C}^n , which extend certain known bounds for various subclasses of normalized analytic univalent functions in the unit disk to higher dimensions.

Keywords: Holomorphic mapping; Toeplitz determinant; Coefficient inequality; bounded starlike circular domain.

AMS Subject Classification: 32H02, 30C45.

1 Introduction

Let \mathcal{S} be the class of normalized univalent holomorphic functions on the unit disk \mathbb{U} in \mathbb{C} of the form

$$g(z) = z + \sum_{n=2}^{\infty} b_n z^n.$$

Let \mathcal{S}^* , $\mathcal{S}^*(\alpha)$ and $\mathcal{SS}^*(\beta)$ represent the subclasses of \mathcal{S} containing the starlike functions, starlike functions of order α ($0 \leq \alpha < 1$) and strongly starlike functions of order β ($0 < \beta \leq 1$), respectively (see [5]). Ali et al. [2] obtained the bounds of Toeplitz determinants formed over the coefficients of $g \in \mathcal{S}$. For $g(z) = z + \sum_{n=2}^{\infty} a_n z^n$, the Toeplitz matrix is given by

$$T_{m,n}(g) = \begin{bmatrix} b_n & b_{n+1} & \cdots & b_{n+m-1} \\ b_{n+1} & b_n & \cdots & b_{n+m-2} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n+m-1} & b_{n+m-2} & \cdots & b_n \end{bmatrix}. \quad (1.1)$$

Thus, the second order Toeplitz determinant is

$$\det T_{2,2}(g) = b_2^2 - b_3^2 \quad (1.2)$$

and the third order Toeplitz determinant is given by

$$\det T_{3,1}(g) = 2b_2^2 b_3 - 2b_2^2 - b_3^2 + 1. \quad (1.3)$$

According to Ye and Lim [29], any $n \times n$ matrix over \mathbb{C} generically can be written as the product of some Toeplitz matrices or Hankel matrices. Toeplitz matrices and Toeplitz determinants have many applications in pure as well as in applied mathematics [24]. For more details of applications in various areas of mathematics, we refer [29] and the references cited therein.

For the class of normalized starlike functions, Ali et al. [2] obtained the following result.

Theorem A. [2] *If $g \in \mathcal{S}^*$, then the following sharp bounds hold:*

$$|\det T_{2,2}(g)| \leq 13 \text{ and } |\det T_{3,1}(g)| \leq 24.$$

Ahuja et al. [1] determined the following estimates for the class of normalized starlike functions of order α .

Theorem B. [1] *If $g \in \mathcal{S}^*(\alpha)$, then*

$$|\det T_{2,2}(g)| \leq (1 - \alpha)^2(4\alpha^2 - 12\alpha + 13)$$

and for $\alpha \in [0, 2/3]$,

$$|\det T_{3,1}(g)| \leq 12\alpha^4 - 52\alpha^3 + 91\alpha^2 - 74\alpha + 24.$$

All these estimates are sharp.

The following result directly follows from [1, Theorem 1] and [1, Theorem 3] for the class $\mathcal{SS}^*(\beta)$.

Theorem C. *If $g \in \mathcal{SS}^*(\beta)$, then for $\beta \in [1/3, 1]$, the following sharp inequalities hold:*

$$|\det T_{2,2}(g)| \leq 9\beta^4 + 4\beta^2 \quad \text{and} \quad |\det T_{3,1}(g)| \leq 15\beta^4 + 8\beta^2 + 1.$$

Cartan [8] stated that the Bieberbach conjecture for the class \mathcal{S} does not hold in case of several complex variables. There are various counterexamples, which show that many results in the Geometric function theory of one complex variable are not applicable for several complex variables (see [8]). Many researchers have focused on generalizing the coefficients inequalities for the subclasses of \mathcal{S} in higher dimensions. Recently, Giri and Kumar [4] generalized the above results on the unit ball in a complex Banach space and on the unit polydisc in \mathbb{C}^n . On the bounded starlike circular domain $\Omega \subset \mathbb{C}^n$, Liu and Xu [21] (see also [28]) solved the Fekete-Szegő problem for a subclass of starlike mappings of order α . Xu [25] generalized the work in [28] to a subclass of holomorphic mappings on the same domain Ω . Contrary to the Fekete-Szegő problems, very few results are known for the inequalities of homogeneous expansions for subclasses of biholomorphic mappings in several complex variables. Results related to the bounds for the coefficients of various subclasses of holomorphic mappings in higher dimensions were obtained by Bracci et al. [3], Graham et al. [6], Graham et al. [7], Graham et al. [9], Hamada and Honda [10], Hamada et al. [11], Kohr [15], Liu and Liu [17, 18], Xu and Liu [26, 27].

In this paper, we find the bounds of second and third order Toeplitz determinants for a class of holomorphic mappings on the bounded starlike circular domain in \mathbb{C}^n , which give an extension of Theorem A, Theorem B and Theorem C to higher dimensions.

2 Preliminaries

By \mathbb{C}^n , we denote the space of n complex variables $z = (z_1, z_2, \dots, z_n)'$ with the Euclidean inner product $\langle z, w \rangle = \sum_{i=1}^n z_i \bar{w}_i$ and the norm $\|z\| = \langle z, z \rangle^{1/2}$. Let \mathbb{U}^n be the Euclidean unit ball in \mathbb{C}^n and $\Omega \subset \mathbb{C}^n$ be a bounded starlike circular domain with $0 \in \Omega$ and its Minkowski functional $\rho(z) \in C^1$ in $\mathbb{C}^n \setminus \{0\}$. Let $\mathcal{H}(\Omega)$ be the set of all holomorphic mappings from Ω into \mathbb{C}^n . If $g \in \mathcal{H}(\Omega)$, then

$$g(w) = \sum_{k=0}^{\infty} \frac{1}{k!} D^k g(z) ((w - z)^k)$$

for all w in some neighborhood of z , where $D^k g(z)$ is the k th Fréchet derivative of g at z . A function $g \in \mathcal{H}(\Omega)$ is said to be biholomorphic if $g(\Omega)$ is a domain in \mathbb{C}^n and inverse of g exists, which is holomorphic on $g(\Omega)$. Let $J_g(z)$ be the Jacobian matrix of g and $\det J_g(z)$ be the Jacobian determinant of g at $z \in \Omega$. A mapping $g \in \mathcal{H}(\Omega)$ is said to be locally biholomorphic if $\det J_g(z) \neq 0$ for all $z \in \Omega$. In higher dimensions, $g \in \mathcal{H}(\Omega)$ is said to be normalized if $g(0) = 0$ and $J_g(0) = I$, where I is the identity matrix. Let $\mathcal{S}^*(\Omega)$ denotes the class of starlike mappings on Ω . When $n = 1$, $\Omega = \mathbb{U}$, the class $\mathcal{S}^*(\mathbb{U})$ is denoted by \mathcal{S}^* .

On a bounded circular domain $\Omega \subset \mathbb{C}^n$, the first and the m th Fréchet derivative of a holomorphic mapping $g \in \mathcal{H}(\Omega)$ are written by $Dg(z)$ and $D^m g(z)(a^{m-1}, \cdot)$, respectively. The matrix representations are

$$Dg(z) = \left(\frac{\partial g_j}{\partial z_k} \right)_{1 \leq j, k \leq n},$$

$$D^m g(z)(a^{m-1}, \cdot) = \left(\sum_{p_1, p_2, \dots, p_{m-1}=1}^n \frac{\partial^m g_j(z)}{\partial z_k \partial z_{p_1} \cdots \partial z_{p_{m-1}}} a_{p_1} \cdots a_{p_{m-1}} \right)_{1 \leq j, k \leq n},$$

where $g(z) = (g_1(z), g_2(z), \dots, g_n(z))'$, $a = (a_1, a_2, \dots, a_n)' \in \mathbb{C}^n$.

According to Liu and Lu [19] (see also [14]), we have the following definition.

Definition 2.1. Let Ω be a bounded starlike circular domain in \mathbb{C}^n with $0 \in \Omega$ and its Minkowski functional $\rho \in C^1$ in $\mathbb{C}^n \setminus \{0\}$. A normalized locally biholomorphic mapping $g : \Omega \rightarrow \mathbb{C}^n$ is said to be starlike of order α ($0 \leq \alpha < 1$) if

$$\left| \frac{2}{\rho(z)} \frac{\partial \rho}{\partial z} J_g^{-1}(z) g(z) - \frac{1}{2\alpha} \right| < \frac{1}{2\alpha}, \quad \forall z \in \Omega \setminus \{0\}.$$

Equivalently, the above equation can be written as

$$\operatorname{Re} \left\{ \frac{\rho(z)}{2 \frac{\partial \rho(z)}{\partial z} J_g^{-1}(z) g(z)} \right\} > \alpha, \quad \forall z \in \Omega \setminus \{0\}.$$

Clearly, when $\Omega = \mathbb{U}^n$, the aforementioned inequality is equivalent to

$$\operatorname{Re} \left\{ \frac{\|z\|^2}{\langle J_g^{-1}(z) g(z), z \rangle} \right\} > \alpha, \quad \forall z \in \mathbb{U}^n \setminus \{0\}.$$

In case of $n = 1$, $\Omega = \mathbb{U}$ and the above relation is equivalent to

$$\operatorname{Re} \frac{zg'(z)}{g(z)} > 0, \quad z \in \mathbb{U}.$$

We denote by $\mathcal{S}_\alpha^*(\Omega)$ the set of all starlike mappings of order α on Ω .

Definition 2.2. [8] (also see [16, 12]) Let Ω be a bounded starlike circular domain in \mathbb{C}^n with $0 \in \Omega$ and its Minkowski functional $\rho \in C^1$ in $\mathbb{C}^n \setminus \{0\}$. A normalized locally biholomorphic mapping $g : \Omega \rightarrow \mathbb{C}^n$ is said to be strongly starlike of order β ($0 < \beta \leq 1$) if

$$\left| \arg \frac{2}{\rho(z)} \frac{\partial \rho}{\partial z} J_g^{-1}(z) g(z) \right| < \frac{\pi}{2} \beta, \quad \forall z \in \Omega \setminus \{0\}.$$

Clearly, when $\Omega = \mathbb{U}^n$, the aforementioned inequality is equivalent to

$$|\arg \langle J_g^{-1}(z) g(z), z \rangle| < \frac{\pi}{2} \beta, \quad \forall z \in \mathbb{U}^n \setminus \{0\}.$$

In case of $n = 1$, $\Omega = \mathbb{U}$ and the above relation is equivalent to

$$\left| \arg \frac{zg'(z)}{g(z)} \right| < \frac{\pi}{2} \beta, \quad \forall z \in \mathbb{U}.$$

We denote by $\mathcal{SS}_\beta^*(\Omega)$ the set of all strongly starlike mappings of order β on Ω .

Next, we recall the class \mathcal{M} , which plays a fundamental role in the study of Loewner chains and Loewner differential equation in several complex variables (see [8, 22]).

$$\mathcal{M} = \left\{ p \in \mathcal{H}(\Omega) : p(0) = 0, J_p(0) = I, \operatorname{Re} \frac{\partial \rho}{\partial z} p(z) > 0, z \in \Omega \setminus \{0\} \right\},$$

where $\partial \rho(z)/\partial z = (\partial \rho(z)/\partial z_1, \partial \rho(z)/\partial z_2, \dots, \partial \rho(z)/\partial z_n)$.

Kohr [15] introduced the class \mathcal{M}_Φ on \mathbb{U}^n , which is studied by Graham et al. [7] (see also [6]), where $\Phi : \mathbb{U} \rightarrow \mathbb{C}$ is a biholomorphic function such that $\Phi(0) = 1$ and $\operatorname{Re} \Phi(z) > 0$ on \mathbb{U} . Recently, Xu et al. [28] considered the class \mathcal{M}_Φ on $\Omega \subset \mathbb{C}^n$. Here, we add some more conditions on Φ and define the following subsets of \mathcal{M} .

Assumption 2.3. Let $\Phi : \mathbb{U} \rightarrow \mathbb{C}$ be a biholomorphic function such that $\Phi(0) = 1$, $\Phi'(0) > 0$, $\Phi''(0) \in \mathbb{R}$ and $\operatorname{Re} \Phi(z) > 0$ on \mathbb{U} .

Obviously, there are many functions which satisfy this assumption. Let

$$\mathcal{M}_\Phi = \left\{ p \in \mathcal{H}(\Omega) : p(0) = 0, J_p(0) = I, \frac{\rho(z)}{2 \frac{\partial \rho}{\partial z} p(z)} \in \Phi(\mathbb{U}), z \in \Omega \setminus \{0\} \right\}.$$

The class \mathcal{M}_Φ coincides with \mathcal{M} for $\Phi(z) = (1+z)/(1-z)$, $z \in \mathbb{U}$. Also, if $\Omega = \mathbb{U}^n$, then

$$\mathcal{M}_\Phi = \left\{ p \in \mathcal{H}(\mathbb{U}^n) : p(0) = 0, J_p(0) = I, \frac{\|z\|^2}{\langle p(z), z \rangle} \in \Phi(\mathbb{U}), z \in \mathbb{U}^n \setminus \{0\} \right\}.$$

Remark 2.1. Let $g \in \mathcal{H}(\mathbb{U})$ be a normalized locally biholomorphic function. If $J_g^{-1}(z)g(z) \in \mathcal{M}_\Phi$, then for different choices of Φ , we obtain different important classes of $\mathcal{S}(\Omega)$. For instance, if we take $\Phi(z) = (1+z)/(1-z)$, $\Phi(z) = (1+(1-2\alpha)z)/(1-z)$ and $\Phi(z) = ((1+z)/(1-z))^\beta$ (where the branch point is chosen such that $((1+z)/(1-z))^\beta = 1$ at $z = 0$), then we easily obtain $g \in \mathcal{S}^*(\Omega)$, $g \in \mathcal{S}_\alpha^*(\Omega)$ and $g \in \mathcal{SS}_\beta^*(\Omega)$, respectively.

The following lemma helps us to prove the main results.

Lemma 2.1. [20] $\Omega \subset \mathbb{C}^n$ is a bounded starlike circular domain if and only if there exists a unique real continuous functions $\rho : \mathbb{C}^n \rightarrow \mathbb{R}$, called the Minkowski functional of Ω , such that

- (i) $\rho(z) \geq 0$, $z \in \mathbb{C}^n$; $\rho(z) = 0 \Leftrightarrow z = 0$;
- (ii) $\rho(tz) = |t|\rho(z)$, $t \in \mathbb{C}$, $z \in \mathbb{C}^n$;
- (iii) $\Omega = \{z \in \mathbb{C}^n : \rho(z) < 1\}$.

Furthermore, if $\rho(z) \in C^1$ in $\mathbb{C}^n \setminus \{0\}$, then the function $\rho(z)$ has the following properties.

$$\begin{aligned} 2 \frac{\partial \rho(z)}{\partial z} z &= \rho(z), \quad z \in \mathbb{C}^n, \\ 2 \frac{\partial \rho(z_0)}{\partial z} z_0 &= 1, \quad z_0 \in \partial\Omega, \\ \frac{\partial \rho(\lambda z)}{\partial z} &= \frac{\partial \rho(z)}{\partial z}, \quad \lambda \in (0, \infty), \\ \frac{\partial \rho(e^{i\theta} z)}{\partial z} &= e^{-i\theta} \frac{\partial \rho(z)}{\partial z}, \quad \theta \in \mathbb{R}. \end{aligned} \tag{2.1}$$

3 Main Results

The sharp bounds of second and third order Toeplitz determinants for a class of holomorphic mappings on Ω are derived in this section. Later, applications of these results for other interesting subclasses of $\mathcal{S}(\Omega)$ are given.

Theorem 3.1. Let $g \in \mathcal{H}(\Omega, \mathbb{C})$ with $g(0) = 1$ and $G(z) = zg(z)$. If $J_G^{-1}(z)G(z) \in \mathcal{M}_\Phi$ such that Φ satisfies

$$|\Phi''(0) + 2(\Phi'(0))^2| \geq 2\Phi'(0) > 0,$$

then

$$\left| \left(2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right)^2 - \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \leq (\Phi'(0))^2 + \frac{(\Phi'(0))^2}{4} \left(\frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} + \Phi'(0) \right)^2.$$

The bound is sharp.

Proof. Since $J_G^{-1}(z)$ exists, therefore $g(z) \neq 0$, $z \in \Omega$. For fix $z \in \Omega \setminus \{0\}$, let us denote $z_0 = \frac{z}{\rho(z)}$ and define $h : \mathbb{U} \rightarrow \mathbb{C}$ such that

$$h(\zeta) = \begin{cases} \frac{\zeta}{2 \frac{\partial \rho(z_0)}{\partial z} J_G^{-1}(\zeta z_0) G(\zeta z_0)}, & \zeta \neq 0, \\ 1, & \zeta = 0. \end{cases}$$

Using the property $2 \frac{\partial \rho(z_0)}{\partial z} z_0 = 1$ for $z_0 \in \partial\Omega$ of Minkowski functional, we obtain $h \in \mathcal{H}(\mathbb{U})$ and since $J_G^{-1}(z)G(z) \in \mathcal{M}_\Phi$, therefore

$$\begin{aligned} h(\zeta) &= \frac{\zeta}{2 \frac{\partial \rho(z_0)}{\partial z} J_G^{-1}(\zeta z_0) G(\zeta z_0)} \\ &= \frac{\rho(\zeta z_0)}{2 \frac{\partial \rho(\zeta z_0)}{\partial z} J_G^{-1}(\zeta z_0) G(\zeta z_0)} \in \Phi(\mathbb{U}), \quad \zeta \in \mathbb{U} \setminus \{0\}. \end{aligned}$$

Applying the same technique as in [23] (also see [8, Theorem 7.1.14]), we obtain

$$J_G^{-1}(z) = \frac{1}{g(z)} \left(I - \frac{\frac{z J_g(z)}{g(z)}}{1 + \frac{J_g(z)z}{g(z)}} \right).$$

Now, using $G(z) = zg(z)$, we have

$$J_G^{-1}(z)G(z) = \frac{zg(z)}{g(z) + J_g(z)z}, \quad z \in \Omega \setminus \{0\},$$

which together with (2.1) gives

$$\frac{\rho(z)}{2 \frac{\partial \rho(z)}{\partial z} J_G^{-1}(z)G(z)} = 1 + \frac{J_g(z)z}{g(z)}, \quad z \in \Omega \setminus \{0\}.$$

In view of the above equation, we obtain

$$h(\zeta) = \frac{\rho(\zeta z_0)}{2 \frac{\partial \rho(\zeta z_0)}{\partial z} J_G^{-1}(\zeta z_0) G(\zeta z_0)} = 1 + \frac{J_g(\zeta z_0)\zeta z_0}{g(\zeta z_0)},$$

which immediately yields

$$h(\zeta)g(\zeta z_0) = g(\zeta z_0) + J_g(\zeta z_0)\zeta z_0.$$

Based on the Taylor series expansions in ζ , the above equation gives

$$\begin{aligned} &\left(1 + h'(0)\zeta + \frac{h''(0)}{2!}\zeta^2 + \dots \right) \left(1 + J_g(0)(z_0)\zeta + \frac{D^2 g(0)(z_0^2)}{2!}\zeta^2 + \dots \right) \\ &= \left(1 + J_g(0)(z_0)\zeta + \frac{D^2 g(0)(z_0^2)}{2!}\zeta^2 + \dots \right) \left(J_g(0)(z_0)\zeta + D^2 g(0)(z_0^2)\zeta^2 + \dots \right). \end{aligned}$$

By the comparison of homogeneous expansions, we get

$$h'(0) = J_g(0)(z_0).$$

Further, using $z_0 = \frac{z}{\rho(z)}$ in the above relation, we have

$$h'(0)\rho(z) = J_g(0)(z). \tag{3.1}$$

Since, we also have $G(z) = zg(z)$, therefore

$$\frac{D^2 G(0)z^2}{2!} = J_g(0)(z)z,$$

which together with (2.1) leads to

$$2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0) z^2}{2!} = J_g(0)(z) \rho(z). \quad (3.2)$$

Thus, from (3.1) and (3.2), we obtain

$$2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0) z^2}{2! \rho^2(z)} = h'(0).$$

Since $h \prec \Phi$, therefore $|h'(0)| \leq |\Phi'(0)|$, using this fact, we get

$$\left| 2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0) z^2}{2! \rho^2(z)} \right| \leq |\Phi'(0)|. \quad (3.3)$$

For $\lambda \in \mathbb{C}$, Xu et al. [28, Theorem 1] proved that

$$\left\{ \begin{aligned} & \left| 2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} - \lambda \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \\ & \leq \frac{|\Phi'(0)|}{2} \max \left\{ 1, \left| \frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} + (1 - 2\lambda) \Phi'(0) \right| \right\}, \quad z \in \Omega \setminus \{0\}. \end{aligned} \right\} \quad (3.4)$$

Thus, when $|\Phi''(0) + 2(\Phi'(0))^2| \geq 2\Phi'(0)$, the equation (3.4) readily yields

$$\left| 2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right| \leq \frac{\Phi'(0)}{2} \left| \frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} + \Phi'(0) \right|. \quad (3.5)$$

Using the bounds given in (3.3) and (3.5), together with the following inequality

$$\begin{aligned} & \left| \left(2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right)^2 - \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \\ & \leq \left| 2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right|^2 + \left| 2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right|^2, \end{aligned}$$

we find the required bound.

To see the sharpness of the bound consider the function

$$G(z) = z \exp \int_0^{\frac{z_1}{r}} \frac{(\Phi(it) - 1)}{t} dt, \quad z \in \Omega, \quad (3.6)$$

where $r = \sup\{|z_1| : z = (z_1, z_2, \dots, z_n)' \in \Omega\}$. It can be easily showed that $J_G^{-1}(z)G(z) \in \mathcal{M}_\Phi$ and

$$\frac{D^2 G(0)(z^2)}{2!} = i\Phi'(0)\left(\frac{z_1}{r}\right)z \quad \text{and} \quad \frac{D^3 G(0)(z^3)}{3!} = -\frac{1}{2} \left(\frac{\Phi''(0)}{2} + (\Phi'(0))^2 \right) \left(\frac{z_1}{r}\right)^2 z.$$

By applying (2.1) in the above relations, we get

$$2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2!} = i\Phi'(0)\left(\frac{z_1}{r}\right)\rho(z)$$

and

$$2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3!} \rho(z) = -\frac{1}{2} \left(\frac{\Phi''(0)}{2} + (\Phi'(0))^2 \right) \left(\frac{z_1}{r}\right)^2 \rho^2(z).$$

Setting $z = Ru$ ($0 < R < 1$), where $u = (u_1, u_2, \dots, u_n)' \in \partial\Omega$ and $u_1 = r$, we obtain

$$2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} = i\Phi'(0) \quad (3.7)$$

and

$$2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} = -\frac{1}{2} \left(\frac{\Phi''(0)}{2} + (\Phi'(0))^2 \right). \quad (3.8)$$

Thus, from (3.7) and (3.8), we have

$$\begin{aligned} \left| \left(2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right)^2 - \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \\ \leq \frac{(\Phi'(0))^2}{4} \left(\frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} + \Phi'(0) \right)^2 + (\Phi'(0))^2, \end{aligned}$$

which shows the sharpness of the bound and completes the proof.

Theorem 3.2. *Let $g \in \mathcal{H}(\Omega, \mathbb{C})$ with $g(0) = 1$ and $G(z) = zg(z)$. If $J_G^{-1}(z)G(z) \in \mathcal{M}_\Phi$ such that Φ satisfies*

$$2\Phi'(0) - 2(\Phi'(0))^2 \leq \Phi''(0) \leq 6(\Phi'(0))^2 - 2\Phi'(0),$$

then

$$\begin{aligned} |2b_2^2 b_3 - b_3^2 - 2b_2^2 + 1| \\ \leq 1 + 2(\Phi'(0))^2 + \frac{(\Phi'(0))^2}{4} \left(3\Phi'(0) - \frac{\Phi''(0)}{2\Phi'(0)} \right) \left(\frac{\Phi''(0)}{2\Phi'(0)} + \Phi'(0) \right), \end{aligned}$$

where

$$b_3 = 2 \frac{\partial \rho}{\partial z} \frac{2D^3 G(0)(z^3)}{3! \rho^3(z)} \quad \text{and} \quad b_2 = 2 \frac{\partial \rho}{\partial z} \frac{2D^2 G(0)(z^2)}{2! \rho^2(z)}. \quad (3.9)$$

The bound is sharp.

Proof. Since $2\Phi'(0) < \Phi''(0) + 2(\Phi'(0))^2$, the inequality (3.4) gives

$$\left| 2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right| \leq \frac{\Phi'(0)}{2} \left(\frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} + \Phi'(0) \right). \quad (3.10)$$

Also, since $2\Phi'(0) + \Phi''(0) \leq 6(\Phi'(0))^2$, the inequality (3.4) for $\lambda = 2$ gives

$$\left| 2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} - 2 \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \leq \frac{\Phi'(0)}{2} \left(3\Phi'(0) - \frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} \right). \quad (3.11)$$

Using the estimates given in (3.3) and (3.10), and the bound given by (3.11) in the following inequality

$$|2b_2^2 b_3 - b_3^2 - 2b_2^2 + 1| \leq 1 + 2|b_2|^2 + |b_3||b_3 - 2b_2^2|$$

the required bound is established.

The result is sharp for the function $G(z)$ given by (3.6). As for this function, we have $b_2 = i\Phi'(0)$ and $b_3 = -(\Phi''(0) + 2(\Phi'(0))^2)/4$ from (3.7) and (3.8), respectively. Therefore

$$1 - b_3(b_3 - 2b_2^2) - 2b_2^2 = 1 + 2(\Phi'(0))^2 + \frac{(\Phi'(0))^2}{4} \left(3\Phi'(0) - \frac{\Phi''(0)}{2\Phi'(0)} \right) \left(\frac{\Phi''(0)}{2\Phi'(0)} + \Phi'(0) \right),$$

which proves the sharpness of the bound.

In case of $\Omega = \mathbb{U}^n$, Theorem 3.1 and Theorem 3.2 directly give the following results, which we state here without proof.

Theorem 3.3. Let $g \in \mathcal{H}(\mathbb{U}^n, \mathbb{C})$ with $g(0) = 1$ and $G(z) = zg(z)$. If $J_G^{-1}(z)G(z) \in \mathcal{M}_\Phi$ such that Φ satisfies

$$|\Phi''(0) + 2(\Phi'(0))^2| \geq 2\Phi'(0) > 0,$$

then

$$\begin{aligned} & \left| \left(\frac{1}{\|z\|^4} \frac{D^3 G(0)(z^3)}{3!} \bar{z} \right)^2 - \left(\frac{1}{\|z\|^3} \frac{D^2 G(0)(z^2)}{2!} \bar{z} \right)^2 \right| \\ & \leq (\Phi'(0))^2 + \frac{(\Phi'(0))^2}{4} \left(\frac{1}{2} \frac{\Phi''(0)}{\Phi'(0)} + \Phi'(0) \right)^2. \end{aligned}$$

The bound is sharp.

Theorem 3.4. Let $g \in \mathcal{H}(\mathbb{U}^n, \mathbb{C})$ with $g(0) = 1$ and $G(z) = zg(z)$. If $J_G^{-1}(z)G(z) \in \mathcal{M}_\Phi$ such that Φ satisfies

$$2\Phi'(0) - 2(\Phi'(0))^2 \leq \Phi''(0) \leq 6(\Phi'(0))^2 - 2\Phi'(0),$$

then

$$\begin{aligned} & |2d_2^2 d_3 - d_3^2 - 2d_2^2 + 1| \\ & \leq \frac{(\Phi'(0))^2}{4} \left(3\Phi'(0) - \frac{\Phi''(0)}{2\Phi'(0)} \right) \left(\frac{\Phi''(0)}{2\Phi'(0)} + \Phi'(0) \right) + 2(\Phi'(0))^2 + 1, \end{aligned}$$

where

$$d_3 = \frac{1}{\|z\|^4} \frac{D^3 G(0)(z^3)}{3!} \bar{z} \text{ and } d_2 = \frac{1}{\|z\|^3} \frac{D^2 G(0)(z^2)}{2!} \bar{z}. \quad (3.12)$$

The bound is sharp.

In sight of remark 2.1, various choices of Φ in Theorem 3.1 to Theorem 3.4 lead to the following results for different subclasses of $\mathcal{S}(\Omega)$.

Corollary 3.5. If $g : \Omega \rightarrow \mathbb{C}$ and $G(z) = zg(z) \in \mathcal{S}^*(\Omega)$, then

$$\left| \left(2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right)^2 - \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \leq 13$$

and

$$|2b_2^2 b_3 - b_3^2 - 2b_2^2 + 1| \leq 24,$$

where b_2 and b_3 are given by (3.9). All these estimations are sharp.

Corollary 3.6. If $g : \Omega \rightarrow \mathbb{C}$ and $G(z) = zg(z) \in \mathcal{S}_\alpha^*(\Omega)$, then

$$\left| \left(2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right)^2 - \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \leq (1 - \alpha)^2 (4\alpha^2 - 12\alpha + 13)$$

and for $\alpha \in [0, 2/3]$,

$$|2b_2^2 b_3 - b_3^2 - 2b_2^2 + 1| \leq 12\alpha^4 - 52\alpha^3 + 91\alpha^2 - 74\alpha + 24,$$

where b_2 and b_3 are given by (3.9). All these estimations are sharp.

Corollary 3.7. If $g : \Omega \rightarrow \mathbb{C}$ and $G(z) = zg(z) \in \mathcal{SS}_\beta^*(\Omega)$, then for $\beta \in [1/3, 1]$, the following sharp inequalities hold:

$$\left| \left(2 \frac{\partial \rho}{\partial z} \frac{D^3 G(0)(z^3)}{3! \rho^3(z)} \right)^2 - \left(2 \frac{\partial \rho}{\partial z} \frac{D^2 G(0)(z^2)}{2! \rho^2(z)} \right)^2 \right| \leq 9\beta^4 + 4\beta^2$$

and

$$|2b_2^2 b_3 - b_3^2 - 2b_2^2 + 1| \leq 15\beta^4 + 8\beta^2 + 1,$$

where b_2 and b_3 are given by (3.9).

If $\Omega = \mathbb{U}^n$, we obtain the following bounds.

Corollary 3.8. *If $g : \mathbb{U}^n \rightarrow \mathbb{C}$ and $G(z) = zg(z) \in \mathcal{S}^*(\mathbb{U}^n)$, then*

$$\left| \left(\frac{1}{\|z\|^4} \frac{D^3 G(0)(z^3)}{3!} \bar{z} \right)^2 - \left(\frac{1}{\|z\|^3} \frac{D^2 G(0)(z^2)}{2!} \bar{z} \right)^2 \right| \leq 13 \quad (3.13)$$

and

$$|2d_2^2 d_3 - d_3^2 - 2d_2^2 + 1| \leq 24, \quad (3.14)$$

where d_2 and d_3 are given by (3.12). All these bounds are sharp.

Remark 3.1. When $n = 1$, (3.13) and (3.14) reduce to the following:

$$\left| \left(\frac{G^{(3)}(0)}{3!} \right)^2 - \left(\frac{G''(0)}{2!} \right)^2 \right| \leq 13$$

and

$$|2d_2^2 d_3 - d_3^2 - 2d_2^2 + 1| \leq 24,$$

where

$$d_3 = \frac{G^{(3)}(0)}{3!} \quad \text{and} \quad d_2 = \frac{G''(0)}{2!}.$$

which are equivalent to the bounds given in Theorem A.

Corollary 3.9. *If $g : \mathbb{U}^n \rightarrow \mathbb{C}$ and $G(z) = zg(z) \in \mathcal{S}_\alpha^*(\mathbb{U}^n)$, then*

$$\left| \left(\frac{1}{\|z\|^4} \frac{D^3 G(0)(z^3)}{3!} \bar{z} \right)^2 - \left(\frac{1}{\|z\|^3} \frac{D^2 G(0)(z^2)}{2!} \bar{z} \right)^2 \right| \leq (1 - \alpha)^2 (4\alpha^2 - 12\alpha + 13) \quad (3.15)$$

and for $\alpha \in [0, 2/3]$,

$$|2d_2^2 d_3 - d_3^2 - 2d_2^2 + 1| \leq 12\alpha^4 - 52\alpha^3 + 91\alpha^2 - 74\alpha + 24, \quad (3.16)$$

where d_2 and d_3 are given by (3.12). All these estimations are sharp.

Remark 3.2. When $n = 1$, (3.15) and (3.16) reduce to the bounds given in Theorem B.

Corollary 3.10. *If $g : \mathbb{U}^n \rightarrow \mathbb{C}$ and $G(z) = zg(z) \in \mathcal{SS}_\beta^*(\mathbb{U}^n)$, then for $\beta \in [1/3, 1]$, the following sharp inequalities hold:*

$$\left| \left(\frac{1}{\|z\|^4} \frac{D^3 G(0)(z^3)}{3!} \bar{z} \right)^2 - \left(\frac{1}{\|z\|^3} \frac{D^2 G(0)(z^2)}{2!} \bar{z} \right)^2 \right| \leq 9\beta^4 + 4\beta^2 \quad (3.17)$$

and

$$|2d_2^2 d_3 - d_3^2 - 2d_2^2 + 1| \leq 15\beta^4 + 8\beta^2 + 1, \quad (3.18)$$

where d_2 and d_3 are given by (3.12).

Remark 3.3. When $n = 1$, (3.17) and (3.18) reduce to the bounds given in Theorem C.

Declarations

Funding

The work of Surya Giri is supported by University Grant Commission, New Delhi, India under UGC-Ref. No. 1112/(CSIR-UGC NET JUNE 2019).

Conflict of interest

The authors declare that they have no conflict of interest.

Author Contribution

Each author contributed equally to the research and preparation of the manuscript.

Data Availability

Not Applicable.

References

- [1] O. P. Ahuja, K. Khatter and V. Ravichandran, Toeplitz determinants associated with Ma-Minda classes of starlike and convex functions, *Iran. J. Sci. Technol. Trans. A Sci.* **45** (2021), no. 6, 2021–2027.
- [2] M. F. Ali, D. K. Thomas and A. Vasudevarao, Toeplitz determinants whose elements are the coefficients of analytic and univalent functions, *Bull. Aust. Math. Soc.* **97** (2018), no. 2, 253–264.
- [3] F. Bracci, I. Graham, H. Hamada and G. Kohr, Variation of Loewner chains, extreme and support points in the class S^0 in higher dimensions, *Constr. Approx.* **43** (2016), no. 2, 231–251.
- [4] S. Giri, and S. S. Kumar, Toeplitz determinants in one and higher dimensions, *arXiv preprint arXiv:2210.13158* (2022).
- [5] A.W. Goodman, *Univalent Functions*, Mariner, Tampa (1983).
- [6] I. Graham, H. Hamada, T. Honda, G. Kohr and K. H. Shon, distortion and coefficient bounds for Carathéodory families in \mathbb{C}^n and complex Banach spaces, *J. Math. Anal. Appl.* **416** (2014), no. 1, 449–469.
- [7] I. Graham, H. Hamada and G. Kohr, Parametric representation of univalent mappings in several complex variables, *Canad. J. Math.* **54** (2002), no. 2, 324–351.
- [8] I. Graham and G. Kohr, *Geometric function theory in one and higher dimensions*, Monographs and Textbooks in Pure and Applied Mathematics, 255, Marcel Dekker, Inc., New York, 2003.
- [9] I. Graham, G. Kohr and M. Kohr, Loewner chains and parametric representation in several complex variables, *J. Math. Anal. Appl.* **281** (2003), no. 2, 425–438.
- [10] H. Hamada and T. Honda, Sharp growth theorems and coefficient bounds for starlike mappings in several complex variables, *Chinese Ann. Math. Ser. B* **29** (2008), no. 4, 353–368.
- [11] H. Hamada, T. Honda and G. Kohr, Growth theorems and coefficients bounds for univalent holomorphic mappings which have parametric representation, *J. Math. Anal. Appl.* **317** (2006), no. 1, 302–319.
- [12] H. Hamada and G. Kohr, Simple criteria for strongly starlikeness and starlikeness of certain order, *Math. Nachr.* **254/255** (2003), 165–171.
- [13] H. Hamada, G. Kohr and M. Kohr, The Fekete-Szegő problem for starlike mappings and nonlinear resolvents of the Carathéodory family on the unit balls of complex Banach spaces, *Anal. Math. Phys.* **11** (2021), no. 3, Paper No. 115, 22 pp.
- [14] H. Hamada, G. Kohr and P. Liczberski, Starlike mappings of order α on the unit ball in complex Banach spaces, *Glas. Mat. Ser. III* **36(56)** (2001), no. 1, 39–48.

- [15] G. Kohr, On some best bounds for coefficients of several subclasses of biholomorphic mappings in \mathbb{C}^n , Complex Variables Theory Appl. **36** (1998), no. 3, 261–284.
- [16] G. Kohr and P. Liczberski, On strongly starlikeness of order α in several complex variables, Glas. Mat. Ser. III **33(53)** (1998), no. 2, 185–198.
- [17] T. Liu and X. Liu, A refinement about estimation of expansion coefficients for normalized biholomorphic mappings, Sci. China Ser. A **48** (2005), no. 7, 865–879.
- [18] X. Liu and T. Liu, The sharp estimates of all homogeneous expansions for a class of quasi-convex mappings on the unit polydisk in \mathbb{C}^n , Chinese Ann. Math. Ser. B **32** (2011), no. 2, 241–252.
- [19] H. Liu and K. P. Lu, Two subclasses of starlike mappings in several complex variables. Chin Ann. Math. Ser.A. **21(5)**, 533–546 (2000).
- [20] T. Liu and G. Ren, The growth theorem for starlike mappings on bounded starlike circular domains, Chinese Ann. Math. Ser. B **19** (1998), no. 4, 401–408.
- [21] T. Liu and Q. Xu, Fekete and Szegő inequality for a subclass of starlike mappings of order α on the bounded starlike circular domain in \mathbb{C}^n , Acta Math. Sci. Ser. B (Engl. Ed.) **37** (2017), no. 3, 722–731.
- [22] J. A. Pfaltzgraff, Subordination chains and univalence of holomorphic mappings in \mathbb{C}^n , Math. Ann. **210** (1974), 55–68.
- [23] J. A. Pfaltzgraff and T. J. Suffridge, An extension theorem and linear invariant families generated by starlike maps, Ann. Univ. Mariae Curie-Skłodowska Sect. A **53** (1999), 193–207.
- [24] O. Toeplitz, Zur Transformation der Scharen bilinearer Formen von unendlichvielen Veränderlichen. Mathematischphysikalische, Klasse, Nachr. der Kgl. Gessellschaft der Wissenschaften zu Göttingen (1907), pp 110–115.
- [25] Q. Xu, On the coefficient inequality on a bounded starlike circular domain in \mathbb{C}^n , Results Math. **74** (2019), no. 1, Paper No. 60, 13 pp.
- [26] Q. Xu and T. Liu, On coefficient estimates for a class of holomorphic mappings, Sci. China Ser. A **52** (2009), no. 4, 677–686.
- [27] Q.-H. Xu and T.-S. Liu, Biholomorphic mappings on bounded starlike circular domains, J. Math. Anal. Appl. **366** (2010), no. 1, 153–163.
- [28] Q. Xu, T. Liu and W. Zhang, The Fekete and Szegő problem on bounded starlike circular domain in \mathbb{C}^n , Pure Appl. Math. Q. **12** (2016), no. 4, 621–638.
- [29] K. Ye and L.-H. Lim, Every matrix is a product of Toeplitz matrices, Found. Comput. Math. **16** (2016), no. 3, 577–598.

*DEPARTMENT OF APPLIED MATHEMATICS, DELHI TECHNOLOGICAL UNIVERSITY, DELHI–110042, INDIA

E-mail address: spkumar@dtu.ac.in

¹DEPARTMENT OF APPLIED MATHEMATICS, DELHI TECHNOLOGICAL UNIVERSITY, DELHI–110042, INDIA

E-mail address: suryagiri456@gmail.com

TRANSFORMER-BASED NAMED ENTITY RECOGNITION FOR FRENCH USING ADVERSARIAL ADAPTATION TO SIMILAR DOMAIN CORPORA

A PREPRINT

Arjun Choudhry*, Pankaj Gupta*, Inder Khatri, Aaryan Gupta

Biometric Research Laboratory

Delhi Technological University, New Delhi, India

{choudhry.arjun, pankajgupta.dtu, inderkhatri999, aaryan227227}@gmail.com

*These authors contributed equally.

Maxime Nicol

Université du Québec à Montréal
Montréal, QC, Canada

nicol.maxime@courrier.uqam.ca

Marie-Jean Meurs

Université du Québec à Montréal
Montreal, QC, Canada

meurs.marie-jean@uqam.ca

Dinesh Kumar Vishwakarma

Biometric Research Laboratory
Delhi Technological University
New Delhi, India

dinesh@dtu.ac.in

ABSTRACT

Named Entity Recognition (NER) involves the identification and classification of named entities in unstructured text into predefined classes. NER in languages with limited resources, like French, is still an open problem due to the lack of large, robust, labelled datasets. In this paper, we propose a transformer-based NER approach for French using adversarial adaptation to similar domain or general corpora for improved feature extraction and better generalization. We evaluate our approach on three labelled datasets and show that our adaptation framework outperforms the corresponding non-adaptive models for various combinations of transformer models, source datasets and target corpora.

Keywords Named Entity Recognition · French Corpora · Adversarial Adaptation · Information Retrieval · Natural Language Processing

1 Introduction

Named Entity Recognition (NER) is an information extraction task where specific entities are extracted from unstructured text and labelled into predefined classes. While NER models for high-resource languages like English have seen notable performance gains due to improvements in model architectures and availability of large datasets, limited-resource languages like French still face a dearth of openly available, large, labelled datasets. Recent research works use adversarial adaptation frameworks for adapting NER models from high-resource domains to low-resource domains. These approaches have been used for high-resource languages, where robust language models are available. We utilize adversarial adaptation to enable models to learn better, generalized features by adapting them to large, unlabelled corpora for better performance on source test set.

We propose a Transformer-based NER approach for French using adversarial adaptation to counter the lack of large, labelled NER datasets in French. We train transformer-based NER models on labelled source datasets and use larger corpora from similar or mixed domains as target sets for improved feature learning. Our proposed approach helps outsource wider domain and general feature knowledge from easily-available large, unlabelled corpora. While we limit our evaluation to French datasets and corpora, our approach can be applied to other languages too.

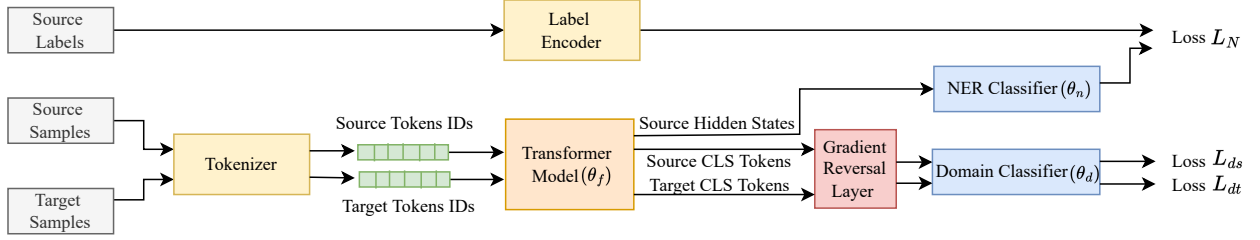


Figure 1: Graphical representation of our adversarial adaptation framework for training NER models on source and target sets.

2 Proposed Methodology

2.1 Datasets and Preprocessing

We use WikiNER French Nothman et al. [2012], WikiNeural French Tedeschi et al. [2021], and Europeana French Neudecker [2016] datasets as the labelled source datasets in our work. Europeana is extracted from historic European newspapers using Optical Character Recognition (OCR), and contains OCR errors, leading to a noisy dataset. For unlabelled target corpora, we use the WikiNER and WikiNeural datasets without their labels, and the Leipzig Mixed French (Mixed-Fr) corpus¹. These enable us to evaluate the impact of adapting models to similar domain, as well as generalized corpora. During preprocessing, we convert all NER tags to Inside-Outside-Beginning (IOB) format Ramshaw and Marcus [1995].

2.2 Adversarial Adaptation to Similar Domain Corpus

Adversarial adaptation helps select domain-invariant features transferable between source and target datasets Ganin et al. [2016]. Based on this premise, we propose that adversarially adapting NER models to large, unlabelled corpora from similar domain as the source helps enable the model to extract more generalizable features. This reduces overfitting on the intricate training set-specific features. We also test the same for the case where target dataset is a mixed-domain, large corpus. We test our approach for three conditions: source and target datasets are from the same domain; source and target datasets are from relatively different domains; and target dataset is a mixed-domain, large-scale, general corpus. Figure 1 illustrates our proposed framework. The domain classifier acts as a discriminator. NER classifier loss, adversarial loss, and total loss are defined as:

$$L_{NER} = \min_{\theta_f, \theta_n} \sum_{i=1}^{n_s} L_n^i \quad (1)$$

$$L_{adv} = \min_{\theta_d} (\max_{\theta_f} (\sum_{i=1}^{n_s} L_{ds}^i + \sum_{j=1}^{n_t} L_{dt}^j)) \quad (2)$$

$$L_{Total} = L_{NER} + \alpha(L_{adv}) \quad (3)$$

where n_s and n_t are number of samples in source and target sets, θ_d , θ_n and θ_f are number of parameters for domain classifier, NER classifier and transformer model, L_{ds} and L_{dt} represent the Negative log likelihood loss for source and target respectively, and α is ratio between L_{NER} and L_{adv} . We found $\alpha = 2$ to provide the best experimental results.

2.3 Language Models for NER

Recent NER research has incorporated large language models due to their contextual knowledge learnt during pretraining Yan et al. [2021], Gong et al. [2019], Lothritz et al. [2020]. We use three French language models for evaluating our proposed approach: CamemBERT-base Martin et al. [2020], CamemBERT-Wiki-4GB (a variant of CamemBERT pretrained on only 4GB of Wikipedia corpus), and FlauBERT-base Le et al. [2020]. Comparing CamemBERT-base

¹<https://wortschatz.uni-leipzig.de/en/download/French>

Model	Source	Target	Precision	Recall	F1
CamemBERT-Wiki-4GB	WikiNER	WikiNeural Mixed-Fr	0.911	0.925	0.918
			0.966	0.963	0.969
			0.956	0.962	0.959
	WikiNeural	WikiNER Mixed-Fr	0.859	0.872	0.866
			0.872	0.891	0.881
			0.870	0.879	0.875
	Europeana	WikiNER Mixed-Fr	0.728	0.642	0.682
			0.738	0.691	0.714
			0.774	0.640	0.701
CamemBERT-base	WikiNER	WikiNeural Mixed-Fr	0.960	0.968	0.964
			0.973	0.976	0.975
			0.972	0.978	0.974
	WikiNeural	WikiNER Mixed-Fr	0.943	0.950	0.946
			0.943	0.953	0.948
			0.946	0.950	0.948
	Europeana	WikiNER Mixed-Fr	0.927	0.933	0.930
			0.911	0.927	0.920
			0.942	0.943	0.943
FlauBERT-base	WikiNER	WikiNeural Mixed-Fr	0.963	0.964	0.963
			0.964	0.968	0.966
			0.974	0.972	0.973
	WikiNeural	WikiNER Mixed-Fr	0.934	0.946	0.940
			0.935	0.950	0.942
			0.941	0.943	0.942
	Europeana	WikiNER Mixed-Fr	0.835	0.863	0.849
			0.855	0.865	0.860
			0.882	0.854	0.867

Table 1: Performance evaluation of our proposed approaches for various combinations of models, source and target sets.

and CamemBERT-Wiki-4GB helps us analyse if we can replace large language models with smaller ones adapted to unlabelled corpora during fine-tuning on a downstream task.

3 Experimental Results and Discussion

We evaluated our approach on various combinations of language models, source and target datasets. Each model was evaluated on the test set of source dataset. Table 1 illustrates our results. Some findings observed are described hereafter.

Adversarial adaptation models outperform their non-adaptive counterparts: We observed that the adaptation models consistently outperformed their non-adaptive counterparts across almost all combinations of datasets and language models on precision, recall and F1-score.

Adversarial adaptation can help alleviate performance loss on using smaller models: Fine-tuning CamemBERT-Wiki-4GB using our adversarial approach helped achieve similar performance to non-adapted CamemBERT-base for certain datasets. CamemBERT-Wiki-4GB adapted to WikiNeural corpus even outperformed unadapted CamemBERT-base for WikiNER dataset. Thus, adversarial adaptation during fine-tuning could act as a substitute for using larger language models.

Adapting models to same domain target corpora leads to slightly better performance than adapting to a mixed corpus: We observed that models adapted to corpora from same domain as source dataset (like for WikiNER and WikiNeural as source and target datasets, or vice versa) showed equal or slightly better performance than models adapted to general domain.

Adapting models to mixed-domain target corpus leads to better performance than adapting to a corpus from a different domain: We observed that models adapted to mixed-domain corpora (Europeana to Mixed-Fr) showed noticeably better performance than models adapted to corpora from different domains (Europeana to WikiNER).

Reproducibility. The source code of the proposed systems is licensed under the GNU GPLv3. The datasets are publicly available.

Acknowledgments. This research was enabled by support provided by Calcul Québec, The Alliance and MITACS.

References

- Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning Multilingual Named Entity Recognition from Wikipedia. *Artificial Intelligence*, 194, 2012. doi:10.1016/j.artint.2012.03.006.
- Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *EMNLP*, 2021.
- Clemens Neudecker. An open corpus for named entity recognition in historic newspapers. In *LREC*, 2016.
- Lance A. Ramshaw and Mitchell P. Marcus. Text Chunking using Transformation-Based Learning, 1995. URL <https://arxiv.org/abs/cmp-lg/9505040>.
- Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1), jan 2016.
- Rongen Yan, Xue Jiang, and Depeng Dang. Named Entity Recognition by Using XLNet-BiLSTM-CRF. *Neural Process. Lett.*, 53(5):3339–3356, oct 2021. ISSN 1370-4621. doi:10.1007/s11063-021-10547-1. URL <https://doi.org/10.1007/s11063-021-10547-1>.
- Cheng Gong, Jiuyang Tang, Shengwei Zhou, Zepeng Hao, and Jun Wang. Chinese Named Entity Recognition with BERT. *DEStech Transactions on Computer Science and Engineering*, 12 2019. doi:10.12783/dtcse/cisnrc2019/33299.
- Cedric Lothritz, Kevin Allix, Lisa Veiber, Tegawendé F. Bissyandé, and Jacques Klein. Evaluating Pretrained Transformer-based Models on the Task of Fine-Grained Named Entity Recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3750–3760, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi:10.18653/v1/2020.coling-main.334. URL <https://aclanthology.org/2020.coling-main.334>.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT: a tasty French language model. In *ACL*, 2020. doi:10.18653/v1/2020.acl-main.645.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. FlauBERT: Unsupervised language model pre-training for French. In *LREC*, May 2020.



Twin core photonic crystal fiber based temperature sensor with improved sensitivity over a wide range of temperature

Vishal Chaudhary¹ · Sonal Singh¹

Received: 5 August 2022 / Accepted: 15 October 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract

This article presents the twin-core photonic crystal fiber (TC-PCF) for temperature sensing application. The twin solid cores in the cross-section of TC-PCF structure, separated by one vertical elliptical air hole, act as two independent waveguides. This novel arrangement of circular and elliptical air holes in our proposed structure gives an edge over other existing structures to attain high sensitivity. The proposed sensor is highly birefringent and operates on the principle of mode coupling between the two fiber cores. It is practically possible to realize this TC-PCF based temperature sensor by coupling one fiber core to a broadband source on the input end while the other fiber core to an optical spectrum analyzer on the output end. The finite element method has been employed to simulate and quantitatively analyze the proposed TC-PCF temperature sensor. Numerical simulation shows that, the 3 cm long TC-PCF sensor has been optimized to have a high temperature sensitivity of around 21.5 pm/°C over a wide temperature sensing range of 0–1200 °C. Also the impact of diameter variation of air holes on sensitivity of proposed model is studied.

Keywords Twin core PCF · Temperature sensitivity · Birefringence · Finite element method · Mode coupling

1 Introduction

Of late, optical fiber sensors have outshined traditional electronic sensors for multi-sensing capabilities such as pressure, stress, temperature, strain and acoustic signals as they offer a great number of advantages over latter such as light weight, high sensitivity, relatively inexpensive, immunity to radio frequency interference and their ability to multiplex sensor networks. In the field of photonics, photonic crystal fibers (PCFs) exhibit a great potential compared to conventional optical fibers. Physical characteristics of PCF's like refractive index, pressure, displacement, curvature, temperature, torsion, vibration, and electric field can be used to analyse optical sensors. Its endless single-mode feature and structural analysis make it a factor to consider for a variety of applications, including sensor, splitter

✉ Sonal Singh
sonalsingh@dtu.ac.in

¹ Department of Electronics and Communication Engineering, Delhi Technological University, Shahbad Daultpur, Main Bawana Road, Delhi 110042, India

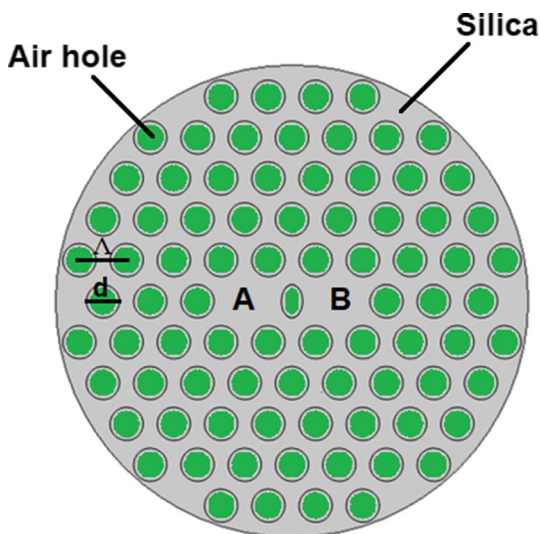
(Dhara and Singh 2015; Liu et al. 2012; Xu et al. 2018), narrow band pass filter (Saitoh et al. 2005), photonic crystal fiber coupler (Koshiba et al. 2003; Lægsgaard et al. 2004), and wavelength MUX and DEMUX (Hameed et al. 2015). A growing number of TC-PCF sensors have been created in recent years, such as pressure (Liu et al. 2012), Refractive Index (R.I.) (Dhara and Singh 2015), bio and temperature sensors (Revathi et al. 2015; Rindorf and Bang 2008). Many research groups have been actively involved in enhancing the temperature sensitivity of dual core-PCF. For ex. S. Revathi et al. (2015) investigated a pressure and temperature sensor based upon Dual-Core Photonic Quasi-Crystal Fiber in 2015 and obtained a temperature sensitivity of about $20 \text{ pm}/^\circ\text{C}$ over the range of $0\text{--}1000^\circ\text{C}$. Similar results have also been observed by S. Jegadeesan et al. (2019) over the same temperature range using different model of PCF. D. Chen and group presented a Dual-Core PCF based pressure/temperature sensor in 2011 achieving a temperature sensitivity of around $20.7 \text{ pm}/^\circ\text{C}$ over the range of $0\text{--}1000^\circ\text{C}$ (Chen et al. 2011).

In this work, we propose a PCF with twin solid cores separated by single vertical elliptical air hole for a wide temperature sensing range of up to 1200°C . When optical light enters the TC-PCF from first core, it travels along the TC-PCF from first core to second core. When temperature is applied to the TC-PCF, the optical light at the edge of output in second core is measured with the help of optical analyzer using the transmission graph. The shift in transmission curve of TC-PCF is observed due to change of R.I. of silica material which thus helps in determining the temperature sensitivity of the TC-PCF. COMSOL software, which is based on the Finite Element Method (FEM), is used to design the proposed TC-PCF setup.

2 PCF structure and result analysis

The proposed TC-PCF is made up of twin-fiber cores separated by single vertical elliptical air hole. Figure 1 illustrates the geometry of the proposed TC-PCF. Two missing holes depicted as A and B represents the twin fiber cores of the PCF. Pitch (Λ) is the

Fig. 1 Geometry of proposed twin core PCF



center-to-center length between air holes. A diameter, d of $1.4 \mu\text{m}$ and a pitch length, Λ of $2 \mu\text{m}$ is chosen for air holes for the proposed PCF and pure silica is taken as a background material. The major-axis of the single elliptical hole in the proposed configuration is $a = 1.4 \mu\text{m}$, the minor axis, $b = 0.9 \mu\text{m}$ and the R.I. of air, $n_{\text{air}} = 1$ is considered. Equation (1) gives the R.I. of pure silica (n_{silica}) which is called Sellmeier equation.

$$n_{\text{silica}}^2(\lambda) = 1 + \frac{c_1 \lambda^2}{\lambda^2 - d_1} + \frac{c_2 \lambda^2}{\lambda^2 - d_2} + \frac{c_3 \lambda^2}{\lambda^2 - d_3} \quad (1)$$

The operational wavelength in μm is denoted by λ and the Sellmeier coefficients are $c_1 = 0.696166300$, $c_2 = 0.407942600$, $c_3 = 0.897479400$, $d_1 = 4.67914826 \times 10^{-3} \mu\text{m}^2$, $d_2 = 1.35120631 \times 10^{-2} \mu\text{m}^2$ and $d_3 = 97.9340025 \mu\text{m}^2$, respectively (Rifat et al. 2016).

In this work, due to small core-to-core length, twin fiber cores inside the PCF form two waveguides, each of which is independent (accompanying with coupling mode) to each other. At 1550 nm , Fig. 2 depicts the electric field vector and amplitude distribution of the four polarized super-modes viz., the x-even, x-odd, y-even and y-odd. The coupling effect observed in TC-PCF is attributed to these super-modes based on FEM.

Figure 3 shows the schematic diagram of the PCF sensing set up. Broadband light source, temperature controller and optical spectrum analyzer are the essential tools. A single-mode fiber (SMF) can be used to feed light into the fiber core A of the proposed TC-PCF from a broadband light source. Through another SMF, the output signal can be

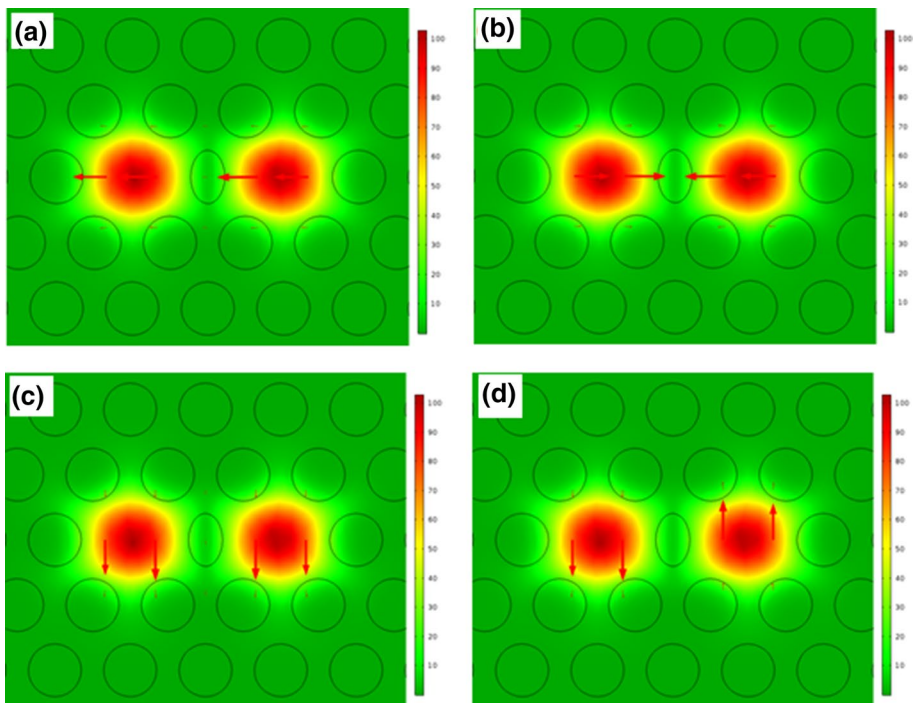


Fig. 2 Electric field vector and amplitude distribution of the four super-modes: **a** x-polarization even **b** x-polarization odd **c** y-polarization even and **d** y-polarization odd

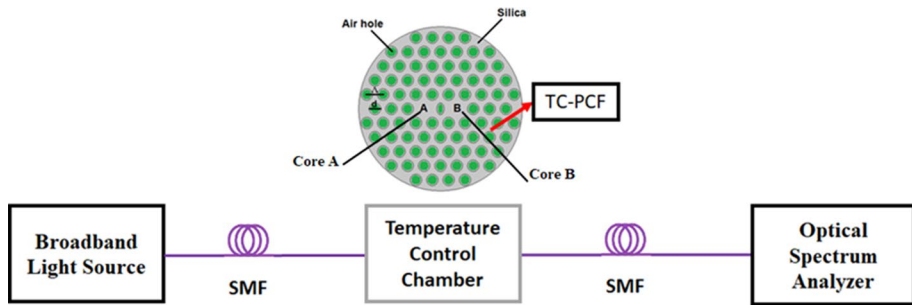


Fig. 3 Schematic of PCF sensing set up for temperature sensor

seen from the fiber core B in an optical spectrum analyzer. The PCF's surrounding environment temperature can be adjusted using the temperature control chamber.

At 1550 nm, the n_{eff}^x even and odd modes are 1.4013807059 and 1.4006623442, respectively and for n_{eff}^y even and odd mode are 1.4009086063 and 1.4001836831, respectively. Figure 4 demonstrates how the n_{eff} of the even and odd modes variation with wavelength for x-polarized and y-polarized modes. The effective refractive index falls as the wavelength increases.

The Δn_{eo} fluctuation, which is $\Delta n_{eo} = |n_e - n_o|$, for x and y polarized modes is depicted in Fig. 5. As the effective index difference between the core and cladding increases, the wavelength increases. The rise in Δn_{eo} of the TC-PCF implies that light is trapped more tightly in the core.

The TC-PCF coupling length is defined as

$$L_c = \frac{\pi}{|\beta_e^i - \beta_o^i|} = \frac{\lambda}{2|n_e^i - n_o^i|} \quad (2)$$

where $i = x, y$, the even and odd mode propagation constants of the TC-PCF are β_e^i and β_o^i , respectively. Even and odd mode effective refractive indices are represented by n_e^i and n_o^i , respectively (Koshiba et al. 2003; Wang et al. 2007). We first estimated the effective refractive indexes of x and y polarized modes, then calculated the difference in effective

Fig. 4 Variation in n_{eff} with wavelength

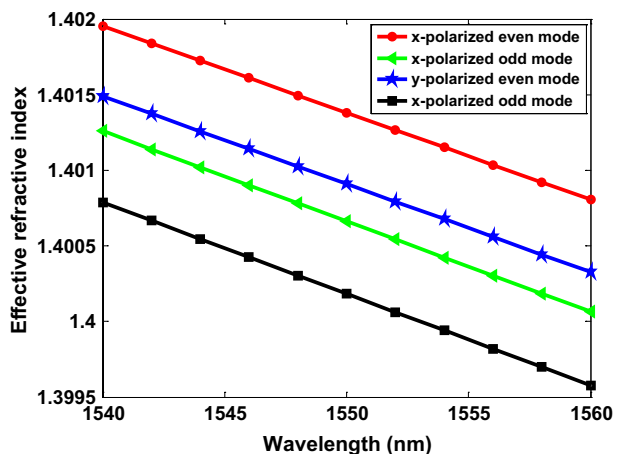
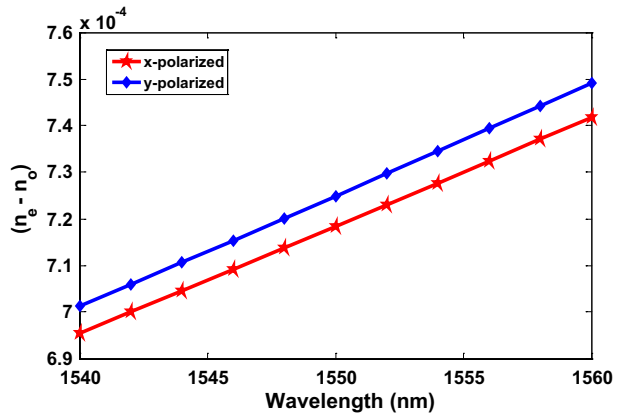


Fig. 5 Variation in Δn_{eo} with wavelength

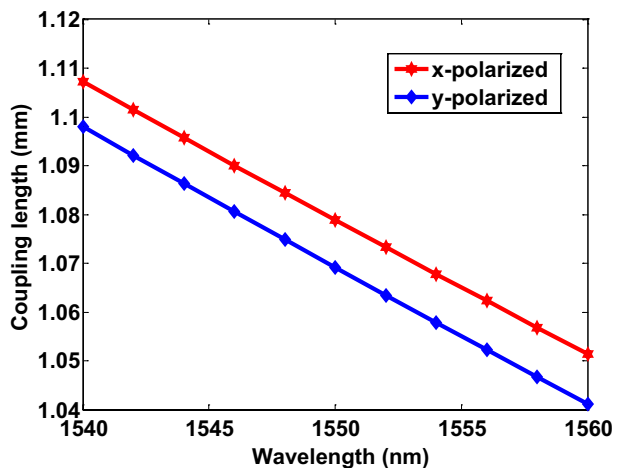
refractive indexes for x and y polarized modes and finally determined their coupling lengths using Eq. (2) for different wavelength.

Figure 6 depicts the change in coupling length as a function of wavelength. To determine the value of Coupling length for different wavelength using Eq. (2), the n_{eff} of four super modes are first computed followed by the Δn_{eo} for x and y polarized mode. It should be noted that the y-polarized coupling length is smaller than that of x-polarized mode and that the coupling length reduces as the wavelength increases.

Confinement loss is the term for the attenuation caused by the waveguide geometry. This type of loss occurs in single material fibers, especially in PCFs because they are generally made of silica and given by

$$\text{Confinement loss} = -20 \log_{10} \varepsilon^{-k \text{Im}[n_{eff}]} = 8.686 k \text{Im}[n_{eff}] \quad (3)$$

where k represents the propagation constant ($k = 2\pi/\lambda$) in free space, λ denotes the wavelength and $\text{Im}[n_{eff}]$ represents the imaginary part of the complex effective index (Russell 2006).

Fig. 6 Coupling length variation with wavelength

The confinement loss of the proposed TC-PCF is calculated using Eq. (3). At transmission wavelength of 1550 nm, the low confinement loss of the proposed PCF comes out to be 1.23×10^{-14} dB/cm, 2.14×10^{-14} dB/cm, 2.77×10^{-14} dB/cm and 5.39×10^{-15} dB/cm for x-polarized even, x-polarized odd, y-polarized even and y-polarized odd modes, respectively.

Birefringence causes double refraction when a light ray strikes a birefringent material, polarization splits it into two different rays that travels in slightly different directions. The birefringence parameter, B of TC-PCF is calculated by using Eq. (4).

$$B = \left| \operatorname{Re} \left(n_{\text{eff}}^x \right) - \operatorname{Re} \left(n_{\text{eff}}^y \right) \right| \quad (4)$$

where the effective R.I. of x and y polarized modes are represented by n_{eff}^x and n_{eff}^y , respectively (Ortigosa-Blanch et al. 2001).

When birefringence is low, the x and y polarized coupling lengths are almost equal, making it tough to split the light. According to Fig. 6, the x polarized coupling length of the proposed TC-PCF is larger than that of y-polarized mode. The high birefringence of 4.721×10^{-4} and 4.787×10^{-4} is achieved for even and odd mode, respectively at $\lambda = 1550$ nm. The changes in birefringence with operational wavelength and at the applied temperature are shown in Fig. (7).

Following a length L, the TC-PCF power transfer is computed by using Eq. (5).

$$I(\lambda) = \sin^2 \left(\frac{\pi}{\lambda} \Delta n_{eo} L \right) \quad (5)$$

The effective R.I. difference is denoted by (Δn_{eo}) and effective R.I. of the even and odd modes are represented by the letters n_e^i and n_o^i respectively (Koshiba et al. 2003; Martynkien et al. 2010).

Figure 8 illustrates the TC-PCF power transmission graph for x-polarized mode when the length is 3 cm at sensing temperature range of 0–1200 °C. It should be noted that the proposed TC-PCF power transmission curve is sinusoidal.

The sensitivity of the proposed twin-core PCF for temperature sensing is determined by shifting of peaks with the rise in temperature detected across the wide range of 0–1200 °C. The twin-core PCF temperature sensitivity is defined as

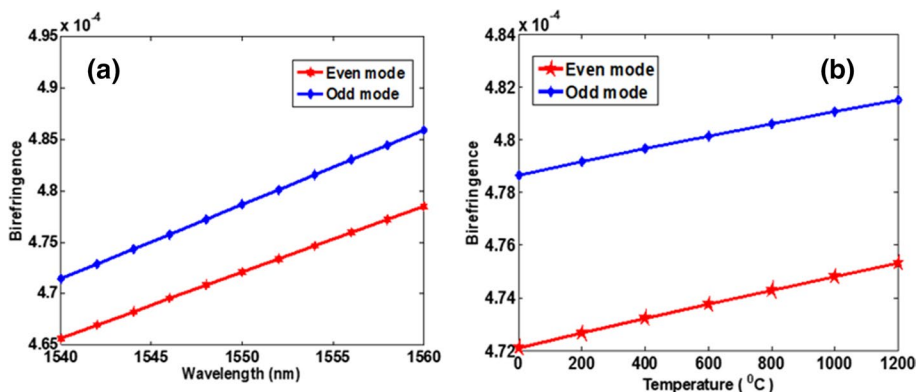
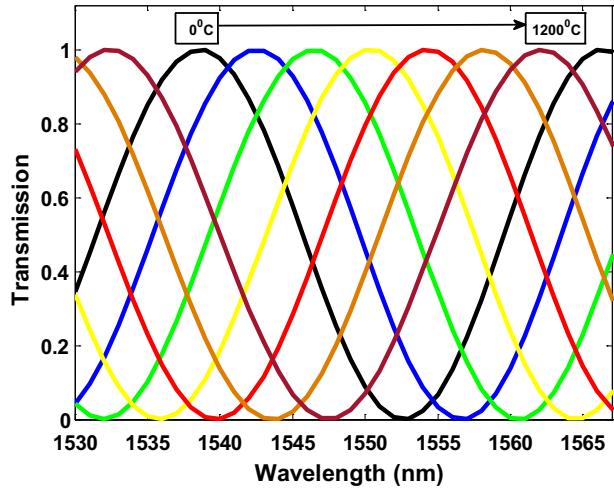


Fig. 7 Variation in birefringence with **a** wavelength **b** temperature applied (°C)

Fig. 8 TC-PCF transmission curve with a length of 3 cm



$$S_{\lambda} (nmRIU^{-1}) = \Delta \lambda_{peak} / \Delta T \quad (6)$$

where $\Delta \lambda_{peak}$ represent the overall shift in wavelength peak due to fluctuation in temperature and ΔT is the shift in temperature (Gauvreau et al. 2007).

The R.I. value of the TC-PCF will vary very little above normal temperature due to the effect of thermo-optic, when heat is applied to the fiber. Due to its almost zero thermal expansion, silica is a ceramic with interesting uses as a filler in composites (Rodrigues and Marinkovic 2022). The relation between refractive index and applied temperature is determined by using Eq. (7).

$$n = n_0 + T \frac{dn}{dT} \quad (7)$$

where n_0 denotes the R.I. of pure silica at zero degree celsius and the thermo-optic coefficient of silica is obtained by using the following formula (Ortigosa-Blanch et al. 2001):

$$\frac{dn}{dT} = 10^{-5} (1/^{\circ}C) \quad (8)$$

For the sensing temperature range of 0–1200 °C, Fig. 9 depicts the shift of the wavelength peak. The transmission graph is linear, shifting to a higher wavelength when the temperature rises.

Figure 10 depicts the numerical fitting line for shift in wavelength spots in the transmission curve of TC-PCF when the temperature is applied. For a 3 cm fiber, the twin-core PCF has a temperature sensitivity of 21.5 pm/°C.

The proposed sensor performance is summarized in Table 1. The presented sensor outperformed the recently published PCF sensors, as shown in Table 1.

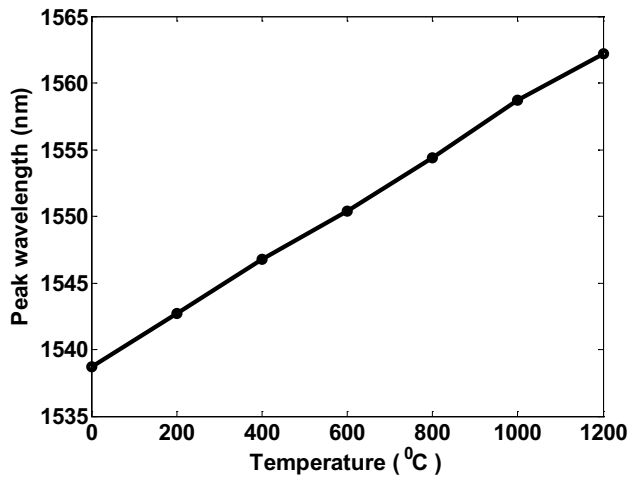


Fig. 9 Transmission spectrum peak wavelength (nm) with applied temperature (°C)

Fig. 10 Numerical line fitting for various wavelength shift values for a 3 cm long TC-PCF at wide temperature range of 0–1200 °C

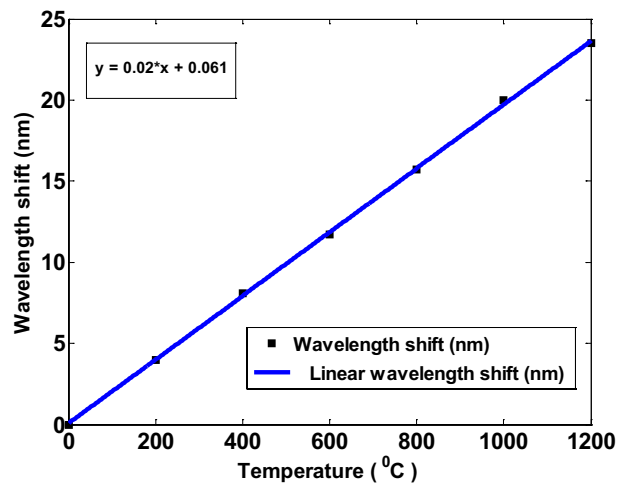


Table 1 Performance comparison of different PCF sensors with our proposed PCF sensor

Reference	Sensitivity (pm/°C)	Sensitivity range (°C)
Revathi et al. (2015)	20	0–1000
Jegadeesan et al. (2019)	20	0–1000
Chen et al. (2011)	20.7	0–1000
Chaudhary et al. (2020)	18.5	0–600
Yu et al. (2018)	15.61	300–1200
Proposed PCF	21.5	0–1200

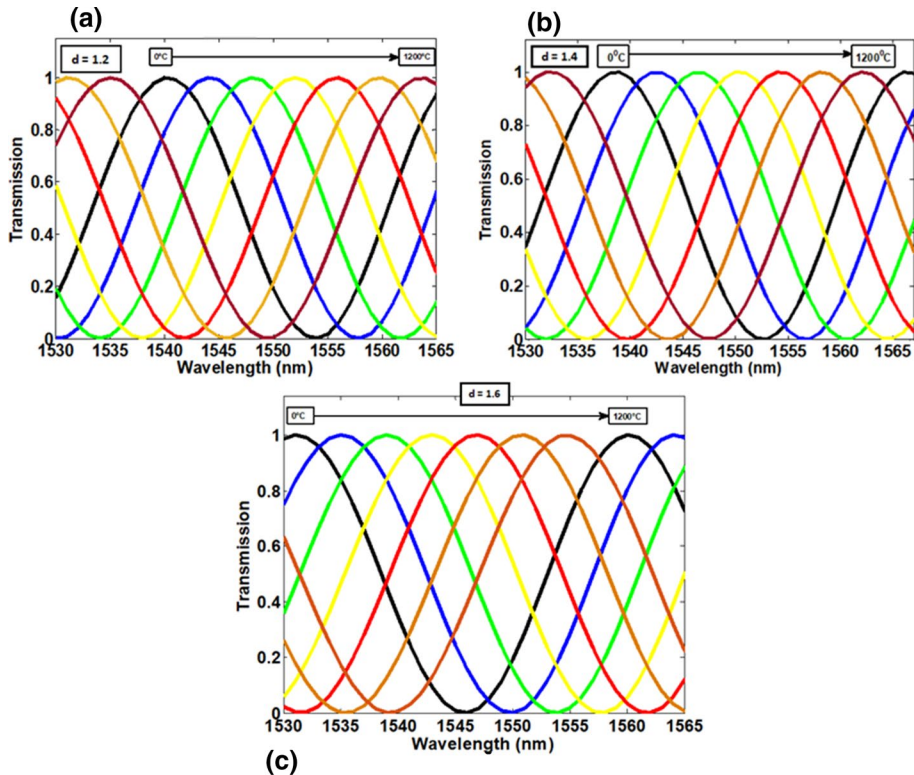


Fig. 11 TC-PCF transmission curve with different diameter of d , **a** $1.2 \mu\text{m}$ **b** $1.4 \mu\text{m}$ and **c** $1.6 \mu\text{m}$

Table 2 Comparison of sensitivity of proposed TC-PCF with different diameters

Diameter (μm)	Sensitivity ($\text{pm}/^\circ\text{C}$)
$d = 1.2$	20
$d = 1.4$	21.5
$d = 1.6$	20.5

3 Impact of diameter variation on sensitivity

Figure 11 shows the TC-PCF power transmission graph for the x-polarized mode with different diameters when the length is 3 cm at a sensing temperature range of 0–1200 °C. According to Fig. 11a, b and c for a temperature sensing range of 0–1200 °C, the twin-core PCF can achieve a temperature sensitivity of 20 $\text{pm}/^\circ\text{C}$ at $d = 1.2 \mu\text{m}$, 21.5 $\text{pm}/^\circ\text{C}$ at $d = 1.4 \mu\text{m}$ and 20.5 $\text{pm}/^\circ\text{C}$ at $d = 1.6 \mu\text{m}$, respectively. As indicated from the graph, when the diameter of hole rings is taken as $1.2 \mu\text{m}$, a blue shift towards higher wavelength is observed, as compared to the diameter of hole rings taken as $1.4 \mu\text{m}$ and when the diameter of hole rings is taken as $1.6 \mu\text{m}$, a red shift towards lower wavelength is observed, as compared to the diameter of hole rings taken as $1.4 \mu\text{m}$.

The variation of sensitivity of proposed TC-PCF with different diameter shown in Table 2.

4 Conclusion

Using the COMSOL software and the Finite Element Method, we have proposed a temperature-sensor based on twin-core PCF highly sensitive over a wide range of temperature. The high sensitivity is ascribed to the unique combination of circular and elliptical air holes arranged in our proposed structure. The R.I. of pure silica at different temperature is determined using the thermo-optic coefficient equation. Furthermore, the proposed PCF temperature sensitivity is evaluated using coupling length and transmission spectrum calculations. Under broad temperature sensing range of 0–1200 °C, the twin-core PCF can achieve a temperature sensitivity of 21.5 pm/°C for 3 cm fiber length with the diameter of 1.4 μm . The key advantages of a TC-PCF based temperature sensor are its low cost, small size, and better stability. Using the current existing known PCF manufacturing techniques, facile fabrication of proposed TC-PCF is possible.

Acknowledgements The authors are grateful to the participants who contributed to this research. The authors have not received any funding for this research.

Author contributions VC: Methodology, Writing-review and editing. SS: Supervision, Writing—original draft, review and editing.

Funding The authors have not received any funding for this research.

Availability of data and materials Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal ties that would have influenced the work presented in this study.

Animal research No bad Impact on animal.

Consent to participate All author is agreed to submit the paper in this journal.

References

- Chaudhary, V.S., Kumar, D., Mishra, R., Sharma, S.: Twin core photonic crystal fiber for temperature sensing. *Mater. Today Proc.* **33**, 2289–2292 (2020). <https://doi.org/10.1016/j.matpr.2020.04.197>
- Chen, D., Hu, G., Chen, L.: Pressure/temperature sensor based on a dual-core photonic crystal fiber. *Asia Commun. Photonics Conf. Exhib. ACP* **2011**(8307), 1–10 (2011). <https://doi.org/10.1117/12.904029>
- Dhara, P., Singh, V.K.: Effect of MMF stub on the sensitivity of a photonic crystal fiber interferometer sensor at 1550 nm. *Opt. Fiber Technol.* **21**, 154–159 (2015). <https://doi.org/10.1016/j.yofte.2014.11.008>
- Gauvreau, B., Hassani, A., Fassi Fehri, M., Kabashin, A., Skorobogatiy, M.A.: Photonic bandgap fiber-based Surface Plasmon Resonance sensors. *Opt. Express.* **15**, 11413–11426 (2007). <https://doi.org/10.1364/oe.15.011413>
- Hameed, M.F.O., et al.: Polarization-independent surface plasmon liquid crystal photonic crystal multiplexer-demultiplexer. *IEEE Photonics J.* **7**, 1–10 (2015). <https://doi.org/10.1109/JPHOT.2015.2480538>

- Jegadeesan, S., Dhamodaran, M., Shanmugapriya, S.S.: Numerical analysis of dual-core photonic crystal fiber based temperature and pressure sensor for oceanic applications. *Opt. Appl.* **49**, 249–264 (2019). <https://doi.org/10.5277/oa190206>
- Koshiba, M., Saitoh, K., Sato, Y.: Coupling characteristics of dual-core photonic crystal fiber couplers. *Opt. Express.* **11**, 3188–3195 (2003)
- Lægsgaard, J., Bang, O., Bjarklev, A.: Photonic crystal fiber design for broadband directional coupling. *Opt. Lett.* **29**, 2473–2475 (2004). <https://doi.org/10.1364/ol.29.002473>
- Liu, Z., Tse, M.-L.V., Wu, C., Chen, D., Lu, C., Tam, H.-Y.: Intermodal coupling of supermodes in a twin-core photonic crystal fiber and its application as a pressure sensor. *Opt. Express.* **20**, 21749–21757 (2012). <https://doi.org/10.1364/oe.20.021749>
- Martynkien, T., Statkiewicz-Barabach, G., Olszewski, J., Wojcik, J., Mergo, P., Geernaert, T., Sonnenfeld, C., Anuszkiewicz, A., Szczurowski, M.K., Tarnowski, K., Makara, M., Skorupski, K., Klimek, J., Poturaj, K., Urbanczyk, W., Nasilowski, T., Berghmans, F., Thienpont, H.: Highly birefringent micro-structured fibers with enhanced sensitivity to hydrostatic pressure. *Opt. Express.* **18**, 15113–15121 (2010). <https://doi.org/10.1364/oe.18.015113>
- Ortigosa-Blanch, A., Knight, J.C., Wadsworth, W.J., Arriaga, J., Mangan, B.J., Birks, T.A., Russell, P.S.J.: Highly birefringent photonic crystal fibers. *Opt. Photonics News.* **12**, 17 (2001). <https://doi.org/10.1364/opn.12.12.000017>
- Revathi, S., Inabathini, S.R., Pal, J.: Pressure and temperature sensor based on a dual core photonicquasi-crystal fiber. *Optik (stuttgart)*. **126**, 3395–3399 (2015). <https://doi.org/10.1016/j.ijleo.2015.07.141>
- Rifat, A.A., Mahdiraji, G.A., Sua, Y.M., Ahmed, R., Shee, Y.G., Adikan, F.R.M.: Highly sensitive multi-core flat fiber surface plasmon resonance refractive index sensor. *Opt. Express.* **24**, 2485–2495 (2016). <https://doi.org/10.1364/oe.24.002485>
- Rindorf, L., Bang, O.: Sensitivity of photonic crystal fiber grating sensors: biosensing, refractive index, strain, and temperature sensing. *J. Opt. Soc. Am. B.* **25**, 310–324 (2008). <https://doi.org/10.1364/josab.25.000310>
- Rodrigues, L.M., Marinkovic, B.A.: Effects of fused silica addition on thermal expansion, density, and hardness of alumix-231 based composites. *Materials.* **15**, 3476 (2022)
- Russell, P.S.J.: Photonic-crystal fibers. *J. Light. Technol.* **24**, 4729–4749 (2006). <https://doi.org/10.1109/JLT.2006.885258>
- Saitoh, K., Florous, N.J., Koshiba, M., Skorobogatiy, M.: Design of narrow band-pass filters based on the resonant-tunneling phenomenon in multi-core photonic crystal fibers. *Opt. Express.* **13**, 10327–10335 (2005). <https://doi.org/10.1364/oe.13.010327>
- Wang, Z., Taru, T., Birks, T.A., Knight, J.C., Liu, Y., Du, J.: Coupling in dual-core photonic bandgap fibers: theory and experiment. *Opt. Express.* **15**, 4795–4803 (2007). <https://doi.org/10.1364/oe.15.004795>
- Xu, Q., Zhao, Y., Xia, H., Lin, S., Zhang, Y.: Ultrashort polarization splitter based on dual-core photonic crystal fibers with gold wire. *Opt. Eng.* **57**, 046104 (2018). <https://doi.org/10.1117/1.oe.57.4.046104>
- Yu, H., Wang, Y., Ma, J., Zheng, Z., uo, Z., Zheng, Y.: Fabry-perot interferometric high-temperature sensing up to 1200 °c based on a silica glass photonic crystal fiber. *Sensors* **18**(1), 273 (2018). <https://doi.org/10.3390/s18010273>

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.