ISSN Print: 2518-4245
ISSN Online: 2518-4253
Vol. 62(4), December 2025

PROCEEDINGS
OF THE PAKISTAN ACADEMY OF SCIENCES:
A. Physical and Computational Sciences

PAKISTAN ACADEMY OF SCIENCES
ISLAMABAD, PAKISTAN

Proceedings of the Pakistan Academy of Sciences: Part A
Physical and Computational Sciences

Proceedings of the Pakistan Academy of Sciences A. Physical and Computational Sciences is the official flagship, the peer-reviewed quarterly journal of the Pakistan Academy of Sciences. This open-access journal publishes original research articles and reviews on current advances in the field of Computer Science (all), Materials Science (all), Physics and Astronomy (all), Engineering Sciences (all), Chemistry, Statistics, Mathematics, Geography, Geology in English. Authors are not required to be Fellows or Members of the Pakistan Academy of Sciences or citizens of Pakistan. The journal is covered by Print and Online ISSN, indexed in Scopus, and distributed to scientific organizations, institutes and universities throughout the country, by subscription and on an exchange basis.

Annual Subscription: **Pakistan:** Institutions, Rupees 8000/-; Individuals, Rupees 4000/- (Delivery Charges: Rupees 300/-)
Other Countries: US$ 200.00 (includes air-lifted overseas delivery)

HEC Recognized; Scopus Indexed

Published by **Pakistan Academy of Sciences**, 3 Constitution Avenue, G-5/2, Islamabad, Pakistan
Email: editor@paspk.org; **Tel:** 92-51-920 7140 & 921 5478; **Websites:** www.paspk.org/proceedings/; www.ppaspk.org

Printed at **Graphics Point.,** Office 3-A, Wasal Plaza, Fazal-e-Haq Road Blue Area Islamabad.
Ph: 051-2806257, **E-mail:** graphicspoint16@gmail.com

PROCEEDINGS OF THE
PAKISTAN ACADEMY OF SCIENCES: PART A
Physical and Computational Sciences

CONTENTS

Volume 62, No. 4, December 2025 **Page**

Submission of Manuscripts: Manuscripts may be submitted as an e-mail attachment at editor@paspk.org or submit online at http://ppaspk.org/index.php/PPASA/about/submissions. Authors must consult the ***Instructions for Authors*** at the end of this issue or at the Website: www.paspk.org/proceedings/ or www.ppaspk.org.

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 271-275 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)705

Pakistan Academy of Sciences

Review Article

Radiation Techniques in Health and Environment†

A.K. Azad Chowdhury[1*], Nusrat Jahan Shawon[2], and Mohammad Mahfujur Rahman[3]

[1]Bangladesh Academy of Sciences & Department of Clinical Pharmacy and Pharmacology, University of Dhaka, Dhaka, Bangladesh

[2]Department of Pharmacy, Independent University, Bangladesh (IUB), Dhaka, Bangladesh

[3]Department of Radiation Oncology, Evercare Hospital Dhaka, Bangladesh

Abstract: Radiation science has become a cornerstone of modern medicine, offering powerful tools for both diagnosis and treatment. Diagnostic imaging technologies such as X-ray, ultrasonography, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and gamma camera systems utilize radiation to provide high-resolution visualization of internal structures. Therapeutic applications have evolved from conventional radiotherapy to highly sophisticated techniques including Photon Beam Radiotherapy using LINAC, Gamma Knife, and CyberKnife systems. Advanced modalities such as Stereotactic Radiosurgery (SRS), and Stereotactic Body Radiation Therapy (SBRT) allow for precise delivery of high-dose radiation to tumors while minimizing exposure to surrounding healthy tissue. Emerging techniques such as FLASH radiotherapy, which delivers radiation at very high speeds, and carbon ion therapy, which is effective against resistant tumors, are bringing major improvements to cancer treatment. Cherenkov radiation is being explored for its role in treatment visualization and dosimetry, while Targeted Radionuclide Therapy (TRT) uses tumor-specific radioactive agents to deliver internal radiation precisely to cancer cells. Adaptive Radiation Therapy (ART) modifies treatment plans during therapy to account for tumor or patient changes. These developments are shaping the future of oncology, with an emphasis on precision, safety, and therapeutic efficiency. Beyond medicine, radiation is also applied in environmental protection. It is used for purifying wastewater through radiolysis, sterilizing hazardous solid waste, facilitating the breakdown of plastics, and detecting pollutants using nuclear analytical methods. These applications highlight the broader utility of radiation in supporting both health and environmental sustainability.

Keywords: Radiation, Gamma Irradiation, FLASH Radiotherapy, Targeted Radionuclide Therapy, Environmental Radiation Applications.

1. INTRODUCTION

Radiation has been a cornerstone of medical science since its discovery in the late 19th century, providing powerful tools for both diagnosis and treatment of diseases, particularly cancer [1]. Radiation therapy, the therapeutic application of ionizing radiation, is a major modality in cancer management, with nearly 50% of patients receiving radiotherapy during their illness to inhibit tumor growth and maximize curative outcomes [2]. The underlying principle of radiotherapy relies on the ability of high-energy radiation to damage the genetic material (DNA) of cancer cells, preventing their proliferation while minimizing exposure to surrounding healthy tissue [3]. The energy transported by radiation is governed by Einstein's mass-energy equivalence equation $E = mc^2$ [4], while the interaction of electromagnetic fields with biological tissues is described by Ampère-Maxwell's law, $\nabla B = \mu_0 (J + \varepsilon_0 \frac{\partial E}{\partial t})$ [5]. Furthermore, the quantum nature of radiation is captured by the Planck-Einstein relation, $E = h\nu$, linking photon energy to frequency [6] and by Einstein's photoelectric equation, $KE = h\nu - \phi$, which describes the kinetic energy of ejected electrons as a function of photon energy and the material's work function [7].

Radiotherapy not only serves curative purposes but also plays a pivotal role in palliative care,

Received: November 2025; Revised: December 2025; Accepted: December 2025
* Corresponding Author: A.K. Azad Chowdhury <akchowdhury2003@yahoo.com>
† This paper was presented in "AASSA-PAS Symposium on Radiation Techniques in Health and Environment" from 18-20 August 2025, held at Islamabad, Pakistan.

alleviating symptoms such as pain, obstruction, or compression caused by tumors. Thus, the integration of physics, imaging, and clinical expertise has made radiation a vital component of modern medical practice, offering both life-saving treatment and improved quality of life for patients [1].

2. MEDICAL IMAGING TECHNIQUES

2.1. X-ray

X-rays are a form of ionizing radiation with wavelengths of 0.01–10 nm, widely used in medical imaging for visualizing internal structures based on differential absorption and transmission through tissues. Modern X-ray systems, including computed radiography, flat-panel detectors, and CT, provide high-resolution 2D and 3D images essential for diagnosing fractures, bone disorders, soft tissue abnormalities, and guiding surgical or interventional procedures. Advances in detector technology and imaging techniques have improved image quality while reducing patient radiation exposure [8, 9].

2.2. Ultrasonography

Ultrasonography has rapidly advanced, offering high-resolution real-time imaging of anatomy, pathology, and blood flow. It is safe, quick, and often superior to CT or MRI in uncooperative or lean patients, though limitations exist with obesity, gas, and bone interfaces. High-quality sonography requires extensive training and expertise, while handheld devices hold promise for screening and enhancing routine clinical diagnosis [10].

2.3. Computed Tomography (CT)

Computed tomography (CT) provides high-resolution, cross-sectional images that accurately distinguish tissues, enabling precise assessment of body composition, including adipose tissue, skeletal muscle, bones, and organs. Modern multidetector CT (MDCT) allows rapid acquisition of three-dimensional volume images with sub-millimeter resolution, improving both speed and reproducibility of measurements. CT can also quantify bone mineral density and fat infiltration in muscles or liver, making it a reliable tool for clinical evaluation and research [11, 12].

2.4. Magnetic Resonance Imaging (MRI)

Magnetic Resonance Imaging (MRI) is a non-invasive technique that produces high-resolution images using strong magnetic fields and radiofrequency radiation, providing excellent soft tissue contrast. It is widely used in clinical diagnostics, radiotherapy planning, and pharmaceutical research to study tissue structure, tumor margins, and in vivo drug delivery [13-15].

2.5. Positron Emission Tomography (PET)

Positron Emission Tomography (PET) is a functional imaging technique widely used in oncology for tumor staging, treatment response assessment, and radiotherapy planning, providing early insights into tumor metabolism beyond anatomical imaging. PET imaging has evolved from early research tools to sophisticated clinical scanners with 3D acquisition, iterative reconstruction, and time-of-flight technology, improving sensitivity, image quality, and quantitative tumor assessment [16, 17].

3. ADVANCED RADIOTHERAPY MODALITIES

3.1. Gamma Knife

Gamma Knife radiosurgery has evolved over the past decades as a minimally invasive alternative for treating intracranial tumors, vascular malformations, and functional disorders, particularly medically refractory tumors. Its advantages include precise high-dose radiation delivery without craniotomy, making it suitable for patients unfit for invasive surgery [18].

3.2. CyberKnife

The CyberKnife system is a frameless, image-guided radiosurgery platform that integrates a compact 6-MV LINAC with a robotic arm to deliver highly precise, non-isocentric radiation beams. Real-time imaging and motion correction allow accurate targeting of both intracranial and extracranial lesions without invasive stereotactic frames. Its treatment planning software supports multimodality imaging fusion, inverse planning, and dose optimization, enabling safe irradiation of complex tumor shapes while sparing adjacent structures. Since FDA approval in 2001, CyberKnife

has been widely adopted as an effective alternative to conventional surgery and radiosurgery systems such as the Gamma Knife [19-22].

3.3. LINAC

LINAC-based radiotherapy uses high-energy X-rays to precisely target tumors while sparing normal tissues. Modern techniques like IMRT and VMAT, combined with image guidance, improve dose accuracy, though CBCT has limitations in soft tissue visualization and motion management. The integration of MRI with LINAC (MR-Linac) allows real-time imaging, adaptive treatment, and better tumor targeting, enhancing efficacy and reducing toxicity [23, 24].

3.4. Stereotactic Radiosurgery (SRS) and Stereotactic Body Radiation Therapy (SBRT)

Stereotactic radiosurgery (SRS) and stereotactic body radiation therapy (SBRT) are noninvasive, high-dose radiotherapy techniques targeting cranial and extracranial tumors, respectively, using image guidance and stereotactic alignment for precise delivery. SRS typically involves a single high-dose session for brain lesions, while SBRT delivers a few large doses to extracranial tumors, including lung, liver, and prostate. Both modalities are effective in local tumor control, with ongoing studies refining their use and exploring combination with targeted systemic therapies [25].

3.5. FLASH Radiotherapy (FLASH-RT)

FLASH radiotherapy (FLASH-RT) delivers ultra-high dose-rate radiation within milliseconds, which has shown the ability to spare normal tissues while maintaining strong antitumor efficacy. Preclinical studies across multiple species and early clinical cases demonstrate reduced toxicity compared to conventional radiotherapy, making FLASH-RT a promising approach for overcoming radio-resistant tumors [26, 27].

3.6. Targeted Radionuclide Therapy (TRT)

TRT delivers cytotoxic radiation to tumor cells using radiolabeled molecules such as antibodies, peptides, or small ligands, minimizing damage to normal tissues. Common applications include I-131 for thyroid cancer, Y-90 ibritumomab tiuxetan and I-131 tositumomab for non-Hodgkin's lymphoma, and Lu-177-DOTA-TATE or Y-90-DOTA-TOC for neuroendocrine tumors [28, 29].

3.7. Adaptive Radiation Therapy (ART)

ART is a closed-loop radiotherapy approach that continuously adapts treatment plans using systematic feedback from patient-specific measurements. Unlike conventional radiotherapy that applies uniform margins based on population averages, ART customizes field margins and radiation doses to individual anatomical and positional variations, thereby enhancing both safety and effectiveness. This process employs advanced technologies such as CT imaging, electronic portal imaging devices, multileaf collimators, and computer-controlled systems to monitor changes and re-optimize treatment in real time. By accounting for organ motion, geometric target shifts, and treatment beam placement errors, ART reduces unnecessary radiation exposure to healthy tissues. It also allows for safer dose escalation by tailoring margins to the actual variability of each patient rather than generalized estimates. Ultimately, ART represents a dynamic, patient-centered strategy that refines radiation delivery and improves therapeutic outcomes [30].

4. ENVIRONMENTAL APPLICATIONS OF RADIATION

4.1. Wastewater Purification

Radiation technology, particularly gamma irradiation, has shown significant potential in purifying municipal wastewater by effectively reducing physical and organic contaminants. Laboratory studies indicate that gamma doses between 100 - 500 krad can degrade up to 88% of organic pollutants while inactivating pathogenic microorganisms, thus lowering biochemical oxygen demand (BOD) and chemical oxygen demand (COD). The method also improves sludge compactness and settling capacity, making it a promising alternative to conventional treatments. With optimized radiation parameters and pilot-scale validation, this technology can provide cost-effective and environmentally compatible wastewater treatment [31, 32].

4.2. Solid Waste Treatment

Radiation technologies have emerged as effective tools for the treatment and disinfection of solid and liquid wastes, addressing growing global concerns over pollution and public health. Techniques such as gamma irradiation, electron-beam, ultraviolet, and X-rays have been applied to sterilize sewage sludge, biomedical wastes, and industrial effluents, while also degrading toxic contaminants in soils.

Gamma irradiation, particularly using cobalt-60, has demonstrated practical efficacy in field-scale applications, providing pathogen-free, nutrient-rich sludge suitable for agricultural use. These technologies offer significant advantages, including odorless, easily handled waste and elimination of withholding periods before crop use, making radiation a promising approach for sustainable waste management [33].

4.3. Pollutant Detection

Radiation techniques, particularly laser-based absorption spectroscopy, are increasingly used to detect and quantify gaseous pollutants in the atmosphere. By targeting specific infrared absorption bands of pollutants such as carbon monoxide, nitric oxide, sulfur dioxide, and ozone, lasers provide high sensitivity and selectivity even at very low concentrations. The collimated, high-power laser beams allow long-distance transmission and multiple-pass absorption, overcoming limitations of traditional light sources and enhancing real-time environmental monitoring [34].

4.4. Plastic Waste Degradation

Radiation processing, using gamma rays or electron beams, effectively modifies the structure of synthetic and natural polymers, enhancing properties such as thermal stability, biodegradability, and mechanical strength. It facilitates plastic waste degradation, accelerates breakdown of cellulose into viscose, and improves chitin/chitosan processing without toxic chemicals.

Electron beam and gamma irradiation offer environmentally friendly alternatives to conventional chemical methods, providing cost-effective and sustainable polymer modification for industrial and environmental applications [35].

5. CONCLUSIONS

Radiation technologies have become indispensable across medicine and environmental management, offering precise, efficient, and versatile solutions. In healthcare, advances in diagnostic imaging and targeted radiotherapy improve tumor control, minimize normal tissue damage, and enable personalized treatment strategies. Environmentally, radiation applications in wastewater purification, solid waste sterilization, and pollutant detection provide sustainable and effective approaches to safeguard public health. Together, these innovations underscore the transformative potential of radiation science in enhancing both human health and environmental protection.

6. CONFLICT OF INTEREST

The authors declare no conflict of interest.

7. REFERENCES

1. D. Abshire and M.K. Lang. The evolution of radiation therapy in treating cancer. *Seminars in Oncology Nursing* 34(2): 151-157 (2018).
2. A. Aggarwal, G. Lewison, D. Rodin, A. Zietman, R. Sullivan, and Y. Lievens. Radiation therapy research: A global analysis 2001-2015. *International Journal of Radiation Oncology* Biology* Physics* 101(4): 767-778 (2018).
3. R. Baskar, K.A. Lee, R. Yeo, and K.W. Yeoh. Cancer and radiation therapy: current advances and future directions. *International Journal of Medical Sciences* 9(3): 193 (2012).
4. M.J. Feigenbaum and N.D. Mermin. E= mc^2. *American Journal of Physics* 56(1): 18-21 (1988).
5. S.E. Hil. Reanalyzing the ampère-maxwell law. *The Physics Teacher* 49(6): 343-345 (2011).
6. P.L. Ward. On the Planck-Einstein Relation. (2020). https://whyclimatechanges.com/relation.pdf.
7. C. Wan, J.M. Vaughn, J.T. Sadowski, and M.E. Kordesch. Scandium oxide coated polycrystalline tungsten studied using emission microscopy and photoelectron spectroscopy. *Ultramicroscopy* 119: 106-110 (2012).
8. M. Berger, Q. Yang, and A. Maier. X-ray Imaging. In: Medical imaging systems: an introductory guide. A. Maier, S. Steidl, V. Christlein, and J. Hornegger (Eds.). *Springer, Cham, Switzerland* pp. 119-145 (2018).

9. X. Ou, X. Chen, X. Xu, L. Xie, X. Chen, Z. Hong, H. Bai, X. Liu, Q. Chen, L. Li, and H. Yang. Recent development in x-ray imaging technology: Future and challenges. *Research* 2021: 9892152 (2021).
10. P.W. Ralls, R.B. Jeffrey, R.A. Kane, and M. Robbin. Ultrasonography. *Gastroenterology Clinics of North America* 31(3): 801-825 (2002).
11. M. Mazonakis and J. Damilakis. Computed tomography: What and how does it measure? *European Journal of Radiology* 85(8): 1499-1504 (2016).
12. D.W. Townsend. Positron emission tomography/ computed tomography. *Seminars in Nuclear Medicine* 38(3): 152-166 (2008).
13. G. Katti, S. Ara, and A. Shireen. A. Magnetic resonance imaging (MRI)–A review. *International Journal of Dental Clinics* 3(1): 65-70 (2011).
14. V.S. Khoo, D.P. Dearnaley, D.J. Finnigan, A. Padhani, S.F. Tanner, and M.O. Leach. Magnetic resonance imaging (MRI): considerations and applications in radiotherapy treatment planning. *Radiotherapy and Oncology* 42(1): 1-15 (1997).
15. J.C. Richardson, R.W. Bowtell, K. Mäder, and C.D. Melia. Pharmaceutical applications of magnetic resonance imaging (MRI). *Advanced Drug Delivery Reviews* 57(8): 1191-1209 (2005).
16. A. Gallamini, C. Zwarthoed, and A. Borra. Positron emission tomography (PET) in oncology. *Cancers* 6(4): 1821-1889 (2014).
17. G. Muehllehner and J.S. Karp. Positron emission tomography. *Physics in Medicine and Biology* 51(13): 117-137 (2006).
18. Y. Higuchi, S. Matsuda, and T. Serizawa. Gamma knife radiosurgery in movement disorders: Indications and limitations. *Movement Disorders* 32(1): 28-35 (2017).
19. W. Cheng and J.R. Adler. An overview of cyberknife radiosurgery. *Chinese Journal of Clinical Oncology 3*(4): 229-243 (2006).
20. C. Ding, C.B. Saw, and R.D. Timmerman. Cyberknife stereotactic radiosurgery and radiation therapy treatment planning system. *Medical Dosimetry* 43(2): 129-140 (2018).
21. W. Kilby, M. Naylor, J.R. Dooley, C.R.M. Jr, and S. Sayeh. A technical overview of the CyberKnife system. In: Handbook of robotic and image-guided surgery. M.H. Abedin-nasab (Ed.). *Elsevier, Amsterdam* pp.15-38 (2020).
22. J.S. Kuo, C. Yu, Z. Petrovich, and M.L. Apuzzo. The CyberKnife stereotactic radiosurgery system: description, installation, and an initial evaluation of use and functionality. *Neurosurgery* 62: 785-789 (2008).
23. J.J. Lagendijk, B.W. Raaymakers, and M. Van Vulpen. The magnetic resonance imaging–linac system. *Seminars in Radiation Oncology* 24(3): 207-209 (2014).
24. X. Liu, Z. Li, and Y. Yin. Clinical application of MR-Linac in tumor radiotherapy: a systematic review. *Radiation Oncology* 18: 52 (2023).
25. B.D. Kavanagh and R.D. Timmerman. Stereotactic radiosurgery and stereotactic body radiation therapy: an overview of technical considerations and clinical applications. *Hematology/Oncology Clinics of North America* 20(1): 87-95 (2006).
26. B. Lin, F. Gao, Y. Yang, D. Wu, Y. Zhang, G. Feng, T. Dai, and X. Du. FLASH radiotherapy: History and Future. *Frontiers in Oncology* 11: 644400 (2021).
27. J. Bourhis, P. Montay-Gruel, P.G. Jorge, C. Bailat, B. Petit, J. Ollivier, W. Jeanneret-Sozzi, M. Ozsahin, F. Bochud, R. Moeckli, and J.F. Germond. Clinical translation of FLASH radiotherapy: Why and how? *Radiotherapy and Oncology* 139: 11-17 (2019).
28. D. Ersahin, I. Doddamane, and D. Cheng. Targeted radionuclide therapy. *Cancers* 3(4): 3838-3855 (2011).
29. S.V. Gudkov, N.Y. Shilyagina, V.A. Vodeneev, and A.V. Zvyagin. Targeted radionuclide therapy of human tumors. *International Journal of Molecular Sciences* 17(1): 33 (2015).
30. D. Yan, F. Vicini, J. Wong, and A. Martinez. Adaptive radiation therapy. *Physics in Medicine & Biology* 42(1): 123-132 (1997).
31. M.Y. Al-Ani and F.R. Al-Khalidy. Use of ionizing radiation technology for treating municipal wastewater. *International Journal of Environmental Research and Public Health* 3(4): 360-368 (2006).
32. R.O.A. Rahman and Y.T. Hung. Application of ionizing radiation in wastewater treatment: an overview. *Water* 12(1): 19 (2019).
33. P.H. Rathod, J.C. Patel, M.R. Shah, and A.J. Jhala. Evaluation of gamma irradiation for bio-solid waste management. *International Journal of Environment and Waste Management* 2(1-2): 37-48 (2008).
34. P.L. Hanst and J.A. Morreal. Detection and measurement of air pollutants by absorptions of infrared radiation. *Journal of the Air Pollution Control Association* 18(11): 754-759 (1968).
35. A.G. Chmielewski, M. Haji-Saeid, and S. Ahmed. Progress in radiation processing of polymers. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 236(1-4): 44-54 (2005).

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 277-288 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)845

Pakistan Academy of Sciences

Research Article

Improving Roman Urdu Topic Classification through Custom Stemming and an SGD-Optimized Machine Learning Pipeline

Muhammad Aqeel[1], Irfan Qutab[2], Khawar Iqbal[3*], Habiba Fiaz[4], and Hira Arooj[5]

[1]School of Software, Northwestern Polytechnical University, Xi'an, China
[2]Department of Engineering, University of Modena and Reggio Emilia, Modena, Italy
[3]Riphah School of Computing and Innovation, Riphah International University, Lahore, Pakistan
[4]School of Mathematics and Statistics, Northwestern Polytechnical University, Xi'an, China
[5]Department of Mathematics and Statistics, The University of Lahore, Sargodha, Pakistan

Abstract: All over social media and internet platforms, Roman Urdu content is extremely casual, inconsistent, and linguistically diversified, which makes it hard to interpret through conventional Natural Language Processing (NLP) techniques. This paper proposes a strong topic-classification framework for Roman Urdu, integrating Stochastic Gradient Descent (SGD)-optimized machine learning, dictionary-assisted stemming, and custom lexical normalization in order to overcome those challenges. The method consists of structured preprocessing, reduction of repeated letters, rule-based normalization, extraction of TF-IDF features, and the evaluation of a few classifiers including Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), K-Nearest Neighbors (KNN), along with the proposed model of SGD. The proposed classifier outperformed all the baseline models with an accuracy of 95%, according to the experimental results on the four-class dataset comprised of Politics, Sports, Education, and Religion. The results depict the importance of stemming and normalization to improve feature quality and reduce orthographic variability in low-resource languages. All things considered, this study provides a repeatable and efficient pipeline for Roman Urdu subject classification and thus lays a concrete foundation for further Roman Urdu NLP research.

Keywords: Roman Urdu Stemmer, TF-IDF, Stochastic Gradient Descent, Topic Classification, Machine Learning.

1. INTRODUCTION

Topic classification using Natural Language Processing (NLP) is a major application, where machines classify texts into predefined categories. Topic classification refers to classifying a document into predefined topics such as social media, news, or reviews. Efficient topic classification systems for multiple languages are becoming more important with the rapid increase in online contents, especially social media contents. Large Language Models [1] or deep learning models for specialized domains [2] are some of the recent advancements that were taken into consideration. Efficient topic classification systems for multiple languages are becoming increasingly important with the rapid growth in online contents, especially social media contents. In South Asia, Roman Urdu which is a form of Urdu written in Latin script is frequently practiced. Roman Urdu undergoes an informal language with limited resources, regardless of its increasing popularity, which leads to substantial challenges for automated text classification [3]. By formulating a high accuracy topic classification system particularly for Roman Urdu, integrating its lexical variation and morphological irregularities, this study aims to address these shortcomings. Roman Urdu is used in a significant portion of South Asian discussion forums because Urdu is one of the languages that are most frequently used in the world [4]. Roman Urdu's lack of standard orthographic structures and a more informal atmosphere of social

Received: December 2023; Revised: November 2025; Accepted: December 2025
* Corresponding Author: Khawar Iqbal <khawar.iqbal@riphah.edu.pk>

media have contributed to the growing number of non-standard spellings, which makes automated text categorization far more challenging [5, 6]. For that reason, it is vital to build such tools that can arrange and classify this massive amount of user-generated content for improved information access and interpretation.

Roman Urdu has received little attention in recent studies, which mainly focused on text classification for high resource languages like English. Techniques using deep learning for text classification have been previously investigated by Minaee *et al.* [7]. These techniques perform well in settings where resources are abundant, but they show limitations when applied to languages with limited resources such as Roman Urdu. In this regard, Gasparetto *et al.* [8] studied algorithms for text categorization and also demonstrated how hard it can be to apply these approaches to unstructured and informal texts such as Roman Urdu. While TF-IDF (Term Frequency-Inverse Document Frequency) is an established feature extraction method [9], it has yet to be studied extensively on Roman Urdu due to the presence of nonstandard spelling and irregular forms in the language that render such methods very difficult to apply. Similarly, Hussain *et al.* [10] carried out a detailed study on Roman Urdu sentiment detection but did not present any preprocessing mechanism, which is considered crucial in topic classification. Similarly, the study carried out by Arshad *et al.* [11] on the recognition of emotions in Roman Urdu text failed to consider the specific preprocessing requirements of the language.

Although, Pakray *et al.* [12] focused on low resource language processing, issues related to Roman Urdu were not sufficiently focused on, where its informal expressions and spelling irregularities make classification a highly challenging job. As far as stemming is concerned, although it has been well studied for languages like English, it does not suffice to handle Roman Urdu, and an efficient stemmer for Roman Urdu remains missing. Adimulam *et al.* [13] focused on transfer learning in languages with very minimal resources. However, the unique morphological constraints pertaining to Roman Urdu were not clearly explored in this work. Avetisyan and Broneske [14] made an effort to review low resource languages but did not provide any customized solution for Roman Urdu, which further gives weight to the importance of effective preprocessing. Similarly, Ògúnremí *et al.* [15], while discussing decolonizing NLP for low resource languages, did not explore those very unique complexities existing in Roman Urdu text.

While the studies of Sandu *et al.* [16] and Chen *et al.* [17] focused on text extraction techniques for social media, they did not cater specifically to Roman Urdu but rather focused their approach on strongly resourced languages. Ghafoor *et al.* [18] studied multilingual text processing, but again, their work did not cover methods that could cater to the rich lexical features of Roman Urdu. Even though TF-IDF is a widespread feature extraction technique, it needs further tuning to deal with informal writing patterns of Roman Urdu. Kumar *et al.* [19] assessed deep learning for hyperspectral image classification, failing to assess the challenge of text classification for low-resourced languages like Roman Urdu. Additionally, Faheem *et al.* [20] investigated part of speech tagging for Roman Urdu but did not expand their work to topic classification and Hussain *et al.* [10] addressed the challenges of emotion recognition in Roman Urdu; however, their work did not discuss topic categorization, which considers a broader perspective of Roman Urdu textual characteristics.

Roman Urdu text categorization has drawn more interest, especially in view of complications linked with the detection of sentiment and emotions. The work of Ilyas *et al.* [21] identified the recognition of emotions in code mixed Roman Urdu-English text, their research has avoided specific challenges that arise when dealing with pure Roman Urdu text, such as the irregular spelling and lack of standardization of the language.

In the same direction, Chandio *et al.* [22] have proposed an attention-driven Residual Unit–Bidirectional LSTM (RU-BiLSTM) framework for sentiment analysis targeting Roman Urdu, but they failed to take into account carefully the difficulty of the topic classification, opening a way to deal with a greater variety of textual structures. Nabeel *et al.* [23] used machine learning (ML) models to classify emotions in Roman Urdu posts but the struggles of classifying topics within this language context were not taken into account by them. Khan *et al.* [24] worked on the sentiment analysis for Roman Urdu from a multilingual point of view, they

predominantly focused on emotion identification, leaving a gap in the establishment of broader topic classification systems. More generalized issue of topic categorization, which has not yet explored, was also ignored by Rana *et al.* [25], who contributed in the area of Roman Urdu language by offering an unsupervised method for analysis of sentiments on social media short text classification.

Tejaswini *et al.* [26] examined social media text interpretation using NLP methods and hybrid deep learning models for detecting depression, and the work of Lavanya and Sasikala [27] explored text classification in social healthcare settings using NLP and deep learning, both of these studies mainly relied on sentiment analysis and did not address the specific challenges of topic classification, which is the focus of our work. The need for improved approaches to Roman Urdu text processing becomes clear when considering that Akhter *et al.* [28] focused on identifying abusive language in both Urdu and Roman Urdu but did not extend the analysis to topic categorization. Similarly, Mehmood *et al.* [29] proposed a discriminant approach for feature spamming and played their role in the analysis of sentiment for Roman Urdu; however, their research work did not incorporate topic classification.

Mehmood *et al.* [30] used a hybrid approach for sentiment analysis of Roman Urdu through the Xtreme multi-channel technique. However, their work still had some shortcomings since it missed the aspect of topic classification. Saeed *et al.* [31] worked on the area of toxic comment classification for Urdu and Roman Urdu by developing the PURUTT corpus, which aimed at enhancing the detection of toxic comments. However, their work does not tackle the key issue of topic classification.

In conclusion, despite some progress made in sentiment analysis and toxic comment detection for Roman Urdu-Urdu, there is still a gap in the application of such techniques to topic classifications. Feature extraction techniques such as TF-IDF and n-gram techniques have gained considerable attention, however, issues such as non-standard spelling, colloquial language use, and small datasets still exist. Therefore, the proposed study strengthens the Stochastic Gradient Descent (SGD) by developing a more accurate topic classification technique and a Roman Urdu stemmer.

2. MATERIALS AND METHODS

Roman Urdu stemming and a vast amount of ML experiments form the basis of this study's methodology. Logistic Regression (LR) [9], Support Vector Machine (SVM) [30], SGD, K-Nearest Neighbors (KNN), Naïve Bayes (NB) and Decision Tree (DT) [32] were among the algorithms whose performances we assessed. The establishment of a method for Roman Urdu text topic classification using SGD is a major accomplishment of this study. Figure 1 is a conceptual illustration of our proposed methodology. Our method incorporates the use of the TF-IDF weighting scheme, but just before inserting the data into the model, a lexical dictionary is utilized to guide a critical stemming process. By contemplating the various spellings and variations in Roman Urdu, this dictionary contributes in standardizing the text. The main purpose of this step is to improve the feature selection process.

It starts with data cleaning, which deletes irrelevant symbols and punctuation marks from the text. Next, lexical normalization is conducted by using a rule-based approach, followed by stemming. Together, these form the preprocessing stage of the work, which is really important to

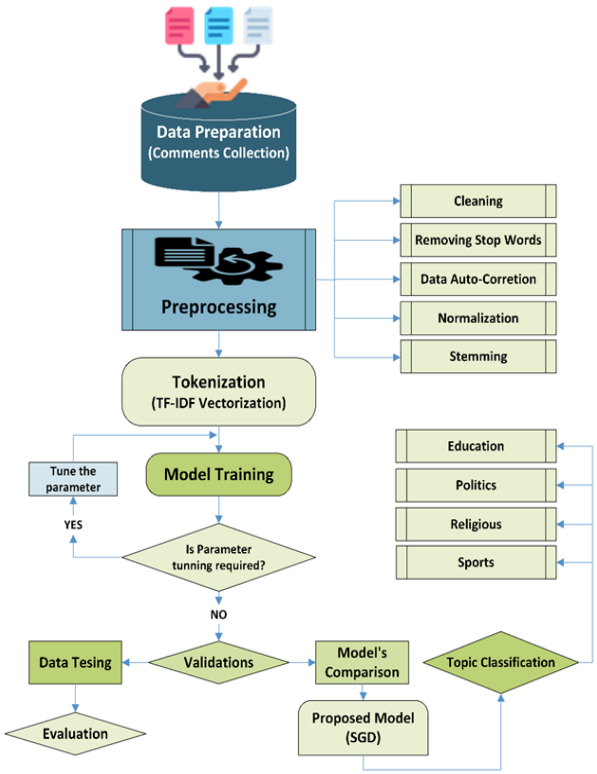


Fig. 1. Proposed model methodology workflow.

handle the irregularities present in Roman Urdu text. A TF-IDF vectorizer was then applied for feature extraction, while a number of ML models were subsequently used for the classification.

2.1. Dataset

The Roman Urdu dataset[1] used in this research has been collected from Kaggle, a well monitored platform acknowledged for its rich dataset repository and data science competitions. This dataset is a very valuable collection of text data, particularly in the Roman Urdu language, which covers a wide range of topics and sentiments. The corpus is collected from online forums and social blogs, hence offering a rich and reliable repository of real-world linguistic interactions and individual opinions. It provides a very useful insight into how people express their sentiments and opinions in Roman Urdu about diverse topics. The dataset consists of 4065 comments, hence, the data is labeled with categories like politics (1398), sports (1092), education (851), and religious (724). The politics and sports categories are most represented, followed by the education and religious comments, as captured in Figure 2, thereby reflecting an imbalanced yet diverse distribution in the corpus.

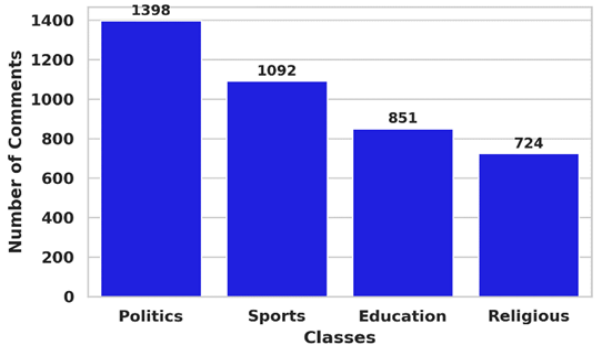


Fig. 2. Roman Urdu Dataset.

2.2. Preprocessing

Preprocessing is important as it retains only the significant words and removes the rid of rest. Filler words like "punch lines," "number characters" and "stop words" were deleted. The data preprocessing decreases computation time and size of the data. Doing that in NLTK library (Python), several operations are performed including removing the unnecessary words and characters, auto correcting and stemming.

2.2.1. Remove Stopping Words

Stop words are those common and repetitive words, which do not appear as useful information for the sentiment prediction. The idea of stop words was first introduced by Luhn [33]. In this paper, we perform a manual selection for these stop words. We will use a curated set of Urdu stop words to efficiently remove irrelevant words, reducing the data processing step. Figure 3 shows the stop words of Roman Urdu.

2.2.2. Data Auto Correction

For the unstructured Roman Urdu used in informal comments over the web, people usually use incorrect syntactical structures, hence the mining process is complicated. Hence, someone might stretch out characters of a word "bohttttttttt khubbbbbb" instead of the desired "boht khub" meaning "well done" in response to this our system attempts to resolve these ill formedness as by identifying the correct syntactic composition of words in order to facilitate better analysis [34].

```
stopwords=['ai', 'ayi', 'hy', 'hai', 'main', 'ki', 'tha', 'koi',
'ko', 'sy', 'woh', 'bhi', 'aur', 'wo', 'yeh', 'rha', 'hota', 'ho',
'ga', 'ka', 'le', 'lye', 'kr', 'kar', 'lye', 'liye', 'hotay',
'waisay', 'gya', 'gaya', 'kch', 'ab', 'thy', 'thay', 'houn', 'hain',
'han', 'to', 'is', 'hi', 'jo', 'kya', 'thi', 'se', 'pe', 'phr',
'wala', 'waisay', 'us', 'na', 'ny', 'hun', 'rha', 'raha', 'ja',
'rahay', 'abi', 'uski', 'ne', 'haan', 'acha', 'nai', 'sent',
'photo', 'you', 'kafi', 'gai', 'rhy', 'kuch', 'jata', 'aye', 'ya',
'dono', 'hoa', 'aese', 'de', 'wohi', 'jati', 'jb', 'krta', 'lg',
'rahi', 'hui', 'karna', 'krna', 'gi', 'hova', 'yehi', 'jana', 'jye',
'chal', 'mil', 'tu', 'hum', 'par', 'hay', 'kis', 'sb', 'gy', 'dain',
'krny', 'tou']
```

Fig. 3. Stop words in Roman Urdu.

2.2.3. Normalization and Stemming

A rule-based approach named hashing with the incorporation of lexical strategies for normalizing the Roman Urdu text is utilized by researchers of [35]. We have developed some guidelines to overcome this issue. These guidelines attempt to minimize the use of shared suffixes and infixes of the Roman Urdu words. In Table 1, an indication of the end of a string or suffix is shown by '$' sign, the start of any string by 'Λ' sign, and repetition of any alphabet is '+'.

So, for example, words such as "khamian" (flaws), "achaaiyaan" (goodness), and "kitabain" (books) become "khami", "achai", and "kitab" respectively. One of the interesting things that can be noticed here is that the suffix "an" is removed when the letter "i" is observed before it. Also, expressions such as "taqreebaat" (ceremonies), "chakkay" (Sixes), and "haqooq" (rights) become "taqreb", "chakka", and "haq" respectively. Moreover, repeated letters are reduced to a single representation, as noticed in the normalizations of "qanooon" to "qanon" and "boohatt" to "bohat". Finally, after the application of these guidelines, the normalized text is then standardized using a human-annotated lexical dictionary.

The stemmer used in the data preprocessing step is intended to reduce words to their root form. Though there could be scenarios where the stem does not match with the root, this is still effective since related words tend to belong to the same stem despite the root not being proper itself. There are numerous stemmers for the English language or any other language that is gifted with rich linguistic resources. Examples of such stemmers include the Porter stemmer [36] and the Snowball stemmer [37]. The situation of stemming words for Roman Urdu is far more complex as compared to other languages.

Table 1 provides some examples of lexically normalized words. It is clear that the words in Table 2 have the same sound or pronunciation but with varying spellings. The stem word generation is dependent on a mapping function that is precisely given by f: N → S, where N denotes a finite set of words against which we strive to link plausible stem words that belong to set S. This function of mapping is set to establish the correct stem word S for the term N, boosting the efficiency of the stem word generation. If the mapping function is unsuccessful in identifying a stem word, then the root word is used. So, for ensuring effective search for the stem word, there is separate indexing of each word by means of a hashing function. Therefore, by using the map function, the entire document is exposed to the stemming process to remove any possibilities of inconsistencies or anomalies.

2.3. Model Training and Validation Phase

The data was divided into model's training and validation subsets as part of the dataset partitioning process [38, 39]. In particular, 70% of the dataset was reserved for model training, and the left over 30% was allocated for validation. Further insights into this division are provided in Table 3, revealing that 2845 comments were incorporated for model's training, and 1220 comments were employed for validation purposes.

2.4. Pipeline

A pipeline combines various estimation procedures into a single step, simplifying the ML process [38]. A pipeline involves the progressive implementation of a set of transformers (data modeling), followed

Table 1. Rules for Lexical Normalization.

Sr. No.	String	Replacement
1.	"ian" $	'i'
2.	"niat" $	"ni"
3.	"iy+"	'i'
4.	"ia"	'i'
5.	"ih"	"eh"
6.	"ay"	'e'
7.	"ie" $	'y'
8.	"ee+"	'e'
9.	"es"	'is'
10.	"ar"	'r'

Table 2. Stemming of Roman Urdu.

Roman Words	Stemming	English
siasat, syasat, sayasat	syast	Politics
parhaye, parhaee, parhai	prhai	Study
kitabain, kitaabain, ketabain	kitab	Books
taqreebaat, tareebat, taqrebaat	taqreeb	Ceremony
achaiyaan, achaian, achaiyan	achai	Goodness

Table 3. Training and Testing Sets Description.

Class	Training Set	Test Set	Total
Politics	994	404	1398
Sports	748	344	1092
Education	596	255	851
Religious	507	217	724
Total	2845	1220	4065

by an estimator at the end (ML model) [39]. The transformation stage includes the methods fit() and transform(), while the estimator includes fit() and predict(). Although an estimator always implements fit(), it may not necessarily implement predict(). Briefly, pipelines are designed with fit(), transform(), and predict() capabilities, allowing the entire pipeline to be fitted to the training data and then applied consistently to the test data without repeating each step manually. A pipeline is then built to convert words into vectors, extract features, and fit the model. In this work, function names such as fit(), transform(), and predict() are written with parentheses to indicate that they refer to callable methods (the () denotes that these are functions that can be executed with arguments), as commonly defined in machine learning libraries.

2.5. Feature Extraction

The step of feature selection involves the utilization of TF-IDF weighting scheme, a widely used method in text classification [32, 34]. This scheme assigns specific weights to individual vocabulary terms, belonging to the set V = $\{v_1, v_2 ... v_n\}$, for each document within the text corpus, in order to estimate their importance [7]. These weights, denoted as W = $\{w_1, w_2 ... w_k\}$, aim to reflect the significance of each vocabulary term. Nevertheless, the term frequency (TF) approach's shortcoming lies in its tendency to give higher weights to frequently appearing terms, which could lead to the neglect of crucial terms and subsequent subpar feature selection. Through the following characteristics, size of the feature can be evaluated.

2.6. TF-IDF Vectorizer

Term Frequency Inverse Document Frequency (TF-IDF) approach has broader utilization to transform text into a numerical illustration for prediction after training the ML models [8]. TF-IDF vectorizer takes into account a word's average prominence in a document [32]. When dealing with the most frequently used words, this is a great method. We can penalize them by using it. TF-IDF vectorizer applies a frequency-based weighting factor to the word counts. Table 4 displays the example of feature extraction using TF-IDF. Equation (1) shows the formulation of TF − IDF value in a particular document 'd' for a specific 't' th term:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \quad (1)$$

The term frequency TF (t, d) is for 't' th term in document 'd'. While Inverse Document Frequency for 't' th term throughout the corpus is represented as IDF (t).

2.7. Classification Scheme

Our classification framework employs a diverse set of ML algorithms to classify topics in Roman Urdu text. These algorithms include Multinomial Logistic Regression (MLR), SVM, Naive Bayes, LR, Decision Tree, and our proposed approach based on SGD to explore the classification schemes that most suit the requirements of Roman Urdu text. The framework we have devised for topic classification is rooted in the utilization of the SGD algorithm [40]. This approach is used for the effective classification of topics in multi-class text reviews. The best algorithm emerged here is SGD, which showed the highest accuracy in categorizing Roman Urdu text. SGD is also an iterative optimization algorithm that plays a key role in the training of ML models [41].

It plays a very contributive role in text classification for Roman Urdu text in our research. The algorithm updates model parameters in an

Table 4. Feature Extraction by using TF-IDF.

Sr. No.	Words	TF-IDF
1.	talem	0.53109389
2.	games	0.57735026
3.	cricket	1.69314718
4.	hamesha	0.29207003
5.	reham	0.41802398
6.	khelta	0.70710678
7.	hifazat	0.26017797
8.	insan	0.24783099
9.	afsos	0.28194161
10.	tawajo	0.33762465

iterative manner, where it considers sometimes a single training example or a small batch every time. Inherent with this stochastic nature, it introduces randomness into the process, allowing the algorithm to avoid local minima and enabling quick convergence, especially in the case of large datasets.

This can be given, mathematically, by an update rule for SGD as:

$$\theta_{t+1} = \theta_t - \eta \nabla f(\theta_t; x_i; y_i) \quad (2)$$

Here θ_t represents the model's parameter vector at iteration t. $\nabla f(\theta_t; x_i; y_i)$ denotes the gradient of the loss function, f with respect to θ_t, evaluated on training example (x_i, y_i). While η, a hyperparameter, is the learning rate and decides the step size in the updates of the parameters. In this scenario, x_i shows input feature vector and y_i displays its respective target label for i-th data point used in the computation of the gradient of the loss function.

We implemented an SGD model based on a well-organized pipeline approach. This was composed of two significant parts: the TF-IDF vectorizer and the SGD classifier. The TF-IDF vectorizer played an important role in converting the text data into a numerical representation by assigning words with numeric values according to their weights in TF-IDF. These weights determine the importance of words within the text corpus. The processed data would then serve as an input to the SGD classifier, which utilizes the SGD optimization technique in training a linear classifier for binary classification problems. The "hinge" choice of loss function played an instrumental role in informing the optimization process, while the "l2" penalty contributed toward regularization. The parameter "max_iter" controlled the maximum number of iterations that should result from the optimization process. Through these components and by combining them in a pipeline configuration, we have successfully engineered a robust and flexible SGD model that can be applied to text classification tasks.

3. RESULTS AND DISCUSSION

The main results of our work demonstrate the efficiency of the proposed methodology for Roman Urdu topic classification. Our model, enhanced through the integration of SGD and a custom Roman Urdu stemmer, outperforms well-established models like LR, SVM, NB, DT, and kNN with regularity, which is also supported by prior works that state that quality preprocessing has a great effect on the classification result in low-resource languages [7, 10]. An achieved accuracy of 95 percent reflected the importance of efficient cleaning and TF-IDF transformation, such a relation is also supported through previous studies on Roman Urdu text processing [32]. A number of factors create this improvement. First of all, Roman Urdu-specific stemming rules and customized normalization reduce spelling inconsistencies and noise, thereby mitigating known limitations in previously reported Roman Urdu classification works [10, 42]. Second, TF-IDF is able to provide a sparse feature space that is efficiently handled by the linear SGD classifier, which further supports the previously found observations regarding the efficiency of linear models for short and informal text [7]. Overall, our results confirm that combining language aware preprocessing with an optimized linear classifier leads to more accurate topic categorization and offers strong potential for broader Roman Urdu text classification applications [32].

3.1. Evaluation Metrics

The efficiency of the classifier's is then assessed by using recall, F1-score and precision. Confusion Matrix of our proposed model is also displayed to illustrate the model's functionality.

3.1.1. Accuracy

From the perspective of examining classification models, accuracy is a fundamental metric. The magnitude of successful predictions of a model is an elementary description of its accuracy. Mathematically, we can formulate it as:

$$Accuracy = \frac{No.\ of\ correct\ predictions}{Total\ no.\ of\ predictions} \quad (3)$$

In the context of binary classification, accuracy is simplified in terms of negatives and positives as:

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad (4)$$

3.1.2. Precision and recall

In the context of information extraction, precision

and recall are most commonly applied. The record numbers that have been reclaimed are considered precision, whereas the total record numbers that have been recovered are termed as recall. Meanwhile Precision and recall are inversely related, this highlights the impact of having a reliable classification system to offer context for their variances.

Mathematical interpretation of both terms in classification task is given as:

$$Precision = \frac{True\ Positive}{False\ Positive\ +\ True\ Positive} \quad (5)$$

$$Recall = \frac{True\ Positive}{False\ Negative\ +\ True\ Negative} \quad (6)$$

3.1.3. F1-score

F-measurement, F-score or F1 are similar calculation of the check. The percentage of correctly recognized positive outcomes is a common way to measure precision p, which are divided by percentage of all samples classified as positive, while recall r is the percentage of correctly identified positive results, which are divided by percentage of all examples categorized as positive.

$$F1 - Score = \frac{2 \times Precison \times Recall}{Precison\ +\ Recall} \quad (7)$$

3.1.4. Confusion matrix

Error matrix is another name for confusion matrix, in ML and classification. It is a table that clearly shows where a model makes mistakes. It helps illustrate model's effectiveness or efficiency by comparing its predictions with the original results. The main goal is to analyze the classifier's efficiency. By depicting both predicted and actual values, the confusion matrix offers a visual representation of disparities. This evaluation draws on insights from the confusion matrix, illustrated in Figure 4. Which encompasses metrics for topic classification. Correct predictions are positioned along the diagonal for visualization with the proper labelling of Politics, Sports, Education and Religious classes.

3.2. Topic Classification

In the context of the experimental study, various ML techniques of classification were used for the task. In order to ensure an unbiased comparison, replication of the earlier proposed solutions was carried out for measurement of the efficiency and validity of the ML models. Table 5 shows the experimental results of various solutions of classification with regard to Roman Urdu topic classification tasks. These experimental results clearly show that the proposed solution of SGD with enhancement of the stemmed solution outperformed all other solutions with its enhanced performance capability. In addition, various other solutions using ML also found effective solutions. It is pertinent to note that solutions by LR and by SVM found solutions equivalent to that of our proposed solution for better understanding with various metrics like recall, precision, F1, and accuracy.

Apparently, the class-wise accuracy of analysis models, as shown in Figure 4, clearly reveals that religious class shows better advancement in terms of each recall, F1-measure, precision, and total accuracy. At the same time, there was a slight drop in precision and recall for politics and support classes. Though Table 5 shows the efficiency of our models relative to other models. When comparing, there was a relative low accuracy of 61% by the SVM model developed by Mehmood *et al.* [30] relative to our fine-tuned models. Notably, even the proposed models by us showed better efficiency relative to the deep learning models Recurrent Convolutional Neural Network (RCNN) with an accuracy of 63%. Moreover, the KNN models [32] showed better efficiency relative to precision with a precision of (70%), though relative to recall, it is ineffective

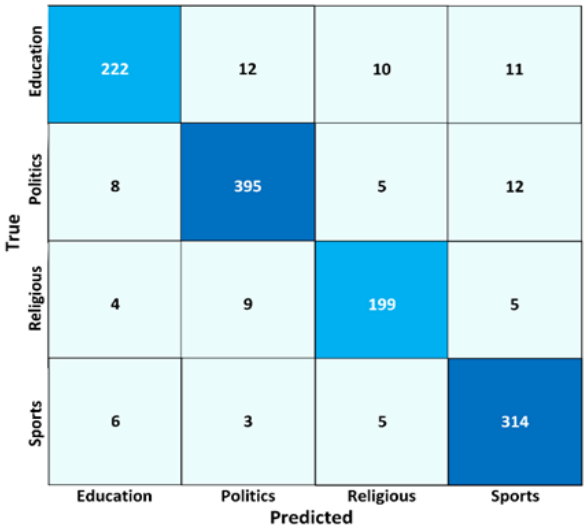


Fig. 4. Confusion Matrix of proposed model (SGD) for Topic Classification of Roman Urdu.

Table 5. Comparative evaluation metrics for proposed and existing models.

Model	Precision	Recall	F1-Score	Accuracy
LR	**0.94**	0.94	0.94	0.94
SVM	**0.94**	0.93	0.94	0.94
Naïve Bayes	**0.90**	0.84	0.86	0.86
Decision Tree	**0.83**	0.83	0.83	0.84
KNN	**0.87**	0.86	0.87	0.87
SVM [30]	0.59	0.58	0.58	0.61
KNN [32]	0.70	0.37	0.48	0.47
LSTM [42]	0.65	0.64	0.65	0.66
Random Forest [43]	0.63	0.61	0.62	0.59
RCNN [44]	0.64	0.62	0.63	0.63
Proposed SGD	**0.95**	0.94	0.94	**0.95**

with low recall that caused the lowest accuracy of 47%. At the same time, the Random Forest models' approach [42] showed relative efficiency relative to NB models, though it gained an accuracy of below 60%, which is unsatisfactory. Additionally, Naive Bayes showed relative efficiency with achieved accuracy of 62%, though it failed to achieve better efficiency relative to the SGD models [43]. At the same time, the efficiency of DT models showed moderate result with the precision of 59%, recall of 57%, and F1-measure of 0.58. Finally, LR and SVM models showed relative efficiency relative to ours with impressive accuracy of 94%. This shows that it is effective relative to regression models as well as classifications.

Figure 5 summarizes the detailed analysis of various models of ML for sentiment classification. This graph is more of a representation of the efficiency of the model in terms of Precision, Recall, F1 Score, and Accuracy of six models: LR, SVM, NB, DT, k-NN, and proposed model. This graph aptly expresses the measures of the models using four bars for each of the models, representing each of the mentioned factors. It is worthy to note that the proposed model gets the maximum number of counts via these factors, highlighting the effectiveness of the proposed model for sentiment analysis.

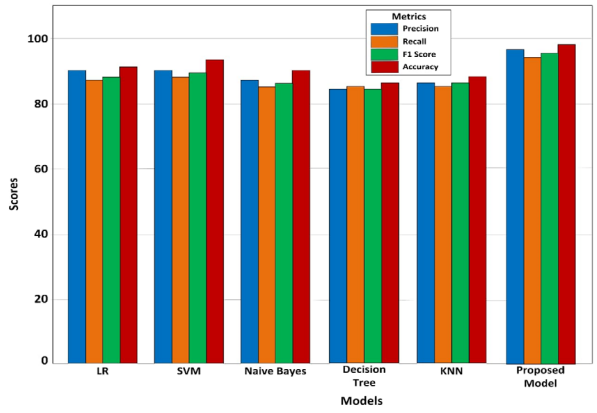


Fig. 5. Comparison of models' performances for Roman Urdu text classification.

4. CONCLUSIONS

In this work, we discussed topic classification for Roman Urdu text with several ML algorithms, including MLR, SVM, NB, Random Forest, DT, and our proposed SGD model supplemented with a Roman Urdu Stemmer. Our approach included extensive data preprocessing and feature extraction so that an optimal classification pipeline was achieved. Among all of the tried models, the SGD model performed best, achieving the maximum accuracy value of 95%. That means the proposed parameter optimization method in the SGD model showed better performance improvement in topic classification for Roman Urdu text. Though promising, we note some limitations of the current study, namely, the adoption of a single train/test split without any evaluation by other measures such as cross-validation that more completely showcases the generalization of the model. Furthermore, further works are needed to address the issues of class imbalance and the application of more advanced methods, such as cross-validation, which could make the results more robust. This study provides validation significant for Roman Urdu topic classification. This could be used in social media monitoring, content categorization, and public discourse studies. Future work will concentrate on refining the SGD model, expanding the dataset, and

integrating additional linguistic features to enhance classification performance further.

5. ACKNOWLEDGMENTS

The authors acknowledge with gratitude the institutional support and resources that enabled the successful execution of this study.

6. CONFLICT OF INTEREST

The authors confirm that they have no conflicts of interest related to this publication. No financial or personal relationships affected how the study was designed, carried out, or interpreted.

7. REFERENCES

1. N. Pangakis and S. Wolken. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, *Mexico City, Mexico*, pp. 113-131 (2024). https://aclanthology.org/2024.nlpcss-1.9.pdf
2. Y. Xie, Z. Li, Y. Yin, Z. Wei, G. Xu, and Y. Luo. Advancing legal citation text classification: A Conv1D-based approach for multi-class classification. *Journal of Theory and Practice of Engineering Science* 4(2): 15-22 (2024).
3. K. Mehmood, D. Essam, K. Shafi, and M.K. Malik. An unsupervised lexical normalization for Roman Hindi and Urdu sentiment analysis. *Information Processing and Management* 57(6): 102368 (2020).
4. G.F. Simons and C.D. Fennig (Eds.). Ethnologue: Languages of the World (20th Edition). *SIL International, Dallas, USA* (2017).
5. G.I. Akabuike and I.C. Onuh. English spelling variations in social media among select students of English language in Nnamdi Azikiwe University. *Ansu Journal of Language and Literary Studies* 5(1): 52-64 (2025).
6. J. Tatemura. Virtual reviewers for collaborative exploration of movie reviews. *Proceedings of the 5th International Conference on Intelligent User Interfaces*, *New Orleans, LA, USA* pp. 272-275 (2000). https://doi.org/10.1145/325737.325870
7. S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys* 54(3): 62 (2021).
8. A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli. A survey on text classification algorithms: From text to predictions. *Information* 13(2): 83 (2022).
9. S. Daud, M. Ullah, A. Rehman, T. Saba, R. Damaševičius, and A. Sattar. Topic classification of online news articles using optimized machine learning models. *Computers* 12(1): 16 (2023).
10. N. Hussain, A. Qasim, G. Mehak, O. Kolesnikova, A. Gelbukh, and G. Sidorov. Hybrid machine learning and deep learning approaches for insult detection in Roman Urdu text. *AI* 6(2): 33 (2025).
11. M.U. Arshad, M.F. Bashir, A. Majeed, W. Shahzad, and M.O. Beg. Corpus for emotion detection on Roman Urdu. *22nd International Multitopic Conference (INMIC 2019), Islamabad, Pakistan* pp. 1-6 (2019). https://doi.org/10.1109/INMIC48123.2019.9022782
12. P. Pakray, A. Gelbukh, and S. Bandyopadhyay. Natural language processing applications for low-resource languages. *Natural Language Processing* 31(2): 183-197 (2025).
13. T. Adimulam, S. Chinta, and S.K. Pattanayak. Transfer learning in natural language processing: Overcoming low-resource challenges. *International Journal of Enhanced Research in Science Technology and Engineering* 11(2): 65-79 (2022).
14. H. Avetisyan and D. Broneske. Large language models and low-resource languages: An examination of Armenian NLP. *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)* pp. 199-210 (2023). https://aclanthology.org/2023.findings-ijcnlp.18.pdf
15. T. Ògúnrèmí, W.O. Nekoto, and S. Samuel. Decolonizing NLP for low-resource languages: Applying Abebe Birhane's relational ethics. *GRACE: Global Review of AI Community Ethics* 1(1): 1-13 (2023). https://ojs.stanford.edu/ojs/index.php/grace/article/view/2584/1546
16. A. Sandu, L.A. Cotfas, A. Stănescu, and C. Delcea. A bibliometric analysis of text mining: Exploring the use of natural language processing in social media research. *Applied Sciences* 14(8): 3144 (2024).
17. Q. Chen, W. Wang, K. Huang, and F. Coenen. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal* 9(12): 9205-9213 (2021).
18. A. Ghafoor, A.S. Imran, S.M. Daudpota, Z. Kastrati, R. Batra, and M.A. Wani. The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*

9: 124478-124490 (2021).

19. V. Kumar, R.S. Singh, M. Rambabu, and Y. Dua. Deep learning for hyperspectral image classification: A survey. *Computer Science Review* 53: 100658 (2024).
20. A. Faheem, F. Ullah, U. Azam, M.S. Ayub, and A. Karim. Part of speech (POS) tagging in Roman Urdu: Datasets and models. *Language Resources and Evaluation* 59(4): 4285-4312 (2025).
21. A. Ilyas, K. Shahzad, and M. Kamran Malik. Emotion detection in code-mixed Roman Urdu-English text. *ACM Transactions on Asian and Low-Resource Language Information Processing* 22(2): 48 (2023).
22. B.A. Chandio, A.S. Imran, M. Bakhtyar, S.M. Daudpota, and J. Baber. Attention-based RU-BiLSTM sentiment analysis model for Roman Urdu. *Applied Sciences* 12(7): 3641 (2022).
23. Z. Nabeel, M. Mehmood, A. Baqir, and A. Amjad. Classifying emotions in Roman Urdu posts using machine learning. *Mohammad Ali Jinnah University International Conference on Computing (MAJICC), (15th-17th July 2021), Karachi, Pakistan* pp. 1-7 (2021). https://doi.org/10.1109/MAJICC53071.2021.9526273
24. I.U. Khan, A. Khan, W. Khan, M.M. Su'ud, M.M. Alam, F. Subhan, and M.Z. Asghar. A review of Urdu sentiment analysis with multilingual perspective: A case of Urdu and Roman Urdu language. *Computers* 11(1): 3 (2021).
25. T.A. Rana, K. Shahzadi, T. Rana, A. Arshad, and M. Tubishat. An unsupervised approach for sentiment analysis on social media short text classification in Roman Urdu. *Transactions on Asian and Low-Resource Language Information Processing* 21(2): 28 (2021).
26. V. Tejaswini, K. Sathya Babu, and B. Sahoo. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Transactions on Asian and Low-Resource Language Information Processing* 23(1): 4 (2024).
27. P.M. Lavanya and E. Sasikala. Deep learning techniques on text classification using Natural Language Processing (NLP) in social healthcare network: A comprehensive survey. *3rd International Conference on Signal Processing and Communication (ICPSC), (13th-14th may 2021), Coimbatore, India* pp. 603-609 (2021). https://doi.org/10.1109/ICSPC51351.2021.9451752
28. M.P. Akhter, Z. Jiangbin, I.R. Naqvi, M. Abdelmajeed, and M.T. Sadiq. Automatic detection of offensive language for Urdu and Roman Urdu. *IEEE Access* 8: 91213-91226 (2020).
29. K. Mehmood, D. Essam, K. Shafi, and M.K. Malik. Discriminative feature spamming technique for Roman Urdu sentiment analysis. *IEEE Access* 7: 47991-48002 (2019).
30. F. Mehmood, M.U. Ghani, M.A. Ibrahim, R. Shahzadi, W. Mahmood, and M.N. Asim. A precisely Xtreme-multi channel hybrid approach for Roman Urdu sentiment analysis. *IEEE Access* 8: 192740-192759 (2020).
31. H.H. Saeed, T. Khalil, and F. Kamiran. Urdu toxic comment classification with PURUTT corpus development. *IEEE Access* 13: 21635-21651 (2025).
32. M. Bilal, H. Israr, M. Shahid, and A. Khan. Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree, and KNN classification techniques. *Journal of King Saud University-Computer and Information Sciences* 28(3): 330-344 (2016).
33. H.P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2(2): 159-165 (1958).
34. S. Tariq, T.A. Rana, and F. Shahzadi. A comparative study of sentiment analysis in Urdu and Roman Urdu: The neglected realms. *CSI Transactions on ICT* 13(2): 193-211 (2025).
35. B. Chandio, A. Shaikh, M. Bakhtyar, M. Alrizq, J. Baber, A. Sulaiman, and W. Noor. Sentiment analysis of Roman Urdu on e-commerce reviews using machine learning. *CMES-Computer Modeling in Engineering and Sciences* 131(3): 1263-1287 (2022).
36. P. Willett. The Porter stemming algorithm: then and now. *Program* 40(3): 219-233 (2006).
37. M.F. Porter. Snowball: A language for stemming algorithms. Snowball Project, *Cambridge, UK* (2001). http://snowball.tartarus.org/texts/introduction.html
38. L. Wratten, A. Wilm, and J. Göke. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature Methods* 18(10): 1161-1168 (2021).
39. Y. Zhou, Y. Yu, and B. Ding. Towards mlops: A case study of ml pipeline platform. *International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, *(23rd-25th October 2020), Beijing, China* pp. 494-500 (2020). https://doi.org/10.1109/ICAICE51518.2020.00102
40. J. Wang and G. Joshi. Cooperative SGD: A unified framework for the design and analysis of local update SGD algorithms. *Journal of Machine*

Learning Research 22(1): 9709-9758 (2021).
41. S.H. Haji and A.M. Abdulazeez. Comparison of optimization techniques based on gradient descent algorithm: A review. *PalArch's Journal of Archaeology of Egypt/Egyptology* 18(4): 2715-2743 (2021).
42. H. Ghulam, F. Zeng, W. Li, and Y. Xiao. Deep learning-based sentiment analysis for Roman Urdu text. *Procedia Computer Science* 147: 131-135 (2019).
43. L.C. Yu, J.L. Wu, P.C. Chang, and H.S. Chu. Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. *Knowledge-Based Systems* 41: 89-97 (2013).
44. Z. Mahmood, I. Safder, R.M.A. Nawab, F. Bukhari, R. Nawaz, A.S. Alfakeeh, and S.U. Hassan. Deep sentiments in Roman Urdu text using recurrent convolutional neural network model. *Information Processing and Management* 57(4): 102233 (2020).

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 289-300 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)693

Pakistan Academy of Sciences

Research Article

Structure Prediction of the *Bombyx mori* Sericin 4 Protein

Khushnudbek Eshchanov*, Dono Babadjanova, and Mukhabbat Baltaeva

Department of Chemistry, Urgench State University, Urgench, Uzbekistan

Abstract: Natural silk (*Bombyx mori*) has been found to contain sericin 1, sericin 2, sericin 3, and sericin 4 proteins. The sequence of amino acid residues in them has also been well studied. However, there is little information on the molecular structure of sericin 4. We conducted studies on the prediction of the sericin 4 molecule's structure using the AlphaFold3 and YASARA computational servers. Molecular dynamics simulations were performed in aqueous solution to evaluate the stability and determine the most favourable conformation of the predicted sericin 4 structure. We mainly used the ProSA-web, Ramachandran Z and Molprobity score to evaluate the predicted structure of sericin 4, and the reliability of the predicted model was determined. The predicted molecular structure serves as a preliminary, yet robust, model of sericin 4.

Keywords: Sericin 4, Silk, Ramachandran Z-Score, Minimum Energy, Solubility, Structure.

1. INTRODUCTION

Proteins extracted from natural silk raw materials are considered as important biomaterials that are the focus of current research. Silk sericin protein is important due to its water solubility, antioxidant properties, biodegradability, and suitability for the preparation of biomaterials for medicine [1-3]. Sericin is often recognised as an "adhesive" protein, enveloping the silk fibroin of *Bombyx mori* and constituting 20–30% of its total mass [4]. In recent years, sericin has been widely employed in nanocomposites, hydrogels, and tissue engineering (for instance, in skin regeneration and wound healing), yielding positive outcomes in its clinical trials [5, 6]. To evaluate and consider the potential uses of sericin, knowledge of its properties, structure, and composition is required [7, 8].

Sericin is a globular protein characterised by the presence of random coils and β-sheet structures. Several external factors, including temperature, humidity, and mechanical stress, can influence the transition of sericin from a random-coil conformation to a β-sheet arrangement. Sericin is highly soluble in water at temperatures of 50 °C and above [9]. This structural transition is thermodynamically linked to a reduction in entropy, and parameters such as pH and ionic strength further affect the kinetics of gel formation [10]. For example, at physiological pH (pH 7), the gelation process can proceed two to three times faster. In contrast, at lower temperatures, the solubility of sericin diminishes, promoting the conversion of random coils into β-sheets and consequently leading to gel formation [11]. Moreover, it has been demonstrated that higher sericin concentrations accelerate the gelation process [12]. Sericin is a hydrophilic protein, distinguished by a high proportion of free hydroxyl (-OH), carboxyl (C=O), and other polar functional groups within its amino-acid residues [13]. Its amino acid composition is dominated by serine (Ser, 37%), glycine (Gly, 16%), and aspartic acid (Asp, 15%), which ensures its high hydrophilicity [14].

It has been found that there are 4 different types of sericin 1, sericin 2, sericin 3, and sericin 4 proteins in *Bombyx mori* silk fiber [4]. These sericin proteins in silk fiber glue together two fibroin fibers. The structure and composition (amino acid sequence) of sericin 1, sericin 2, as well as sericin 3 proteins have been well studied by previous researchers [15, 16]. Komatsu [17] determined the amounts of sericin 1, sericin 2, sericin 3, and sericin 4 proteins in an aqueous solution of sericin

Received: July 2025; Revised: November 2025; Accepted: December 2025
* Corresponding Author: Khushnudbek Eshchanov <xeshchanov77@gmail.com>

extracted from *Bombyx mori* cocoons, and showed that the amount of sericin 4 was 3.1%. The low content of sericin 4 indicates its specific role in interaction with fibroin, it is primarily located in the inner layers and contributes to mechanical strength. This protein serves as a protective and binding component that surrounds the fibroin filaments. Therefore, determining the molecular structure of sericin 4 provides not only insight into its unique physicochemical properties but also a deeper understanding of the surface behaviour of silk-based biomaterials.

The structural uniqueness of sericin 4 is reflected in its amino acid composition and polypeptide chain arrangement. It is rich in polar amino acids such as serine, asparagine, and threonine, which impart a highly hydrophilic character to the protein. As a result, sericin 4 readily interacts with water molecules, thereby contributing to the surface moisture of silk. This property enhances the biocompatibility of silk materials and is particularly important for their biomedical applications, such as in wound dressings, drug delivery systems, and biopolymer films [18].

Information about sericin proteins is also included in the Uniprot and Swiss databases. The Uniprot database accurately describes the 3D molecular structures of sericin proteins and their amino acid sequences [19, 20]. Many scientific publications have been published that fully confirm this information. However, the 3D molecular structure of the sericin 4 protein is poorly understood. It should also be noted that successful work has been carried out to determine the amino acid sequence of sericin 4 [21]. However, the molecular structure of the sericin 4 molecule remains elusive. To some extent, it is possible to predict the formation of the sericin 4 protein to solve this problem. Using the latest AlphaFold3 and RoseTTAFold models, it is possible to predict the approximate 3D structure of sericin 4, which may reveal its β-sheet richness (45%) and potential disulphide bridges [22].

Protein structure prediction relies on the amino acid sequence. The secondary and tertiary structures are inferred from the primary structure. It should be noted, however, that the predicted structure may differ slightly from the protein's actual conformation [23]. The protein chain can adopt numerous conformations due to rotation around the φ and ψ torsion angles at the Cα atom. This conformational freedom contributes to variations in the three-dimensional architecture of proteins. Peptide bonds within the chain are polar, containing carbonyl and -NH- groups that are capable of forming hydrogen bonds. As a result, these groups interact within the protein and play a crucial role in stabilising its structure. Glycine holds a distinctive position in protein architecture, as its minimal side chain grants it increased local flexibility. In contrast, cysteine residues may react with one another to form disulfide bonds, creating cross-links that reinforce the overall stability of the protein. Protein structure is commonly described in terms of secondary structural elements, such as α-helices and β-sheets. Within these motifs, regular hydrogen-bonding patterns arise between the -NH- and C=O groups of neighbouring amino acids, and the residues typically possess similar φ and ψ torsion angles [24].

The development of secondary structural elements enables the hydrogen-bonding potential of peptide bonds to be effectively fulfilled. These secondary structures may be densely packed within the hydrophobic core of a protein, although they may also be found on the surface where the environment is polar. Each amino-acid side chain occupies a finite volume and can engage in only a limited range of interactions with neighbouring residues; such steric and interaction constraints must be carefully considered in molecular modelling and sequence alignment studies [25]. The Ramachandran plot is employed to identify the energetically allowed regions for φ and ψ torsion angles, thereby demonstrating the thermodynamic favourability of β-sheet formation in Sericin 4.

Protein structures can be experimentally identified using methods such as X-ray crystallography, cryo-electron microscopy, and nuclear magnetic resonance (NMR) spectroscopy. However, these approaches are both costly and time-consuming. Over the past six decades, experimental efforts have resolved the structures of approximately 170000 proteins, despite the fact that more than 200 million proteins are known across all forms of life. By 2025, the AlphaFold database had predicted structures for over 214 million proteins, yet certain rare proteins, including sericin 4, have not been fully verified experimentally. Throughout recent decades, numerous computational strategies

have been developed to infer three-dimensional protein structures directly from amino-acid sequences. In the most successful cases, homology-based modelling grounded in molecular evolution has achieved accuracy approaching that of experimental methods, such as NMR spectroscopy [26]. Precise protein-structure prediction holds major importance in fields such as drug discovery and biotechnology [27-29].

Protein structure prediction represents one of the central objectives of computational biology and is closely related to the resolution of the Levinthal paradox. Levinthal's paradox is a conceptual experiment in the context of protein-folding studies, highlighting that protein folding involves identifying the most energetically stable conformation. Exhaustively searching all possible structural conformations to locate the lowest-energy state would be computationally impractical. Yet, in nature, proteins fold extremely rapidly - even when adopting highly complex topologies - indicating that folding proceeds through a rugged energy landscape that guides the molecule efficiently towards a stable configuration [30]. Levinthal also demonstrated that, in cases where the global minimum energy state is not kinetically accessible, proteins may adopt a metastable conformation with slightly higher energy [31]. The most effective approaches in structural bioinformatics tend to be those that build upon existing biological and structural knowledge, rather than attempting to model protein folding entirely from first principles.

When predicting a protein structure or evaluating the quality of a homology model, it is highly beneficial to first examine a large number of experimentally determined structures to gain an understanding of what the actual protein may look like. This comparative insight facilitates a more accurate assessment of the model's reliability and structural validity. Many servers have been created for protein structure prediction. The AlphaFold3 server occupies a special place in protein structure prediction and is the leading server. AlphaFold3 is not limited to single-chain proteins, as it can also predict the structures of RNA, DNK, post-translational modifications, and protein complexes with selected ligands and ions. The AlphaFold3 server allows for structure prediction of proteins consisting of sequences of up to 5000 amino acid residues [32-34].

The Ramachandran Z-score is also regarded as a reliable indicator for the overall assessment of protein structures. Hooft *et al*. introduced this numerical measure, known as the Ramachandran Z-score (Rama-Z), to characterise the distribution of φ and ψ torsion angles in the Ramachandran plot. Its primary significance lies in its ability to indicate the structural credibility of newly determined protein models. The Rama-Z score functions as a global metric, offering an overall evaluation of model quality, although it does not identify local deviations in main-chain geometry. In addition to the single global score, separate Rama-Z values are also computed for β-strands, α-helices, and loop regions. Nevertheless, the global Rama-Z score remains the most informative measure for general structural validation. The value of the Rama-Z score correlates with the proportion of residues that fall within the favourable regions of the Ramachandran plot. Analyses of models resolved at 1.2–5 Å resolution demonstrated that 28% exhibited Rama-Z < -2, 14% had Rama-Z < -3, 0.19% displayed Rama-Z > 2, and only 0.01% had Rama-Z > 3. Based on these observations, a protein structure is considered acceptable when its Rama-Z score lies within the range -3 to 3 [34].

We attempted to demonstrate the 3D molecular structure of sericin 4 based on the latest information on its amino acid sequence, and studies have been conducted. In this work, the potential conformations of sericin 4 are analysed using AlphaFold3 and molecular dynamics (MD) simulations, which may reveal its novel applications as a biomaterial.

2. MATERIALS AND METHODS

Using the AlphaFold3 server, CIF and JSON files were generated (by entering the amino acid residue sequences of sericin 4) for five distinct models of the predicted protein structure. However, the generated models contain structural errors. The model with the fewest errors was identified using dedicated evaluation servers. ProSA-web and Ramachandran Z-scores were employed to provide an overall assessment of the protein structures. The ProSA-web server determines the similarity of protein structures to those characterised by X-ray and NMR analyses; low similarity may indicate the presence of structural errors [25, 26]. The sericin 4 structure was evaluated using MolProbity, one of the most reliable validation tools available. To

achieve favourable validation metrics, defects in the protein structure were minimised using the YASARA minimization server [35]. This server performs an energy minimisation using the YASARA force field. Iterative refinement of the sericin 4 molecular model was performed via this server to optimise the structure. Subsequently, the stability of the sericin 4 model in aqueous solution was investigated through molecular dynamics (MD) simulations. Computations were conducted using the OPLS-AA/L force field and the SPCE water model within the GROMACS MD package, as implemented in the BioExcel Building Blocks Workflows platform. The reliability of the optimised model was reassessed using MolProbity.

3. RESULTS AND DISCUSSION

The presence of four sericin proteins in *Bombyx mori* silk has been reported in the literature [4, 17]. UniProt, Swiss-Prot, and other protein databases contain extensive information on the composition, structure, and other properties of sericin 1, sericin 2, and sericin 3. These databases do not contain information about sericin 4. However, studies have been conducted to determine the structure of sericin 4, and positive results have been reported. Ping Zhao et al. have published research on the sequence of amino acid residues in the sericin 4 molecule. They analysed sericin 4 in terms of its chain segments based on the amino acid residue sequence [20]. This study did not, however, provide information on the complete structure of sericin 4.

The three-dimensional structure of Sericin 4 was predicted using the AlphaFold server based on its amino acid sequence, and comparative analyses were performed to select the most reliable structural model. The sericin 4 protein consists of 2296 amino acid residues, with the largest proportions being Lys (9.7%), Thr (9.4%), Ser (9.4%), Glu (8.9%), and Gly (7.4%). The theoretically calculated isoelectric point (pI) is 6.25. As shown in Figure 1, the following structural models were predicted by the AlphaFold server based on the amino acid residue sequence of sericin 4.

Calculations were carried out using the ProSA-web server to evaluate which of the derived sericin 4 molecular models was the most reliable. ProSA-web determines an overall quality score for the submitted structure. If this score falls outside the range typical of native proteins, the structure may contain errors. The local quality score diagram highlights problematic regions within the model. A three-dimensional molecular representation

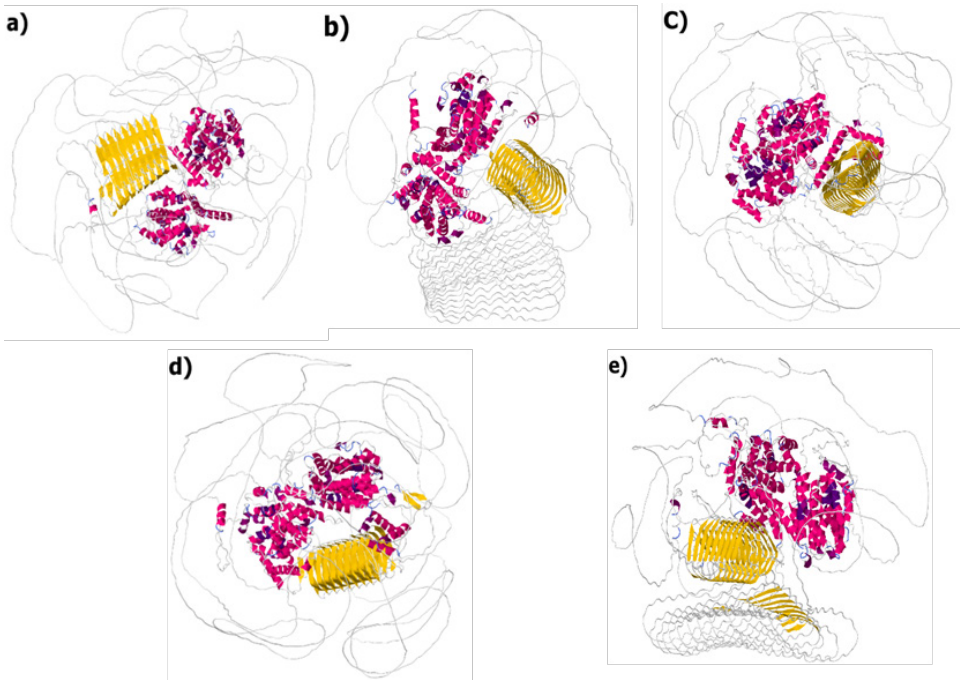


Fig. 1. Models of the sericin 4 molecule created using the AlphaFold3 computational server (Five different molecular models: (a) Compact β-barrel-rich globular model, (b) Extended loop-dominant unfolded-like model, (c) Intermediate partially folded β-sheet model, (d) Globular model with central β-barrel core', and (e) Elongated multi-domain flexible model).

can also be generated to aid in the identification of such areas. ProSA-web is applicable to both low-resolution structures and approximate models obtained during the early stages of structural determination.

The Z-score reflects the overall quality of the model. Its value is displayed on a graph containing the Z-scores of all experimentally determined protein chains, with those derived from different experimental techniques (X-ray and NMR) indicated in distinct colours [25, 26]. The Z-score of a protein is defined as the energy separation between the local fold and the mean value of an ensemble of misfolded folds, expressed in units of the ensemble's standard deviation. It has been reported that calculated Z-scores are generally smaller than experimental values [32, 33].

The results showing the Z-scores for the sericin 4 models generated by the AlphaFold server, and indicating chain segments with relatively higher energy, are presented in Figure 2. The Z-scores for models "a", "b", "c", "d", and "e" of sericin 4 were 0.53, -7.55, -1.52, -6.3, and -8.51, respectively. Examination of these values reveals that the lowest score (-8.51) corresponds to the "e" model structure.

In Figure 2(I-V), illustrating problematic or erroneous regions of the structures, positive values indicate faulty areas. The single-residue energy diagram typically exhibits large fluctuations and is therefore of limited use in model assessment. The greater the number of lines representing negative energy regions, the fewer the structural defects, and thus the more reliable the model. Based on these results, the "e" model of sericin 4 (Z-score -8.51) can be regarded as the most reliable structure.

The sericin 4 models were also evaluated using the global Ramachandran Z score (Rama-Z). The results obtained are presented in Table 1.

The Rama-Z score serves as a global indicator for assessing the overall quality of a protein model and does not provide information on local backbone alignment issues. It is important to highlight that, in addition to the single global Rama-Z value, individual Rama-Z scores are also determined for coils, helices, and β-sheets. A model is generally considered accurate and reliable when its Rama-Z score falls within the range of -3 to 3 [34]. Based on the structural evaluation of sericin 4, it can be observed that the Rama-Z score for the "e" model lies relatively close to -3.

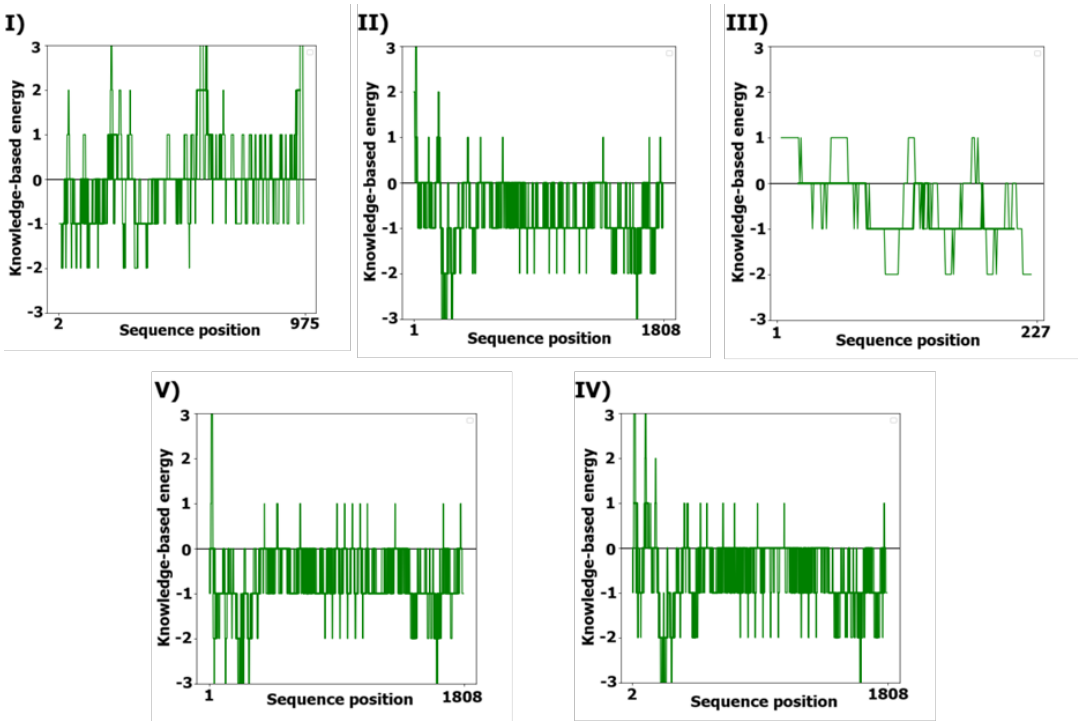


Fig. 2. Diagrams showing high-energy chain segments in models of the sericin 4 molecule: (I) a-Compact β-barrel-rich globular model, (II) b-Extended loop-dominant unfolded-like model, (III) c-Intermediate partially folded β-sheet model, (IV) d-Globular model with central β-barrel core', and (V) e-Elongated multi-domain flexible model.

Table 1. Ramachandran Z score values of sericin 4 molecular models.

Molecular model	Ramachandran Z-score	Side-chain Z-score
a) Compact β-barrel-rich globular model	-6.08	-2.27 ± 0.22
b) Extended loop-dominant unfolded-like model	-4.52	-1.14 ± 0.22
c) Intermediate partially folded β-sheet model	-5.31	-1.80 ± 0.22
d) Globular model with central β-barrel core'	-5.30	-1.91 ± 0.21
e) Elongated multi-domain flexible model	-4.51	-0.89 ± 0.22

The YASARA minimisation server was used to correct energetically unfavourable regions in the "e" model chain of the sericin 4 molecule and to improve its geometry. The YASARA minimisation server is invaluable in protein structure determination, as it provides a realistic impression of the protein's native conformation and demonstrates how to assess the accuracy of the refined model [35]. Using the YASARA minimisation server, the energy of the "e" model of sericin 4 was reduced to its minimum state (Figure 3).

The model was energy-minimised using the YASARA minimisation server for 57 cycles. The Rama-Z score was again used to evaluate the overall structure of the energy-minimised model. The model exhibiting the best Rama-Z score of -2.72 and a minimum energy value of -1069996.7 kJ/mol is presented in Figures 3 and 4. However, according to the MolProbity analysis, among all energy-minimised structures, the model obtained after 51 optimisation cycles in the YASARA program demonstrated the highest quality score, indicating the lowest level of structural errors (Figure 5).

MolProbity is a widely recognised platform for evaluating the geometrical and all-atom quality of three-dimensional macromolecular models, including proteins, nucleic acids, and ligands. It provides detailed validation metrics such as clash scores, Ramachandran plot and rotamer outliers, Cβ deviations, and the overall MolProbity score [36]. The model optimised 51 times achieved a MolProbity score of 1.25, suggesting a high-quality and well-refined structure. The summarised validation results are presented in Table 2.

MolProbity analysis reveals that the protein structure is of high quality: Clashscore 0.45 (99th percentile) and MolProbity score 1.25 (99th percentile) - placing it within the top 1% of PDB entries. Steric clashes and overall geometry are excellent. Ramachandran favoured 88.49% (<98%) - slightly low, but outliers (0.96%) remain within acceptable limits. CaBLAM (6.1%) and CA outliers (3.14%) are acceptable for lower-resolution structures.

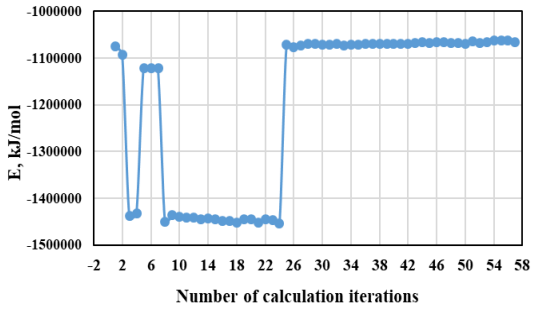


Fig. 3. Minimum energy results of the "e" model of sericin 4 in iterative calculations using the YASARA minimisation server.

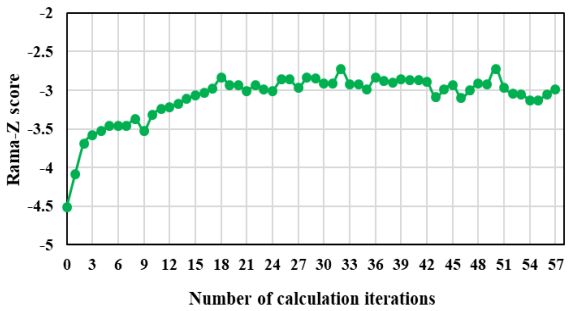


Fig. 4. Rama-Z scores of "e" model sericin 4 that were re-minimised 57 times in the YASARA minimisation server.

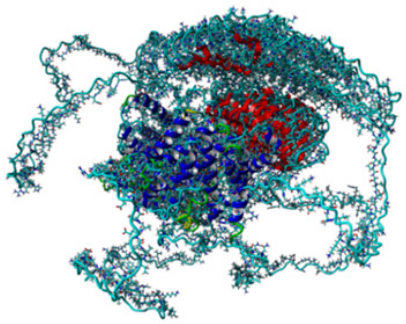


Fig. 5. Energy minimised model of sericin 4 by the YASARA minimisation server.

To mitigate structural inconsistencies observed in the sericin 4 model, the Rosetta Relax refinement was applied [37]. This approach resulted in a notable improvement in the overall structural quality, as evidenced by the evaluation metrics presented in Table 3.

Molecular dynamics (MD) simulation is one of the most powerful computational techniques for investigating the structural and functional properties of proteins at the atomic level. Unlike static crystallographic structures, MD provides a realistic description of the time-dependent dynamic

Table 2. Molprobity analysis of Sericin 4 molecular structures optimised 51 times using YASARA minimisation server.

Clashscore, all atoms	**0.45**	**99th percentile*(N=1784, all resolutions)**
Poor rotamers	0.90%	Goal: <0.3%
Favored rotamers	96.65%	Goal: >98%
Ramachandran outliers	0.96%	Goal: <0.05%
Ramachandran favored	88.49%	Goal: >98%
Rama distribution Z-score	-2.24 ± 0.15	Goal: abs(Z score) < 2
MolProbity score^	**1.25**	**99th percentile* (N=27675, 0Å - 99Å)**
Cβ deviations >0.25Å	0.19%	Goal: 0
Bad bonds:	0.25%	Goal: 0%
Bad angles:	0.39%	Goal: <0.1%
Cis Prolines:	8.70%	Expected: ≤1 per chain, or ≤5%
Twisted Peptides:	0.04%	Goal: 0
CaBLAM outliers	6.1%	Goal: <1.0%
CA Geometry outliers	3.14%	Goal: <0.5%
Chiral volume outliers	0/2720	
Waters with clashes	0.00%	See UnDowser table for details

Table 3. MolProbity analysis of sericin 4 structures refined with Rosetta Relax.

Clashscore, all atoms:	**1.96**	**99th percentile*(N=1784, all resolutions)**
Poor rotamers	0.00%	Goal: <0.3%
Favored rotamers	99.95%	Goal: >98%
Ramachandran outliers	1.05%	Goal: <0.05%
Ramachandran favored	94.07%	Goal: >98%
Rama distribution Z-score	-0.78 ± 0.16	Goal: abs(Z score) < 2
MolProbity score^	**1.36**	**99th percentile* (N=27675, 0Å - 99Å)**
Cβ deviations >0.25Å	0.00%	Goal: 0
Bad bonds:	0.07%	Goal: 0%
Bad angles:	0.13%	Goal: <0.1%
Cis Prolines:	8.70%	Expected: ≤1 per chain, or ≤5%
Twisted Peptides:	0.00%	Goal: 0
CaBLAM outliers	5.4%	Goal: <1.0%
CA Geometry outliers	2.49%	Goal: <0.5%
Chiral volume outliers	0/2720	
Waters with clashes	0.00%	See UnDowser table for details

behaviour of biomolecules. Through MD, the motion of each atom within the protein is computed based on Newtonian mechanics, allowing the exploration of energetically favourable conformations, internal flexibility, and vibrational motions within the system. By evaluating the stability of a protein structure, MD simulation helps to identify the lowest potential energy conformation, which often corresponds to its biologically active form. Therefore, it significantly contributes to energy minimisation and a more accurate representation of the native structural state. Moreover, the simulation enables the analysis of a protein's flexibility, its response to environmental conditions such as temperature and pH, and its interaction mechanisms with ligands or substrates.

Additionally, molecular dynamics complements experimental methods such as X-ray crystallography and NMR spectroscopy by providing time-resolved atomic-level information. The combination of MD data with experimental results allows researchers to construct a more complete and realistic molecular model that explains the functional mechanism, stability, and conformational transitions of the protein. Based on this data, calculations were performed using the MD method for the sericin 4 molecule.

Molecular dynamics (MD) simulations were performed on the BioExcel Building Blocks Workflows platform using the GROMACS MD package with the OPLS-AA/L force field and the SPCE water model [38]. In the simulation setup, a single protein molecule was solvated with 10000 water molecules, 956 Na^+ ions, and 910 Cl^- ions. The net charge of the protein was -46. The simulation lasted for 100 nanoseconds (ns), and the molecular structure was optimised.

The RMSD (Root mean square deviation) graph shows how the shape of the molecule changes over time (Figure 6). In the graph, the RMSD increases from 0 ps to 500 ps and stabilises around 0.4 nm. This indicates that the molecule initially underwent a rapid conformational adjustment (adaptation phase) and subsequently reached a stable state. The RMSD value suggests that the molecule has deviated to some extent from its initial conformation; however, this does not imply instability. Rather, it is associated with the molecule's transition to a new, energetically favourable conformation. Structural stability was achieved after approximately 200-300 ps, and the system remained stable overall.

The radius of gyration (Rg) was also analysed, and the corresponding results are shown in the graph. Rg reflects the compactness or degree of expansion of the molecule. The overall Rg value remained nearly constant at around 4.8 nm. The RgX, RgY, and RgZ values along the three axes also showed very little fluctuation. This indicates that the molecule maintained its general shape, meaning that it neither compressed nor expanded noticeably. Therefore, compactness and structural stability were preserved throughout the entire simulation. Conformational changes were minimal, and the molecule remained in a stable configuration (Figure 7).

The energetic states of sericin 4 were assessed based on the "GROMACS Energies" plot, which shows the potential and total energy (Figure 8). Both energy values remained nearly constant over 500 ps, with only minor fluctuations. The potential energy stabilised around $-16 \cdot 10^6$ kJ/mol, and the

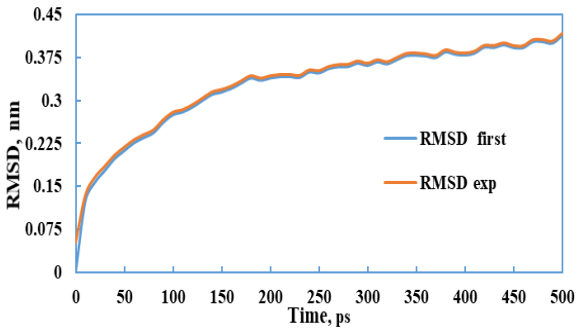


Fig. 6. Root mean square deviation plot of sericin 4 molecule.

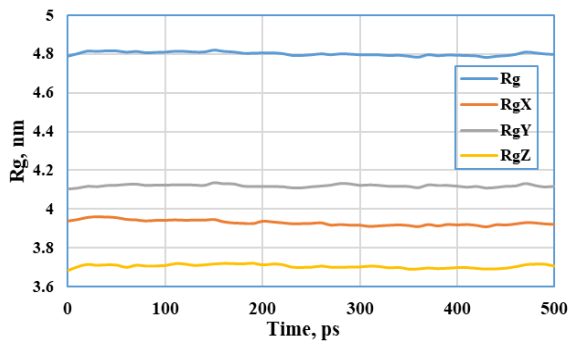


Fig. 7. Stability analysis of sericin 4 based on radius of gyration (Rg).

total energy around -13.5·10^6 kJ/mol. The very small fluctuations indicate that the system reached thermal equilibrium. No significant variations or signs of instability were observed in the results (Figure 9).

The molecular weight, isoelectric point, and other parameters of sericin 4 were determined using the ExPASy (ProtParam) server. The results are presented in Table 4. This server can help to accurately calculate many protein parameters [39-41].

The CamSolpH computational server was used to theoretically study the dependence of the solubility of the improved model of sericin 4 on the pH value of the medium in the YASARA minimization server. CamSolpH provides a solubility profile, where regions with a score greater than 1 indicate highly soluble regions and regions with a score less than -1 indicate poorly soluble regions. The entire sequence is given an overall solubility score. This score can be used to rank different protein variants with high accuracy according to their solubility [42].

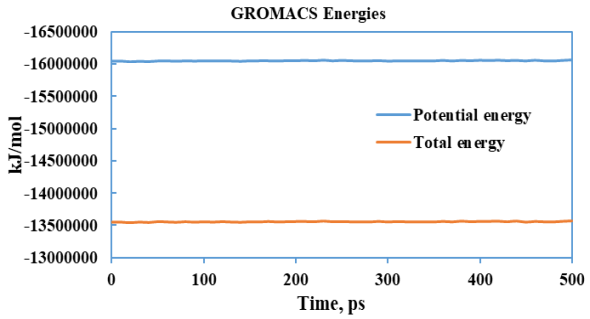


Fig. 8. Potential and total energy stability of the sericin 4 protein during MD simulation.

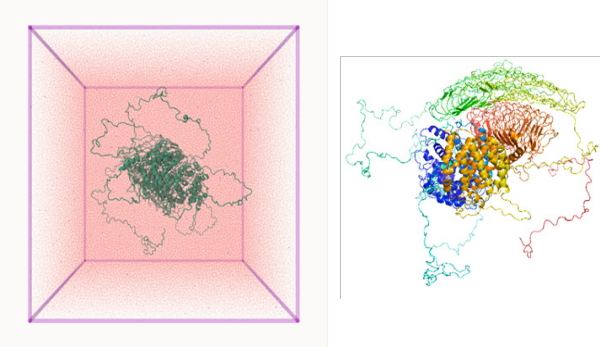


Fig. 9. Conformational state of the Sericin 4 molecule resulting from molecular dynamics simulation.

If we look at Figure 10, the CamSolpH score is greater than 1 in the range of pH values in the solvent (water) medium from 1 to 14. This value theoretically confirms that sericin 4 has good solubility. When comparing the relative solubility at different pH values, it can be seen that the solubility is lowest at pH = 10. It can be assumed that the solubility of sericin 4 is highest in solvents with a pH value of up to 4. However, an increase in solubility can be observed in solvents with a pH value higher than 10.

Table 4. Some calculated parameters of sericin 4.

Molecular model	Parameters
Amino acid number	2296
Molecular weight	254369.63 Da
Isoelectric point	6.25
Extinction coefficients (in water, 280 nm)	175395 $M^{-1}\cdot cm^{-1}$
The instability index	43.88

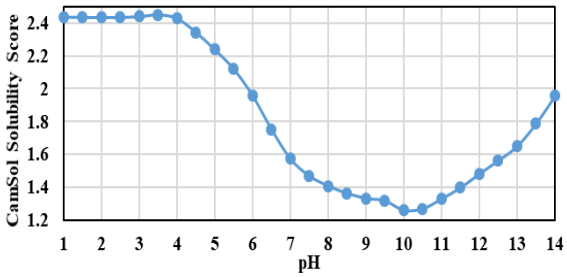


Fig. 10. Solubility index of sericin 4 in solvents (water) with different pH values.

4. CONCLUSIONS

In this study, a comprehensive computational investigation was carried out to predict and analyse the structural and dynamic properties of the sericin 4 protein from *Bombyx mori*. Since no experimental data are available in protein databases, structural prediction was initially performed using the AlphaFold server, yielding five possible molecular conformations. Comparative evaluation through ProSA-web analysis identified the "e" model (elongated multi-domain flexible model) as the most reliable structure, with the lowest Z-score (-8.51). Further refinement using the YASARA minimisation server reduced the overall potential energy of the structure to its minimum state and improved its geometry. Furthermore,

refinement with the Rosetta Relax resulted in an additional improvement of the sericin 4 structure. MolProbity validation confirmed the high quality of the optimised model (MolProbity score 1.36, Clashscore 1.96, 99th percentile, Rama distribution Z-score -0.78 ± 0.16, favored rotamers 99.95%), suggesting that the refined model accurately represents the likely native conformation of sericin 4. Molecular dynamics (MD) simulations performed with GROMACS (OPLS-AA/L force field and the SPCE water model) demonstrated the structural stability of the sericin 4 molecule over a 100 ns trajectory. The RMSD and radius of gyration (Rg) analyses indicated that the protein achieved a stable conformational equilibrium after approximately 200-300 ps, maintaining compactness and structural integrity throughout the simulation. Potential and total energy profiles remained constant, confirming thermal and conformational stability. Solubility profiling performed using the CamSolpH calculation server revealed that sericin 4 exhibits high solubility across a wide pH range (1-14), with a slight decrease observed around pH 10.

Overall, these results provide the first detailed computational insight into the structure, stability, and solubility properties of the sericin 4 protein. The findings not only contribute to filling the existing knowledge gap regarding this protein but also establish a reliable structural model that can serve as a foundation for future experimental studies on its biological functions, material properties, and potential biotechnological applications.

5. CONFLICT OF INTEREST

The authors declare no conflict of interest.

6. REFERENCES

1. R. Suryawanshi, J. Kanoujia, P. Parashar, and S. Saraf. Sericin. A versatile protein biopolymer with therapeutic significance. *Current Pharmaceutical Design* 26(42): 5414-5429 (2020).
2. G. Das, H.S. Shin, E.V.R. Campos, L.F. Fraceto, M.D.P. Rodriguez-Torres, K.C.F. Mariano, D.R. Araujo, F. Fernández-Luqueño, R. Grillo, and J.K. Patra. Sericin based nanoformulations: a comprehensive review on molecular mechanisms of interaction with organisms to biological applications. *Journal of Nanobiotechnology* 19: 30 (2021).
3. A.A. Sarymsakov, S.S. Yarmatov, and K.E. Yunusov. Extraction of Sericin from Cocoons of the Silkworm *Bombyx Mori*, Its Characteristics, and a Dietary Supplement on Its Basis to Prevent Diabetes Mellitus. *Polymer Science Series B* 66(1): 89-96 (2024).
4. M.N. Padamwar and A.P. Pawar. Silk sericin and its applications: A review. *Journal of Scientific & Industrial Research* 63(4): 323-329 (2004).
5. L. Lamboni, Y. Li, and Y. Zhang. Silk sericin-enhanced hydrogel for tissue engineering and wound healing. *Biomaterials Science* 7(11): 4567-4578 (2019).
6. Z. Wang, Y. Zhang, and Y. Yang. Sericin-based biomaterials for regenerative medicine: Current insights and future directions. *Advanced Healthcare Materials* 10(15): 2100456 (2021).
7. C.J. Park, J. Ryoo, C.S. Ki, J.W. Kim, I.S. Kim, D.G. Bae, and I.C. Um. Effect of molecular weight on the structure and mechanical properties of silk sericin gel, film, and sponge. *International Journal of Biological Macromolecules* 119: 821-832 (2018).
8. H. Yun, H. Oh, M.K. Kim, H.W. Kwak, J.Y. Lee, I.Ch. Um, S.K. Vootla, and K.H. Lee. Extraction conditions of Antheraea mylitta sericin with high yields and minimum molecular weight degradation. *International Journal of Biological Macromolecules* 52: 59-65 (2013).
9. H.Y. Kweon, J.H. Yeo, K.G. Lee, Y.W. Lee, Y.H. Park, J.H. Nahm, and C.S. Cho. Effects of poloxamer on the gelation of silk sericin. *Macromolecular Rapid Communications* 21(18): 1302-1305 (2000).
10. Y.N. Jo, B.D. Park, and I.C. Um. Effect of storage and drying temperature on the gelation behavior and structural characteristics of sericin. *International Journal of Biological Macromolecules* 81: 936-941 (2015).
11. R.I. Kunz, R.M.C. Brancalhão, L.D.F.C. Ribeiro, and M.R.M. Natali. Silkworm Sericin: Properties and Biomedical Applications. *BioMed Research International* 2016: 8175701 (2016).
12. R. Aad, I. Dragojlov, and S. Vesentini. Sericin Protein: Structure, Properties, and Applications. *Journal of Functional Biomaterials* 15(11): 322 (2024).
13. Q. Xia, Z. Zhou, C. Lu, D. Cheng, F. Dai, B. Li, P. Zhao, X. Zha, T. Cheng, C. Chai, *et al*. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306(5703): 1937-1940 (2004).
14. H. Yun, M. K. Kim, and H.W. Kwak. Structural characterization and biological activities of sericin from different silkworm races. *International Journal*

of Industrial Entomology 27(1): 135-140 (2013).
15. H. Okamoto, F. Ishikawa, and Y. Suzuki. Structural analysis of sericin genes. Homologies with fibroin gene in the 5'flanking nucleotide sequences. *Journal of Biological Chemistry* 257(24): 15192-15199 (1982).
16. B. Kludkiewicz, Y. Takasu, R. Fedic, T. Tamura, F. Sehnal, and M. Zurovec. Structure and expression of the silk adhesive protein Ser2 in *Bombyx mori*. *Insect Biochemistry and Molecular Biology* 39(12): 938-946 (2009).
17. K.I. Komatsu. Chemistry and structure of silk. *Jarq-Japan Agricultural Research Quarterly* 13(1): 64-72 (1979).
18. Y. Takasu, H. Yamada, and K. Tsubouchi. Isolation of three main sericin components from the cocoon of the silkworm, *Bombyx mori*. *Bioscience, Biotechnology, and Biochemistry* 66(12): 2715-2718 (2002).
19. Y. Takasu, H. Yamada, T. Tamura, H. Sezutsu, K. Mita, and K. Tsubouchi. Identification and characterization of a novel sericin gene expressed in the anterior middle silk gland of the silkworm *Bombyx mori*. *Insect Biochemistry and Molecular Biology* 37(11): 1234-1240 (2007).
20. Z. Dong, K. Guo, X. Zhang, T. Zhang, Y. Zhang, S. Ma, H. Chang, M. Tang, L. An, Q. Xia, and P. Zhao. Identification of *Bombyx mori* sericin 4 protein as a new biological adhesive. *International Journal of Biological Macromolecules* 132: 1121-1130 (2019).
21. D.W. Mount (Ed.). Bioinformatics: Sequence and Genome Analysis (2nd Edition). *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, United States of America* (2004).
22. P. Chakrabarti and D. Pal. The interrelationships of side-chain and main-chain conformations in proteins. *Progress in Biophysics and Molecular Biology* 76(1-2): 1-102 (2001).
23. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583 (2021).
24. R.H. Yousif, H.A. Wahab, K. Shameli, and N.B. Khairudin. Exploring the Molecular Interactions between Neoculin and the Human Sweet Taste Receptors Through Computational Approaches. *Sains Malaysiana* 49(3): 517-525 (2020).
25. R.F. Service. The game has changed. AI triumphs at protein folding. *Science* 370(6521): 1144-1145 (2020).
26. H.A. Mesrabadi, K. Faez, and J. Pirgazi. Drug-target interaction prediction based on protein features, using wrapper feature selection. *Scientific Reports* 13: 3594 (2023).
27. K.K. Barani, M. Mohammadi, M. Ghambarian, and Z. Azizi. Fe_3O_4/ZnO@ MWCNT promoted green synthesis of biological active of new azepinooxazepine derivatives: Combination of experimental and theoretical study. *Polycyclic Aromatic Compounds* 44(1): 528-554 (2024).
28. H.A. Guvenilir and T. Doğan. How to approach machine learning-based prediction of drug/compound–target interactions. *Journal of Cheminformatics* 15: 16 (2023).
29. L.N. David, M.C. Michael, and L.L. Albert (Eds.). Lehninger Principles of Biochemistry (7th Edition.). *W.H. Freeman, New York, USA* (2017).
30. P. Hunter. Into the fold. Advances in technology and algorithms facilitate great strides in protein structure prediction. *EMBO Reports* 7(3): 249-252 (2006).
31. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A.J. Ballard, J. Bambrick, S.W. Bodenstein, D.A. Evans, Ch. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A.I. Cowen-Rivers, A. Cowie, M. Figurnov, F.B. Fuchs, H. Gladman, R. Jain, Y.A. Khan, C.M.R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E.D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J.M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold3. *Nature* 630: 493-500 (2024).
32. M. Wiederstein and M.J. Sippl. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* 35: W407-W410 (2007).
33. M.J. Sippl. Recognition of Errors in Three-Dimensional Structures of Proteins. *Proteins* 17(4): 355-362 (1993).
34. O.V. Sobolev, P.V. Afonine, N.W. Moriarty, M.L. Hekkelman, R.P. Joosten, A. Perrakis, and P.D. Adams. A global Ramachandran score identifies protein structures with unlikely stereochemistry. *Structure* 28(11): 1249-1258 (2020).

35. E. Krieger, K. Joo, J. Lee, J. Lee, S. Raman, J. Thompson, M. Tyka, D. Baker, and K. Karplus. Improving physical realism, stereochemistry, and side-chain accuracy in homology modeling: Four approaches that performed well in CASP8. *Proteins* 77(9): 114-22 (2009).
36. L. Zhang and J. Skolnick. What should the Z-score of native protein structures be? *Protein Science* 7(5):1201-1207 (1998).
37. S. Lyskov, F.C. Chou, S.Ó. Conchúir, B.S. Der, K. Drew, D. Kuroda, J. Xu, B.D. Weitzner, P.D. Renfrew, P. Sripakdeevong, B. Borgo, J.J. Havranek, B. Kuhlman, T. Kortemme, R. Bonneau, J.J. Gray, and R. Das. Serverification of Molecular Modeling Applications: The Rosetta Online Server That Includes Everyone (ROSIE). *PLoS One* 8(5): e63906 (2013).
38. G. Bayarri, P. Andrio, A. Hospital, M. Orozco, and J.L. Gelpí. BioExcel Building Blocks Workflows (BioBB-Wfs), an integrated web-based platform for biomolecular simulations. *Nucleic Acids Research* 50(W1): W99–W107 (2022).
39. M.R. Wilkins, E. Gasteiger, A. Bairoch, J.C. Sanchez, K.L. Williams, R.D. Appel, and D.F. Hochstrasser. Protein identification and analysis tools in the ExPASy server. *Methods in Molecular Biology* 112: 531-552 (1999).
40. M. Naveed, K. Javed, T. Aziz, A. Zafar, M. Fatima, H.M. Rehman, A.A. Khan, A.S. Alamri, W.F. Alsanie, and M. Alhomrani. Innovative Approach of High-Throughput Screening in the Drug Discovery Quest for Chronic Bronchitis Treatment. *Journal of Computational Biophysics and Chemistry* 24(02): 173-187 (2025).
41. M. Naveed, I. Ali, T. Aziz, A. Saleem, Z. Rajpoot, S. Khaleel, A.A. Khan, M. Al-harbi, and T.H. Albekairi. Computational and GC-MS screening of bioactive compounds from Thymus Vulgaris targeting mycolactone protein associated with Buruli ulcer. *Scientific Reports* 15(1): 131 (2025).
42. M. Oeller, R. Kang, R. Bell, H. Ausserwöger, P. Sormanni, and M. Vendruscolo. Sequence-based prediction of pH-dependent protein solubility using CamSol. *Briefings in Bioinformatics* 24(2): 1-7 (2023).

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 301-312 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)697

Pakistan Academy of Sciences

Research Article

A Flexible-Scalar Splitting Iterative Method for Linear Inverse Problems with Complex Symmetric Matrix

Ruiping Wen[1], Dongqi Li[1], Zubair Ahmed[2], Jinrui Guan[1], and Owais Ali*[2, 3]

[1]School of Mathematics and Statistics, Taiyuan Normal University, Jinzhong, China
[2]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan
[3]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan

Abstract: This paper introduces a flexible scalar-splitting (f-SCSP) iterative scheme and examines its convergence properties. The approach also yields a straightforward matrix-splitting preconditioner for the original linear system. To confirm the theoretical results and evaluate practical performance, comprehensive numerical examinations are performed on various test cases. The findings indicate that the proposed method is practical, reliable, and more efficient than existing techniques for handling demanding classes of complex symmetric linear systems.

Keywords: Inverse Problem, Complex Symmetric Matrix, Splitting; Iterative Method, Flexible-scalar.

1. INTRODUCTION

We focus on the iterative resolution of linear systems.

$$Ax = b \tag{1}$$

where $A \in C^{n\times n}$ and $x,\ b \in C^n$. In Equation (1), $A = W + iT$ is a matrix which is non-Hermitian and symmetric $(A \neq A^*, A = A^T)$ with $W, T \in \mathbb{R}^{n\times m}$ are real and symmetric, and W and T are positive definite and positive semidefinite matrices, respectively. In this text, the imaginary quantity iota, $i^2 = -1$, is denoted by the symbol i. Let there be a splitting $A = M - N$ of the matrix $A \in \mathbb{C}^{n\times n}$ i.e., $M \in \mathbb{C}^{n\times n}$ is nonsingular and $N \in \mathbb{C}^{n\times n}$. This splitting gives rise to a fixed-point iterative method of the following form.

$$x^{k+1} = M^{-1}Nx^k + M^{-1}b, \qquad k = 0,1,2,\ldots \tag{2}$$

where $x^0 \in \mathbb{C}^n$ is a given starting vector.

Systems corresponding to Equation (1) appear frequently throughout computational science and in numerous branches of engineering, where they form a core component of many modelling and simulation tasks. A few notable examples include Diffuse Optical Tomography (DOT); very helpful for small animal imaging, breast cancer detection, and functional brain imaging [1]. Because of the nature of light propagation in scattering media and the usage of complex coefficients to simulate absorption and diffusion, the mathematical modelling and numerical computation required in DOT frequently result in complex symmetric linear systems. When time-dependent PDEs are treated with FFT-driven schemes, the resulting discretisations commonly lead to complex symmetric linear algebraic systems, particularly in frequency-domain formulations or in spectral and pseudo-spectral frameworks [2].

Advanced scientific applications in structural dynamics, especially those involving damping, frequency-domain analysis, or non-proportional damping models, the governing equations lead to complex symmetric linear systems [3]. Lattice Quantum Chromo Dynamics (Lattice-QCD)

Received: July 2025; Revised: November 2025; Accepted: December 2025
* Corresponding Author: Owais Ali <owais@g.ecc.u-tokyo.ac.jp>

[4] is a computational approach for examining QCD. Complex symmetric linear systems emerge naturally in various formulations of Lattice-QCD, particularly in fermion discretization such as staggered fermions or twisted mass fermions [5]. Numerical computations in molecular scattering is a crucial subject in quantum chemistry, chemical physics, and dynamics. The foundational theory relies on quantum scattering theory, resulting in extensive linear algebraic systems that are frequently complex and occasionally symmetrical under certain conditions [6].

Recently, Ahmed *et al.* [7] and Kanwal *et al.* [8] suggested that if the forward operator A is symmetric, iterative over-relaxation can solve (1) efficiently. Axelsson and Kucherov [9] presented an iterative method for real matrices, Benzi and Bertaccini [10] proposed a block preconditioning for real-valued iterative algorithms, Bai [11] and Bai *et al.* [12, 13] introduced a modified Hermitian and skew-Hermitian splitting (MHSS) as well as preconditioned-MHSS (PMHSS) iterative methods and Wang *et al.* [14] improved the PMHSS method. Various preconditioning techniques have been developed to enhance the convergence rate of these iterative methods. For instance, Salkuyeh *et al.* [15], Hezari *et al.* [16], Axelsson and Salkuyeh [17], Xie and Li [18], Xiang and Zhang [19], and Salkuyeh [20], Zhao *et al*. [21] put forward a Single-Step-MHSS method (SMHSS) and its variants with a flexible-shift (f-SMHSS). Wen *et al*. [22, 23] also suggested some iterative methods and respective preconditioning techniques. Vorst and Melissen [24], Freund [25], while, Bunse-Gerstner and Stöver [26] presented the conjugate gradient-type methods; Clements *et al*. [27] introduced Krylov-type methods. In particular, Hezari *et al*. [28] proposed the Scale-Spliting (SCSP) method employing a scaling approach. Later Salkuyeh [29] suggested a two-step SCSP method, while Salkuyeh and Siahkolaei [30] introduced a two-parameter SCSP (TSCSP). Zheng *et al.* [31] also introduced a double-step scale splitting iterative method. Li *et al*. [32, 33] put forward a dual-parameter double-step splitting iteration method, and two iterative methods with quasi-combining real and imaginary parts. However, the scaled parameters mentioned above are given in advance. Motivated by the optimization models given by Zhao *et al*. [21], this study introduced a flexible-scalar strategy based on the SCSP iterative method, which the scaled parameters α are determined by minimizing the residuals at each iteration.

Following we present the essential notations. The set of $p \times p$ real (complex) arrays and the p -dimensional real (complex) vector space are represented as $\mathbb{R}^{p\times p}$ and $\mathbb{R}^{p}$ ($\mathbb{C}^{p\times p}$ and $\mathbb{C}^{p}$) respectively. The conjugate and transpose of a matrix or a vector x is x^{*} and x^{T} repectively. A matrix $A \in \mathbb{C}^{p\times p}$ ($A \in \mathbb{R}^{p\times p}$) is said to be Hermitian (symmetric) positive definite (or semidefinite), denoted by $A > 0$ (or ≥ 0); if it is Hermitian (or symmetric) and for all $x \in \mathbb{C}^{n},\ x \neq \bar{0},\ x^{*}Ax > 0$ ($x^{*}Ax \geq 0$) holds true. The real and imaginary parts of a complex number x are denoted by $\mathfrak{Re}(x)$ and $\mathfrak{Im}(x)$, respectively. $\rho(A)$ is used to represent the spectral radius of a matrix A and $\Sigma(A)$ represents the spectrum set of the matrix. The condition number of a matrix A is denoted by $\kappa(A)$. The splitting of A , defined as $A = M - N,$ is said to be convergent if $\rho(M^{-1}N) < 1$.

A broad range of preconditioning strategies has been introduced in past to accelerate the convergence behavior of such iterative schemes. For instance, a double-step scale splitting iterative method employing a scaling approach given by Salkuyeh and Siahkolaei [30]. By multiplying two parameters $(\alpha - i)$ and $(1 - i\alpha)$ both sides of the Equation (1), two equivalent systems can be respectively yielded, i.e., $(\alpha - i)Ax = (\alpha - i)b$ and $(1 - i\alpha)Ax = (1 - i\alpha)b$, where α is a real positive number. Then two fixed-point equations can be generated as follows:

$$\big((\alpha W + T) + i(\alpha T - W)\big)x = (\alpha - i)b\,, \qquad (3)$$

$$\big((\alpha W + T) + i(\alpha T - W)\big)x = (1 - i\alpha)b. \qquad (4)$$

Zheng *et al.* [31] expanded on the PMHSS iterative method, suggested by Bai *et al*. [13], and proposed the following alternative iterative scheme:

$$\begin{cases} (\alpha W + T)x^{k+\frac{1}{2}} = i(W - \alpha T)x^{k} + (\alpha - i)b \\ (\alpha W + T)x^{k+\frac{1}{2}} = i(W - \alpha T)x^{k} + (1 - i\alpha)b \end{cases}, \ k = 0,1,2,\ldots$$

whereas the Equations (3) and (4) are in fact two preconditioned systems in Equation (2) when $P = (\alpha - 1)I$ and $P = (1 - i\alpha)I$, that is to say, the preconditioned matrices are both the scalar matrices. Equations (3) and (4) are one when $\alpha = 1$, therefore, the alternation of the DSS iterative method

was only carried out in twins of two preconditioned systems. This work focuses on linear systems whose coefficient matrices are complex symmetric yet not Hermitian. We focus on the scaled preconditioned splitting iterative methods generally and consider the systems in Equation (2) when $P = (\alpha - \beta i)I$ with α, β are both real numbers in this study.

2. MATERIALS AND METHODS

To provide context and completeness, this section begins with a brief overview of existing methods for solving linear systems whose coefficient matrices are complex symmetric but non-Hermitian, as in Equation (1). We then introduce the Flexible-Scalar Splitting (f-SCSP) scheme.

2.1. The Relevant Methods

2.1.1. MHSS method [12, 13]:

The MHSS iteration method: Let $x^{(0)} \in \mathbb{C}^n$ be an initial guess. For $k = 0, 1, 2, \cdots$, until $\{x^{(k)}\}$ converges, compute $x^{(k+1)}$ according to the following sequence:

$$\begin{cases} (\alpha I + W)x^{k+\frac{1}{2}} = (\alpha I - iT)x^k + b, \\ (\alpha I + T)x^{k+1} = (\alpha I + iW)x^{k+\frac{1}{2}} - ib, \end{cases}$$

where α is a given positive constant.

2.1.2. The SMHSS and f-SMHSS methods [21]:

(1) The SMHSS iteration method: Let $x^{(0)} \in \mathbb{C}^n$ be an initial guess. For $k = 0, 1. 2. \cdots$, until $\{x^{(k)}\}$ converges, compute $x^{(k+1)}$ according to the following sequence $(\alpha I + W)x^{k+1} = (\alpha I - iT)x^k + b.$

(2) The f-SMHSS iteration method: Let $x^{(0)} \in \mathbb{C}^n$ be an initial guess, for $\varepsilon > 0$, $k = 0, 1, 2, \cdots$, until $\{x^{(k)}\}$ converges, the single-step iteration formula for computing the next $x^{(k+1)}$ is as follows.

Step 1: Compute $r_k = b - Ax_k$.

Step 2: Solve the equation
$(\alpha_{k+1} I + W)x^{k+1} = (\alpha_{k+1} I - iT)x^k + b,$
where the flexible shift α_{k+1} is the solution to the following optimization problem: $\min_{\alpha} \|(\alpha I + W)^{-1} r_{k+1}\|_2^2$, with $r_{k+1} = b - Ax_{k+1}$.

Step 3: If $\|r_k\|_2 \leq \epsilon$, stop; otherwise, set $k = k + 1$ and return to Step 1.

2.1.3. The scale-splitting (SCSP) method [28]:

Let α be a real positive constant and the matrix $\alpha W + T$ be nonsingular. By multiplying the complex number $(\alpha - i)$ through both sides of Equation (1), the following equivalent system can be obtained.

$$A_\alpha x = (\alpha - i)b \tag{5}$$

Where $A_\alpha = (\alpha W + T) + i(\alpha T - W)$. By rewriting it as the system of fixed-point equations:
$(\alpha W + T)x = i(W - \alpha T)x + (\alpha - i)b,$
the SCSP iteration method can be summarized as follows.

The SCSP iteration method: Let $x^{(0)} \in \mathbb{C}^n$ be an initial guess. For $k = 0, 1, 2, \cdots$, until $\{x^{(k)}\}$ converges, compute $x^{(k+1)}$ according to the following sequence:

$$(\alpha W + T)x^{k+1} = i(W - \alpha T)x^k + (\alpha - i)b, \tag{6}$$

where α is a given positive constant.

2.2. Proposed Iterative Method: The Flexible-Scalar Splitting (f-SCSP)

The variant system can be obtained by multiplying the complex number $\alpha - i$,
$[(\alpha W + T) - i(W - \alpha T)]x = (\alpha - i)b.$
To use the flexible-scalar strategy, the f-SCSP method is formulated as follows:

$$(\alpha_{k+1} W + T)x^{k+1} = i(W - \alpha_{k+1} T)x^k + (\alpha_{k+1} - i)b \tag{7}$$

where,

$$\alpha_{k+1} = \arg\min_{\alpha} \frac{1}{2} r_k^* (\alpha W + T)^{-1} r_k \tag{8}$$

with $r_k = b - Ax^k$, $k = 0,1,2, \ldots$.
Remark: In fact, the exact solutions of the quadratic programming models in Equation (8) can be given theoretically by simple computing. To avoid the tedious computation of $(\alpha_k W + T)^{-1}$, we can use the inexact line search to find the approximations of α.
In matrix-vector form, the scheme presented in Equation (7) can be equivalently rewritten as:

$$x^{k+1} = T_{\alpha_{k+1}} x^k + G_{\alpha_{k+1}} b, \quad k = 0,1,2, \ldots \tag{9}$$

where,

$$T_{\alpha_{k+1}} = i(\alpha_{k+1}W + T)^{-1}(W - \alpha_{k+1}T), \text{ and } G_{\alpha_{k+1}} = (\alpha - i)(\alpha W + T)^{-1} \quad (10)$$

Here, $T_{\alpha_{k+1}}$ is the iteration matrix of the f-SCSP method. In fact, Equation (9) is also generated by the splitting, $A_{\alpha_k} = M_{\alpha_k} - N_{\alpha_k}$, with

$$M_{\alpha_k} = \frac{\alpha + i}{\alpha^2 + 1}(\alpha W + T), \quad \text{and } N_{\alpha_k} = \frac{-1 + i\alpha}{\alpha^2 + 1}(W - iT)$$

Moreover,

$T_{\alpha_{k+1}} = M_{\alpha_{k+1}}^{-1} N_{\alpha_{k+1}}$, and M_{α_k} can be identified as a preconditioner to all linear systems of type Equation (1).

Consequently, the preconditioned system can be expressed as follows.

$$M_{\alpha_k}^{-1} A x = M_{\alpha_k}^{-1} b. \quad (11)$$

We now investigate the optimal parameter selection and the spectral radius characteristics of the iteration matrix, and assess the convergence behavior of the previously described f-SCSP method.

Theorem 2.1: Let be a non-Hermitian but symmetric matrix $A = W + iT \in \mathbb{C}^{n\times n}$, $(A \neq A^*,\ A = A^T)$ with both $W, T \in \mathbb{R}^{n\times n}$ being symmetric, W and T being both positive definite positive. Let α be positive real numbers and $\lambda_{\min}$ and $\lambda_{\max}$ be the extremal eigenvalues of the matrix $W^{-1}T$. Then the following statements hold true:

(i) In the f-SCSP method, the upper bound of the spectral radius $\rho(T_{\alpha_k})$ is:

$$\delta_{\alpha_k} = \max\left\{\frac{-\alpha_k\lambda_{\min}}{\alpha_k+\lambda_{\min}}, \frac{\alpha_k\lambda_{\max}}{\alpha_k+\lambda_{\max}}\right\} \quad (12)$$

(ii) The sequence $\{x^k\}$ produced by Method 2.1

$$\begin{cases} \dfrac{1-\lambda_{\min}}{1+\lambda_{\min}} < \alpha_k < \dfrac{1-\lambda_{\max}}{1+\lambda_{\max}}, & \lambda_{\max} \in (1, +\infty) \\ \alpha_k > \dfrac{1-\lambda_{\min}}{1+\lambda_{\min}}, & \Sigma(W^{-1}T) \subset [0, 1] \end{cases}$$

In particular, the iterative scheme presented in Equation (6) is convergent if α for the case that T is a positive semidefinite matrix.

Proof (i): By Equation (12) and direct calculations, we have:

$$\begin{aligned} &\rho(T_{\alpha_k}) = \rho\left(i(\alpha_k W + T)^{-1}(W - \alpha_k T)\right) \\ &\le \|i(\alpha_k W + T)^{-1}(W - \alpha_k T)\|_2 \\ &\le \|(\alpha_k W + T)^{-1}\|_2 \|W - \alpha_k T\|_2 \\ &= \|(\alpha_k I + W^{-1}T)^{-1}\|_2 \|I - \alpha_k W^{-1}T\|_2 \\ &= \max_{\lambda\in\Sigma(W^{-1}T)} \left|\frac{-\alpha_k\lambda}{\alpha_k + \lambda}\right|. \end{aligned}$$

In the last step, the equality holds since $W^{-1}T$ is a symmetric positive definite matrix, and then so is $(\alpha I + W^{-1}T)^{-1}$.

It is known that λ is positive. By introducing the following function:

$$f(\lambda) = \frac{-\alpha\lambda}{\alpha + \lambda},$$

it is obtained that $f(\lambda)$ is a decreasing function

with respect to λ since $f'(\lambda) = -\frac{1+\alpha^2}{(\alpha+\lambda)^2} < 0$.

Thus Equation (12) provides the upper bound of $\rho(T_{\alpha_k})$.

Proof (ii): For the case that $\lambda_{\max} > 1$, $\delta_{\alpha_k} < 1$ is equivalent to $\alpha > \frac{1-\lambda_{\min}}{1+\lambda_{\min}}$ by simple calculations.

And then $\rho(T_{\alpha_k}) < 1$, so the sequence $\{x^k\}$ produced by the f-SCSP method converges to the unique solution to Equation (1) for any initial guess x^*.

For the case that $\Sigma(W^{-1}T) \subset [0,\ 1]$, then $\lambda_{\max} < 1$ at that time. Thus, $\delta_{\alpha_k} < 1$ is only equivalent to $\alpha_k > \frac{1-\lambda_{\min}}{1+\lambda_{\min}}$.

It is well-known that $\lambda_{\min} = 0$ if T is a positive semidefinite matrix. And then $\rho(T_{\alpha_k}) \le \alpha^{-1}$, the iterative scheme in Equation (6) is convergent if $\alpha > 1$. The proof is completed.

Corollary 2.1: Assuming the conditions of Theorem 2.1 hold, the optimal the parameters α that minimises the upper bound δ_{α_k} of the spectral radius $\rho(T_{\alpha_k})$ is given by:

$$\alpha = \frac{1-\lambda_{\min}\lambda_{\max}+\sqrt{(1+\lambda_{\min}^2)(1+\lambda_{\max}^2)}}{\lambda_{\min}+\lambda_{\max}} \quad (13)$$

A similar proof is presented in [28, theorem 1], which is omitted here.

Theorem 2.2: Let be a non-Hermitian, symmetric matrix $A = W + iT \in \mathbb{C}^{n\times n}$, with $W, T \in \mathbb{R}^{n\times n}$ being both symmetric, also, W being positive-definite and T positive definite or semidefinite. Then $\rho(T_{\alpha_k}) < 1$ if for all $x \in \mathbb{C}^n$, it holds that $\alpha > \frac{x^*Wx - x^*Tx}{x^*Wx + x^*Tx}$.

Proof: Let an eigenvalue of the matrix T_{α_k}

be λ with the corresponding eigenvector x, i.e., $M_{\alpha_k}^{-1}N_{\alpha_k}x = \lambda x$, which means, $\lambda(\alpha W + T)x = i(W - \alpha T)x$. Then we have from the assumptions that:

$$|\lambda| = \left|\frac{x^*Wx - \alpha x^*Tx}{\alpha x^*Wx + x^*Tx}\right|$$

We obtain $\alpha > \frac{x^*Wx - x^*Tx}{x^*Wx + x^*Tx}$ by direct calculations under $|\lambda| < 1$. The theorem is proved.

Remark: Theorem 2.2 implies that all eigenvalues of the matrix T_{α_k} lie along the imaginary axis.

The last of this section, a property of the matrix $M_{\alpha_k}^{-1}A$ can be given.

Theorem 2.3: Let $A = W + iT \in \mathbb{C}^{n\times n}$ be a non-Hermitian but symmetric matrix ($A \neq A^*$, $A = A^T$) with $W, T \in \mathbb{R}^{n\times n}$ be real, symmetric, and W being positive-definite and T positive definite or semidefinite. Assuming that μ is any eigenvalue of the matrix $M_{\alpha_k}^{-1}A$ defined by Theorem (2.2), the $\Re e(\mu) = 1$.

Proof: Let λ be an eigenvalue of the matrix $M_{\alpha_k}^{-1}A$ and x be the corresponding eigenvector of the eigenvalue λ with $\|x\|_2 = 1$. It is known that:

$$(\alpha_k - i)(W + iT)x = \lambda(\alpha_k W + T)x.$$

So, we have:

$$\lambda = \frac{\alpha_k x^*Wx + x^*Tx + i(\alpha_k x^*Tx - x^*Wx)}{\alpha_k x^*Wx + x^*Tx}.$$

From assumptions, $x^*Wx = c > 0$, $x^*Tx = d \geq 0$. Then we yield $\Re e(\mu) = 1$.

3. RESULTS AND DISCUSSION

This section presents a series of numerical experiments designed to evaluate the practicality, reliability, and computational efficiency of the proposed f-SCSP method in comparison with existing approaches. The evaluation is based on three key performance metrics: the number of iterations to convergence (IT), the total processing time taken by our computer in seconds for convergence (CPU), and the final residual norm (RES). These measures provide a comprehensive assessment of both the convergence characteristics and computational cost of each method.

The performance of f-SCSP is assessed in comparison with four well-known iterative techniques. The MHSS method [12, 13], SMHSS method [21], the f-SMHSS method [21], and the SCSP method [28], which were introduced and discussed in Section 2. In all numerical experiments, the initial guess is taken as the zero vector, and the iterations are terminated once the relative residual norm meets the predefined stopping criterion, set here as an ℓ_2-norm of the residual $\leq 10^{-6}$. The iteration process is considered unsuccessful if convergence is not achieved within a maximum of 8000 iterations. This limit guarantees an equitable assessment among all techniques and aids in avoiding excessive computation time when convergence is not reached as expected. All these experiments are done with different vector space sizes m given $A : \mathbb{C}^m \to \mathbb{C}^m$; the results provide empirical validation of the theoretical analysis and demonstrate the performance of the proposed method.

Example 3.1 [28]: The linear system of equations in (1) represents the form $(W + iT)x = b$, with

$$W = 10(I \otimes V_c + V_c \otimes I) + 9(e_1 e_m^T + e_m^T e_1) \otimes I$$

and $T = I \otimes V + V \otimes I$ where $V = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m\times m}$, $V_c = V - e_1 e_m^T + e_m^T e_1 \in \mathbb{R}^{m\times m}$, $e_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^m$ and $e_m = (0, 0, \dots, 1)^T \in \mathbb{R}^m$. The vector b on the right-hand side can be choosen as $b = (1 + i)A1$, where 1 is the vector with all entries equal to 1.

Example 3.2 [28]**:** The complex linear systems (1) is of the form:

$$[(-\omega^2 M + K) + i(\omega C_V + C_H)]x = b$$

where ω denote the driving circular frequency, with M and K representing the inertia and stiffness matrices, and C_V and C_H are denoting the viscous and hysteretic damping matrices. The viscous damping is modelled as $C_H = \mu K$ where μ is given as the damping coefficient, $M = I$, $C_V = 10I$, $K = I \otimes B_m + B_m \otimes I$, with $B_m = \frac{1}{h^2}\text{tridiag}(-1, 2, -1) \in \mathbb{R}^{m\times m}$, and mesh size $h = \frac{1}{m+1}$. Accordingly, K takes the form of an $n \times n$ block-tridiagonal matrix with block dimension $n = m^2$. We further specify $\omega = \pi$, $\mu = 0.02$, and construct the right-hand vector $b = (1 + i)A1$, where 1 denotes the vector with all components equal to 1. To standardise the system, we pre-multiply both sides by h^2thereby obtaining a normalised formulation.

Example 3.3: Consider the two-dimensional

convection-diffusion equation:

$$-(u_{xx} + u_{yy}) + \eta(u_x + u_y) = g(x, y),$$

the region of interest is considered over the unit square domain $[0,\ 1] \times [0,\ 1]$ assuming constant coefficient η and imposing Dirichlet boundary conditions. Employing the five-point central difference discretisation leads to the linear system (1), characterised by the following coefficient matrix:

$W = T_1 \otimes I + I \otimes T_1$ and $T = I \otimes V + V \otimes I$,

where the matrices T_1 and V are given by:

$$T_1 = \text{tridiag}(-1 - R_e, 2, -1 + R_e), V = \text{tridiag}(2, -1, -1)$$

with $R_e = \mu h/2$, being the mesh Reynolds number, and $h = 1/(m + 1)$ being the equidistant step-size. Moreover, the right-hand side vector b is taken to be $b = Ax$, with $x^* = (1, 1, 1, \ldots, 1)^T \in \mathbb{R}^n$ being the true solution.

In the conducted experiments, matrices with dimensions approaching $270{,}000$ (i.e., $n = m^2 = 512 \times 512 = 262{,}144$) were examined. The numerical results are summarized in Tables 1–3. Evidently, the SCSP and f-SCSP methods perform commendably; the f-SCSP method achieves convergence in the fewest iterations, whereas the SCSP method demonstrates superior computational efficiency in most tests. The challenge of balancing iteration count and execution time to develop an enhanced method constitutes a key direction for forthcoming research.

When compared against its counterparts, SCSP, f-SMHSS, SMHSS and MHSS, the proposed f-SCSP method exhibits a compelling balance between iteration count and computational cost. Table 1 shows results from Example 3.1, and that SCSP is achieving convergence in 10-103 iterations across increasing problem sizes, closely matching the iteration efficiency of flexible f-SCSP but requiring only approximately half the CPU time (e.g., 0.0153s vs. 0.0592s for $m = n = 16$), highlighting its lower overhead in parameter selection. Although f-SCSP attains marginally fewer iterations in some cases, its per-iteration optimization of α_k sustain a significant time penalty. In contrast, classical SMHSS and MHSS methods demand up to an order of magnitude more iterations and substantially longer runtimes, often exceeding SCSP by factors of 5-10, reflecting the superior conditioning induced by the scaled preconditioning. Overall, f-SCSP converges in fewer iterations with better efficiency in all system sizes compared to MHSS, SMHSS, and f-SMHSS. The comparison between f-SCSP and SCSP is however subtle; f-SCSP converges with fewer iterations and a slightly better relative residual in larger system sizes, but the CPU time shows that SPSC is the most efficient throughout. Similarly, Table 2 shows results from Example 3.2, and again f-SPSC and SPSC are very close, with f-SPSC convergeing in fewer iterations and with better relative residual, and SPSC being faster in terms of CPU computational time. All the other methods follow f-SCSP and SCSP. In Table 3, we see results from Example 3, which show that f-SCSP performs superior to all of the existing methods, including SCSP, in terms of all, number of iterations required to converge, the relative residual, and the required CPU time for computation, while SMHSS variants exceed hundreds to thousands of iterations. This consistent performance highlights SCSP's robustness and its practical advantage for large-scale complex symmetric systems.

A catch is the use of the initial guess. All our experiments use $x_0 = \vec{0}$, but many practical solvers benefit from warm starts. Finally, while the convergence proofs (Theorems 2.1-2.3) guarantee $\rho(T_\alpha) < 1$ under stated assumptions, the potential for combining f-SCSP with Krylov acceleration can be addressed, representing an opportunity for further speed-ups in challenging regimes.

Our numerical results presented in the tables are given in line plots. Figure 1 shows the CPU time of taken by the respective methods plotted vs the vector space size m in Example 3.1. The f-SCSP is much faster than most other methods, and it performs very close to the existing SCSP. Similarly, Figure 2 show that in 3.2, as the system size increases, the SCSP performs better than the proposed method. However, it can be seen in Figure 3 for Example 3.3 that both methods perform equally well for all system sizes. Figure 4 show the convergence behavior of the proposed method in Example 3.1 with different system sizes. The residual error is plotted vs the number of iterations, and f-SCSP outperforms the existing methods in all tests, as demonstrated. Similarly, Figure 5 shows how f-SCSP outperforms all of the existing methods in convergence in Example 3.2. In Figure 6, the difference in convergence between f-SCSP and SCSP looks tight, especially in figure 6(b),

Table 1. Tests from Example 3.1. The first column lists the system sizes in $\mathbb{C}^m$. The second column shows iteration count, CPU time, and residual error. Columns 3-7 present the results from SCSP, f-SCSP, f-SMHSS, and MHSS respectively.

m		SCSP	f-SCSP	f-SMHSS	SMHSS	MHSS
16	Iter. Count	10	10	16	18	54
	CPU Time (s)	0.015	0.059	0.048	0.026	0.160
	Res. Err.	5.218e-7	9.694e-7	8.918e-7	6.845e-7	8.238e-7
32	Iter. Count	16	16	26	24	131
	CPU Time (s)	0.150	0.243	0.306	0.202	1.840
	Res. Err.	9.321e-7	4.443e-7	9.043e-7	7.460e-7	9.525e-7
48	Iter. Count	22	20	32	36	171
	CPU Time (s)	0.419	0.608	1.113	0.683	5.053
	Res. Err.	5.287e-07	8.892e-07	7.711e-07	9.641e-07	9.716e-07
64	Iter. Count	28	26	49	55	191
	CPU Time (s)	0.809	1.063	2.681	1.680	5.954
	Res. Err.	6.803e-07	8.043e-07	9.987e-07	8.918e-07	9.875e-07
128	Iter. Count	63	46	119	108	306
	CPU Time (s)	5.971	6.999	13.862	9.638	52.724
	Res. Err.	8.541e-07	9.464e-07	9.601e-07	8.168e-07	9.893e-07
256	Iter. Count	63	60	325	332	997
	CPU Time (s)	28.805	42.680	199.335	302.205	804.894
	Res. Err.	8.258e-07	8.122e-07	9.949e-07	9.929e-07	9.981e-07
512	Iter. Count	103	84	1093	7080	3345
	CPU Time (s)	252.411	510.790	3640.200	17965.00	22926.00
	Res. Err.	9.731e-07	9.537e-07	9.962e-07	9.995e-07	9.993e-07

Table 2. Tests from Example 3.2. The first column lists the system sizes in $\mathbb{C}^m$. The third column shows iteration count, CPU time, and residual error. Columns 3-7 present the results from SCSP, f-SCSP, f-SMHSS, and MHSS respectively.

m		SCSP	f-SCSP	f-SMHSS	SMHSS	MHSS
16	Iter. Count	37	40	268	268	34
	CPU Time (s)	0.053	0.104	0.772	0.372	0.094
	Res. Err.	8.345e-07	8.514e-07	9.782e-07	9.667e-07	9.539e-07
32	Iter. Count	42	38	245	244	49
	CPU Time (s)	0.243	0.364	1.729	1.107	0.557
	Res. Err.	8.969e-07	9.367e-07	9.600e-07	9.878e-07	8.624e-07
48	Iter. Count	44	39	231	231	82
	CPU Time (s)	0.584	0.808	2.900	1.795	1.310
	Res. Err.	8.230e-07	9.204e-07	9.940e-07	9.771e-07	8.920e-07
64	Iter. Count	45	40	222	222	128
	CPU Time (s)	1.147	1.101	6.738	3.955	6.312
	Res. Err.	7.628e-07	7.895e-07	9.781e-07	9.625e-07	9.766e-07
128	Iter. Count	46	41	200	199	440
	CPU Time (s)	4.321	5.710	49.653	30.106	138.168
	Res. Err.	7.429e-07	7.176e-07	9.574e-07	9.870e-07	9.928e-07
256	Iter. Count	46	41	177	177	835
	CPU Time (s)	18.428	26.657	225.347	143.796	1118.5
	Res. Err.	8.145e-07	7.801e-07	9.778e-07	9.643e-07	9.998e-07
512	Iter. Count	46	41	153	152	3160
	CPU Time (s)	140.608	186.084	581.892	355.640	17228.00
	Res. Err.	8.371e-07	7.998e-07	9.613e-07	9.838e-07	9.987e-07

Table 3. Tests from Example 3.3. The first column lists the system sizes in $\mathbb{C}^m$. The third column shows iteration count, CPU time, and residual error. Columns 3-7 present the results from SCSP, f-SCSP, f-SMHSS, and MHSS respectively.

m		SCSP	f-SCSP	f-SMHSS	SMHSS	MHSS
16	Iter. Count	6	3	131	131	150
	CPU Time (s)	0.009	0.010	0.468	0.201	0.443
	Res. Err.	5.645e-07	2.396e-07	9.467e-07	9.374e-07	9.449e-07
32	Iter. Count	6	3	226	226	238
	CPU Time (s)	0.040	0.036	1.938	1.241	1.934
	Res. Err.	5.645e-07	2.396e-07	9.974e-07	9.955e-07	9.782e-07
48	Iter. Count	6	3	345	323	347
	CPU Time (s)	0.080	0.080	5.208	3.519	5.710
	Res. Err.	5.645e-07	2.396e-07	9.934e-07	9.809e-07	9.956e-07
64	Iter. Count	6	3	437	428	624
	CPU Time (s)	0.158	0.155	11.377	7.933	15.952
	Res. Err.	5.645e-07	2.396e-07	9.876e-07	9.973e-07	9.848e-07
128	Iter. Count	6	3	815	751	912
	CPU Time (s)	0.887	0.852	121.434	83.248	177.660
	Res. Err.	5.645e-07	2.396e-07	9.942e-07	9.898e-07	9.914e-07
256	Iter. Count	6	3	1426	1350	1905
	CPU Time (s)	3.062	2.553	1091.90	814.111	1906.40
	Res. Err.	5.645e-07	2.396e-07	9.955e-07	9.950e-07	9.973e-07
512	Iter. Count	6	3	4712	4421	5233
	CPU Time (s)	15.015	13.637	11507.4	17269.0	42689.00
	Res. Err.	5.645e-07	2.396e-07	9.9966e-07	9.994e-07	9.9989e-07

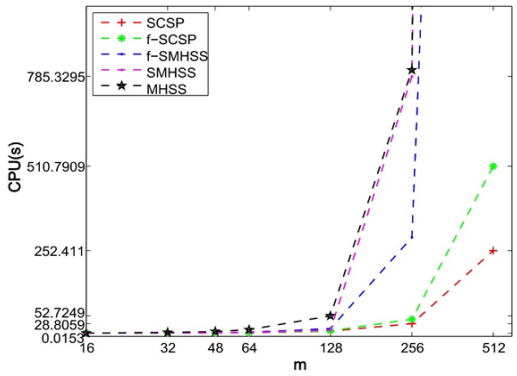


Fig. 1. Comparing f-SCSP with the existing methods in terms of CPU time from Example 3.1. f-SCSP performs better than its most counterparts.

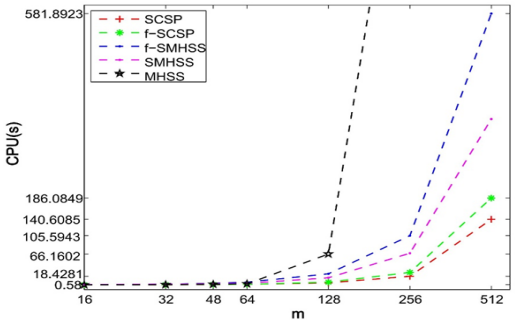


Fig. 2. Comparing f-SCSP with the existing methods in terms of CPU time from Example 3.2. f-SCSP performs better than its counterparts, and is close to SCSP, if not matches its performance. f-SCSP takes a little longer to converge for larger system sizes.

but f-SCSP outperforms SCSP both in number of iterations and final residual error, taking half the number of iterations.

Moreover, Figure 7 shows the eigenvalues spread of the preconditioned matrix vs the actual system matrix in Examples 3.1 for a system size of 48×48 . The real part of an eigenvalue is directly related to how a system behaves over time. If the real part is positive, the system grows exponentially, meaning it becomes unstable over time. If the real part is negative, the system decays exponentially, meaning it settles down to zero. In all preconditioned cases, we see that the eigenvalues have a real part of one and that the system has no fast growing or decaying. Instead, it might oscillate or stay at a constant amplitude. This doesn't guarantee that the matrix is strictly stable, but it demonstrates that the matrix is not unstable either. The same behaviour of strong clustering of the spectrum resulting due to preconditioning can also be observed in Figures 8 and 9 for Example 3.2 and 3.3, respectively, where the preconditioned matrix $M_{\alpha_k}^{-1}A$ evidently has a faster convergence compared to the original matrix A.

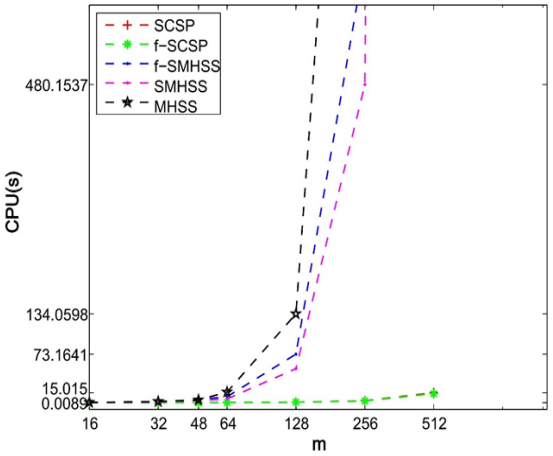


Fig. 3. Comparing f-SCSP with the existing methods in terms of CPU time from Example 3.3. f-SCSP performs better than its counterparts, and performs equally well as SCSP, matching its performance.

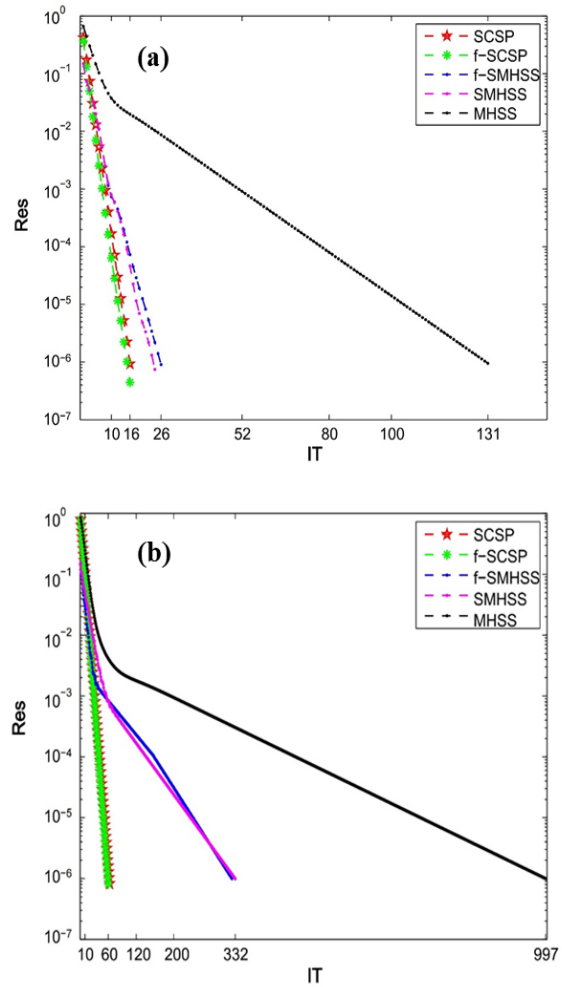


Fig. 4. The convergence behavior of f-SCSP vs its counterparts. (a) show tests from Example 3.1 with vector space $\mathbb{C}^{32}$ and (b) shows $\mathbb{C}^{256}$. Clearly, the convergence in f-SCSP dominates others with a margin.

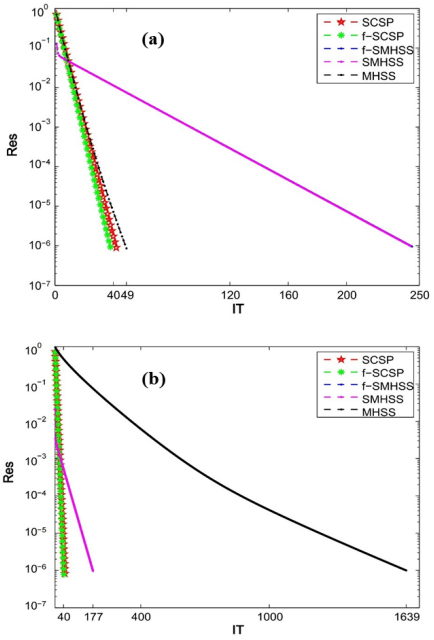


Fig. 5. The convergence behavior of f-SCSP vs its counterparts. (a) test results from Example 3.2 with vector space $\mathbb{C}^{32}$ and (b) shows results with vector space $\mathbb{C}^{256}$. f-SCSP dominates others in convergence with a margin. (a) show the dominance of f-SCSP clearly; whereas (b) shows convergence line of f-SCSP close to SCSP; however, f-SCSP convergence in fewer iterations and with lower residual error.

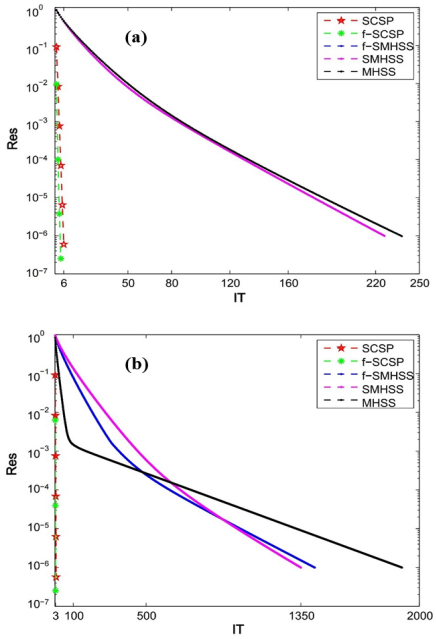


Fig. 6. The convergence behavior of f-SCSP vs its counterparts. (a) test results from Example 3.3 with vector space $\mathbb{C}^{32}$ and (b) shows results with vector space $\mathbb{C}^{256}$. f-SCSP dominates others in convergence with a margin. (a) show the dominance of f-SCSP clearly; however, (b) shows almost overlapping lines for f-SCSP and SCSP; but f-SCSP convergence in half the number of iterations required by SCSP and with lower residual error.

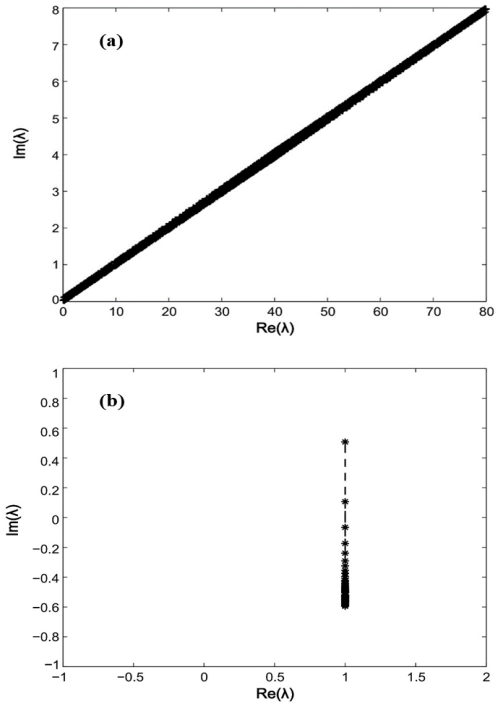


Fig. 7. The eigenvalues of the matrices $\boldsymbol{A}$ compared (a), and the preconditioned matrix $\boldsymbol{M}_{\alpha_k}^{-1}\boldsymbol{A}$ (b), from the system matrix in Example 3.1. The eigenvalues spread in preconditioned system matrix (b) shows the eigenvalues clustered much closer compared to the original matrices (a). Note that the axes ranges are not consistent.

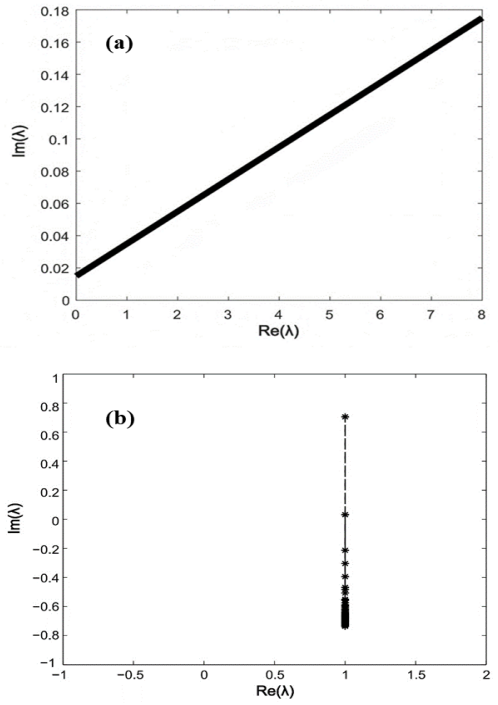


Fig. 8. The eigenvalues of the matrices $\boldsymbol{A}$ compared (a), and the preconditioned matrix $\boldsymbol{M}_{\alpha_k}^{-1}\boldsymbol{A}$ (b), from the system matrix in Example 3.2. The eigenvalues spread in preconditioned system matrix (b) shows the eigenvalues clustered much closer compared to the original matrices (a). Note that the axes ranges are not consistent.

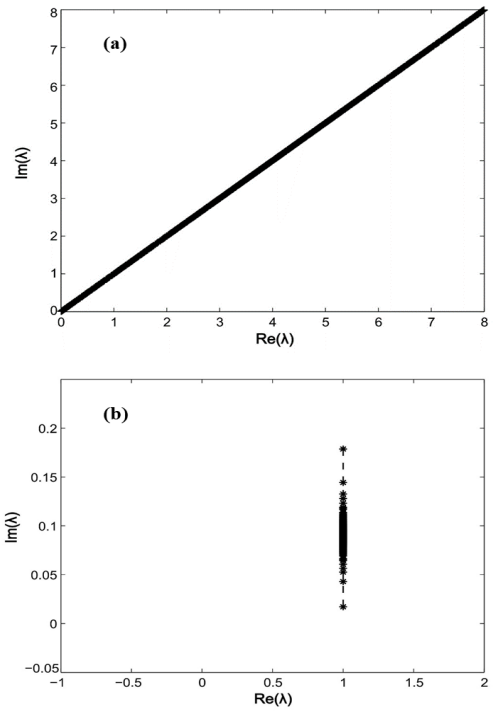


Fig. 9. The eigenvalues of the matrices $\boldsymbol{A}$ compared (a), and the preconditioned matrix $\boldsymbol{M}_{\alpha_k}^{-1}\boldsymbol{A}$ (b), from the system matrix in Example 3.3. The eigenvalues spread in preconditioned system matrix (b) shows the eigenvalues clustered much closer compared to the original matrices (a). Note that the axes ranges are not consistent.

4. CONCLUSIONS

In this paper, we have presented a flexible-scalar splitting iterative methods based on the SCSP method for effectively solving a broad category of complex symmetric linear systems. Special attention is given to the structure and properties of the equivalent systems $(\alpha - i)Ax = (\alpha - i)b$ particularly in cases where the parameters α is chosen to preserve the symmetry and improve the conditioning of the original system. Theoretical analyses have been conducted to demonstrate that the proposed method is convergent under reasonable and practically relevant assumptions. Moreover, explicit expressions linking the optimal parameters α to the spectral radius of the associated iteration matrix have been established, offering a rigorous theoretical basis for parameter tuning and enhanced convergence rates.

To evaluate the practical efficacy of the proposed approaches, extensive numerical experiments were performed comparing the f-SCSP method against four established algorithms from the literature [28]. The findings consistently highlight the proposed

method' reliability, robustness, and computational efficiency. Notably, the f-SCSP method exhibit equal or superior convergence rates and iteration counts, thereby confirming their suitability for tackling complex symmetric linear systems.

5. ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China under Grant No. 12001395, and by the Natural Science Foundation of Shanxi Province, China under Grant No. 202403021222270.

6. CONFLICT OF INTEREST

The authors declare that they have no financial or personal conflicts of interest that could have influenced the research presented in this paper. This study was carried out independently, with funding sources having no role in the design, execution, or interpretation of the results.

7. REFERENCES

1. S.R. Arridge. Optical tomography in medical imaging. *Inverse Problems* 15(2): R41-R93 (1999).
2. D. Bertaccini. Efficient preconditioning for sequences of parametric complex symmetric linear systems. *Electronic Transactions on Numerical* 18: 49-64 (2004).
3. A. Feriani, F. Perotti, and V. Simoncini. Iterative system solvers for the frequency analysis of linear mechanical systems. *Computational Methods in Applied Mechanics and Engineering* 190: 1719-1739 (2000).
4. A. Frommer, T. Lippert, B. Medeke, and K. Schilling (Eds.). Numerical challenges in lattice quantum chromodynamics. *Proceedings, Joint Interdisciplinary Workshop, Wuppertal, Germany* (2000).
5. B. Poirier. Efficient preconditioning scheme for block partitioned matrices with structured sparsity. *Numerical Linear Algebra with Applications* 7: 715-726 (2000).
6. W.V. Dijk and F.M. Toyama. Accurate numerical solutions of the time-dependent Schrödinger equation. *Physical Review E* 75: 036707 (2007).
7. Z. Ahmed, Z.A. Kalhoro, A.W. Shaikh, M.S.R. Baloch, and O.A. Rajput. An improved iterative scheme using successive over-relaxation for solution of linear system of equations. *Proceedings of the Pakistan Academy of Sciences: Part A. Physical and Computational Sciences* 59(3): 35-43 (2022).
8. M. Kanwal, Z. Ahmed, and S. Jamali. Development of generalized refinement strategies in composite stationary iterative solvers for linear systems. *Southern Journal of Research* 5(02(01)): 1-13 (2025).
9. O. Axelsson and A. Kucherov. Real valued iterative methods for solving complex symmetric linear systems. *Numerical Linear Algebra with Applications* 7: 197-218 (2000).
10. M. Benzi and D. Bertaccini. Block preconditioning of real valued iterative algorithms for complex linear systems. *IMA Journal of Numerical Analysis* 28: 598-618 (2007).
11. Z.Z. Bai. On preconditioned iteration methods for complex linear systems. *Journal of Engineering Mathematics* 93: 41-60 (2014).
12. Z.Z. Bai, M. Benzi, and F. Chen. Modified HSS iteration methods for a class of complex symmetric linear systems. *Computing* 87: 93-111 (2010).
13. Z.Z. Bai, M. Benzi, and F. Chen. On preconditioned MHSS iteration methods for complex symmetric linear systems. *Numerical Algorithms* 56: 297-317 (2011).
14. T. Wang, Q. Zheng, and L. Lu. A new iteration method for a class of complex symmetric linear systems. *Journal of Computational and Applied Mathematics* 325: 188-197 (2017).
15. D.K. Salkuyeh, D. Hezari, and V. Edalatpour. Generalized successive overrelaxation iterative method for a class of complex symmetric linear system of equations. *International Journal of Computer Mathematics* 92: 802-815 (2015).
16. D. Hezari, V. Edalatpour, and D.K. Salkuyeh. Preconditioned GSOR iterative method for a class of complex symmetric system of linear equations. *Numerical Linear Algebra with Applications* 22: 761-776 (2015).
17. O. Axelsson and D.K. Salkuyeh. A new version of a preconditioning method for certain two-by-two block matrices with square blocks. *BIT Numerical Mathematics* 59: 321-342 (2019).
18. X. Xie and H. Li. On preconditioned Euler-extrapolated single-step Hermitian and skew-Hermitian splitting method for complex symmetric linear systems. *Japan Journal of Industrial and Applied Mathematics* 8: 503-518 (2020).
19. Y. Xiang and N.M. Zhang. On the preconditioned conjugate gradient method for complex symmetric systems. *Applied Mathematics Letters* 120: 107250 (2021).
20. D.K. Salkuyeh. A Preconditioner for Complex

Symmetric System of Linear Equations with Indefinite Hermitian Part. *Bulletin of the Iranian Mathematical Society* 51: 25-25 (2025).

21. P.P. Zhao, S.D. Li, and R.P. Wen. Single-step HSS methods for a class of complex symmetric linear systems. *Communications in Applied Mathematics and Computation* 31: 200-212 (2017).
22. R.P. Wen, F.J. Ren, and Y.Q. Gao. On convergence of splitting iteration methods for the complex symmetric linear systems. *Mathematica Applicanda* 29: 173-182 (2016).
23. R.P. Wen, S.D. Li, and F.J. Ren. A new splitting and preconditioner for iteratively solving a class of complex symmetric linear systems. *Mathematica Applicanda* 27: 65-72 (2014).
24. H.A. van-der-Vorst and J.B.M. Melissen. A Petrov–Galerkin type method for solving $Ax = b$, where A is symmetric complex. *IEEE Transactions on Magnetics* 26: 706-708 (1990).
25. R.W. Freund. Conjugate gradient-type methods for linear systems with complex symmetric coefficient matrices. *SIAM Journal on Scientific and Statistical Computing* 13: 425-448 (1992).
26. A. Bunse-Gerstner and R. Stöver. On a conjugate gradient–type method for solving complex symmetric linear systems. *Linear Algebra and Its Applications* 287: 105-123 (1999).
27. M. Clemens, T. Weiland, and U. Van-Rienen. Comparison of Krylov-type methods for complex linear systems applied to high-voltage problems. *IEEE Transactions on Magnetics* 34: 3335-3338 (1998).
28. D. Hezari, D.K. Salkuyeh, and V. Edalatpour. A new iterative method for solving a class of complex symmetric system of linear equations. *Numerical Algorithms* 73: 927-955 (2016).
29. D.K. Salkuyeh. Two-step scale-splitting method for solving complex symmetric system of linear equations. *Arxiv Preprint Arxiv*: 1705.02468 (2017).
30. D.K. Salkuyeh and T.S. Siahkolaei. Two-parameter TSCSP method for solving complex symmetric system of linear equations. *Calcolo* 55: 8 (2018).
31. Z. Zheng, F.L. Huang, and Y.C. Peng. Double-step scale splitting iteration method for a class of complex symmetric linear systems. *Applied Mathematics Letters* 73: 91-97 (2017).
32. B. Li, J. Cui, Z. Huang, and X. Xie. A dual-parameter double-step splitting iteration method for solving complex symmetric linear equations. *Applications of Mathematics* 69: 311-337 (2024).
33. B. Li, J. Cui, Z. Huang, and X. Xie. Two Quasi-combining Real and Imaginary Parts Iteration Methods for Solving Complex Symmetric System of Linear Equations. *Communications on Applied Mathematics and Computation* (2024). https://doi.org/10.1007/s42967-024-00448-0

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 313-321 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)699

Pakistan Academy of Sciences

Research Article

A Modified Twentieth-Order Iterative Method for Solving Nonlinear Physicochemical Models: Convergence, Physical Models and Basin of Attraction Analysis

Sanaullah Jamali[1*], Zubair Ahmed Kalhoro[2], Saifullah Jamali[3], Baddar ul ddin Jamali[4], Abdul Wasim Shaikh[2], and Muhammad Saleem Chandio[2]

[1]Department of Mathematics, University of Sindh, Laar Campus, Badin, Sindh, Pakistan
[2]Institute of Mathematics and Computer Science, University of Sindh, Jamshoro-76080, Sindh, Pakistan
[3]Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China
[4]Dr. M.A. Kazi Institute of Chemistry, University of Sindh, Jamshoro-76080, Sindh, Pakistan

Abstract: This paper introduces a modified twentieth-order method for solving nonlinear equations that commonly arise in physicochemical models. The proposed method is designed to efficiently handle the complex problems that normally occur in the van der Waals equation for real gases, Planck's radiation law, and chemical equilibrium conditions. The traditional method has a lower order of convergence and uses higher-order derivatives. However, proposed method has twentieth-order convergence with only one first derivative used in each iteration. A detailed convergence order has been carried out to demonstrate the theoretical order of accuracy. Various numerical experiments have also been carried out to validate the performance of the proposed method. The results show the significantly improve the accuracy and taking a smaller number of iterations, number of function evaluations, and CPU time when applied to nonlinear equations arises in van der Waals equation for real gases, Planck's radiation law, and chemical equilibrium conditions and basin of attraction further validate the stability of proposed method.

Keywords: Nonlinear Physicochemical Models, Iterative Method, Convergence Analysis, Weight Function, Hermite Interpolation, Basin of Attraction.

1. INTRODUCTION

One of the key challenges in numerical analysis is solving nonlinear equations that arise in engineering problems, specifically in arises in van der Waals equation for real gases, Planck's radiation law, and chemical equilibrium conditions. Iterative methods, like newton's method, are commonly employed for this purpose. In this context, this article focuses on iterative techniques aimed at finding a simple root a, such that $\psi(a) = 0$ and $\psi'(a) \neq 0$, for a nonlinear equation $\psi(x) = 0$ [1]. High precision is most significant for numerical computation, highlighting the importance of higher-order numerical methods [2]. Many scholars have proposed higher-order methods for solving nonlinear algebraic and transcendental equations [3-5]. Similarly, a number of researchers have also introduced a higher-order convergence optimal method [6-8]. Bracketing/closed method [9-13] have also have their importance because they have always been convergent, but their convergence is very slow. So now the researchers are more intend to introduce higher order method using weight function techniques [14-16].

2. DERIVATION

We use the Newton technique [1] as the first step in the suggested approach.

Received: October 2025; Revised: November 2025; Accepted: December 2025
* Corresponding Author: Sanaullah Jamali <sanaullah.jamali@usindh.edu.pk>

$$v_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} \tag{1}$$

In the second step of the proposed method, we utilize a variant of the double Newton method [17] and modify it by substituting $\psi(\varkappa_n)$ with $\psi'(\varkappa_n)$ in this step.

$$\xi_n = v_n - \left[1 + \left(\frac{\psi(v_n)}{\psi'(\varkappa_n)}\right)^2\right]\left(\frac{\psi(v_n)}{\psi'(v_n)}\right) \tag{2}$$

From Equations (1) and (2) we get:

$$\left.\begin{aligned} &\text{Step 1. } v_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} \\ &\text{Step2. } \xi_n = v_n - \left[1 + \left(\frac{\psi(v_n)}{\psi'(\varkappa_n)}\right)^2\right]\left(\frac{\psi(v_n)}{\psi'(v_n)}\right) \end{aligned}\right\} \tag{3}$$

To enhance the accuracy and convergence, introduce the weight function L see in Thukral [18] in the step 2 of Equation (3).

$$\left.\begin{aligned} &\text{Where } L = K - 2a + 2ab(a-1)^2 + 2a^3b^{-1} \\ &\text{And } K = (1 - a^2 - 10a^4)^{-1}, \quad a = \frac{\psi(v_n)}{\psi(\varkappa_n)}, \quad b = \frac{\psi'(\varkappa_n)}{\psi'(v_n)} \end{aligned}\right\}$$

We get

$$\left.\begin{aligned} &\text{Step 1. } v_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} \\ &\text{Step 2. } \xi_n = v_n - L\left[1 + \left(\frac{\psi(v_n)}{\psi'(\varkappa_n)}\right)^2\right]\left(\frac{\psi(v_n)}{\psi'(v_n)}\right) \end{aligned}\right\} \tag{4}$$

And add one more step of newton by using $\psi(\xi_n)$ and $\psi'(\xi_n)$, $\psi'(\xi_n) \approx h_3'(\xi_n)$

$$\left.\begin{aligned} &\text{Step 1. } v_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} \\ &\text{Step 2. } \xi_n = v_n - L\left[1 + \left(\frac{\psi(v_n)}{\psi'(\varkappa_n)}\right)^2\right]\left(\frac{\psi(v_n)}{\psi'(v_n)}\right) \\ &\text{Step 3. } o_n = \xi_n - \frac{\psi(\xi_n)}{\psi'(\xi_n)} \end{aligned}\right\} \tag{5}$$

In three-step formula mentioned in Equation (5) we estimate $\psi'(\xi_n)$ using existing data, thereby reducing the number of function evaluations needed per iteration. At the nodes $\varkappa, v$, and ξ, we have four values $\psi(\varkappa), \psi'(\varkappa), \psi(v)$ and $\psi(\xi)$. In the third step of the iterative scheme in Equation (5), we use the approximation $\psi'(\xi) \approx H_3'(\xi)$ to approximate ψ using Hermite's interpolating polynomial of degree 3. This algorithm takes the following form.

$$H_3(\eta) = a_0 + a_1(\eta - \varkappa) + a_2(\eta - \varkappa)^2 + a_3(\eta - \varkappa)^3 \tag{6}$$

And its derivative is:

$$H_3'(\eta) = a_1 + 2a_2(\eta - \varkappa) + 3a_3(\eta - \varkappa)^2 \tag{7}$$

The unknown coefficients will be determined using available data from the conditions:

$$H_3(\varkappa) = \psi(\varkappa), \quad H_3(v) = \psi(v), \quad H_3(\xi) = \psi(\xi)$$
$$\& \quad H_3'(\varkappa) = \psi'(\varkappa)$$

Putting $\eta = \varkappa$ into Equations (6) and (7) we get $a_0 = \psi(\varkappa)$ and $a_1 = \psi'(\varkappa)$. The coefficients a_2 and a_3 are obtained from the system of two linear equations formed by using the remaining two conditions $\eta = v \,\&\, \eta = \xi$ in Equation (6) and we obtain:

$$a_2 = \frac{(\xi-\varkappa)\psi[v,\varkappa]}{(\xi-v)(v-\varkappa)} - \frac{(v-\varkappa)\psi[\xi,\varkappa]}{(\xi-v)(v-\varkappa)} - \psi'(\varkappa)\left(\frac{1}{\xi-\varkappa} - \frac{1}{v-\varkappa}\right)$$
$$\& \quad a_3 = \frac{\psi[\xi,\varkappa]}{(\xi-v)(\xi-\varkappa)} - \frac{\psi[v,\varkappa]}{(\xi-v)(v-\varkappa)} + \frac{\psi'(\varkappa)}{(\xi-\varkappa)(v-\varkappa)}$$

By putting the values of $a_1, a_2, a_3 \,\&\, \eta = \xi$ in Equation (7) we get:

$$H_3'(\xi) = 2(\psi[\varkappa,\xi] - \psi[\varkappa,v]) + \psi[v,\xi] + \frac{v-\xi}{v-\varkappa}\left(\psi[\varkappa,v] - \psi'(\varkappa)\right) \tag{8}$$

We replace $\psi'(\xi_n)$ in third step of Equation (5) by Equation (8) H_3Hermite we get:

$$\left.\begin{aligned} &\text{Step 1. } v_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} \\ &\text{Step 2. } \xi_n = v_n - L\left[1 + \left(\frac{\psi(v_n)}{\psi'(\varkappa_n)}\right)^2\right]\left(\frac{\psi(v_n)}{\psi'(v_n)}\right) \\ &\text{Step 3. } o_n = \xi_n - \frac{\psi(\xi_n)}{h_3'(\xi_n)} \end{aligned}\right\} \tag{9}$$

Now add one more step of newton by using $\psi(o_n)$ and $\psi'(o_n)$.
And finally, we got:

$$\left.\begin{aligned} &\text{Step 1. } v_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} \\ &\text{Step 2. } \xi_n = v_n - L\left[1 + \left(\frac{\psi(v_n)}{\psi'(\varkappa_n)}\right)^2\right]\left(\frac{\psi(v_n)}{\psi'(v_n)}\right) \\ &\text{Step 3. } o_n = \xi_n - \frac{\psi(\xi_n)}{h_3'(\xi_n)} \\ &\text{Step 4. } \varkappa_{n+1} = o_n - \frac{\psi(o_n)}{\psi'(o_n)} \end{aligned}\right\} \tag{10}$$

Equation (10) is the twentieth-order method with four function evaluations and three first derivatives.

3. CONVERGENCE ANALYSIS

Theorem: D represents an open interval containing $\varkappa_0$ as a first estimate of $\sigma \in D$. Let $\sigma \in D$ be a simple root of a function $\psi : D \subset R \to R$ that is suitably differentiable. Under these conditions, Equation (10) yields Twentieth-order of convergence and requires only four function evaluations along with

three first derivative calculations in each complete iteration, with no need for second or higher-order derivatives.

Proof.
The Taylor series expansion for the function $\psi(\varkappa_n)$ can be expressed as:

$$\psi(\varkappa_n) = \sum_{m=0}^{\infty} \frac{\psi^m(\sigma)}{m!}(\varkappa_n - \sigma)^m = \psi(\sigma) + \psi'(\sigma)(\varkappa_n - \sigma) + \frac{\psi''(\sigma)}{2!}(\varkappa_n - \sigma)^2 + \frac{\psi'''(\sigma)}{3!}(\varkappa_n - \sigma)^3 + \cdots \quad (11)$$

For simplicity, we assume that

$$R_k = \left(\frac{1}{k!}\right)\frac{\psi^k(\sigma)}{\psi'(\sigma)}, k \geq 2.$$

and assume that $\varepsilon_n = \varkappa_n - \sigma$. Thus, we have:
For step one:

$$\psi(\varkappa_n) = \psi'(\sigma)\begin{pmatrix}\varepsilon_n + R_2\varepsilon_n^2 + R_3\varepsilon_n^3 + \\ R_4\varepsilon_n^4 + \cdots + R_{21}\varepsilon_n^{21}\end{pmatrix} \quad (12)$$

$$\psi'(\varkappa_n) = \psi'(\sigma)\begin{pmatrix}1 + 2R_2\varepsilon_n + 3R_3\varepsilon_n^2 + \\ 4R_4\varepsilon_n^3 \ldots + 21R_{21}\varepsilon_n^{20}\end{pmatrix} \quad (13)$$

From Equations (12) and (13):

Step 1. $\nu_n = \varkappa_n - \frac{\psi(\varkappa_n)}{\psi'(\varkappa_n)} = R_2\varepsilon_n^2 + (2R_3 - 2R_2^2)\varepsilon_n^3 + (4R_2^3 - 7R_3R_2 + 3R_4)\varepsilon_n^4 + \cdots + O(\varepsilon_n^{21})$ (14)

Step 2. $\xi_n = \nu_n - L * \left(1 + \left(\frac{\psi(\nu_n)}{\psi'(\varkappa_n)}\right)^2\right)\left(\frac{\psi(\nu_n)}{\psi'(\nu_n)}\right) = R_2^2(3R_2^3 - (7R_3 + 1)R_2 + R_4)\varepsilon_n^6 - 2\left(R_2\begin{pmatrix}R_2^5 - (36R_3 + 5)R_2^3 + 9R_4R_2^2 + \\ (R_3(20R_3 + 3) - R_5)R_2 - 2R_3R_4\end{pmatrix}\right)\varepsilon_n^7 + \ldots + O(\varepsilon_n^{21})$ (15)

Step 3. $o_n = \xi_n - \frac{\psi(\xi_n)}{h_3'(\xi)} = R_2^3R_4(3R_2^3 - (7R_3 + 1)R_2 + R_4)\varepsilon_n^{11} + R_2^2\begin{pmatrix}2R_2(3R_2^3 - (7R_3 + 1)R_2 + 2R_4)R_5 - \\ 2R_4\begin{pmatrix}4R_2^5 - 2(23R_3 + 3)R_2^3 + 10R_4R_2^2 + \\ R_3(27R_3 + 4)R_2 - 3R_3R_4\end{pmatrix}\end{pmatrix}\varepsilon_n^{12} + \ldots + O(\varepsilon_n^{21})$ (16)

Step 4. $\varkappa_{n+1} = o_n - \frac{\psi(o_n)}{\psi'(o_n)} = R_2^7R_4^2(3c_2^3 - (7R_3 + 1)R_2 + R_4)^2\varepsilon_n^{20} + O(\varepsilon_n^{21})$ (17)

Lastly, the efficiency index of the proposed approach mentioned in Equation (10) is 1.534127405, the rate of convergence is twenty, and each iteration requires three first derivative evaluations and four function evaluations.

4. NUMERICAL EXPERIMENT AND DISCUSSION

Problem 1. A chemical equilibrium problem (see [19-21])

$$\varkappa^4 - 7.79075\varkappa^3 + 14.7445\varkappa^2 + 2.511\varkappa - 1.674 = 0$$

Table 1. Numerical results for problem 1 for first four iterations and their absolute function values at $\varkappa_0 = 0.6$.

Method	Root & absolute function value	1st iteration	2nd iteration	3rd iteration	4th iteration
PM	$\varkappa$	0.2777 ...	0.2777 ...	0.2777 ...	0.2777 ...
	$\lvert\psi(\varkappa)\rvert$	3.9356E − 13	2.9239E − 267	7.6755E − 5350	1.8529E − 107001
A1 20th	$\varkappa$	0.2777 ...	0.2777 ...	0.2777 ...	0.2777 ...
	$\lvert\psi(\varkappa)\rvert$	5.0042E − 11	1.2188E − 221	6.5800E − 4434	2.9086E − 88679
A2 20th	$\varkappa$	0.2777 ...	0.2777 ...	0.2777 ...	0.2777 ...
	$\lvert\psi(\varkappa)\rvert$	2.2287E − 10	5.0928E − 208	7.6768E − 4161	2.8154E − 83217
A3 20th	$\varkappa$	0.2777 ...	0.2777 ...	0.2777 ...	0.2777 ...
	$\lvert\psi(\varkappa)\rvert$	1.6868E − 10	1.4682E − 210	9.1397E − 4212	6.9775E − 83236

Table 2. Numerical results for the problem 1, error fixed at $\delta = 1 \times 10^{-5}$.

Method	IG	N	FE	CPU Time
PM	0.6	4	28	2.78×10^{0}
A1 20th	0.6	5	35	8.39×10^{0}
A2 20th	0.6	5	35	9.56×10^{0}
A3 20th	0.6	5	35	1.02×10^{1}

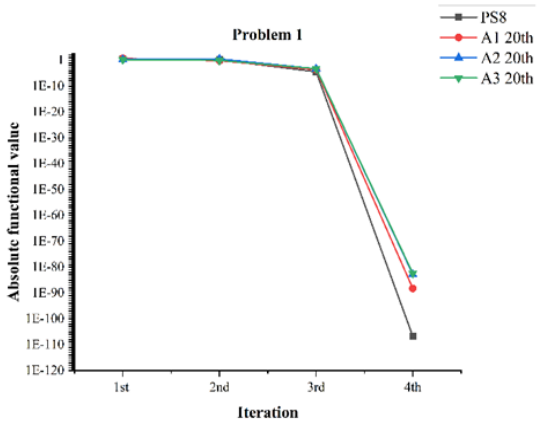


Fig. 1. Graphical Representation of $|\psi(\varkappa)|$ of Table 1. by assuming the scale $1 \times 10^{-3} = 1 \times 10^{-1}$.

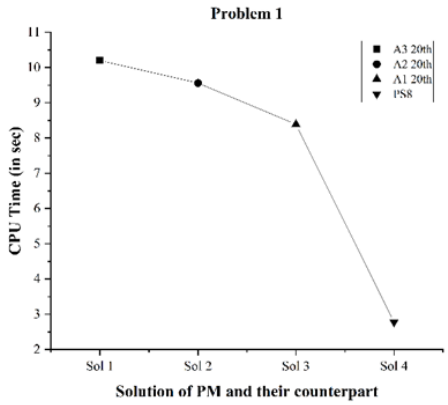


Fig. 2. CPU time (in sec) versus solution of problem 1 by the proposed scheme and its counterparts.

The performance of the PM method in solving problem 1 is evaluated against A1 20th, A2 20th, and A3 20th up to the fourth iteration. Results presented in Table 1 indicate that PM achieves higher accuracy and faster convergence, as depicted in Figure 1, which illustrates PM's quicker convergence relative to the other methods. Table 2 provides detailed metrics, showing that PM requires only 4 iterations and 28 function evaluations, whereas the other methods necessitate 5 iterations and 35 evaluations. Additionally, PM consumes less CPU time to achieve a tolerance of 1×10^{-5}, with Figure 2 reinforcing its superior CPU time performance compared to alternative methods.

Problem 2. Volume from van der Waals equation (see [8]**)**

$$\psi(\varkappa) = 40\varkappa^3 - 95.26535116\varkappa^2 + 35.28\varkappa - 5.6998368$$

Table 3. Numerical results for problem 2 for first four iterations and their absolute function values at $\varkappa_0 = 2.5$.

Method	Root & absolute functional value	1st iteration	2nd iteration	3rd iteration	4th iteration
PM	$\varkappa$	1.9707 ...	1.9707 ...	1.9707 ...	1.9707 ...
	$\|\psi(\varkappa)\|$	2.7230E − 7	7.3008E − 207	1.3896E − 4996	7.1118E − 119950
A1 20th	$\varkappa$	1.9707 ...	1.9707 ...	1.9707 ...	1.9707 ...
	$\|\psi(\varkappa)\|$	8.3409E − 5	1.4913E − 118	1.6624E − 2393	1.4603E − 47892
A2 20th	$\varkappa$	1.9707 ...	1.9707 ...	1.9707 ...	1.9707 ...
	$\|\psi(\varkappa)\|$	4.2265E − 5	8.9428E − 125	2.8928E − 2518	4.5534E − 50388
A3 20th	$\varkappa$	1.9707 ...	1.9707 ...	1.9707 ...	1.9707 ...
	$\|\psi(\varkappa)\|$	5.1315E − 5	5.3469E − 123	1.2172E − 2482	1.7007E − 49675

Table 4. Numerical results for problem 2, error fixed at $\delta = 1 \times 10^{-5}$.

Method	IG	N	FE	CPU Time
PM	2.5	4	28	7.08×10^{0}
A1 20th	2.5	5	35	7.32×10^{0}
A2 20th	2.5	5	35	7.94×10^{0}
A3 20th	2.5	5	35	7.78×10^{0}

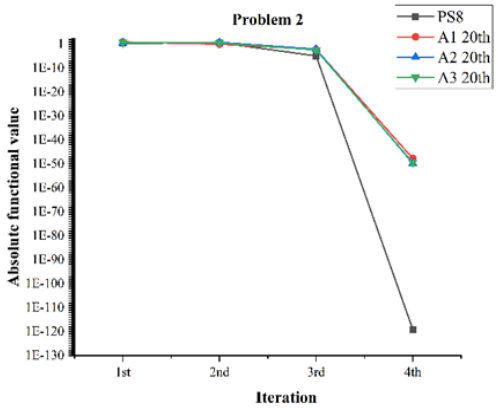


Fig. 3. Graphical Representation of $|\psi(\varkappa)|$ of Table 3. by assuming the scale $1 \times 10^{-3} = 1 \times 10^{-1}$.

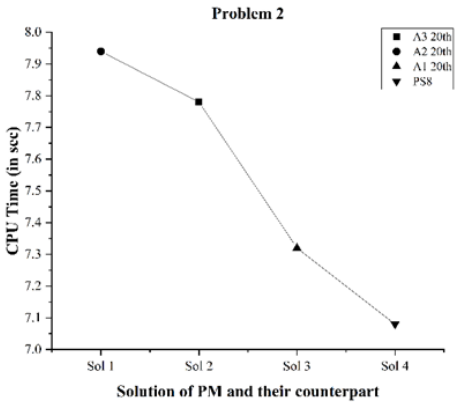


Fig. 4. CPU time versus the solution of problem 2 with the proposed scheme and its counterparts.

Table 3 shows that PM is more accurate and converges quickly than its counterpart approaches in problem 2. And Table 4 shows the iterations, function evaluations, and CPU time (in seconds), where A1, A2, and A3 need 5 iterations and 35 function evaluations, whereas PM requires 4 and 28. PM achieves a tolerance of $\delta = 1 \times 10^{-5}$ more effectively than comparable approaches because of its decreased CPU time (in seconds). However, Figures 3 and 4 are graphical representations of Tables 3 and 4, also demonstrating that the proposed method is more accurate.

Problem 3. Planck's radiation law (see [20, 22-25, 27])

$$e^{-\varkappa} - 1 + \frac{\varkappa}{5} = 0.$$

Table 5. Numerical results for problem 3 for first four iterations and their absolute function values at $\varkappa_0 = -0.5$.

Method	Root & absolute functional value	1st iteration	2nd iteration	3rd iteration	4th iteration
PM	$\varkappa$	$-5.9344E-14$	$-1.6768E-269$	$-1.7657E-5380$	$-4.9576E-107600$
	$\lvert\psi(\varkappa)\rvert$	$4.7475E-14$	$4.7475E-269$	$4.7475E-5380$	$4.7475E-107600$
A1 20th	$\varkappa$	$-5.4708E-10$	$-2.0950E-187$	$-9.6359E-3736$	$-1.7293E-74702$
	$\lvert\psi(\varkappa)\rvert$	$4.3767E-10$	$1.6760E-187$	$7.7087E-3736$	$1.3835E-74702$
A2 20th	$\varkappa$	$-7.6741E-11$	$-2.5011E-205$	$-2.5011E-4095$	$-8.0702E-81890$
	$\lvert\psi(\varkappa)\rvert$	$6.1393E-11$	$2.0009E-205$	$3.6606E-4095$	$6.4562E-81890$
A3 20th	$\varkappa$	$-1.5682E-10$	$-8.2960E-199$	$-2.4446E-3964$	$-5.9562E-79275$
	$\lvert\psi(\varkappa)\rvert$	$1.2545E-10$	$6.6368E-199$	$1.9556E-3964$	$1.9556E-79275$

Table 6. Numerical results for problem 3, error fixed at $\delta = 1 \times 10^{-5}$.

Method	IG	N	FE	CPU Time
PM	-0.5	4	28	5.18×10^2
A1 20th	-0.5	5	35	5.25×10^2
A2 20th	-0.5	5	35	5.20×10^2
A3 20th	-0.5	5	35	5.24×10^2

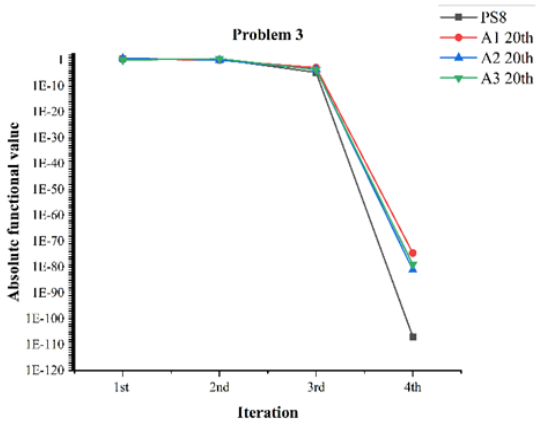


Fig. 5. Graphical Representation of $|\psi(\varkappa)|$ of Table 5. by assuming the scale $1 \times 10^{-3} = 1 \times 10^{-1}$.

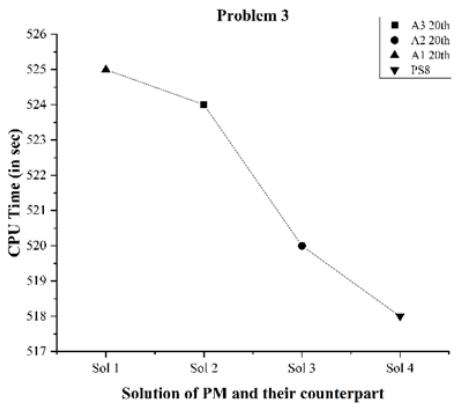


Fig. 6. CPU time (in sec) versus solution of problem 3 with the proposed scheme and its counterparts.

Compared to its counterpart approaches in problem 3, PM is more accurate and converges faster, as Table 5 demonstrates. Additionally, Table 6 displays the CPU time (in seconds), number of iterations, function evaluations. A1, A2, and A3 require five iterations and thirty-five function evaluations, while PM needs four and twenty-eight. PM's reduced CPU time (in seconds) allows it to achieve a tolerance of $\delta = 1 \times 10^{-5}$ more efficiently than similar methods. Figures 3 and 4, on the other hand, are graphical depictions of Tables 5 and 6, further proving the validity of the suggested approach.

The visuals show that PM is more accurate, efficient, and consistent than alternative approaches.

5. BASIN OF ATTRACTION

The stability of the solutions (roots) for the nonlinear function $\psi(\varkappa) = 0$.The concept of basins of attraction can be used to facilitate an iterative method [26]. MATLAB R2014a was used to generate a depiction of all basins within the range $R = [-5 \times 5] \times [-5 \times 5]$, with a total of 360,000 points at a 600×600 density. There were two criteria established: An error threshold of 1×10^{-10} or a maximum iteration count of 10. Each point in the R-range served as the starting condition for the iterative algorithms that are initiated.

The iterative algorithm assigned a unique color number k (other than black) to the initial point if the sequence converged to a root x_k^* of the polynomial $P_n(x)$ of degree k within 10 iterations and a predetermined tolerance. On the other hand, if the iterative process started at a point $x \in C$ and surpassed the maximum iteration limit of 10 without converging to any root x_k or converged to a different value p such that $|p - x^*| < 1 \times 10^{-10}$, the starting point was classified as diverging. In these instances, the starting point was marked with the color black. The number of iterations for each point in another basin is represented, accompanied by a color scale for reference.

The visual representations presented in Figure 7 show that PM has significantly higher stability than alternative methods.

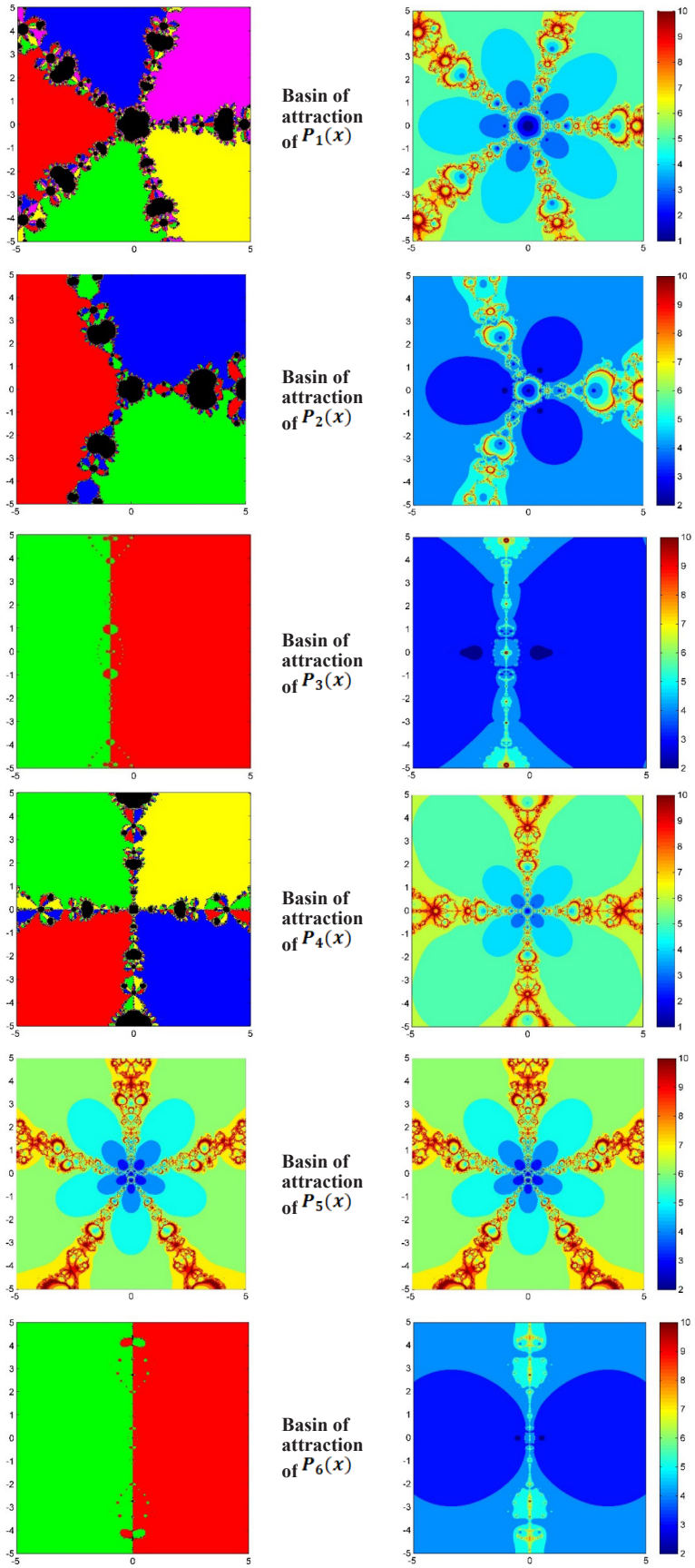


Fig. 7. The left Figures shows roots, while the right Figures. shows the number of iterations at each initial point of $P_n(x)$ of problems 4 obtained by the proposed Twentieth-order method.

Problem 4. Below problems were taken from the literature [26].

S. No.	Functions $(P(x))$	Roots $(x_k : k = 1, 2, 3, \ldots)$
1.	$P_1(x) = x^5 + 1$	$x_k = -1, -\frac{305}{987} \pm \frac{855}{899}i, \frac{1292}{1597} \pm \frac{4456}{7581}i$
2.	$P_2(x) = x^3 + 1$	$x_k = 1, \frac{1 \pm \sqrt{3}i}{2}$
3.	$P_3(x) = x^2 + 2x - \frac{1}{2}$	$x_k = \frac{-2 \pm \sqrt{6}}{2}$
4.	$P_4(x) = x^4 + \frac{1}{64}$	$x_k = \frac{1 \pm 1i}{4}, \frac{-1 \pm 1i}{4}$
5.	$P_5(x) = x^5 - \frac{1}{2}ix^4 + \frac{1}{64}x - \frac{1}{128}i$	$x_k = \frac{1 \pm 1i}{4}, \frac{-1 \pm 1i}{4}, \frac{1}{2}i$
6.	$P_6(x) = x^2 - \frac{1}{4}$	$x_k = \frac{1}{2}, -\frac{1}{2}$

6. CONCLUSIONS

The proposed fourth step, the twenty-order method based on the weight function, is introduced for the solution of nonlinear equations arising in Physicochemical Models. In conclusion, we have derived the convergence order (theoretical) of the proposed method, various application problems from the Physicochemical Models have been tested and compared with counterparts A1, A2, and A3. In all cases proposed method outperforms existing methods in terms of accuracy, number of iterations, number of function evaluations, and CPU time. Furthermore, the Basin of attraction in the complex plane confirms the stability of the proposed method.

7. CONFLICT OF INTEREST

The authors declare no conflict of interest.

8. REFERENCES

1. H. Susanto and N. Karjanto. Newton's method's basins of attraction revisited. *Applied Mathematics and Computation* 215(3): 1084-1090 (2009).
2. M. Grau and J.L. Díaz-Barrero. An improvement to Ostrowski root-finding method. *Applied Mathematics and Computation* 173(1): 450-456 (2006).
3. F.A. Lakho, Z.A. Kalhoro, S. Jamali, A.W. Shaikh, and J. Guan. A three steps seventh order iterative method for solution nonlinear equation using Lagrange Interpolation technique. *VFAST Transactions on Mathematics* 12(1): 46-59 (2024).
4. Z. Abbasi, Z.A. Kalhoro, S. Jamali, A.W. Shaikh, and O.A. Rajput. A novel approach for real-world problems based on Hermite interpolation technique and analysis using basins of attraction. *Science* 5(3): 112-126 (2024).
5. S. Jamali, Z.A. Kalhoro, and I.Q. Memon. An efficient four step fifteenth order method for solution of non-linear models in real-world problems. *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences* 61(3): 273-281 (2024).
6. A.S. Alshomrani, R. Behl, and V. Kanwar. An optimal reconstruction of Chebyshev–Halley type methods for nonlinear equations having multiple zeros. *Journal of Computational and Applied Mathematics* 354: 651-662 (2019).
7. M.U.D. Junjua, F. Zafar, and N. Yasmin. Optimal derivative-free root finding methods based on inverse interpolation. *Mathematics* 7(2): 164 (2019).
8. O.S. Solaiman and I. Hashim. Efficacy of optimal methods for nonlinear equations with chemical engineering applications. *Mathematical Problems in Engineering* 2019: 1728965 (2019).
9. V. Kodnyanko. Improved bracketing parabolic method for numerical solution of nonlinear equations. *Applied Mathematics and Computation* 400: 125995 (2021).
10. B.M. Faraj, S.K. Rahman, D.A. Mohammed, B.M. Hussein, B.A. Salam, and K.R. Mohammed. An improved bracketing method for numerical solution

of nonlinear equations based on Ridders method. *Matrix Science Mathematics* 6(2): 30-33 (2022).

11. S. Jamali, Z.A. Kalhoro, A.W. Shaikh, and M.S. Chandio. An iterative, bracketing & derivative-free method for numerical solution of non-linear equations using Stirling interpolation technique. *Journal of Mechanics of Continua and Mathematical Sciences* 16(6): 13-27 (2021).
12. A. Suhadolnik. Combined bracketing methods for solving nonlinear equations. *Applied Mathematics Letters* 25(11): 1755-1760 (2012).
13. M.I. Soomro, Z.A. Kalhoro, A.W. Shaikh, S. Jamali, and O. Ali. Modified bracketing iterative method for solving nonlinear equations. *VFAST Transactions on Mathematics* 12(1): 105-120 (2024).
14. A. Cordero, N. Garrido, J.R. Torregrosa, P. Triguero-Navarro, M. Moscoso-Martínez, and J.R. Torregrosa. Three-step iterative weight function scheme with memory for solving nonlinear problems. *Mathematical Methods in the Applied Sciences* 48(7): 8024-8036 (2023).
15. S. Jamali, Z.A. Kalhoro, A.W. Shaikh, M.S. Chandio, and S. Dehraj. A new three step derivative free method using weight function for numerical solution of non-linear equations arises in application problems. *VFAST Transactions on Mathematics* 10(2): 164-174 (2022).
16. M.Q. Khirallah and A.M. Alkhomsan. A new fifth-order iterative method for solving non-linear equations using weight function technique and the basins of attraction. *Journal of Mathematics and Computer Science* 28(3): 281-293 (2023).
17. M. Kumar, A.K. Singh, and A. Srivastava. Various Newton-type iterative methods for solving nonlinear equations. *Journal of the Egyptian Mathematical Society* 21(3): 334-339 (2013).
18. R. Thukral. Two-step iterative methods with sixth-order convergence for solving nonlinear equations. *British Journal of Mathematics and Computer Science* 4(14): 1941-1950 (2014).
19. A. Naseem, M.A. Rehman, and T. Abdeljawad. Numerical methods with engineering applications and their visual analysis via polynomiography. *IEEE Access* 9: 99287 (2021).
20. A. Naseem, M.A. Rehman, and T. Abdeljawad. Computational methods for non-linear equations with some real-world applications and their graphical analysis. *Intelligent Automation and Soft Computing* 30(3): 805-819 (2021).
21. A. Naseem, M.A. Rehman, and T. Abdeljawad. Real-world applications of a newly designed root-finding algorithm and its polynomiography. *IEEE Access* 9: 160868 (2021).
22. D. Jain. Families of Newton-like methods with fourth-order convergence. *International Journal of Computer Mathematics* 90(5): 1072-1082 (2013).
23. P. Sivakumar and J. Jayaraman. Some new higher order weighted Newton methods for solving nonlinear equation with applications. *Mathematical and Computational Applications* 24(2): 59 (2019).
24. F. Soleymani. Efficient optimal eighth-order derivative-free methods for nonlinear equations. *Japan Journal of Industrial and Applied Mathematics* 30(2): 287-306 (2013).
25. I.K. Argyros, M. Kansal, V. Kanwar, and S. Bajaj. Higher-order derivative-free families of Chebyshev–Halley type methods with or without memory for solving nonlinear equations. *Applied Mathematics and Computation* 315: 224-245 (2017).
26. J. Li, X. Wang, and K. Madhu. Higher-order derivative-free iterative methods for solving nonlinear equations and their basins of attraction. *Mathematics* 7(1): 1052 (2019).
27. R. Meghwar, Z.A. Kalhoro, and S. Jamali. Computationally Efficient Three-Step Derivative-Free Iterative Scheme for Nonlinear Algebraic and Transcendental Equations. *Quest Research Journal* 23(01): 38-45 (2025).

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 323-336 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)696

Pakistan Academy of Sciences

Research Article

Hybrid Supervised Machine Learning Models for Enhanced Alzheimer's Disease Classification

Muazzam Ali[1*], M.U. Hashmi[1], Zakeesh Ahmad[2], Noor Ul Ain Kazmi[2], Asifa Ittfaq[2], and Amna Ashraf[2]

[1*]Department of Computer Sciences, Superior University, Lahore, Pakistan
[2]Department of Basic Sciences, Superior University, Lahore, Pakistan

Abstract: This research aims to facilitate the early and precise identification of Alzheimer's disease (AD), which remains one of the most prevalent neurodegenerative diseases impacting people's health and quality of life around the world. Employing machine learning algorithms, this study aims to develop reliable and effective models that support clinical workflows and streamline processes, thereby reducing the burden on patients and their families and ultimately enhancing patient-centric diagnostic frameworks. An approach to data cleaning, involving data imputation, encoding categorical variables, normalization of certain features, and stratified training and testing data splitting with hyperparameter tuning, was employed. This approach utilized both grid search and stratified k-fold cross-validation. Traditional models, ensemble techniques, and hybrid models were tested, including Lasso + LightGBM, XGBoost + SVM, and blended models such as LightGBM, CatBoost, Logistic Regression, and XGBoost. Lasso + LightGBM outperformed others in hybrid models. Lasso + LightGBM achieved an accuracy of 0.961240, precision of 0.943231, recall of 0.947368, and F1score of 0.945295, Cohen's Kappa of 0.915284, Hamming Loss of 0.038760, and Jaccard Index with the value of 0.896266. This research contributes to UNSDG 3, "Good Health and Well-being", by enhancing data-driven health education and resources, and an efficient diagnostic and management system for Alzheimer's. It also promotes healthy aging globally among the population.

Keywords: Predictive Modeling, Biomedical Data Analysis, Feature Engineering, Gradient Boosting, Clinical Decision Support, Cross-Validation, Diagnostic Accuracy.

1. INTRODUCTION

Alzheimer's is a behavior and progressive dementia disorder that impacts behavior, and thinking to a major extent and memory. It is the most common form of dementia, which induces tremendous loss of cognitive ability as people grow older [1]. Diagnosing Alzheimer disease is challenging as it can resemble the aging process or other brain-related diseases. In modern times, diagnosis is made through cognitive tests, brain scans, as well as clinical examinations, which are subjective and time-consuming [2, 3]. It does not have a single conclusive test, which is why detecting it early is a challenge, as it is crucial to the treatment and management of the condition. Following advances in machine learning (ML), a potent tool has emerged for enhancing the diagnosis of Alzheimer's disease by analyzing large and complex medical data. Patterns in the patient data have been drawn using traditional statistical methods and simple ML algorithms like the Naive Bayes and K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). These methods produce fast results; however, as with high-dimensional data, such as brain scans and genetic data, the methods are not particularly effective, which restricts their accuracy [4]. Ensembles and deep learning are sophisticated machine learning methods that help to mitigate these challenges. Cloud random Forests and gradient boosting are ensemble models that involve using a combination of models to improve the accuracy of predictions [5-7]. Deep learning-based models, such as Convolutional Neural Networks

Received: July 2025; Revised: November 2025; Accepted: December 2025
* Corresponding Author: Muazzam Ali <muazzamali@superior.edu.pk>

(CNNs), are indeed powerful tools that enable the processing of medical images and the detection of subtle changes in medical imaging (e.g., MRI, PET) associated with Alzheimer's disease. It is possible to improve patient outcomes by enhancing diagnosis accuracy and reducing the diagnosis period, thereby decreasing the risks of human error and leading to a better situation for clients. By providing better, more accurate, and timely diagnostics, researchers will be able to improve both treatment strategies and disease prevention [8, 9].

Current techniques of Alzheimer's disease (AD) diagnosis predominantly focus on genetic factors that involve machine learning and deep learning models, particularly by analyzing gene expression data for early detection of the disease. Studies have shown that deep learning (DL) models, including DGS-TabNet, outperform traditional ML algorithms by selecting more precise and efficient meaningful genes, obtaining superior classification performance (up to 93.8% accuracy and 98.53% Area under Curve (AUC) in binary classification tasks). Moreover, some key genes may also have biological significance by revealing their roles in other diseases, which could partly confirm that the use of network-based analyses in conjunction with traditional methods is valuable for identifying genetic markers related to AD [10]. Alzheimer's disease prediction has been significantly enhanced by recent machine learning algorithms, particularly those utilizing ensemble models (e.g., LightGBM and Random Forest), which can achieve accuracies exceeding 96.35% on several databases [11]. The use of Shapley Additive Explanation (SHAP) and Local Interpretable Model-agnostic Explanation (LIME) enhances artificial intelligence (AI) explainability, and as a result, the model's transparency leads to higher clinician trust in it. Compared to existing methods that are restricted by the number of datasets, data type, or interpretability, this method has improved efficiency and usability in AD diagnosis [12]. Mahamud *et al*. [13] developed a framework that uses data on handwriting to detect Alzheimer's disease, which involves a two-phase forward-backward selection of features via XGBoost. This strategy limits the workflow to a minimal set of tasks to increase interpretability to achieve 91.37% accuracy. The robust performance by using the leave-one-out cross-validation indicates that the sample size was adequate and transforms towards more friendly AD diagnosis. The present study also provides autography as a more reliable and straightforward strategy for early detection of AD.

The proposed research problem in the present study is the Computer-Aided Diagnosis (CAD) of Alzheimer's disease, which is addressed by designing and testing hybrid supervised machine learning models that combine adaptive feature selection, blended probability fusion, and gradient boosting. Responses to existing research have proven encouraging with the use of individual classifiers and the simple ensemble technique; however, they often fail to address high-dimensional, imbalanced, and heterogeneous clinical data, which ultimately results in poor generalizability and reduced clinical interpretability. To address these weaknesses, this work generalizes gradient boosting in a meta-modeling system, which has enhanced the robustness, discrimination, and interpretability of both linear and nonlinear learners.

The dataset used in the present study is the result of less controlled environments, specifically community-based and non-specialist clinical environments, where the data may be noisier, less standardized, and even completely missing, compared to strictly controlled research studies. This feature drove the adoption of hybrid designs that can tolerate uncertainty and variability while preserving the performance of diagnosis. In this connection, the objectives of this study will be the following:

- To build and test a set of hybrid machine learning models to classify Alzheimer's disease, which incorporate feature selection (i.e., Lasso) with effective gradient-boosting algorithms (i.e., LightGBM, XGBoost, CatBoost).
- To evaluate the capabilities of such hybridization in terms of predictive reliability and robustness, in comparison with standalone methods and conventional ensemble methods reported in recent literature.
- To ensure that the final models can be interpreted clinically, where interpretability is measured by the sparsity of the chosen features and the transparency of the linear elements in the hybrid structures.

The present study focuses on integrating and benchmark existing strategies to address the issue in the Alzheimer's CAD system. These issues

include data heterogeneity, small sample size and transparency of the model. Rather than proposing the new model, the approach in the present study aims to increase the effectiveness of current models, by developing the ML models that are clinically viable and applicable in practice.

2. METHODOLOGY

2.1. Dataset and Preprocessing

The Alzheimer's disease dataset, which was submitted to Kaggle by Rabie El Kharoua in 2024 and is released under the Attribution 4.0 International (CC BY 4.0) license (DOI: 10.34740/KAGGLE/DSV/8668279), is utilized in this research. 35 variables, including demographic, lifestyle, medical history, cognitive evaluation, symptoms, and diagnostic information pertaining to Alzheimer's disease, are included in the dataset, which includes 2,149 patient records (IDs 4751-6900). Because it is a binary variable that indicates whether Alzheimer's disease is present (1) or not (0), the diagnosis column is the target variable.

2.1.1. Handling missing values

Missing values in the dataset can compromise the reliability of model predictions. Therefore, all missing data are imputed using the mode (i.e., the most frequent value) for each column [14]. This approach is mathematically expressed as:

$$\hat{q}_i = mod(q_{i,1}, q_{i,2}, q_{i,3}, \dots \dots \dots \dots q_{i,n}) \quad (1)$$

Where $\hat{q}_i$ denotes the imputed value for feature i, while n represents samples. This method ensures the categorical and numerical integrity of the dataset, preserving both the sample size and variance structure.

2.1.2. Categorical encoding

To transform categorical variables into a numerical format, Label Encoding is applied to all features except the target column [15]. Each category is mapped to a unique integer, enabling the models to process categorical features mathematically:

$$\text{Encoded}(x) = i, \text{where } x \in \text{Categories}, i \in N \quad (2)$$

2.1.3. Normalization

For all continuous features, normalization using the Standard Scalar is performed, transforming the data to have a zero mean and unit variance [16].

$$z = \frac{x-\mu}{\sigma} \quad (3)$$

where σ is the standard deviation, μ is the mean, and x is the initial value for each feature. To guarantee that feature-scaling-sensitive models (like SVM and KNN) operate at their best, this step is essential.

2.1.4. Feature importance

The features of the Alzheimer's disease dataset have been ranked based on the scores of feature importance from the model using Random Forests, as illustrated in Figure 1. Random Forest has been used because the dataset is not very large, and it is capable of handling a large number of features without any problem. Functional Assessment and ADL (Activities of Daily Living) were the

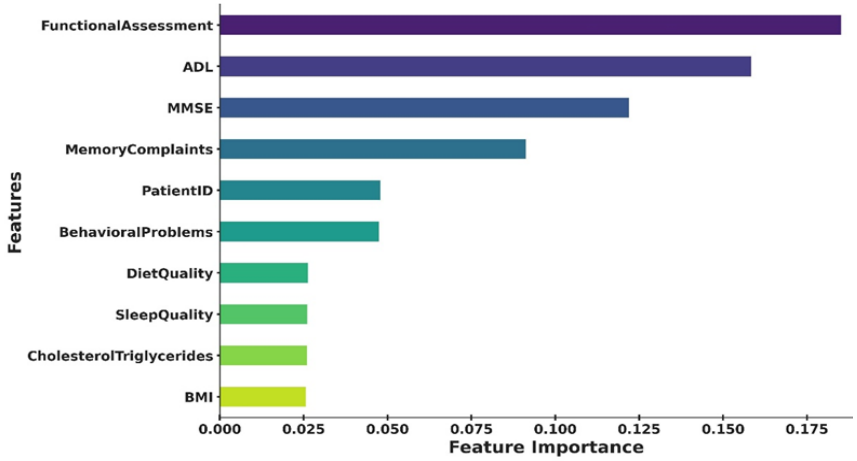


Fig. 1. Top 10 feature importance for Alzheimer's classification.

most significant factors. Therefore, they are the most important in predicting whether a case is Alzheimer's disease or non-Alzheimer's disease.

The other characteristics, such as the MMSE (Mini-Mental State Examination) and Memory Complaints, also play a significant role, showing that they are important in the clinical assessment of cognitive abilities. Conversely, the importance of features such as Cholesterol/Triglycerides, Sleep Quality, and Diet Quality is lower, which is a sign of weakness in these variable predictors in the dataset. This distribution is logical, given that functional and cognitive assessments are primary constituents for diagnosing Alzheimer's disease, thereby confirming the dataset's primary clinical relevance.

2.2. Data Splitting

A stratified train-test split is utilized to maintain class distribution in both sets. 70% of the data is allocated for training (Xtrain,ytrainX_{train}, y_{train}Xtrain,ytrain), and 30% for testing (Xtest,ytestX_{test}, y_{test}Xtest,ytest), ensuring that performance metrics generalize to unseen data.

$$(X_{train}, Y_{train}), (X_{test}, Y_{test}) = \text{StratifiedSplit}(X, Y, \text{test}_{score} = 3.0) \quad (4)$$

2.3. Model Training and Hyperparameter Tuning

A variety of supervised learning models are compared, with a particular focus on hybrid models developed by combining model outputs or feature selection pipelines. We performed hyperparameter optimization using GridSearchCV with stratified k-fold cross-validation (k = 5) to optimize precision and recall. We aimed to optimize the F1-score as the basic criterion for the model selection. In this process, stratified fold cross-validation was used to preserve the properties of class, decreasing the risk of overfitting. Moreover, this strategy ensured that hyperparameter estimation remains robust.

2.4. Used Models

We trained models using grid search with traditional classifiers, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression, and boosting and bagging techniques (XGBoost, LightGBM, CatBoost, AdaBoost, and Bagging Classifier). These may be used as standalone benchmarks or in conjunction with hybrid approaches. The model parameters are listed in Table 1.

2.4.1. K-nearest neighbors (KNN)

KNN is a non-parametric, instance-based algorithm where classification is based on the majority vote among the k closest training samples in the feature space [17]. The value of k is selected via grid search. The distance metric, typically Euclidean, is calculated as:

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2} \quad (5)$$

The size of the data affects this approach; hence, the previously mentioned normalization step is required. The curse of dimensionality can cause KNN's performance to deteriorate in high-dimensional environments, yet it is still a useful baseline for tabular datasets with modest complexity [18].

2.4.2. AdaBoost

Adaptive Boosting, also known as AdaBoost, is a technique that builds a powerful classifier by repeatedly training weak learners, typically decision stumps. However, each new learner is modeled after its predecessors, focusing on their mistakes [19]. The last model is the weighted sum of such learners:

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \quad (6)$$

Where α_t is the weight assigned to weak classifier $h_t(x)$. AdaBoost is especially robust to overfitting in many practical cases, but can be sensitive to noisy data and outliers.

2.4.3. Bagging (bootstrap aggregating)

Bagging trains multiple base estimators on different bootstrap samples of the dataset and averages their predictions to reduce variance. For binary classification:

$$\hat{y} = majority\_vote(h_1(x), h_1(x), \ldots, h_n(x)) \quad (7)$$

This strategy makes the models more stable especially, when using high variance base

Table 1. Hyperparameters tuned and their grid search values for each machine learning model.

Model	Hyperparameter Name	Hyperparameter Values
RandomForestClassifier	n_estimators, max_depth, min_samples_split, min_samples_leaf, bootstrap	100, 10, 2, 1, True
SVM (Support Vector Machine)	C, kernel, gamma, degree, coef0, tol	1, rbf, scale, 3, 0.0, 1e-3
KNN (K-Nearest Neighbors)	n_neighbors, weights, algorithm, leaf_size, p	5, uniform, auto, 30, 2
LogisticRegression	C, penalty, solver, max_iter, tol	1, l2, lbfgs, 100, 1e-3
XGBoost	n_estimators, learning_rate, max_depth, subsample, colsample_bytree, gamma	100, 0.1, 6, 0.8, 0.8, 0.1
LightGBM	n_estimators, learning_rate, max_depth, num_leaves, min_child_samples, subsample	100, 0.1, 6, 31, 20, 0.8
CatBoost	iterations, learning_rate, depth, l2_leaf_reg, subsample, colsample_bylevel	100, 0.1, 6, 3, 0.8, 0.8
AdaBoost	n_estimators, learning_rate, algorithm	100, 1.0, SAMME.R
Bagging	n_estimators, max_samples, max_features, bootstrap, n_jobs	100, 1.0, 1.0, True, -1
StackingClassifier	estimators, final_estimator, cv	RandomForestClassifier, XGBClassifier, LogisticRegression, 5
RF + Logistic Regression (Stacked)	rf__n_estimators, rf__max_depth, rf__min_samples_split, rf__min_samples_leaf, lr__C, lr__penalty, lr__solver	100, 10, 2, 1, 1, l2, lbfgs
XGBoost + SVM (Stacked)	xgb__n_estimators, xgb__learning_rate, xgb__max_depth, svm__C, svm__kernel, svm__gamma	100, 0.1, 6, 1, rbf, scale
Lasso + LightGBM (Hybrid)	lasso__alpha, lgbm__n_estimators, lgbm__learning_rate, lgbm__max_depth, lgbm__num_leaves, lgbm__min_child_samples	0.1, 100, 0.1, 6, 31, 20
RF-FeatureSelection + LR (Hybrid)	rf__n_estimators, rf__max_depth, rf__min_samples_split, rf__min_samples_leaf, lr__C, lr__penalty, lr__solver	100, 10, 2, 1, 1, l2, lbfgs
Blended Probabilities (LGBM + CatBoost + XGB) + LR	lgbm__n_estimators, lgbm__learning_rate, catboost__iterations, catboost__learning_rate, xgb__n_estimators, lr__C	100, 0.1, 100, 0.1, 100, 1

This rigorous methodology underpins both the fairness and scientific validity of model comparison, ensuring that reported results are robust, replicable, and meaningful for biomedical decision-making.

learners like decision trees. Hyperparameters (e.g. estimators) are optimized with the help of cross-validation [20].

2.4.4. Logistic regression

Standard Logistic Regression is used as a core linear baseline [21]. It estimates the probability of the binary outcome using the logistic function:

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(\beta_0 + \beta^T x))} \quad (8)$$

When the coefficients of β are estimated using the maximum likelihood method. C is a parameter that is regularized to control the model's complexity. Despite being linear, Logistic Regression is likely to compete with biomedical data and provide understandable coefficients.

It is not new to use some of the models employed in the present study; however, when applied to a comparatively strict and data-driven technique for Alzheimer's disease, which has high dimensionality and noise, they are instructive in science. Not merely accumulating, but this choice supposes the potential of an orderly examination

of model action and hybrid synergy, by which empirically information on what architectures will be evident in most clinically diverse situations. This is the gap, which is negatively addressed in the literature.

2.5. Hybrid Model Architectures

The study constructs and evaluates five advanced hybrid models, each leveraging the strengths of its constituent algorithms to address the nonlinearity, feature interaction, and potential collinearity within the dataset.

2.5.1. Hybrid 1: Random forest probabilities as features for logistic regression (RF + LR)

First, a Random Forest classifier is trained on the original feature set, outputting class probabilities for each sample:

$$P_{RF}(y = 1 \mid x) = \frac{1}{n_{trees}} \sum_{t=1}^{n_{trees}} h_t(x) \quad (9)$$

Where $h_t(x)$ is the prediction probability from tree t. The predicted probability P_{RF} is then appended as a new feature to both the training and test datasets:

$$X' = [X, P_{RF}] \quad (10)$$

The hybrid RF+LR model follows a two-stage stacking formulation. Consider $f_{RF}(x)$ is the random forest probability estimator then:

$$f_{RF} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \quad (11)$$

We produce out of fold (OOF) predictions by using:

$$\hat{p}_{RF,i} = f_{RF}^{(-k)}(x_i) \quad (12)$$

The meta feature matrix becomes:

$$X^{RF} = [X, \hat{p}_{RF}] \quad (13)$$

Now the logistic regression function for the decision is given by:

$$f_{LR}(X^{RF}) = \sigma(\beta_0 + \beta^T X + \gamma \hat{p}_{RF}) \quad (14)$$

γ represent the weight assigned to RF-derived probability, so the final hybrid prediction is computed using:

$$\hat{y} = 1\{f_{RF}(X^{RF}) > 0.5\} \quad (15)$$

A Logistic Regression model is subsequently trained on X′, learning a linear boundary in the enriched feature space. This hybridization combines the nonlinear feature extraction capability of Random Forests with the interpretability and regularization strength of Logistic Regression. The hybrid model can potentially address nonlinearity and feature interactions missed by Logistic Regression alone. However, there is a risk of overfitting if the new probability feature is highly correlated with the target, particularly in small or unbalanced datasets. In this study, cross-validation and the use of the test set mitigate such risks [22].

2.5.2. Hybrid 2: XGBoost probabilities as features for SVM (XGBoost + SVM)

An XGBoost model, known for its gradient-boosted tree structure and robustness to feature collinearity, is first trained. The predicted probabilities for each sample, P_{XGB}, are calculated:

$$P_{XGB}(y = 1 \mid x) = \sigma(f_{XGB}(X)) \quad (16)$$

Where σ denotes the sigmoid function. These probabilities are appended as an additional feature to the input matrix, after which a Support Vector Machine (SVM) classifier is trained and OFF probabilities $\hat{p}_{XGB}$ are concatenated with the input features:

$$X'' = [X, P_{XGB}] \quad (17)$$

The SVM with a radial bases function (RBF) kernel learns separating hyperplane in the augmented space:

$$f_{SVM}(X'') = sign(w^T X) + \gamma \hat{p}_{XGB} + b \quad (18)$$

The term $\gamma \hat{p}_{XGB}$ quantifies the contribution of initial stage boosted the probabilities to SVM margin. This hybrid combines XGBoost's nonlinear learning capacity with the margin-maximizing properties of SVMs. This approach can significantly enhance performance if XGBoost probabilities encapsulate a high-level structure that is not easily captured by SVM alone. However, SVMs are sensitive to irrelevant features, so the benefit depends on the informativeness of the probability feature [23].

2.5.3. Hybrid 3: Lasso feature selection followed by LightGBM (Lasso + LightGBM)

A Logistic Regression model with L1 regularization (Lasso) is employed to perform feature s A Logistic Regression model with L_1 regularization (Lasso) is employed to perform feature selection:

$$min_\beta = (-logL(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|) \quad (19)$$

Where $L(\beta)$ is the likelihood, β_j are the coefficients, and λ is the regularization parameter. Only features with nonzero coefficients are retained:

$$S = \{j : \beta \neq 0\} \quad (20)$$

The reduced feature matrix is:

$$X^{Lasso} = X[:, S] \quad (21)$$

LightGBM is trained on the reduced space:

$$\hat{y} = f_{LGBM}(X^{Lasso}) \quad (22)$$

This hybrid is a sequential architecture an optimizing based selector followed by the gradient boosting. LightGBM, a fast and efficient gradient boosting implementation, is trained on the selected features. This hybrid is especially effective in high-dimensional data, as it removes redundant and noisy variables before applying a strong tree-based learner. The risk is that overly aggressive feature selection can discard weak but informative features, potentially lowering overall model capacity [24].

2.5.4. Hybrid 4: Top N random forest feature importance with logistic regression (RF-Feature Selection + LR)

Random Forests naturally provide feature importance measures based on mean decrease in impurity (MDI) or mean decrease in accuracy (MDA). Random forest computed the importance values by:

$$I_j = \sum_{t=1}^{T} \sum_{s \in S_{t,j}} \Delta i(s) \quad (23)$$

The top N features with the highest importance scores are selected:

$$S_N = argsort(Importance_{RF})[:N] \quad (24)$$

Logistic regression is trained on:

$$X^{RF} = X[:, S] \quad (25)$$

The model is then given by:

$$f_{LR}(X^{RF}) = \sigma(\beta_0 + \beta^T X^{RF}) \quad (26)$$

This hybrid is featuring selection driven linear model contrasting with fully nonlinear boosters. Logistic Regression is then trained on this reduced feature set. Selecting the most predictive variables reduces dimensionality and may improve generalization, especially for linearly separable relationships. However, feature importance scores can be unstable in the presence of multicollinearity or redundant predictors, and choosing N is somewhat heuristic [25].

2.5.5. Hybrid 5: Blended probabilities of multiple boosting models with logistic regression (Blended Probabilities + LR)

LightGBM, CatBoost, and XGBoost models are independently trained on the original dataset. For each sample, the predicted probabilities from each model are extracted:

$$P_{LGBM}(y = 1 \mid X) \quad (27)$$

$$P_{CAT}(y = 1 \mid X) \quad (28)$$

$$P_{XGB}(y = 1 \mid X) \quad (29)$$

These probabilities are concatenated with the original features to create a new, augmented feature space:

$$X''' = [X, P_{LGBM}, P_{CAT}, P_{XGB}] \quad (30)$$

Let the blended meta feature vector be:

$$z_i = \begin{bmatrix} P_{LGBM,i} \\ P_{CAT,i} \\ P_{XGB,i} \end{bmatrix} \quad (31)$$

The final model is:

$$f_{LR}(X'') = \sigma(\beta^T X + \alpha_1 P_{LGBM} + \alpha_2 P_{CAT} + \alpha_3 P_{XGB} + b) \quad (32)$$

This is the probabilistic blending architecture that combines diverse gradient boosting models.

A Logistic Regression model is trained on X‴, learning how to combine the output of diverse boosting models optimally. This method synthesizes predictions from heterogeneous boosting frameworks, enabling the final model to exploit differences in model behavior [26]. While potentially powerful, this approach increases the risk of overfitting if the boosting models themselves are highly correlated or overfit the training data.

The benefits of these hybrid models extend beyond the advantages of conventional classifiers (such as Random Forest and Logistic Regression) to more complex algorithms, including feature selection with Lasso, boosting on XGBoost, LightGBM, and CatBoost, as well as ensemble learning methods like Stacking and Blended Probabilities. The hybrid models that use the probabilities generated by one model as input for the other model are helpful for the consideration of complexities like intricate feature interactions and nonlinearity that providing a novel approach to increase the model performance. A stronger decision is achieved using combined models, such as RF + LR, XGBoost + SVM, and Lasso + LightGBM, which present a novel perspective for processing high-dimensional imbalanced data.

2.6. Evaluation Metrics

The performance and robustness of these classification models are evaluated using specific metrics. These provide complementary information, accurately reflecting the overall correctness of the model, while precision measures how many of the predicted positives are truly positive. Recall shows how many actual positives are identified correctly and the F1-score balances the tradeoff between false positive and false negative. Cohen's Kappa, Hamming loss, and Jaccard Index capture the nuances of agreement and multi-label performance. The use of these measures enables a more advanced and less biased assessment of predictive models in various situations under different data distributions [27].

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (33)$$

$$precision = \frac{TP}{TP + FP} \quad (34)$$

$$recall = \frac{TP}{TP + FN} \quad (35)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (36)$$

$$Cohen's\ Kappa = \frac{P_0 - P_e}{1 - P_e} \quad (37)$$

$$Hamming\ Loss = \frac{1}{N}\sum_{i=1}^{N} 1(y_i \neq \hat{y}_i) \quad (38)$$

$$Jaccard\ Index = \frac{|A \cap B|}{|A \cup B|} \quad (39)$$

This rigorous methodology underpins both the fairness and scientific validity of model comparison, ensuring that reported results are robust, replicable, and meaningful for biomedical decision-making.

3. RESULTS AND DISCUSSION

The cross-evaluation of model benchmarks reveals reasonable differences in various measures, indicating the impact of different machine learning and hybrid methods for classifying Alzheimer's disease. Table 2 presents the evaluation metrics values for all models. The best accuracy is reported for , CatBoost, and Lasso + LightGBM, both scoring 0.961240, closely followed by XGBoost 0.961041, LightGBM and stacking at 0.958140 and Blended Probabilities (LGBM + CatBoost + XGB) + LR at 0.956589. This identifies the better performance of gradient boosting-based and ensemble hybrid methods for classifying the disease status. On the other hand, the KNN (0.737984) and RF-FeatureSelection + LR (0.846512) models exhibit relatively lower accuracy, which stems from high dimensionality and the sensitivity to feature selection, respectively. The accuracy achieved in this research is slightly higher than previous values of 0.9635 reported by Mahamud *et al.* [13] and 0.9380 recorded by Jin *et al*. [10].

Table 2 shows that the highest precision is recorded for CatBoost (0.951111) and XGBoost + SVM (0.950893), which are higher than the previous values of 0.95 stated by Mahamud *et al*. [13] and 0.9396 (with proposed model), reported by Jin *et al*. [10]. Both are effective in minimizing false positive rates and thereby curtailing diagnosis overestimation, which is crucial for less invasive procedures in clinical practice. Traditional classifiers, such as SVM (0.774336) and KNN (0.680982), perform markedly worse and are often unable to manage the class imbalance and complexity of features, despite normalization.

Table 2. Performance metrics (Accuracy, Precision, Recall, F1-Score) for various machine learning models evaluated in Alzheimer's disease classification.

Model	Accuracy	Precision	Recall	F1-Score
RandomForest	0.941085	0.943925	0.885965	0.914027
SVM	0.838760	0.774336	0.767544	0.770925
KNN	0.737984	0.680982	0.486842	0.567775
LogisticRegression	0.838760	0.787037	0.745614	0.765766
XGBoost	0.961041	0.941831	0.943468	0.941155
LightGBM	0.958140	0.938865	0.942982	0.940919
CatBoost	0.961240	0.951111	0.938596	0.944812
AdaBoost	0.927132	0.891775	0.903509	0.897603
Bagging	0.947287	0.925439	0.925439	0.925439
Stacking	0.958140	0.942731	0.938596	0.940659
RF + LR	0.945736	0.940639	0.903509	0.921700
XGBoost + SVM	0.959690	0.950893	0.934211	0.942478
Lasso + LightGBM	0.961240	0.943231	0.947368	0.945295
RF-FeatureSelection + LR	0.846512	0.781659	0.785088	0.783370
Blended Probabilities (LGBM + CatBoost + XGB) + LR	0.956589	0.942478	0.934211	0.938326

The XGBoost and Lasso + LightGBM achieved the highest values of recall, that are 0.943468 and 0.947368, respectively, that is higher than the value of 0.9380, reported by Jin *et al*. [10]. This aspect is crucial in clinical practice, where this kind of performance is needed to minimize the number of missed cases. Models such as the KNN model (score = 0.486842) have vast potential for further improvement, indicating that a simple model is underfitted in the presence of complex data.

As shown in Table 2, XGBoost (0.941155), CatBoost (0.944812), and Lasso + LightGBM (0.945295) achieved the highest F1-score, indicating that they can balance the precision-recall tradeoff better than other models, which is crucially important for medical diagnosis. The error spread is small; therefore, we can expect good accuracy from these algorithms.

Table 3 presents the Cohen's Kappa values of hybrid and ensemble approaches, including XGBoost (0.915284), CatBoost (0.914946), and Lasso + LightGBM (0.915284), which demonstrate considerable reliability in model classification consistency and performance, as well as reasonable performance. With Kappa point classification, the SVM (0.646527) and KNN (0.387154) are considered too soft, indicating that both have insufficient reliability to validate incomplete agreement. The hamming loss value is decreased with perfect classification and is particularly low when models XGBoost (0.038760), Catboost (0.038760) and Lass + LightGBM (0.038760) outperform the other models. As expected, KNN, due to its loss, suffers significant losses, which remain at 0.262016, primarily due to poor recall and precision, resulting in numerous mismatches. The three algorithms, XGBoost, CatBoost, and Lasso + LightGBM, scored the best with scores of 0.896266, 0.895397, and 0.896266, respectively, indicating that they have better predictive ability than other models and align more closely with the predicted true label. Many traditional and hybrid strategies like KNN (0.39642) and RF-feature Selection + LR (0.64388) performed below the chance level as expected due to their lower overall classification performance.

These results support the reasoning behind the methodology's focus on ensembles of hybrid models, as the integration of feature selection with probabilistic augmentation and gradient boosting is expected to improve performance significantly. The dataset underwent extensive preprocessing, including the meticulous imputation of missing values, label encoding, normalization, and stratified train-test splitting, which preserved class

Table 3. Cohen's Kappa, Hamming Loss, and Jaccard Index scores for different machine learning models in Alzheimer's disease classification.

Model	Cohen Kappa	Hamming Loss	Jaccard Index
RandomForest	0.869284	0.058915	0.841667
SVM	0.646527	0.161240	0.627240
KNN	0.387154	0.262016	0.396429
LogisticRegression	0.642971	0.161240	0.620438
XGBoost	0.915284	0.038760	0.896266
LightGBM	0.908506	0.041860	0.888430
CatBoost	0.914946	0.038760	0.895397
AdaBoost	0.841049	0.072868	0.814229
Bagging	0.884671	0.052713	0.861224
Stacking	0.908324	0.041860	0.887967
RF + LR	0.880208	0.054264	0.854772
XGBoost + SVM	0.911455	0.040310	0.891213
Lasso + LightGBM	0.915284	0.038760	0.896266
RF-FeatureSelection + LR	0.664523	0.153488	0.643885
Blended Probabilities (LGBM + CatBoost + XGB) + LR	0.904834	0.043411	0.883817

proportions to ensure the data's integrity while enhancing model generalizability. Grid search with stratified cross-validation for class-preserved folds enabled extensive multi-criteria hyperparameter optimization, minimizing the risk of overfitting and further augmenting model performance through fine-tuned hyperparameter adjustment.

The complicated nonlinear correlations observed in clinical and demographic data cannot be fully represented by simpler models such as KNN and Logistic Regression, in addition to the more traditional boundary-defining approximations and closest neighbor assumptions. The successful use of feature engineering and hyperparameter tuning has led to the development of clinical decision support tools for testing, highlighting the potential of complicated ensemble models for early Alzheimer's disease identification.

Figure 2 illustrates the pairwise distributions and interrelations between the significant predictors (Functional Assessment, ADL, MMSE, Memory Complaints, Behavioral Problems, and Sleep Quality) by diagnosis class. It is also easy to note clear differences between the Alzheimer and non-Alzheimer groups of the Functional Assessment, ADL, and MMSE, which indicates their great discriminative power. Contrastingly, Memory Complaints and Behavioral Problems have a higher overlap, meaning a lower predictive ability independently. Such visual trends are reflected in the rankings of feature importance gained with the help of Random Forest and Lasso selection, with functional and cognitive measures prevailing. Feature selection methods like mRMR and mutual information have also explained their efficiency in enhancing the prediction of Alzheimer's disease with an accuracy of 0.9908 [28].

More importantly, the figure also presents qualitative data on why the hybrid and ensemble models (e.g., Lasso + LightGBM) performed well: these models can learn nonlinear and partially collinear relationships between features, especially between cognitive and behavioral variables. Such curved or overlapping boundaries are not easily modeled using standard linear classifiers (e.g., Logistic Regression), which is why such classifiers achieve relatively low recall and F1 scores. That is why a pair-plot is not only justifying feature selection, but also the models' success, as it sets up the data structure visually and demonstrates where simple models may fail.

3.1. Model Behavior and Error Analysis

Lasso + LightGBM. L1 selection yielded a sparse

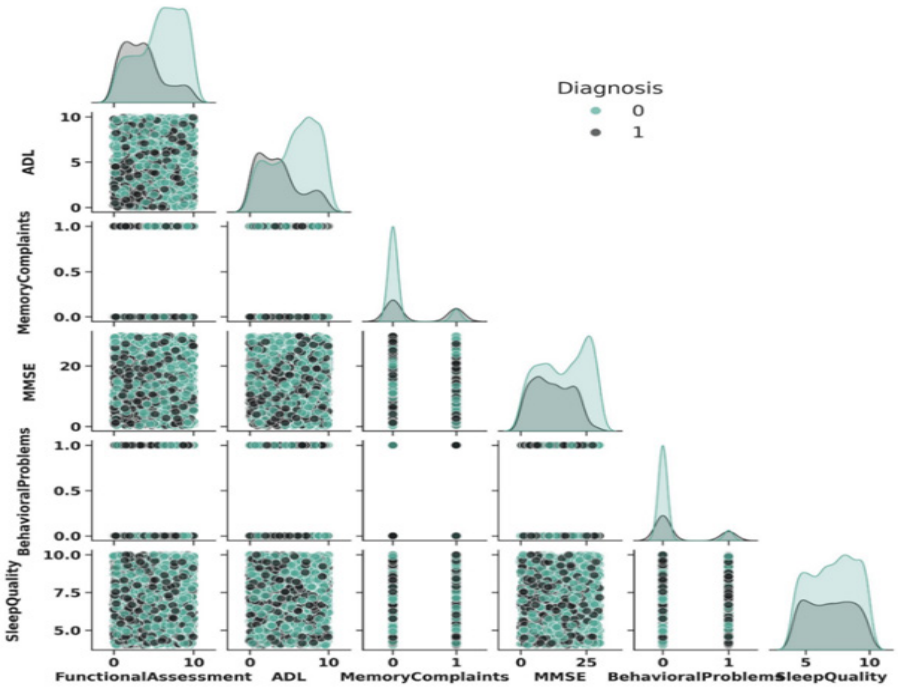


Fig. 2. Pairwise feature relationships by Alzheimer's diagnosis.

and lower-correlation subsample that reduces noise and redundancy; LightGBM then learned nonlinear interactions in this low-dimensional space, which are consistent with the more evident separations in the cases of Functional Assessment, ADL, and MMSE in Figure 2. Blended probabilities + LR. The base boosters had a high prediction correlation due to theoretical gains, which constrained the meta-learner's ability to be diverse. In a small sample size, the inclusion of correlated probability enhanced variance and decreased net benefit; moreover, variation in probability calibration was likely a restraining factor for the LR combiner. The combination of RF with Adaboost achieved 0.9255 accuracy which explained the benefits of ensemble learning in boosting model performance. The combination of DT, Adaboost and LR achieved highest accuracy of 0.9546 which shows the effectiveness of blending different models [29].

The study relies on a single dataset from Kaggle that may limit the generalizability of the model to clinical datasets. The models in the present study were evaluated only on provided dataset and external validation on an independent dataset was not performed. It is difficult to confirm the robustness and real-world applicability of the proposed models. Hybrid models such as Lasso + LightGBM and blended probabilities show strong performance; these may remain complex and less interpretable. This can limit their practical use in clinical settings where model transparency and interpretability are very important for clinical trust and decision-making.

RF-FeatureSelection + LR. RF importances based on impurity can be unstable under collinearity and biased against specific types of features; in a top-N heuristic, weak yet informative variables can be discarded. A linear LR fitted on this subset underfits the nonlinear structure, shown in Figure 2, which explains the gap between the accuracy and recall. Practical note: Future variants will (i) apply permutation/Boruta or stability selection to features, (ii) impose out-of-fold predictions and temperature/Platt calibration in blending, and (iii) take into account Elastic-Net LR or monotone-constrained boosting to make the thus far observed structure more like reality.

To evaluate the robustness, consistency and adaptability of the models, we used many established mechanisms. Robustness and generalization were assessed by using the stratified 5-folds cross validation, where models were trained and validate on multiple class preserving split ad by using was the out-of-fold (OOF) predictions to avoid the information leakage in hybrid stacking. Consistency was verified by using a various set of metrics like accuracy, precision, recall, F1score,

kappa, hamming loss, Jaccard index that showed stable rankings within the Table 2 and Table 3. Adaptability was evaluated testing the models on heterogeneous mix of demographic. Cognitive, behavioral and clinical features. Lastly, all results were confirmed on a held 30% unseen test set to ensure the valid generalization.

3.2. Comparative Discussion

Direct and cross-paper comparisons of point estimations (e.g., accuracy or F1) are necessarily constrained since results are highly dependent on the particular dataset (size, difficulty, feature set, and class balance) and preprocessing options, as well as the evaluation protocol. We therefore do not claim that we are better than previous studies solely because our point estimates (e.g., accuracy 0.961) are numerically larger than those obtained with other datasets and setups (e.g., 0.938). Rather, we place our findings on a par with ranges reported in recent literature on classifying ADs using gradient-boosted and hybrid ensemble classifiers, with overall similar levels of accuracy and F1 where tasks and data are similar [10, 12, 13].

Future research must incorporate evaluation on common publicly available benchmarks (e.g., using the same train/test splits with ADNI, OASIS, or the same Kaggle dataset). It also incorporates the standardization of preprocessing pipelines to reduce variability and measurement of uncertainty (e.g. per-split results and 95% CIs through bootstrapping) and paired-sample tests (e.g., McNemar test to establish accuracy, DeLong test to establish AUC). Calibration and decision-curve analyses to supplement the results are indicated within these limits, we find that hybrid strategies (e.g., Lasso + LightGBM) can produce state-of-the-art dataset competitive performance and practical interpretability in line with the trends of previous work [10, 12, 13].

4. CONCLUSIONS

The paper compared conventional, ensemble, and hybrid supervised classifiers in the classification of Alzheimer's disease using tabular clinical data. CatBoost and Lasso + LightGBM (accuracy = 0.96124) were the closest as the strongest point estimate, and XGBoost was considered the third closest (accuracy = 0.96104). All with a strong F1 (0.94 - 0.95). Since we did not report any measures of variance or formal tests of significance, we do not claim to have been statistically better than the other models; instead, the models can be viewed as those that perform best and are statistically equivalent, given the evidence at hand. On a methodological level, the results are congruent with the hypothesis that, with L1-Based selection, features may be denoised and decorrelated, allowing a gradient-boosting learner (LightGBM) to represent nonlinear feature interactions more effectively. Nevertheless, we have seen that the Lasso + LightGBM hybrid cannot be readily interpreted: Lasso produces sparse selections, but the black box model of the final boosted model remains a black box. Future studies will (i) quantify the uncertainty (per-fold results, bootstrap CIs, paired tests such as McNemar/DeLong) to find out whether small metric deltas are statistically significant; (ii) provide explanatory analyses (e.g. SHAP global summaries, local explanations, partial dependence/ICE, and calibration curves) to describe how the output of functional and cognitive measures drives the predictions; (iii) assess blending/stacking on out-of-fold meta-features and probability calibration to increase the diversity among base learners. These criteria suggest that gradient-boosted and hybrid studies are dataset-competitive in AD classification on structured clinical data, and that an additional investigation into uncertainty and explainability is necessary to make comparative or clinical assertions.

5. ACKNOWLEDGMENT

The authors owe a great deal to Superior University, Lahore, as it has facilitated this research work by providing the necessary materials and financial support. We appreciate the valuable comments and direction provided by colleagues from the Department of Basic Sciences and the Department of Computer Sciences.

6. ETHICAL STATEMENT

The protocols put in place ensured that all research work involving medical data was carried out ethically, employing the best practices. The protocols of this research were reviewed by the Research committee at Superior University Lahore that was chaired by Dr. Muhammad Azam. All materials in this study were confidential and required anonymity.

7. CONFLICT OF INTEREST

The authors have no conflict of interest regarding this article.

8. REFERENCES

1. D. Jadhav, N. Saraswat, N. Vyawahare, and D. Shirode. Targeting the molecular web of Alzheimer's disease: unveiling pathways for effective pharmacotherapy. *The Egyptian Journal of Neurology, Psychiatry and Neurosurgery* 60(1): 7 (2024).
2. M. Wang, Y. Lin, F. Gu, W. Xing, B. Li, X. Jian, C. Liu, D. Li, Y. Li, T. Wu, and D. Ta. Diagnosis of cognitive and motor disorder levels in stroke patients through explainable machine learning based on MRI. *Medical Physics* 51(3):1763-1774 (2024).
3. F. Öhman, J. Hassenstab, D. Berron, M. Schöll, and K. Papp. Current Advances in Digital Cognitive Assessment for Preclinical Alzheimer's Disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 13(1): 1-19 (2021).
4. P. El Kafrawy, H. Fathi, M. Qaraad, A.K. Kelany, and X. Chen. An efficient SVM-based feature selection model for cancer classification using high-dimensional microarray data. *IEEE Access* 9: 155353-69 (2021).
5. Q.U. Hamza, M.A. Baloch, M.A. Rajwana, A. Raza, and Z.U. Zia. Hybrid ensemble learning approaches for high-accuracy dementia detection: integrating deep learning models. *Kashf Journal of Multidisciplinary Research* 2(05): 66-83 (2025).
6. N. Rane, S.P. Choudhary, and J. Rane. Ensemble Deep Learning and Machine Learning: Applications, Opportunities, Challenges, and Future Directions. *Studies in Medical and Health Science* 1(2): 18-41 (2024).
7. H.T. Hoc, R. Silhavy, Z. Prokopova, and P. Silhavy. Comparing stacking ensemble and deep learning for software project effort estimation. *IEEE Access* 11: 60590-60604 (2023).
8. M.J. Iqbal, Z. Javed, H. Sadia, I.A. Qureshi, A. Irshad, R. Ahmed, K. Malik, S. Raza, A. Abbas, R. Pezzani, and J. Sharifi-Rad. Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. *Cancer Cell International* 21(1): 270-280 (2021).
9. S. Asif, Y. Wenhui, S. Ur-Rehman, Q. Ul-Ain, K. Amjad, Y. Yueyang, S. Jinhai, and M. Awais. Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision. *Archives of Computational Methods in Engineering* 32(2): 853-883 (2024).
10. Y. Jin, Z. Ren, W. Wang, Y. Zhang, L. Zhou, X. Yao, and T. Wu. Classification of Alzheimer's disease using robust TabNet neural networks on genetic data. *Mathematical Biosciences and Engineering MBE* 20(5): 8358-8374 (2023).
11. M. Chakraborty, N. Naoal, S. Momen, and N. Mohammed. ANALYZE-AD: A Comparative Analysis of Novel AI Approaches for Early Alzheimer's Detection. *Array* 22: 100352 (2024).
12. A.S. Alatrany, W. Khan, A. Hussain, H. Kolivand, and D. Al-Jumeily. An Explainable Machine Learning Approach for Alzheimer's Disease Classification. *Scientific Reports* 14(1): 2637-2654 (2024).
13. E. Mahamud, M. Assaduzzaman, J. Islam, N. Fahad, M.J. Hossen, and T.T. Ramanathan. Enhancing Alzheimer's disease detection: An explainable machine learning approach with ensemble techniques. *Intelligence-Based Medicine* 11(11): 100240 (2025).
14. T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona. A survey on missing data in machine learning. *Journal of Big Data* 8(1): 140 (2021).
15. M.K. Dahouda and I. Joe. A deep-learning embedding technique for categorical features encoding. *IEEE Access* 9: 114381-114391 (2021).
16. V. Sharma. A study on data scaling methods for machine learning. *International Journal for Global Academic & Scientific Research* 1(1): 31-42 (2022).
17. R.K. Halder, M.N. Uddin, M.A. Uddin, S. Aryal, and A. Khraisat. Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *Journal of Big Data* 11(1): 113-125 (2024).
18. A.A. Amer, S.D. Ravana, and R.A. Habeeb. Effective k-nearest neighbor models for data classification enhancement. *Journal of Big Data* 12(1): 86 (2025).
19. C.K. Reddy, P.A. Reddy, P.S. Reddy, M. Shuaib, S. Alam, S. Ahmad, and A. Rajaram. Twined ensemble framework for network security: integrating Random Forest, AdaBoost, and Gradient Boosting for enhanced intrusion detection. *Discover Internet of Things* 5(1): 107 (2025).
20. H. Şevgin. A comparative study of ensemble methods in the field of education: Bagging and boosting algorithms. *International Journal of Assessment Tools in Education* 10(3): 544-562 (2023).
21. J.C. Timoneda. Estimating group fixed effects in panel data with a binary dependent variable: How

the LPM outperforms logistic Regression in rare events data. *Social Science Research* 93(1): 102486 (2021).
22. A. Demircioğlu. Applying oversampling before cross-validation will lead to high bias in radiomics. *Scientific Reports* 14(1): 11563 (2024).
23. N.S. Nafis and S. Awang. An enhanced hybrid feature selection technique using term frequency-inverse document frequency and support vector machine-recursive feature elimination for sentiment classification. *IEEE Access* 9: 52177-52192 (2021).
24. T. Mahmood, A. Rehman, T. Saba, T.J. Alahmadi, M. Tufail, S.A. Bahaj, and Z. Ahmad. Enhancing Prognosis of Coronary Artery Disease: A Novel Dual-Class Boosted Decision Trees Strategy for Robust Optimization. *IEEE Access* 12: 107119-107143 (2024)**.**
25. R. Vishraj, S. Gupta, and S. Singh. Evaluation of feature selection methods utilizing random forest and logistic regression for lung tissue categorization using HRCT images. *Expert Systems* 40(8): e13320 (2023).
26. K.A. Ahmed, I. Humaira, A.R. Khan, M.S. Hasan, M. Islam, A. Roy, M. Karim, M. Uddin, A. Mohammad, and M.D. Xames. Advancing breast cancer prediction: Comparative analysis of ML models and deep learning-based multi-model ensembles on original and synthetic datasets. *PLOS One* 20(6): e0326221 (2025).
27. I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, and A. Borodulin. Boosting-based machine learning applications in polymer science: A review. *Polymers* 17(4): 499 (2025).
28. H. Alshamlan, A. Alwassel, A. Banafa, and L Alsaleem. Improving Alzheimer's Disease Prediction with different Machine Learning Approaches and Feature Selection Techniques. *Diagnostics* 14(19): 2237 (2024).
29. R.A. Gad and A. Abdelhafeez. Alzheimer's Disease Prediction using Hybrid Machine Learning Techniques. *SciNexuses* 1: 174-83 (2024).

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): 337-347 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253
http://doi.org/10.53560/PPASA(62-4)704

Pakistan Academy of Sciences

Research Article

Cd(II) Derivatives of Substituted Phenylacetic Acids, Synthesis, Spectroscopic Characterization and Binding Studies with DNA

Haleema Bibi[1,†], Aneeqa Shamim[1,†], Saba Naz[1], Moazzam Hussain Bhatti[2], Mahboob-ur-Rehman[3], Ali Haider[1], and Saqib Ali[1*]

[1]Department of Chemistry Quaid-i-Azam University, 45320, Islamabad, Pakistan
[2]Department of Chemistry, Allama Iqbal Open University, Islamabad, Pakistan
[3]Department of Cardiology, Pakistan Institute of Medical Sciences (PIMS), Islamabad, Pakistan

Abstract: The methoxy substituted phenylacetic acid (MeOPhA) and chloro substituted phenoxyacetic acid (ClPhA) were used to synthesize eight new Cd(II) based complexes. The nitrogen donor 2,2′- bipyridine (MeOPhA2, ClPhA2) and 1,10-phenanthroline (MeOPhA3, ClPhA3) were used as auxiliary ligands for the synthesis of mixed ligand complexes. These complexes were characterized by FT-IR and multinuclear NMR (^{1}H and ^{13}C-NMR) spectroscopic techniques. The FT-IR spectra of the complexes showed characteristic COO^- asymmetric and COO^- symmetric vibrational bands indicating metal coordination through oxygen. Moreover, their difference, i.e., Δν reveal that the selected ligands are coordinated to the Cd(II) center in a bidentate manner. The ^{1}H-NMR and ^{13}C-NMR data recorded in deuterated solvents also supported successful synthesis in pure form as well as metal coordination through carboxylate group. The nature of the complex–DNA interaction was examined, and the impact of hetero ligand attachment on binding strength and reactivity was assessed using UV–visible spectroscopy. The obtained data confirmed the effective binding ability through partial intercalation and groove binding through spontaneous process for all the complexes.

Keywords: Mixed Ligands, Spectroscopic Techniques, Auxiliary Ligands, Surface Binding, Multinuclear NMR.

1. INTRODUCTION

Metal complexes have been used in medicinal industry since ancient times; however, their pharmacological significance was firmly recognized after Rosenberg's 1969 discovery of cisplatin's anticancer activity [1]. Cisplatin's distinct method of action, which involves covalent interaction with DNA, has been attributed to its exceptional therapeutic success. This interaction ultimately inhibits the growth of cancerous cell by blocking the mechanisms required for their replication [2].
DNA is regarded as the blueprint of life, controlling and regulating a wide range of cellular metabolic activities [3]. Many other anticancer drugs such as Actinomycin D and Doxorubicin exert their effect by binding with DNA [4-6]. Nitrogenous bases of DNA show distinct preferences for metal cations, general stability order for 3d transition series is given as: M–guanine > M–adenine and M–cytosine > M–thymine [6, 7]. Chelation results in increase in drug absorption across cells by reducing metal ion polarity through orbital overlap and resonance. Hence, understanding these selective interactions of metal and DNA bases and the right selection of metal and ligands is necessary for developing advanced metallodrugs [8, 9]. Cadmium (Cd) is a d^{10} metal belonging to group 12 of periodic table with zero crystal field stabilization energy. It has no strong geometric preference due to filled d orbitals and can easily be identified through spectroscopy [10]. This is, however, categorized as a highly toxic heavy metal due to its strong affinity for sulfhydryl groups in protein which results in oxidative stress, enzyme inhibition and tissue damage. Recent studies have revealed that the toxicity of a metal is not a fixed property, it is influenced by various factors such as the ligand environment, oxidation

Received: November 2025; Revised: December 2025; Accepted: December 2025
* Corresponding Author: Saqib Ali <saqibali@qau.edu.pk>
† Both authors contributed equally to the work

state, and coordination geometry [11, 12]**.** Egorova and Ananikov [13] highlighted the role of the metal in the living systems, which is intrinsically linked to the specific molecular form in which the metal exists. Thus, toxicity of Cd(II) can be reduced by its complexation with suitable oxygen and nitrogen donor ligands which stabilize the Cd(II) center and can direct its biomolecular interaction in a controlled way [14]. A variety of important functionalities are associated with Cadmium complexes such as anti-microbial, anti-cancer**,** catalytic and anti-bacterial properties [15, 16].

In coordination chemistry, the choice of ligand is crucial because it affects the coordination behavior, stability, geometry, and biological activity of the desired complex [17]. Carboxylic acids are organic ligands of choice on account of their favorable chemistry especially the versatile coordination ability [18]. Carboxylic acids can form complex and stable structures by coordinating with metals in many ways such as ionic, monodentate, and bidentate. In biological and catalytic processes, their coordination flexibility is crucial [19]. Utilizing these medicinally active ligands in metal complexation has become a developing trend to create more potent and focused therapeutic agents because carboxylic acids are essential structural elements of many already available therapeutic agents [8, 20]. Cadmium carboxylates display flexible coordination geometries due to the large ionic radius of Cd^{2+}, i.e., 109 pm [21]. Due to this flexibility, these complexes find their applications in bio-sensing, bio-imaging, nanomedicine and drug delivery [22, 23].

Naturally occurring phenylacetic acid and its derivatives are known for their bioactivity and significant contribution to improving the flavor and scent of food and cosmetic items [24]. 2-methoxyphenylacetic acid contains a methoxy group in addition to carboxylic group attached to phenyl ring. The presence of these strong electron-donating and coordinating groups significantly enhance its reactivity as well as metal binding capabilities. The commonly used NSAIDs like diclofenac etc., with an aromatic carboxylate group, exhibited significant pharmacological and coordinating properties. This chemical resemblance allows the formulation of metal based pharmacologically active compounds by the incorporation of active functional groups [20, 25, 26]. 2,4-Dichlorophenoxyacetic acid contains a phenoxy oxygen atom in addition to carboxylate group offering versatile coordination modes thus making it suitable for forming stable metal complexes [27]. The synthesis of complexes with different donors and heterocyclic ligands is a current trend, inspired by biomacromolecules. Overall efficiency can be improved by using a heterocyclic donor as an auxiliary ligand and a carboxylate group as the main ligand.

N-donor heterocycles are regarded as stable and adaptable co-ligands because the lone pair on their sp^2-hybridized nitrogen [28, 29]. Both 2,2′-bipyridine and 1,10-phenanthroline are plannar ligands having sp^2 hybridized nitrogen as well as extended π-conjugation system enabling π–π stacking and other non-covalent interactions in resulting complexes. The ligand 2,2′-bipyridine contains trans-oriented nitrogen atoms, mostly forms slightly strained cis complexes which show diverse electronic and biological activity [30]. Whereas, 1,10-Phenanthroline contains cis-oriented nitrogen atoms that favor bidentate chelation with metal centers, thus making it significant in bioinorganic and therapeutic chemistry [31].

According to data found in the literature, the overall effectiveness of the resultant complexes is greatly increased by the addition of active structural motifs including metal centers, carboxylate ligands, and heterocycles containing nitrogen. Numerous mixed ligand complexes based on substituted aromatic carboxylic acids such as methoxyphenylacetic and dichlorophenoxyacetic derivatives have been reported. Consistently, both ligand form structurally unique and biologically relevant heteroleptic metal complexes when coordinated with N-donor co-ligands [8, 18, 20, 32].

The present research project is an attempt to synthesize mixed ligand complexes of cadmium by using substituted phenylacetic acids and N-donor ligands and to evaluate their ability to bind with DNA. The FT-IR and multi-nuclear NMR (^1H, ^{13}C) were employed for their characterization.

2. MATERIALS AND METHODS

Reactant used in the synthesis such as 2-methoxyphenylacetic acid (MeOPhA), 2,4-dichlorophenoxyacetic acid (ClPhA),

sodium bicarbonate, cadmium chloride, nitrogen donor ligands e.g. 2,2′- bipyridine and 1,10-phenanthroline were acquired from Sigma-Aldrich (USA) and were used as such. The solvents used during the synthesis, recrystallization and for NMR data collections include some organic solvents and n-hexane, etc., were of absolute purity and were acquired from Merck (Germany). They were utilized in all the experiments without any further purification processes. Gallen Kamp (UK) electrothermal apparatus was employed to find out the melting point of all complexes by using the capillary tubes. FT-IR Spectrophotometer of Thermo Nicolet-6700 was used to record FT-IR spectra of complexes in the range of 4000-400 cm^{-1}. Multi-nuclear NMR (^{1}H and ^{13}C) spectra of ligands and complexes were taken by Bruker Advanced Digital instrument having frequency of 300 MHz at room temperature in deuterated dimethyl sulfoxide (DMSO). Chemical shifts and coupling constants were noted in parts per million (ppm) and Hertz (Hz), respectively. The UV-Visible spectrophotometer (Shimadzu 1800) served to record the absorption spectra of the complexes for DNA binding analysis.

2.1. Synthetic Protocols

2.1.1. Procedure for ligand's sodium salts

To prepare sodium salts (see scheme 1) the aqueous solution of sodium bicarbonate **(**3 mmol, 0.252 g**)** was added dropwise to the aqueous solution of each ligand, i.e., 2-methoxyphenylacetic acid (3 mmol, 0.498 g) and 2,4-dichlorophenylacetic acid (3 mmol, 0.615 g) under continuous stirring. The mixtures were stirred maximum until neutralization at room temperature. The solvents were then evaporated under reduced pressure to get the solid sodium salts, which were collected and stored in glass vials. This synthesis procedure for the sodium salt is in accordance with the previously reported method [33]. The scheme 2 represents the structure of ligands used in synthesis along with the numbering scheme for NMR interpretation.

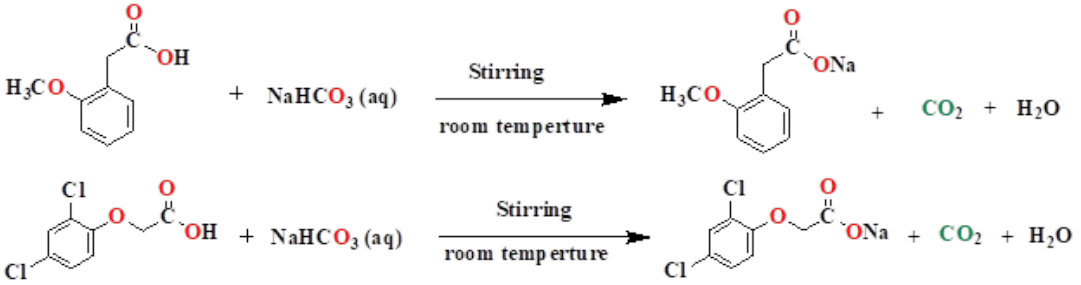
2.1.2. Synthesis of single ligand cadmium carboxylates

2.1.2.1. Synthesis of MeOPhA1

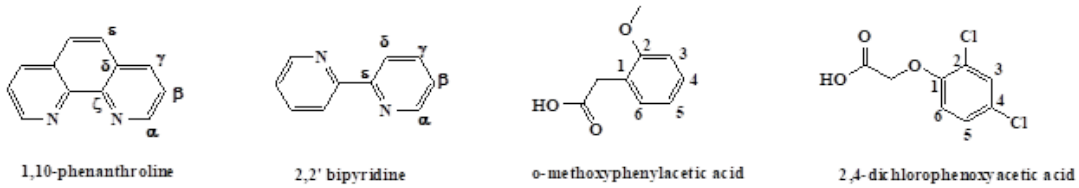
Under constant stirring, methanolic solutions of sodium salt of ligand MeOPhA (3 mmol, 0.564 g) were added into aqueous solution of cadmium chloride (1.5 mM, 0.275 g). The reaction mixtures were stirred for 5 hours at 50 °C, the resulting precipitates were obtained through filtration. They were washed with water to remove any impurity/ residual reactants and was dried in air. The procedure is presented in Scheme 3.

2.1.2.2. Synthesis of ClPhA1

The synthesis of the ClPhA1 was carried out by following the synthetic procedure as discussed for complex MeOPhA1. However, the sodium salt of ClPhA (3 mmol, 0.729 g) were added into aqueous solution of cadmium chloride (1.5 mM, 0.275 g) instead of MeOPhA. The product was also



Scheme 1: Synthesis of sodium salts of substituted phenylacetic acids.



Scheme 2: Numbering pattern for the ligands and nitrogen donor heterocycles.

processed in the same way and the synthetic route is presented in Scheme 3.

2.1.3. Synthesis of mixed ligand cadmium carboxylates

The sodium salt of ligand 2-methoxyphenylacetic acid (MeOPhA, 3 mmol, 0.564 g) and 2,4-dichlorophenoxyacetic acid (ClPhA, 3 mmol, 0.729 g) were dissolved separately in methanol. To this, an aqueous solution of cadmium chloride (1.5 mmol, 0.275 g) and bipyridine (1.5 mmol, 0.234 g) were added simultaneously for the synthesis of complexes MeOPhA2 and ClPhA2 respectively. The resulting mixtures were stirred for about 8 hours at 50 °C. The same procedure was used for the synthesis of complexes MeOPhA3 and ClPhA3 except that the addition of bipyridine was replaced by the phenanthroline (1.5 mmol, 0.270 g). The resulting solutions were filtered, extra solvents were removed through rotary evaporation, and the solid products were washed several times with water and dried in air. The obtained products were recrystallized from combination of appropriate solvents. Melting points were recorded for all the synthesized complexes. The synthetic route for complexes of both ligands and the corresponding NMR numbering scheme is presented in Scheme 4.

Cd(MeOPhA)$_2$: (MeOPhA1)

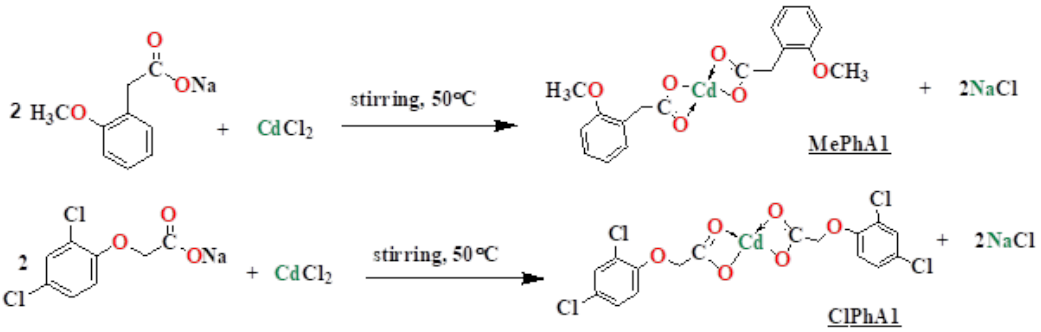
Solubility: Chloroform, DMSO, Methanol; M.P: 73-75 °C; % Yield: 78.1; FT-IR (cm^{-1}): 1582 (COO$_{asym}$), 1410 (COO$_{asym}$), 172 (Δν), 526 (Cd-O); ^{1}H NMR (DMSO-d$_6$, ppm): 3.20 (s, 2H, –CH$_2$), 3.70 (br, 3H, –OCH$_3$), 6.77-6.86 (m, 2H, H3, 5), 7.06-7.66 (m, 1H, H4), 7.09-7.15 (m, 1H, H6); ^{13}C NMR (DMSO-d$_6$, ppm): 175.4 (C=O), 39.0 (–CH$_2$), 55.6 (–OCH$_3$), 126.7 (C1), 157.5 (C2), 110.6 (C3), 128.6 (C4), 120.1 (C5), 131.2 (C6).

Cd(ClPhA)$_2$: (ClPhA1)

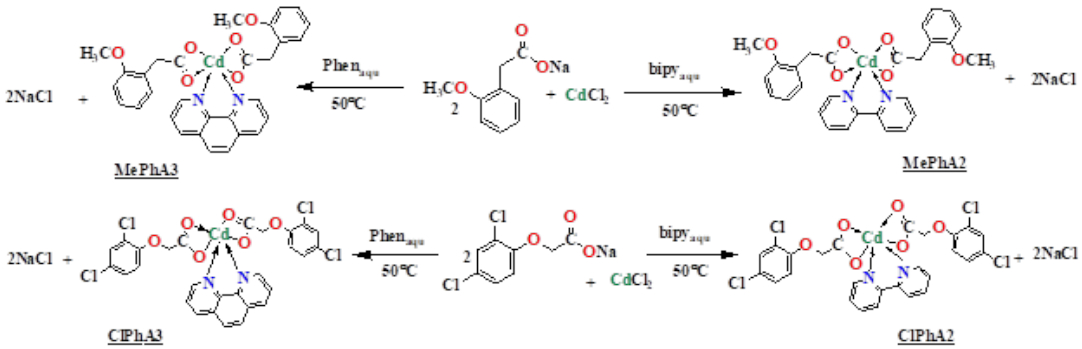
Solubility: DMSO, Ethanol, Methanol; M.P: 294–296 °C; %Yield: 78.3; FT-IR (cm^{-1}): 1598 (COO$_{asym}$), 1422 (COO$_{asym}$), 176 (Δν), 460 (Cd-O); ^{1}H NMR (DMSO-d$_6$, ppm): 4.28 (s, 4H, –OCH$_2$), 7.47 (s, 2H, H3), 7.24-7.27 (d, 2H, H5, J = 9 Hz), 6.84-6.87 (d, 2H, H6, J = 9 Hz);^{13}C NMR (DMSO-d$_6$, ppm): 170.7 (C=O), 68.7 (–OCH$_2$), 154.0 (C1), 129.2 (C2), 123.6 (C3), 127.9 (C4), 122.2 (C5), 115.4 (C6).

Cd(MeOPhA)$_2$(bipy): (MeOPhA2)

Solubility: Chloroform, DMSO, Methanol; M.P :68-70 °C, % Yield: 76.5; FT-IR (cm^{-1}): 1560 (COO$_{asym}$), 1386 (COO$_{asym}$) ,174 (Δν), 590 (Cd-N), 486 (Cd-O); ^{1}H NMR (DMSO-d$_6$, ppm): 3.22 (s, 2H, –CH$_2$), 3.70 (s, 3H, –OCH$_3$), 6.77-6.86 (m, 2H,



Scheme 3: Synthesis of single ligand complex derived from substituted phenylacetic acids.



Scheme 4: Synthesis of mixed Cd(II) carboxylates derived from 2-methoxyphenylacetic acids.

H3, 5), 7.06-7.15 (m, 2H, H4, 6), 8.68-8.69 (d, 2H, Hα, *J* = 4.8), 7.44-7.47 (m, 2H, Hβ), 7.92-7.98 (m, 2H, Hγ), 8.37-8.39 (d, 2H, Hδ, *J* =8.1Hz), ^{13}C NMR (DMSO-d$_6$, ppm): 175.7 (C=O), 39.1 (–CH$_2$), 55.6 (–OCH$_3$), 124.7 (C1), 149.7 (C2), 110.6 (C3), 126.8 (C4), 120.1 (C5), 128.4 (C6), 157.5 (Cα),120.9 (Cβ), 131.1 (Cγ), 137.8 (Cδ), 157.5 (Cε).

Cd(MeOPhA)$_2$(1,10-phen): (MeOPhA3)
Solubility: Chloroform, DMSO, Methanol; M.P: 75-77 °C; % Yield: 73.7; FT-IR (cm^{-1}): 1570 (COO$_{asym}$), 1390 (COO$_{sym}$), 180 (Δν), 607 (Cd-N), 517 (Cd-O); ^{1}H NMR (DMSO-d$_6$, ppm): 3.69 (br, 5H, –CH$_2$, –OCH$_3$), 6.76-6.86 (m, 2H, H3, 5), 7.06-7.14 (m, 2H, H4, 6), 9.08-9.10 (dd, 2H, Hα *J* = 1.5 Hz, 4.2 Hz), 7.79-7.83 (m, 2H, Hβ), 8.52-8.55 (dd, 2H, Hγ *J* = 1.5 Hz, 8.1 Hz), 8.02 (s, 2H, H); ^{13}C NMR (DMSO-d$_6$, ppm): 175.4 (C=O), 39.0 (–CH$_2$), 55.5 (–OCH$_3$), 126.7 (C1), 150.5 (C2), 110.6 (C3), 127.1 (C4), 120.1 (C5), 128.5 (C6), 157.5 (Cα), 124.0 (Cβ), 131.1 (Cγ), 137.0 (Cδ), 128.9 (Cε), 157.5 (Cζ).

Cd(ClPhA)$_2$(bipy): (ClPhA2)
Solubility: DMSO, Ethanol, Methanol; M.P: 125-127 °C;% Yield: 71.1; FT-IR (cm^{-1}): 1609 (COO$_{asym}$), 1419 (COO$_{sym}$), 190 (Δν), 556 (Cd-N), 475 (Cd-O); ^{1}H NMR (DMSO-d$_6$, ppm): 4.25 (s, 4H, –OCH$_2$), 7.43-7.48 (m, 4H, H3, Hβ), 7.23-7.27 (dd, 2H, H5, *J* = 2.7 Hz, 9Hz), 6.83-6.86 (d, 2H, H6, *J* = 9 Hz), 8.68-8.69 (d, 2H, Hα *J* = 3.9 Hz), 7.92-7.98 (td, 2H, Hγ, *J* = 1.8 Hz, 7.8 Hz), 8.37-8.40 (d, 2H, Hδ, 7.8 Hz); ^{13}C NMR (DMSO-d$_6$, ppm): 170.3 (C=O), 68.9 (–OCH$_2$), 154.0 (C1), 137.8 (C2), 124.6 (C3), 129.1 (C4), 123.4 (C5), 115.5 (C6). 149.7 (Cα), 120.8 (Cβ), 122.1 (Cγ), 127.9 (Cδ), 149.7 (Cε).

Cd(ClPhA)$_2$(1,10-phen): (ClPhA3)
Solubility: DMSO, Ethanol, Methanol; M.P = 140-142 °C; % Yield: 72.8; FT-IR (cm^{-1}): 1588 (COO$_{asym}$), 1422 (COO$_{sym}$), 166 (Δν), 584 (Cd-N), 461(Cd-O); ^{1}H NMR (DMSO-d$_6$, ppm): 4.29 (s, 2H, –OCH$_2$), 7.45-7.46 (d, 2H, H3, J = 2.4 Hz), 7.21-7.25 (dd, 2H, H5, J = 2.4 Hz, 9 Hz), 6.85 -6.88 (d, 2H, H6, J = 9 Hz), 9.08-9.10 (dd, 1H, Hα J = 1.5 Hz, 2.7 Hz), 7.80-7.84 (dd, 2H, Hβ 4.2 Hz, 8.1 Hz), 8.55-8.58 (dd, 1H, Hγ J = 1.5 Hz, 8.1 Hz), 8.03 (s, 2H, Hε,) ;^{13}C NMR (DMSO-d$_6$, ppm):171.1 (C=O), 68.8 (–OCH$_2$), 154.0 (C1), 137.3 (C2), 127.9 (C3), 128.9 (C4), 123.6 (C5), 115.5 (C6), 150.7 (Cα), 122.5 (Cβ), 129.2 (Cγ), 127.1 (Cδ), 124.1 (Cε), 145.6 (Cζ$_{phen}$).

2.1.4. DNA interaction study through UV visible spectroscopy

In order to evaluate the ability of the synthesized complexes to interact with the DNA the binding experiments were performed. Here at first the solution of SS-DNA was prepared by dissolving 20 mg of the respective sodium salt in water and by stirring it for 24 hours. Concentration of DNA solution calculated by using Beer-Lambert was found to be 153 µM. The absorbance of the resulting solution was noted at 259 nm to 260 nm and was adjusted at appropriate intensity, i.e., in between 0.9 to 1.3 a.u. The ratio of the absorbance at 260/280 was found to fall around 1.7, assuring the solution purity from any other interrupting proteins. Solutions of all the complexes were made in analytical grade ethanol. Concentration of test complexes was kept fixed and SS-DNA concentration was varied. Equivalent amount of SS-DNA was added into reference cell and sample cell to nullify the absorption of DNA. The complex-DNA solution was incubated for 5-7 minutes and then absorbance was recorded at room temperature [32-34].

3. RESULTS AND DISCUSSION

3.1. FT-IR Spectral Interpretation

Infrared spectral analysis served as a crucial technique for confirming complex formation since observable shifts or disappearance of absorption bands indicate interactions between the ligand and metal ion. FT-IR data of all synthesized complexes is given in Table S1. Assignment of bands was made by comparison with spectra of free ligands and already reported similar data.

The FT-IR spectra of both free ligands MeOPhA and ClPhA consist of wide O–H stretching band between 3400 and 2700 cm^{-1} region. After complexation, this band totally vanishes, demonstrating that the ligand is deprotonated [20, 35]. Similarly, both free ligands showed strong bands in the 1680-1740 cm^{-1} region for C=O stretch and around 1240–1260 cm^{-1} region corresponding to C–O stretching vibrations [33]. In the complexes, these strong vibrational bands were replaced by a new pair of bands, i.e., ν(COO)$_{asym}$ in the range of 1550-1610 and ν(COO)$_{sym}$ in the range of 1370-1430 cm^{-1} region. This is because electronic density of carbonyl oxygen is pulled towards metal

upon coordination thus the symmetry of C=O bond decreases and the strong C=O band replaced by two resonance stabilized COO^- bands [33, 36, 37].

The mode of coordination of carboxylate ligand was decided by $\Delta\nu$ ($\nu_{asymm} - \nu_{symm}$) according to the Deacon-Phillips description which they made after studying a big number of complexes [38]. The $\Delta\nu$ values for all complexes were less than 200 cm^{-1} which suggest that carboxylate ligand is coordinated through bidentate mode. In the fingerprint region, two new prominent bands appear in the 425–620 cm^{-1} region due to Cd–N and Cd–O bonds which confirm the coordination of acid ligand and N-donor moiety to metal center. The similar finding has also been discussed by Singh *et al* [39] about the M-O and M-N bonds. All the heteroleptic Cd complexes showed strong vibrational bands in the region of 750-860 cm^{-1} corresponding to C–H out-of-plane vibrations from the N-donor heterocyclic ligands [33, 40]. MeOPhA2 and ClPhA2 showed an intense band near 650 cm^{-1} due to ring bending vibrations of bipyridine which confirmed the formation of both pyridine containing complexes [41].

3.2. ^{1}H NMR Spectroscopy

A 300 MHz spectrometer was used to record the ^{1}H NMR spectra of the ligands and their associated metal complexes in deuterated dimethyl sulfoxide (DMSO). ^{1}H NMR data of all complexes is given in Table S2 and S3.

NMR spectra of the free ligands showed O–H signals at 11-12 ppm that vanished in the spectra of complexes confirming deprotonation of acid ligands [20, 42]. All the other protons of ligands appeared in their characteristic regions, i.e., methoxy and aliphatic methyl proton, methylene proton and the aromatic protons. These protons showed negligible shift upon complexation indicating their non-involvement in metal coordination [33]. In the case of heteroleptic complexes, additional signals were observed for N-donor ligands. Four distinctive aromatic proton signals in the range of 7.1-9.1 ppm are seen in MeOPhA2 and ClPhA2 complexes containing 2,2′-bipyridine. The chemical shift values were assigned to the protons following the order: $H_\alpha > H_\delta > H_\gamma > H_\beta$. Similarly, four additional protons signal in the range of 7.1-8.8 ppm were spectra of complexes MeOPhA3 and ClPhA3 containing 1,10-phenanthroline confirms its attachment. The chemical shift of proton was assigned the following order: $H_\alpha > H_\gamma > H_\delta > H_\beta$. These signals shift to higher ppm values as compared to free ligand upon coordination indicating a decrease in electron density on the nitrogen atoms and consequent deshielding thus confirming the formation of heteroleptic cadmium carboxylates containing N-donor heterocyclic ligands. As the distance of proton from coordinating nitrogen increases, the effect of deshielding also decreases so only small shift in frequency on coordination [33, 43, 44].

3.3. ^{13}C NMR Spectroscopy

^{13}C NMR helps in identification and quantification of different types of carbon atoms; methyl (CH_3), methylene (CH_2), methine (CH), aromatic carbons and carbons of N-donor ligands. It is a useful mean to directly observe a molecule's carbon structure. It provides important details regarding the hybridization states of individual carbon atoms, particularly those that are directly linked to a metal center [45]. ^{13}C NMR data of all complexes is given in Table S4 and S5.

^{13}C NMR spectra of free ligands MeOPhA and ClPhA show resonance signal of C=O group at 172.3 ppm and 167 ppm respectively. Within spectra of complexes, this resonance signal was shifted toward a downfield (higher ppm) region which indicates the deprotonation of ligands and their coordination to the metal center. This deshielding effect occurs due to the electropositive nature of Cd(II), which withdraws electron density from the carboxylate group so it resonates downfield [33, 46]. Aliphatic methylene carbon in MeOPhA and ClPhA appearing at frequency 55 ppm-67 ppm in free form showed noticeable downfield shift in spectra of complexes. Aromatic carbons of MeOPhA and ClPhA ligand appeared in their respective regions in the spectra of metal complexes thus providing strong evidence of desired complexes formation. Appearance of five additional peaks in the spectra of complexes MeOPhA2 and ClPhA2 confirms the coordination of bipyridine to metal center and chemical shifts values were assigned in the following order $C_{\varepsilon \geq} C_\alpha > C_\delta > C_\gamma > C_\beta$. The coordination of 1,10-phenanthroline is confirmed by six peaks in spectra of complexes MeOPhA3 and ClPhA3 and assignment of chemical shift values was done in following order $C_\alpha \geq C_\zeta > C_\delta > C_\gamma > C_\varepsilon > C_\beta$ [20, 47].

3.4. DNA Interaction Studies

A drug's biological action is greatly influenced by how it interacts with DNA, which has an impact on vital cellular functions like transcription and translation. Understanding these interactions is an important field of study in medicinal chemistry. Here, UV-visible spectroscopy was used to observe the interaction of SS-DNA with synthesized complexes in an ethanolic solution, using an aqueous solution of DNA. The mode of interaction is revealed by variations in absorbance and wavelength [18]. The binding constant K_b (M^{-1}) is used to measure the binding strength, whereas the Gibbs free energy (ΔG) is used to measure the spontaneity of interaction. Both parameters were calculated using Benesi- hildebrand equation [48].

$$\frac{A_\circ}{A-A_\circ} = \frac{\varepsilon_G}{\varepsilon_{H-G}-\varepsilon_G} + \frac{\varepsilon_G}{\varepsilon_{H-G}-\varepsilon_G} \times \frac{1}{K[DNA]} \quad (1)$$

A is the absorbance of complex in the presence of DNA and A_o is the absorbance of complex without DNA addition. For each complex, A and A_o are noted and $A_o/A\text{-}A_o$ ratio is plotted on y-axis and inverse of DNA concentration is taken on x-axis. $\mathcal{E}_{H-G}$ and $\mathcal{E}_G$ represent the molar absorptivity of synthesized complexes without DNA and after DNA addition [42].

Binding constant K_b is calculated by taking intercept to slope ratio. Gibbs free energy is calculated for each complex by using Equation 2:

$$\Delta G = -RT\ln K_b \quad (2)$$

In homoleptic complexes MeOPhA1 and ClPhA1, hypochromism is observed with a minor blue shift after incremental additions of DNA and strong absorption bands at 271 nm and 284 nm respectively. Hypochromic effect is observed due to binding of partially filled π* orbital of complex with π orbital of DNA hence the probability of excitation is getting less so decrease in absorbance [49]. Shoulder peak is also seen in both complexes due to n-π transitions. Both homoleptic complexes bind to DNA by groove binding mode [50].

All heteroleptic complexes MeOPhA2, MeOPhA3, ClPhA2, ClPhA3 showed strong absorption bands that appear at 278nm, 264 nm, 264 nm, and 282 nm, respectively. After incremental additions of DNA, hypochromism is observed along with a blue shift of 3 to 4 nm in Λ_{max}. These complexes bind to DNA by groove binding. The interacting molecule creates a parallel stacking arrangement by occupying a location where it sits on the DNA chromophore's floor. This configuration produces a parallel interaction (at a 90° angle where transitions are restricted for forbidden states and permitted for upper states), which raises the energy needed for the transition and, consequently, the blue shift. Additionally, there is a slight hypochromic impact from this face-to-face position [8] The binding constant values measured for the homo and heteroleptic cadmium carboxylates with ligand o-methoxyphenylacetic acid were found to be in the order:
MeOPhA3 > MeOPhA2 > MeOPhA1

The binding constant values measured for the homo and heteroleptic cadmium carboxylates with ligand 2,4-dichlorophenoxyacetic acid were found to be in the order:
ClPhA3 > ClPhA2 > ClPhA1

K_b values for heteroleptic complexes are high because they contain intercalating agent 1,10-phenenthroline and 2,2′-bipyridine which strongly intercalate with DNA and provide more area of interaction thus increasing reactivity [34]. Homoleptic complexes have no such intercalating agents thus the value of binding constants is low. Negative Gibbs free energy shows the spontaneous nature of interaction with DNA [51]. The UV-Visible spectra and graphs showing DNA binding studies of all the synthesized complexes are given in Figure 1. The binding constant, λ_{max} and ΔG values for all synthesized complexes are given in Table 1.

4. CONCLUSIONS

The synthesis of mixed ligand Cd(II) complexes by using the already bio-active moieties like substituted phenylacetic acid and nitrogen heterocycles was carried out over here. The synthesis was carried out by keeping in view their application as a drug which could target DNA, which is considered to be the main house of disease cause, propagation, its treatment and disgnosis. The ligand 2-methoxyphenyl acetic acid and 2,4-dichlorophenoxy acetic acid used as primary and 2,2-bipyridine as well 1,10-phenanthroline posses structural and electronic characteristic which effectively tuned the geometric environment around

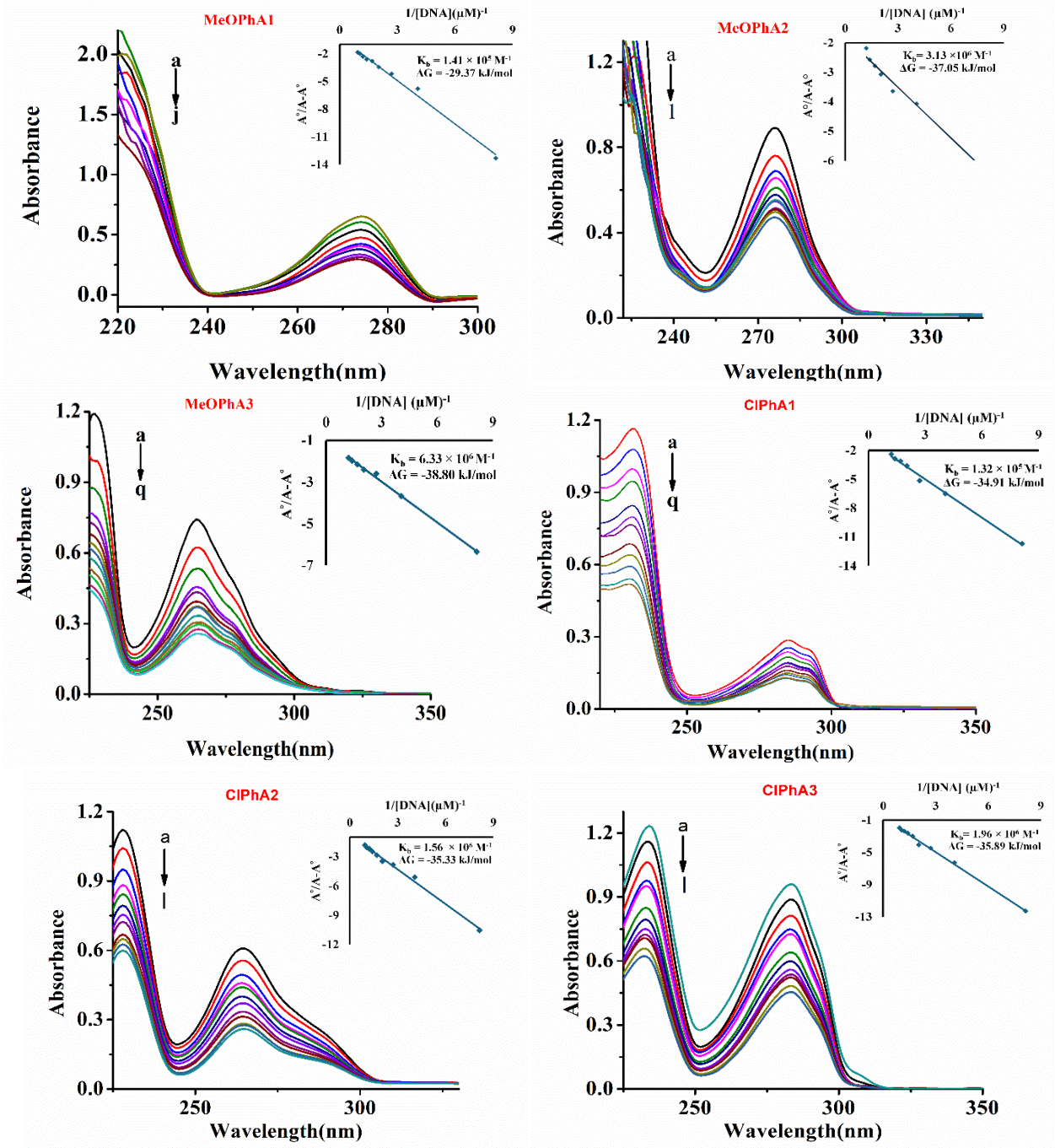


Fig. 1. Absorption spectra of complexes showing the effect of addition of DNA.

Table 1. Binding constant K_b and Gibbs Free energy ΔG values for all synthesized complexes.

Complex	λ_{max} (nm)	Binding Constant K_b (M^{-1})	Gibbs Free Energy ΔG (kJ/mol)	Mode of interaction
MeOPhA1	271	$1.41 \times 10^5\ M^{-1}$	-29.37 kJ/mol	Groove binding
MeOPhA2	278	$3.13 \times 10^6\ M^{-1}$	-37.05 kJ/mol	Groove binding
MeOPhA3	264	$6.33 \times 10^6\ M^{-1}$	-38.80 kJ/mol	Groove binding
ClPhA1	284	$1.32 \times 10^5\ M^{-1}$	-34.91 kJ/mol	Groove binding
ClPhA2	264	$1.56 \times 10^6\ M^{-1}$	-35.33 kJ/mol	Groove binding
ClPhA3	282	$1.96 \times 10^6\ M^{-1}$	-35.89 kJ/mol	Groove binding

Cd(II) center. The FT-IR data reveal the bidentate coordination mode adopted by the primary ligands. The ^{1}H and ^{13}C spectra reveal the presence of clear resonance signal attributed to proton and carbon of the complex under study. The DNA binding study through absorption spectroscopy indicate the success of the synthesized complexes. The plannar moieties and other characteristics of the plannar moieties founds to play a significant role in binding with DNA. The data indicate that such kind of structural design could provide significant help in the search for novel, effective therapeutic agents against diseases relevant to DNA.

5. CONFLICT OF INTEREST

The authors have no conflict of interest.

6. ACKNOWLEDGEMENT

S.A. is grateful to the Pakistan Academy of Sciences and Quaid-i-Azam University, Islamabad for the financial assistance.

7. REFERENCES

1. K.J. Franz and N. Metzler-Nolte. Introduction: metals in medicine. *Chemical Reviews* 119(2): 727-729 (2019).
2. S. Hamaya, K. Oura, A. Morishita, and T. Masaki. Cisplatin in liver cancer therapy. *International Journal of Molecular Sciences* 24(13): 10858 (2023).
3. A.V. Ciurea, L-A. Glavan, H.P. Costin, R-A. Covache-Busuioc and F.M. Brehar. Celebrating 70 years of DNA discovery: exploring the Blueprint of Life. *Journal of Medicine and Life* 17(4): 387 (2024).
4. U. Hollstein. Actinomycin. Chemistry and mechanism of action. *Chemical Reviews* 74(6): 625-652 (1974).
5. M. Kciuk, A. Gielecińska, S. Mujwar, D. Kołat, Ż. Kałuzińska-Kołat, I. Celik, and R. Kontek. Doxorubicin an agent with multiple mechanisms of anticancer activity. *Cells* 12(4): 659 (2023).
6. A.C. Hangan, L.S. Oprean, L. Dican, L.M. Procopciuc, B. Sevastre, and R.L. Lucaciu. Metal-based drug DNA interactions and analytical determination methods. *Molecules* 29(18): 4361 (2024).
7. S. Muthaiah, A. Bhatia, and M. Kannan. Stability of Metal Complexes. In: Stability and Applications of Coordination Compounds. A.N. Srivastva (Ed.). *London, United Kingdom* pp. 23-40 (2020).
8. S. Naz, S. Ullah, U. Iqbal, S. Yousuf, S. Rahim, N. Muhammad, R. Fatima, I.U. Haq, A. Haider, and S. Ali. Homo-and heteroleptic 3-methylbenzoates of zinc (II) ion based on N-donor heterocycles; structure, DNA binding and pharmacological evaluation. *Journal of Molecular Liquids* 368: 120792 (2022).
9. P.K. Panchal, H.M. Parekh, P.B. Pansuriya, and M.N. Patel. Bactericidal activity of different oxovanadium (IV) complexes with Schiff bases and application of chelation theory. *Journal of Enzyme Inhibition and Medicinal Chemistry* 21(2): 203-209 (2006).
10. P. Pandey, G. Manibalan, and R. Murugavel. Controlling metal coordination geometry in dinuclear zinc and cadmium hydroxy aryl carboxylates incorporating five-membered aromatic cyclic amine co-ligands. *Inorganica Chimica Acta* 551: 121461 (2023).
11. M.K. Abd Elnabi, N.E. Elkaliny, M.M. Elyazied, S.H. Azab, S.A. Elkhalifa, S. Elmasry, M.S. Mouhamed, E.M. Shalamesh, N.A. Alhorieny, *et al*. Toxicity of heavy metals and recent advances in their removal: a review. *Toxics* 11(7): 580 (2023).
12. M.R. Rahimzadeh, M.R. Rahimzadeh, S. Kazemi, and A-A. Moghadamnia. Cadmium toxicity and treatment: an update. *Caspian Journal of Internal Medicine* 8(3): 135-145 (2017)
13. K.S. Egorova and V.P. Ananikov. Toxicity of metal compounds: knowledge and myths. *Organometallics* 36(21): 4071-4090 (2017)
14. Z. Zhang, C. Bi, D. Buac, Y. Fan, X. Zhang, J. Zuo, P. Zhang, N. Zhang, L. Dong, and Q.P. Dou. Organic cadmium complexes as proteasome inhibitors and apoptosis inducers in human breast cancer cells. *Journal of Inorganic Biochemistry* 123: 1-10 (2013).
15. K. Kanude and P. Jain. Biosynthesis of CdS nanoparticles using Murraya Koenigii leaf extract and their biological studies. *International Journal of Scientific Research in Multidisciplinary Studies* 3(7): 5-10 (2017).
16. H.M. Jirjes, A.A. Irzoqi, L.A. Al-Doori, M.A. Alheety, and P.K. Singh. Nano cadmium (II)-benzyl benzothiazol-2-ylcarbamodithioate complexes: synthesis, characterization, anti-cancer and antibacterial studies. *Inorganic Chemistry Communications* 135: 109110 (2022).
17. K.N. Aziz, K.M. Ahmed, R.A. Omer, A.F. Qader, and E.I. Abdulkareem. A review of coordination compounds: structure, stability, and biological

significance. *Reviews in Inorganic Chemistry* 45(1): 1-19 (2025).
18. F. Mazhar, S. Naz, S. Muzaffar, R. Fatima, S. Yousuf, S. Ali, A. Haider, and K.S. Munawar. Mixed ligand triorganotin (IV) complexes based on oxygen and nitrogen heterocycles; exploration of the geometry and DNA binding potential. *Journal of Molecular Structure* 1345: 143004 (2025).
19. M-L. Hu, A. Morsali, and L. Aboutorabi. Lead (II) carboxylate supramolecular compounds: coordination modes, structures and nano-structures aspects. *Coordination Chemistry Reviews* 255(23-24): 2821-2859 (2011).
20. S. Muzaffar, S. Naz, F. Mazhar, Z. Rashid, M. Bibi, S. Yousuf, S. Ali, A. Haider and K.S. Munawar. Heteroleptic Zn (II) complexes; synthesis, spectral characterization, DNA interaction, enzyme inhibition and docking studies. *Journal of Molecular Structure* 1326: 141078 (2025).
21. D. Rixson, G.G. Sezer, E. Alp, M.F. Mahon, and A.D. Burrows. Synthesis, structures and properties of metal–organic frameworks prepared using a semi-rigid tricarboxylate linker. *Crystal Engineering Communication* 24(4): 863-876 (2022).
22. J. Rockenberger, L. Tröger, A. Kornowski, T. Vossmeyer, A. Eychmüller, J. Feldhaus, and H. Weller. EXAFS studies on the size dependence of structural and dynamic properties of CdS nanoparticles. *The Journal of Physical Chemistry B* 101(14): 2691-2701 (1997).
23. K. Shivaji, S. Mani, P. Ponmurugan, C.S. De Castro, M.L. Davies, M.G. Balasubramanian, and S. Pitchaimuthu. Green-synthesis-derived CdS quantum dots using tea leaf extract: antimicrobial, bioimaging, and therapeutic applications in lung cancer cells. *ACS Applied Nano Materials* 1(4): 1683-1693 (2018).
24. M. Oelschlägel, S.R. Kaschabek, J. Zimmerling, M. Schlömann, and D. Tischler. Co-metabolic formation of substituted phenylacetic acids by styrene-degrading bacteria. *Biotechnology Reports* 6: 20-26 (2015).
25. M. Mazik and P. Seidel. Synthesis of 2-[(3, 4, 5-Triphenyl) phenyl] acetic acid and derivatives. *Molbank* 2024(2): M1837 (2024).
26. H. Singh, P. Pinacho, D.A. Obenchain, M.M. Quesada-Moreno, and M. Schnell. The many forms of alpha-methoxy phenylacetic acid in the gas phase: flexibility, internal dynamics, and their intramolecular interactions. *Physical Chemistry Chemical Physics* 24(44): 27312-27320 (2022).
27. G. Smith. Low-dimensional coordination polymeric structures in alkali metal complex salts of the herbicide (2,4-dichlorophenoxy) acetic acid (2,4-D). *Crystal Structure Communications* 71(2): 140-145 (2015).
28. C. Sen, M. Kumar, Z. Nisa, N.A. Ashashi, A. Frontera, S.C. Sahoo, and H.N. Sheikh. Coordination polymers of manganese (II), cobalt (II), nickel (II) and cadmium (II) decorated with rigid pyrazine-2,3-dicarboxylic acid linker: synthesis, structural diversity, DFT study and magneto-luminescence properties. *Polyhedron* 187: 114629 (2020).
29. A. Kumar, A.K. Singh, H. Singh, V. Vijayan, D. Kumar, J. Naik, S. Thareja, J.P. Yadav, P. Pathak, *et al*. Nitrogen containing heterocycles as anticancer agents: a medicinal chemistry perspective. *Pharmaceuticals* 16(2): 299 (2023).
30. E.C. Constable and C.E. Housecroft. The early years of 2,2′-bipyridine A ligand in its own lifetime. *Molecules* 24(21): 3951 (2019).
31. P.G. Sammes and G. Yahioglu. 1, 10-Phenanthroline: a versatile ligand. *Chemical Society Reviews* 23(5): 327-334 (1994).
32. S. Naz, M. Sirajuddin, I. Hussain, A. Haider, A. Nadhman, A. Gul, S. Faisal, S. Ullah, S. Yousuf, and S. Ali. 2-Phenylbutyric acid based organotin (IV) carboxylates; synthesis, spectroscopic characterization, antibacterial action against plant pathogens and in vitro hemolysis. *Journal of Molecular Structure* 1203: 127378 (2020).
33. A. Sarfraz, S. Naz, A. Haider, K.S. Munawar, R. Fatima, S. Yousuf, M.N. Tahir, Y. Wasti, I.U. Haq, and S. Ali. Synthesis, Characterization, and in vitro pharmacological evaluation of Zinc (II) complexes of cycloalkanes and bioactive nitrogen donor heterocycles: *Journal of Molecular Liquids* 418: 126713 (2025).
34. S. Naz, N. Uddin, K. Ullah, A. Haider, A. Gul, S. Faisal, A. Nadhman, M. Bibi, S. Yousuf, and S. Ali. Homo- and heteroleptic Zinc (II) carboxylates: synthesis, structural characterization, and assessment of their biological significance in in vitro models. *Inorganica Chimica Acta* 511: 119849 (2020).
35. I. Samfira, S. Rodino, P. Petrache, R. Cristina, M. Butu, and M. Butnariu. Characterization and identity confirmation of essential oils by mid infrared absorption spectrophotometry. *Digest Journal of Nanomaterials and Biostructures* 10(2): 557-566 (2015).
36. K.I. Hadjiivanov, D.A. Panayotov, M.Y. Mihaylov, E.Z. Ivanova, K.K. Chakarova, S.M. Andonova, and N.L. Drenchev. Power of infrared and Raman

spectroscopies to characterize metal-organic frameworks and investigate their interaction with guest molecules. *Chemical Reviews* 121(3): 1286-1424 (2020).

37. P.N. Nelson and R.A. Taylor. Theories and experimental investigations of the structural and thermotropic mesomorphic phase behaviors of metal carboxylates. *Applied Petrochemical Research* 4(3): 253-285 (2014).
38. G.B. Deacon and R.J. Phillips. Relationships between the carbon-oxygen stretching frequencies of carboxylato complexes and the type of carboxylate coordination. *Coordination Chemistry Reviews* 33(3): 227-250 (1980).
39. V.P. Singh, S. Singh, D.P. Singh, K. Tiwari, and M. Mishra. Synthesis, spectroscopic (electronic, IR, NMR and ESR) and theoretical studies of transition metal complexes with some unsymmetrical Schiff bases. *Journal of Molecular Structure* 1058: 71-78 (2014).
40. K. Ullah, S. Ali, A. Haider, S. Naz, S. Yousuf, K.S. Munawar, M.S. Jan, R. Zafar, and R. Kumar. Investigation of pivalic acid-derived organotin (IV) carboxylates: Synthesis, structural insights, interaction with biomolecules, and computational studies. *Journal of Molecular Structure* 1322: 140444 (2025).
41. J. Strukl and J. Walter. Infrared and Raman spectra of heterocyelic compounds IV: The infrared studies and normal vibrations of some 1: 1 transition metal complexes of 2, 2'-bipyridine. *Spectrochimica Acta Part A: Molecular Spectroscopy* 27(2): 223-238 (1971).
42. L. Qadeer, S. Ali, A. Haider, N. Uddin, K.S. Munawar, M. Ashfaq, M.N. Tahir, and M.U. Rehman. Synthesis, spectral elucidation and DNA binding studies of cadmium (II) carboxylates with nitrogen donor heteroligands. *Inorganic Chemistry Communications* 168: 112894 (2024).
43. M. Shiotsuka, Y. Ueno, D. Asano, T. Matsuoka, and K. Sako. Synthesis and photophysical characterization of ruthenium (II) and platinum (II) complexes with bis-pyridylethynyl-phenanthroline ligands as a metalloligand. *Transition Metal Chemistry* 40(6): 673-679 (2015).
44. L. Pazderski, T. Pawlak, J. Sitkowski, L. Kozerski, and E. Szłyk. ^{1}H NMR assignment corrections and ^{1}H, ^{13}C, ^{15}N NMR coordination shifts structural correlations in Fe (II), Ru (II) and Os (II) cationic complexes with 2,2′-bipyridine and 1,10-phenanthroline. *Magnetic Resonance in Chemistry* 48(6): 450-457 (2010).
45. P. Jolly and R. Mynott. The application of ^{13}C-NMR spectroscopy to organo-transition metal complexes. *Advances in Organometallic Chemistry* 19: 257-304 (1981).
46. M. Tahir, M. Sirajuddin, M. Zubair, A. Haider, A. Nadman, S. Ali, F. Perveen, H.B. Tanveer, and M.N. Tahir. Designing, spectroscopic and structural characterization and evaluation of biological potential as well as molecular docking studies of Zn (II)-based metallo-pharmaceuticals. *Journal of the Iranian Chemical Society* 18(7): 1689-1702 (2021).
47. S. Hemalatha, J. Dharmaraja, S. Shobana, P. Subbaraj, T. Esakkidurai, and N. Raman. Chemical and pharmacological aspects of novel hetero MLB complexes derived from NO_2 type Schiff base and N_2 type 1,10-phenanthroline ligands. *Journal of Saudi Chemical Society* 24(1): 61-80 (2020).
48. A. Altaf, U. Hashmat, M. Yousaf, B. Lal, S. Ullah, A. Holder, and A. Badshah. Synthesis and characterization of azo-guanidine based alcoholic media naked eye DNA sensor. *Royal Society Open Science* 3: 160351 (2016).
49. S. Ullah, M. Sirajuddin, Z. Ullah, A. Mushtaq, S. Naz, M. Zubair, A. Haider, S. Ali, M. Kubicki, T.A. Wani, S. Zargar, and M.U. Rehman. Synthesis, Structural Elucidation and Pharmacological Applications of Cu(II) Heteroleptic Carboxylates. *Pharmaceuticals* 16(5): 693 (2023).
50. J.M. Kelly, A.B. Tossi, D.J. McConnell, and C. OhUigin. A study of the interactions of some polypyridylruthenium (II) complexes with DNA using fluorescence spectroscopy, topoisomerisation and thermal denaturation. *Nucleic Acids Research* 13(17): 6017-6034 (1985).
51. A. Tarushi, G. Psomas, C.P. Raptopoulou, and D.P. Kessissoglou. Zinc complexes of the antibacterial drug oxolinic acid: structure and DNA-binding properties. *Journal of Inorganic Biochemistry* 103(6): 898-905 (2009).

Proceedings of the Pakistan Academy of Sciences: A Physical and Computational Sciences 62(4): S1-S3 (2025)

ISSN (Print): 2518-4245; ISSN (Online): 2518-4253

Pakistan Academy of Sciences

Supplementary Data

Cd(II) Derivatives of Substituted Phenylacetic Acids, Synthesis, Spectroscopic Characterization and Binding Studies with DNA

Haleema Bibi[1,†], Aneeqa Shamim[1,†], Saba Naz[1], Moazzam Hussain Bhatti[2], Mahboob-ur-Rehman[3], Ali Haider[1], and Saqib Ali[1*]

[1]Department of Chemistry Quaid-i-Azam University, 45320, Islamabad, Pakistan
[2]Department of Chemistry, Allama Iqbal Open University, Islamabad, Pakistan
[3]Department of Cardiology, Pakistan Institute of Medical Sciences (PIMS), Islamabad, Pakistan

Table S1. The FT-IR data (cm^{-1}) of ligands and Cd(II) carboxylates.

Code	Compound	-OH	C=O/C-O		Δν	Cd-O	Cd-N	C-H
MeOPhA	**2-methoxy phenylacetic acid**	**3400-2600**	**1682/1297**					
			COO_{asymm}	**COO_{symm}**				
MeOPhA1	$Cd(MeOPhA)_2$	-	1582	1410	172	526		
MeOPhA2	$Cd(MeOPhA)_2(bipy)$	-	1560	1386	174	490	590	734
MeOPhA3	$Cd(MeOPhA)_2(phen)$	-	1570	1390	180	507	607	856
ClPhA	**2,4-dichlorophenyl acetic acid**	**3300-2400**	**1733/1290**		-	-	-	-
ClPhA1	$Cd(ClPhA)_2$	-	1611	1422	189	460		
ClPhA2	$Cd(ClPhA)_2(bipy)$	-	1613	1419	194	475	556	752
ClPhA3	$Cd(ClPhA)_2(phen)$	-	1611	1422	189	461	584	872

* Corresponding Author: Saqib Ali <saqibali@qau.edu.pk>
† Both authors contributed equally to the work

Table S2. [1]H-NMR data in ppm of o-methoxyphenylacetic acid and synthesized complexes.

Proton	MeOPhA	MeOPhA1	MeOPhA2		MeOPhA3	
-OH	11.0	-	-		-	
$-CH_2$	3.49	3.20 s	3.22 s		3.69 br	
$-OCH_3$	3.73	3.70 br	3.70 s		3.69 br	
H3	6.65	6.77-6.86 m	6.77-6.86 m		6.76-6.86 m	
H4	6.96	7.06-7.66 m	7.06-7.15 m		7.06-7.14 m	
H5	6.70	6.77-6.86 m	6.77-6.86 m		6.76-6.86 m	
H6	6.95	7.09-7.15 m	7.06-7.15 m		7.06-7.14 m	
			Bipy (free)	**Bipy (bound)**	**Phen (free)**	**Phen (bound)**
Hα	-	-	8.59	8.68-8.69 d J=4.8	8.81	9.08-9.10 dd
						J = 1.5 Hz, 4.2 Hz
Hβ	-	-	7.12	7.44-7.47 m	7.26	7.79-7.83 m
Hγ	-	-	7.66	7.92-7.98 m	8.00	8.52-8.55 dd
						J = 1.5 Hz, 8.1 Hz
Hδ	-	-	8.50	8.37-8.39	-	-
				J =8.1Hz		
Hε			-		7.55	8.02 s

Table S3. [1]H-NMR data in ppm of 2,4-chlorophenoxyacetic acid and synthesized complexes.

Proton	ClPhA	ClPhA1	ClPhA2		ClPhA3	
-OH	11.0	-	-		-	
$-OCH_2$	4.88	4.28 s	4.25 s		4.29 s	
H3	7.17	7.47 s	7.43-7.48 m		7.45-7.46 d	
						J = 2.4 Hz
H5	7.04	7.24-7.27 d	7.23-7.27 dd		7.21-7.25 dd	
		J=9Hz	*J* = 2.7 Hz, 9Hz		*J* = 2.4 Hz, 9 Hz	
H6	6.65	6.84-6.87 d	6.83-6.86 d		6.85 -6.88	
		J=9Hz	*J* = 9 Hz		*J* = 9 Hz	
			Bipy (free)	**Bipy (bound)**	**Phen (free)**	**Phen (bound)**
Hα	-	-	8.81	8.68-8.69 d	8.59	9.08-9.10 dd
				J=3.9 Hz		J = 1.5 Hz, 2.7 Hz
Hβ	-	-	7.26	7.43-7.48 m	7.12	7.80-7.84 dd
						J= 4.2 Hz, 8.1 Hz
Hγ	-	-	8.00	7.92-7.98 td	7.66	8.55-8.58 dd
				J = 1.8 Hz, 7.8 Hz		J = 1.5 Hz, 8.1 Hz)
Hδ	-	-	7.55	8.37-8.40 d	-	-
				J = 7.8 Hz		
Hε					8.50	8.03 s

Table S4. ^{13}C-NMR data in ppm of o-methoxyphenylacetic acid (MeOPhA) and synthesized complexes.

Carbon	MeOPhA	MeOPhA1	MeOPhA2		MeOPhA3	
C=O	172.3	175.4	175.7		175.4	
-CH_2	38.3	39.0	39.1		39.0	
-OCH_3	55.1	55.6	55.6		55.5	
C1	124.1	126.7	124.7		126.7	
C2	159.1	157.5	149.7		150.5	
C3	114.7	110.6	110.6		110.6	
C4	128.6	128.6	126.8		127.1	
C5	121.5	120.1	120.1		120.1	
C6	130.8	131.2	128.4		128.5	
			Bipy (free)	**Bipy (bound)**	**Phen (free)**	**Phen (bound)**
Cα	-	-	149.3	157.5	150.0	157.5
Cβ	-	-	121.0	120.9	121.5	124
Cγ	-	-	137.2	131.1	136.4	131.1
Cδ	-	-	123	137.8	129.1	137
Cε	-	-	155.4	157.5	127.5	128.9
Cζ	-	-			144.5	157.5

Table S5. ^{13}C-NMR data in ppm of 2,4-chlorophenoxyacetic acid (ClPhA) and synthesized complexes.

Carbon	ClPhA	ClPhA1	ClPhA2		ClPhA3	
C=O	167.0	170.7	170.3		171.1	
-OCH_2	67.1	68.7	68.9		68.8	
C1	152.8	154.0	154.0		154.0	
C2	124.0	129.2	137.8		137.3	
C3	131.4	123.6	124.6		127.9	
C4	128.0	127.9	129.1		128.9	
C5	128.0	122.2	123.4		123.6	
C6	117.1	115.4	115.5		115.5	
			Bipy (free)	**Bipy (bound)**	**Phen (free)**	**Phen (bound)**
Cα	-	-	149.3	149.7	150.0	150.7
Cβ	-	-	121.0	120.8	121.5	122.5
Cγ	-	-	137.2	122.1	136.4	129.2
Cδ	-	-	123.0	127.9	129.1	128.9
Cε	-	-	155.4	149.7	127.5	124.1
Cζ	-	-			144.5	145.6

Instructions For Authors

Manuscript Writing

The manuscript may contain a Title, Abstract, Keywords, INTRODUCTION, MATERIALS AND METHODS, RESULTS, DISCUSSION (or RESULTS AND DISCUSSION), CONCLUSIONS, ETHICAL STATEMENT (if applicable), ACKNOWLEDGEMENTS, CONFLICT OF INTEREST and REFERENCES, *and any other information that the author(s) may consider necessary*.

Title (Bold and font size 16): The title should be expressive, concise, and informative to the entire readership of the journal. It may include common terms, to make it more identifiable when people search online. Please avoid the use of long pervasive terms and non-standard or obscure abbreviations, acronyms, or symbols.

Abstract (font size 10, max 250 words)**:** Must be self-explanatory**,** stating the rationale, objective(s), methodology, main results, and conclusions of the study. Abbreviations, if used, must be defined on the first mention in the Abstract as well as in the main text. Abstracts of review articles may have a variable format.

Keywords (font size 10)**:** Provide five to eight keywords consisting of words and phrases that are closely associated with the topic depicting the article.

INTRODUCTION (font size 11)**:** Provide a clear and concise statement of the problem, citing relevant recent literature, and objectives of the investigation. Cite references in the text by number in square brackets**,** the reference must be cited in a proper English sentence [1]. or "... as previously described [3, 6–8]". For a single author: Bednorz [2] investigated the environmental pollution ... When there are only two authors: Bednorz and Allan [2] investigated the environmental pollution ... and for three or more authors: Bednorz *et al.* [2] investigated the environmental pollution ..; and list them in the REFERENCES section, in the order of citation in the text.

MATERIALS AND METHODS (font size 11)**:** Provide an adequate account of the procedures or experimental details, including statistical tests (if any), concisely but sufficiently enough to replicate the study. Relevant references to methodology must be cited.

RESULTS (font size 11)**:** Be clear and concise with the help of appropriate Tables, Figures, and other illustrations. Data should not be repeated in Tables and Figures but must be supported with statistics. The data presented in Tables and Figures must be elaborated in the main text.

DISCUSSION (font size 11)**:** Provide interpretation of the RESULTS in the light of previous relevant studies, citing published references.

CONCLUSIONS (font size 11)**:** Briefly state the implication of your study findings, and carefully address the study questions. Confine your conclusions according to the objectives of your study and the aspects covered in the abstract. Discuss both positive and negative findings.

ETHICAL STATEMENT (font size 10)**:** The statement of ethical approval by an appropriate ethics committee or review board must be included in the manuscript (if applicable), as per the Journal's policy.

ACKNOWLEDGEMENTS: (font size 10)**:** In a brief statement, acknowledge the financial support and other assistance.

CONFLICT OF INTEREST (font size 10)**:** State if there is any conflict of interest.

REFERENCES (font size 10)**:** References must be listed in numerical order as listed in the main text. Only published (and accepted for publication) journal articles, books and book chapters, conference proceedings, online reports, a degree thesis, and materials available on the website qualify for REFERENCES. Give online link/doi for published articles.

Declaration: Provide a declaration that: (i) the results are original, (ii) the same material is neither published nor under consideration for publication elsewhere, (iii) approval of all authors has been obtained, and (iv)

in case the article is accepted for publication, its copyright will be assigned to the *Pakistan Academy of Sciences*. Authors must obtain permission to reproduce, where needed, copyrighted material from other sources and ensure that no copyrights are infringed upon.

Manuscript Formatting

Manuscripts must be submitted in Microsoft Word (Latest Version .doc or .docx format); pdf files are not acceptable. Figures can be submitted separately in TIFF, GIF, JPEG, EPS, or PPT. Manuscripts, in *Times New Roman*, 1.15 spaced (but use single-space for Tables, long headings, and long captions of tables and figures). The Manuscript sections must be numbered, i.e., **1. INTRODUCTION, 2. MATERIALS AND METHODS,** and so on… (a) **Title** of the article **(**Capitalize the initial letter of each main word, font-size 16, **bold**), max 160 characters (no abbreviations or acronyms), depicting article's contents; (b) Author's complete name (font size 12, **bold**), and professional affiliation (i.e., each author's Department, Institution, Mailing address, and Email and Contact number, but no position titles) (font size 12); (c) Indicate the corresponding author with *; and (d) **Short running title**, max 50 characters (font size 10).

Headings and Subheadings (font size 11): All flush left

LEVEL-1: **ALL CAPITAL LETTERS; Bold**

Level-2: **Capitalize Each First Letter (**Except prepositions**); Bold**

Level-3: Capitalize the first letter only (Sentence case)***; Bold, Italic***

Level-4: *Run-in head*; *Italics*, *in the normal paragraph position. Capitalize the first letter only and end in a colon (i.e., **:**)*

A list of REFERENCES must be prepared as under:

a. Journal Articles *(Name of journals must be stated in full)*

1. J. Rashid, A. Ahsan, M. Xu, I. Savina, and F. Rehman. Synthesis of cerium oxide embedded perovskite type bismuth ferrite nanocomposites for sonophotocatalysis of aqueous micropollutant ibuprofen. *RSC Advances* 13(4): 2574-2586 (2023). DOI: 10.1039/d2ra07509a

2. A. Fayyaz, N. Ali, Z.A. Umar, H. Asghar, M. Waqas, R. Ahmed, R. Ali, and M.A. Baig. CF-LIBS based elemental analysis of Saussurea simpsoniana medicinal plant: a study on roots, seeds, and leaves. *Analytical Sciences* 40(3): 413-427 (2024). DOI: 10.1007/s44211-023-00480-9

3. W. Bialek and S. Setayeshgar. Cooperative sensitivity and noise in biochemical signaling. *Physical Review Letters* 100: 258-263 (2008). https://doi.org/10.1103/PhysRevLett.100.258101

b. Books

4. W.R. Luellen (Ed.). Fine-Tuning Your Writing. *Wise Owl Publishing Company, Madison, WI, USA* (2001).

5. U. Alon and D.N. Wegner (Eds.). An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, Boca Raton, FL, USA* (2006).

c. Book Chapters

6. M.S. Sarnthein, J.E. Smolen, and J.D. Stanford. Basal sauropodomorpha: historical and recent phylogenetic developments. In: The Northern North Atlantic: A Changing Environment. P.R. Schafer and W. Schluter (Eds.). *Springer, Berlin, Germany* pp. 365-410 (2000).

7. S. Brown and L.A. Boxer. Functions of Europhiles. In: Hematology, (4^{th} ed). W.J. Williams, E. Butler, and M.A. Litchman (Eds.). *McGraw Hill, New York, USA* pp. 103-110 (1991).

d. Reports

8. M.D. Sobsey and F.K. Pfaender. Evaluation of the H_2S method for Detection of Fecal Contamination of

Drinking Water. Report No.-WHO/SDE/WSH/02.08. *Water Sanitation and Health Programme, WHO, Geneva, Switzerland* (2002).

e. Online References

These should specify the full URL for reference, please check again to confirm that the work you are citing is still accessible:

9. UNESCO. Global Education Monitoring Report 2024/5: Leadership in education−Lead for learning. *United Nations Educational, Scientific and Cultural Organization, Paris, France* (2024). https://digitallibrary.un.org/record/4066661?ln=en&v=pdf

10. L.M. Highland and P. Bobrowsky. The landslide handbook−A guide to understanding landslides. Circular 1325. *US Geological Survey, Reston, Virginia* (2008).

    https://pubs.usgs.gov/circ/1325/pdf/C1325_508.pdf

f. Conference Proceedings

11. M. Khalid, A.B. Majid, F. Mansour, and C.R. Smith. Word Representations with Recursive Neural Networks for Morphology. *27th European Conference on Signal Processing, (2nd - 6th September 2021), Madrid, Spain* (2021).

g. A Degree Thesis

12. M. Afzal. Investigation of structural and magnetic properties of nanometallic Fe-Mn Alloys. Ph.D. Thesis. *Quaid-i-Azam University, Islamabad, Pakistan* (2023).

Tables: Insert all tables as editable text, not as images. Number tables consecutively following their appearance in the text. A concise but self-explanatory heading must be given. Tables should be numbered according to the order of citation (like **Table 1.**, **Table 2.** (font size 10)). *Do not* abbreviate the word "Table" to "Tab.". Round off data to the nearest three significant digits. Provide essential explanatory footnotes, with superscript letters or symbols keyed to the data. Do not use vertical or horizontal lines, except for separating column heads from the data and at the end of the Table.

Figures: In the main text write Figure, not Fig. Figures may be printed in two sizes: column width of 8.0 cm or page width of 16.5 cm; In the Figure caption, number them as **Fig. 1.**, **Fig. 2.** Captions to Figures must be concise but self-explanatory (font size 10). Laser-printed line drawings are acceptable. Do not use lettering smaller than 9 points or unnecessarily large. Photographs must be of high quality. A scale bar should be provided on all photomicrographs. All Figures should have sufficiently high resolution (minimum 300 dpi) to enhance the readability. Figures as separate files in JPG or TIFF format may be provided.

SUBMISSION CHECKLIST

The following list will be useful during the final checking of an article before submission to the journal.

1. Manuscript in MS Word format
2. Cover Letter
3. Novelty Statement
4. Copyright Form
5. Figures in JPG or TIFF format

In case of any difficulty while submitting your manuscript, please get in touch with:

Editor-in-Chief

Pakistan Academy of Sciences
3-Constitution Avenue,
G-5/2, Islamabad, Pakistan
Email: editor@paspk.org
Tel: +92-51-920 7140
Websites: http://www.paspk.org/proceedings/; http://ppaspk.org/

PROCEEDINGS OF THE
PAKISTAN ACADEMY OF SCIENCES: PART A
Physical and Computational Sciences

CONTENTS

Volume 62, No. 4, December 2025 **Page**

PAKISTAN ACADEMY OF SCIENCES, ISLAMABAD, PAKISTAN

HEC Recognized; Scopus Indexed

Websites: http://www.paspk.org/proceedings/; http://ppaspk.org