ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ISSN: 3041-0894 (Print)
ISSN: 3029-2387 (Online)
Volume 2 · Issue 1
January 2025

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

Print ISSN: 3041-0894
Online ISSN: 3029-2387

Artificial Intelligence in Health aims to provide a freely accessible multidisciplinary and comprehensive platform for researchers, scientists, and AI in health and medicine sciences practitioners to publish and exchange cutting-edge advancements, insights, technological development and innovations at the intersection of artificial intelligence (AI) and health. The journal seeks to explore the transformative potential of AI in improving and understanding health and medicine research outcomes, enhancing clinical decision-making, optimizing resource allocation, and addressing various challenges in the multidisciplinary field of health.

About the Publisher

AccScience Publishing is a publishing company based in Singapore. We publish a range of high-quality, open-access, peer-reviewed journals and books from a broad spectrum of disciplines.

Contact Us

Managing Editor
aih.office@accscience.sg

AccScience Publishing
8 Burn Road, #15-03 Trivex, Singapore 369977.

Volume 2 • Issue 1 • January 2025
ISSN 3041-0894 (print) ISSN 3029-2387 (online)

ARTIFICIAL INTELLIGENCE IN HEALTH

Editor-in-Chief
Andrzej Cichocki
Systems Research Institute of Polish Academy of Science, Poland

ACCSCIENCE PUBLISHING

Access Science Without Barriers

Full issue copyright © 2025 AccScience Publishing
All rights reserved. Without permission in writing from the publisher, this full issue publication in its entirety may not be reproduced or transmitted for commercial purposes in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system. Permissions may be sought from aih.office@accscience.sg.

Article copyright © Respective Author(s)
See articles for copyright year. All articles in this full issue publication are open-access. There are no restrictions in the distribution and reproduction of individual articles, provided the original work is properly cited. However, permission to reuse copyrighted materials of an article for commercial purposes is applicable if the article is licensed under Creative Commons Attribution-NonCommercial License. Check the specific license before reusing.

Artificial Intelligence in Health

ISSN: 3041-0894 (print)
ISSN: 3029-2387 (online)

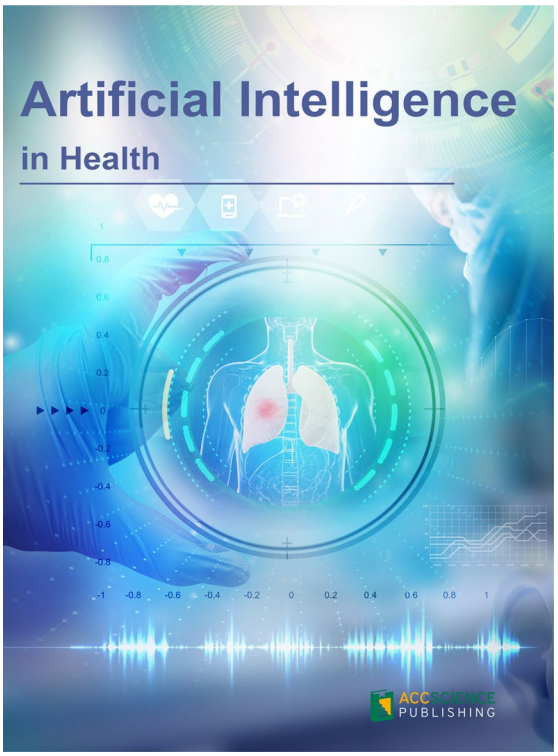
Editorial and Production Credits

Publisher: AccScience Publishing
Managing Editor: Irene Zhao
Production Editor: Sharmila Velapasamy
Article Layout and Typeset: Sinjore Technologies (India)

For all advertising queries, contact aih.office@accscience.sg.

Supplementary file

Supplementary files of articles can be obtained at https://accscience.com/journal/AIH/2/1.



Disclaimer

AccScience Publishing is not liable to the statements, perspectives, and opinions contained in the publications. The appearance of advertisements in the journal shall not be construed as a warranty, endorsement, or approval of the products or services advertised and/or the safety thereof. AccScience Publishing disclaims responsibility for any injury to persons or property resulting from any ideas or products referred to in the publications or advertisements. AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Artificial Intelligence in Health

Editorial Board

Editor-in-Chief

Andrzej Cichocki, *Poland*

Executive Editors

Adrian David Cheok, *China*
Hongcai Shang, *China*
Xiaobo Zhou, *USA*

Associate Editor

Weiping Ding, *China*

*Editorial Board Members**

Adel Al-Jumaily, *Australia*
Zeeshan Ali, *China*
Ahmed Bouridane, *UAE*
Joaquim Carreras, *Japan*
Faouzi Alaya Cheikh, *Norway*
Xiaojun Chen, *China*
Krzysztof Jozef Cios, *USA*
Alfredo Cuzzocrea, *Italy*
Di Dong, *China*
Anastasios Dounis, *Greece*
Włodzislaw Duch, *Poland*
Ayman El-Baz, *USA*
Adel Elmaghraby, *USA*
Manuel Francisco González Penedo, *Spain*
Rémy Guillevin, France
Andrew A. Gumbs, *France*
A. Ben Hamza, *Canada*
Alexander Hramov, *Russia*
Bin Hu, *China*
Donato Impedovo, Italy
S. M. Riazul Islam, *UK*
Ankush D. Jamthikar, *India*
Jay Kalra, *Canada*
Uzay Kaymak, *Netherlands*
Fahmi Khalifa, *USA*
Antonio Lanata, *Italy*
Xueping Li, USA
Zihuai Lin, *Australia*
Wing-Kuen Ling, *China*
Nicola Luigi Bragazzi, *Canada*
Xiaoke Ma, *China*
Xuelei Ma, *China*
George D. Magoulas, *UK*
Mrinal Mandal, *Canada*
Francesco Mercaldo, *Italy*
Reza Mirnezami, *UK*
Jianwei Niu, *China*
George Notas, *Greece*
JungHwan Oh, USA
Peichen Pan, *China*
Alexander N. Pisarchik, *Spain*
Dawid Polap, *Poland*
Mihail Popescu, *USA*
Mukesh Prasad, *Australia*
Marek Reformat, *Poland*
Hassan Rivaz, *Canada*
José Santamaría López, *Spain*
Paulo Adriano Schwingel, Brazil
Wei Shao, *China*
Chao Shen, *China*
Patricia A. Shewokis, *USA*
Qiongfeng Shi, *China*
Lampros Stergioulas, *Netherlands*
Jasjit S. Suri, *USA*
Kenji Suzuki, *Japan*
Abdelmalik TALEB-AHMED, *France*
Miguel Garcia Torres, *Spain*
Ricardo Vardasca, *Portugal*
Eugenio Vocaturo, *Italy*
Alan Wang, *New Zealand*
Guotai Wang, *China*
Yanfeng Wang, *China*
Fangxiang Wu, *Canada*
Jian Yang, *China*
Qi Yang, *China*
Zhewei Ye, *China*
Xujiong Ye, UK
Yudong Zhang, *UK*
Yu Zhang, *USA*
Wensheng Zhang, *China*
Zhuhuang Zhou, *China*
Shang-Ming Zhou, *UK*

*Youth Editorial Board Members**

Yankai Chen, *USA*
Afify Heba, Egypt
Hongxin Pan, *China*

*Editorial Board Members as of January 15, 2025

ACCSCIENCE PUBLISHING

Volume 2 | Issue 1 | January 2025 | Page 1 – 120 |

Artificial Intelligence in Health

https://accscience.com/journal/AIH

CONTENTS

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

REVIEW ARTICLE

The role of artificial intelligence in higher medical education and the ethical challenges of its implementation

Mark Perkins[1,2†] and Agnieszka Pregowska[3†*]

[1]Collegium Prometricum, The Business School for Healthcare, Sopot, Poland

[2]Royal Society of Arts, London, United Kingdom

[3]Department of Information and Computational Science, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland

Abstract

Artificial intelligence (AI) is penetrating higher medical education; however, its adoption remains low. A PRISMA-S search of the Web of Science database from 2020 to 2024, utilizing the search terms "artificial intelligence," "medicine," "education," and "ethics," reveals this trend. Four key areas of AI application in medical education are examined for their potential benefits: Educational support (such as personalized distance education), radiology (diagnostics), virtual reality (VR) (visualization and simulations), and generative text engines (GenText), such as ChatGPT (from the production of notes to syllabus design). However, significant ethical risks accompany AI adoption, and specific concerns are linked to each of these four areas. While AI is recognized as an important support tool in medical education, its slow integration hampers learning and diminishes student motivation, as evidenced by the challenges in implementing VR. In radiology, data-intensive training is hindered by poor connectivity, particularly affecting learners in developing countries. Ethical risks, such as bias in datasets (whether intentional or unintentional), need to be highlighted within educational programs. Students must be informed of the possible motivation behind the introduction of social and political bias in datasets, as well as the profit motive. Finally, the ethical risks accompanying the use of GenText are discussed, ranging from student reliance on instant text generation for assignments, which can hinder the development of critical thinking skills, to the potential danger of relying on AI-generated learning and treatment plans without sufficient human moderation.

Keywords: Artificial intelligence; Metaverse; Medical education; Education system; Ethics

[†]These authors contributed equally to this work.

***Corresponding author:**
Agnieszka Pregowska
(aprego@ippt.pan.pl)

Citation: Perkins M, Pregowska A. The role of artificial intelligence in higher medical education and the ethical challenges of its implementation. *Artif Intell Health*. 2025;2(1):1-13.
doi: 10.36922/aih.3276

Received: March 26, 2024

Revised: April 29, 2024

Accepted: July 1, 2024

Published Online: October 21, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Medical practice, which heavily relies on advancements in medical education, is one of the fastest-moving fields, frequently testing technological innovations through pilot trials and proof-of-concept studies.[1] Artificial intelligence (AI) now stands at the forefront of these innovations, offering many benefits, such as effective tools for analyzing and processing large datasets quickly – tasks that would be impossible for humans to accomplish.

One important area in healthcare involves electronic health records, which can serve as input data for AI and be processed quickly. However, such datasets not only contain sensitive content but also constitute ethical risks, especially when data collection is subject to various forms of bias[2] and is exposed to a large number of hostile attacks.[3] More concerning is that most medical researchers treat AI as a black box, leaving its ethical risks concealed.[4] A strong argument can be made that the successful application of AI in medical practice will depend on addressing legitimate concerns about misunderstandings of its principles and data risks, in accordance with evolving bioethical principles.[5] In a field such as medicine, which is critically related to issues of life and health,[6] it is particularly important to explain and address the impact of AI on its essence and principles in medical educational programs, both in terms of how it works and its underlying ethical assumptions. For medical practitioners to use AI-based solutions effectively in their work, they must first learn how to use them correctly during their training.

Moreover, AI-based solutions may be more vulnerable to attacks compared to other approaches, such as statistical methods.[7] It is also worth stressing that, especially in the field of medicine, deep neural networks with many layers (such as highly complex architectures) are commonly applied. This may contribute to AI models being more susceptible to overfitting, where the neural network memorizes the training data rather than generalizing from it. In this context, statistical methods are composed of simpler models with fewer parameters, which may lead to easier interpretation of the model.[8]

A significant limitation of AI is its dependence on data.[9] In particular, the essence of AI, comprising algorithms for learning complex patterns and making accurate predictions, has a core sensitivity feature: the quality and representativeness of the training data. Inaccuracies in the training data significantly affect the efficiency and accuracy of the results obtained, potentially skewing outcomes and leading to ethical consequences that oppose the institution's goal. Indeed, it can be said that the quality and output of AI algorithms are directly dependent on the medical data used to develop, test, and validate them. Therefore, a key issue in using AI in medicine is the reliability of biomedical data obtained from patients, which must be compiled and categorized in an ethical manner. Unlike AI-based models, statistical methods can work with smaller datasets, and the optimal selection of data may help minimize data errors more efficiently. The heavy data dependence on AI-based solutions also makes them vulnerable to developing learning patterns based on biased and faulty training data. If the input data is not representative of the real-world population or reflects historical biases and inequalities, AI can learn and perpetuate these biases. For example, a language model trained on text from certain online communities may accidentally learn and replicate the biases expressed in that community. A lack of diversity in the ethical standpoint of AI researchers may also contribute to bias issues. Moreover, the algorithms themselves may introduce or amplify algorithmic errors due to their inherent operational principles.

Another challenge related to data is security. Compared to traditional statistical methods, AI-based algorithms are more susceptible to adversarial attacks that exploit security vulnerabilities, such as sensitivity to even low noise in the input data.[10] Traditional methods are more deterministic, making them more resistant to such attacks.

In this paper, we analyze the technical and ethical risks associated with certain AI applications in medical education, exploring the potential benefits and risks of these technologies in practice, the awareness of students and practitioners regarding these issues, and the latest scientific research in this area.

2. An overview of current research activity in AI, medical education, and ethics

In this paper, we conducted a systematic review of research on AI, medical education, and ethics based on the PRISMA academic review process and its extensions, including PRISMA-S.[11] Resources written in English from the Web of Science (WoS) database were considered, excluding PhD theses and any material not related to AI or education. Our searches for the terms "AI," "education," and "medicine" yielded 488 resources, of which 34 addressed ethical issues. Figure 1 presents the participation rate in % of individual areas of the world in research relating to AI in medical education (Figure 1A) and AI in medical education, taking into account ethical issues (Figure 1B). These results highlight both the very low participation of low-income countries in research and a lack of focus on ethics. However, the study also included searches involving the search terms "artificial intelligence," "medicine," and "ethics" (AI+med+ethics), which yielded 328 results, giving a higher result when education as a whole is considered. The sources included were selected to answer the research question, "What multi-criteria impact will AI have on higher education in the field of medicine?" First, duplicate records in the database were excluded. In the second step, records whose titles and abstracts were not related to the subject of the analysis were excluded. Then, records that were not accessible were disabled. In the final stage of the search, records without information concerning the topic of consideration were excluded from the analysis. Finally,

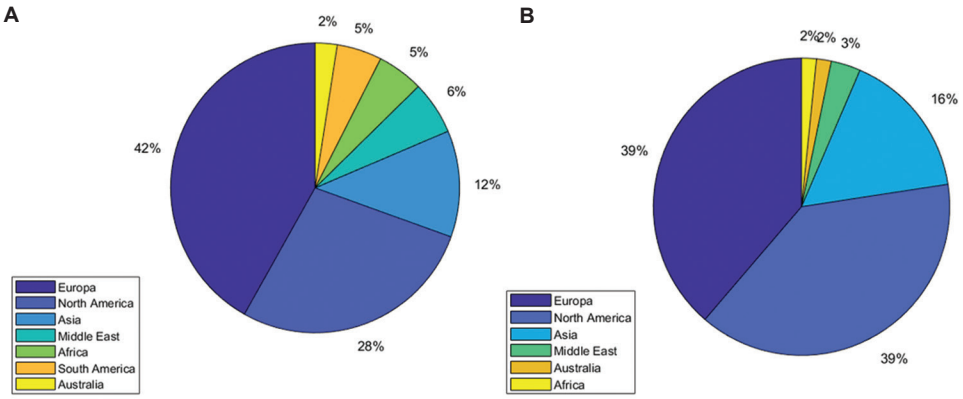


Figure 1. Geographical distribution of papers related to artificial intelligence (AI) in medical (med) education (ed) on the Web of Science. (A) Papers including the terms AI+med. (B) Papers including the terms AI+med.

94 documents were taken into account. This investigation has two limitations. First, the study only takes into account the WoS database, which is the most restricted of its type (although this ensures the integrity of the dataset), and second, only publications written in English were included in the systematic review, which may cause a potential language bias.

The results of 328/488 resources may seem low. This view is supported by Lee *et al.*,[12] who noted that AI is a relatively new concept in medical education. More recently, as a result of an exhaustive search in four databases (PubMed, Embase, Scopus, and WoS) during the period 2020–2024, Weidener and Fischer[13] affirmed that there is "a scarcity of literature on teaching AI ethics in medical education, with most of the available literature being recent and theoretical" (ibid., p. 399). The study shows that the major studies (about 90%) in the field of AI ethics were published in the years 2020–2024, which coincides with the dynamic development of AI. This is largely due to the fact that currently solutions based on AI can be implemented in practice, and there is a need to consider all risks, both ethical and practical (technical). Since we analyze the status of development and implementation of the general guidance on the ethics of AI in the field of medical education (with special emphasis on practical implications), in this study, we concentrate on the time frame in which the most dynamic development of the field of AI ethics occurs. In addition, the analysis highlighted a research gap in low-income countries. One of the reasons may be the lack of access to the latest technologies, which often involves significant costs. The lack of research in this area also translates into a potentially low level of implementation of AI in practice. Indeed, it is evident that the results of AI+med+ed are a small proportion of those for AI+med, and that the results for AI+med+ed+ethics are an even smaller proportion of AI+med (Figures 2A and 2B). Overall, there is an underrepresentation of research in the developing world, despite the recognized importance of AI+med+ed+ethics.

A similar situation was found when a search for the terms AI+med+radiology and AI+med+XR was conducted (Figure 3A and B).

Almost half of the occurrences of AI+med+radiology were found in North America (49%), compared to only 7% in Europe. On the other hand, almost the opposite was found regarding AI+med+XR: North America (47%) and Europe (19%). This suggests that research and awareness of AI and radiology are more advanced in North America than in Europe and that the opposite is true concerning AI and XR. Asian results were similar in both cases (25% and 21%), but as a large developing territory, Africa was substantially underrepresented (2% and 3%).

3. AI in medical education-some practical applications

AI is increasingly seen as a significant resource for medical education that will permeate all areas and become integral. AI is being applied in several different types of medical fields, including technical support and distance learning, data analysis and interpretation, 3D modeling and remote virtual surgery, and text production by AI-powered text generation engines (GenText) such as ChatGPT (Chat Generative Pre-trained Transformer). A study by Civaner *et al.*[14] showed that 80% of medical students perceive AI as a technology supporting both the process of education and health care, although the study also revealed concerns among the medical community about AI undermining their skills and negatively impacting the patient-doctor relationship (50% and 40% of respondents, respectively). These concerns are not shared by biomedical physicians, more than 80% of whom see AI as a support tool, not a risk

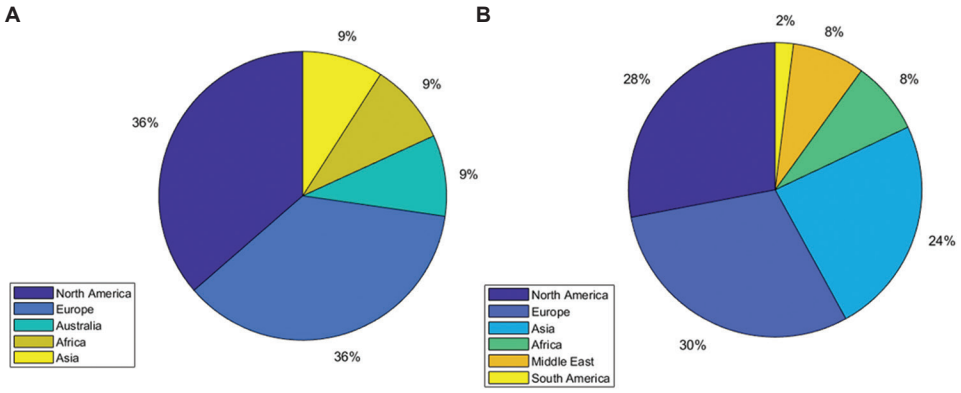


Figure 2. Geographical distribution of papers related to artificial intelligence (AI) in medical (med) education (ed) and ethics (ethics) based on the Web of Science. (A) Distribution of papers on AI+med+ed. (B) Distribution of papers on AI+med+ed+ethics.

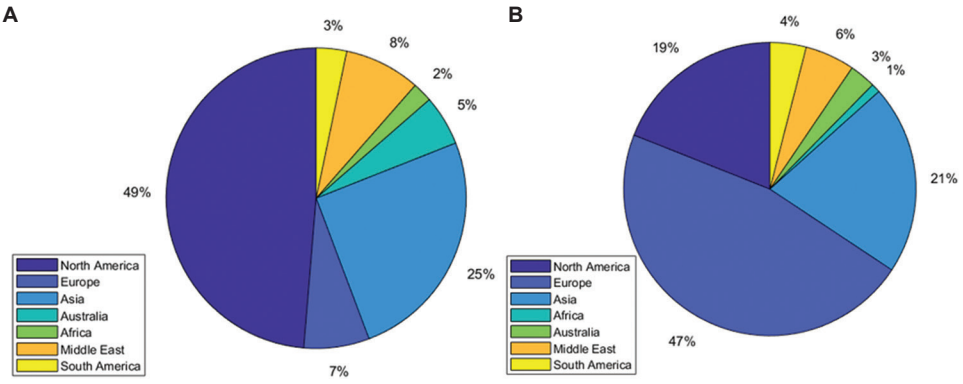


Figure 3. Geographical distribution of papers related to artificial intelligence (AI) in medical (med) education (ed) and XR based on the Web of Science. (A) Distribution of papers on AI+med in radiology. (B) Distribution of papers on AI+med in Extended Reality (i.e., virtual reality, augmented reality, mixed reality, and metaverse).

tool. Similarly, Andersson *et al.*[15] and Mosh *et al.*[16] concluded that AI helps reduce physicians' workload. In addition, during the COVID-19 pandemic, distance learning developed significantly and came to cover the whole world, not just people living in remote and inaccessible areas. This was made possible by the development of technology, including AI-based solutions. Furthermore, the high cost of practical offline classes, especially in the field of medicine, and the consequently limited opportunities for participation make remote learning solutions appear to be an accessible and natural development of the educational sector. Indeed, the potential of AI in distance education has been demonstrated by Garlinska *et al.*,[17] who pointed out that AI-based algorithms can be applied to the personalization of learning content, automated grading, and virtual touring, including as they do features such as speech-to-text and text-to-speech.

More specifically, AI has the potential to improve the diagnostic process, which is crucial for education in this field.[18] Radiology is a significant area of medicine. In July and August 2023, Gordon *et al.*[19] conducted a thorough search of publications found in PubMed/MEDLINE, EMBASE, and MedEdPublish, looking for keywords connected to the use of AI in medical education. The largest preponderance occurred in the field of radiology (11.2%), followed by surgery (8.7%). Then, Pinto dos Santos *et al.*[20] conducted an anonymous survey concerning the attitude of radiology students to AI applications, whereby 83% of participants believed AI-based algorithms could potentially detect pathological changes in radiological images, while 56% felt they were unable to correctly interpret these changes. At the same time, 68% of respondents admitted they were not aware of the technologies and risks associated with AI implementation in radiology. This indicates a significant gap among medical staff regarding AI technology, which may hinder its effective use.

Visualization constitutes a major area where AI-based solutions play a significant role in medical education. These solutions enable the creation of realistic 3D visualizations of organs, their abnormalities, and the entire human

body. They can then be used to develop virtual reality (VR) medical simulators in the metaverse and medical holograms (mixed reality).[21-23] Thus, AI has an integral role in the development of the virtual environment. Similarly, surgical simulation provides physicians with richer, more realistic tools for training and improving their skills.[24] It enables the production of 3D interactive objects from medical data such as magnetic resonance imaging, computer tomography, and other techniques.[25,26] This helps students understand medical content, as demonstrated in a pilot study by Sariciliar *et al.*[27] The use of solutions based on AI and XR also allows the operating table to be observed from an unlimited perspective and from anywhere in the world without disrupting the course of the operation (although this is naturally dependent on a stable internet connection.[28,29] Furthermore, the attending physician also has the opportunity to consult with another physician in a different location without leaving the patient and the operating table. Operations can also be recorded and played back from different observation perspectives, which also has a significant educational dimension. An interesting AI-based application field has also been pointed out by Winker-Schwartz *et al.*,[30] namely the use of machine learning to analyze texts to assess knowledge in the area of surgery. AI can also help assess the level of knowledge and skills of the user performing surgery in a virtual environment.[31] For this purpose, electroencephalography recording was used as input data. It turned out that the records differ depending on qualifications and experience, and this can be used to develop a classifier of their skills (markers of experience). However, a major contrast is to be made between the application of AI in medical practice in high and low-income countries, where the latter find the requirement for expensive equipment and high-capacity internet connectivity challenging.

In contrast to the visual side, text is a vast area where AI has the potential for extensive use in many use cases. A GenText engine such as ChatGTP based on large language models (LLMs) is capable of quickly producing large amounts of convincing natural language text of many types and purposes, from chatbots[32] to scientific writing.[33] Although its use as a production tool has led to bans and restrictions in the higher education sector due to fears that students will lose the capability of writing original work or thinking for themselves, the sector is now beginning to accept that there will be a place for ChatGPT and it is better to educate users than to ban it outright.[34] Apart from a tool to assist in the production of text (ranging from reports, notes, and summaries to discursive pieces), ChatGPT can be used to create learning programs themselves. The potential of AI in distance education has been demonstrated by Garlinska *et al.*,[17] who pointed out that AI-based algorithms can be applied to the personalization of learning content, automated grading, and virtual touring, including as they do features such as speech-to-text and text-to-speech. Then, GenText can offer an AI-human interactive experience on top. For example, Beilby *et al.*[35] evaluated the use of ChatGTP for providing fertility information that could inform decision-making. They found that while ChatGPT effectively sorted through a large amount of information and generated summaries, medical practitioners were still needed for further explanations and counseling. Moreover, Funk and coauthors demonstrated statistically significant differences between ChatGPT versions.[36] Specifically, ChatGPT-4 was found to be more consistent, providing 44.9% more correct answers to medical questions than version 3.5.

While AI-based tools such as ChatGTP offer many benefits, they are not without drawbacks.[37] An undoubted advantage is the fact that ChatGPT is a tool that provides the user with output at any time; the only requirement is to have an account and good internet access. Moreover, through the way the user asks questions and based on the context of the conversation, ChatGTP adapts the content to its interlocutor.[38]

There is also an option to include in the question the level at which the user would like to get an answer. This works well even when working on the basic free level rather than the paid (Doctor GPT). However, the biggest limitation is the strict dependence on the nature of the input data applied. In the case of ChatGTP, it is also worth noting the fact that databases have not been updated since January 2022, which is an important issue considering the rapid development of medicine. Furthermore, this operation results from the very principles of its internal structure, its ability to create answers based on patterns within data sequences. In addition, it may provide incomplete information,[39] which may cause ChatGPT to create literature sources that do not exist in reality. Another significant disadvantage of LLMs is the lack of consideration for the context of medical concepts or the nuances of patient care. In addition, there is an ethical issue related to the security of patient data. A further limitation is the fact that ChatGPT operates in English, but data added in other languages could be converted into English by automated translation. In the future, this can improve communication with both patients and students who do not speak English fluently. Furthermore, a limitation is the fact that ChatGPT does not have a built-in data reliability assessment module. It is also possible to overtrain LLMs. It is worth noting that ChatGPT is a content aggregator and analyzer based on a language model, not a source of knowledge in itself. It should rather be treated as a supporting tool or kind of

guide whose answers should be examined critically. On the other hand, combining ChatGPT with another tool, such as virtual simulators, can be extremely beneficial for medical students.[40] However, it is during this time that ChatGPT should be thoroughly tested against possible errors that can be made in medical education processes. It is also worth emphasizing that the long-term impact of AI tools, including ChatGPT, on learning outcomes, especially in the field of medicine, should be examined.[41]

On the other hand, an interesting study[42] analyzed medical students' readiness for AI-based solutions. The findings revealed that students who believed AI technologies would contribute to their profession and reduce workload outnumbered those who held a different view. In addition, a study[43] proposed a Persian version of the Medical AI Readiness Scale to evaluate the readiness of medical students to work with AI, including factors, such as cognition, ability, vision, and ethics.

4. Ethical risks in the implementation of AI in medical education

Each of the four examples of AI's significant role in medicine and medical education offers great hope for rapid improvements in medical practice. However, these advancements come with ethical risks that, if not addressed, could result in a curse of malpractice and bad outcomes for educationalists and their students as well as for practitioners. There has been a discussion regarding AI and ethics for many years, as illustrated by Dennett's vision of a novel-writing machine and the dilemmas it raises about the notion of self.[44] Yet, it is only recently that a focus on ethical risks, AI, and medical education has appeared, no doubt in tandem with the rapid development of technology. Indeed, on the general level, as noted above, Weidener and Fischer[13] demonstrated that there is a lack of discussion concerning AI and medical education overall, even though, as Civaner *et al.*[14] pointed out, there is a recognition amongst many medical students that AI needs to play a role in medical education. This shows that there is a student (or consumer) demand for AI in educational curricula and a need for educators to fill that gap. There is thus a clear requirement for AI to be integrated into medical education programs, but reasons can be advanced for the slow pace of adoption. For example, such programs are extensive and well-established, and there may be resistance from course designers and managers, educators, and other stakeholders.[45] On the other hand, the integration of AI into medical education is likely inevitable, paving the way for serious disruption and commercial opportunities. Indeed, it is necessary since a lack of integration will constitute a further type of broad ethical risk: if students are not equipped with AI knowledge, they will be less able to cope with the various and detailed types of ethical risk as practitioners. However, advances are being made even while calls for a faster pace of change are being made.[46,47] An outline model for the application of AI in medical education is provided by Zarei *et al.*,[48] along with an assessment of challenges such as the current lack of infrastructure. Krive *et al.*[49] designed and tested a model comprising a modular 4-week AI course, which proved to be successful.

As a specific area, radiology, for example, depends heavily on data.[50-52] It is immediately apparent that the successful manipulation of information-intensive radiological data using AI requires significant computational resources. This raises concerns about energy use, costs, and environmental impact, where developing countries may be at a disadvantage, thus increasing ethical risk for them. Another extremely important issue concerns how the accuracy of AI predictions using various types of metrics is to be evaluated.[53] This is connected with algorithmic fairness:[54] If one method of evaluation produces a different metric than another, the outcome could result in being unfair to one or another cohort, an ethical issue. The most popular algorithms in the field of medicine are the Dice coefficient and accuracy.[4,55] However, there is no accepted standardization for the assessment of such algorithms in medicine. Turning to the issue of data biases, the extensive account provided by Ueda *et al.*[56] broadly separated into machine and human-originated, and the discussion of biases identified by Pregowska and Perkins[5] (passim) prompts the need for two underlying dimensions of bias to be highlighted in addition. The first is intentional and unintentional. The introduction of bias into a dataset (such as the over-representation of one demographic cohort at the expense of another or incorrect,[57] and model and interpretation bias[58]) may be intentional on the part of the human agent or unintentional (due to accident, neglect, human error, or subconscious attitude). Once intentional bias has been identified, the question of motivation arises as a second underlying dimension. Bias can be introduced into dataset selection, and datasets can be manipulated due to social and political attitudes in some societies. The profit motive may also raise issues of control, ownership, deployment, and use of data, and even falsification.[59] The increasing role of AI, along with its ability to create and amplify biases or distort information – complicated by the need for radiological data between institutions and across borders[60] – highlights the importance of transparently identifying agents within the system and their access to AI tools. This transparency should be integrated into medical education from the outset.[61] In addition, convincing practitioners of the significant benefits AI offers to

radiology[62] presents serious challenges to those engaged in curriculum and syllabus design. Moreover, there is an external dimension of malign intention represented by cybersecurity threats. Medicine is under increasing attack, and practically no field is more greatly exposed than radiology.[63] Cyber-attacks can range from malicious insider activity to data theft, credential harvesting, and phishing. They can occur at various points in the radiological landscape, including medical devices, wireless systems, data warehouses, and social networks, and the increasing use of AI on both sides has created vulnerabilities.[63] Moreover, there is also no clear overview of approved AI-based medical devices. This leads to inconsistency and increased ethical risk. However, the problem is recognized, and investigations are currently underway by the Food and Drug Administration in the USA,[64,65] and the Medicines and Healthcare Products Regulatory Agency,[65] which is developing guidelines for such devices.[66] Here, broadly understood, cybersecurity is an important issue.[5] AI systems are vulnerable to adversarial attacks,[67] such as the introduction of minor modifications to input data in changing training labels that lead to invalid predictions. Each such attack is a breach of sensitive patient information, and any wrong decision in the medical field has potentially disastrous consequences. This vulnerability extends not only to patient data but also to student data. Tsai and Lin[68] proposed a procedure to evaluate the resistance of AI models based on medical images against these attacks. There are various techniques to defend against adversarial attacks, including data augmentation, adversarial training, and robust optimization. However, establishing effective protection protocols remains a challenge.[69]

In the development of two further contrasting and specific areas of AI, there is evidence of AI being used in education, VR, and GenText.[70] In the case of VR, although some evidence of adoption has been found to be sparse, as in the database search by Lie *et al*.[71] covering November and December 2021, a subsequent more extensive literature search study in the period January 2017 to March 2022 demonstrated a rapid and increasing take-up perhaps in the latter part of this period, although this is not stated in the research.[72] Moreover, students trained using VR produce better results than those conventionally taught. Kim and Kim[73] identified and examined 24 studies and a sub-group of 18 on the use versus non-use of VR in medical education and found that "there was a significant improvement in the VR group's skill and satisfaction levels, and that less immersive VR was more efficacious for knowledge outcomes than fully immersive VR" (ibid., p. 13). Greater student satisfaction in using AI is also confirmed by Leng,[74] who found that in the case of learning anatomy, ChatGPT has increased student engagement. Then, in a small-scale study of 44 students aimed at validating VR-based medical training, Pedram *et al*.[75] not only found a user acceptance level of 75% but also an outperformance by those using VR of the control group that did not. These studies reinforce the view that there is a greater ethical risk in a sluggish implementation of AI in medical education than in a rapid one. Slow implementation will result in inferior education. In turn, this will lead to slower and possibly deficient deployment of AI in the clinic, with consequently worse patient outcomes. While the fast deployment of AI in medical education will bring lower ethical risk, another aspect of risk may be avoided, that of the vulnerability of data. In a clinical setting, real patient data will be used. In a VR scenario, simulated data are sufficient. Mergen *et al*.[76,77] have developed a project tool entitled "medical tr.AI.ning," an immersive VR learning platform based on AI that generates simulated patient data, thus obviating ethical concerns.

Regarding ChatGPT and other potential GenText engines, there are many points of ethical risk in medical education. Once more, the output quality depends on the input datasets. Very often, data, especially medical data, is burdened with various types of bias.[46] There is also a further question of whether ChatGPT is biased as a collection of algorithms or whether algorithmic bias could be introduced unethically.[78] If bias can occur at these two levels, there is a further systemic ethical threat in the vulnerability of GenText and other engines to jailbreak, where an AI system acts outside the restrictions placed on it by its designers.[79] Further alarming consequences may arise when an AI that has broken free can create other AIs that may produce harmful output, such as producing a set of instructions for synthesizing methamphetamine.[80] However, it is at the day-to-day level that ChatGPT causes a great deal of concern: at face value, ChatGPT can be used to generate substantial amounts of convincing text. Such text can be used for framework and content infill for curricula and syllabi (by course managers and designers), teaching material (by educators), and assignments (by students). However, where there is a risk that the content generated is at risk of being out of date (depending at least on the latency of input protocols and difficulty in ensuring the provision of the latest academic material (due to secrecy concerns) and that ChatGPT is capable of hallucinating with it comes to references,[81] the validity of the output will be variable and at times questionable. In addition, Májovský *et al*., 2023[82] considered ChatGTP as a tool for the generation of fake medical papers. The whole process took an hour, and it turned out that the text looked convincing. Although references and specific errors raised doubts, these errors could only be detected by an experienced reader, here a medical doctor. This creates a

risk that a student using ChatGTP or a student generating, for example, text for his or her work will not be able to find irregularities that may be important. Chio *et al.*[83] put forward the same argument, raising ChatGPT's lack of critical reflection in the case of nurse education. ChatGTP does not act in such a way as to assess the credibility of sources; when asked to provide literature, in many cases, it creates references that do not exist. This all amounts to a substantial ethical risk. Simply put, the output from GenText cannot be completely relied upon and needs human moderation. In a study conducted by Tsegay *et al.*[84] on writing in an undergraduate medical degree course in Darussalam, they found substantial inclusion of ChatGPT-generated text and citations to non-existent references. As a result, they propose that educators should be more aware of Gen Text detection tools. Another ethical issue arising here concerns the fact that ChatGPT can give answers to students very quickly: The normal study processes of critical thinking (assessing information, making connections, and drawing conclusions) are thus bypassed. Another ethical issue is the possibility of students writing potentially fraudulent assignments. This indicates the need for teaching how ChatGPT can be used appropriately in the learning situation. Apart from discursive documents and reports, there are many other types of text that can be generated. These include personalized learning plans and treatment plans. If these are relied on without scrutiny or moderation, adverse results could occur, such as misdiagnosis and bad treatment (of oneself and others). This needs to be regulated on a high level and accounted for in local ethical policies and educational practices,[85] and calls are emerging for the development of new educational governance of AI in higher medical education.[86]

The existing system of text types and uses provides a benchmark against which to assess ChatGPT, but the components of that benchmark are not 100% accurate, acceptable, correct, or free from contestable interpretation. Indeed, the production of error (however defined) and the principle of falsifiability are a necessary condition for the advancement of knowledge.[87] In that case, if ChatGPT is to be criticized as falling short, as it does, to what extent can it serve as a useful tool? The immediate answer lies in a case-by-case detailed evaluation and benchmarking process where each instance is allocated a point on a scale of usefulness and risk. Specific cases include that discussed by Abdelhady and Davis,[88] who reported that ChatGPT was able to record operative notes extremely quickly and to a high level of accuracy compared with manual procedures and was deemed acceptable by surgeons and patients alike. Furthermore, several research studies have been conducted where ChatGPT was required to take a variety of medical tests, such as the UK BMAT, TMUA, LNAT, and TSA examinations, the United States Medical Licensing Exam (USMLE), and certain university tests.[5] In all cases, ChatGPT came out sufficiently well to be deemed able to set and mark tests, although Giannos and Delardas[89] found that it had a poor knowledge of science and mathematics. Indeed, in many areas of medical education, this tool does not offer specialized knowledge, as in the case of pediatric cardiology education.[90] In a further study, Danesh *et al.* (2024) tested both the free and the premium versions of ChatGTP in terms of its ability to pass professional examinations (excluding questions containing imaging data). ChatGTP was able to answer 50% of the questions correctly in the free version and 70% of the questions in the paid version. Similar examination results were obtained in the fields of orthopedics[91] and health professional exams.[92] Finally, Sevgi *et al.*[93] proposed an evaluation of ChatGTP in the field of neurosurgery by asking it to create questions at the level of a neurosurgery board exam. The question format was to be multiple choices and the answers were also to be generated. Next, it was asked to devise artificial neurosurgical cases with examinations and treatment histories. The final stage involved an evaluation of the tool's ability to create articles in this area. It turned out that the proposed cases did serve to help neurosurgery students develop their knowledge. However, it transpired that a correct assessment of the solutions proposed by ChatGTP was only possible under the supervision of a person with appropriate medical knowledge, in this case, an experienced neurosurgeon.

5. Conclusion

The application of AI allows the efficient analysis of huge amounts of data in a finite time. It can be considered a powerful computational tool for solving complex problems related to pattern recognition, classification, grouping, behavior prediction, or, more generally, approximation of functions and processes. Consequently, AI is becoming a highly precise tool in medicine. It is worth stressing that compared to statistical methods; it is more susceptible to various types of threats in comparison, which is a result of its complexity, data dependence, and susceptibility to adversarial attacks. Although AI offers many benefits in medical education, ethical concerns about its accessibility, validity, use, and implementation raise many questions. AI can be implemented in medical education in a variety of beneficial and relatively uncontroversial ways. These include the rapid analysis of large-scale simulated datasets (thus obviating requirements of real-life patient data regulation), pattern recognition and diagnostics (as in radiology), general educational support in the design of personalized learning programs (at least on a basic level), models built in VR as teaching aids, and uses of GenText such as rapid assembly of post-operative notes.

On the other hand, barriers to implementation and use in developing countries, such as limited internet connectivity, have resulted in lower levels of discussion around global fairness as an ethical issue. In addition, input data latency and potential dataset and algorithmic bias raise genuine concerns about the output validity, especially regarding GenText and statistical analytical output. A particular concern in discursive production is the ability of GenText to hallucinate (create non-existent references) and create output text of a biased nature (opinions and accounts ultimately derived from bias found in input datasets and algorithmic structures) that could distort the nature of medical education, leading to bad ethical and practical outcomes in the future. Furthermore, the intensive development of computer hardware, including quantum computers, and the algorithms themselves, and in particular their learning methods, which is the heart of AI, is likely to significantly shorten the time needed for more precise analysis, which is crucial in the context of medical data.

Acknowledgments

None.

Funding

This study was partially supported by the National Center for Research and Development (research grant Infostrateg I/0042/2021-00).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: All authors
Writing – original draft: All authors
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

Further disclosure

None.

References

1. Dolega-Dolegowski D, Proniewska K, Dolega-Dolegowska M, *et al.* Application of holography and augmented reality based technology to visualize the internal structure of the dental root – a proof of concept. *Head Face Med*. 2022;18(1):12.

   doi: 10.1186/s13005-022-00307-4

2. Baker RS, Hawn A. Algorithmic bias in education. *Int J Artif Intell Educ*. 2022;32:1052-1092.

   doi: 10.1007/s40593-021-00285-9

3. Albahri AS, Duhaim AM, Fadhel MA, *et al*. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inform Fusion*. 2023;96:156-191.

   doi: 10.1016/J.INFFUS.2023.03.008

4. Rudnicka Z, Szczepanski J, Pregowska A. Artificial intelligence-based algorithms in medical image scan segmentation and intelligent visual content generation-a concise overview. *Electronics* (*Basel*). 2024;13(4):746.

   doi: 10.3390/electronics13040746

5. Pregowska A, Perkins M. *Artificial Intelligence in Medical Education: Technology and Ethical Risk*. Available from: https://ssrn.com/abstract=4643763 [Last accessed on 2024 Oct 18].

   doi: 10.2139/ssrn.4643763

6. Naik N, Hameed BMZ, Shetty DK, *et al*. Legal and ethical consideration in artificial intelligence in healthcare: Who takes responsibility? *Front Surg*. 2022;9:862322.

   doi: 10.3389/fsurg.2022.862322

7. Bae CY, Im Y, Lee J, *et al*. Comparison of biological age prediction models using clinical biomarkers commonly measured in clinical practice settings: AI techniques Vs. traditional statistical methods. *Front Anal Sci*. 2021;1:709589.

   doi: 10.3389/frans.2021.709589

8. Hassija V, Chamola V, Mahapatra A, *et al*. Interpretability of black-box models: A review on explainable artificial intelligence (XAI). *Cognit Comput*. 2024;16:45-74.

   doi: 10.1007/s12559-023-10179-8

9. Kunze KN, Williams RJ 3rd, Ranawat AS, *et al*. Artificial intelligence (AI) and large data registries: Understanding the advantages and limitations of contemporary data sets for use in AI research. *Knee Surg Sports Traumatol Arthrosc*. 2024;32(1):13-18.

   doi: 10.1002/ksa.12018

10. Baniecki H, Biecek P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. Information Fusion. 2023;107:102303.

    doi: 10.1016/j.inffus.2024.102303

11. Rethlefsen ML, Kirtley S, Waffenschmidt S, *et al*. PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews. *Syst Rev*.

2021;10:39.

doi: 10.1186/s13643-020-01542-z

12. Lee J, Wu AS, Li D, Kulasegaram K (Mahan). Artificial intelligence in undergraduate medical education: A scoping review. *Acad Med*. 2021;96(11S):S62-S70.

doi: 10.1097/ACM.0000000000004291

13. Weidener L, Fischer M. Teaching AI ethics in medical education: A scoping review of current literature and practices. *Perspect Med Educ*. 2023;12:399-410.

doi: 10.5334/pme.954

14. Murat Civaner M, Uncu Y, Bulut F, Chalil G, Tatli A. A three-arm single blind randomised control trial of naïve medical students performing a shoulder joint clinical examination. *BMC Med Educ*. 2021;22;772.

doi: 10.1186/s12909-022-03852-3

15. Andersson J, Nyholm T, Ceberg C, *et al*. Artificial intelligence and the medical physics profession - a Swedish perspective. *Phys Med*. 2021;88:218-225.

doi: 10.1016/J.EJMP.2021.07.009

16. Mosch L, Fürstenau D, Brandt J, *et al*. The medical profession transformed by artificial intelligence: Qualitative study. *Digit Health*. 2022;8:20552076221143903.

doi: 10.1177/20552076221143903

17. Garlinska M, Osial M, Proniewska K, Pregowska A. The influence of emerging technologies on distance education. *Electronics* (*Switzerland*). 2023;12(7):1550.

doi: 10.3390/electronics12071550

18. Klauschen F, Dippel J, Keyl P, *et al*. Toward explainable artificial intelligence for precision pathology. *Annu Rev Pathol Mech Dis*. 2024;19:541-570.

doi: 10.1146/annurev-pathmechdis

19. Gordon M, Daniel M, Ajiboye A, *et al*. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. *Med Teach*. 2024;46:446-470.

doi: 10.1080/0142159X.2024.2314198

20. Pinto Dos Santos D, Giese D, Brodehl S, *et al*. Medical students' attitude towards artificial intelligence: A multicentre survey. *Eur Radiol*. 2019;29:1640-1646.

doi: 10.1007/s00330-018-5601-1

21. Chengoden R, Victor N, Huynh-The T, *et al*. Metaverse for healthcare: A survey on potential applications, challenges and future directions. *IEEE Access*. 2023;PP:1-1.

doi: 10.1109/ACCESS.2023.3241628

22. Musamih A, Yaqoob I, Salah K, *et al*. Metaverse in healthcare: Applications, challenges, and future directions. *IEEE Consum Electron Mag*. 2023;12(4):33-46.

doi: 10.1109/MCE.2022.3223522

23. Zechner O, Guirao DG, Schrom-Feiertag H, *et al*. Multimodal technologies and interaction NextGen training for medical first responders: Advancing mass-casualty incident preparedness through mixed reality technology. *Multimodal Technol Interact*. 2023;7:113.

doi: 10.3390/mti7120113

24. Balak N, Ganau M, Tsianaka E, Park JJ, Tiefenbach J, Demetriades AK. The role of artificial intelligence in surgical simulation. *Front Med Technol*. 2022;4:1076755.

doi: 10.3389/fmedt.2022.1076755

25. Neves CA, Tran ED, Kessler IM, Blevins NH. Fully automated preoperative segmentation of temporal bone structures from clinical CT scans. *Sci Rep*. 2021;11:116.

doi: 10.1038/s41598-020-80619-0

26. Hamabe A, Ishii M, Kamoda R, *et al*. Artificial intelligence-based technology to make a three-dimensional pelvic model for preoperative simulation of rectal cancer surgery using MRI. *Ann Gastroenterol Surg*. 2022;6:788-794.

doi: 10.1002/ags3.12574

27. Saricilar EC, Burgess A, Freeman A, *et al*. A pilot study of the use of artificial intelligence with high-fidelity simulations in assessing endovascular procedural competence independent of a human examiner. *ANZ J Surg*. 2023;93:1525-1531.

doi: 10.1111/ans.18484

28. Mirchi IN, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *PLoS One*. 2020;15:e0229596.

doi: 10.1371/journal.pone.0229596

29. Proniewska K, Dolega-Dolegowski D, Kolecki R, Osial M, Pregowska A. The 3D operating room with unlimited perspective change and remote support. In: *Applications of Augmented Reality - Current State of the Art [Working Title]*. London: Intechopen; 2023.

doi: 10.5772/intechopen.1002252

30. Winkler-Schwartz A, Bissonnette V, Mirchi N, *et al*. Artificial intelligence in medical education: Best practices using machine learning to assess surgical expertise in virtual reality simulation. *J Surg Educ*. 2019;76(6):1681-1690.

doi: 10.1016/J.JSURG.2019.05.015

31. Natheir S, Christie S, Yilmaz R, *et al*. Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task. *Comput Biol Med*. 2023;152:106286.

doi: 10.1016/j.compbiomed.2022.106286

32. Wölfel M, Taecharungroj V. "What Can ChatGPT Do?" Analyzing early reactions to the innovative AI Chatbot on twitter. *Big Data Cogn Comput*. 2023;7:35.

doi: 10.3390/bdcc7010035

33. Alkaissi H, Mcfarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15:e35179.

doi: 10.7759/cureus.35179

34. Available from: https://lawlibguides.sandiego.edu/c.php?g=1317323&p=9686671 [Last accessed on 2024 Oct 18].

35. Beilby K, Hammarberg K. ChatGPT: A reliable fertility decision-making tool? *Hum Reprod*. 2024;39:443-447.

doi: 10.1093/humrep/dead272

36. Funk PF, Hoch CC, Manuel F, *et al*. Citation: ChatGPT's response consistency: A study on repeated queries of medical examination questions. *J Investig Health Psychol Educ*. 2024;14:657-668.

doi: 10.3390/ejihpe14030043

37. Mu Y, He D. The potential applications and challenges of ChatGPT in the medical field. *Int J Gen Med*. 2024;17:817-826.

doi: 10.2147/ijgm.s456659

38. Carr SE, Canny BJ, Wearn A, *et al*. Twelve tips for medical students experiencing an interruption in their academic progress. *Med Teach*. 2022;44(10):1081-1086.

doi: 10.1080/0142159X.2021.1921134

39. Shen Y, Heacock L, Elias J, *et al*. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307(2):e230163.

doi: 10.1148/radiol.230163

40. Lee H. The rise of ChatGPT: Exploring its potential in medical education. 2023.

doi: 10.1002/ase.2270

41. Saleem N, Mufti T, Sohail SS, Madsen DØ. ChatGPT as an innovative heutagogical tool in medical education. Cogent Education. 2024;11(1).

doi: 10.1080/2331186X.2024.2332850

42. Emir B, Yurdem T, Ozel T, *et al*. Artificial intelligence readiness status of medical faculty students. *Konuralp Med J*. 2024;16(1):88-95.

doi: 10.18521/ktd.1387826

43. Rezazadeh H, Ahmadipour H, Salajegheh M. Psychometric evaluation of Persian version of medical artificial intelligence readiness scale for medical students. *BMC Med Educ*. 2023;23(1):527.

doi: 10.1186/s12909-023-04516-6

44. Dennett D. *The Self as a Center of Narrative Gravity. Self and Consciousness: Multiple Perspectives*. Hillsdale, NJ: Lawrence Erlbaum; 1992. p. 53.

45. Acharya V, Padhan P, Bahinipati J, *et al*. Artificial intelligence in medical education. *J Integr Med Res*. 2023;1(3):87-91.

46. Katznelson G, Gerke S. The need for health AI ethics in medical school education. *Adv Health Sci Educ*. 2021;26:1447-1458.

doi: 10.1007/s10459-021-10040-3

47. Ötleş E, James CA, Lomis KD, Woolliscroft JO. Teaching artificial intelligence as a fundamental toolset of medicine. *Cell Rep Med*. 2022;3(12):100824.

doi: 10.1016/J.XCRM.2022.100824

48. Zarei M, Eftekhari Mamaghani H, Abbasi A, Hosseini MS. Application of artificial intelligence in medical education: A review of benefits, challenges, and solutions. *Med Clín Práct*. 2024;7(2):100422.

doi: 10.1016/J.MCPSP.2023.100422

49. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R. Grounded in reality: Artificial intelligence in medical education. *JAMIA Open*. 2023;6:ooad037.

doi: 10.1093/jamiaopen/ooad037

50. Piorkowski A, Obuchowicz R, Najjar R. Redefining radiology: A review of artificial intelligence integration in medical imaging. *Diagnostics* (*Basel*). 2023;13:2760.

doi: 10.3390/diagnostics13172760

51. Brady AP, Allen B, Chong J, *et al*. STATEMENT Open Access. *J Med Imaging Radiat Oncol*. 15.

doi: 10.1186/s13244-023-01541-3

52. Choudhury A, Elkefi S. Acceptance, initial trust formation, and human biases in artificial intelligence: Focus on clinicians. *Front Digit Health*. 2022;4:966174.

doi: 10.3389/fdgth.2022.966174

53. Pagano TP, Loureiro RB, Lisboa FVN, *et al*. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data Cogn. Comput*. 2023;7:15.

doi: 10.3390/bdcc7010015

54. Pessach D, Shmueli E. A review on fairness in machine learning. *ACM Comput Surv*. 2022;55(3):51.

doi: 10.1145/3494672

55. Rudnicka Z, Proniewska K, Perkins M, Pregowska A. Cardiac healthcare digital twins supported by artificial intelligence-based algorithms and extended reality-a systematic review. *Electronics* (*Basel*). 2024;13(5):866.

doi: 10.3390/electronics13050866

56. Ueda D, Kakinuma T, Fujita S, *et al*. Fairness of artificial intelligence in healthcare: Review and recommendations. *Jpn J Radiol*. 2024;42:3-15.

doi: 10.1007/s11604-023-01474-3

57. Shen D, Liu T. Grand challenges in AI in radiology. *Front Radiol*. 2021;1:629992.

doi: 10.3389/fradi.2021.629992

58. Park Y, Hu J. Bias in artificial intelligence: Basic primer. *Clin J Am Soc Nephrol*. 2023;18(3):394-396.

doi: 10.2215/CJN.0000000000000078

59. Meyers PM, Gabelloni M, Wagner M, Schweitzer M, Khosravi P. Artificial intelligence in neuroradiology: A scoping review of some ethical challenges. *Front Radiol.* 2023;3:1149461.

doi: 10.3389/fradi.2023.1149461

60. Bell LC, Shimron E. Sharing data is essential for the future of AI in medical imaging. *Radiol Artif Intell*. 2023;6(1):e230337.

doi: 10.1148/ryai.230337

61. Bernstein MH, Atalay MK, Dibble EH, *et al*. Can incorrect artificial intelligence (AI) results impact radiologists, and if so, what can we do about it? A multi-reader pilot study of lung cancer detection with chest radiography Chest radiograph DSI Data Science Institute FN False negative FP False positive GLMM Generalized linear mixed modeling PACS Picture archiving and communication system. *Eur Radiol*. 2023;33:8263-8269.

doi: 10.1007/s00330-023-09747-1

62. Eltawil FA, Atalla M, Boulos E, Amirabadi A, Tyrrell PN. Analyzing barriers and enablers for the acceptance of artificial intelligence innovations into radiology practice: A scoping review. *Tomography*. 2023;9(4):1443-1455.

doi: 10.3390/tomography9040115

63. Kelly BS, Quinn C, Belton N, *et al*. Cybersecurity considerations for radiology departments involved with artificial intelligence. *Eur Radiol*. 2023;33:8833-8841.

doi: 10.1007/s00330-023-09860-1

64. Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices [Last accessed on 2024 Oct 08].

65. Available from: https://www.gov.uk/government/organisations/medicines-and-healthcare-products-regulatory-agency[Last accessed on 2024 Oct 08].

66. UK Digital Health - the Future of Software as a Medical Device. Available from: https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme-roadmap [Last accessed on 2024 Oct 18].

67. Khazane H, Ridouani M, Salahdine F, Kaabouch N. A holistic review of machine learning adversarial attacks in IoT networks. *Future Internet*. 2024;16:32.

doi: 10.3390/fi16010032

68. Tools M, Tsai MJ, Lin PY. Medical images under tampering. *Multimed Tools Appl*. 2024;83:65407-65439.

doi: 10.1007/s11042-023-17968-1

69. Prajapati JB, Kumar A, Singh S, *et al*. Artificial intelligence-assisted generative pretrained transformers for applications of ChatGPT in higher education among graduates. *SN Soc Sci*. 2024;4:19.

doi: 10.1007/s43545-023-00818-0

70. Narayanan S, Ramakrishnan R, Durairaj E, Das A. Artificial intelligence revolutionizing the field of medical education. *Cureus*. 2023;15(11):e49604.

doi: 10.7759/cureus.49604

71. Lie SS, Helle N, Sletteland NV, Dubland Vikman M, Bonsaksen T. Implementation of virtual reality in health professions education: Scoping Review. *JMIR Res Protoc*. 2022;11:e37222.

doi: 10.2196/37222

72. Dhar E, Upadhyay U, Huang Y, *et al*. A scoping review to assess the effects of virtual reality in medical education and clinical care. *Digit Health*. 2023;9:20552076231158022.

doi: 10.1177/20552076231158022

73. Kim HY, Kim EY, Dominguez-Morales M, Billis A, Kim HY, Kim EY. Effects of medical education program using virtual reality: A systematic effects of medical education program using virtual reality: A systematic review and meta-analysis. *Int J Environ Res Public Health*. 2023;20:3895.

doi: 10.3390/ijerph20053895

74. Leng L. Challenge, integration, and change: ChatGPT and future anatomical education. *Med Educ Online*. 2024;29(1):2304973.

doi: 10.1080/10872981.2024.2304973

75. Pedram S, Kennedy G, Sanzone S. Assessing the validity of VR as a training tool for medical students. *Virtual Real*. 2024;28:15.

doi: 10.1007/s10055-023-00912-x

76. Mergen M, Meyerheim M, Graf N. Reviewing the current state of virtual reality integration in medical education - a scoping review protocol. *Syst Rev*. 2023;12(1):97.

doi: 10.1186/s13643-023-02266-6

77. Mergen M, Meyerheim M, Graf N. Towards integrating virtual reality into medical curricula: A single center student survey. *Educ Sci*. 2023;13:477.

doi: 10.3390/educsci13050477

78. Available from: https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases [Last accessed on 2024 Oct 08].

79. Feng Y, Chen Z, Kang Z, *et al*. JailbreakLens: Visual Analysis of Jailbreak Attacks against Large Language Models.

doi: 10.48550/arXiv.2404.08793

80. Shah R, Feuillade-Montixi Q, Pour S, Tagade A, Casper S, Rando J. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation.

doi: 10.48550/arXiv.2311.03348

81. Eysenbach G. The role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Med Educ*. 2023;9:e46885.

doi: 10.2196/46885

82. Májovský M, Černý M, Kasal M, Komarc M, Netuka D. Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *J Med Internet Res*. 2023;25:e46924.

doi: 10.2196/46924

83. Choi EPH, Lee JJ, Ho MH, Kwok JYY, Lok KYW. Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. *Nurse Educ Today*. 2023;125:105796.

doi: 10.1016/J.NEDT.2023.105796

84. Tsegay SM, Imafuku R, Alam F, Lim MA, Zulkipli IN. Integrating AI in medical education: Embracing ethical usage and critical understanding. *Front Med*. 2023;10:1279707.

doi: 10.1080/14779072.2023.2223978

85. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires Human Judgment. *Radiology*. 2023;307(2):e230171.

doi: 10.1148/radiol.230171

86. Filgueiras F. Artificial intelligence and education governance. *Educ Citizsh Soc Justice*. 2023.

doi: 10.1177/17461979231160674

87. Taran S, Adhikari NKJ, Fan E. Falsifiability in medicine: What clinicians can learn from Karl Popper. *Intensive Care Med*. 2021;47:1054-1056.

doi: 10.1007/s00134-021-06432-z

88. Abdelhady AM, Davis CR. Plastic surgery and artificial intelligence: How ChatGPT improved operation note accuracy, time, and education. *Mayo Clin Proc Digit Health*. 2023;1(3):299-308.

doi: 10.1016/J.MCPDIG.2023.06.002

89. Giannos P, Delardas O. Performance of ChatGPT on UK standardized admission tests: Insights from the BMAT, TMUA, LNAT, and TSA examinations. *JMIR Med Educ*. 2023;9:e47737.

doi: 10.2196/47737

90. Gritti MN, Hussain A, Farid P, Morgan CT. Progression of an artificial intelligence chatbot (ChatGPT) for pediatric cardiology educational knowledge assessment. *Pediatr Cardiol*. 2024;45:309-313.

doi: 10.1007/s00246-023-03385-6

91. Rizzo MG, Cai N, Constantinescu D. The performance of ChatGPT on orthopaedic in-service training exams: A comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. *J Orthop*. 2024;50:70-75.

doi: 10.1016/J.JOR.2023.11.056

92. Davies NP, Wilson R, Winder MS, *et al*. ChatGPT sits the DFPH exam: Large language model performance and potential to support public health learning. *BMC Med Educ*. 2024;24(1):57.

doi: 10.1186/s12909-024-05042-9

93. Sevgi UT, Erol G, Doğruel Y, *et al*. The role of an open artificial intelligence platform in modern neurosurgical education: A preliminary study. *Neurosurg Rev*. 1998;46:86.

doi: 10.1007/s10143-023-01998-2

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Diagnosis of COVID-19 from computed tomography slices using flower pollination algorithm, k-nearest neighbor, and support vector machine classifiers

Betshrine Rachel Jibinsingh[1], Khanna Nehemiah Harichandran[1]*, Kabilasri Jayakannan[2], Rebecca Mercy Victoria Manoharan[3], and Anisha Isaac[1]

[1]Ramanujan Computing Centre, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

[2]Department of Information Science and Technology, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

[3]Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

***Corresponding author:**
Khanna Nehemiah Harichandran (nehemiah@annauniv.edu)

Citation: Jibinsingh BR, Harichandran KN, Jayakannan K, Manoharan RMV, Isaac A. Diagnosis of COVID-19 from computed tomography slices using flower pollination algorithm, k-nearest neighbor, and support vector machine classifiers. *Artif Intell Health*. 2025;2(1):14-28. doi: 10.36922/aih.3349

Received: April 3, 2024

1st revised: May 22, 2024

2nd revised: June 17, 2024

Accepted: June 24, 2024

Published Online: October 23, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Coronavirus disease 19 (COVID-19), caused by the severe acute respiratory syndrome-coronavirus-2 virus, is commonly diagnosed through imaging techniques such as computed tomography (CT) scans, which reveal characteristic lung lesions. In this study, we propose a computer-aided diagnosis (CAD) system to assist in the early detection of COVID-19 from CT lung slices, leveraging advanced machine-learning algorithms for precise and efficient analysis. To achieve this, we developed a CAD system that diagnoses COVID-19 from CT lung slices. An adaptive Wiener filter was applied to remove noise from the CT images. The chest tissues were then segmented using an optimal thresholding method to extract regions of interest, which represent the COVID-19 lesions under investigation. The feature vectors were divided into training and testing with an 80/20 ratio. A wrapper-based flower pollination algorithm was employed alongside the k-nearest neighbor classifier to select the optimal feature set. These selected features were subsequently used to train a support vector machine (SVM) classifier. With feature selection, the SVM achieved an accuracy of 91.30% on a real-time dataset, outperforming seven other machine learning classifiers (radial basis function-SVM, k nearest neighbor, linear discriminant analysis, random forest, naïve Bayes, AdaBoost, extreme gradient boosting) and four deep learning classifiers (convolutional neural network, recurrent neural network, long short term memory, Bidirectional long short term memory). For the publicly available COVID-19 CT dataset, an accuracy of 88.18% was achieved. In conclusion, our COVID-19 CAD system improves diagnostic accuracy, with future work aimed at enhancing efficiency and expanding to covariant detection and severity assessment.

Keywords: Support vector machine; Flower pollination algorithm; k-nearest neighbor; Coronavirus disease 19; Coronavirus disease 19 computed tomography dataset

1. Introduction

The lungs are a pair of spongy, air-filled organs located on either side of the chest. Each lung is roughly cone-shaped, with its base resting on the diaphragm.[1] The lung has two parts: the right lung, which is larger and has three lobes (superior, middle, and inferior), and the left lung, which is smaller and divided into superior and inferior lobes.[2] Lung diseases are conditions that obstruct normal lung function.[3] These include a variety of conditions such as chronic obstructive pulmonary disease, pneumonia, asthma, acute bronchitis, Coronavirus disease 19 (COVID-19), pulmonary edema, idiopathic pulmonary fibrosis, sarcoidosis, pleural effusion, pleurisy, bronchiectasis, cystic fibrosis, lymphangioleiomyomatosis, interstitial lung diseases, lung cancer, tuberculosis, acute respiratory distress syndrome (ARDS), and coccidioidomycosis, and so on.[4] In this research, early detection of COVID-19 is the key focus.

COVID-19 is an infectious disease caused by severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2), which transmits between humans through physical contact, respiratory droplets, and aerosols. The disease is identified by lung lesions detected through imaging techniques, such as X-rays and computed tomography (CT) scans. CT scans make it easier to assess the presence and severity of COVID-19 nodules. Moreover, considering the structural or anatomical details of the lung that are essential for the detection analysis, CT imaging outperforms X-ray radiography in providing knowledge on these.[5] The typical signs of lung lesions, such as ground glass opacity (GGO) in the early stages and consolidation in the later stages, could be observed from CT slices.[6] Studies have reported that radiological imaging, such as CT and X-rays, may be helpful in supporting the early screening of COVID-19.[7,8] Although real-time polymerase chain reaction (RT-PCR) is considered the gold standard for diagnosing COVID-19, recent advancements in medical imaging have significantly improved the diagnosis and quantification of various diseases. Using RT-PCR results as a reference, a study of 1,014 patients in Wuhan, China, achieved an accuracy of 0.68, a sensitivity of 0.97, and a specificity of 0.25 for CT slices indicating COVID-19 infection.[9]

Segmentation is the process of partitioning lung tissues with accurate boundaries from CT slices by eliminating surrounding anatomical structures, such as bones and fat tissues.[10] The objective of segmentation is to extract regions of interest (ROIs) within the lung region to differentiate abnormality from anatomical background. There are 10 different segmentation techniques for lung imaging,[11] including: thresholding,[12-15] region growing method,[16-18] watershed algorithm,[19-21] active contour model,[22,23] clustering techniques,[24-26] level set techniques,[27,28] graph cut techniques,[29,30] genetic algorithms,[31,32] artificial intelligence-based segmentation,[33,34] and hybrid algorithm.[35] In our work, an optimal thresholding approach has been used to locate a value acceptable for segmenting the lung CT slice.

Computer-aided diagnosis (CAD) systems play an important role in assisting physicians in the process of clinical decision-making.[36] In the domain of diagnostic radiology, the CAD system is designed to diagnose abnormalities in images created by imaging modalities. The imaging modalities are X-rays, CT, high-resolution CT, positron emission tomography, single-photon emission CT, and magnetic resonance imaging. A CAD system helps medical professionals by simplifying the process of interpreting numerous images created by different types of imaging, where manual involvement is time-consuming.[37] In the domain of diagnosing pulmonary disorders, the CAD system takes the input image obtained from the imaging modalities, employs computational techniques to locate suspected abnormalities present in the image, and leads to a precise diagnosis. Techniques such as machine learning (ML), image processing, pattern recognition, and deep learning (DL) are commonly employed to enhance abnormality detection in medical images.[38]

In this research, we developed a CAD system to detect the presence or absence of COVID-19. First, an adaptive Wiener filter was used to eliminate the additive noises. Then, optimal thresholding was used to segment the lungs, and relevant features were extracted. To select the optimal feature set, a bio-inspired wrapper-based flower pollination technique was employed, using the accuracy of the k-NN classifier as the fitness function. The support vector machine (SVM) classifier was then trained using the selected optimal subset of features.

This framework can be generalized for applications in biomedical lung imaging diagnosis. This manuscript is structured as follows: **S**ection 2 discusses the relevant literature; Section 3 outlines the system's methodology; Section 4 summarizes the dataset, compares classifiers, evaluates other state-of-the-art approaches, and presents the experimental findings; Section 5 offers conclusions and recommendations for future work.

2. Related works

2.1. Segmentation techniques for CAD to detect COVID-19

Segmentation is an essential step in image processing and analysis for the assessment and quantification of COVID-19. It delineates the ROIs, namely, lung, lobes, lesions or infected regions, and bronchopulmonary

segments, in the chest X-ray or CT slices. Segmented regions could be further used to extract features for diagnosis and other applications. This subsection summarizes the related segmentation works in COVID-19.

Khin *et al.*[39] have proposed a segmentation algorithm to detect COVID-19 in chest CT slices using Deeplab v3+. The dataset used was the COVID-19 radiography database, which contains a total of 15,153 images, including 10,192 normal images, 3,616 COVID-19 images, and 1,345 pneumonia images. Since the dataset was highly imbalanced, five different approaches were employed. The ensemble of convolutional neural network (CNN) with image augmentation achieved an accuracy of 99.23%.

Venkatesan *et al.*[40] have introduced an automated image processing scheme to extract the COVID-19 lesions from lung CT scan images. In their work, the firefly algorithm and Shannon Entropy-based multi-threshold were used to enhance the pneumonia lesions, followed by Markov-Random-Field segmentation to extract the lesions with better accuracy. The dataset was obtained from the COVID-19 database, which includes 100 images for training and 45 images for testing. The proposed scheme was tested and validated using a class of COVID-19 CT images, achieving a mean accuracy >92% in lesion segmentation.

Chandra[41] has demonstrated a segmentation approach using the Cuckoo search algorithm with Otsu's image thresholding for the extraction of COVID-19 pneumonia infection. The proposed approach used Otsu's/Kapur to enhance the value with a threshold of three and employed Level Set techniques to extract ROIs. The dataset included COVID-19 images from 20 patients, and the approach achieved a segmentation accuracy of 97.62.

Mohammed *et al.*[42] have proposed a CAD system for the diagnosis of COVID-19 disease from chest X-ray images. This system can be used to differentiate COVID-19 from other viral pneumonia-like Middle East respiratory syndrome, SARS, and ARDS. Segmentation was performed using Li's[42] method, followed by the application of Law's[42] masks to enhance secondary details in the segmented chest images. Texture features were then extracted using the gray-level co-occurrence matrix (GLCM). The obtained feature vectors were used to build SVM ensemble models. Then, the choices of ensemble classifiers were put together using a weighted voting method. The proposed CAD system achieved an accuracy of 98.04%.

Bhargava *et al.*[43] have introduced an automatic detection system for the diagnosis of COVID-19 from CXR and chest CT slices. Segmentation was done using the FCM algorithm. Four types of features, namely, histograms of gradients, textural, statistical, and discrete wavelet transforms, were extracted using the method of principal component analysis. In the classification, k-NN, sparse representation classifier (SRC), artificial neural network (ANN), and SVM classifiers were used for normal, pneumonia, and COVID-19 classifications. Nine different datasets collected from various sources were examined. The accuracies achieved were 91.70%, 94.40%, 96.16%, and 99.14% by k-NN, SRC, ANN, and SVM, respectively, for COVID-19 diagnosis.

Shankar *et al.*[44] have suggested a CAD system for diagnosing COVID-19 using chest X-ray images. Initially, the Wiener filter was used to pre-process images. The fusion-based feature extraction method was subsequently carried out using GLCM, gray level run length matrix, and local binary patterns. The ideal feature subset was then determined using the Salp swarm algorithm. The images were divided as infected or healthy using an ANN. The obtained outcomes outperformed state-of-the-art techniques. The proposed CAD model's experimental results showed 95.1% and 95.65% accuracy for binary and multiple classes, respectively.

Kadry *et al.*[45] have proposed a classification technique using a machine learning system (MLS) to classify the CT slices as healthy or affected by COVID-19. The MLS includes five steps, namely, tri-level thresholding, segmentation of the image, feature extraction, feature ranking, implementation of serial fusion, and classifier implementation and validation. This proposed system was tested with 500 images, which includes 250 normal and 250 COVID-19-affected images obtained from benchmark datasets (Table 1). The proposed MLS achieved an accuracy of 89.80%.

2.2. CAD system to detect COVID-19 using supervised and un-supervised techniques

Wu *et al.*[46] have proposed a classification system using a random forest (RF) classifier for the diagnosis of COVID-19 disease. The dataset description has been given in Table 1. In the proposed system, 11 key features were selected from 49 features. The model was trained with 11 key features and achieved an accuracy of 96.95%.

Banerjee *et al.*[47] have suggested a binary classification model utilizing ANN, Logistic regression (LR), and LASSO Elastic Net Regularized Generalized Linear Models. The dataset comprised 598 full blood count results obtained from COVID-19 patients. The model with LR achieved an accuracy of 87% for the diagnosis of COVID-19 disease.

Moutaz *et al.*[48] have demonstrated an artificial intelligence technique based on deep CNN to detect COVID-19 disease. The dataset was obtained from the Kaggle dataset, which has 128 images, including 28 healthy

Table 1. Comparison of computer-aided diagnosis systems for diagnosing COVID-19 and the dataset used

References	Contribution	Dataset used	Number of images	Balanced/Unbalanced	Techniques used
Khin *et al.*[39]	Deeplab v3+for diagnosing COVID-19 achieved an accuracy of 99.23%	COVID-19 radiography database	15,153 images, including 10,192 normal, 3,616 COVID-19, and 1,345 pneumonia	Highly unbalanced	Weighted loss, image augmentation, undersampling, oversampling, and hybrid resampling
Kadry *et al.*[45]	Machine learning system using SVM with an accuracy of 89.80%	LIDC-IRDI dataset, RIDER-TCIA dataset, and COVID-19 images from the Radiopedia database	500 images, including 250 normal and 250 COVID-19	Balanced	Balanced dataset from benchmark datasets
Wu *et al.*[46]	Random forest classifier with 11 key features achieved an accuracy of 96.95%	Real-time dataset	253 samples	Balanced	-
Banerjee *et al.*[47]	LR achieved an accuracy of 87%	COVID-19 Data sharing/BR initiative	5644 images, in which 598 samples are considered	Unbalanced	Tested separated for specificity and sensitivity
Moutaz *et al.*[48]	VGG16 with an accuracy of 94.80%	Kaggle dataset	128 images, including 28 healthy and 70 COVID-19 images	Balanced	Data augmentation
Najjar *et al.*[50]	Feature extraction using GLCM and classification using k-NN and SVM classifier. k-NN classifier achieved 99.96%	COVID-19 radiography database	2,399 chest X-ray images, which include 1,577 normal and 822 COVID-19 images	Unbalanced	Using the performance metrics
Maryam *et al.*[51]	Ensemble learning model	COVID-19 Data Sharing/BR initiative	5644 images	Unbalanced	Ensemble model using performance metrics
Atta *et al.*[52]	CSDC-SVM model with an accuracy of 98%	Real-time	547 samples that are classified through the SVM K-fold cross-validation method	Unbalanced	The area under the receiver operating characteristics curve, G-mean, and the F1-score
Tongxue *et al.*[54]	U-Net-based segmentation network using attention mechanism achieved a specificity of 99.3%	Italian Society of Medical and Interventional Radiology: COVID-19 CT segmentation dataset	Dataset 1: 100 axial CT slices from 60 patients with COVID-19 with pleural effusion Dataset 2: 373 slices of COVID-19 with consolidation	Unbalanced	Because of the small data in both datasets, they combine the two datasets as the final training dataset
Mobiny *et al.*[55]	Detail-oriented capsule network architecture with 83.2% accuracy	COVID-19 CT dataset	746 images, which includes 349 COVID-19 and 397 non-COVID-19 images	Unbalanced	Image-to-Image (pi×2pix) conditional GAN architecture augmentation
Hasoon *et al.*[56]	LBP-k-NN, HOG-k-NN, Haralick-k-NN, LBP-SVM, HOG-SVM, and Haralick-SVM. Achieved an accuracy of 89.2% and 98.66%	Github repository	5,000 normal and pneumonia COVID-19 images	Unbalanced	Feature-based balancing

Abbreviations: CSDC-SVM: Cloud-based smart detection algorithm using support vector machine; CT: Computed tomography; GLCM: Gray level co-occurrence matrix; HOG-KNN: Histogram of gradients k nearest neighbor; KNN: K-nearest neighbor; LBP-KNN: Local binary pattern k nearest neighbor; RF: Random forest; SVM: Support vector machine; LR: Logistic regression.

and 70 COVID-19 images. The forecasting methods, namely, the prophet algorithm, auto-regressive algorithm, integrating moving average model, and long short-term memory (LSTM), were used to predict the number of

COVID-19 confirmations. The proposed system achieved an accuracy of 94.80%.

Feng *et al.*[49] have proposed a predictive model using four classifiers, namely LR with LASSO, LR with ridge regularization, decision tree, and adaptive boosting (AB) algorithms, for the early detection of COVID-19 disease. The strength of this proposed model lies in the 46-feature selection. Based on the results, the LR with the LASSO classifier selected only 18 features and achieved an accuracy of 93.80%.

Najjar *et al.*[50] have presented a cutting-edge solution for classifying COVID-19 from chest radiography slices using the SVM and k-NN classifiers. The dataset was obtained from the COVID-19 radiography database, which included 1577 normal and 822 COVID-19 images. The proposed work produced five matrices, namely, GLCM1, GLCM2, GLCM3, GLCM4, and GLCMA, and achieved an accuracy of (95.83 – 97.07%), (95.21 – 97.03%), (95.52 – 96.87%), (95.57 – 97.24%), and (95.94 – 96.87%) with SVM and k-NN classifiers, respectively.

Maryam *et al.*[51] have proposed an ensemble learning model for the diagnosis of COVID-19 from a blood routine test. This proposed model was trained and evaluated using a publicly available dataset in Brazil, which includes 5644 images. This proposed model achieved an accuracy of 99.88% in diagnosing COVID-19 disease.

Atta *et al.*[52] have demonstrated a supervised approach named the cloud-based smart detection algorithm using SVM (CSDC-SVM), tested with 5, 10, 15, and 20 cross-fold validation. The dataset included 547 samples, which were classified using the SVM K-fold cross-validation method. The proposed CSDC-SVM model classifies COVID-19 into four categories, namely, negative, mild, moderate, and severe. The virus can be classified as negative, mild, moderate, or severe, indicating its presence at various levels. The proposed system with CSDC-SVM achieved an accuracy of 98.4% with a 15-fold cross-validation strategy.

The results presented in Table 1 show that to identify the COVID-19 infection more accurately, image-aided diagnosis is important. In addition, by providing the necessary details about the patient who had been admitted with a COVID-19 infection, this system could significantly reduce the pulmonologist's diagnostic burden. The infection rate may be precisely identified when there is an image processing system that is properly developed and implemented.

The aforementioned results were obtained by reviewing this pertinent literature. To begin, ROIs are along lung boundaries; segmenting the lung tissues is essential. Second, training the CAD system with the best ROI features promotes classification performance. Third, a wrapper-based feature selection strategy that uses bio-inspired algorithms is more robust and performs better in a variety of optimization challenges when compared to conventional approaches to feature selection.

3. Methods

The proposed CAD system illustrated in Figure 1 consists of five main steps: (i) segmentation with image enhancement, optimal thresholding, cavity filling, and background removal process; (ii) ROI extraction; (iii) GLCM feature extraction; (iv) selection of features; and (v) classification by building a set of SVM models to classify the chest image into either positive (COVID-19) or negative type (non-COVID-19).

3.1. Segmentation

The objective of segmentation is to partition lung tissues from each lung CT slice. To eliminate additive noise and improve edge sharpness, a Laplacian filter is applied. Next, lung parenchyma is partitioned using an optimal

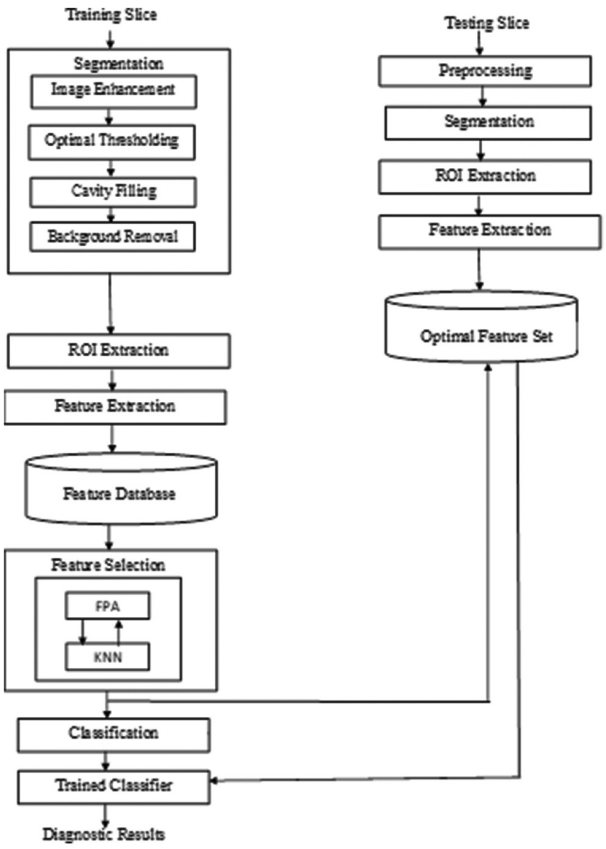


Figure 1. The proposed COVID-19 CAD system. Image created using MS Word application.
Abbreviations: CT: Computed tomography; FPA: Flower pollination algorithm; k-NN: k-nearest neighbor; ROI: Region of interest.

thresholding technique. Morphological techniques are then used to eliminate the background and holes of the lung. An adaptive Wiener filter is used to remove noises from the input CT slices. After removing the noises, optimal thresholding is applied to segment the left and right lung tissues. Optimal thresholding is a method that divides the histogram into two parts to minimize variance within the same class while maximizing separation between different classes.[53] There are two distinct types of pixels that can be seen in a CT slice of the lung, namely, high- and low-intensity pixels. Since their intensity distributions differ, an optimum thresholding approach is used to locate a value acceptable for segregating the lung slice.[57] In the cavity-filling process, the appearance of airways, small holes, or cavities in the binary slice, which represent pathogenic regions, is addressed. Morphological techniques are used to fill these cavities with intensity levels similar to those of neighboring pixels. Pixels with lower intensity values outside the chest cavity are classified as background pixels.[58] In addition, morphological operations are employed to eliminate all connected components smaller than 1000 pixels in the area from the slice.

3.2. Region of interest extraction

The ROIs considered for the COVID-19 CAD model are crazy paving, interlobular septal-thickening, patchy GGO, bilateral GGO, traction bronchiectasis, sub-pleural GGO, peripheral GGO, consolidation, bronchovascular thickening in the lesion, and GGO with consolidation. The ROIs with the pixel intensity score on the scale from 125 to 255 were extracted. The pixel intensity scores <125 are not considered. Each ROI was annotated and labeled by an expert radiologist. Then, Class Label 1 was given to the ROI diagnosed with the presence of COVID-19, and Class Label 2 was given to the ROI diagnosed with the absence of COVID-19.

3.3. Feature extraction using gray level co-occurence matrix

The GLCM-based features were extracted to differentiate between the CT slices with positive and negative cases of COVID-19. The GLCM matrix uses pixel pairs of a joint probability distribution (JPD). The JPD between pixel pairs is calculated by using angle "θ" and the distance "d." The value will be the $(i,j)^{th}$ entry in the GLCM matrix.[42] The features that are extracted from each ROI, as well as the class label that is associated with each ROI, are saved as a feature vector in a database that stores features. From the class labeled ROI, geometrical and textural features were extracted. In our work, 12 geometrical features and 17 textural features, along with four orientations (0°, 45°, 90°, and 135°) were extracted. Then, the feature vector pertaining to each ROI from the 80 extracted features (12 geometric features and 68 texture features) along with the class label were stored in the feature database.[59-61] The features that were extracted from each ROI are outlined in Table 2.

3.4. Feature selection

The goal of this step is to select the optimal feature subset from the extracted features to improve the classifier's predicted performance. The subset of features has been chosen using the Wrapper technique, which combines the Flower pollination algorithm (FPA) and the accuracy of the k-NN classifier as the fitness function.

Table 2. Outline of features extracted from each region of interest

Geometric features
1. Euler number
2. Major axis length
3. Eccentricity
4. Orientation
5. Convex area
6. Filled area
7. Solidity
8. Extent
9. Perimeter
10. Equivalent diameter
11. Minor axis length
12. Area
Texture features (0°, 45°, 90°, and 135°)
1. Sum of squares variance
2. Autocorrelation
3. Cluster prominence
4. Cluster shade
5. Information measure of correlation
6. Energy
7. Correlation
8. Difference variance
9. Dissimilarity
10. Difference entropy
11. Entropy
12. Homogeneity
13. Contrast
14. Inverse difference
15. Maximum probability
16. Sum average
17. Sum entropy

3.4.1. FPA or flower algorithm

Yang[64] introduced the FPA in 2012, inspired by the way blossoming plants attract pollinators.[62-65]

A flower is an angiosperm's bisexual reproductive shoot, with reproductive organs encircled by whorls of sterility organs. Angiosperms are distinguished by a number of characteristics, of which the flower is only one. Sepals, stamens, petals, and carpels are the four whorls that make up the flower. The sepals, which morphologically resemble a whorl of leaves, are the first whorl of the flower. The sepals, which are usually green in color, are formed as lateral extensions from the floral meristem. The petals, which are morphologically identical to leaves, make up the second whorl. The third outer whorl is the sexual organ named the stamens. Stamens and leaves share traits in common in the presence of chlorophyll and their growth form, which is elongation in a single plane (with little or no laminar growth). Female sexual parts, known as carpels, are found in the fourth and innermost whorl of the flower.

Flowers are labeled as bisexual or unisexual depending on the presence of male (gametes) as well as female (gametophyte) reproductive organs. Bisexual or hermaphrodite flowers have both male and female sexual parts. Unisexual flowers have either male sexual parts or female sexual parts. The main function of the sexual organ is to prepare seeds and fruits. The first and foremost step is achieving seeds and fruits, which is possible through pollination. Pollination is divided into two types: self-pollination and cross-pollination. Self-pollination occurs when pollen from one flower's anther is transferred to the stigma of the same bloom (autogamy) or to the stigma of another bloom on the same plant (geitonogamy). Cross-pollination occurs when pollen travels from the anthers in one individual's flower to the stigma in another. Because plants are immobile, pollen movement from plant to plant requires the use of a pollen vector, which can be abiotic or biotic.

Abiotic pollen vectors are primarily caused by water and wind. In wind pollination, the stamen filaments of wind-pollinated flowers are typically long, exposing the locules to the wind and causing an aeroelastic release of pollen as an energy that is transferred from the wind to the stamen through the long filament. Water pollination, also known as hydrophily, is a rather unusual method of gamete transfer used by a few grass and waterweed species. Most hydrophilous species release pollen below the water's surface, where it is passively conveyed by currents to female reproductive structures. Many maritime plants use this process of water pollination.

Biotic pollen vectors, on the other hand, comprise a wide range of species, particularly insects, but also birds, bats, and a small number of other vertebrates that observe flowers as a source of food. Insects are the most common biotic pollen transporter. They receive a complimentary sample of nectar, a sugar solution containing varying amounts of various different sugars as well as other nutrients, and pollen, which is high in amino acids. When animals and insects collect this food, they unintentionally touch the flower's reproductive organs, transferring pollen from stamens to their bodies and from their bodies to stigmatic surfaces. Many flowers have structural elements that promote this unintended interaction.

Pollen grains settle on the stigma's surface and germinate, forming pollen tubes. One of the pollen tubes continues to develop downward. This tube transports male gametes to the ovary. After reaching the ovule, the male gametes get released from the pollen tube and mate with the egg cell. The process of merging the male and female gametes is called fertilization. After fertilization, the ovary becomes larger and develops into a fruit. The fertilized ovule, which results from the fusion of gametes, matures into a seed. The fertilized gamete is referred to as a fertilized ovule. Other aspects of the flower, such as the sepals and the petals, will detach themselves after fertilization has taken place in the bloom. The developed ovary of the flower serves as the primary component of the fruit. The FPA parameters and their respective values are outlined in Table 3.

Output: Feature vectors.

The FPA algorithm has been outlined as follows:

Input: Feature vectors.

Process:

Step 1: Generate a random initial population that is evaluated to determine the current optimal solution.

Initialize the size of the population n, MaxGeneration and p.

Step 2: Initialize the population of n pollen gametes $x = (x_1, x_2, \ldots x_n)$ with random solutions. By calculating the fitness value of each pollen gamete in the population using the k-NN classifier, where the k-NN classifier's accuracy is regarded as the fitness function, the best solution g_* in the initial population is found.

Table 3. Parameters outlined in the flower pollination algorithm

Parameter	Value	Definition
n	10	Initial population
MaxGeneration	100	Maximum no. of iterations
p	(0, 1)	Switch probability
λ	1.5	Control parameter
γ	0.01	Scaling factor

Step 3: Determine the type of pollination based on a predetermined probability p. Generate a random number $r \in [0,1]$, and if $r < p$, where p is the switching probability, then global pollination and flower constancy take place, as described by Equation I:

$$x_i^{t+1} = x_i^t + \gamma L\left(g^* - x_i^t\right) \quad \text{(I)}$$

Where x_i^t denotes the solution of i at iteration t, γ is a scaling factor, and g_* is the current optimal solution at iteration . The parameter L is the pollination strength, in which essentially a step size is drawn from Levy flight, which is given by Equation II:

$$L \sim \frac{\lambda * \Gamma(\lambda) * \sin(\frac{\pi\lambda}{2})}{\pi} * \frac{1}{S^{1+\lambda}}, (s \gg 0) \quad \text{(II)}$$

Where, $\Gamma(\lambda)$ denotes the standard gamma function, and this distribution is for $S > 0$. λ is the tail amplitude of the distribution's control parameter. Commonly, it is recommended to use $\lambda = 1.5$, which is followed in all simulations.

Step 4: Otherwise, if $r > p$, then the local pollination and the flower constancy are performed, as described by Equation III:

$$x_i^{t+1} = x_i^t + \varepsilon\left(x_j^t - x_k^t\right) \quad \text{(III)}$$

Where, x_j^t and x_k^t are pollens from other flowers of the same plant species, with j and k chosen at random from all the solutions. $\varepsilon \in [0,1]$ is a random number.

Step 5: Evaluate each new solution x_i^{t+1} in the population and update the population according to their fitness value.

Step 6: Calculate the current best solution g^* by ranking the solution.

Step 7: Repeat Steps 3 through 6 until MaxGeneration is reached or until convergence is achieved.

In this feature selection Step 24 features have been selected namely Area, Minor Axis Length, Convex Area, Eccentricity, Cluster Prominence 2, Cluster Prominence 3, Contrast 1, Contrast 3, Correlation 2, Correlation 3, Difference Variance 4, Dissimilarity 2, Dissimilarity 4, Energy 1, Entropy 1, Entropy 2, Entropy 4, Homogeneity 4, Information Measure of Correlation 1, Information Measure of Correlation 4, Inverse Difference 1, Sum Average 1, Sum entropy 1, Sum of Squares Variance 4.

3.5. Classification

The SVM algorithm was employed to train the optimal feature subset. A SVM is a supervised learning algorithm that uses hyperplanes to separate different classes. The distance of a feature vector from these hyperplanes indicates how likely it is to belong to a specific class.[66,67] In ML, SVM is a model that classifies and predicts outcomes based on training data. The main goal of SVM is to identify the best hyperplane that divides two classes within a feature set. In other words, the SVM training method builds a model that assigns new examples into one of the two classes based on a set of training examples for binary classification. A SVM assigns training samples in a spatial arrangement that maximizes the separation between the two classes. When new samples are introduced, they are similarly positioned in this space, and their class is forecasted based on which side of the hyperplane they fall. The SVM classifier was trained using the set of FPA-selected features. Then, the performance of the trained SVM classifier was validated using the test dataset.[67]

4. Results

This section includes a description of the real-time dataset and public dataset used in this research, as well as performance evaluation, comparison results of ML and DL classifiers, and experimental results.

4.1. Dataset outline

The research utilized two datasets: a real-time dataset collected from Bharat Scan Centre, Chennai, India, and a COVID-19 CT dataset obtained from the GitHub repository. In the real-time dataset, CT slices were labeled by an expert radiologist as either "normal" or "COVID-19." This dataset includes images from 41 individuals, comprising 26 with COVID-19 and 15 with healthy lungs. Among the COVID-19 patients, 19 exhibited mild severity while seven had moderate severity; the cohort included 17 females and nine males. The ages of the COVID-19 patients ranged from 23 to 49 years, with an average of 36 years. The images in the dataset have a pixel size of 512 × 512 and are in jpg format. The nodule size ranges from 3 to 30 mm, with lesions primarily located in the sub-pleural and posterior respiratory zones. The ROIs were patchy GGO, bilateral GGO, subpleural GGO, peripheral GGO, broncho-vascular thickening, traction bronchiectasis, consolidations, and GGO with consolidations.

The datasets have been divided into training and testing datasets, with training datasets comprising 80% of the total and testing datasets 20%. To preserve privacy, we have masked all personal information from CT slices. Each ROI has been differentiated based on the opinion of an experienced radiologist. In addition, the radiologist manually identified and described each ROI. Table 4 gives an overview of the real-time dataset.

Table 4. Overview of the experimental dataset

Patient cases	Total no. of patients	Total COVID-19 CT slices considered	ROIs	Training set ROIs	Testing set ROIs
COVID-19	26	342	343	242	101
Normal	15	446	452	394	58
Total	41	788	795	636	159

Abbreviations: CT: Computed tomography; ROI: Region of interest.

For the COVID-19 CT database, a publicly available dataset was utilized to train and test the proposed model. It contains a total of 349 COVID-19 CT images from 216 patients and 463 non-COVID-19 CTs, which have been divided into two classes, namely, COVID-19 and non-COVID-19. The COVID-19 CT dataset was divided into training and testing datasets, with training datasets comprising 80% of the total and testing datasets 20%. A pre-processed version of the dataset is available at https://github.com/UCSD-AI4H/COVID-CT.

4.2. Performance evaluation

The aim of this work is to decrease the false negative and false positive values, that is, to increase the sensitivity and specificity, respectively. However, there is often a tradeoff between sensitivity and specificity; as one increases the other decreases. In the proposed research, we obtained inferences from the radiologist. He reviewed the model and provided feedback, suggesting that although it works well, more CT slices should be included so that it may be used to diagnose different lung diseases. Figures 2 and 3 display the effectiveness of the CAD system's implementation for patients with and without COVID-19. The algorithm's optimization performance was compared in terms of accuracy, precision, recall or sensitivity, and specificity, with results obtained using Equations IV–VII:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} \quad \text{(IV)}$$

$$\text{Precision} = \frac{a}{a+b} \quad \text{(V)}$$

$$\text{Sensitivity} = \frac{a}{a+c} \quad \text{(VI)}$$

$$\text{Specificity} = \frac{d}{b+d} \quad \text{(VII)}$$

Where a, b, c, and d denote actual positives, predicted positives, predicted negatives, and predicted positives, respectively. The confusion matrix obtained for FPA is shown below in Table 5.

The extraction of COVID-19 lesions from a chest CT slice demonstrating the presence of the COVID-19 disease is depicted in Figure 4A-D. Figure 4A displays the reference chest CT slice. Figure 4B and C illustrate the segmentation and feature extraction processes necessary for effectively isolating the nodules. Figure 4D displays the peripheral GGO lesion that was excised, indicating the presence of COVID-19.

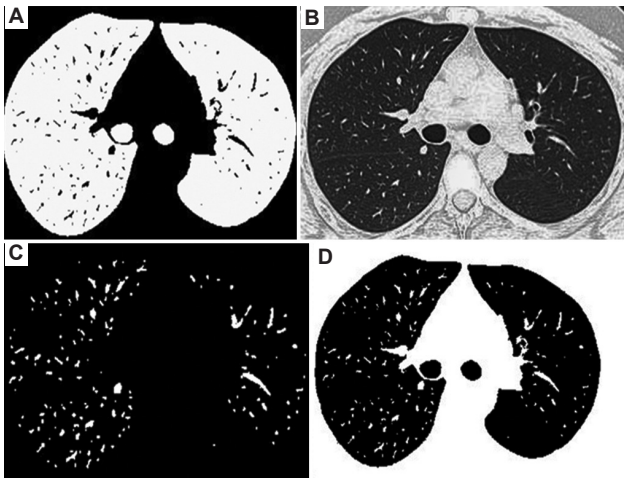


Figure 2. Experimental images of a normal lung CT slice. (A) Non-COVID-19 input CT slice. (B) Segmented image. (C) Extracted ROI. (D) Non-COVID-19 nodules. These images were generated using Python
Abbreviations: CT: Computed tomography; ROI: Region of interest.

Figures 2A-D depict the steps involved in the extraction of ROIs that indicate the absence of COVID-19 disease. The input CT slice of the lung is displayed in Figure 2A. The output image of various steps involved in extracting the nodules is shown in Figures 2B and C. The nodules extracted are shown in Figure 2D.

The CAD system that utilizes FPA for feature selection with 100 iterations produced a greater accuracy of 91.30% for the real-time dataset and 88.18% for the COVID-19 CT dataset. The performance comparison using the real-time and COVID-19 CT datasets is outlined in Table 6.

4.3. Comparison with machine learning and DL classifiers

The proposed CAD system was compared against seven traditional ML classifiers and four DL classifiers. The ML classifiers included radial basis function SVM, k-NN,

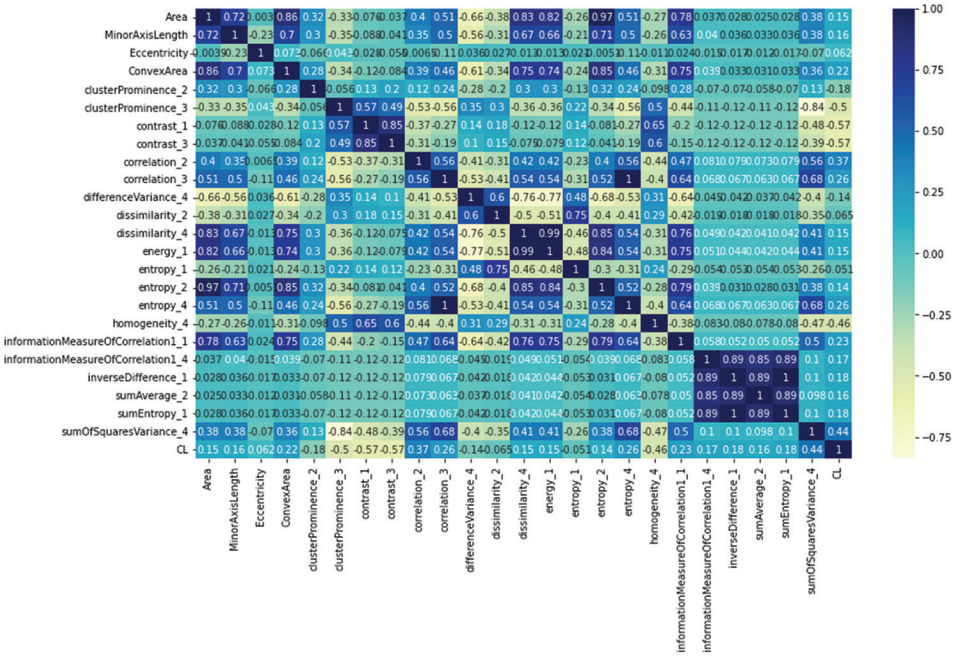


Figure 3. Kendal's rank correlation map. Output generated using the Python application

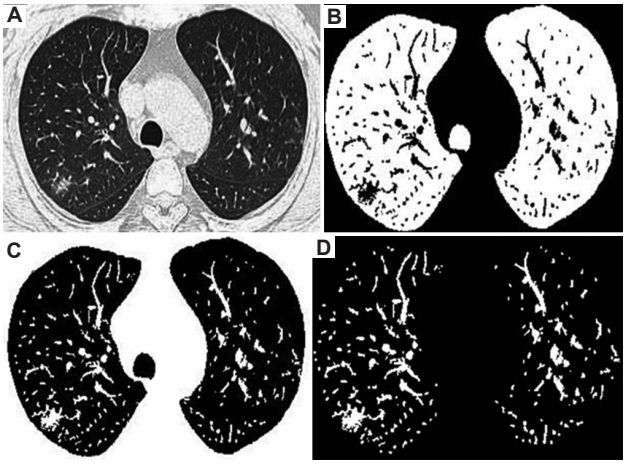


Figure 4. Experimental images obtained for COVID-19 CT slices. (A) COVID-19 input CT slice. (B) Segmented image. (C) Extracted ROI. (D) COVID-19 nodules. These images were generated using Python
Abbreviations: CT: Computed tomography; ROI: Region of interest.

linear discriminant analysis, RF, naïve bias, extreme gradient boosting, and AB. The four DL classifiers used for comparison were CNN, recurrent neural network, LSTM, and bidirectional LSTM, respectively. Our system outperformed these ML classifiers with an accuracy of 91.30%. For each model, average (± standard deviation) performance was reported over 30 iterations. The comparison of ML and DL classifiers in terms of accuracy, precision, recall, and specificity, along with mean and standard deviation values, is presented in Tables 7 and 8.

Table 5. Generated confusion matrix

Actual/predicted	Predicted positive	Predicted negative
Actual positive	94	9
Actual negative	5	51

Table 6. Performance comparison using real-time and COVID-19 dataset

Performance metrics average	Real-time dataset	COVID-19 CT dataset
Accuracy (*M*±*SD*)	0.9130±0.0177	0.8818±0.0180
Precision (*M*±*SD*)	0.8989±0.0324	0.9192±0.0280
Recall (*M*±*SD*)	0.8003±0.0340	0.8956±0.0305
Specificity (*M*±*SD*)	0.9374±0.0218	0.8574±0.0538
F1 score (*M*±*SD*)	0.9302±0.0217	0.9065±0.0140
Selected features	24	22

Abbreviation: CT: Computed tomography

4.4. Comparison with other state-of-the-art approaches using the COVID-19 CT dataset

Our proposed CAD system metrics using the COVID-19 CT dataset obtained from the GitHub repository were compared with other state-of-the-art approaches[55,69-72] for diagnosing COVID-19 disease (Table 9). A maximum accuracy of 89.36% in this comparison was achieved by Ali and Assadi[71], whereas our CAD system using the COVID-19 CT dataset produced an accuracy of 88.18%.

Table 7. Machine learning classifier comparison

Classifier/ performance metrics	RBF-SVM	k-NN	LDA	RF	NB	EB	AB	Our proposed system using real-time dataset
Accuracy (*M*±*SD*)	0.6329±0.0387	0.8572±0.0243	0.8706±0.0210	0.8996±0.0180	0.7541±0.0403	0.9044±0.0232	0.8753±0.0220	0.9130±0.0177
Precision (*M*±*SD*)	0.9189±0.0660	0.8779±0.0481	0.8861±0.0341	0.9135±0.0337	0.9093±0.0551	0.9113±0.0388	0.8697±0.0404	0.8989±0.0324
Recall (*M*±*SD*)	0.1815±0.0515	0.7855±0.0495	0.8095±0.0435	0.8524±0.0432	0.4883±0.0722	0.8673±0.0403	0.8434±0.0381	0.8003±0.0340
Specificity (*M*±*SD*)	0.9867±0.0111	0.9149±0.0320	0.8095±0.0435	0.9373±0.0235	0.9625±0.0230	0.9342±0.0281	0.9009±0.0321	0.9302±0.0217

Abbreviations: AB: AdaBoost; EB: Extreme boosting; k-NN: k-nearest neighbor; LDA: Linear discriminant analysis; NB: Naïve bias; RBF-SVM: Radial basis function-support vector machine; RF: Random forest.

Table 8. Deep learning classifier comparison

Classifiers/ performance metrics	CNN (%)	RNN (%)	LSTM (%)	BLSTM (%)
Training accuracy	89.15	84.74	80.66	83.64
Testing accuracy	89.31	85.53	83.01	83.67
Training precision	88.54	81.29	80.57	81.27
Testing precision	84.81	84.50	83.58	81.94
Training recall	85.61	83.39	71.95	80.07
Testing recall	93.05	82.33	77.77	81.94
Training specificity	87.05	82.33	76.02	80.67
Testing specificity	88.74	83.91	80.57	81.94

Abbreviations: CNN: Convolutional neural network; BLSTM: Bidirectional LSTM; LSTM: Long short-term memory; RNN: Recurrent neural network.

Table 9. Comparison of the proposed CAD system with state-of-the-art approaches for the COVID-19 CT dataset

State-of-the-art approaches	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
Mobiny *et al.* using Inception V3.[55]	85.3	84.4	74	85.3	78.1
Mobiny *et al.* using DenseNet 121.[55]	82.5	81.5	79.4	83.9	80.1
Xingyi *et al.* using DenseNet-169.[69]	79.5	-	-	-	76
Polsinelli *et al.*[70]	85.03	85.01	81.44	88.23	83.98
Xingyi *et al.* using ResNet-50[69]	77.4%	-	-	-	74.6
Ali and Assadi[71]	89.26	-	-	-	89.18
Pedro *et al.*[72]	87.6	-	-	-	86.19
Our proposed CAD system	88.18	91.92	89.56	85.74	90.65

Abbreviations: CAD: Computer-aided diagnosis; CT: Computed tomography.

Our proposed system achieved higher precision, recall, and F1 score values, as shown in Table 9.

4.5. Statistical test

The Mann–Whitney U test revealed significant differences between the variables and the dependent variable ($P < 0.001$). The difference is statistically significant ($P < 0.001$). The $P = 0.001$, which is less than the minimum value of 0.05 for significance. Kendal's rank correlation coefficient map examines sample correlation. Kendal's correlation map for the selected attributes in the dataset is given in Figure 3.

5. Conclusion

Our proposed COVID-19 CAD system achieved an accuracy of 91.30% on a real-time dataset and 88.18% accuracy on the COVID-19 CT Public Dataset. Notably, our system demonstrated significant superiority over seven state-of-the-art ML classifiers and four DL classifiers. This shows that our COVID-19 model excels in generating robust and highly discriminative features. The primary goal of our research is to improve classification accuracy and aid physicians in clinical decision-making. Hence, time and space complexity are not the primary interests of this research work. The suggested CAD system exhibited improved accuracy when employing FPA with k-NN and SVM classifiers because it increased the test accuracy and time efficiency. Since the FPA algorithm is larger than some algorithms, more memory is needed. In addition, since this is a classification system, it does not provide information on disease severity.

In the future, this work can be extended to identify the covariants of COVID-19 and the assessment of COVID-19's severity. Optimizing the system's architecture and integrating other feature selection methods are two excellent methods to improve the rapidity of the COVID-19 CAD system. Importantly, for the COVID-19 CAD system to be clinically validated, it should be implemented in real-world settings, such as by training it on a hospital's private

database. This would allow for a thorough evaluation and improvement of its clinical validity.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Betshrine Rachel Jibinsingh, Khanna Nehemiah Harichandran
Formal Analysis: Betshrine Rachel Jibinsingh
Investigation: Betshrine Rachel Jibinsingh, Kabilasri Jayakannan, Rebecca Mercy Victoria Manoharan
Methodology: Betshrine Rachel Jibinsingh, Khanna Nehemiah Harichandran, Anisha Isaac
Writing – original draft: Betshrine Rachel Jibinsingh
Writing – review & editing: Betshrine Rachel Jibinsingh, Khanna Nehemiah Harichandran

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data used in this work are available from the corresponding author on reasonable request.

References

1. Fuller HW. Lea HC, editor. *On Diseases of the Lungs and Air-Passages: Their Pathology, Physical Diagnosis, Symptoms, and Treatment*. Kissingert: Publishing; 1867.
2. Tomashefski JF, Farver CF. Anatomy and histology of the lung. In: *Dail and Hammar's Pulmonary Pathology*. Germany: Springer; 2008. p. 20-48.
3. Demedts M, Wells AU, Anto JM, *et al*. Interstitial lung diseases: An epidemiological overview. *Eur Respir J*. 2001;18(32):2S-16S.
4. Schwarz T, Johnson V. Lungs and bronchi. In: *Veterinary Computed Tomography*. United States: John Wiley and Sons; 2011. p. 261-262.
5. Fan DP, Zhou T, Ji GP, *et al*. Inf-net: Automatic COVID-19 lung infection segmentation from CT images. *IEEE Trans Med Imaging*. 2020;39(8):2626-2637.
6. Carotti M, Salaffi F, Puttini PS, *et al*. Chest CT features of coronavirus disease 2019 (COVID-19) pneumonia: Key points for radiologists. *J Natl Public Health Emerg Collect*. 2020;125:636-646.

   doi: 10.1007/s11547-020-01237-4
7. Fang Y, Zhang H, Xie J, *et al*. Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. *Radiology*. 2020;296(2):E115-E117.

   doi: 10.1148/radiol.2020200432
8. Ng MY, Lee EY, Yang J, *et al*. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiol Cardiothorac Imaging*. 2020;2(1):e200034.

   doi: 10.1148/ryct.2020200034
9. Ai T, Yang Z, Hou H, *et al*. Correlation of chest CT and RT-PCR testing in COVID-19 in China: A report of 1014 cases. *Thorac Imaging Radiol*. 2020;296(2):E32-E40.

   doi: 10.1148/radiol.2020200642
10. Pal NR, Pal SK. A review on image segmentation techniques. *Pattern Recognit*. 1993;26(9):1277-1294.

    doi: 10.1016/0031-3203(93)90135-J
11. Norouzi A, Rahim MS, Altameem A, *et al*. Medical image segmentation methods, algorithms, and applications. *IETE Tech Rev*. 2014;31(3):199-213.

    doi: 10.1080/02564602.2014.906861
12. Sitanggang S, Sonang S, Yuhandri Y, Setiawan A. Image transformation with lung image thresholding and segmentation method. *J RESTI* (*Rekayasa Sist Teknol Inform*). 2023;7(2):278-285.

    doi: 10.29207/resti.v7i2.4321
13. Yu T, Huang L. An Adaptive Thresholding Method for Automatic Lung Segmentation in CT Images. In: *IEEE AFRICON Conference*; 2009. p. 1-5.
14. Sweetlin JD, Nehemiah HK, Kannan A. Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images. *Comput Methods Programs Biomed*. 2017;145:115-125.

    doi: 10.1016/j.cmpb.2017.04.009
15. Sweetlin JD, Nehemiah HK, Kannan A. Computer aided diagnosis of drug sensitive pulmonary tuberculosis with cavities, consolidations and nodular manifestations on lung CT images. *Int J Bio Inspired Comput*. 2019;13(2):71-85.

    doi: 10.1504/IJBIC.2019.098405
16. Dehmeshki J, Amin H, Valdivieso M, Ye X. Segmentation of pulmonary nodules in thoracic CT scans: A region growing approach. *IEEE Trans Med Imaging*. 2008;27(4):467-480.

    doi: 10.1109/TMI.2007.907555
17. Nabipour S, Khorshidi A, Noorian B. Lung tumor segmentation using improved region growing algorithm.

Nucl Eng Technol. 2020;52(10):2313-2319.

doi: 10.1016/j.net.2020.03.011

18. Prabin A, Veerappan J. Automatic segmentation of lung CT images by CC based region growing. *J Theor Appl Inform Technol*. 2014;68(1):63-69.

19. Avinash S, Manjunath K, Kumar SS. An Improved Image Processing Analysis for the Detection of Lung Cancer Using Gabor Filters and Watershed Segmentation Technique. In: *IEEE International Conference on Inventive Computation Technologies*; 2016.

20. Kumar SL, Swathy M, Sathish S, Sivaraman J, Rajasekar M. Identification of lung cancer cell using watershed segmentation on CT images. *Indian J Sci Technol*. 2016;9:1-4.

doi: 10.17485/ijst/2016/v9i1/85765

21. Shojaii R, Alirezaie J, Babyn P. Automatic Lung Segmentation in CT Images Using Watershed Transform. In: *IEEE International Conference on Image Processing*; 2005.

22. Nithila EE, Kumar SS. Segmentation of lung from CT using various active contour models. *Biomed Signal Process Control*. 2019;47:57-62.

23. Kasinathan G, Jayakumar S, Gandomi AH, *et al*. Automated 3-D lung tumor detection and classification by an active contour model and CNN classifier. *Expert Syst Appl*. 2019;15(134):112-119.

doi: 10.1016/j.eswa.2019.05.041

24. Sangamithraa PB, Govindaraju S. Lung Tumor Detection and Classification Using EK-Mean Clustering. In: *IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*; 2016. p. 2201-2206.

25. Joon P, Bajaj SB, Jatain A. Segmentation and detection of lung cancer using image processing and clustering techniques. In: *Advanced Computing and Intelligent Engineering*. Germany: Springer Nature; 2019. p. 13-23.

26. Xu M, Qi S, Yue Y, *et al*. Segmentation of lung parenchyma in CT images using CNN trained with the clustering algorithm generated dataset. *Biomed Eng Online*. 2019;18:2.

doi: 10.1186/s12938-018-0619-9

27. Farag AA, Munim HE, Graham JH, Farag AA. A novel approach for lung nodules segmentation in chest CT using level sets. *IEEE Trans Image Process*. 2013;22:5202-5213.

doi: 10.1109/TIP.2013.2282899

28. Swierczynski P, Papież BW, Schnabel JA, Macdonald C. A level-set approach to joint image segmentation and registration with application to CT lung imaging. *Comput Med Imaging Graph*. 2018;65:58-68.

doi: 10.1016/j.compmedimag.2017.06.003

29. Wei J, Deihui X, Zhang B, Wang L, Kopriva I, Chen X. Random walk and graph cut for co-segmentation of lung tumor on PET-CT images. *IEEE Trans Image Process*. 2015;24(12):5854-5867.

doi: 10.1109/TIP.2015.2488902

30. Ali AM, Farag AA. Automatic Lung Segmentation of Volumetric Low-dose CT Scans Using Graph Cuts. In: *International Symposium on Visual Computing*; 2008. p. 258-267.

31. Bhuvaneswari P, Therese BA. Detection of cancer in lung with k-NN classification using genetic algorithm. *Procedia Mater Sci*. 2015;10:433-440.

doi: 10.1016/j.mspro.2015.06.077

32. Filho DC, Silva AO, Paiva AC, Nunes RA, Gattass M. Computer-aided diagnosis system for lung nodules based on computed tomography using shape analysis, a genetic algorithm, and SVM. *Med Biol Eng Comput*. 2017;55:1129-1146.

doi: 10.1007/s11517-016-1577-7

33. Herrmann P, Busana M, Cressoni M, *et al*. Using artificial intelligence for automatic segmentation of CT lung images in acute respiratory distress syndrome. *Front Physiol*. 2021;12:76118.

doi: 10.3389/fphys.2021.676118

34. Shi F, Wang J, Shi J, *et al*. Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Rev Biomed Eng*. 2021;14:4-15.

doi: 10.1109/RBME.2020.2987975

35. Van EM, Hoop D, Viergever MA, Prokop M, Ginneken BV. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Med Phys*. 2009;36:2934-2947.

doi: 10.1118/1.3147146

36. Doi K. Computer-aided diagnosis in medical imaging: Historical review, current status, and future potential. *Comput Med Imaging Graph*. 2007;31(5):198-211.

doi: 10.1016/j.compmedimag.2007.02.002

37. Choi YJ, Baek JH, Park HS, *et al*. A computer-aided diagnosis system using artificial intelligence for the diagnosis and characterization of thyroid nodules on ultrasound: Initial clinical assessment. *Thyroid*. 2017;27(4):546-552.

doi: 10.1089/thy.2016.0372

38. Isaac A, Nehemiah HK, Dunston DS, Christo VRE, Kannan A. Feature selection using competitive coevolution of bio-inspired algorithms for the diagnosis of pulmonary emphysema. *Biomed Signal Process Control*. 2022;72:103340.

doi: 10.1016/j.bspc.2021.103340

39. Khin Y, Maneerat N, Sreng S, Hamamoto K. Ensemble deep learning for the detection of COVID-19 in unbalanced chest X-ray dataset. *Appl Sci*. 2021;11(22):10528.

doi: 10.3390/app112210528

40. Venkatesan R, Kadry R, Thanaraj KP, Kamalanand K, Seo S.

Firefly-Algorithm Supported Scheme to Detect COVID-19 Lesion in Lung CT Scan Images using Shannon Entropy and Markov Random Field. [arXiv Preprint].

41. Chandra SC. Segmentation and evaluation of COVID-19 lesion from CT scan slices-a study with Kapur/Otsu function and Cuckoo Search Algorithm. 2020.

    doi: 10.21203/rs.3.rs-40148/v1

42. Mohammed SN, Alkinani FS, Hassan YA. Automatic computer-aided diagnostic for COVID-19 based on chest X-ray image and particle swarm intelligence. *Int J Intell Eng Syst*. 2020;13(5):63-73.

43. Bhargava A, Bansal A, Goyal V. Machine learning-based automatic detection of novel coronavirus (COVID-19) disease. *Multimed Tools Appl*. 2022;81(10):13731-13750.

    doi: 10.1007/s11042-022-12508-9

44. Shankar K, Mohanty SN, Yadav K, Gopalakrishnan T, Elmisery AM. Automated COVID-19 diagnosis and classification using convolutional neural network with fusion based feature extraction model. *Cogn Neurodyn*. 2021;10:1-4.

    doi: 10.1007/s11571-021-09712-y

45. Kadry S, Rajinikanth V, Rho S, *et al*. *Development of a Machine-learning System to Classify Lung CT Scan Images into Normal/COVID-19 Class*. arXiv [Preprint]

46. Wu G, Zhou S, Wang Y, *et al*. A prediction model of outcome of SARS-CoV-2 pneumonia based on laboratory findings. *Sci Rep*. 2020;10(1):14042.

    doi: 10.1038/s41598-020-71114-7

47. Banerjee A, Ray S, Vorselaars B, *et al*. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. *Int Immunopharmacol*. 2020;86:106705.

    doi: 10.1016/j.intimp.2020.106705

48. Moutaz A, Awajan A, Mesleh A, Alhyari S. COVID-19 prediction and detection using deep learning. *Int J Comput Inf Syst Ind Manag Appl*. 2020;12:11-14.

49. Feng C, Wang L, Chen X, *et al*. *A Novel Triage Tool of Artificial Intelligence-assisted Diagnosis Aid System for Suspected COVID-19 Pneumonia in Fever Clinics*. MedRxiv; 2020.

50. Najjar FH, Kadhim KA, Kareem MH, *et al*. Classification of COVID-19 from X-ray images using GLCM features and machine learning. *Malays J Fundam Appl Sci*. 2023;19(6):389-398.

    doi: 10.11113/mjfas.v19n3.2911

51. Maryam A, Ahmad I, Imtiaz A, Mohammed A. Ensemble learning model for diagnosing COVID-19 from routine blood tests. *Inform Med Unlocked*. 2020;21:100449.

    doi: 10.1016/j.imu.2020.100449

52. Atta A, Sultan K, Naseer I, *et al*. Supervised machine learning-based prediction of COVID-19. *Comput Mater Contin*. 2021;69(1):21-34.

    doi: 10.32604/cmc.2021.013453

53. Rajinikanth V, Dey N, Raj AN, *et al*. *Harmony Search and Otsu based System for Coronavirus Disease (COVID-19) Detection Using Lung CT Scan images*. arXiv [Preprint].

54. Tongxue Z, Canu S, Ruan S. *An Automatic COVID-19 CT Segmentation Network Using Spatial and Channel Attention Mechanism*. [arXiv Preprint].

55. Mobiny A, Cicalese PA, Zare S, *et al*. *Radiologist-level COVID-19 Detection Using CT Scans with Detail-oriented Capsule Networks*. [arXiv Preprint].

56. Hasoon JN, Fadel AH, Hameed RS, *et al*. COVID-19 anomaly detection and classification method based on supervised machine learning of chest X-ray image. *Results Phys*. 2021;31:105045.

57. Mahdy LN, Ezzat KA, Elmousalami HH, *et al*. *Automatic X-ray COVID-19 Lung Image Classification System Based on Multi-level Thresholding and Support Vector Machine*. MedRxiv; 2020. p. 2020-2023.

58. Elizabeth DS, Raj CS, Nehemiah HK, Kannan A. A novel segmentation approach for improving diagnostic accuracy of CAD systems for detecting lung cancer from chest computed tomography images. *J Data Inf Qual*. 2012;3:1-16.

59. Rachel RB, Nehemiah HK, Marishanjunath CS, Manoharan RM. Diagnosis of pulmonary edema and COVID-19 from CT slices using squirrel search algorithm, support vector machine and back propagation neural network. *J Intell Fuzzy Syst*. 2023;44:1-4.

    doi: 10.3233/JIFS-222564

60. Rachel RB, Nehemiah HK, Singh VK, Manoharan RM. Diagnosis of COVID-19 from CT slices using whale optimization algorithm, support vector machine and multi-layer perceptron. *J Xray Sci Technol*. 2023;32:253-269.

    doi: 10.3233/XST-230196

61. Anisha I, Nehemiah HK, Anubha I, Kannan A. Computer-Aided Diagnosis system for diagnosis of pulmonary emphysema using bio-inspired algorithms. *Comput Biol Med*. 2020;124:103940.

    doi: 10.1016/j.compbiomed.2020.103940

62. Glover B. *Understanding Flowers and Flowering: An Integrated Approach*. Oxford: Oxford University Press; 2007.

63. Kalra S, Arora S. Firefly Algorithm Hybridized with Flower Pollination Algorithm for Multimodal Functions. In: *Proceedings of the International Congress on Information and Communication Technology*. Germany: Springer; 2016. p. 207-219.

64. Yang XS. Flower Pollination Algorithm for Global Optimization. In: *International Conference on*

Unconventional Computing and Natural Computation. Germany: Springer; 2012. p. 240-249.

65. Pavlyukevich I. Levy flights, non-local search and simulated annealing. *J Comput Phys*. 2007;226(2):1830-1844.

66. Fred AL, Daniel A, Carol JJ. SFCM for efficient brain tumour segmentation. *Int J Adv Eng Technol*. 2019.

67. He B, Zhao W, Pi JY, *et al*. A biomarker basing on radiomics for the prediction of overall survival in non-small cell lung cancer patients. *Respir Res*. 2018;19:199.

doi: 10.1186/s12931-018-0887-8

68. Isaac A, Nehemiah HK, Kannan A. Computer-aided diagnosis system for diagnosis of cavitary and miliary tuberculosis using improved artificial bee colony optimization. *IETE J Res*. 2021;69:1-20.

doi: 10.1080/03772063.2021.1946440

69. Xingyi Y, Xuehai H, Jinyu Z, *et al*. *COVID-CT Dataset: A CT Image Dataset about COVID-19*. [arXiv Preprint].

70. Polsinelli M, Cinque L, Placidi G. A light CNN for detecting COVID-19 from CT scans of the chest. *Pattern Recognit Lett*. 2020;140:95-100.

doi: 10.1016/j.patrec.2020.10.001

71. Ali AE, Assadi TA. GLCMs based multi-inputs 1D CNN deep learning neural network for COVID-19 texture feature extraction and classification. *Karbala Int J Mod Sci*. 2022;8(1):28-39.

doi: 10.33640/2405-609X.3201

72. Pedro S, Luz E, Silva G, *et al*. COVID-19 detection in CT images with deep learning: A voting-based scheme and cross-datasets analysis. *Inform Med Unlocked*. 2020;20:100427.

doi: 10.1016/j.imu.2020.100427

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Deep learning on chest X-ray and computed tomography scans for detection of COVID-19 as a part of a network-centric digital health stack for future pandemics

Ajay Kumar Gogineni[1], Madapathi Hitesh[1], Prashant Kumar Jha[2], Soumya Suvashish Sen[3], Shreeja Das[2], and Kisor Kumar Sahu[2,4,5]*

[1]School of Electrical Sciences, Indian Institute of Technology, Bhubaneswar, Odisha, India

Abstract

Developing a reliable rapid screening protocol for highly infectious diseases like COVID-19 is of paramount interest since it facilitates the isolation of infected patients from the rest of the population. Reverse-transcription polymerase chain reaction (RT-PCR) test is presently the most widely accepted gold-standard test to detect COVID-19. In this method, the RNA of the virus is duplicated by a process called reverse transcription to form DNA for facilitating the copying process. Fluorescent dye is attached to the viral genetic material and copied billions of times through the process called polymerase chain reaction. Enhanced fluorescence is used to identify the presence of genetic material of the virus. These tests are time-consuming and have significant false negatives, *i.e.*, a person with COVID-19 might be categorized as not having the virus. Large-scale RT-PCR testing has its own share of problems such as logistics, availability and affordability in underdeveloped nations, and reliability of the test results. Machine learning algorithms can act as a cheaper supplementary/alternative diagnostic tool for the testing process. In the current study, using publicly available chest X-ray image datasets, different convolutional neural network (CNN)-based models were developed for efficient identification of COVID-19 infected patients, and their efficacies were compared. Key innovations in training the CNNs are discussed. Our results indicate that EfficientNet, SeResNext, and ResNet are best at classifying normal, pneumonia and COVID-19 cases, respectively. The ResNet architecture with transfer learning performed best at detecting COVID-19 with an accuracy of 94%, a rate far superior to that in the RT-PCR test, which is typically in the range of 70 – 80%. This is particularly attractive as an additional noninvasive protocol since such technology-augmented detection is likely to help in reducing the psychological refractory period due to COVID-19 infections. Toward the healthy lung initiative in the post-COVID-19 era, we propose close coupling of the present diagnostic protocols with digital approaches to ensure more reliable personal care within the ambit of large-scale pandemic control mechanisms. Such integration with emerging technological tools can create a benchmark for the first line of defense against future global pandemics.

Keywords: COVID-19; Machine learning; Deep learning; EfficientNet; ResNet; SeResNext; Network-centric digital health stack

***Corresponding author:**
Kisor Kumar Sahu
(kisorsahu@iitbbs.ac.in)

Citation: Gogineni AK, Hitesh M, Jha PK, Sen SS, Das S, Sahu KK. Deep learning on chest X-ray and computed tomography scans for detection of COVID-19 as a part of network-centric digital health stack for future pandemics. *Artif Intell Health*. 2025;2(1):29-41. doi: 10.36922/aih.2888

Received: February 5, 2024

1st revised: April 17, 2024

2nd revised: May 15, 2024

3rd revised: July 3, 2024

Accepted: July 17, 2024

Published Online: October 7, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[2]School of Minerals, Metallurgical and Materials Engineering, Indian Institute of Technology, Bhubaneswar, Odisha, India
[3]Department of Pulmonary Medicine, Srirama Chandra Bhanja Medical College and Hospital, Cuttack, Odisha, India
[4]Centre of Excellence for Novel Energy Materials (CENEMA), Indian Institute of Technology, Bhubaneswar, Odisha, India
[5]Virtual and Augmented Reality Centre of Excellence, Indian Institute of Technology, Bhubaneswar, Odisha, India

1. Introduction

In December 2019, clusters of pneumonia-like cases of unknown origin were first reported in the city of Wuhan in China. Upon investigations, it was found that the disease was caused by a new type of single-stranded RNA virus, which was officially named by the World Health Organization (WHO) as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[1,2] The disease caused by SARS-CoV-2 was subsequently named COVID-19. The rate of spread and severity of COVID-19 all around the world forced the WHO to declare it as a pandemic on March 11, 2020. It had caused unparalleled disruptions in the post-internet modern era of global civilization and, in a way, demonstrated the serious shortcomings and vulnerabilities of traditional approaches for tackling such issues.[2] It was depicted that, although we had the bits and pieces to construct a robust technology-armed first line of defense against global pandemics based upon smart integration of the new age tools, we did not invest enough to construct the requisite technological architecture to realize this. Therefore, various countries attempted to do what they could do best under the prevailing situations. Some implemented nationwide lockdowns, limiting movement of the population, commanding social distancing, and expanding the consciousness of cleanliness and good hygiene as a range of preventive measures against the pandemic; an early analysis of such measures has been reported in the literature.[3]

SARS-CoV-2 can cause symptoms of fever, fatigue, headache, cough, sore throat, myalgia (muscle pain), anosmia (loss of smell), and other respiratory symptoms. Most people recover from the disease without requiring any special treatment but those who have comorbid medical conditions such as chronic kidney disease, autoimmune disease, cancer, heart conditions, obesity, diabetes, and respiratory disease are more prone to develop serious illness.[3] COVID-19 causes serious damage to the lungs, causing oxygen deficiency in the patient, a condition referred to as hypoxia. This can be easily identified by lower oxygen saturation level, typically less than 94%. Infections due to bacterial pneumonia and pulmonary embolism present similar characteristics. Furthermore, other underlying conditions such as chronic obstructive pulmonary disease typically complicate the identification of COVID-19. Atelectasis (indicating a partial or complete collapse of the lungs caused by COVID-19) can be partially reversed by maintaining continuous positive airway pressure, which forces the collapsed alveoli to remain open with oxygen-rich air. A far worse situation demands the use of ventilators, where breathing is assisted with a life-support device. Thus, since it is a common symptom across various respiratory diseases, depleting oxygen saturation cannot be a definitive indicator of COVID-19 infection.

To arrest the spread of the disease, affected countries have adopted reverse-transcription polymerase chain reaction (RT-PCR) tests as the gold-standard diagnostic method to detect COVID-19 infection and isolate the infected individuals as early as possible to limit further transmission of the infection. Despite having many positive aspects, this testing method requires the setting-up of entirely new facilities with specially trained personnel for sample collection and analysis. Such facilities are largely authority-regulated and many underdeveloped countries find it difficult to procure sufficient numbers of test kits. Moreover, it is time-consuming and sometimes gives non-negligible false negative (infected person is tested negative) as well as false positive (uninfected person is tested positive) results.

Machine learning (ML) has made a significant impact in many disciplines of science and technology.[4-11] An ML tool, based on chest X-ray and/or computed tomography (CT) scan images of COVID-19 suspected individuals, can be very attractive in this scenario, primarily because of the very limited resources required and short diagnosis time. Computer-aided biomedical image analysis might turn out to be an additional useful tool to assist medical practitioners in correct and quick decision-making. For example, Habib *et al.*[7] introduced a novel modified lightweight SqueezeNet (SQN-MF) model (demonstrated in non-medical application) coupled with continuous wavelet transformation for converting acoustic emission signals into two-dimensional (2D) images, achieving 100% classification accuracy (surpassing traditional techniques by 20.8%). The lightweight model (0.5 MB) is suitable for field programmable gate array implementation, enabling real-time monitoring. Their success in achieving high accuracy with a memory-efficient model demonstrates the potential for similar approaches in medical diagnostics that will help to build reliable, rapid screening protocols for infectious diseases like COVID-19. This kind of lightweight

computer-based automated tools can be extremely useful, particularly when faced with a logistics bottleneck or when the entire medical infrastructure is overwhelmed for some reason or other, which can be very handy during the outbreak of some other potent infectious diseases.

The method outlined in the present article could be a very important building block in an emerging integrated digital medical doctrine that can ensure a more reliable, personalized, and targeted medical intervention even within the ambit of a very large-scale pandemic control initiative spanning across prefectures/territories/states/countries as previously outlined.[2] One of the foundational principles of modern-era scientific practices is to decouple critical processes to maximize control over them and create provisions for more efficient resource allocation. In the RT-PCR test, the processes of physical examination of the patient by the doctor, sample collection, and actual evaluation of the testing results are highly integrated and codependent. A decoupling approach in this scenario is difficult to achieve. In the present ML-based method, the physical examination of the patient in the form of taking an X-ray radiograph and the evaluation of the testing results from the ML model can be effectively decoupled, thus offering a modular approach that tremendously enhances flexibility in operational and clinical protocols for pandemic control. This effective decoupling and modular approach will be a crucial component for the network-centric digital 'health stack' for the future pandemic control by effectively predicting its geographical occurrences and this will be dealt with in the second article (in writing progress) in this series. Let us illustrate this point further with a simple example for the sake of brevity (the comprehensive architecture will be outlined in the follow-up article): say an economically weaker country "A" lacks enough resource pool of highly trained medical professionals and/or RT-PCR test kits reserves. However, medical X-ray radiography is one of the oldest and most common diagnostic techniques. As a result, most countries (including "A") are already well-equipped with X-ray devices. The approach proposed in this article requires minimal training of personnel who can supervise X-ray radiography tests satisfactorily (that is, just to take a satisfactory radiograph; in the case it is not satisfactory, it can be automatically flagged by a computer/professional located in another country to re-do the test). However, the difficult part of this approach is to train a large number of medical professionals in analyzing/getting familiar with X-ray images for COVID-19 or future pandemic detection. Here, the great advantage of decoupling the critical process becomes clear. By using internet-enabled technologies, the tests and results can then be transmitted online for further evaluation by automated ML algorithms. Only a few confusing cases need to be cross-examined by trained medical professionals in another country, say "B," who can undertake the post-ML decision-making with the help of well-developed resources natively available in "B." It is important to note that the present protocol is performed by an ML algorithm implemented on a computer, and the actual job of a medical expert is minimal, limited to only overviewing/double-checking the assessment of the computer to eliminate the remote possibility of mistakes (since the machine accuracy is more than 94% even without human intervention, as demonstrated in a later section). Such global alliances can be a game changer in providing equitable diagnostics to underdeveloped regions of the world (Global South), as already demonstrated in the case of vaccination under the COVID-19 Vaccines Global Access, (COVAX).[12] Moreover, free exchange of valuable medical data across the boundaries will have the potential to create huge synergistic effects, for example, in faster identification of newer strains, proper epidemiological analysis and consequent prevention strategy, and most importantly, a reliable prediction about the future trends based on hard data obtained through data-driven approach. Developing an automated analysis system can therefore save valuable time for the medical professionals in the country "A," which could be best utilized to address some other critical conditions. This will also ensure optimal resource utilization and advanced preparation for the pandemic in country "B." As pointed out earlier, the detailed digital architecture and infrastructure planning will be presented in the follow-up article. It is important to note that such a design will not be COVID-19-specific but flexible enough to handle a broad spectrum of other diseases.

COVID-19 radiograph (CORAD) scores from CT scans are considered a definite diagnosis of COVID-19 even in the case of negative RT-PCR results. Chest X-ray or CT scan of a COVID-19-infected patient generally reports abnormal findings, such as ground glass opacities, and coarse horizontal linear opacities scattered throughout the lungs, often with consolidation.[13] They represent tiny air sacs getting filled with fluid. Another finding is called the "crazy paving" pattern, which is caused due to swelling of the interstitial septum along the walls of the lung lobes superimposed on the background of ground glass opacities. The latter finding is observed in the advanced stage of infection.[14] Although RT-PCR remains the gold standard procedure for COVID-19 detection, these findings in X-ray images can help in the initial screening of the suspected patients. There are six patterns indicative of COVID-19 infection that can be observed in a patient's X-ray report. They are: (i) reverse batwing, (ii) multifocal lower lobe predominant consolidation,

(iii) peribronchial rounded consolidations, (iv) multifocal bilateral consolidations, (v) ball pattern or round pneumonia, and (vi) bilateral symmetrical diffuse lung involvement. Out of these patterns, pattern (vi) is the most severe, suggesting acute respiratory distress syndrome in the patient.

Several research groups have used deep learning-based techniques for COVID-19 and pneumonia detection.[15-31] Since most of these techniques are well discussed and debated in the literature, we will very briefly review only a few of them. Wang *et al.*[16] used deep learning techniques on CT images to screen COVID-19 patients with an accuracy, specificity, and sensitivity of 89.5%, 88%, and 87%, respectively. A novel convolutional neural network (CNN), COVID-Net for detecting COVID-19 using chest X-ray images presented by a different research team has an accuracy of 83.5%.[15] Transfer learning offers several benefits, primarily in saving training time, significantly improving the performance of ML models, and requiring smaller data sets for training. For example, Sohaib *et al.*[8] proposed a novel approach (demonstrated in non-medical applications) that used step transfer learning (STL) combined with extreme learning machine (ELM), primarily focusing on improving the generalization power of deep learning models for autonomous inspection. By leveraging STL to extract generalized abstract features from diverse source images and utilizing ELM to overcome optimization limitations of traditional neural networks, their model achieved significant improvements in accuracy (2.5%), recall (4.8%), and precision (0.8%) compared to existing studies. This approach enhanced generalization demonstrating the usefulness of transfer learning techniques for increasing the robustness of the detection models and making the model more generalizable. Transfer learning has found many applications in the biomedical arena.[32] Joaquin[17] used a small dataset of 339 images for training and testing by utilizing the ResNet50-based deep transfer learning technique and obtained a validation accuracy of 96.2%. Ahmmed *et al.*[32] conducted an in-depth analysis of brain tumor classification using transfer learning across multiple classes, utilizing robust frameworks, such as ResNet 50 and Inception V3 for MRI images. Their research meticulously curated paired datasets and incorporated advanced techniques such as Early Stopping, ReduceLROnPlateau, and hyperparameter optimization. These strategies significantly improved model accuracy, achieving exceptional classification rates for various types of brain tumors. Similarly, Podder *et al.*[31] developed a deep learning model using optimized DenseNet architectures to diagnose infectious diseases from chest X-rays, achieving high detection rates for COVID-19 and other conditions. Their modifications to the DenseNet architecture and hyperparameter tuning demonstrated the potential of deep learning models in improving diagnostic accuracy and early disease detection.

In this study, we developed deep transfer learning-based approaches for the automatic detection of COVID-19 using various CNN-based models on chest X-ray images and implemented several innovations in the training protocol. Beyond the immediate applicability of the present method as an additional diagnostic tool for identifying COVID-19 infections, it also might have far-reaching consequences. An earlier big challenge was to identify and detect the infected individual; however, slowly but steadily, the major focus has shifted toward the long-term care of the lungs of the infected persons. All seven human coronaviruses affect the respiratory tracts and lungs. However, three of these virus types, *i.e.*, SARS-CoV, Middle East respiratory syndrome coronavirus, and SARS-CoV-2 are known to severely affect the lungs. Unfortunately, the long-term impact of SARS-CoV-2 infection is neither clearly understood nor widely studied as the viruses are dynamically evolving across the world. Most severe infections are known to cause post-COVID-19 sequelae, *i.e.*, fibrosis of the lungs in the future. A worthwhile initiative toward recording post-COVID-19 symptoms in this direction has been reported in the literature.[33] An ML-based system that relies on the chest X-ray and/or CT scan image analysis might prove to be highly valuable and instrumental in the long-term care of the lungs by appropriately creating a highly scalable and easily integrable digital infrastructure through curating, cataloging, classifying and most importantly, creating an appropriate data repository of the vital information about different stages of the lungs as a function of time (using automated time-stamps), significantly augmenting the healthy lung initiative in the post-COVID-19 era.

2. Data and methods

2.1. Data

The dataset of chest X-ray and CT scan images[34,35] used in this study were sourced from a freely available GitHub repository maintained by Cohen.[36] X-ray images containing pneumonia and normal images were obtained from the Kaggle dataset.[37] We used 1763 images to analyze the performance of the neural networks used for this study and utilized 1260 images to train the model along with 251 and 252 images for validating and testing the model, respectively. The dataset consists of 563, 947, and 223 images that belong to COVID-19, normal, and pneumonia categories, respectively. Figure 1 shows some examples from the dataset.

2.2. Methods: Different CNN architectures

Various CNN models such as ResNet, SeResNext, DenseNet, and EfficientNet were used for classification. Here we briefly

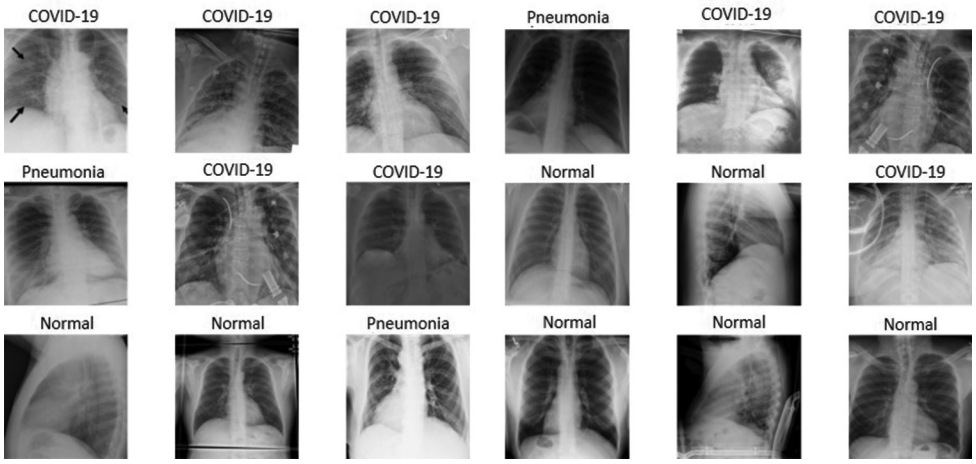


Figure 1. Sample images are taken from the dataset.[34,35,37] The indication above each image corresponds to the associated label. Image created by the author.

sketch their outlines, although complete details can be found in the reference mentioned in the corresponding sections. Fundamental building blocks are schematically depicted in Figure 2A-D.

2.2.1. Method 1: ResNet34

Developing a proper training protocol is a matter of serious concern for the implementation of any deep neural network. One such major issue is the divergence from local minima, leading to improper training. To address this, in ResNet,[38] a modification in the network architecture was introduced by incorporating skip connections, expressed as H(x) + x, between layers. This alteration facilitated quicker and more efficient model training. The smooth loss landscape of ResNet prevents the model from becoming trapped in local minima or saddle points, resulting in improved training speed and accuracy. In our study, we utilized a variant of ResNet, specifically ResNet34, consisting of a total of 34 convolutional layers.

2.2.2. Method 2: SeResNext50

SeNet was originally proposed by Hu *et al.*[39] SeNet differs from conventional neural network designs by emphasizing the exploration of channel-wise features rather than solely focusing on spatial features. In its fundamental structure, the squeeze-and-excitation (SE) block transforms the input, denoted as x, into a feature map U through convolution. This map undergoes a squeeze operation, consolidating feature maps across spatial dimensions to produce channel descriptors. These descriptors encapsulate the global distribution of channel-wise feature responses. Following this, an excitation operation converts the descriptors into per-channel modulation weights. These weights are then applied to the feature map U to yield the output of the SE

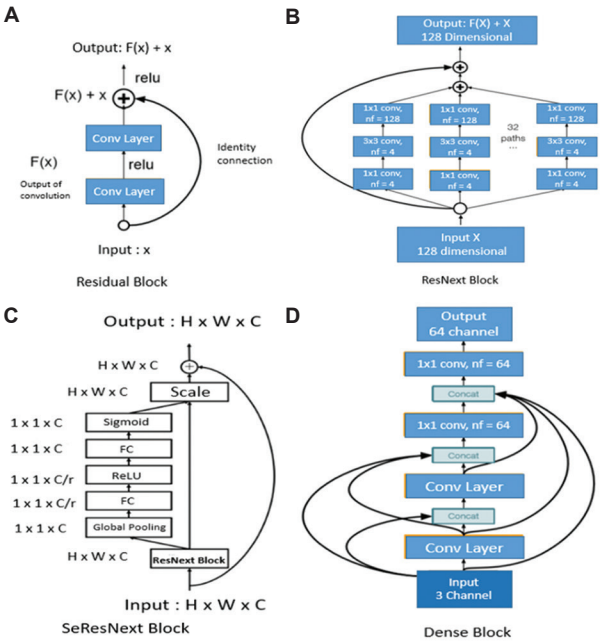


Figure 2. Building blocks of different convolutional neural network-based architectures: (A) ResNet, (B) ResNext, (C) SeResNext, and (D) DenseNet. Copyright © 2020 Springer International Publishing. Reprinted with permission of Springer International Publishing.

block. In the SeResNext model,[40,41] SE block is integrated into every non-identity branch of the ResNext block—a variant of the ResNet block characterized by multiple convolution layers and skip connections. A ResNext block consists of several convolution layers, each having a distinct set of filter sizes and dimensions, and incorporates a single skip connection for mitigating the vanishing gradient problem during training. A SeResNext block is shown in Figure 2C.

2.2.3. Method 3: DenseNet

DenseNet architecture was proposed by Xie *et al.*[40] Its architecture is very similar to that of ResNet, where there are feed-forward connections from each layer to the next layer.[38] In DenseNet, feature maps of one layer are concatenated with feature maps of all the following layers. This approach offers a benefit by leveraging features extracted from early layers for subsequent layers. Convolution blocks are sequentially stacked, and interspersed with basic convolution layers to preserve dimensionality across the network's depth. It consists of various "dense blocks." A simple "dense block" is depicted in Figure 2D.

2.2.4. Method 4: EfficientNet

Most of the architectures such as ResNet, VGG Net, and Inception Net, are created manually by researchers where they specify the complete network architecture upfront, for example, number of layers, filter size, and number of channels based on previous experiments/experience. EfficientNets[42] are created using neural architecture search where the complete model is built algorithmically by keeping a constraint on the number of parameters. It uses ResNet as a baseline model and modifies the number of layers, number of channels, and input image dimensions in the baseline model to create the desired model. The smallest model, *i.e.*, with a minimum number of parameters in EfficientNet is called b0, and seven other models are generated by changing the constraints, such as the number of parameters and the number of Floating Point Operations Per Second. EfficientNet b7 is the largest model among EfficientNets. EfficientNets have very less inference time when compared to other models with a similar number of parameters. As the image size increases, larger EfficientNet models are preferred since they have a greater number of layers and the channel size also increases. This helps in obtaining useful features from the larger image.

Overall, the rational for choosing this architecture is as follows: ResNet was selected for its skip connection architecture, which facilitates stable learning during training; DenseNet for its dense connectivity facilitating feature reuse across layers; SeResNext for its integration of SE modules for enhanced feature recalibration; and EfficientNet for its efficient model scaling strategy, which collectively provides a diverse range of architectural innovations for achieving accurate and reliable image classification. It should be noted that it is always possible (and sometimes more desirable) to make an ensemble of these models for further improving the overall prediction. Since the present article focuses on the fundamental aspects of the implementation of these models, the ensemble strategy is not explored in this work.

Transfer learning uses the initial weights of the neural network pre-trained on a different database, which is ImageNet[43] for the present study. Transfer learning plays a crucial role in improving the model performance when working on a dataset where the number of images are limited. The amount of labeled data available in the biomedical domain is limited mainly due to the time taken to annotate the dataset. Initializing the model weights in this manner helps the model to capture important information from the images. The initial layers of the network capture very generic information from an image such as horizontal and vertical edges, whereas the later layers of the network capture patterns in an image that are very specific to the dataset of study as previously described.[41] This pre-training on a large dataset provides a solid foundation, allowing the model to start with a better understanding of general visual patterns. The fine-tuning process adapts the model to the unique features and characteristics of the target dataset, allowing it to specialize in recognizing patterns relevant to the biomedical images at hand.

2.3. Training

We employed a learning rate scheduler[44] to determine the most effective learning rate for our specific dataset. During this process, the learning rate is cautiously increased after each mini-batch, with the corresponding loss recorded at each increment. Subsequently, we plotted the loss versus learning rate, as illustrated in Figure 3, revealing how different learning rates impacted the model's performance. Notably, for a very low learning rate, the loss diminishes at a slower pace. As the learning rate increased, the loss showed a rapid decline, indicating the optimal range. Beyond this point, further increases in the learning rate caused the loss to rise sharply, suggesting overshooting. By identifying the point of the steepest loss decline (0.002 in our case), we determined the optimal learning rate for

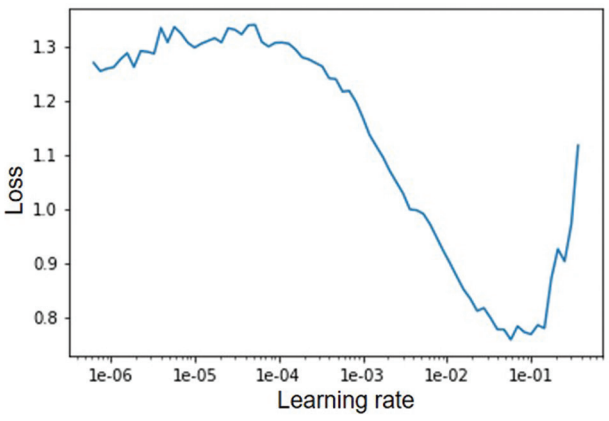


Figure 3. Scheduling the learning rate by investigating its impact on the loss function. Image created by the authors.

training. It is crucial to avoid exceeding this optimal value, as it may lead to overshooting the global minima of the loss function. This balanced approach ensures efficient training and better model generalization by avoiding both slow convergence and the risk of overshooting the global minima.

Adjusting the learning rate plays a critical role in training neural networks. A smaller value of the learning rate leads to gradual changes in the loss function but can prolong convergence due to small gradients. Conversely, a larger value of the learning rate may cause overshooting of the global minimum. Hence, striking a balance between the two extremes is essential for efficient training and better generalization. Furthermore, employing a fixed learning rate can lead to challenges such as becoming trapped in local minima or saddle points where gradients are insufficient for optimization. To mitigate these issues, we set upper and lower bounds on the learning rate, typically differing by a factor of ten. For instance, in our experiments, the upper bound is set at 10^{-4} and the lower bound at 10^{-5}. This range allows the model to explore the loss landscape effectively, avoiding stagnation in local minima or saddle points. In addition, in lieu of employing a uniform learning rate throughout the entire network during the training process, this study adopts a strategy known as discriminative learning rates. Here, learning rates are tailored for different layers of the classifier, typically ranging between 0.0001 and 0.01. This approach acknowledges that various network layers capture distinct types of information, thus warranting diverse learning rates. Initial layers receive lower learning rates compared to later layers, reflecting their differing roles in feature extraction and abstraction.

The training methodology incorporates a one-cycle training policy, characterized by a dynamic adjustment of learning rates across epochs. Initially, a higher learning rate is applied, gradually decreasing toward the final epoch. This technique promotes improved model performance and stability, facilitating parameter updates at an appropriate pace. By mitigating the risk of local minima entrapment, the model's generalizability is enhanced.

Implementation details involve Python programming utilizing the fastai library,[45] a PyTorch-based open-source platform tailored for deep learning model development. Execution takes place on Google Collaboratory, leveraging its provision of a free K-80 GPU with 12GB RAM, ideal for executing ML algorithms. The source code and neural network weights utilized in this study are publicly available for reference.[46] In the following, we summarize some of the key novelties in the training strategy in the present study. We used training strategies to optimize the performance of EfficientNet, ResNet, and SeResNext for COVID-19 detection using X-ray images. Key innovations include the use of a dynamic learning rate scheduler to identify the optimal learning rate, setting adaptive boundaries (10^{-4} – 10^{-5}) to avoid local minima, and employing discriminative learning rates tailored for different network layers (ranging from 0.0001 to 0.01) to enhance feature extraction and abstraction. In addition, we implemented a one-cycle training policy, dynamically adjusting learning rates across epochs to improve model performance and stability. These training methodologies significantly enhance the model's robustness and accuracy, providing a distinctive contribution to the field. Altogether, our proposed model introduces several key innovations that enhance the robustness and accuracy of these models, making our work distinct from prior studies.

3. Results

A total of 1763 images were used to build the ML model, of which 1260 images were used to train the model and 251 and 252 images were employed for validating and testing the model, respectively. After the screening, the 563 images indicating COVID-19 were split into train (450 images), valid (71 images), and test datasets (72 images).

Figure 4A-D shows the confusion matrix for various CNN architectures. ResNet and DenseNet obtained the best accuracy, at 94.09%. The confusion matrix corresponds to the predictions of the model on the test dataset. The model is able to distinguish clearly among various classes. The confusion matrices indicate that EfficientNet is best at classifying normal images and SeResNext is best at classifying pneumonia. ResNet performs best for classifying images pertaining to COVID-19.

Figure 5 shows the predicted class, actual class, loss, and probability of actual class for a set of misclassified images for each model in the format "Prediction/Actual/Loss/Probability" on the top of each image. Each image illustrates the target class results for the model, highlighting areas where the model's predictions did not align with the true labels.

Figure 6 shows the sensitivity as well as specificity of the CNN models used in this work. The sensitivity of a class measures the proportion of images belonging to a particular class that are correctly classified by the model. The specificity of a class measures the proportion of images that do not belong to the class of interest and are correctly classified by the model.

4. Discussion

EfficientNet has the highest specificity toward images of class COVID-19 and ResNet has the highest sensitivity

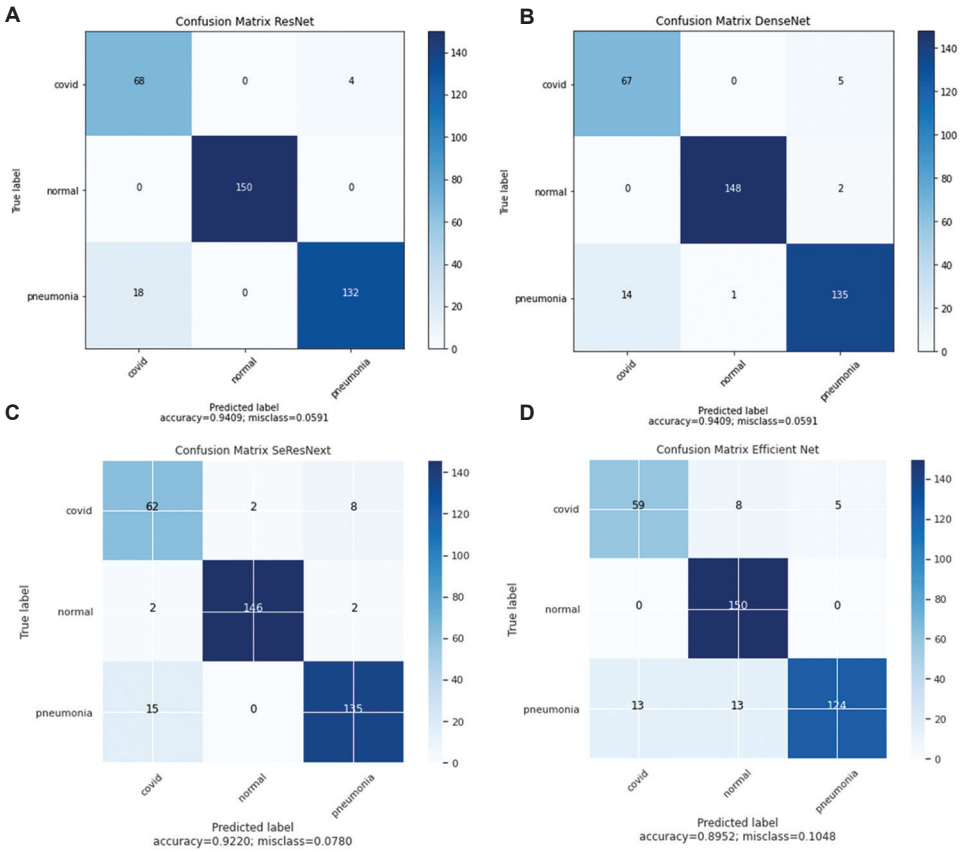


Figure 4. Confusion matrices for (A) ResNet (34), (B) DenseNet, (C) SeResNext, and (D) EfficientNet
Note: The number 34 indicates the number of convolution layers).

toward images of class COVID-19. The sensitivity and specificity values in Figure 6A and B, respectively, were computed by evaluating the performance of each CNN model (ResNet, DenseNet, SeResNext, EfficientNet) on a labeled test set consisting of three classes: COVID-19, pneumonia, and normal X-ray images. Sensitivity (true positive rate) was calculated as the ratio of correctly identified positive cases to the total actual positive cases for each class. Specificity (true negative rate) was determined by the ratio of correctly identified negative cases to the total actual negative cases for each class. For each model, predictions were compared against the ground truth labels to calculate these metrics, providing a comprehensive assessment of each model's ability to correctly identify and distinguish between the three classes.

RT-PCR's sensitivity and specificity are typically in the range of 70 – 80% and 99 – 100%, respectively. Therefore, the 94% sensitivity achieved in the present study with very limited numbers of training images is an indication of good performance, which is expected to get better with more training images. Therefore, our method is significantly more accurate. Subsequently, doctors may be consulted for a confirmed diagnosis. Some examples of COVID-19 scans are described in the literature.[47] In addition to the tremendous promise, the efficacy of the present method can be significantly enhanced, if it is supplemented with blood lymphocyte count (as lymphopenia—a lower count of lymphocytes—is mostly associated with COVID-19 and indicates severe form) and RT-PCR test data from nasopharyngeal samples collected through swabs. The efficacy of the present method will significantly improve, even when not assisted by other methods, with more and more usage (as the method progressively learns and gets better at the job), as is the case for any ML method.

It is imperative to mention that collecting data from diverse geographical regions can significantly improve the performance of our model by introducing a wider variety of image characteristics and potential variations in COVID-19 presentation. Different regions may have variations in imaging equipment, patient demographics, and prevalence of comorbidities, all of which can influence the appearance of X-ray images. By incorporating a more diverse dataset, our model can learn to generalize better across different populations and imaging conditions,

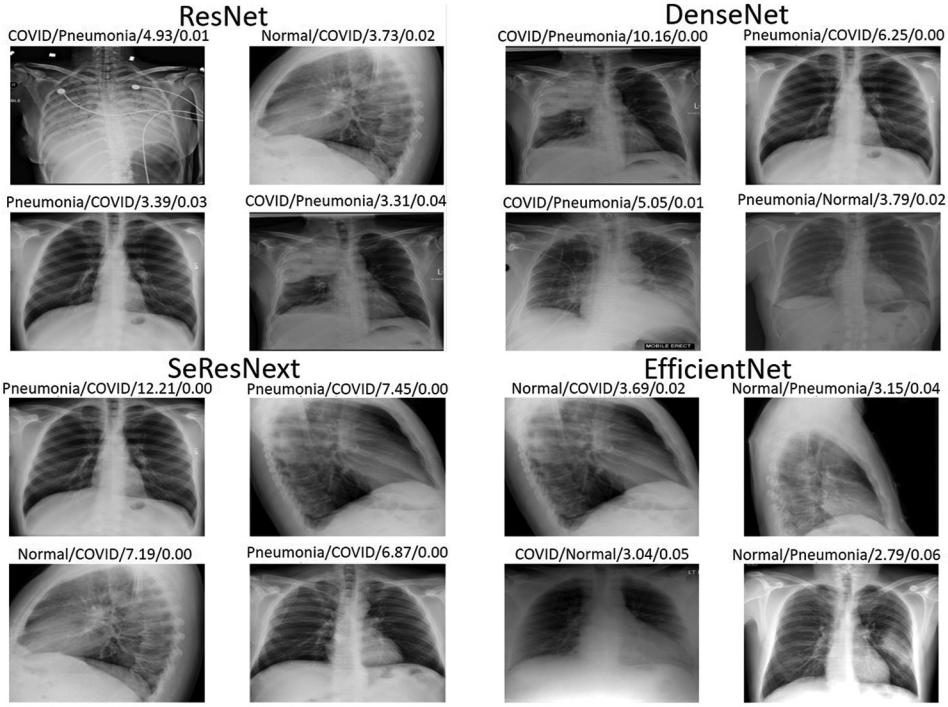


Figure 5. Misclassified X-ray images with prediction details using ResNet, DenseNet, SeResNext, and EfficientNet models. On the top of each subplot, "Prediction/Actual/Loss/Probability" details for each individual image are shown.

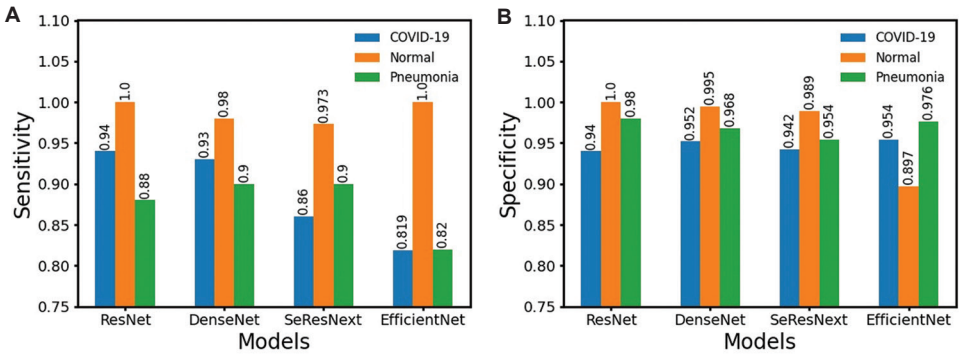


Figure 6. Sensitivity (A) and specificity (B) of ResNet (34), DenseNet, SeResNext, and EfficientNet for COVID-19, normal, and pneumonia class detection
Note: The number 34 indicates the number of convolution layers).

leading to more robust and accurate predictions. This enhanced diversity helps mitigate biases and ensures that the model performs well in real-world, heterogeneous environments, ultimately improving its reliability and effectiveness in detecting COVID-19, pneumonia, and normal cases at the global scale and for future pandemics, as well.

5. Conclusion

In this work, we implemented various CNN models with a transfer learning-based approach to classify COVID-19 and pneumonia from normal patients through chest X-rays. This research article introduces several key innovations in the training methodology that distinguish it from prior works:

(1) *Dynamic learning rate optimization.* We developed a novel dynamic learning rate scheduler that adaptively identifies the optimal learning rate within carefully set boundaries (10^{-4} – 10^{-5}). This approach mitigates the risk of getting trapped in local minima, a common challenge in neural network training.

(2) *Layer-specific discriminative learning.* This study implemented a discriminative learning rate strategy, applying different learning rates (ranging from 0.0001

to 0.01) to various network layers. This nuanced approach enhances both the low-level feature extraction and high-level abstraction capabilities of the models.

(3) *One-cycle policy implementation.* We integrated a one-cycle training policy that dynamically adjusts learning rates across epochs. This method has been shown to significantly improve model convergence, stability, and overall performance.

(4) *Comparative analysis of advanced architectures.* While individual CNN architectures have previously been applied to COVID-19 detection, our study provides a comprehensive comparison of EfficientNet, ResNet, and SeResNext using these advanced training strategies. This comparison offers valuable insights into the relative strengths of these architectures for this specific task.

(5) *Reproducibility and benchmarking.* By using a publicly available dataset and clearly documenting our methodologies, we provide a robust benchmark for future studies in medical image analysis, which extends beyond the scope of COVID-19 detection.

These methodological innovations collectively enhance the robustness, accuracy, and generalizability of CNN models for medical image analysis. While the immediate application to COVID-19 may seem less pressing now, the techniques we developed have broader implications for improving deep learning approaches in medical imaging across various conditions that might be very useful in our fight against future pandemics.

Among the models implemented in the present study, ResNet and DenseNet have achieved more than 94% accuracy. This is far superior to the typical sensitivity of 70 – 80% for RT-PCR. Our results indicate that EfficientNet is best at classifying normal images, and SeResNext is best at classifying pneumonia. ResNet performs best for classifying images pertaining to COVID-19. While the accuracy of the present method is expected to get better with increasing usage, which is an inherent feature of artificial intelligence, there is no such chance for RT-PCR, since this traditional method is not a smart protocol. The model is able to learn the inherent features of pneumonia and COVID-19 from a relatively small dataset. The performance of the model can be improved further by collecting data from diverse geographical regions. This will also improve the generalizability of the model.

We strongly believe that this ML-aided diagnostic protocol can help in detecting individuals suspected of carrying infections with greater speed and accuracy, and more importantly, it charts out the blueprint to rapidly develop a new med-tech protocol for quick screening of future pandemics. It is pertinent to point out here that the decoupling of physical examination of the patient and analytical pathology leads to an effective and modular approach. This is likely to significantly enhance detection speed, accuracy, and sensitivity, expected to form the fundamental cornerstone that will be pivotal for an extensive digital architecture to safeguard against many future pandemics (to be elaborated in the follow-up article). Furthermore, these models help in the early screening of suspects in remote places in countries where the health care providers as well as resources (such as RT-PCR kits and CT scan machines) are limited.

Acknowledgments

The authors acknowledge the sincere help from the authorities of VARCoE, IIT BBS, and SCBMCH for the research support and encouragement.

Funding

None.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: Ajay Kumar Gogineni, Kisor Kumar Sahu
Formal analysis: All authors
Investigation: Ajay Kumar Gogineni, Madapathi Hitesh
Methodology: Ajay Kumar Gogineni, Kisor Kumar Sahu
Writing – original draft: All authors
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data are available at the following resources:

(1) Cohen JP, Morrison P, Dao L. COVID-19 Image Data Collection. arXiv.org. Accessed April 9, 2021. https://doi.org/10.48550/arXiv.2003.11597

(2) Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P. COVID-CT-Dataset: A CT Scan Dataset about COVID-19. *arXiv:200313865 [cs, eess, stat]*. Published online June 17, 2020. http://arxiv.org/abs/2003.13865

(3) Cohen JP. ieee8023/covid-chestxray-dataset. GitHub. Published June 10, 2020. https://github.com/ieee8023/covid-chestxray-dataset

(4) Chest X-Ray Images (Pneumonia). www.kaggle.

com. Accessed April 9, 2021. https://kaggle.com/paultimothymooney/chest-xray-pneumonia

(5) Gogineni A. AjayKumarGogineni777/covid_cnn. GitHub. Published October 16, 2020. Accessed April 10, 2021. https://github.com/AjayKumarGogineni777/covid_cnn

References

1. World Health Organization. *Determinants of Health*; 2023. Available from: https://www.who.int [Last accessed on 2024 Sep 16].

2. Pathak AD, Saran D, Mishra S, Hitesh M, Bathula S, Sahu KK. Smart war on COVID-19 and global pandemics: Integrated AI and blockchain ecosystem. Panigrahi CR, Pati B, Rath M, Buyya R, editors. *Computational Modeling and Data Analysis in COVID-19 Research*. United States: CRC Press; 2021.

   doi: 10.1201/9781003137481-5

3. Kishore R, Jha PK, Das S, Agarwal D, Maloo T, Pegu H, *et al*. A Kinetic Model for Qualitative Understanding and Analysis of the Effect of Complete Lockdown Imposed by India for Controlling the COVID-19 disease spread by the SARS-CoV-2 virus.

   doi: 10.48550/arXiv.2004.05684

4. Varadi M, Anyango S, Deshpande M, *et al*. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high accuracy models. *Nucleic Acids Res*. 2021;50(D1):D439-D444.

   doi: 10.1093/nar/gkab1061

5. Chowdhery A, Narang S, Devlin J, *et al*. *PaLM: Scaling Language Modeling with Pathways*. arXiv:220402311; 2022. Available from: https://arxiv.org/abs/2204.02311 [Last accessed on 2023 Mar 05].

6. Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S. *A ConvNet for the 2020s*. arXiv:220103545; 2022. Available from: https://arxiv.org/abs/2201.03545 [Last accessed on 2023 Mar 05].

7. Habib A, Hasan J, Kim J. A Lightweight deep learning-based approach for concrete crack characterization using acoustic emission signals. *IEEE Access*. 2021;9:104029-50.

   doi: 10.1109/access.2021.3099124

8. Sohaib M, Hasan MJ, Chen J, Zheng Z. Generalizing infrastructure inspection: Step transfer learning aided extreme learning machine for automated crack detection in concrete structures. *Meas Sci Technol*. 2024l;35:055402.

   doi: 10.1088/1361-6501/ad296c

9. Liu Z, Lin Y, Cao Y, *et al*. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*. arXiv:210314030; 2021. Available from: https://arxiv.org/abs/2103.14030 [Last accessed on 2023 Mar 05].

10. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is Worth 16x16 Words: Transformers for Image Recognition at Scale; 2020.

    doi: 10.48550/arXiv.2010.11929

11. Radford A, Kim JW, Hallacy C, *et al*. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv:210300020; 2021. Available from: https://arxiv.org/abs/2103.00020 [Last accessed on 2023 Mar 05].

12. WHO. *COVAX: Working for Global Equitable Access to COVID-19 Vaccines*. World Health Organization; 2020. Available from: https://www.who.int/initiatives/act-accelerator/covax [Last accessed on 2021 Apr 09].

13. Cleverley J, Piper J, Jones MM. The role of chest radiography in confirming covid-19 pneumonia. *BMJ*. 2020;370:m2426.

    doi: 10.1136/bmj.m2426

14. Zu ZY, Jiang MD, Xu PP, *et al*. Coronavirus disease 2019 (COVID-19): A perspective from China. *Radiology*. 2020;296(2):E15-E25.

    doi: 10.1148/radiol.2020200490

15. Wang L, Lin ZQ, Wong A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep*. 2020;10(1):19549.

    doi: 10.1038/s41598-020-76550-z

16. Wang S, Kang B, Ma J, *et al*. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *Eur Radiol*. 2021;31:6096-6104.

    doi: 10.1007/s00330-021-07715-1

17. Joaquin AS. *Using Deep Learning to Detect NCOV-19 from X-Ray Images*. Medium; 2020. Available from: https://towardsdatascience.com/using-deep-learning-to-detect-ncov-19-from-x-ray-images-1a89701d1acd [Last accessed 2020 Jun 27].

18. Ismael AM, Şengür A. Deep learning approaches for COVID-19 detection based on chest X-ray images. *Expert Syst Appl*. 2021;164:114054.

    doi: 10.1016/j.eswa.2020.114054

19. Zhang J, Xie Y, Pang G, *et al*. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. *IEEE Trans Med Imaging*. 2021;40(3):879-890.

    doi: 10.1109/tmi.2020.3040950

20. Hemdan EE, Shouman MA, Mohamed Esmail Karar. *COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images*. *arXiv*. Ithaca: Cornell University; 2020.

    doi: 10.48550/arxiv.2003.11055

21. Jain R, Gupta M, Taneja S, Hemanth DJ. Deep learning based detection and analysis of COVID-19 on chest X-ray

images. *Appl Intell*. 2020;51:1690-1700.

doi: 10.1007/s10489-020-01902-1

22. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput Biol Med*. 2020;121:103792.

doi: 10.1016/j.compbiomed.2020.103792

23. Apostolopoulos ID, Aznaouridis SI, Tzani MA. Extracting possibly representative COVID-19 biomarkers from X-ray images with deep learning approach and image data related to pulmonary diseases. *J Med Biol Eng*. 2020;40(3):462-469.

doi: 10.1007/s40846-020-00529-4

24. Rahaman MM, Li C, Yao Y, *et al*. Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. *J Xray Sci Technol*. 2020;28(5):821-839.

doi: 10.3233/xst-200715

25. Ouchicha C, Ammor O, Meknassi M. CVDNet: A novel deep learning architecture for detection of coronavirus (Covid-19) from chest x-ray images. *Chaos Solitons Fractals*. 2020;140:110245.

doi: 10.1016/j.chaos.2020.110245

26. Ayan E, Ünver HM. Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning. Istanbul, Turkey: Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT); 2019. p. 1-5.

doi: 10.1109/EBBT.2019.8741582

27. Asnaoui, Khalid El, Chawki Y, Idri A. *Automated Methods for Detection and Classification Pneumonia based on X-Ray Images Using Deep Learning*. arXiv. Ithaca: Cornell University; 2020.

doi: 10.48550/arxiv.2003.14363

28. Elshennawy NM, Ibrahim DM. Deep-pneumonia framework using deep learning models based on chest X-ray images. *Diagnostics*. 2020;10(9):649.

doi: 10.3390/diagnostics10090649

29. Ibrahim AU, Ozsoz M, Serte S, Al-Turjman F, Yakoi PS. Pneumonia classification using deep learning from chest X-ray images during COVID-19. *Cogn Comput*. 2021;16:1589-1601.

doi: 10.1007/s12559-020-09787-5

30. Kundu R, Das R, Geem ZW, Han GT, Sarkar R. Pneumonia detection in chest X-ray images using an ensemble of deep learning models. *PLoS One*. 2021;16(9):e0256630.

doi: 10.1371/journal.pone.0256630

31. Podder P, Alam FB, Mondal MR, Hasan MJ, Rohan A, Bharati S. Rethinking densely connected convolutional networks for diagnosing infectious diseases. *Computers*. 2023;12(5):95.

doi: 10.3390/computers12050095

32. Ahmmed S, Podder P, Mondal M, *et al*. Enhancing brain tumor classification with transfer learning across multiple classes: An in-depth analysis. *BioMedInformatics*. 2023;3(4):1124-1144.

doi: 10.3390/biomedinformatics3040068

33. Kostka K, Roel E, Trinh NT, *et al*. The burden of post-acute COVID-19 symptoms in a multinational network cohort analysis. *Nat Commun*. 2023;14:7449.

doi: 10.1038/s41467-023-42726-0

34. Cohen JP, Morrison P, Dao L. COVID-19 Image Data Collection; 2020.

doi: 10.48550/arXiv.2003.11597

35. Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P. *COVID-CT-Dataset: A CT Scan Dataset about COVID-19*; 2020. Available from: http://arxiv.org/abs/2003.13865 [Last accessed on 2021 Apr 09].

36. Cohen JP. *GitHub*; 2020. Available from: https://github.com/ieee8023/covid-chestxray-dataset [Last accessed on 2021 Apr 09].

37. Chest X-Ray Images (Pneumonia). Available from: https://kaggle.com/paultimothymooney/chest-xray-pneumonia [Last accessed on Apr 09].

38. He K, Zhang X, Ren S, Sun J. *Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016. p. 770-778.

doi: 10.1109/cvpr.2016.90

39. Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-Excitation Networks; 2017.

doi: 10.48550/arxiv.1709.01507

40. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep Neural Networks. Honolulu, HI, USA: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p. 5987-5995.

doi: 10.1109/CVPR.2017.634

41. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps; 2013.

doi: 10.48550/arXiv.1312.6034

42. Tan M, Le QV. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks; 2019.

doi: 10.48550/arXiv.1905.11946

43. Simon M, Rodner E, Denzler J. ImageNet Pre-trained Models with Batch Normalization; 2016.

doi: 10.48550/arXiv.1612.01452

44. Smith LN. *Cyclical Learning Rates for Training Neural Networks*. Piscataway: IEEE Xplore.

doi: 10.1109/WACV.2017.58

45. Fastai-Text Learner. *Fastai*. Available from: https://docs.fast.ai/text.learner.html [Last accessed on 2021 Apr 09].

46. Gogineni A. *AjayKumarGogineni777/covid_cnn. GitHub*; 2020. Available from: https://github.com/ajaykumargogineni777/covid_cnn [Last accessed on 2021 Apr 10].

47. Yasin R, Gouda W. Chest X-ray findings monitoring COVID-19 disease course and severity. *Egypt J Radiol Nucl Med*. 2020;51(1):193.

doi: 10.1186/s43055-020-00296-x

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Enhancing spinal MRI segmentation with an asymmetric U-Net architecture

Longfei Zhou[1†*], Xingyu Chen[2†], Weihao Cheng[3], Zhanghao Qin[2], Tianao Shen[4], Pingyu Cao[5], Zebo Huang[2], Xiangyu Wu[6], and Yiyao Zhang[7]

[1]Department of Biomedical, Industrial and Systems Engineering, College of Engineering and Business, Gannon University, Erie, Philadelphia, United States of America

[2]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, Jiangsu, China

[3]School of Physics and Optoelectronic Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

[4]School of Internet of Things Engineering, Jiangnan University, Wuxi, Jiangsu, China

[5]School of Remote Sensing and Surveying Engineering, Nanjing University of Information Science and Technology, Nanjing, Jiangsu, China

[6]School of Mechanical Engineering, Jiangnan University, Wuxi, Jiangsu, China

[7]College of Robot and Engineering, Guangzhou City University of Technology, Yinchuan, Ningxia, China

[†]*These authors contributed equally to this work.*

***Corresponding author:**
Longfei Zhou
(zhou009@gannon.edu)

Citation: Zhou L, Chen X, Cheng W, *et al*. Enhancing spinal MRI segmentation with an asymmetric U-Net architecture. *Artif Intell Health*. 2025;2(1):42-52. doi: 10.36922/aih.3889

Received: June 7, 2024

1st revised: July 5, 2024

2nd revised: July 12, 2024

Accepted: August 1, 2024

Published Online: October 21, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Spinal diseases are among the most prevalent health issues in modern society, significantly impacting patients' quality of life. Diagnosing conditions such as disc herniation and spinal deformity requires advanced medical imaging techniques, including X-rays, magnetic resonance imaging (MRI), computed tomography, and nuclear magnetic resonance. Spine MRI is particularly crucial due to its ability to provide high-resolution images of soft tissues, essential for accurate diagnosis. However, the manual segmentation of spine MRI images is labor-intensive and inadequate for large-scale quantitative analysis. Thus, developing automated spinal MRI segmentation methods is critical to alleviating doctors' workload and enhancing diagnostic efficiency. In this study, we propose a novel asymmetric U-Net architecture designed to improve the precision of reconstructing complex structures and details by increasing the depth of the upsampling side. The model incorporates adjacent-scale skip connections to control parameters while maintaining high segmentation accuracy. In addition, residual connections on the upsampling side prevent gradient vanishing, thereby enhancing the network's feature learning and representation capabilities. Experimental results indicate that this method significantly reduces training time and increases model accuracy compared to traditional approaches, marking a substantial advancement in automated spinal MRI segmentation. This innovative approach holds promise for improving clinical outcomes and optimizing the workflow in medical imaging departments.

Keywords: Spinal magnetic resonance imaging; Automated segmentation; Asymmetric U-Net; Medical imaging; Deep learning

1. Introduction

Spinal diseases are among the most prevalent health issues in modern society. The pathogenesis of spine diagnostics involves understanding the underlying mechanisms and origins of spinal disorders, which is crucial for accurate diagnosis and effective treatment. Modern spine diagnostics has evolved significantly with advancements in imaging technology, genetic research, and molecular biology, providing deeper insights into spinal pathologies. Pathogenesis in spine diagnostics refers to the study of how spinal diseases develop and progress. This includes degenerative diseases such as osteoarthritis and intervertebral disc degeneration, as well as inflammatory conditions such as ankylosing spondylitis.[1] Degenerative spinal disorders often involve the breakdown of intervertebral discs and facet joints. Factors such as aging, mechanical stress, and genetic predisposition contribute to the degeneration process. Studies have shown that mechanical loading and biochemical changes play significant roles in disc degeneration.[2] Diagnostic tools such as magnetic resonance imaging (MRI) and computed tomography (CT) scans allow for detailed visualization of these changes.[3] Inflammatory spinal diseases, such as ankylosing spondylitis, involve chronic inflammation of the spinal joints, which leads to pain and stiffness. The pathogenesis of these diseases is linked to genetic markers such as *HLA-B27*.[4] Advanced diagnostic methods, including MRI and blood tests for inflammatory markers, are essential for early detection and monitoring.[5] Spinal tumors' pathogenesis includes genetic mutations and environmental factors that lead to abnormal cell growth. Diagnostic imaging, biopsy, and molecular testing are crucial in identifying and characterizing spinal tumors, guiding treatment decisions.[6,7]

The advancements in spine diagnostics have a profound impact on modern healthcare, influencing clinical practice, patient outcomes, and healthcare systems. Improved imaging technologies, such as high-resolution MRI and three-dimensional (3D) CT scans, provide detailed visualization of spinal structures, enhancing diagnostic accuracy.[8,9] The integration of genetic and molecular diagnostics enables personalized treatment plans. By understanding the genetic and molecular basis of spinal diseases, clinicians can tailor therapies to individual patients, improving efficacy and reducing adverse effects.[10] Advances in diagnostic imaging have facilitated the development of minimally invasive surgical techniques. Real-time imaging guidance during procedures minimizes tissue damage, reduces recovery time, and lowers the risk of complications.[11] Early detection of spinal disorders through advanced diagnostics allows for timely intervention, potentially preventing disease progression and reducing the burden of chronic spinal conditions on patients and healthcare systems. Studies have shown that early intervention can significantly improve long-term outcomes for patients with spinal conditions.[12] Accurate diagnostics and early interventions can reduce healthcare costs by decreasing the need for extensive surgeries and long-term care. Efficient diagnostic processes also streamline patient management, optimizing resource utilization within healthcare systems.

Accurate diagnosis of spinal conditions is critically dependent on the analysis provided by MRI.[9,13-14] High-quality spine MRI segmentation is crucial for enabling doctors to precisely locate and examine spinal structures, thereby facilitating the diagnosis of various spine-related diseases such as disc herniation and spinal deformities.[15,16] Accurate segmentation results are essential for assessing the severity of these conditions and developing effective treatment plans.

To address the labor-intensive nature of medical imaging tasks, there is a growing trend toward data-driven approaches in contemporary medical imaging technology.[17,18] Traditional image processing techniques, such as thresholding,[19] edge detection,[20] and mathematical morphology,[21] have yielded some positive outcomes. However, significant advancements have been made in recent years with the advent of deep learning methods, particularly convolutional neural networks, which have revolutionized spine medical image segmentation.[22,23] Models such as U-Net,[24] DeepLab,[25] and Fully Convolutional Networks[26] have been extensively applied to spine image segmentation tasks, with the U-Net architecture achieving notable success. This model excels in extracting and representing feature information in MRI medical image analysis.

To further enhance U-Net's ability to capture multi-scale information, Huang *et al.* introduced U-Net++,[27] which incorporates full-scale skip connections. This design effectively aggregates low-level and high-level semantic information, improving segmentation performance. However, the practical application of U-Net++ is challenged by the large size of spine medical images and the network's complex structure, leading to prolonged training times and reduced accuracy due to the intricate nature of spinal images.

This study proposes an innovative and efficient spine segmentation method called "J-Unet" network to overcome these limitations and improve model accuracy while reducing training costs. Our approach includes optimizing multi-scale skip connection paths and deepening the network depth of the upsampling component to capture

finer features across different depths. The key contributions of this study are as follows:

(1) Optimization of multi-scale skip connection paths: By refining these paths, we can better control model parameters while ensuring high segmentation accuracy.
(2) Increased network depth in the upsampling process: This modification enables the model to reconstruct complex structures and details more accurately, enhancing overall precision.
(3) Introduction of residual connections: These connections are incorporated locally to mitigate the problem of gradient vanishing, accelerate training, and strengthen the network's feature learning and representation capabilities.

In summary, the proposed method addresses the challenges of excessive training time and model accuracy in spine MRI segmentation. By optimizing the architecture of U-Net++, we aim to provide a more efficient and precise tool for medical professionals, thereby improving diagnostic and treatment outcomes for spinal diseases.

2. Data and methods

In this section, the dataset used in this study and the processing methods employed are introduced. A detailed description of the proposed J-Unet architecture, designed specifically for spine MRI image segmentation, is also provided.

2.1. Dataset and processing methods

The dataset utilized in this study consists of T2-weighted MRI scans obtained from a cohort of 215 patients, sourced from multiple medical institutions to ensure diversity and robustness in the data.[28,29] All images are provided in Nifti format, a widely recognized standard for medical imaging, which facilitates comprehensive 3D visualization and analysis.

Initially, the dataset's labels included 21 distinct pixel values representing various anatomical structures and features. For the purposes of this study, these original labels were reclassified into three primary categories: vertebral bone, intervertebral disc, and background regions. This reclassification simplifies the segmentation task, focusing on the most clinically relevant structures. The reclassified pixel values for segmentation are detailed in Table 1, providing a clear framework for subsequent image processing and analysis.

We configured the training and validation datasets with a 4:1 ratio. To ensure a balanced and representative distribution of data, the allocation was performed through random sampling, which mitigates potential biases and ensures that the training dataset captures a comprehensive range of variations present in the images. Each input spine image is standardized to a size of 512 × 512 pixels to maintain consistency and optimize the computational efficiency during model training. Figure 1 shows an example of an original spinal MRI image and its corresponding label image, illustrating the visual clarity and distinct boundaries of the segmented regions.

Table 1. Reclassified pixel values for segmentation

Vertebral bone	Intervertebral disc	Background regions
(100 100 100)	(255 255 255)	(0 0 0)

2.2. J-Unet architecture

The J-Unet architecture introduced in this study represents a significant evolution from traditional U-Net and U-Net3+ designs, addressing some of the limitations inherent in these models and incorporating advanced features to enhance performance in spine MRI image segmentation. Figure 2 illustrates the architecture of the proposed J-Unet, highlighting the innovative structural elements that differentiate it from its predecessors.

2.2.1. Asymmetric network architecture

One of the most notable advancements in the J-Unet model is its asymmetric network architecture, which diverges from traditional U-Net structures by extending the upsampling pathway with three additional layers compared to its downsampling counterpart. This asymmetry is deliberately engineered to enhance the network's ability to capture and reconstruct complex, hierarchical features more comprehensively. By increasing the depth on the upsampling side, the model can progressively refine the spatial resolution of the feature maps, thereby improving the accuracy of the segmentation output.

In this design, the input and output channels of these additional upsampling layers are set to 120, effectively reducing the number of parameters without compromising accuracy. The advantages of an asymmetric network architecture are evident in its enhanced ability to precisely reconstruct complex structures and details. Traditional symmetric architectures often struggle with images characterized by irregular shapes and intricate edges. By increasing the depth on the upsampling side, the asymmetric network better adapts to these complexities. It also utilizes skip connections between specific scales to fine-tune model parameters while maintaining segmentation accuracy.

The depth of the upsampling pathway in the J-Unet is crucial for reconstructing the finer details of spinal

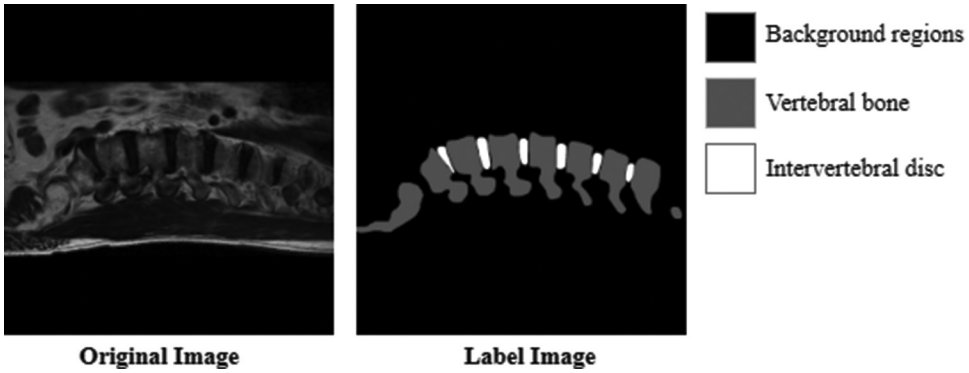


Figure 1. Example original spinal magnetic resonance imaging image and its corresponding label image

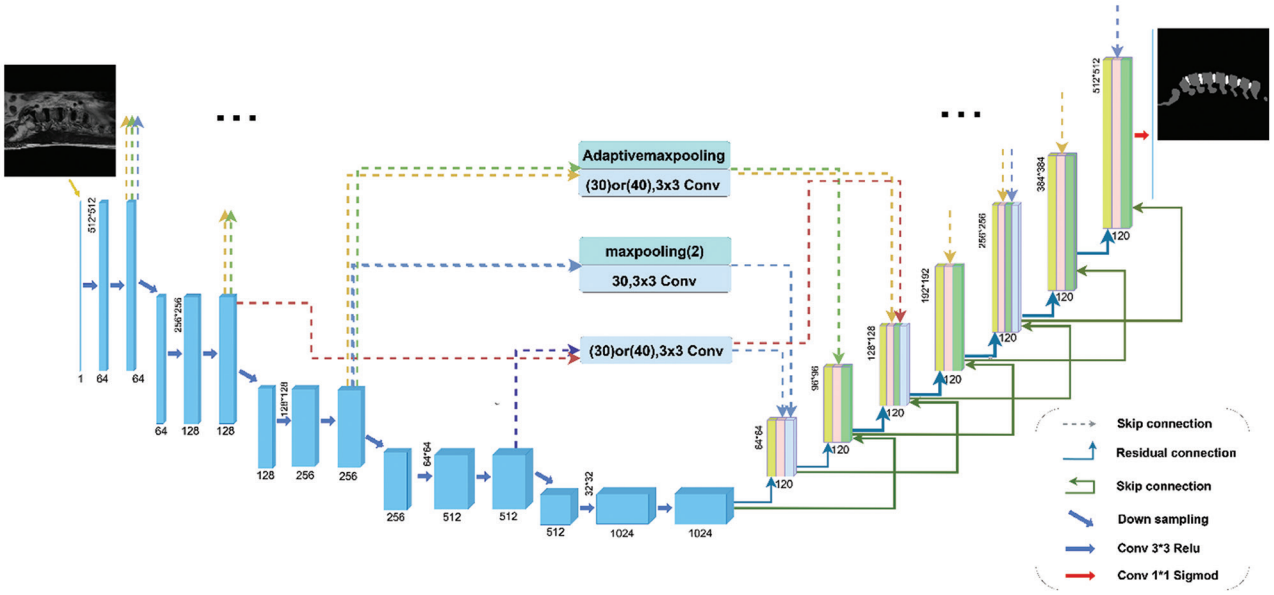


Figure 2. The architecture of the proposed J-Unet

structures, which are often lost in traditional symmetric architectures. This extended pathway allows the network to learn more detailed and nuanced representations of the spinal structures, crucial for tasks that require high precision, such as medical image segmentation. The ability to capture and reconstruct hierarchical features ensures that the segmentation of the spine is both accurate and reliable, which is essential for clinical applications where precise anatomical delineation is required for diagnosis and treatment planning.

2.2.2. Adjacent-scale skip connections

Unlike the U-Net3+ architecture, which utilizes fully-scaled skip connections, the J-Unet adopts adjacent-scale skip connections. This approach strategically reduces the redundancy and overall parameter count associated with the skip connections while maintaining the ability to leverage multi-scale feature information effectively. Fully-scaled skip connections can introduce excessive redundancy, leading to an unnecessary increase in computational load and model complexity. In contrast, adjacent-scale skip connections streamline the network, reducing computational overhead without sacrificing the richness of the multi-scale features. This optimization not only simplifies the network architecture but also enhances computational efficiency and model scalability, making the J-Unet more practical for large-scale applications and real-time processing.

In the J-Unet model, the size of feature maps does not uniformly change by integer multiples. For instance, adaptive max pooling is employed to resize a 256 × 256 feature map from the encoder to a 192 × 192 feature map in the decoder, whereas standard fixed window max pooling is used for regular size adjustments. This method enhances

internal information flow and integration within the network, improves the perception of detailed structures and edges, and ultimately boosts the accuracy of segmentation tasks. Figure 3 illustrates these tailored connections in the J-Unet architecture, showcasing how the model efficiently integrates multi-scale information without the overhead of fully-scaled connections.

Adjacent-scale skip connections allow the network to capture both fine-grained details and coarse-grained semantics across different scales but with fewer parameters. This design retains the ability to capture detailed features necessary for accurate segmentation while reducing the computational burden, making the model more efficient and scalable. The strategic use of these connections ensures that the model can effectively integrate information from different scales, enhancing its ability to accurately segment complex spinal structures.

2.2.3. Partial residual connections (PRCs)

The addition of three upsampling layers not only deepens the network architecture but also introduces the risk of gradient vanishing. To counteract this and boost the network's feature representation capabilities, residual connections are strategically employed during the upsampling process. Residual connections, introduced by He *et al.*[30] in their seminal work on deep residual networks, are designed to preserve and reuse the information captured in earlier layers, addressing the vanishing gradient problem and facilitating the training of deeper networks.

In the J-Unet model, each decoder layer is connected to the largest-scale feature map from the most adjacent layer through a residual connection, allowing for the construction of an even deeper network. The innovation of PRCs offers several advantages. Unlike traditional global residual connections, PRCs are more selective, maintaining and transmitting essential feature information and thereby enhancing the network's ability to learn complex feature representations more efficiently.

This selective connection strategy not only accelerates the training process but also reduces resource consumption. Overall, PRCs significantly improve training stability, enhance learning capacity, and bolster the generalization performance of neural networks, making them a vital component in the design of advanced deep learning models. By preserving critical information from earlier layers and reintroducing it at later stages, PRCs facilitate a more robust learning process, enabling the network to capture intricate details and complex patterns within the spinal MRI images.

2.2.4. Integration of advanced structural elements

The integration of these advanced structural elements in J-Unet, including extended asymmetry in the upsampling path, optimized skip connections, and strategic use of residual connections, aims to improve the accuracy and efficiency of spine MRI image segmentation. These enhancements enable the model to handle the intricacies of medical imaging data, which often involve complex anatomical variations and subtle pathological features.

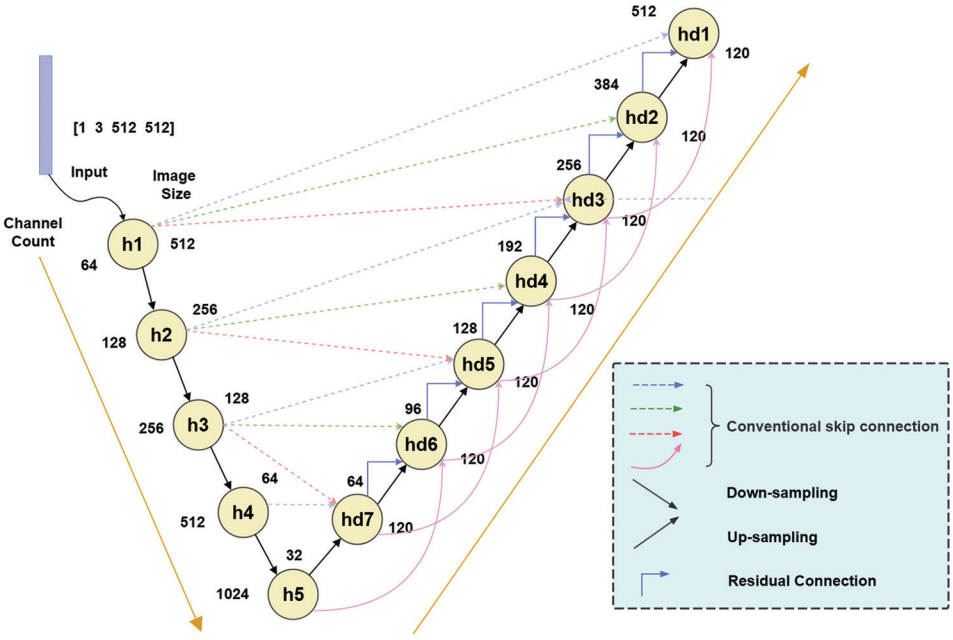


Figure 3. Detailed illustration of connections in J-Unet

By effectively managing these complexities, the J-Unet architecture supports more precise diagnostic outcomes and potentially informs better clinical decision-making.

Accurate segmentation of spine MRI images can aid in the early detection and diagnosis of spinal pathologies such as disc herniation, spinal stenosis, and tumors. Precise delineation of these structures is essential for planning surgical interventions, assessing disease progression, and evaluating treatment efficacy. The improved segmentation performance offered by the J-Unet model can therefore contribute to better patient outcomes by enabling more targeted and effective treatments.

Moreover, the computational efficiency and scalability of the J-Unet architecture make it suitable for deployment in clinical settings where rapid processing of large volumes of imaging data is required. This is particularly important in modern healthcare environments, where the demand for advanced imaging techniques is increasing, and the ability to process and analyze data quickly can significantly impact the quality of care provided.

3. Results and analysis

In this section, we detail the organization and methodology of our experimental study using magnetic resonance (MR) imaging data. The dataset, comprising MR image sequences from 215 patients, was stratified into training and test sets in a 4:1 ratio. This separation was carefully designed to ensure both sets were representative of the overall dataset, supporting the generalizability of our findings. Using the J-Unet architecture, we developed a model capable of automatically segmenting vertebrae and intervertebral discs within spinal MR images. The performance of this model was rigorously evaluated to ascertain its efficacy in medical imaging tasks. In addition, to provide a comprehensive analysis of our model's capabilities, we compared its performance with several other established neural network architectures: Unet, Unet++, Unet+++, and Res-Unet. This comparative study aimed to highlight the strengths and potential areas for improvement in the J-Unet architecture relative to other models in handling complex segmentation tasks in spinal MR images.

The computational experiments were conducted using the Pytorch framework and cuDNN library, optimized for deep neural network operations. All models were trained on a robust hardware setup featuring a single NVIDIA GTX 3090 graphics processing unit. This high-performance computing environment ensured efficient processing of large-scale data, facilitating timely training and evaluation of the models. This systematic approach allowed us to not only assess the specific advantages of the J-Unet model but also to establish a baseline for performance against other prominent architectures in the field, thereby providing a clear perspective on the current state of the art in spinal MR image segmentation.

3.1. Experimental setup

We utilized three evaluation metrics to assess the segmentation performance of our models: accuracy, mean intersection over union (mIOU), and dice coefficient. These metrics provide a comprehensive view of the model's effectiveness in segmenting vertebrae and intervertebral discs in spinal MR images. The formulas for these metrics are as follows:

(1) Accuracy: Measures the proportion of true results (both true positives [TP] and true negatives [TN]) among the total number of cases examined.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{I}$$

(2) mIOU: Calculates the average IOU across all classes, providing an overall measure of segmentation performance.

$$mIOU = \frac{cIOU + ucIOU}{2} \tag{II}$$

(3) Dice coefficient: Evaluates the overlap between the predicted and ground truth segments, often used in medical image analysis.

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{III}$$

Definitions of confusion matrix terms are given below:

- TP: A positive class instance correctly predicted as positive.
- False negative (FN): A positive class instance incorrectly predicted as negative.
- False positive (FP): A negative class instance incorrectly predicted as positive.
- TN: A negative class instance correctly predicted as negative.

In these definitions, TP and TN indicate correct predictions of the instance class, while FN and FP indicate incorrect predictions. Table 2 outlines the specific parameter settings used for the segmentation model.

These settings were carefully chosen to optimize model performance and ensure reliable evaluation metrics. They

Table 2. Parameter settings of the segmentation model

Parameter	Learning rate	Optimizer	Batch size	Epoch
Value	1e-4	Adam	2	100

include details on learning rate, batch size, number of epochs, and other hyperparameters critical for training deep learning models. The segmentation performance of J-Unet was benchmarked against several established models: Unet, Unet++, Unet+++, and Res-Unet. Each model was trained and evaluated using the same dataset and parameter settings to ensure a fair comparison. The results were analyzed using the aforementioned metrics to determine the relative strengths and weaknesses of each model.

3.2. Optimization and loss function selection

In this study, we utilized the Adam optimizer due to its significant advantages in deep learning model training. Adam's ability to apply different scaling factors to parameter updates facilitates the discovery of the global optimum more efficiently during the training process. This tailored approach to parameter adjustment helps navigate the complex loss landscapes often encountered in deep learning models, ensuring that each parameter evolves at an appropriate pace. By adaptively adjusting the learning rate for each parameter, Adam accelerates convergence, enhancing the overall efficiency of the training process. In addition, Adam is known for its computational efficiency and ease of implementation. It requires less memory compared to other optimizers, making it ideal for handling large-scale data and complex models. To ensure stable convergence towards a local optimum, we set the learning rate to 0.0001. This small learning rate helps in fine-tuning the model parameters, preventing overshooting and ensuring a smooth descent in the loss landscape.

The choice of Dice Loss as the loss function for this study is driven by its effectiveness in addressing common challenges in medical image segmentation. Dice Loss is particularly robust in scenarios with imbalanced classes, a common occurrence in medical imaging. Its calculation involves the intersection and union of the predicted and true values, making it less sensitive to the disproportionate pixel counts of different classes. This robustness ensures that the model performs well even when certain classes are underrepresented.

Moreover, Dice Loss provides smoother gradients compared to other loss functions, contributing to a more stable training process. This stability helps mitigate issues such as exploding or vanishing gradients, which can hinder the training of deep neural networks. By emphasizing the similarity between predicted and true segmentations, Dice Loss encourages the model to produce accurate and refined segmentation results. This focus on overlap and accuracy is crucial for tasks requiring precise delineation of structures, such as in medical imaging where detail and precision are paramount.

The proven applicability of Dice Loss in medical imaging further validates its use in this study. Dice Loss is widely employed in medical image segmentation tasks, particularly those demanding high accuracy, such as tumor segmentation. Its sensitivity to fine structures and boundaries makes it an excellent choice for medical applications, where the accurate segmentation of anatomical structures is critical for diagnosis and treatment planning. The combination of the Adam's optimizer and Dice Loss provides a robust framework for training our segmentation model. Adam's efficiency and adaptability, coupled with Dice Loss's robustness to class imbalance and emphasis on accurate segmentation, ensure a high-performance model suitable for the complexities of medical image analysis.

3.3. Model training

During the training process, Dice Loss was employed as the loss function for our deep learning model. Dice Loss is a widely used metric in image segmentation tasks, designed to measure the similarity between the predicted segmentation and the ground truth.

The segmentation models, namely UNET, UNET++, UNET+++, ResUNET, and J-UNET, were trained using the training set. Figure 4 illustrates the variation of loss and Dice coefficient during training and validation for these five models. The graphs provide a clear comparison of how each model's performance evolved over the training epochs. During the training process, we encountered several instances of gradient explosions, particularly with the UNET+++ model. These gradient explosions resulted in a sharp increase in loss values. To address this issue, we employed a strategy of resuming training from checkpoints saved before the occurrence of the gradient explosions. This approach allowed us to continue training effectively, leading to eventual convergence. The training of the remaining network models exhibited normal convergence patterns without such disruptions.

Upon analyzing the final convergence ranges, it was evident that J-UNET and UNET achieved lower loss values on the training set compared to the other models. On the test set, although the differences in loss values among the various models were not substantial, J-UNET and ResUNET consistently demonstrated smaller loss values. This indicates that J-UNET, in particular, exhibited superior generalization ability, aligning well with the task objectives. The combination of Dice Loss and the robust architecture of J-UNET contributed to its superior performance in segmenting vertebrae and intervertebral discs in spinal MR images. The J-UNET model not only converged effectively but also showed a strong ability to generalize from the training set to the test set, making it a highly effective tool for medical image segmentation tasks.

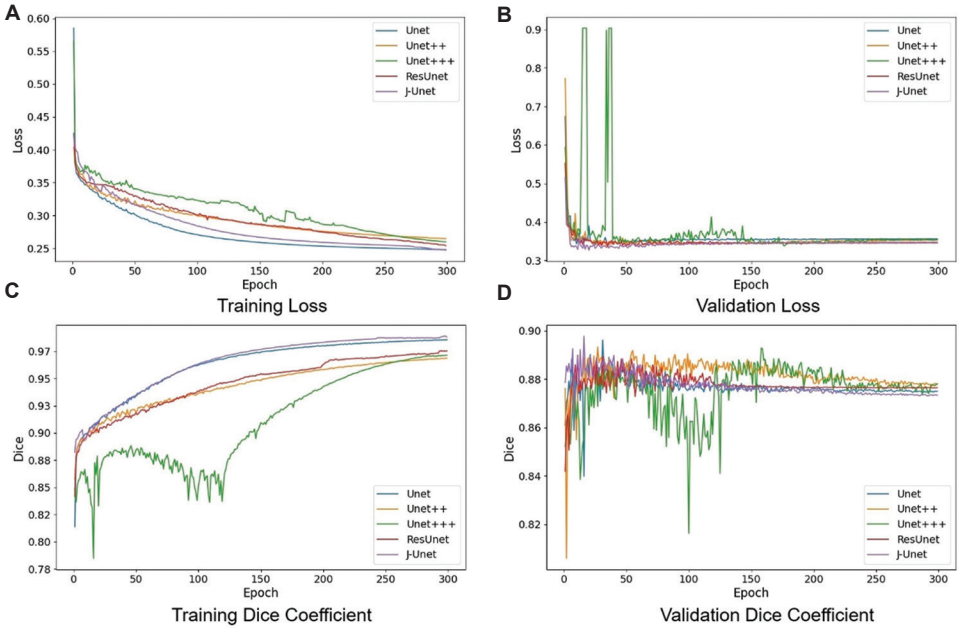


Figure 4. Loss values and evaluation metrics in model training and validation. (A) Training Loss; (B) Validation Loss; (C) Training Dice Coefficient; and (D) Validation Dice Coefficient

3.4. Results

Upon completion of the training phase for the various network models, we computed several key metrics on the test set to assess their performance. These metrics include dice, accuracy, and mIOU, which are standard measures for evaluating segmentation quality. For detailed formulas of these metrics, refer to section 3. The results, presented in Table 3, highlight the superior performance of the J-Unet model across all evaluation criteria.

As illustrated in Table 3, the J-Unet model achieved notable improvements in comparison to other models. Specifically, J-Unet improved the dice score by at least 0.24%, accuracy by 0.74%, and mIOU by at least 0.24%. These improvements underscore the effectiveness of the J-Unet architecture in accurately segmenting spinal MR images.

In addition to evaluating segmentation performance, we compared the number of parameters across the different models, as shown in Table 4. Our J-Unet model, while having a parameter count slightly higher than UNET, boasts approximately 78.1% fewer parameters than Res-UNET, 17.2% fewer than UNET+++, and roughly 2.3% fewer than Res-UNET. Despite its relatively smaller parameter count, J-Unet consistently achieved the highest performance in all evaluation metrics. This efficiency indicates that our model is not only accurate but also resource-effective.

The experimental results clearly demonstrate that the design choices of J-Unet, such as the asymmetric network structure, adjacent-scale skip connections, and PRCs, contribute significantly to its enhanced accuracy and generalization capabilities. These architectural innovations enable J-Unet to maintain high performance with fewer parameters, reducing the computational resources and time required for both training and inference. Figure 5 shows the segmentation maps of different models. The superior performance and efficiency of J-Unet affirm the benefits of its innovative design. The model's ability to achieve high segmentation accuracy with a smaller parameter footprint makes it valuable for medical image

Table 3. Scores obtained of all models

Model	Dice	Accuracy	mIOU
Unet	0.8791	0.9613	0.8093
Unet++	0.8885	0.9655	0.8264
Unet+++	0.8866	0.9606	0.8248
Res-Unet	0.8889	0.9642	0.8241
J-Unet	**0.8913**	**0.9729**	**0.8288**

The values in boldface indicate the best performance among all models in each metric.
Abbreviation: mIOU: Mean intersection over union.

Table 4. The number of parameters of different models

Model	Unet	Unet++	Unet+++	Res-Unet	J-Unet
Total parameters	17267523	24423232	25659999	101942977	22331979

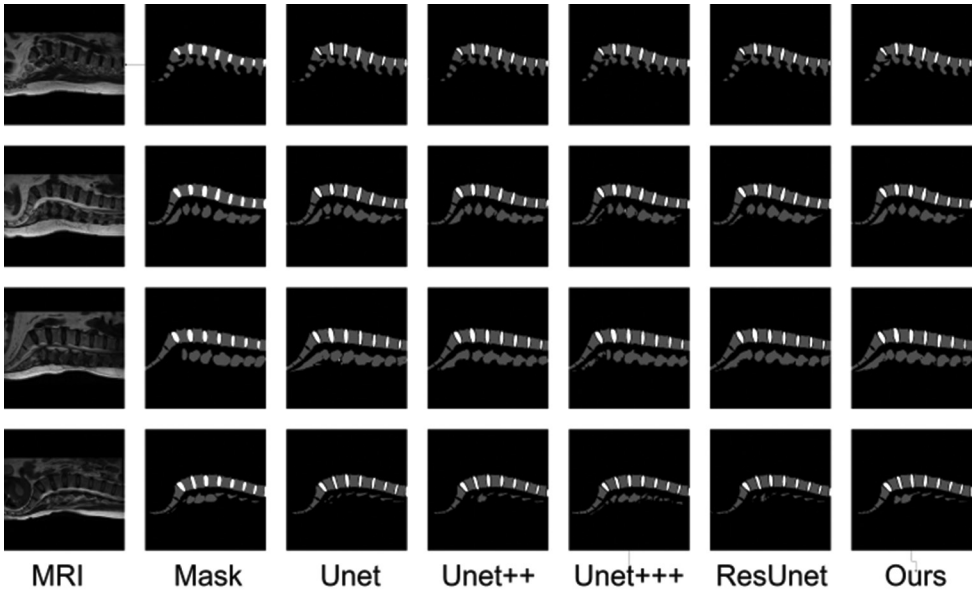


Figure 5. Segmentation maps of different models

segmentation, providing a practical solution that balances performance with computational efficiency.

4. Conclusion

In this study, we introduce a novel asymmetric U-Net architecture, termed J-Unet, designed for efficient and accurate spine MRI image segmentation. J-Unet distinguishes itself from conventional U-Net and its variants through its asymmetric encoder-decoder structure, featuring a deeper upsampling path that enhances the precise reconstruction of anatomical details. The incorporation of adjacent-scale skip connections and PRCs allows J-Unet to reduce the number of model parameters while maintaining the flow of multi-scale contextual information. We rigorously evaluated J-Unet using a dataset comprising 215 spine MRI images. The results indicate that J-Unet significantly outperforms other models, including U-Net, U-Net++, and U-Net+++, achieving at least a 0.24% improvement in both dice score and mIoU. Furthermore, J-Unet operates with substantially fewer parameters compared to U-Net+++ and Res-UNET, demonstrating superior performance and efficiency. In conclusion, this study presents an innovative asymmetric U-Net architecture specifically tailored for spine MRI image segmentation. J-Unet's unique design enables precise localization and segmentation while optimizing parameter efficiency, making it a highly accurate and resource-effective solution. Our model offers a promising advancement in automating spine segmentation in medical image analysis, potentially enhancing diagnostic processes and treatment planning.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Longfei Zhou

Formal analysis: Xingyu Chen, Weihao Cheng, Zhanghao Qin, Tianao Shen, Pingyu Cao, Zebo Huang, Xiangyu Wu, Yiyao Zhang

Investigation: Longfei Zhou, Xingyu Chen

Methodology: Xingyu Chen, Weihao Cheng, Zhanghao Qin, Tianao Shen, Pingyu Cao, Zebo Huang, Xiangyu Wu, Yiyao Zhang

Writing–original draft: Xingyu Chen, Weihao Cheng, Zhanghao Qin, Tianao Shen, Pingyu Cao, Zebo Huang, Xiangyu Wu, Yiyao Zhang

Writing–review & editing: Longfei Zhou

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data are available from the corresponding author upon reasonable request.

References

1. Kubaszewski Ł, Wojdasiewicz P, Rożek M, *et al*. Syndromes with chronic non-bacterial osteomyelitis in the spine. *Reumatologia*. 2015;53(6):328-336.

   doi: 10.5114/reum.2015.57639

2. Rajasekaran S, Bajaj N, Tubaki V, Kanna RM, Shetty AP. ISSLS prize winner: The anatomy of failure in lumbar disc herniation: An *in vivo*, multimodal, prospective study of 181 subjects. *Spine* (*Phila Pa 1976*). 2013;38(17):1491-1500.

   doi: 10.1055/s-0034-1376749

3. Adams MA, Roughley PJ. What is intervertebral disc degeneration, and what causes it? *Spine* (*Phila Pa 1976*). 2006;31(18):2151-2161.

   doi: 10.1097/01.brs.0000231761.73859.2c

4. Pedersen SJ, Maksymowych WP. The pathogenesis of ankylosing spondylitis: An update. *Curr Rheumatol Rep*. 2019;21(10):58.

   doi: 10.1007/s11926-019-0856-3

5. Aouad K, Maksymowych WP, Baraliakos X, Ziade N. Update of imaging in the diagnosis and management of axial spondyloarthritis. *Best Pract Res Clin Rheumatol*. 2020;34(6):101628.

   doi: 10.1016/j.berh.2020.101628

6. Hassan I, Wietfeldt ED. Presacral tumors: Diagnosis and management. *Clin Colon Rectal Surg*. 2009;22(2):84-93.

   doi: 10.1055/s-0029-1223839

7. Hashimoto K, Nishimura S, Miyamoto H, Toriumi K, Ikeda T, Akagi M. Comprehensive treatment outcomes of giant cell tumor of the spine: A retrospective study. *Medicine* (*Baltimore*). 2022;101(32):e29963.

   doi: 10.1097/MD.0000000000029963

8. Kalra MK, Maher MM, Toth TL, *et al*. Strategies for CT radiation dose optimization. *Radiology*. 2004;230(3):619-628.

   doi: 10.1148/radiol.2303021726

9. Winn A, Martin A, Castellon I, *et al*. Spine MRI: A review of commonly encountered emergent conditions. *Top Magn Reson Imaging*. 2020;29(6):291-320.

   doi: 10.1097/RMR.0000000000000261

10. Castaldo G, Lembo F, Tomaiuolo R. Molecular diagnostics: Between chips and customized medicine. *Clin Chem Labo Med*. 2010;48(7):973-982.

    doi: 10.1515/CCLM.2010.182

11. Arslantaş A, Dalbayrak S, Şimşek S, *et al*. Minimally Invasive Spine Surgery Current Aspects. Turkey: Ali Arslantaş. 2016.

12. Haldeman S, Kopansky-Giles D, Hurwitz EL, *et al*. Advancements in the management of spine disorders. *Best Pract Res Clin Rheumatol*. 2012;26(2):263-280.

    doi: 10.1016/j.berh.2012.03.006

13. Azimi P, Yazdanian T, Benzel EC, *et al*. A review on the use of artificial intelligence in spinal diseases. *Asian Spine J*. 2020;14(4):543.

    doi: 10.31616/asj.2020.0147

14. Da Costa RV, Moore SA. Differential diagnosis of spinal diseases. *Vet Clin North Am Small Anim Pract*. 2010;40(5):755-763.

    doi: 10.1007/978-981-16-9759-3_11

15. Cohen-Adad J, Alonso-Ortiz E, Abramovic M, *et al*. Generic acquisition protocol for quantitative MRI of the spinal cord. *Nat Protocols*. 2021;16(10):4611-4632.

16. Sollmann N, Löffler MT, Kronthaler S, *et al*. MRI-based quantitative osteoporosis imaging at the spine and femur. *J Magn Reson Imaging*. 2021;54(1):12-35.

    doi: 10.1002/jmri.27260

17. Willemink MJ, Koszek WA, Hardell C, *et al*. Preparing medical imaging data for machine learning. *Radiology*. 2020;295(1):4-15.

    doi: 10.1148/radiol.2020192224

18. Patel V. A framework for secure and decentralized sharing of medical imaging data via blockchain consensus. *Health Informatics J*. 2019;25(4):1398-1411.

    doi: 10.1177/1460458218769699

19. Senthilkumaran N, Vaithegi S. Image segmentation by using thresholding techniques for medical images. *Comput Sci Eng Int J*. 2016;6(1):1-13.

20. Song Y, Ma B, Gao W, Fan S. Medical image edge detection based on improved differential evolution algorithm and prewitt operator. *Acta Microscopica*. 2019;28(1).

21. Zhao F, Zhang J, Ma Y. Medical image processing based on mathematical morphology. In: *Proceedings of the 2012 International Conference on Computer Application and System Modeling (ICCASM 2012)*; 2012. p. 948-950.

    doi: 10.2991/iccasm.2012.241

22. Yamanakkanavar N, Choi JY, Lee B. MRI segmentation and classification of human brain using deep learning for diagnosis of Alzheimer's disease: A survey. *Sensors* (*Basel*). 2020;20(11):3243.

    doi: 10.3390/s20113243

23. Li H, Luo H, Huan W, *et al*. Automatic lumbar spinal MRI image segmentation with a multi-scale attention network. *Neural Comput Appl*. 2021;33:11589-11602.

    doi: 10.1007/s00521-021-05856-4

24. Ronneberger O, Fischer P, Brox T. *U-net: Convolutional Networks for Biomedical Image Segmentation*. Berlin: Springer; 2015. p. 234-241.

doi: 10.1007/978-3-319-24574-4_28

25. Hempe H, Yilmaz EB, Meyer C, Heinrich MP. Opportunistic CT screening for degenerative deformities and osteoporotic fractures with 3D DeepLab. In: *Medical Imaging 2022: Image Processing*. Bellingham, DC: SPIE; p. 127-134.

doi: 10.1117/12.2612848

26. Miao S, Piat S, Fischer P, *et al*. Dilated FCN for Multi-agent 2D/3D Medical Image Registration. In: *Proceedings of the AAAI Conference on Artificial Intelligence*.

doi: 10.1609/aaai.v32i1.11576

27. Huang H, Lin L, Tong R, *et al*. *Unet 3+: A full-Scale Connected UNet for Medical Image Segmentation*. United States: IEEE; p. 1055-1059.

doi: 10.1109/ICASSP40776.2020.9053405

28. Pang S, Pang C, Zhao L, *et al*. SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. *IEEE Trans Med Imaging*. 2020;40(1):262-273.

doi: 10.1109/TMI.2020.3025087

29. Pang S, Pang C, Su Z, *et al*. "DGMSNet: Spine segmentation for MR image by a detection-guided mixed-supervised segmentation network. *Med Image Anal*. 2022;75:102261.

doi: 10.1016/j.media.2021.102261

30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; 2016. p. 770-778.

doi: 10.1109/CVPR.2016.90

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Algorithm development and metal oxide nanoparticle analysis in magnetic resonance imaging: Advancing neurodegenerative disease diagnostics

Daniela Gomes Bornal[1†], Hulder Henrique Zaparoli[1†], Marina Piacenti-Silva[2], Paulo Noronha Lisboa-Filho[2], and Marcela de Oliveira[2*]

[1]Postgraduate Program in Science and Technology of Materials - POSMAT, School of Sciences/São Paulo State University, Bauru, São Paulo, Brazil

[2]Department of Physics and Meteorology, School of Sciences/São Paulo State University, Bauru, São Paulo, Brazil

(This article belongs to the *Special Issue: Artificial intelligence for diagnosing brain diseases*)

Abstract

Magnetic resonance imaging (MRI) is critical in the diagnosis of neurodegenerative diseases, enabling the detection of brain lesions. Recent research has examined metallic nanoparticles (NPs) as MRI contrast agents (CAs) that can enhance lesion visibility by altering relaxation times. This study investigates the effects of metal oxide NPs on MRI relaxation times and brain lesion signals and proposes an algorithm for automated relaxation time determination using these NPs. The utilized NPs were synthesized using the sol–gel method and characterized using Fourier-transform infrared spectroscopy and X-ray diffraction. MRI scans were performed on a phantom infused with varying concentrations of each metal oxide NP to assess changes in pixel signal intensities and relaxation rates. Our analysis involved segmenting the MRI images to focus on regions with different NP concentrations. The algorithm computed the longitudinal relaxation time for each region, revealing that Fe_2O_3 NPs exhibited the most substantial effect on signal intensity and relaxation time. The results indicated a high correlation ($r = 0.9977$), demonstrating strong agreement and confirming the reliability of our method. Our findings suggest that metallic oxide NPs, particularly Fe_2O_3, can considerably alter magnetization and act as effective negative CAs in MRI. These capabilities can improve the monitoring and treatment efficacy of neurodegenerative diseases. Our method for quantifying longitudinal relaxation times can potentially enhance routine clinical MRI assessments, offering a promising tool for future clinical applications.

Keywords: Magnetic resonance imaging; Algorithm; Longitudinal relaxation time (T1); Signal intensity

1. Introduction

Magnetic resonance imaging (MRI) is a vital diagnostic imaging tool in the medical field, particularly for diagnosing various neurodegenerative diseases.[1,2] MRI can differentiate

†*These authors contributed equally to this work.*

***Corresponding author:**
Marcela de Oliveira
(marcela.oliveira@unesp.br)

Citation: Bornal DG, Zaparoli HH, Piacenti-Silva M, Lisboa-Filho PN, de Oliveira M. Algorithm development and metal oxide nanoparticle analysis in magnetic resonance imaging: Advancing neurodegenerative disease diagnostics. *Artif Intell Health*. 2025;2(1):53-67. doi: 10.36922/aih.3947

Received: June 14, 2024

Revised: August 1, 2024

Accepted: August 28, 2024

Published Online: October 9, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

between neurodegenerative conditions such as multiple sclerosis (MS), which is characterized by brain lesions primarily in the white matter. These lesions are identified through demyelination, inflammation, and axonal loss.[3,4] A comprehensive understanding of brain MRI findings is essential for accurate MS diagnosis.[5]

In MRI, a portion of proton nuclei within the body aligns parallel to an external magnetic field (B_0) to generate images.[6] These nuclei precess at a Larmor frequency (w_0) and are excited to an antiparallel state by a radio frequency (RF) pulse. On removing the RF pulse, the nuclei return to their equilibrium state, a process involving longitudinal (*T1*) and transverse (*T*2) relaxations.[6,7] *T*1 denotes the time required to reach 63% longitudinal magnetization, while *T*2 is defined as the time required for a decrease in the transverse magnetization by 37% its initial value.[8] Standard MRI sequences, including *T*1-weighted (*T*1-w), *T*2-weighted (*T*2-w), fluid-attenuated inversion recovery (FLAIR), and *T*1-weighted contrast modalities, are employed to detect overt lesions and assess tissue atrophy in MS.[2,9] MS lesions typically manifest as hyperintensities in *T*2-w and FLAIR images and as hypointensities in *T1*-w images.[2]

Recent research has revealed an imbalance in the metal levels among individuals suffering from MS, suggesting a link between metal levels and neurodegenerative diseases. This imbalance may contribute to brain injuries.[10-15] Metallic elements are considered potential causes of brain lesions. Furthermore, these metals are hypothesized to accumulate within lesions, altering the MRI contrast signal similar to contrast agents (CAs). However, the mechanism by which these metals influence the MRI signals of lesions remains underexplored.

Advancements in nanotechnology and the unique properties of metallic nanoparticles (NPs) that influence MRI relaxation times have facilitated the use of NPs as CAs in MRI.[1,8,16-19] Metallic NPs can reduce *T1* or *T2* by accelerating relaxation rates and inducing magnetic field inhomogeneity.[20] Regions containing these NPs appear bright in *T1*-w images, and NPs act as negative CAs, reducing *T2* signals.[16] CAs are essential for enhancing the contrast and sensitivity in MRI diagnostics. For instance, Gd is widely used as a CA in MRI, favored for its prolonged magnetic relation time and large magnetic moment.[21,22] Studies have also explored MRI CAs based on iron oxide (Fe_2O_3), gadolinium oxide, and manganese oxide NPs.[7,22-25] Cai *et al.* (2019)[21] highlighted advancements in the utilization of Mn oxide as a CA in MRI, while Blanco-Andujar *et al.*[24] emphasized the design of Fe_2O_3-based magnetic NPs that enable the optimization of their relaxivity for use as CAs in *T2*-w MRI.

The vast amount of data generated during MRI presents challenges for visual analysis, necessitating advanced analytical methods. Artificial intelligence-based algorithms are gaining prominence in the biomedical field and medical image analysis.[26] Automated image analysis enables the handling of extensive datasets with consistent precision, overcoming the limitations of manual methods. AI applications serve as decision support systems, although their development poses challenges.[27]

AI algorithms are widely used for targeting specific regions (organs or tissues), classifying disease stages, and diagnosing tumors.[28-31] For instance, Chang *et al.*[32] explored the use of a deep learning algorithm for the automated segmentation and quantification of the myocardial *T1* values, while Bidhult *et al.*[33] developed algorithms for *T1* and *T2* relaxation mapping in cardiac imaging. Specifically, for brain regions, Jibon *et al.*[34] improved a classification method to distinguish between cancerous and noncancerous tumors from brain MRI using log polar transformation and convolutional neural networks. In addition, the improved algorithm developed by Oliveira *et al.*[11] demonstrated the effectiveness of convolutional neural networks for detecting brain lesions in individuals with MS. In general, the primary role of AI is to create tools that automatically learn from data and produce accurate results,[35] potentially minimizing medical errors and aiding clinicians.[36]

Given the role of metal oxide NPs as CAs and importance of algorithms in medical image analysis, developing an algorithm to study NP signals in MRI is essential. This study investigates the relationship between different metal oxide NP concentrations and relaxation times, hypothesizing the following. (1) Various metal oxides affect signal intensity, (2) different metal oxide NP concentrations alter signal intensity, and (3) metal oxide NPs influence the longitudinal relaxation time in MRI. Moreover, we present an algorithm to analyze the signal intensity and autonomously determine relaxation times in MRI using metal oxide NPs.

2. Methods

2.1. Chemicals and reagents

Five distinct NPs were synthesized using the sol–gel method, a bottom-up chemical approach enabling enhanced control over procedural steps and the chemical compositions of the final products.[37] All reagents were sourced from Sigma-Aldrich, including cobalt(II) nitrate hexahydrate ($Co(NO_3)\cdot 6H_2O$, 98%), copper(II) nitrate tetrahydrate ($Cu(NO_3)_2\cdot 3H_2O$, 99%), iron(III) nitrate nonahydrate ($Fe(NO_3)_3.9H_2O$, 98%), nickel(II) nitrate hexahydrate ($Ni(NO_3)_2\cdot 6H_2O$, 97%), and

zinc oxide (ZnO, 99%). Additional reagents employed were ethylene glycol, citric acid, and nitric acid.

2.2. Synthesis of NPs

To synthesize Co_3O_4 NPs, $Co(NO_3)\cdot 6H_2O$ and urea were separately dissolved in 30 mL Milli-Q water at a molar ratio of 1:1, following which the solutions were combined and maintained at 30°C for 2 h for homogenization. The mixed solution was then heated to 80°C with vigorous stirring until complete solvent evaporation. Co_3O_4 NPs were finally obtained post-calcination at 400°C.[38]

Copper oxide (CuO) NPs, in the form of a Cu_2O/CuO nanocomposite, were prepared by dissolving 5 g $Cu(NO_3)_2\cdot 3H_2O$ in 20 mL ethylene glycol. After stirring the solution for 1 h, it was allowed to form a gel over 24 h, followed by drying at 200°C and calcination at 300°C for 1 h each. A final heat treatment was performed at 500°C for 1 h.

Fe_2O_3 NPs were synthesized by dissolving 7.2 g $Fe(NO_3)_3\cdot 9H_2O$ in 200 mL Milli-Q water and 32.6 g citric acid in 800 mL Milli-Q water. The iron nitrate solution was gradually added to the citric acid solution under constant stirring. The mixture was then heated to 90°C until gel formation, dried in an oven at 100°C for 24 h, and calcined at 400°C for 2 h. The resulting gel was ground into a powder.[39]

Nickel oxide (NiO) NPs were prepared by dissolving 3 g $Ni(NO_3)_2\cdot 6H_2O$ in 100 mL Milli-Q water. To this solution, 0.5 M sodium hydroxide solution was added drop-wise under continuous stirring until the pH reached 11, at which point precipitates were formed. These precipitates were washed 5 times with Milli-Q water, dried at 95°C to completely remove the solvent, and finally calcined at 550°C for 3 h.[40]

Zinc oxide (ZnO) NPs were synthesized using amorphous ZnO powder. Initially, 100 mL Milli-Q water and 15 mL nitric acid were mixed using a magnetic stirrer at 90°C, following which 5.5 g amorphous ZnO was added gradually. A secondary solution containing 190 mL deionized water and 5.5 g citric acid was prepared and mixed with the first solution. After 15 min of stirring and adding 10.5 mL ethylene glycol, the mixture was maintained at 290°C until reaching a basic pH. The temperature was then lowered to 180°C and subsequently to 70°C until the solvent was evaporated. Post-crystallization, the product was dried for 2 h at 350°C and further heated at 500°C for 30 min.[41]

2.3. Characterization of the NPs

2.3.1. X-ray diffraction (XRD)

XRD was used to determine the crystalline structures and phases of the NPs by comparing the results with the entries in the Joint Committee on Powder Diffraction Standards (JCPDS) database.[42] The crystallite sizes of the NPs were estimated using the Scherrer equation.[43] This size estimation was also conducted by analyzing XRD peaks, employing the Debye–Scherrer equation.[43] XRD spectra were acquired using a D/MAX-2100/PC (Rigaku) apparatus, equipped with a Cu Kα radiation source (λ = 1.5418 Å). The spectra were scanned in a 2θ range of 20 – 80°, with a scanning speed of 2°/min, a step size of 0.02°/min, and operating conditions of 40 kV and 20 mA.

2.3.2. Fourier-transform infrared spectroscopy (FTIR)

FTIR spectroscopy was performed to analyze the functional groups and molecular bonds within the NPs, facilitating the detection of compositional changes in the samples. This analysis was performed using a Vertex 70-Bruker spectrometer, supported by a diamond crystal. Spectra were acquired in the infrared region using the attenuated total reflectance method, with a scanning range of 3000 – 400 cm^{-1}, comprising 32 scans at a resolution of 4 cm^{-1}.

2.4. MRI

An acrylic phantom, simulating a brain, was filled with paramagnetic aqueous solutions to mimic different biological tissues, such as gray matter, white matter, and cerebrospinal fluid.[44] To assess changes in MRI signal intensity and relaxation time, four distinct concentrations of NPs (Table 1) were prepared and added to specific compartments of the phantom. The effects of these varying NP concentrations on MRI signal characteristics were subsequently evaluated.

The acquisition of MRIs followed a protocol recommended by the Consortium of MS Centers, tailored specifically for patients with MS. These images were acquired using a 3.0 Tesla Siemens Verio MRI scanner. To investigate the impact of varying NP concentrations on the MRI signals, signal quantification was performed across three different imaging sequences: *T1*-w, *T2*-w, and FLAIR. Furthermore, a study was performed to assess the influence of echo time (*TE*) variations on signal intensity, using TEs of 11, 32, 43, 64, and 86 ms. Figure 1 illustrates the phantom infused with different concentrations of the five metallic oxide NPs (Figure 1A) and a representative MRI slice of the phantom (Figure 1B).

Table 1. Concentrations of nanoparticles (NPs) in the phantom

Hole	Co_3O_4 (g/L)	CuO (g/L)	Fe_2O_3 (g/L)	NiO (g/L)	ZnO (g/L)
1	0.20	0.27	0.11	0.35	0.19
2	0.49	0.65	0.53	0.63	0.52
3	1.39	2.07	1.83	1.97	1.41
4	3.59	3.29	3.33	3.59	3.32

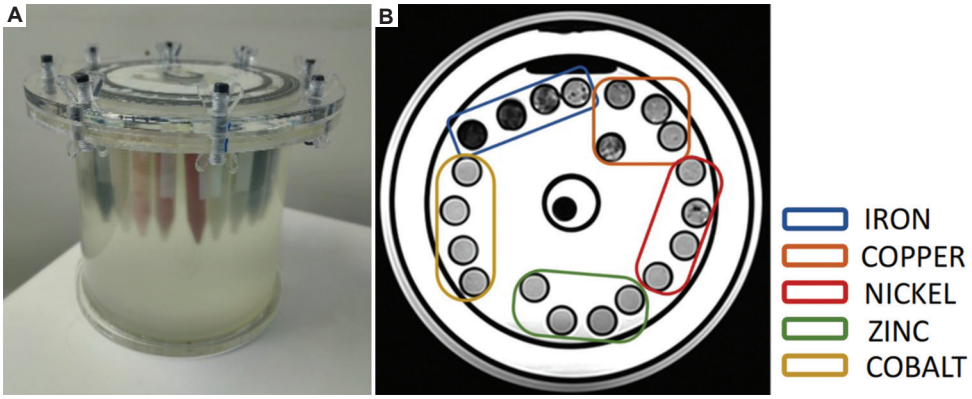


Figure 1. Phantom and its respective magnetic resonance imaging (MRI) for signal intensity analysis of the metal oxide nanoparticles (NPs). (A) Phantom used in the experiment. (B) MRI slice, with various metal oxide NPs being marked using different colors

2.5. Manual evaluation: MRI signal quantification and relaxation time evaluation

The MRI scans, obtained in the digital imaging and communications in medicine (DICOM) format, were manually analyzed using Image J and 3D Slicer software. A quantitative assessment of the mean signal intensity was performed for each compartment of the phantom, across various concentrations and *TE*. The manual calculation of *T1* relaxation time was executed using Equation I:

$$Y = A - Be^{\left(-\frac{TI}{T1}\right)} \quad \text{(I)}$$

where Y denotes the signal intensity, TI represents the inversion time, A signifies a scaling factor for each signal intensity, and B denotes the quality of inversion.[45] Furthermore, *T2* relaxation times were determined by fitting the *ET* signal curve using Equation II:

$$S(TE) = S_{máx}e^{\frac{TE}{T2}} + S_n, \quad \text{(II)}$$

where S_n corresponds to the noise level in the image.[45]

2.6. Automatic evaluation: Algorithm development

The MRI images, obtained in the DICOM format, underwent preprocessing in three stages: (I) Reconfiguration to 1-mm^3 resolution, (II) application of an anisotropic diffusion filter, and (III) intensity correction for magnetic field inhomogeneity.[11] First, *T1*-weighted and *T2*-weighted images were resliced to an isotropic resolution of 1 mm^3 using cubic spline interpolation. Second, an anisotropic diffusion filter was employed to mitigate potential noise in the images.[46,47] Finally, image homogeneity was enhanced through bias correction using an N4ITK filter.[48]

After preprocessing, the algorithm segmented the region of interest (ROI) for the automatic determination of the longitudinal relaxation time (*T1*). Following this ROI segmentation, the algorithm autonomously calculated *T1* based on Equation I. For this, the algorithm detected and stored the signal intensity value Y by treating A and B as constants, and the TI value was input by the operator. Using these parameters, the algorithm automatically calculated and outputted the *T1* value. To validate and ensure the reproducibility of quantification, a Bland–Altman plot was created to compare automated *T1* quantification, with the manual approach using Equation I and Image J.[49,50]

3. Results

3.1. Characterization of the NPs

The chemical structures, crystal lattice indices, and crystallite sizes of the five NPs were determined using XRD and FTIR analyses. Figure 2 illustrates the combined chemical (FTIR) and structural (XRD) results of all NPs. The characteristic absorption bands of the metal-oxygen bond[51] were prominently observed at 1500 – 400 cm^{-1}. Furthermore, in all FTIR spectra, absorption bands were consistently observed at 2341 and 2358 cm^{-1}, likely attributable to atmospheric CO_2 absorption on metallic cations, a phenomenon that may have occurred within the apparatus during analysis.

The XRD spectrum of Co_3O_4 NPs in Figure 2A reveals distinct peaks at 2θ values of 31.45°, 37.17°, 38.79°, 45.10°, 59.72°, 65.54°, and 74.43°, corresponding to the lattice planes (220), (311), (222), (400), (511), (440), and (620), respectively. These peaks confirm the cubic phase of Co_3O_4 (JCPDS: 65-3103), with lattice parameters $a = b = c = 8.056$ Å.[38,52] The FTIR spectrum of these NPs in Figure 2B displays two vibrational modes at 667 and 561 cm^{-1}, indicative of Co–O bonds.[53,54]

Figure 2C presents the XRD spectrum of CuO NPs, where 2θ peaks aligned with the lattice planes (110), (111), (−112), (−202), (020), (202), (−113), (−311), (113), (311),

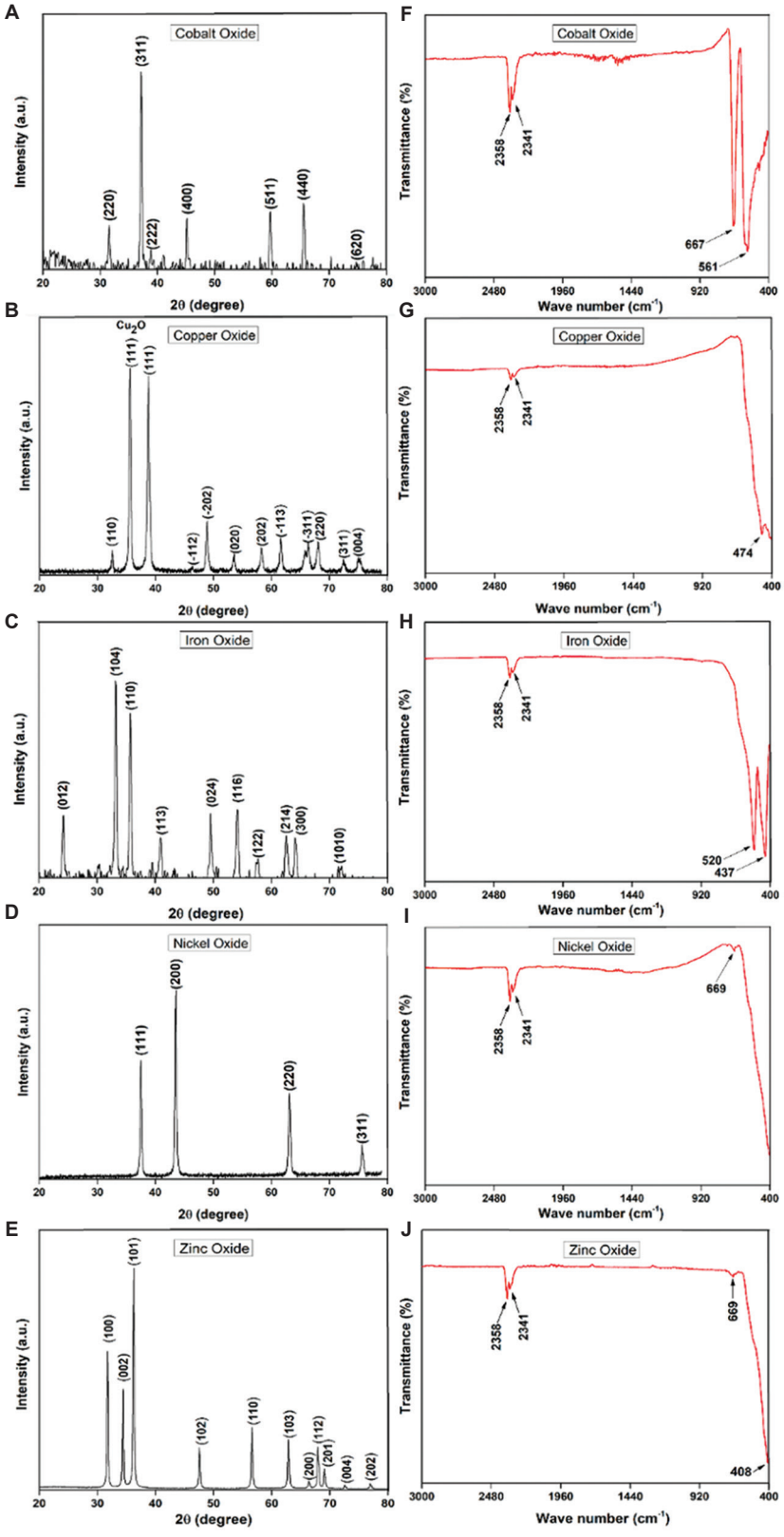


Figure 2. Characterization of nanoparticles by X-ray diffraction (XRD) and Fourier-transform infrared (FTIR). XRD patterns of (A) Co_3O_4, (C) CuO, (E) Fe_2O_3, (G) NiO, and (I) ZnO. FTIR spectra of (B) Co_3O_4, (F) Fe_2O_3, (H) NiO, and (J) ZnO

and (004). An additional peak at 35.58° corresponded to the (111) plane of the Cu_2O phase. These peaks were consistent with the monoclinic phase of CuO (JCPDS: 89-5895), with lattice parameters a = 4.682 Å, b = 3.424 Å, and c = 5.127 Å.[55] The FTIR spectrum of these NPs in Figure 2D presents a band at 474 cm^{-1}, denoting Cu–O formation.[56,57]

Figure 2E presents the XRD pattern of the Fe_2O_3 NPs, with peaks corresponding to rhombohedral Fe_2O_3 (JCPDS: 79-0007) and lattice parameters $a = b$ = 5.0285 Å and c = 13.7360 Å.[39,58,59] The FTIR spectrum of these NPs in Figure 2F reveals absorption bands at 520 cm^{-1} and 437 cm^{-1}, attributed to the Fe–O stretching and bending modes in hematite.[60]

In the XRD spectrum of NiO NPs displayed in Figure 2G, the peaks at 37.50°, 43.50°, 63.05°, and 75.58° corresponded to the lattice planes (111), (200), (220), and (311) of the cubic phase of NiO (JCPDS: 01-1239), with lattice parameters $a = b = c$ = 4.1710 Å.[40,61] The FTIR spectrum of these NPs in Figure 2H presents a prominent absorption band at 669 cm^{-1}, signifying Ni–O formation.[62]

In the XRD spectrum of ZnO NPs displayed in Figure 2I, the diffraction peaks aligned with the lattice planes (100), (002), (101), (102), (110), (103), (200), (112), (201), (004), and (202) of hexagonal ZnO (JCPDS: 65-3411), with lattice parameters $a = b$ = 3.249 Å and c = 5.206 Å.[63,64] The FTIR spectrum of these NPs in Figure 2J illustrates a band at 408 cm^{-1}, corresponding to Zn–O bonding, and a band at 669 cm^{-1}, ascribed to C–H stretching in the alkyne group.[65,66]

The average crystallite size was estimated using the Debye–Scherrer equation

$$d = 0.89\lambda/\beta\cos\theta, \quad \text{(III)}$$

where 0.89 is the Debye constant, λ represents the X-ray wavelength (1.5406 Å), β denotes the full-width at half-maximum of the peak, and θ represents the Bragg angle. The estimated average crystallite sizes for Co_3O_4, CuO, Fe_2O_3, NiO, and ZnO NPs were 12, 26, 21, 38, and 25 nm, respectively.

3.2. Manual evaluation: MRI signal and relaxation time evaluation

The acquired MRIs were manually evaluated to quantify the mean signal intensity in each compartment of the phantom across varying NP concentrations. Figure 3 illustrates the signal intensity as a function of the NP concentration for the *T1*, *T2*, and FLAIR sequences. Circles in Figure 3 represent the cross-sections of each compartment from which the mean pixel intensity was extracted. Notably, different NPs altered the MRI signal characteristics, with Fe_2O_3 NPs exerting particularly notable effects.

Figure 4 presents a series of curves demonstrating the relationship between signal intensity and NP concentrations, as well as the dependency of signal intensity on *TE* for these NPs in MRI. The results revealed that the pixel signal intensity is inversely proportional to the *ET* and NP concentration, a trend that was consistent across all the tested NPs. As the *ET* increased, a reduction in signal intensity was observed for all NPs. Furthermore, this reduction in signal intensity was more pronounced for NP concentrations with greater metal concentrations. Moreover, a steeper slope in the *T2* relaxation curve correlated with a greater decrease in the signal intensity, enhancing the effectiveness of the NPs as *T2* negative CAs.

In addition, Figure 4A-E illustrates variations in the signal intensities for the *T1*, *T2*, and FLAIR scan sequences as a function of the metal NP concentration. All metal oxide NPs exhibited a decrease in the signal intensity with increasing NP concentration across all three sequences consistently. Notably, Fe_2O_3 NPs exhibited a more pronounced signal reduction, while ZnO NPs demonstrated a slightly divergent behavior in the FLAIR sequence. Consequently, a steeper relaxation curve slope correlated with increased signal reduction, enhancing the efficacy of metal oxide NPs as CAs.

The magnitude of the *T2* contrast effect was quantitatively represented by spin–spin relaxivity R_2; an increase in the R_2 values indicated a corresponding increase in the contrast effect. The relaxation rate R_2 (=1/*T2*) is plotted against the *ET* in Figure 5A. An analysis of these curves revealed that NP concentrations directly impact the relaxivity time, thereby influencing the pixel intensity. In Figure 5B-D, which corresponds to the *T1*, *T2*, and FLAIR MRI sequences, respectively, the pixel intensity trends for each metal oxide NP were distinctly observed in relation to their concentrations. In particular, Fe_2O_3 NPs (depicted by the blue curve) demonstrated the most significant intensity variation across all sequences, showing a notable intensity decrease with increasing concentrations, which highlights their substantial impact on MRI imaging. Conversely, Co_3O_4 NPs (black curve) exhibited a more gradual decline, suggesting a less pronounced impact of their concentrations on MRI signal intensities. Other metal oxide NPs, including CuO (red curve), NiO (green curve), and ZnO (purple curve), also exhibited reductions in the pixel intensities with increasing concentrations; however, these changes vary, reflecting the distinct reactivity of each metal oxide NP in the MRI sequences. These differences underscore the importance of considering the unique properties of different metal oxide NPs when employing them as CAs in MRI to optimize image acquisition and interpretation.

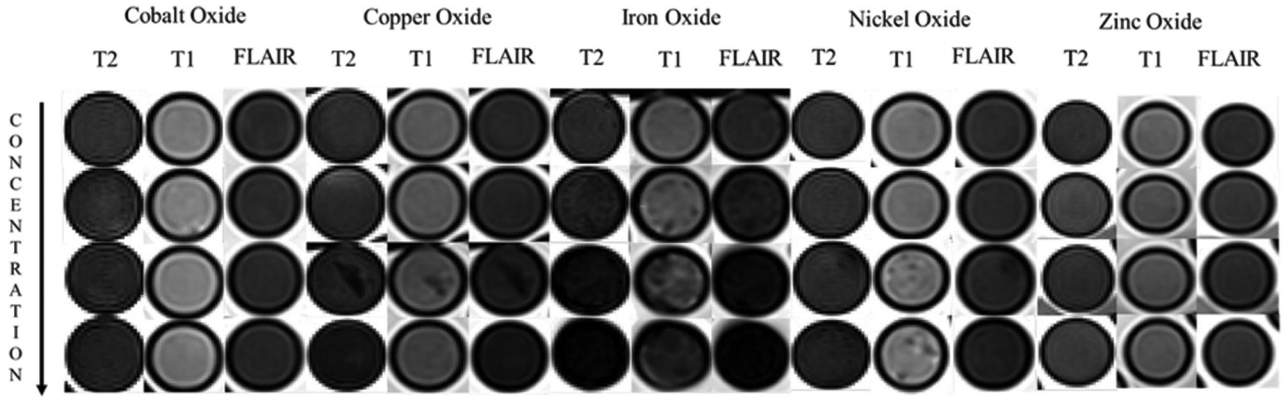


Figure 3. Magnetic resonance imaging: signal intensity as a function of nanoparticle concentration

3.3. MRI preprocessing, segmentation, quantification, and validation

Figure 6 offers a comprehensive and sequential overview of the entire MRI image analysis protocol. During the preprocessing stage, the images were subjected to three bias correction steps, uniformly applied to all slices within each specific examination sequence. An operator facilitated the selection of the particular slice for the ROI segmentation using a specialized algorithm designed to load all slices from the examination. Following this initial selection, meticulous segmentation of the ROIs was executed on the chosen slice. The subsequent stage involved quantifying the longitudinal relaxation time. This was accomplished by utilizing the signal intensity data extracted from the segmented ROIs in conjunction with Equation I. The final step of the algorithm computed and stored the average signal intensity for each of the segmented ROIs, thereby completing the intricate process of MRI image analysis.

To assess the reproducibility of our algorithm and validate it, a comparative analysis was conducted between the automatically computed *T1* values (derived from our algorithm) and manually determined *T1* values by experts. This comparison utilized the Bland–Altman plot (Figure 7A). The results revealed a close correspondence between the *T1* values obtained through the automated and manual methods, indicating a high degree of concordance. This was quantitatively supported by a correlation coefficient (r) of 0.9977, as illustrated in Figure 7B, thereby validating the accuracy and reliability of the algorithm in *T1* quantification.

4. Discussion

4.1. Signal intensities of the NPs and algorithmic analysis for MRI

Recent investigations into the signal intensities of NPs in MRI have highlighted notable advancements in the utilization of metal oxides, such as those of cobalt, copper, iron, nickel, and zinc, as CAs. Our findings indicated a decrease in pixel intensity across the *T1*, *T2*, and FLAIR sequences with increasing concentrations of these metal oxide NPs, suggesting enhanced proton relaxation. This effect was particularly prominent in the *T2* sequence, underscoring the potential of these NPs in influencing *T2* relaxation times and their effectiveness as CAs.

Further analysis of the effective *TE* graphs corroborated that all metallic oxide NPs under study resulted in decreased pixel intensity as *TE* increased. This characteristic aligns with the desirable attributes of CAs, highlighting the ability of these NPs to alter proton relaxation times in adjacent tissues. Notably, due to the different in local magnetic field disortions, each metal oxide NP exhibited different perturbations in the MRI signal, reflecting their different efficacies as CAs.

Our study demonstrated that different metallic oxide NPs interfere with the MRI signal intensity and variations in their concentrations alter the signal intensity and relaxation time, thus confirming our hypothesis. Specifically, varying concentrations of the Fe_2O_3 NPs displayed significant variations in both signal intensity and relaxation time, resulting in high contrast. This behavior was expected due to the magnetic properties of Fe. Furthermore, the Fe_2O_3 NPs exhibited CA characteristics in *T2*-w images, with higher concentrations resulting in lower signal intensities in *T1*-w sequences.

NPs synthesized in this study altered relaxation times in MRI, thereby modifying the pixel signal intensity. Our results revealed that these NPs displayed negative contrast, a characteristic influenced by the particle size. Specifically, NPs increased the signal intensity in *T1*-w images while decreasing the contrast intensity in *T2*-w images. This behavior is attributed to the fact that under a magnetic field, a magnetic dipole moment is induced

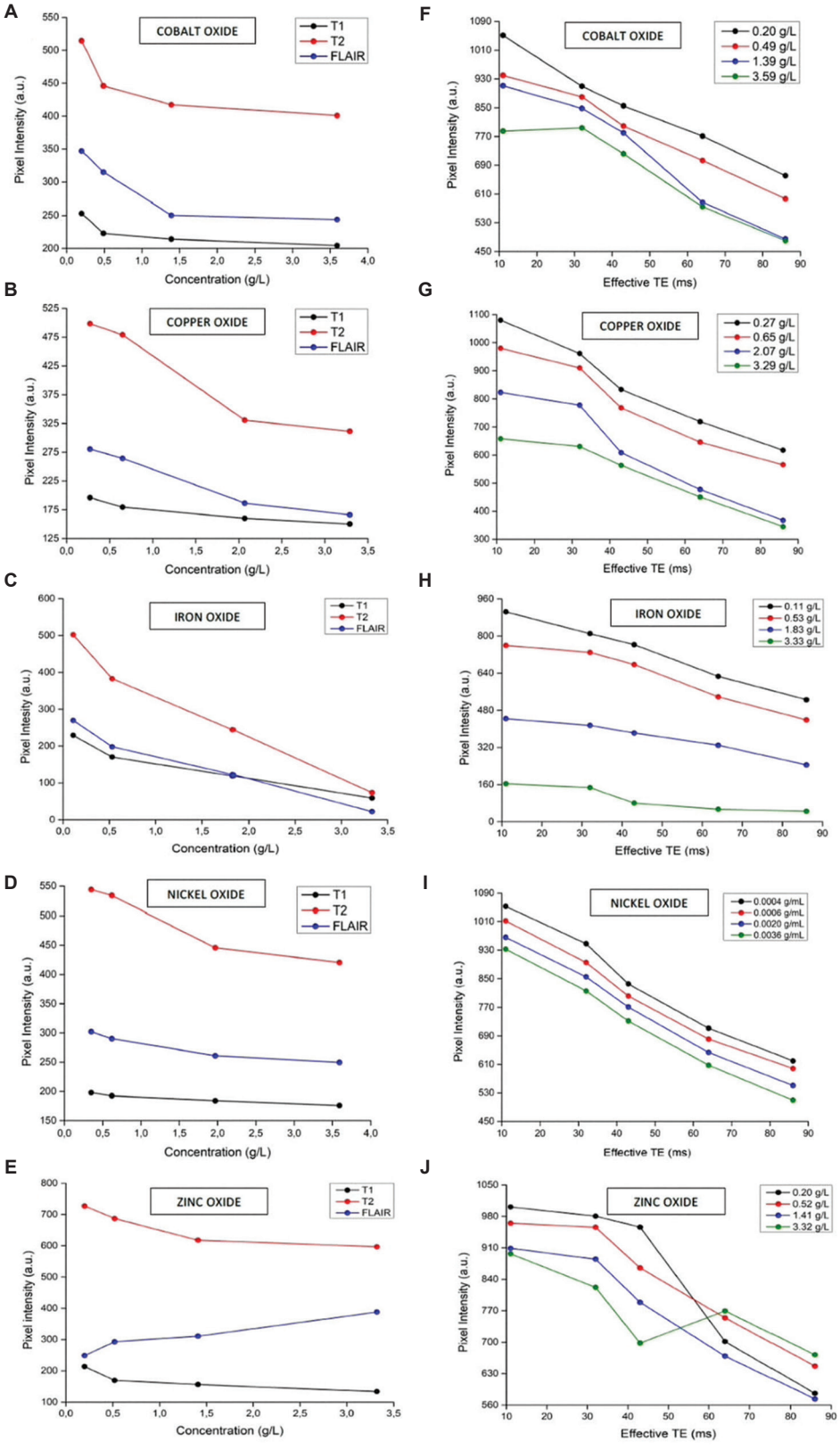


Figure 4. Signal intensity as a function of metal oxide nanoparticle concentration: (A) Co_3O_4, (C) CuO, (E) Fe_2O_3, (G) NiO, and (I) ZnO. Signal intensity as a function of *TE*: (B) Co_3O_4, (D) CuO, (F) Fe_2O_3, (H) NiO, and (J) ZnO

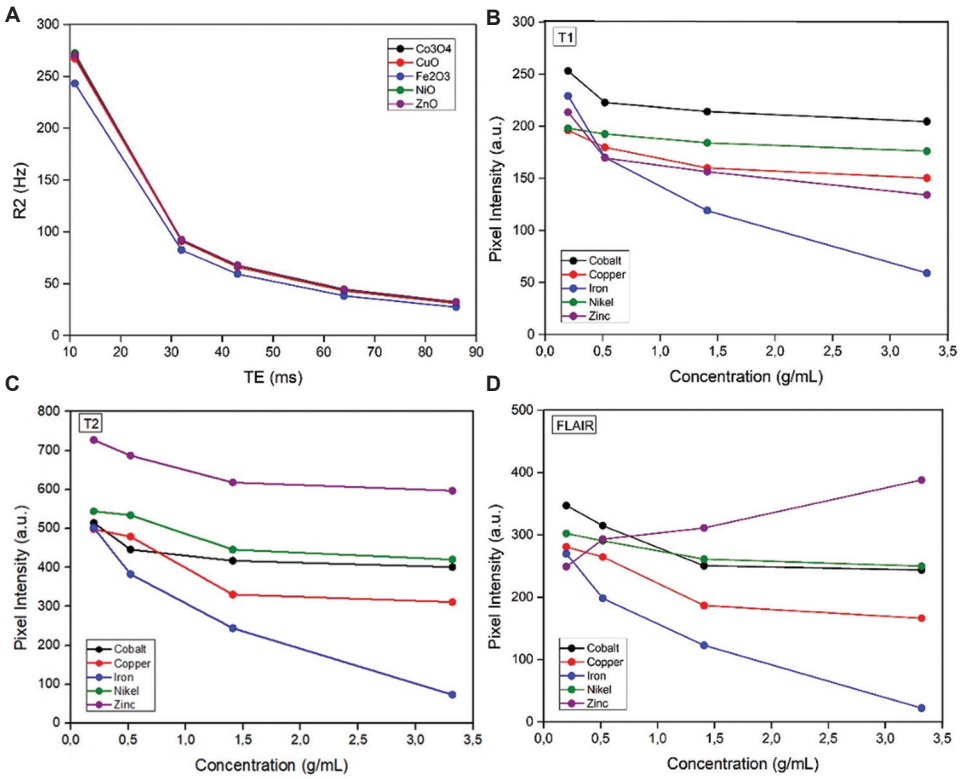


Figure 5. Comparison between different metal oxide nanoparticles for relaxation rate, T1, T2, and fluid-attenuated inversion recovery (FLAIR). (A) Relaxation rate R_2 as a function of echo time, (B) concentration as a function of *T1*, (C) concentration as a function of *T2*, and (D) concentration as a function of FLAIR for different NPs

in superparamagnetic NPs.[20] As water molecules diffuse toward the outer sphere of the induced dipole moment, the magnetic relaxation processes of water protons are disturbed, decreasing the spin-spin relaxation time (*T2*).[53] This disturbance causes darkening in the *T2*-w MRI images, corroborating our findings. In addition, Chen *et al.* (2022)[67] showed that the dimensions of nanomaterial's influence contrast, with NPs larger than 12 nm producing negative contrast. Our synthesized NPs, particularly Fe_2O_3 NPs, exhibited such a negative contrast characteristic, wherein higher concentrations resulted in darker areas. The signal reduction effect observed with increasing Fe_2O_3 NP concentration may be attributed to the disturbance of local magnetic field homogeneity caused by Fe_2O_3-core NPs.[68] This perturbation causes protons to lose energy owing to spin-spin interactions in an aqueous medium, increasing the loss of coherence and consequently decreasing the T2 time.[68]

As anticipated, Fe_2O_3 NPs substantially impacted the MRI signal intensity and longitudinal relaxation time. However, the other metal oxide NPs also demonstrated potential for specific applications based on their signal-altering properties in MRI.

Simultaneously, our algorithm for MRI analyses offers a transformative approach to MRI image interpretation. Traditional visual analysis by specialists often faces challenges with respect to subtle variations in the signal intensities associated with relaxation times. However, the implemented algorithm enables the automatic quantification of the longitudinal relaxation time, effectively overcoming these limitations. This tool segments different ROIs, analyzes signal intensity, and quantifies the longitudinal relaxation time for various NPs.

The reproducibility analysis of this method, validated against manual quantitative evaluations, revealed that the algorithm effectively assists specialists in identifying subtle variations in the signal intensity. This capability is particularly valuable when using different metal oxide NPs NPs as CAs. For example, Fe_2O_3 NPs exhibited substantial variations in the signal intensity and relaxation time, yielding high contrast, which aligned with the Fe magnetic properties. Moreover, the synthesized NPs altered the relaxation time in MRI, modifying the pixel signal intensity and displaying negative contrast influenced by the particle size.[67]

This algorithmic approach is particularly valuable in situations wherein changes in the signal intensity

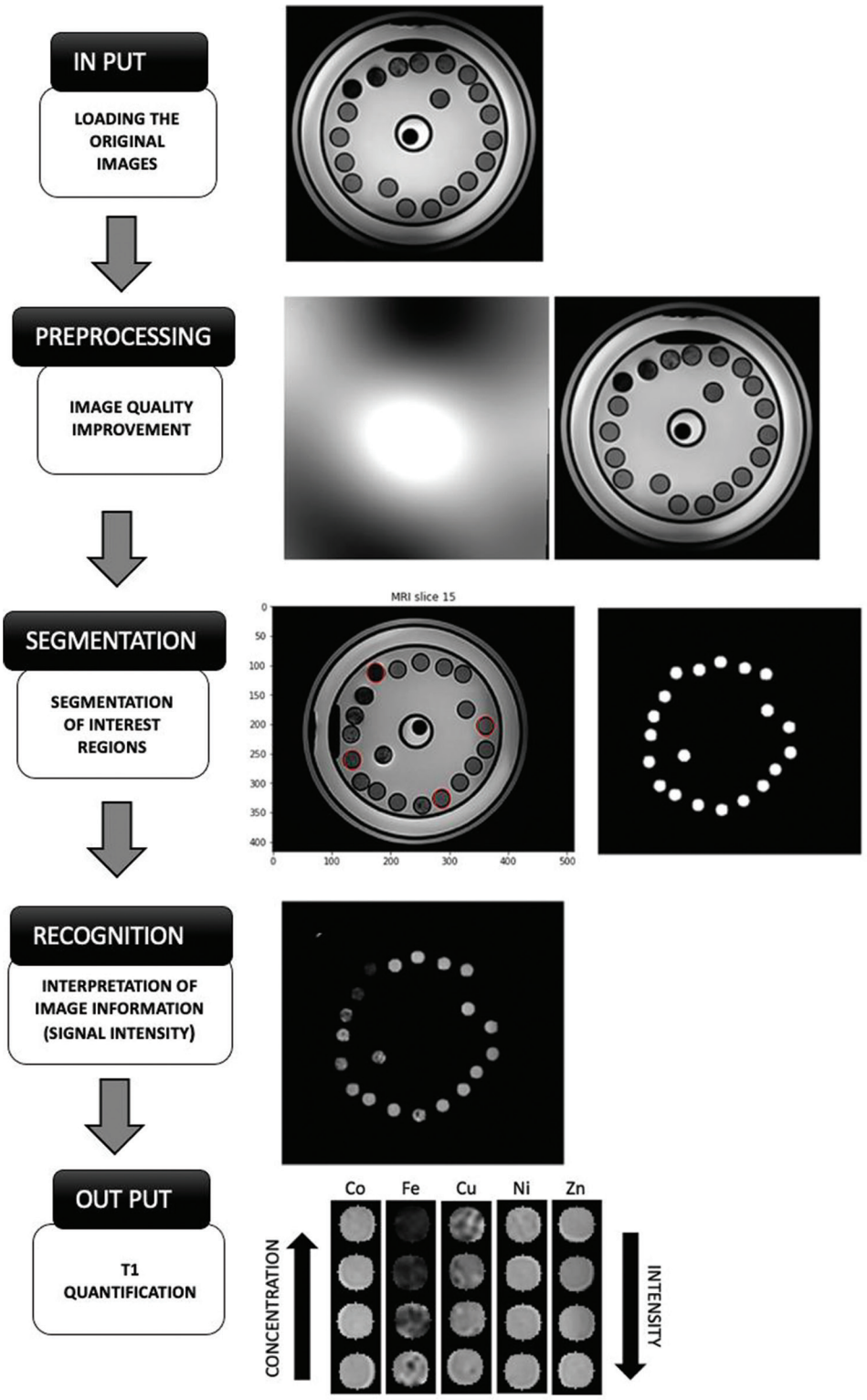


Figure 6. Flowchart of the proposed method. Input: Original magnetic resonance imaging. Processing: application of the bias correction filler. Segmentation: segmentation of all region of interests (ROIs). Recognition: signal intensity of all ROIs. Output: Quantification of the signal intensity for *T1* as a function of the metal oxide nanoparticle concentration

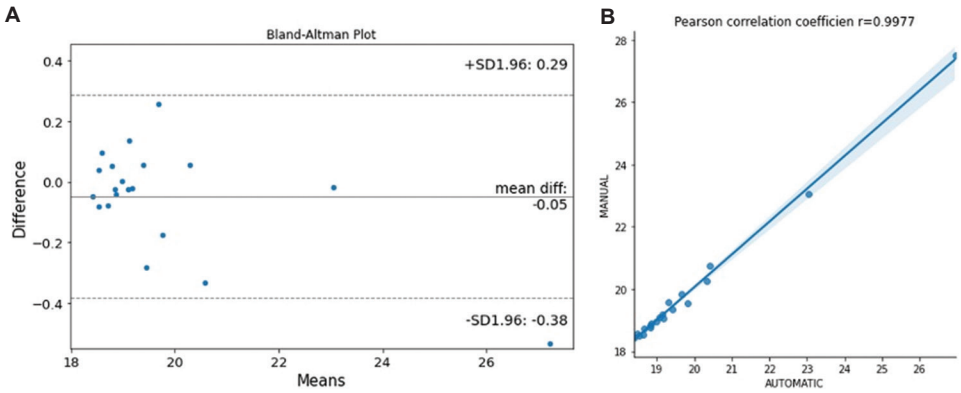


Figure 7. Validation. (A) Bland-Altman and (B) scatter plots comparing the *T1* values obtained from the automatic and manual quantification methods

are imperceptible to the naked eye. By automating the quantification process, this strategy enables a more precise and nuanced understanding of the impacts of different metal oxide NP concentrations on the signal intensity and relaxation time. The integration of our empirical data on the NP signal intensity, with the advancements facilitated by this algorithm, presents significant potential for MRI applications. This combined approach not only enhances the visual representation of anatomical structures and facilitates the detailed detection and characterization of diseases but also establishes a groundbreaking method for the quantitative assessment of signal intensity variations. Furthermore, this innovation represents a paradigm shift in imaging techniques, setting the stage for substantial advancements in the field of diagnostic medicine.

4.2. NPs as potential biomarkers

Studies on the effects of metal NPs on MRI signal intensities, particularly in the context of neurodegenerative processes, have garnered considerable attention. This is attributed to challenges encountered in identifying toxic metals that contribute to pathological changes in brains with neurodegenerative lesions. In addition, studies that have monitored cerebral lesions and analyzed metal elements involved in disease progression remain notably limited.

Existing studies indicate that metal elements can be detected in the brains of individuals with neurodegenerative lesions.[69-73] This suggests that the timing of image acquisition is crucial for patients as early active or acute lesions can influence the signal intensity. For instance, Tham *et al.* (2021)[69] displayed that early lesions contain substantially higher metal concentrations than acute lesions. If metal elements indeed impact the signal intensity in lesion regions, our study demonstrated that different metals can serve as biomarkers for monitoring brain lesions in patients with MS during disease progression.

Although the impact of NPs on the relaxation time may often be imperceptible in the visual analysis of the signal intensity due to subtle differences in contrast, the proposed automatic detection algorithm facilitates their use as biomarkers. Indeed, the reduction in the signal intensity with increasing NP concentration observed across the three image sequences (*T1*, *T*2, and FLAIR) in Figure 4 indicates that the largest differences occur at lower NP concentrations, where contrast changes are minimal. Hence, future research should focus on investigating NP signals at low concentrations and exploring the mechanism by which these metals relate to brain lesion progression in longitudinal studies.

4.3. Potential for clinical applications and diagnosis of neurodegenerative diseases

Our results offer new insights into the use of MRI for clinical applications, particularly for the detection and monitoring of brain lesions and neurodegenerative diseases. Notably, the contrast afforded by NPs, along with quantification achieved through our algorithm, enhances diagnostic capabilities. The ability to modulate contrast and signal intensities with different types of NPs can be beneficial across various medical diagnoses.

Despite these promising results, our study has some limitations. The nanometer scale of materials is highly sensitive, and the handling of nanomaterials requires meticulous attention, which complicates operations in low concentration ranges. Future research focusing on low concentration ranges may yield better correlations with neurodegenerative disease levels. In addition, given that the observed signals do not differentiate between the types of metals, signal specificity remains a limitation.

5. Conclusions

This study demonstrates that various concentrations of metallic NPs considerably influence MRI signal intensity,

impacting longitudinal relaxation time. In addition, we present an algorithm to analyze signal intensity and automatically determine relaxation times in MRI using metal oxide NPs. This innovative quantification holds considerable potential for enhancing treatment monitoring and routine clinical assessments in MRI analyses, particularly through the use of metallic oxide NPs as CAs. In contrast to the existing literature that predominantly focuses on a single type of NP, such as Fe_2O_3 NPs, our research encompasses a broader spectrum of metallic NPs, including Co_3O_4, CuO, Fe_2O_3, NiO, and ZnO NPs. This comprehensive exploration offers deeper insights into the versatile roles of different metallic NPs as CAs in MRI. Our study represents a substantial interdisciplinary achievement, integrating aspects of chemistry, physics, materials science, and medicine. It lays the groundwork for future innovations in MRI technology, particularly in tailoring CAs to meet the diverse requirements of clinical applications.

Acknowledgments

None.

Funding

This study was funded by the Brazilian agency Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP 2020/03022-9, 2019/16362-5 and 2017/20032-5).

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Marcela de Oliveira, Marina Piacenti-Silva
Formal analysis: Hulder Henrique Zaparoli, Marcela de Oliveira, Marina Piacenti-Silva
Investigation: All authors
Methodology: Daniela Gomes Bornal, Hulder Henrique Zaparoli
Writing – original draft: Daniela Gomes Bornal, Hulder Henrique Zaparoli, Marcela de Oliveira
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The data utilized in this study are available from the corresponding author on reasonable request.

References

1. Carr DH, Gadian DG. Contrast agents in magnetic resonance imaging. *Clin Radiol*. 1985;36(6):561-568.
doi: 10.1016/S0009-9260(85)80234-8
2. Bakshi R. Magnetic resonance imaging advances in multiple sclerosis. *J Neuroimaging*. 2005;15(4 Suppl):10-14.
doi: 10.1177/1051228405283362
3. Valverde S, Cabezas M, Roura E, *et al.* Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. *Neuroimage*. 2017;155:159-168.
doi: 10.1016/j.neuroimage.2017.04.034
4. Compston A, Coles A. Multiple sclerosis. *Lancet*. 2008;372(9648):1502-1517.
doi: 10.1016/S0140-6736(08)61620-7
5. Wang KY, Chetta J, Bains P, *et al.* Spectrum of MRI brain lesion patterns in neuromyelitis optica spectrum disorder: A pictorial review. *Br J Radiol*. 2018;91(1086):20170690.
doi: 10.1259/bjr.20170690
6. Zhang Y, Yang H, Zhou Z, Huang K, Yang S, Han G. Recent advances on magnetic relaxation switching assay-based nanosensors. *Bioconjug Chem*. 2017;28(4):869-879.
doi: 10.1021/acs.bioconjchem.7b00059
7. Na HB, Song IC, Hyeon T. Inorganic nanoparticles for MRI contrast agents. *Adv Mater*. 2009;21(21):2133-2148.
doi: 10.1002/adma.200802366
8. Lee N, Hyeon T. Designed synthesis of uniformly sized iron oxide nanoparticles for efficient magnetic resonance imaging contrast agents. *Chem Soc Rev*. 2012;41(7):2575-2589.
doi: 10.1039/C1CS15248C
9. Bakshi R, Thompson AJ, Rocca MA, *et al.* MRI in multiple sclerosis: Current status and future prospects. *Lancet Neurol*. 2008;7(7):615-625.
doi: 10.1016/S1474-4422(08)70137-6
10. de Oliveira M, Gianeti TMR, da Rocha FCG, Lisboa-Filho PN, Piacenti-Silva M. A preliminary study of the concentration of metallic elements in the blood of patients with multiple sclerosis as measured by ICP-MS. *Sci Rep*. 2020;10(1):13112.
doi: 10.1038/s41598-020-69979-9
11. de Oliveira M, Piacenti-Silva M, da Rocha FCG, Santos JM, Cardoso JS, Lisboa-Filho PN. Lesion volume quantification using two convolutional neural networks in MRIs of multiple sclerosis patients. *Diagnostics* (*Basel*). 2022;12(2):230.
doi: 10.3390/diagnostics12020230
12. Giacoppo S, Galuppo M, Calabrò RS, *et al.* Heavy metals

and neurodegenerative diseases: An observational study. *Biol Trace Elem Res*. 2014;161(2):151-160.

doi: 10.1007/s12011-014-0094-5

13. Forte G, Alimonti A, Pino A, *et al.* Metals and oxidative stress in patients with Parkinson's disease. *Ann Ist Super Sanita*. 2005;41(2):189-195.

14. Roos PM, Vesterberg O, Syversen T, Flaten TP, Nordberg M. Metal concentrations in cerebrospinal fluid and blood plasma from patients with amyotrophic lateral sclerosis. *Biol Trace Elem Res*. 2013;151(2):159-170.

doi: 10.1007/s12011-012-9547-x

15. Squadrone S, Brizio P, Abete MC, Brusco A. Trace elements profile in the blood of Huntington' disease patients. *J Trace Elem Med Biol*. 2020;57:18-20.

doi: 10.1016/j.jtemb.2019.09.006

16. Xiao YD, Paudel R, Liu J, Ma C, Zhang ZS, Zhou SK. MRI contrast agents: Classification and application (Review). *Int J Mol Med*. 2016;38(5):1319-1326.

doi: 10.3892/ijmm.2016.2744

17. Liu Z, Zhao M, Wang H, *et al.* High relaxivity Gd3+-based organic nanoparticles for efficient magnetic resonance angiography. *J Nanobiotechnology*. 2022;20(1):170.

doi: 10.1186/s12951-022-01363-3

18. Elhabiri M, Abada S, Sy M, *et al.* Importance of outer-sphere and aggregation phenomena in the relaxation properties of phosphonated gadolinium complexes with potential applications as MRI contrast agents. *Chemistry*. 2015;21(17):6535-6546.

doi: 10.1002/chem.201500155

19. Zheng R, Guo J, Cai X, *et al.* Manganese complexes and manganese-based metal-organic frameworks as contrast agents in MRI and chemotherapeutics agents: Applications and prospects. *Colloids Surfaces B Biointerfaces*. 2022;213:112432.

doi: 10.1016/j.colsurfb.2022.112432

20. Jun YW, Lee JH, Cheon J. Chemical design of nanoparticle probes for high-performance magnetic resonance imaging. *Angew Chem Int Ed Engl*. 2008;47(28):5122-5135.

doi: 10.1002/anie.200701674

21. Cai X, Zhu Q, Zeng Y, Zeng Q, Chen X, Zhan Y. Manganese oxide nanoparticles as MRI contrast agents in tumor multimodal imaging and therapy. *Int J Nanomedicine*. 2019;14:8321-8344.

doi: 10.2147/IJN.S218085

22. Olchowy C, Cebulski K, Łasecki M, *et al.* The presence of the gadolinium-based contrast agent depositions in the brain and symptoms of gadolinium neurotoxicity - A systematic review. *PLoS One*. 2017;12(2):e0171704.

doi: 10.1371/journal.pone.0171704

23. Wang J, Mei T, Liu Y, *et al.* Dual-targeted and MRI-guided photothermal therapy via iron-based nanoparticles-incorporated neutrophils. *Biomater Sci*. 2021;9(11):3968-3978.

doi: 10.1039/D1BM00127B

24. Blanco-Andujar C, Walter A, Cotin G, *et al.* Design of iron oxide-based nanoparticles for MRI and magnetic hyperthermia. *Nanomedicine*. 2016;11(14):1889-1910.

doi: 10.2217/nnm-2016-5001

25. Norouzi A, Rahim MSM, Altameem A, *et al.* Medical image segmentation methods, algorithms, and applications. *IETE Tech Rev*. 2014;31(3):199-213.

doi: 10.1080/02564602.2014.906861

26. Castiglioni I, Rundo L, Codari M, *et al.* AI applications to medical images: From machine learning to deep learning. *Phys Med*. 2021;83:9-24.

doi: 10.1016/j.ejmp.2021.02.006

27. Sharma N, Ray A, Shukla K, *et al.* Automated medical image segmentation techniques. *J Med Phys*. 2010;35(1):3-14.

doi: 10.4103/0971-6203.58777

28. Liu H, Ren L, Fan B, Wang W, Hu X, Zhang X. Artificial intelligence algorithm-based MRI in the diagnosis of complications after renal transplantation. *Contrast Media Mol Imaging*. 2022;2022:8930584.

doi: 10.1155/2022/8930584

29. Guo YY, Huang YH, Wang Y, Huang J, Lai QQ, Li YZ. Breast MRI tumor automatic segmentation and triple-negative breast cancer discrimination algorithm based on deep learning. *Comput Math Methods Med*. 2022;2022:2541358.

doi: 10.1155/2022/2541358

30. Liu B, Tan B, Huang L, *et al.* Intelligent algorithm-based picture archiving and communication system of mri images and radiology information system-based medical informatization. *Contrast Media Mol Imaging*. 2021;2021:4997329.

doi: 10.1155/2021/4997329

31. Wang X, Li X, Chen H, Peng Y, Li Y. Pulmonary MRI radiomics and machine learning: Effect of intralesional heterogeneity on classification of lesion. *Acad Radiol*. 2022;29:S73-S81.

doi: 10.1016/j.acra.2020.12.020

32. Chang S, Han K, Lee S, *et al.* Automated measurement of native T1 and extracellular volume fraction in cardiac magnetic resonance imaging using a commercially available deep learning algorithm. *Korean J Radiol*. 2022;23(12):1251-1259.

doi: 10.3348/kjr.2022.0496

33. Bidhult S, Kantasis G, Aletras AH, Arheden H, Heiberg E, Hedström E. Validation of T1 and T2 algorithms for quantitative MRI: Performance by a vendor-independent software. *BMC Med Imaging*. 2016;16(1):46.

doi: 10.1186/s12880-016-0148-6

34. Jibon FA, Khandaker MU, Miraz MH, *et al.* Cancerous and non-cancerous brain MRI classification method based on convolutional neural network and log-polar transformation. *Healthcare* (*Basel*). 2022;10(9):1801.

doi: 10.3390/healthcare10091801

35. Holzinger A, Plass M, Kickmeier-Rust M, *et al.* Interactive machine learning: Experimental evidence for the human in the algorithmic loop. *Appl Intell*. 2019;49(7):2401-2414.

doi: 10.1007/s10489-018-1361-5

36. Topol EJ. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44-56.

doi: 10.1038/s41591-018-0300-7

37. Modan EM, Plăiaşu AG. Advantages and disadvantages of chemical methods in the elaboration of nanomaterials. *Ann Dunarea Jos Univ Galati Fascicle IX Metall Mater Sci*. 2020;43(1):53-60.

doi: 10.35219/mms.2020.1.08

38. Priyadharsini CI, Marimuthu G, Pazhanivel T, *et al.* Sol-Gel synthesis of Co3O4 nanoparticles as an electrode material for supercapacitor applications. *J Sol Gel Sci Technol*. 2020;96(2):416-422.

doi: 10.1007/s10971-020-05393-x

39. Alagiri M, Hamid SBA. Sol-gel synthesis of α-Fe2O3 nanoparticles and its photocatalytic application. *J Sol Gel Sci Technol*. 2015;74(3):783-789.

doi: 10.1007/s10971-015-3663-y

40. Marlin V, Lugo C, Manuel P, *et al.* Synthesis and characterization of magnetic nickel used in dry rerforming of methane. *Revista Ciencia e Ingeniería.*. 2017;38:31-40.

41. Pires LA, de Azevedo Silva LJ, Ferrairo BM, *et al.* Effects of ZnO/TiO_2 nanoparticle and TiO_2 nanotube additions to dense polycrystalline hydroxyapatite bioceramic from bovine bones. *Dent Mater*. 2020;36(2):e38-e46.

doi: 10.1016/j.dental.2019.11.006

42. Gates-Rector S, Blanton T. The Powder Diffraction File: A quality materials characterization database. *Powder Diffr*. 2019;34(4):352-360.

doi: 10.1017/S0885715619000812

43. Ahammed KR, Ashaduzzaman M, Paul SC, *et al.* Microwave assisted synthesis of zinc oxide (ZnO) nanoparticles in a noble approach: Utilization for antibacterial and photocatalytic activity. *SN Appl Sci*. 2020;2(5):955.

doi: 10.1007/s42452-020-2762-8

44. Zaparoli HH, De Oliveira M, Lisboa-Filho PN, Piacenti-Silva M. Using zinc particles in a phantom to simulate multiple sclerosis lesions on magnetic resonance imaging. *Rev Bras Física Méd*. 2021;15:619.

doi: 10.29384/rbfm.2021.v15.19849001619

45. Oliveira E, Rocha M, Froner AP, Basso N, Zanini M, Papaléo R. Synthesis and nuclear magnetic relaxation properties of composite iron oxide nanoparticles. *Quim Nova*. 2018;42:57-64.

doi: 10.21577/0100-4042.20170309

46. Perona P, Malik J. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell*. 1990;12(7):629-639.

doi: 10.1109/34.56205

47. Gerig G, Kbler O, Kikinis R, Jolesz FA. Nonlinear anisotropic filtering of MRI data. *IEEE Trans Med Imaging*. 1992;11(2):221-232.

doi: 10.1109/42.141646

48. Tustison NJ, Avants BB, Cook PA, *et al.* N4ITK: Improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-1320.

doi: 10.1109/TMI.2010.2046908

49. García-Lorenzo D, Francis S, Narayanan S, Arnold DL, Collins DL. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med Image Anal*. 2013;17(1):1-18.

doi: 10.1016/j.media.2012.09.004

50. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-310.

51. Wu S, He M, Yang M, Zhang B, Wang F, Li Q. Near-infrared spectroscopy study of serpentine minerals and assignment of the OH group. *Crystals*. 2021;11(9):1130.

doi: 10.3390/cryst11091130

52. Packiaraj R, Devendran P, Venkatesh KS, Asath Bahadur S, Manikandan A, Nallamuthu N. Electrochemical investigations of magnetic Co3O4 nanoparticles as an active electrode for supercapacitor applications. *J Supercond Nov Magn*. 2019;32(8):2427-2436.

doi: 10.1007/s10948-018-4963-6

53. Binitha NN, Suraja PV, Yaakob Z, Resmi MR, Silija PP. Simple synthesis of Co3O4 nanoflakes using a low temperature sol-gel method suitable for photodegradation of dyes. *J Sol Gel Sci Technol*. 2010;53(2):466-469.

doi: 10.1007/s10971-009-2098-8

54. Farhadi S, Pourzare K, Sadeghinejad S. Simple preparation of ferromagnetic Co3O4 nanoparticles by thermal dissociation

of the [CoII(NH3)6](NO3)2 complex at low temperature. *J Nanostructure Chem*. 2013;3(1):16.

doi: 10.1186/2193-8865-3-16

55. Sundar S, Venkatachalam G, Kwon S. Biosynthesis of copper oxide (CuO) nanowires and their use for the electrochemical sensing of dopamine. *Nanomaterials*. 2018;8(10):823.

doi: 10.3390/nano8100823

56. Usha V, Kalyanaraman S, Thangavel R, Vettumperumal R. Effect of catalysts on the synthesis of CuO nanoparticles: Structural and optical properties by sol-gel method. *Superlattices Microstruct*. 2015;86:203-210.

doi: 10.1016/j.spmi.2015.07.053

57. Zayyoun N, Bahmad L, Laânab L, Jaber B. The effect of pH on the synthesis of stable Cu2O/CuO nanoparticles by sol-gel method in a glycolic medium. *Appl Phys A*. 2016;122(5):488.

doi: 10.1007/s00339-016-0024-9

58. Abdulkadir I, Abdallah HMI, Jonnalagadda SB, Martincigh BS. The effect of synthesis method on the structure, and magnetic and photocatalytic properties of hematite (α-Fe2O3) nanoparticles. *South Afr J Chem*. 2018;71:68-78.

59. Raja K, Mary Jaculine M, Jose M, *et al.* Sol-gel synthesis and characterization of α-Fe2O3 nanoparticles. *Superlattices Microstruct*. 2015;86:306-312.

doi: 10.1016/j.spmi.2015.07.044

60. Namduri H, Nasrazadani S. Quantitative analysis of iron oxides using Fourier transform infrared spectrophotometry. *Corros Sci*. 2008;50(9):2493-2497.

doi: 10.1016/j.corsci.2008.06.034

61. Shamim A, Ahmad Z, Mahmood S, Ali U, Mahmood T, Nizami ZA. Synthesis of nickel nanoparticles by sol-gel method and their characterization. *Open J Chem*. 2019;2(1):16-20.

doi: 10.30538/psrp-ojc2019.0009

62. Gogoi P, Saikia BJ, Dolui SK. Effects of nickel oxide (NiO) nanoparticles on the performance characteristics of the jatropha oil based alkyd and epoxy blends. *J Appl Polym Sci*. 2015;132(8):n/a-n/a.

doi: 10.1002/app.41490

63. Jung HJ, Lee S, Yu Y, Hong SM, Choi HC, Choi MY. Low-temperature hydrothermal growth of ZnO nanorods on sol-gel prepared ZnO seed layers: Optimal growth conditions. *Thin Solid Films*. 2012;524:144-150.

doi: 10.1016/j.tsf.2012.10.007

64. Patel M, Mishra S, Verma R, Shikha D. Synthesis of ZnO and CuO nanoparticles via sol gel method and its characterization by using XRD and FT-IR analysis. *Research Square. 2022*;1:1-13.

doi: 10.21203/rs.3.rs-1234162/v1

65. Kaningini AG, Azizi S, Sintwa N, *et al.* Effect of optimized precursor concentration, temperature, and doping on optical properties of ZnO nanoparticles synthesized via a green route using bush tea (*Athrixia phylicoides* DC.) leaf extracts. *ACS Omega*. 2022;7(36):31658-31666.

doi: 10.1021/acsomega.2c00530

66. Nayan MB, Jagadish K, Abhilash MR, Namratha K, Srikantaswamy S. Comparative study on the effects of surface area, conduction band and valence band positions on the photocatalytic activity of ZnO-MxOy heterostructures. *J Water Resour Prot*. 2019;11(3):357-370.

doi: 10.4236/jwarp.2019.113021

67. Chen C, Ge J, Gao Y, *et al.* Ultrasmall superparamagnetic iron oxide nanoparticles: A next generation contrast agent for magnetic resonance imaging. *Wiley Interdiscip Rev Nanomed Nanobiotechnol*. 2022;14(1):e1740.

doi: 10.1002/wnan.1740

68. Zottis ADA, Beltrame JM, Lara LRS, *et al.* Pheomelanin-coated iron oxide magnetic nanoparticles: A promising candidate for negative T_2 contrast enhancement in magnetic resonance imaging. *Chem Commun* (*Camb*). 2015;51(56):11194-11197.

doi: 10.1039/C5CC02536B

69. Tham M, Frischer JM, Weigand SD, *et al.* Iron heterogeneity in early active multiple sclerosis lesions. *Ann Neurol*. 2021;89(3):498-510.

doi: 10.1002/ana.25974

70. Butterworth RF. Metal toxicity, liver disease and neurodegeneration. *Neurotox Res*. 2010;18(1):100-105.

doi: 10.1007/s12640-010-9185-z

71. Kanda T, Nakai Y, Aoki S, *et al.* Contribution of metals to brain MR signal intensity: Review articles. *Jpn J Radiol*. 2016;34(4):258-266.

doi: 10.1007/s11604-016-0532-8

72. Chen P, Miah MR, Aschner M. Metals and neurodegeneration. *F1000Research*. 2016;5:366.

doi: 10.12688/f1000research.7431.1

73. Dales JP, Desplat-Jégo S. Metal imbalance in neurodegenerative diseases with a specific concern to the brain of multiple sclerosis patients. *Int J Mol Sci*. 2020;21(23):9105.

doi: 10.3390/ijms21239105

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Vision transformers for glioma classification using T1 magnetic resonance imaging

W. M. S. P. B. Wickramasinghe, and Maheshi B. Dissanayake*

Department of Electrical and Electronics Engineering, Faculty of Engineering, University of Peradeniya, Peradeniya, Sri Lanka

(This article belongs to the *Special Issue: Artificial intelligence for diagnosing brain diseases*)

Abstract

Automated image analysis and classification have increasingly advanced in recent decades owing to machine learning and computer vision. In particular, deep learning (DL) architectures have become popular in resource-limited and labor-restricted environments such as the health-care sector. Transformer architecture, a DL method with self-attention mechanism, excels in natural language processing; however, its application in image-based diagnosis in health-care sector remains limited. Herein, the feasibility, bottlenecks, and performance of transformers in magnetic resonance imaging (MRI)-based brain tumor classification were investigated. To this end, a vision transformer (ViT) model was trained and tested using the popular Brain Tumor Segmentation (BraTS) 2015 dataset for glioma classification. Owing to limited data availability, domain adaptation techniques were used to pretrain the ViT model and the BraTS 2015 dataset was used for its fine-tuning. With the model only trained for 100 epochs, the confusion matrix for the two-class problem of tumor and nontumor classification showed an overall classification accuracy of 81.8%. In conclusion, although convolutional neural networks are traditionally used for DL-based medical image classification owing to their attention mechanism and long-range dependency-capturing capability, ViTs can outperform them in MRI-based brain tumor classification.

Keywords: Vision transformers; Medical image analysis; Deep neural networks; Magnetic resonance imaging; Convolutional neural network; Glioma detection

***Corresponding author:**
Maheshi B. Dissanayake
(maheshid@eng.pdn.ac.lk)

Citation: Wickramasinghe, WMSPB and Dissanayake MB. Vision transformers for glioma classification using T1 magnetic resonance imaging. *Artif Intell Health*. 2025;2(1):68-80. doi: 10.36922/aih.4155

Received: July 5, 2024

1st revised: August 28, 2024

2nd revised: September 9, 2024

Accepted: September 19, 2024

Published Online: November 6, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Medical imaging is crucial in the health-care sector for noninvasive diagnostic procedures because it can provide functional and visual representations of internal organs. X-ray imaging, nuclear imaging, magnetic resonance imaging (MRI), mammography, computed tomography (CT), and ultrasound imaging are some popular imaging techniques.[1] The four primary phases of medical image analysis include image formation, reconstruction, processing, and analysis. These phases help to create two-dimensional and three-dimensional (3D) images and enhance them; quantitative data are used for segmentation, classification, and object identification.[2,3] Modern advancements in artificial intelligence (AI), computer vision (CV), machine learning, and deep learning (DL) techniques can qualitatively and quantitatively improve medical image analysis.

MRI is a noninvasive medical imaging technique used for diagnosing brain tumors, injuries, and bleedings[4]. Over 120 types of brain tumors have been identified using MRI, which begin with primary tumors and are followed by secondary tumors. Glioma is a type of brain tumor that originates from glial cells and can be visualized through MRI or CT. Hence, automated image analysis techniques can positively contribute to the diagnosis of glioma.

Automated image analysis has rapidly evolved in the past few years owing to the introduction of AI and CV into traditional image processing techniques.[2] DL has been used in medical imaging to recognize cells of various sizes and shapes, locate organs and body parts, and automatically identify local anatomical features.[3-5] Owing to the intrinsic locality of convolution operations, popular architectures such as convolutional neural network (CNNs) have shown limitations in modeling straightforward long-range relations. Therefore, CNNs with attention mechanisms that assist AI models to focus on specific pixels, regions, or features have gained research attention in image analysis.

Transformer architecture, a DL method with self-attention mechanism, has become vital in natural language processing (NLP) tasks.[6] Recently, it has considerably impacted text classification, machine translation, and query responding. However, its application in CV problems requires further research. In CV, attention can either work in tandem with CNNs or replace some of its components while maintaining the overall network structure. Thus, this architecture largely has the potential to provide promising results in object detection, video classification, image classification, and image generation.

1.1. Contribution

This study mainly focused on evaluating the transformer architecture for MRI image classification when applied directly to the sequences of image patches. It concentrates on the classification of MRI images based on the presence and absence of glioma while overcoming the persistent class imbalance within a dataset to obtain feasible and resource-optimized solutions. We focus on glioma as it is a malignant (cancerous) brain tumor, which is treatable with high prognosis if detected early. The classification of brain tumors before segmentation is beneficial for saving time and resources, improving accuracy, and providing valuable information for treatment planning. Moreover, organized data aid in analysis and model training. However, medical data are mostly biased toward the absence of disease (negative outcome) and require careful implementation of algorithms to avoid model overfitting. This study aims to present a comprehensive ground-up mechanism for glioma classification using vision transformers (ViTs).

The primary contributions of this study are as follows:

(a) A ViT was proposed as an alternative to CNN for glioma classification using MRI.
(b) A pretraining method was proposed for ViTs when working with small datasets.
(c) An effective approach of intensity uniformization for MRI images as a preprocessing technique was introduced.
(d) The performance of CNN models and ViTs was compared for two grades of glioma classification as well as tumorous and nontumorous MRI classification.
(e) The effects of class imbalance in the medical dataset were discussed.

2. Background literature review

With the advancement in AI technologies, computer-aided diagnoses have been extensively studied in medical sciences for different disease diagnoses. In particular, noninvasive image-based diagnosis has garnered the attention of researchers and medical practitioners owing to its high accuracy, high precision, and auxiliary capabilities in applications such as brain tumor classification and segmentation using AI and DL models. Moreover, this field has gained popularity among medical image analysis researchers owing to well-established open challenges such as the BraTS challenge and publicly available large MRI datasets.[7-9] For instance, in their brain tumor classification and segmentation study, Kaldera *et al.*[5] proposed a simple CNN-based classifier for classifying glioma, meningioma, and the absence of a tumor using MRI. One of the main bottlenecks faced when using DL architectures for medical domain are data scarcity. This bottleneck was addressed using general data augmentation techniques such as flipping, rotation, and translation. Alsaif *et al.*[8] presented an improved ResNet50 architecture, which incorporated data augmentation techniques for effective brain tumor classification.

Because of the intrinsic locality of convolution operations, CNN-based approaches are generally inadequate for modeling straightforward long-range relations. Therefore, CNN-based architectures exhibit weak performances, particularly for target structures with varying textures, shapes, and sizes across patients. In previous studies, self-attention mechanisms with CNN features were used to overcome these limitations.[9]

Transformers, intended for sequence-to-sequence prediction, have emerged as ideal candidates to replace CNNs. These were first proposed for machine translation by Vaswani *et al.*[6] It was then established as the state-of-the-art method for many NLP tasks. It has the capacity to substitute attention mechanisms in place of convolution.[9-13]

Transformers also exhibit superior transferability for downstream tasks through extensive pretraining and superior performance in modeling global contexts. In many applications of machine translation and NLP, long short-term memory and artificial neural network have been successfully substituted by transformers.[10]

The results of transformers have matched or surpassed those of the state-of-the-art methods in various image recognition tasks.[12-16] The original design of transformers presented by Dosovitskiy *et al.*[11] has undergone several changes for suitability with CV tasks. For instance, Parmar *et al.*[12] modified transformers that used the self-attention mechanism only in local neighborhood of each query pixel. A novel transformer model, known as sparse transformers, was proposed by Child *et al.*[14] which attained global self-attention using scalable approximations. Wu *et al.*[15] introduced convolutions into ViTs to achieve best results on both convolutions and Transformers.

In general, large amounts of data and powerful computers are required for training ViTs, limiting their application in medical imaging diagnostics.[17-20] Hence, the research presented exploits the possibility of utilizing transformer-based attention features along with DL for the classification of brain tumors with a relatively small clinical dataset. We proposed mechanisms to tackle data scarcity and high processing power requirements while achieving sufficient model performance.

After image classification, MRI images with tumor underwent segmentation. Although segmentation generally provides detailed information about the spatial extent of tumors, classification offers insights into their nature. Therefore, segmentation was not researched and only image classification was focused on herein. However, as segmentation and classification work in tandem to provide a comprehensive understanding of disease diagnosis, existing studies can be referred to for more information on medical image segmentation.[21-25]

2.1. ViT model

ViTs, as presented by Dosovitskiy *et al.*,[11] mimic the original transformer model developed for NLP tasks using image patches as words for the input. ViTs can be used for image classifications primarily because they reduce architectural complexity and have enhanced exploring scalability and training efficiency. Recent studies have shown that the direct application of transformers with global self-attention to input images provided excellent results on ImageNet classification.[15] Moreover, ViTs can achieve high training accuracy with less computational time.[16] The success of transformers in medical image segmentation and classification was proven in the diagnosis of breast cancer using biopsy images and an end-to-end holistic attention network.[17] ViT-based medical image classification and segmentation continues to be a popular topic among researchers.

ViT contains stacks of encoder and decoder layers in its core, which will be hereinafter referred to as an encoder and a decoder, respectively. The encoder comprises two sublayers, namely multihead attention and feed-forward layers. The decoder comprises three sublayers, where the masked multihead attention layer is followed by the multihead attention layer and feed-forward layer. The encoder maps an input sequence $x = (x_1, x_2,\ldots, x_n)$ to a sequence $z = (z_1, z_2,\ldots, z_n)$. Based on z, the decoder generates an output sequence $y = (y_1, y_2,\ldots, y_n)$, with one element at a time. The model is auto-regressive at every step and uses already generated data as the additional input to create a new data instance. For more detail on the implementation of the ViT architecture, please refer to "An image is worth 16×16 words" by Dosovitskiy *et al.*[11]

There are two approaches to ViTs: hybrid and transformer-only architectures.[20] Hybrid architectures use a CNN to produce an embedding for an image or subregion of an image (patch). Encoding is used as the input for a subsequent transformer. In hybrid method, a CNN was used to process lower-level features in the input. In transformer-only architectures, a trainable part of the architecture projects patches to an embedding space and a hand-coded or convolutional architecture is not used. The transformer architecture learns only lower- and higher-level features.[16] Herein, transformer-only architecture is focused on and the model developed by Dosovitskiy *et al.*[11] was used for image classification.

Transformers have been used for tumor analysis in several studies. For instance, Asiri *et al.* used fine-tuned ViT model with the CE-MRI dataset containing only 5712 images for brain tumor classification.[25] The lack of diversity and limited number of images in the dataset affected the generalizability of ViT to real-world scenarios, suggesting further research to improve its accuracy and reliability, particularly for complex cases. Overall, the current ViT model used for brain tumor classification might not be fully optimized, and further research is required to enhance its diversity, reliability, and accuracy. This study focused on addressing this research gap in brain tumor classification using diverse BraTS datasets that primarily contain glioma MRIs. This dataset offered a benchmarked set of ground truth labels for glioma classification, addressing the limitations of existing studies. Moreover, potential model optimization techniques and MRI preprocessing techniques were discussed for their use in improving the model results.

3. Methodology

This section presents in detail, dataset preparation, including data preprocessing, ViT architecture, and model training with special attention to pretraining and fine-tuning approaches.

3.1. Dataset preparation and preprocessing

The BraTS 2015 dataset[7] containing 220 MRI images of high-grade gliomas (HGGs) and 54 images of low-grade gliomas (LGGs) was used for model training, validation, and testing. The dataset also contained MRI images of a patient in four modalities: T1 (spin-lattice relaxation), T1Gd (postcontrast T1-weighted), T2 (spin–spin relaxation), and T2-Flair (fluid attenuation inversion recovery). The analysis was restricted to the axial plane images of T1-MRIs, and the file format of the dataset was ".mha," which primarily is associated with the insight segmentation and registration toolkit. The DL architecture used the ".png" as the input image format. Hence, the T1-MRIs of a patient were converted to ".png" using "mha2png." Each patient's record resulted in 154 independent ".png" files, corresponding to brain slices in the coronal plane. Therefore, this resulted in a ".png" image dataset containing 42,196 images. Using the tumor mask of the BraTS 2015 dataset, each slice was first labeled based on the presence or absence of brain tumors. Then, slices with tumors were categorized into HGG or LGG tumors using the auxiliary data available in the BraTS 2015 dataset.

Intensity uniformization is another essential step in the preprocessing of MRI images. The pixel intensity of MRI images in BraTS ranges from −1000 to +1000, with more than 2000 levels. To aid image handling in limited resource environment, this pixel intensity range was decreased and scaled to match the intensity levels of 0–255, i.e., 8 bits/pixel grayscale. During preprocessing, the values above the upper gray level (G_u) and below the lower gray level (G_d) were assigned white and black, respectively. The center, also known as the window level (WL) and window width (WW), was changed based on the upper and lower gray levels. The upper gray level (G_u) was calculated as $G_u = WL + \left(\frac{WW}{2}\right)$, and the lower gray level (G_d) was calculated as $G_d = WL - \left(\frac{WW}{2}\right)$. Table 1 summarizes the effect of different values for WL, WW, and Range (G_u, G_d) on the preprocessed images. For instance, input images preprocessed considering WL = 0, WW = 400, and Range (−200 – 200) failed to show fine details of brain MRI images. After few trial and error iterations, range (−200 – 100), WW = 1200, and WL = 400 were chosen as the best parameters for 8 bit/pixel grayscale conversion.

Moreover, as the data on one patient record holds 154 (or 155) images, each image was considered a single input in the analysis and classified into one of the three classes: HGG, LGG, and nontumorous. The dataset was first developed using 120 patient records comprising 18,480 images, which were subgrouped into 3 subsets with 40 patients each. The dataset was further separated into two subsets, namely training and testing; approximately 70% of data were used for training and 30% for testing.

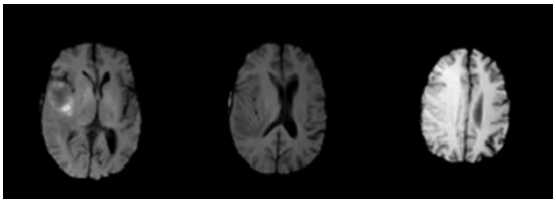
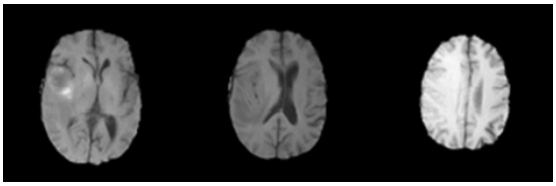
3.2. ViT architecture

ViTs are a group of neural network architectures that convert one sequence of images into another sequence. During preprocessing in ViTs, the input image is split into fixed-size patches and an input sequence is generated by linearly embedding each image into a sequence vector by adding position embedding information (Figure 1). The encoder transforms the input sequence into an embedding space, which is a vector representation of the image. Subsequently, the decoder receives the data in the embedding space and converts this into an output vector. An embedding layer generally proceeds each encoder or decoder to process their respective input, and an output layer is used toward the end of the architecture to generate the final output. ViTs perform classification using an extra learnable layer, i.e., classifier.[20] Figure 1 summarizes the process of image classification using the ViT for image recognition. Herein, a modified version of the model[26] was used for the classification of MRI images from the BraTS 2015 dataset. The classification operation flow of ViT is shown in Figure 2, and the performance of the proposed system was analyzed using accuracy, training, and validation loss and confusion matrix.

3.3. Model pretraining and fine-tuning

ViT is a DL model that requires considerably large dataset for model training. As BraTS is a relatively small dataset to train the ViT effectively, pretraining was performed to generate initial weights. CIFAR-10, a simple dataset, can serve as a foundation for pretraining models for medical image analysis.[27] The ViT was pretrained using the grayscale images of CIFAR-10, comprising 60,000 32 × 32 images belonging to 10 classes. All classes in CIFAR-10 are mutually exclusive, without any overlap between each class, which are well defined and bounded. For pretraining, the dataset was split into five training batches and one test batch, with each batch comprising 10,000 images. The test batch of CIFAR-10 was created using exactly 10,000 randomly selected images, and the training batches contained the remaining 50,000 images. Some training batches contained more images from one class than the other because the remaining images were added to the training batches in a random order.

Table 1. Comparison of preprocessed magnetic resonance imaging images with different intensity ranges, WLs, and WWs

Range	WW	WL	Image
from−1000 to−200	800	−600	
from−200 to 200	400	0	
from 200 to 1000	800	600	
from−200 to 1000	1200	400	

Abbreviation: WL: Window level; WW: Window width.

After the ViT model was successfully pretrained using CIFAR-10, transfer learning was used to initiate starting weights for the brain tumor classification task. The BraTS dataset with 15,000 images generated was split into training and testing datasets with a 70:30 ratio. Using the pretrained initial weights obtained using CIFAR-10, the model was warm started and its weights were fine tuned for brain tumor classification using BraTS dataset.

3.4. Statistical analysis

The analysis performed herein was simulated using Google Colab Jupyter notebook and Python 3.6 programming language. To evaluate the performance of the proposed ViT architecture, its training and validation accuracies and loss curves were analyzed. Thereafter, the model's performance was compared against a simple CNN network. Also, performance of the model was tested further using the accuracy, precision, and recall metrics. These metrics were calculated from the confusion matrix.[28]

4. Results

The performance of the ViT model in classifying glioma from MRI images was evaluated herein and compared with that of the conventional CNN. Its performance was evaluated for the task of handling two- and three-class problems under the class imbalance problem.

4.1. Training the ViT model

4.1.1. Pretraining the ViT model

In medical image analysis, collecting a considerably large dataset is a practically infeasible task. However, to achieve desirable performance with the ViT model, the DL architecture must be trained using a large dataset. To address this shortcoming, the customized ViT model was pretrained using a large general dataset, specifically CIFAR-10, and later fine-tuned with BraTS. Figure 3 shows the performance of the ViT model during pretraining using CIFAR-10, indicating that the model stabilized over time under 100 epochs.

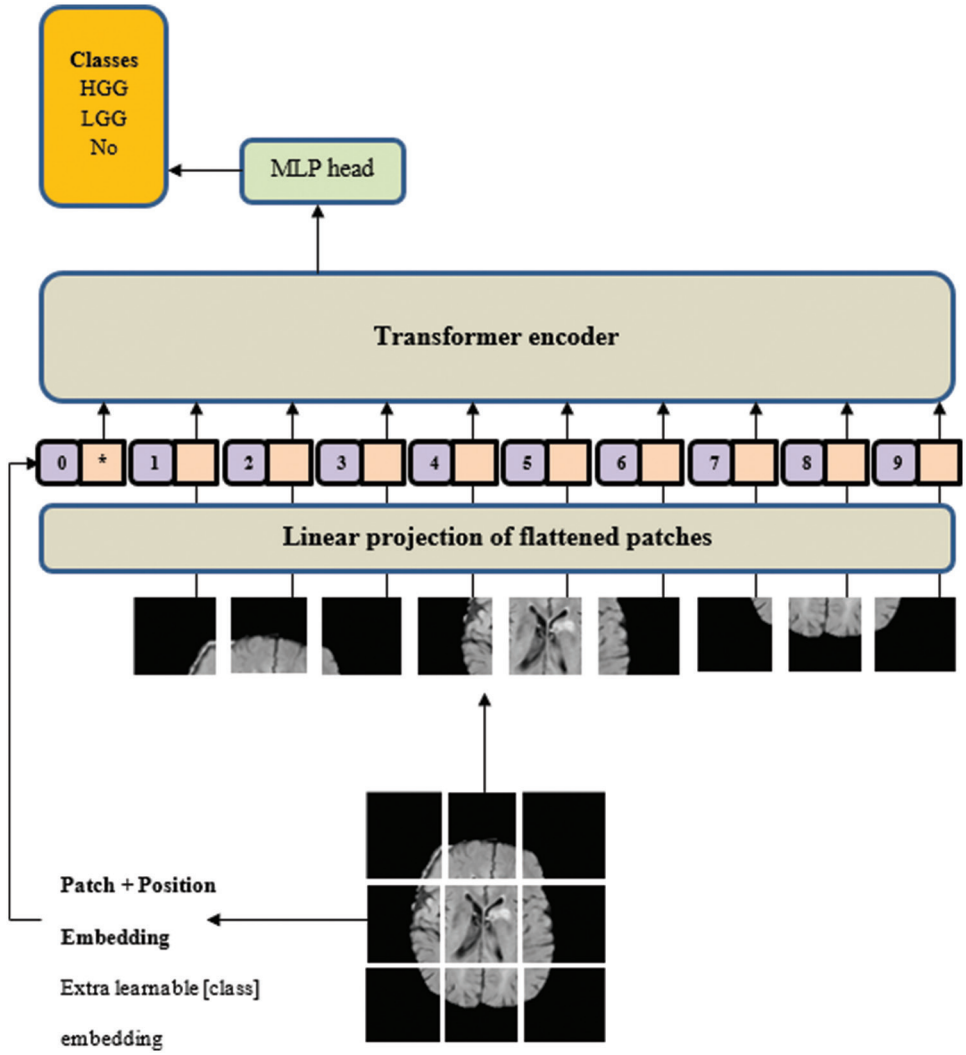


Figure 1. Basic block diagram of the vision transformer model

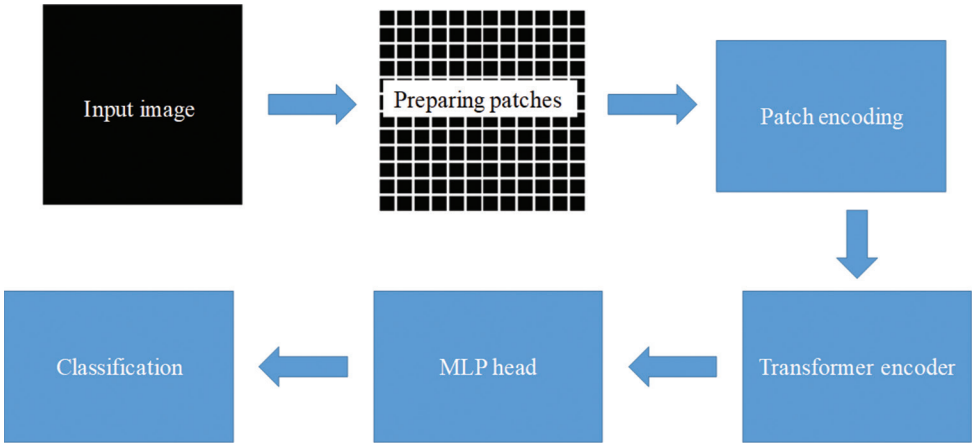


Figure 2. Operation flow of vision transformer-based classification

4.1.2. Fine-tuning the ViT model under different patch sizes

One of the distinct novelties associated with ViT model is the patch architecture. The pertained ViT model was fine tuned under different patch resolutions using the BraTs 2015 dataset. The objective of this approach was to find the most suitable patch size for a given application. The performance of each patch size was analyzed using model accuracy, loss performance, and confusion matrix. Figures 4-6 demonstrate the performance variation of ViT with patch sizes of 6 × 16, 8 × 8, and 4 × 4, respectively. In these figures, subplot (a) presents the training and validation loss, subplot (b) presents the training and validation accuracy while subplot (c) presents the confusion matrix, for the respective patch size. Table 2 summarizes the performance of ViT model under each patch size. As shown in Figures 4-6 and Table 2, the 4 × 4 patch resolution shows acceptable performance with 62.56% accuracy and lower level of fluctuation in the validation curves. The model could accurately detect the nontumorous MRI images, as shown in Figure 6C. However, the 4 × 4 patch resolution drastically increased the model tuning time.

4.2. Comparison of ViT model performance against CNN architecture

The traditional CNN architecture was used as the reference model for performance comparison of the ViT model

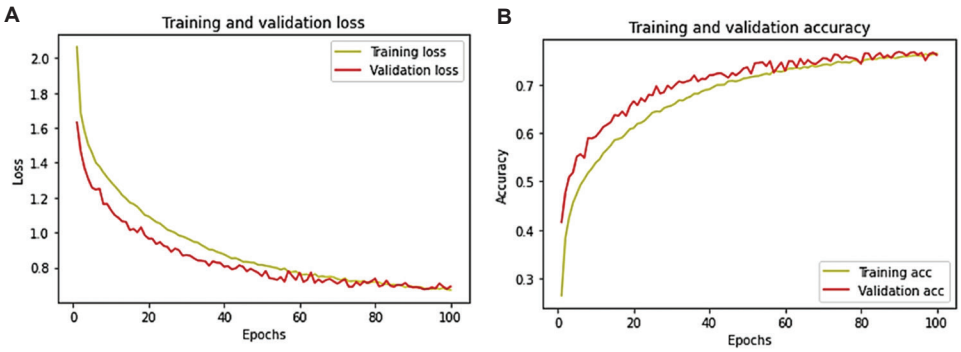


Figure 3. Vision transformer model performance during pretraining. (A) Training and validation losses. (B) Training and validation accuracy.

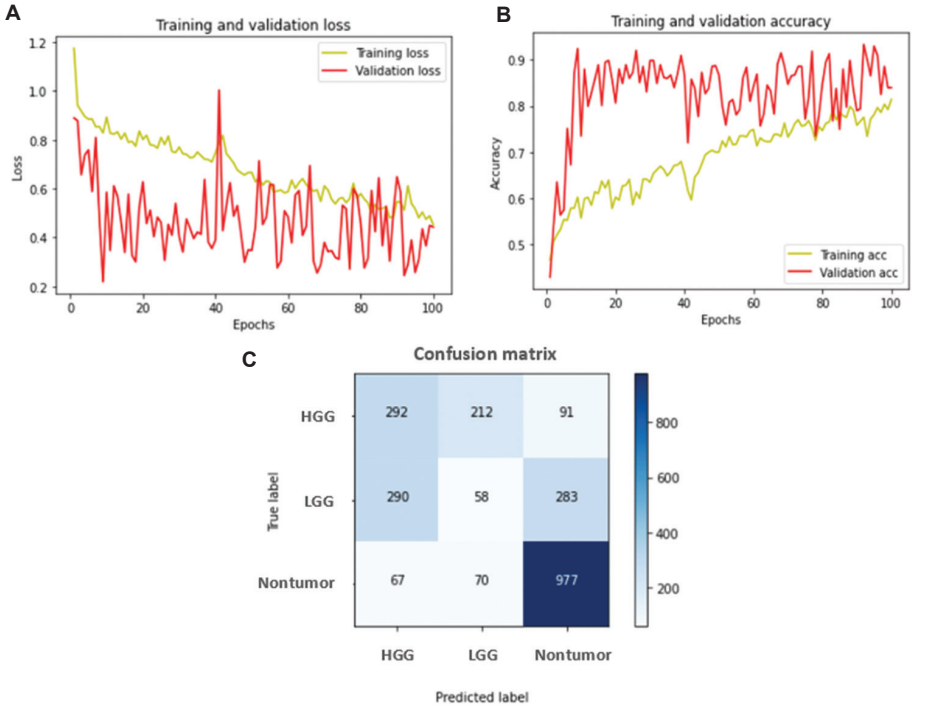


Figure 4. Performance of model fine-tuning using 16 × 16 patches. (A) Variation of model loss versus epoch. (B) Variation of model accuracy versus epoch. (C) Classification performance of the model presented using the confusion matrix.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma.

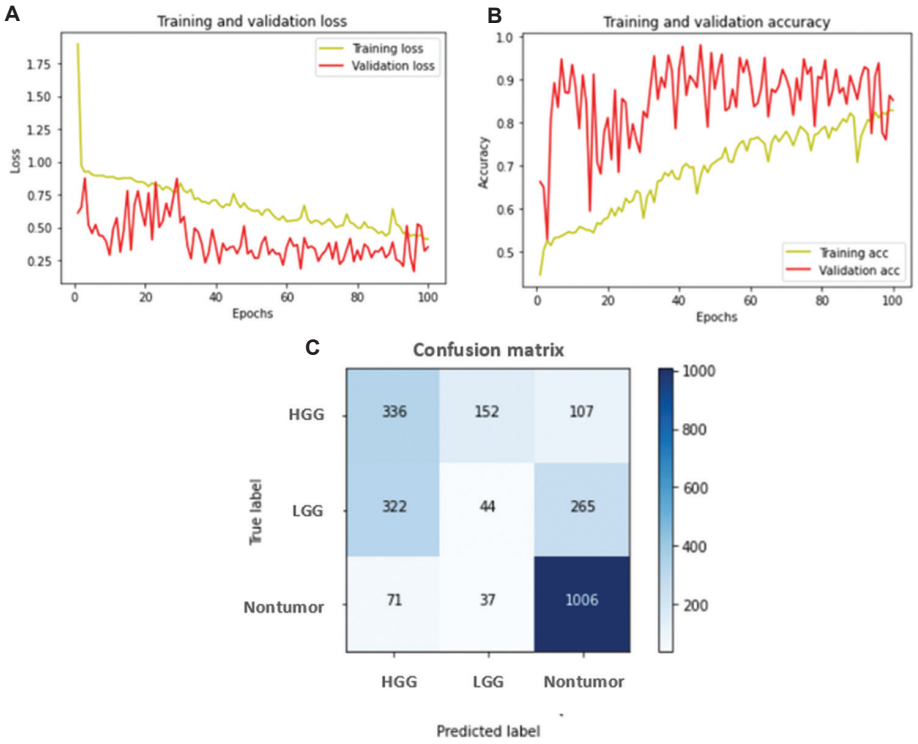


Figure 5. Performance of model fine-tuning using 8 × 8 patches. (A) Variation of model loss versus epoch. (B) Variation of model accuracy versus epoch. (C) Classification performance of the model presented using the confusion matrix.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma.

with 4 × 4 patch resolution. This CNN model has four convolution layers with only 8 million trainable parameters and was trained using the same dataset as ViT to classify the brain tumors.

Figure 7A shows the training and validation accuracy for the CNN model while Figure 7B shows that of the ViT model for brain tumor classification. Both models were trained using the same dataset under optimized settings. Although the CNN model showed nearly prefect accuracy with training, it underperformed during validation and indicated model overfitting. By contrast, the ViT model exhibited better performance in training and validation settings. As shown in Figure 7B, the ViT model exhibited a considerable level of instability. To stabilize the ViT model, it needs to be further trained using a large dataset. However, one of the critical factors in medical image analysis is the limitations in dataset; therefore, stabilizing the ViT model with small datasets is challenging.

4.3. Model performance under two-class problem

Furthermore, the accuracy of the ViT model with 4 × 4 patch size was analyzed for the task of classifying MRI images as with tumor or without tumor. According to the confusion matrix shown in Figure 8, the overall accuracy of classification for a three-class problem was 63.2% (Figure 8A), whereas that of two-class problem was 81.8% (Figure 8B), i.e., the trained and fine-tuned ViT model could detect the presence and absence of tumors with higher accuracy than classifying the different grades of tumors. The main reason behind this observation is the restriction in the number of images belonging to each class. For the three-class problem, the dataset showed a class imbalance, whereas it was balanced for the two-class problem. This observation indicated that the dataset used was suitable for tumor identification with two classes: with tumor and without tumor.

5. Discussion

CNN-based approaches are a popular choice for brain tumor classification using MRI images. They are highly effective in processing and analyzing medical data owing to their ability to automate feature extraction, capture hierarchical features, perform end-to-end learning, and yield high-accuracy output. However, transformers are emerging as leading contenders for this task, mainly because of their global context modeling features. In particular, their capacity to capture long-range dependencies and ability to focus on relevant parts of the input images are noteworthy.

CNN-based architectures perform weakly, particularly with datasets that show large variation in terms of texture,

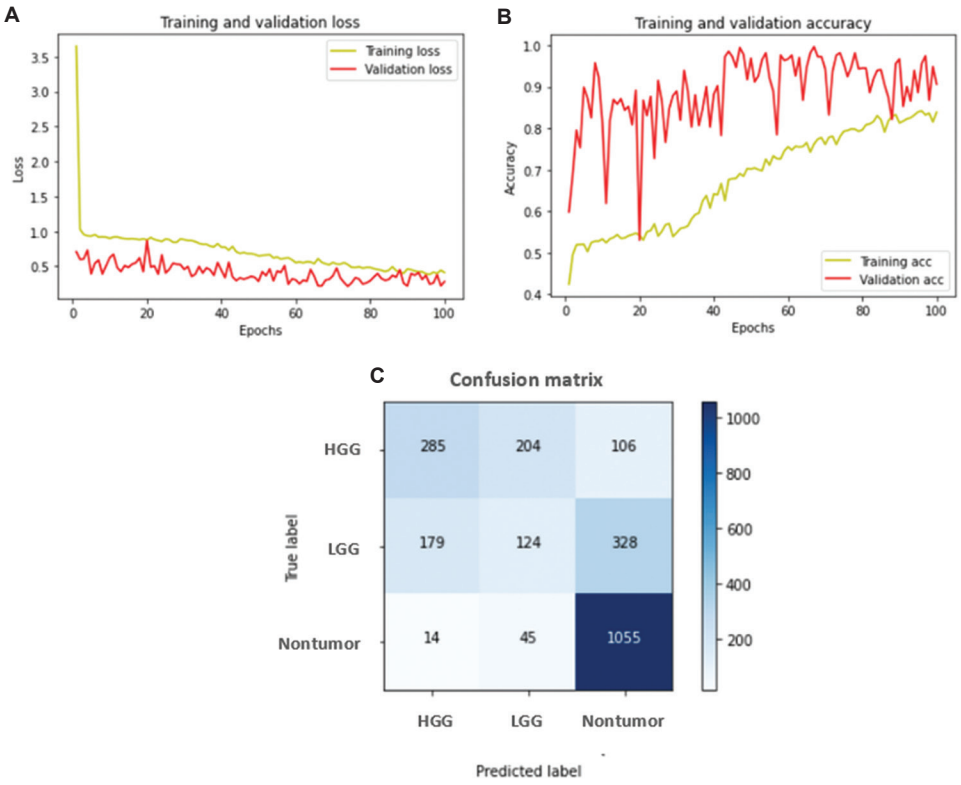


Figure 6. Performance of model fine-tuning using 4 × 4 patches. (A) Variation of model loss versus epoch. (B) Variation of model accuracy versus epoch. (C) Classification performance of the model presented using the confusion matrix.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma

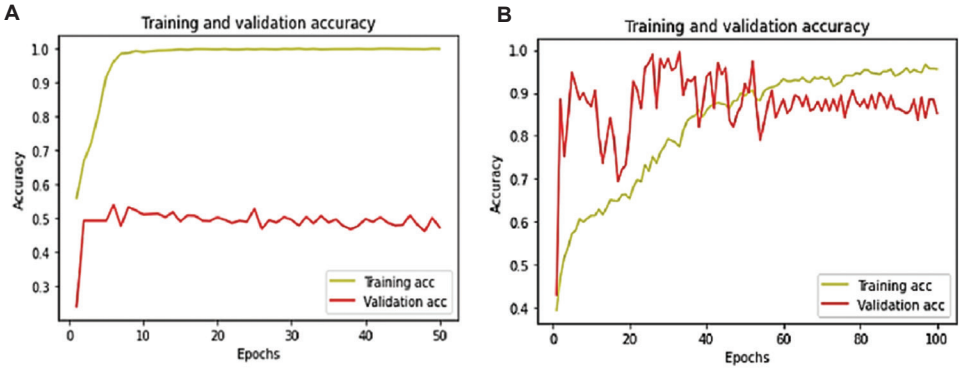


Figure 7. Model performance comparison between CNN and ViT models. (A) Performance accuracy versus epochs for CNN model. (B) Performance accuracy versus epochs for the ViT model.
Abbreviations: CNN: Convolutional neural network; ViT: Vision transformer.

shape, and size. The newly emerged transformer-based DL architectures, especially ViTs, show promising capacity to overcome these limitations. Although ViTs are a new concept for medical imaging, the accuracy of medical image classification can be improved using self-attention. For instance, the model can be trained to focus on abnormal cells in MRI by dynamically adjusting the weight assigned to these areas using attention mechanisms. This eventually improves the overall model performance. Moreover, the model can capture the relationship between tumors that are far apart owing to the inherent long-range dependency of ViTs. This introduces a provision for the model to learn dependencies between different slices of different planes of MRI images. Figure 7 shows the performance improvements achieved owing to these inherent characteristics of the ViT model in comparison with those of simple CNN.

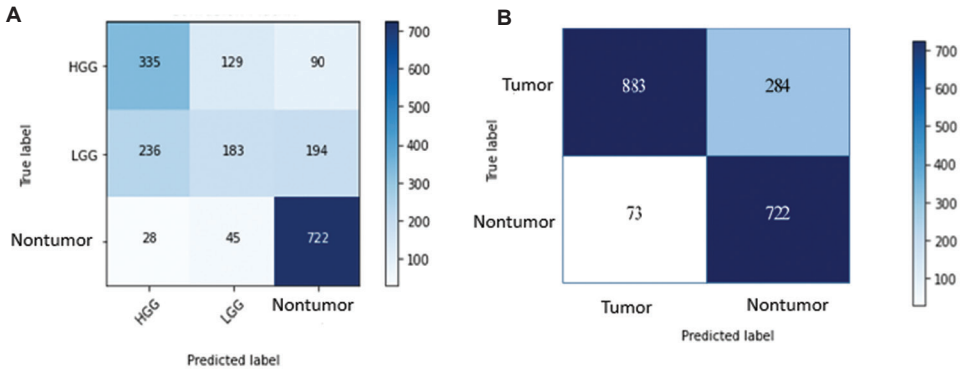


Figure 8. Comparison between three- and two-class classification problem. (A) Confusion matrix for a three-class problem with HGG, LGG, and nontumor. (B) Confusion matrix for a two-class problem with tumor (HGG and LGG tumors) and nontumor.
Abbreviations: HGG: High-grade glioma; LGG: Low-grade glioma.

Table 2. Comparison of the model performance for different patch sizes with learning rate=0.001 and weight decay=0.0001 and Adam optimizer

Patch size	Number of patches	Overall accuracy	Time taken to process (s)
16×16	4	56.70%	700
8×8	16	59.23%	1,900
4×4	64	62.56%	8,600
2×2	256	-	5,6100 (Estimated)

To combat the negative performance of ViTs owing to data scarcity, a pretraining approach coupled with transfer learning is presented herein. Moreover, the effects of the patch resolution on the overall performance accuracy and the loss curve behavior are discussed. With a 4 × 4 patch resolution, the stability of the model increased at the expense of inference time. Experimental results showed that the model performed better on the two-class problem of tumor and nontumor detection than on the three-class problem of HGG, LGG, and nontumor detection owing to class imbalance present in the BraTS 2015 dataset.

Moreover, the proposed model achieved an average classification accuracy of 81.8% for the BraTS 2015 dataset for the two-class problem. The confusion matrix in Figure 8 shows a model accuracy of 75.6% in detecting tumors and 90.8% in detecting nontumors. These results agreed well with previous studies using the BraTS 2015 data. For instance, the DL ensemble model that concatenates the weighted outputs of the cascaded anisotropic CNN (CA-CNN), DFKZ Net, and 3D U-Net achieved a classification accuracy of 46.4% during validation and 61% during testing with the BraTS 2018/2015 dataset.[29] The multiclass glioma tumor classification architecture presented in a previous study[30] achieved a 96.3% classification accuracy on a custom-built dataset that mainly used the BraTS 2015 dataset along with the other MRI images collected from different sources. The same custom-built dataset achieved a classification accuracy of 80.85% using 10 statistical features along with random forest[31] and 84.9% with dual-path residual CNNs.[32] The classification algorithm presented by Amin *et al.*[33] used discrete wavelet transform (DWT) to fuse MRI image sequences during preprocessing. The fused images followed the pipeline of denoising with a partial differential diffusion filter, segmentation using a global thresholding method, and classification of the segmented output into glioma, meningioma, and sarcoma using a CNN. This algorithm yielded a very high accuracy of nearly 100% in image fusion of all four MRI sequences, 89% in Flair + T1 fused images, and 78% in T1 images used herein. However, this algorithm first segmented tumor regions and then applied classification on the segmented region. Therefore, the results do not clearly present the detection accuracy on the initial dataset before segmentation.

Moreover, the BraTS datasets yielded better model performance. For instance, B. Maram and P. Rana achieved a quick and accurate image classification with a training accuracy of 98.485% using a U-Net architecture and BraTS 2020 dataset.[34] The novel linear-complexity data-efficient image transformer[35] achieved a classification accuracy of 97.86% with BraTS 2021 dataset. The ViT model discussed herein achieved a substantial level of classification accuracy using the BraTS 2015 dataset compared with those reported in the literature. However, if the input was preprocessed[33] or tested on an improved dataset such as BraTS 2021,[35] the performance accuracy of ViTs may increase compared with the current classification accuracy of 81.8%. Thus, the ViT model will be tested using the BraTS 2021 dataset and image preprocessing will be performed to facilitate better comparison and understanding on the performance of transformers for brain tumor classification.

6. Conclusion

Herein, the ViT architecture was studied for MRI image classification, focusing on glioma. To address the issues of data scarcity and class imbalance, ViT was pretrained using the CIFAR-10 dataset and fine-tuned using the BraTS 2015 dataset. The fine-tuned ViT could accurately and effectively identify glioma compared with the popular CNN architecture. Moreover, the effects of the patch resolution on the overall performance accuracy and the behavior of the loss curve were discussed. Overall, this study proposed a feasible and resource-optimized solution for the early detection and better prognosis of brain tumors. Further research is required to improve the predictions of the model while making the results understandable with explainable AI techniques for the advancement of automated systems for brain tumor detection and diagnosis.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Maheshi B. Dissanayake
Formal analysis: All authors
Investigation: All authors
Methodology: All authors
Writing – original draft: All authors
Writing – review & editing: Maheshi B. Dissanayake

Ethics approval and consent to participate

The data collection was not part of this research. We use publicly available BRATS dataset. Ethical clearance had already been obtained before the upload of the medical dataset BRATS onto the public domain by Menze BH *et. al*.[7]

Consent for publication

Not applicable.

Availability of data

The data utilized in this research are publicly available. The authors have released the code on the GitHub page (https://github.com/SaneruW/Vision-transformers-for-glioma-classifications-using-T1-magnetic-resonance-images). Regarding materials or details related to the implementation, please contact Mr. Saneru Wickramasinghe (saneruw@gmail.com).

References

1. Pal A, Chaturvedi A, Garain U, Chandra A, Chatterjee R. *Severity Grading of Psoriatic Plaques using Deep CNN Based Multi-task Learning*. Mexico: ICPR; 2016.

   doi: 10.1109/ICPR.2016.7899846

2. Wang G. A perspective on deep imaging. *IEEE Access*. 2016;4:8914-8924.

   doi: 10.1109/ACCESS.2016.2624938

3. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. *IEEE Access*. 2018;6:9375-9389.

   doi: 10.1109/ACCESS.2017.2788044

4. Kabir Anaraki A, Ayati M, Kazemi F. Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms. *Biocybern. Biomed. Eng*. 2019;39(1):63-74.

   doi: 10.1016/j.bbe.2018.10.004

5. Kaldera HNTK, Gunasekara SR, Dissanayake MB. Brain Tumor Classification and Segmentation Using Faster R-CNN. In: *Proceedings ASET*. United States: IEEE; 2019.

   doi: 10.1109/ICASET.2019.8714263

6. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. In: *Advances in Neural Information Processing Systems*. United States: The MIT Press; 2017. p. 5998-6008.

   doi: 10.48550/arXiv.1706.03762

7. Menze BH, Jakab A, Bauer S, *et al*. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imaging*. 2014;34(10):1993-2024.

   doi: 10.1109/TMI.2014.2377694

8. Alsaif H, Guesmi R, Alshammari BM, *et al*. A novel data augmentation-based brain tumor detection using convolutional neural network. *Appl Sci*. 2022;12(8):3773.

   doi: 10.3390/app12083773

9. Pan X, Ge C, Lu R, *et al*. *On the Integration of Self-Attention and Convolution*. United States: IEEE/CVF; 2022. p. 815-825.

   doi: 10.1109/CVPR52688.2022.0008

10. Devlin J, Chang MW, Lee K, Toutanova K. Pre-training of Deep Bidirectional Transformers for Language Understanding; 2018.

    doi: 10.48550/arXiv.1810.04805

11. Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale; 2020.

    doi: 10.48550/arXiv.2010.11929

12. Parmar N, Vaswani A, Uszkoreit J, *et al*. Image Transformer. In: *JMLR Workshop and Conference Proceedings*; 2018. p. 4055-4064.

doi: 10.48550/arXiv.1802.05751

13. Zheng S, Lu J, Zhao H, *et al*. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers; 2020.

doi: 10.48550/arXiv.2012.15840

14. Child R, Gray S, Radford A, Sutskever I. Generating long sequences with sparse transformers; 2019.

doi: 10.48550/arXiv.1904.10509

15. Wu H, Xiao B, Codella N, *et al*. *Introducing Convolutions to Vision Transformers. CVF 2021*. United States: IEEE. p. 22-31.

doi: 10.1109/ICCV48922.2021.00009

16. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *European Conference on Computer Vision 2020 Aug 23*. Cham: Springer International Publishing; 2020. p. 213-229.

doi: 10.1007/978-3-030-58452-8_13

17. Aloraini M, Khan A, Aladhadh S, Habib S, Alsharekh MF, Islam M. Ombining the transformer and convolution for effective brain tumor classification using MRI Images. *Appl Sci*. 2023;13:3680.

doi: 10.3390/app13063680

18. Mehta S, Lu X, Weaver D, Elmore JG, Hajishirzi H, Shapiro L. HATNet: An end-to-end holistic attention network for diagnosis of breast biopsy images; 2007.

doi: 10.48550/arXiv.2007.13007

19. Lan YL, Zou S, Qin B, Zhu X. Potential roles of transformers in brain tumor diagnosis and treatment. *Brain X*. 2023;1:e23.

doi: 10.1002/brx2.23

20. Courant R, Edberg M, Dufour N, Kalogeiton V. Transformers and visual transformers. In: Colliot O, editors. *Machine Learning for Brain Disorders. Neuromethods*. vol. 197. United States: Humana; 2023.

doi: 10.1007/978-1-0716-3195-9_6

21. Zunair H, Ben Hamza A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Comput Biol Med*. 2021;136:104699.

doi: 10.1016/j.compbiomed.2021.104699

22. Dasanayaka C, Dharmasena B, Bandara WR, Dissanayake MB, Jayasinghe R. Segmentation of Mental Foramen in Dental Panoramic Tomography Using Deep Learning. In: *2019 IEEE 14th Conference on Industrial and Information Systems (ICIIS)*. IEEE; 2019. p. 81-84.

doi: 10.1109/ICIIS47346.2019.9063312

23. Wang P, Yang Q, He Z, Yuan Y. Vision transformers in multi-modal brain tumor MRI segmentation: A review. *Meta Radiol*. 2023;1:100004.

doi: 10.1016/j.metrad.2023.100004

24. Marathe A, Kadam V, Chaumal A, Kodilkar S, Joshi A, Sawant S. Performance analysis of memory-efficient vision transformers in brain tumor segmentation. In: *Artificial Intelligence-Based Healthcare Systems*. Cham: Springer Nature Switzerland; 2023. p. 125-133.

doi: 10.1007/978-3-031-41925-6_9

25. Asiri AA, Shaf A, Ali T, *et al*. Exploring the power of deep learning: Fine-tuned vision transformer for accurate and efficient brain tumor detection in MRI Scans. *Diagnostics*. 2023;13(12):2094.

doi: 10.3390/diagnostics13122094

26. Salama K. *Image Classification with Vision Transformer*; 2022. Available: https://keras.io/examples/vision/image_classification_with_vision_transformer [Last accessed on 2022 Oct 10].

27. Mabu S, Atsumo A, Kido S, Kuremoto T, Hirano Y. Investigating the effects of transfer learning on ROI-based classification of chest CT images: A case study on diffuse lung diseases. *J Signal Process Syst*. 2020;92:307-313.

doi: 10.1007/s11265-019-01499-w

28. Kanesamoorthy K, Dissanayake MB. Prediction of treatment failure of tuberculosis using support vector machine with genetic algorithm. *Int J Mycobacteriol*. 2021;10(3):279-284.

doi: 10.4103/ijmy.ijmy_130_21

29. Sun L, Zhang S, Chen H, Luo L. Brain tumor segmentation and survival prediction using multimodal MRI scans with deep learning. *Front Neurosci*. 2019;13:810.

doi: 10.3389/fnins.2019.00810

30. Latif G. DeepTumor: Framework for brain MR image classification, segmentation and tumor detection. *Diagnostics* (*Basel*). 2022;12(11):2888.

doi: 10.3390/diagnostics12112888

31. El-Melegy MT, El-Magd KMA. A Multiple Classifiers System for Automatic Multimodal Brain Tumor Segmentation. In: *Proceedings of the 2019 15th International Computer Engineering Conference (ICENCO), Giza, Egypt. 29-30 December 2019*. New York, NY, USA: IEEE; 2019.

doi: 10.1109/ICENCO48310.2019.9027389

32. Xue Y, Yang Y, Farhat FG, *et al*. Brain tumor classification with tumor segmentations and a dual path residual convolutional neural network from MRI and pathology images. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Germany: Springer; 2020. p. 360-367.

doi: 10.1007/978-3-030-46643-5_36

33. Amin J, Sharif M, Gul N, Yasmin M, Shad SA. Brain tumor classification based on DWT fusion of MRI sequences using convolutional neural network. *Pattern Recognit Lett*. 2020;129:115-122.

doi: 10.1016/j.patrec.2019.11.016

34. Maram B, Rana P. Brain Tumour Detection on BraTS 2020 using U-Net. In: *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, Noida, India; 2021. p. 1-5.

doi: 10.1109/ICRITO51393.2021.9596530

35. Ferdous GJ, Sathi KA, Hossain MA, Hoque MM, Dewan MAA. LCDEiT: A linear complexity data-efficient image transformer for MRI brain tumor classification. *IEEE Access*. 2023;11:20337-20350.

doi: 10.1109/ACCESS.2023.3244228

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Benchmarking machine learning missing data imputation methods in large-scale mental health survey databases

Preethi Prakash[1], Kelly Street[2], Shrikanth Narayanan[3], Bridget A. Fernandez[4,5], Yufeng Shen[6], and Chang Shu[7]*

[1]Department of Computer Science, Fu Foundation School of Engineering and Applied Science, Columbia University, New York, NY, United States of America

[2]Department of Population and Public Health Sciences, Division of Biostatistics, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America

[3]Viterbi School of Engineering, University of Southern California, Los Angeles, CA, United States of America

[4]Department of Pediatrics, Division of Medical Genetics, Children's Hospital Los Angeles and The Saban Research Institute, Los Angeles, CA, United States of America

[5]Department of Pediatrics, Keck School of Medicine of USC, University of Southern California, Los Angeles, CA, United States of America

[6]Departments of Systems Biology and Biomedical Informatics, and JP Sulzberger Columbia Genome Center, Columbia University Irving Medical Center, New York, NY, United States of America

[7]Department of Population and Public Health Sciences, Center for Genetic Epidemiology, Division of Epidemiology and Genetics, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America

***Correspondence author:**
Chang Shu
(april.shu@usc.edu)

Citation: Prakash P, Street K, Narayanan S, Fernandez BA, Shen Y, Shu C. Benchmarking machine learning missing data imputation methods in large-scale mental health survey databases. *Artif Intell Health.* 2025;2(1):81-92. doi: 10.36922/aih.4406

Received: August 1, 2024

Revised: September 17, 2024

Accepted: October 14, 2024

Published Online: November 7, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

Databases tied to mental and behavioral health surveys suffer from the issue of missing data when participants skip the entire survey, which affects the data quality and sample size. These missing data patterns were investigated and the imputation performance was evaluated in Simons Foundations Powering Autism Research for Knowledge, a large-scale autism cohort consists of over 117,000 participants. Four common methods were assessed – Multiple imputation by chained equations (MICE), K-nearest neighbors (KNN), MissForest, and multiple imputation with denoising autoencoders (MIDAS). In a complete subset of 15,196 autism participants, three types of missingness patterns were simulated. We observed that MIDAS and KNN performed the best as the random missingness rate increased and when blockwise missingness was simulated. The average computational times were each 10 min for MIDAS and KNN, 35 min for MissForest, and 290 min for MICE. MIDAS and KNN both provide promising imputation performance in mental and behavioral health survey data that exhibit blockwise missingness patterns.

Keywords: Missing data; Mental health survey; Imputation methods; Machine learning

1. Introduction

Large-scale biobank databases in mental and behavioral health such as Simons Foundations Powering Autism Research for Knowledge (SPARK), UK Biobank, and All

of Us have empowered researchers to investigate the genetic and environmental risk factors associated with mental and behavioral disorders among more than 100,000 subjects.[1-3] Self-reported surveys and questionnaires such as the social communication questionnaire (SCQ),[4] repetitive behavior scale-revised (RBS-R),[5] and developmental coordination disorder questionnaire (DCDQ)[6] are commonly used to quantify mental and behavioral functions at scale. These questionnaires typically consist of a series of related questions and measure responses using ordinal scales with a natural order or rank to indicate the level of agreement known as Likert scales.[7]

However, missingness commonly occurs in the responses to these surveys and questionnaires. The reasons include non-inapplicable or ambiguous questions, and characteristics of the participants themselves including reluctance to answer sensitive questions, incomplete knowledge, and lack of time.[8] Missingness can also arise at the source level. Specifically, data may have been curated from varying sources with different administered instrument protocols. Certain questions in the survey also may not be relevant to specific demographic groups, such as those that might not apply to young children.

Common types of missing data include missing completely at random (MCAR) and missing not at random (MNAR), with either specific parts of surveys or entire surveys being incomplete.[9] In MCAR, the probability of missingness is independent of the observed and unobserved data. MAR is a broader class than MCAR in which the missing data is related to the observed but not the unobserved data. On the other hand, the probability of missingness in MNAR data depends on the unobserved missing values. Typically, participants tend to skip entire questionnaires due to unobserved factors, and a form of MNAR missingness referred to as blockwise missingness arises. Blockwise missingness occurs when all responses belonging to the same survey are missing simultaneously for the same participants, forming clustered missing blocks in the overall phenotypic data.

The simplest solution to address blockwise missingness in mental and behavioral questionnaires is to drop participants with missing surveys.[10] However, this option leads to a significant loss of information, reduced sample size, and loss of statistical power when analyzing mental and behavioral questionnaires in biobank data. Another commonly used approach is to impute missing data using statistical and computational methods. Mean, median, and mode substitutions are basic imputation approaches that maintain the original sample size but can lead to biased inferences.[11] Specifically, participants who skip certain questionnaires may exhibit different characteristics than those who complete the questionnaires.[12]

More advanced imputation approaches using statistical and computational methods are needed to accurately impute mental and behavioral surveys with blockwise missingness. Here, four commonly used missing data imputation methods were employed – Multivariate imputation by chained equations (MICE), K-nearest neighbors (KNN), non-parametric missing value imputation using random forest (MissForest), and multiple imputation with denoising autoencoders (MIDAS).[13-16] MICE is one of the most popular methods of multiple imputation originally developed in the early 2000s.[13] This approach uses a series of regression models to predict each variable with missingness using the remaining variables in the data.[14] KNN is a supervised machine learning algorithm commonly used when the distribution of the data is unknown or difficult to determine.[15] This method performs predictions on the missing data by averaging the K-nearest data points. MissForest is a missing data imputation method based on a random forest developed in 2012. It predicts missing values based on random forest models trained on the complete dataset and imputes missing values iteratively.[16] MIDAS uses a type of unsupervised neural network to predict missing values in the data by reducing the dimensions in the observed data and reconstructing the missing data. MIDAS was recently developed in 2022 and has proven its high accuracy and computational efficiency through systematic tests on simulated and real social science data.[17]

Previous studies have not systematically reviewed machine learning-based imputation methods recently developed for the databases tied to mental and behavioral health surveys.[18-22] Most psychiatric studies use multiple imputation for handling missing data and have not taken advantage of the latest machine learning-based imputation techniques.[18-22] In addition, they have not focused on assessing imputation accuracy in surveys with blockwise missing structures.[18-22] This study systematically examines the imputation performance and computational time of these four commonly used missing data imputation methods (MICE, KNN, MissForest, and MIDAS) in the presence of blockwise missingness in mental and behavioral surveys. It uses data from the SPARK, a large-scale autism research study that collects social functioning and behavioral surveys from over 117,000 participants. This study assesses imputation models on both MCAR and MNAR data, identifying the optimal method for each type of missingness pattern. This study conducts a novel exploration of these methods while also addressing the commonly encountered blockwise missingness pattern.

2. Methods

Figure 1 outlines the sample selection and workflow of the study. The four major steps included: (1) preprocessing the

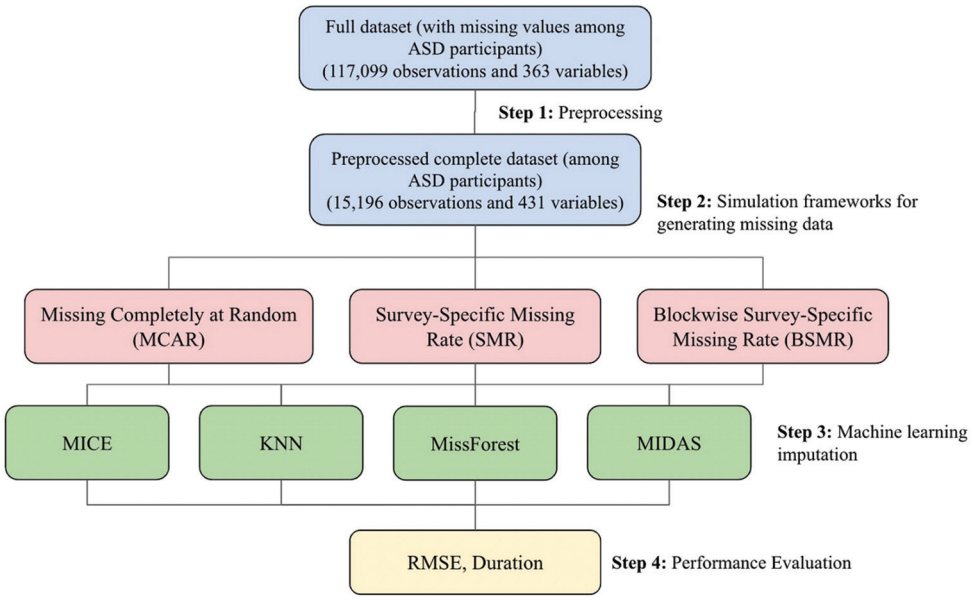


Figure 1. Overview of workflow and study design. (A) The full dataset refers to the original data filtered to only include autism spectrum disorder (ASD) participants. The preprocessed complete dataset refers to the original dataset after filtering to only include ASD participants, dropping incomplete rows, removing variables with extreme rates of missingness, and conducting one-hot-encoding on the categorical variables (which increases the number of variables). (B) Missing completely at random refers to the simulation scenario that randomly converts a specified fraction of the input dataset to missing. Survey-specific missing rate refers to the simulation environment that is tailored to the missingness of the original dataset. Blockwise survey-specific missing rate refers to the simulation environment that is also tailored to the missingness of the original dataset but converts all rows of a given column to missing at once. (C) Multiple imputation by chained equations is an imputation method that employs a series of regression models; MissForest is an imputation method that is based on random forests; Multiple Imputation with Denoising Autoencoders is an imputation method that uses denoising autoencoders; K-nearest neighbors is an imputation method that uses neighboring data points in the feature space. (D) RMSE corresponds to root mean squared error.

data to generate a dataset comprising complete observations, (2) setting up the simulation scenarios for three missing data mechanisms including random missingness, survey-specific missing rates, and blockwise missingness with survey-specific missing rates, (3) conducting the missing data imputation, and (4) evaluating the performance of each model.

2.1. Data source and preprocessing

The dataset used in this study is based on SPARK phenotype V8, consisting of 117,099 participants with autism and 363 variables. It contains information extracted from standardized surveys and parent-reported medical history regarding children with autism. The following eight surveys with <80% missing rates in the full dataset (Table 1) were included in the missing data imputation assessment: individuals registration, basic medical screening, background history, SCQ, RBS-R, DCDQ, Child Behavior Checklist, and area deprivation index.

This dataset was first filtered to remove variables with extreme rates of missingness (~90% or greater), resulting in a drop of 22 variables. The dataset was then modified to remove any rows with missing information. This resulted in 15,196 participants with autism and 347 variables.

Table 1. Percentage of subjects who did not complete each individual survey among all 117,099 participants with autism in SPARK

Survey name	Percentage of subjects who did not complete corresponding survey (%)
Individuals registration	0
Basic medical screening	39.9
Background history	59.3
Area deprivation index	35.1
SCQ	51.3
RBS-R	63.8
DCDQ	72.9
Vineland	82.2
Intelligence quotient	95.3
CBCL	99.6

Note: SCQ: Social communication questionnaire; RBS-R: Repetitive behavior scale-revised; and DCDQ: Developmental coordination disorder questionnaire; are surveys commonly used to quantify the mental and behavioral functions at scale.
Abbreviation: CBCL: Child Behavior Checklist.

One-hot encoding was used to transform the categorical variables in this dataset, resulting in 15,196

participants with autism and 431 variables. The preprocess method from the caret package in R was used to normalize each variable with a mean of 0 and a standard deviation of 1. This was mainly to allow for comparable root mean squared error (RMSE) metrics across all variables that are commonly used in similar studies.[21,23,24]

This preprocessed complete dataset of participants with autism was used to simulate different missing data mechanisms and assess the accuracy or various imputation methods.

2.2. Three simulation scenarios for missing data mechanisms

Three simulation scenarios were constructed for missing data mechanisms in mental and behavioral surveys as outlined in Figure 2.

2.2.1. MCAR

The first missing data simulation scenario, referred to as MCAR, introduces missingness completely at random by converting a specific percentage of the preprocessed complete dataset to missing. To observe the imputation performance as the missing rate gradually increases, MCAR was implemented with missing rates from 10% to 90% in 10% intervals for all variables in the dataset.

2.2.2. MNAR: SMR

The second missing data simulation scenario is SMR, in which the proportion of missing values in each column is dependent on the survey type that it belongs to. SMR is tailored to mirror the missing rates in the full SPARK dataset by reusing the same proportions of missing values for each survey (Table 1).

2.2.3. MNAR: BSMR

The last missing data simulation scenario, referred to as BSMR, incorporates blockwise missingness with survey-specific missing rates. Instead of randomly selecting a specific portion of each column to be converted to missing as in SMR, a proportion of participants is randomly selected to have completely missing values for all surveys of a particular survey type. In other words, every column of a specific survey type contains the same missing rows. This resembles real data more closely when subjects skip the entire survey.

2.3. Machine learning imputation

For each missing data simulation scenario described in the previous section, multiple machine learning models were used to impute the missing values. The generated incomplete datasets were passed through the following imputation algorithms to compute the predicted values. A separate set of 10 datasets with 20% randomly selected missing values was used to conduct hyperparameter tuning on each of these models.

2.3.1. MICE

This study used the MICE[13] (version 3.16.0) package in R which employs a multiple imputation model. It uses a

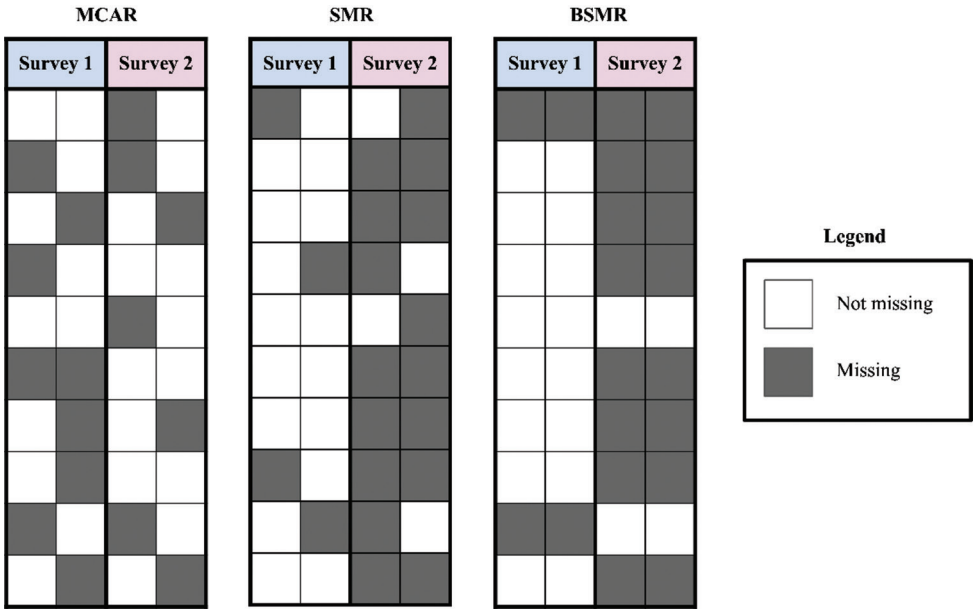


Figure 2. Visualization of the three missing data simulation scenarios explored in this study. On the left is Missing Completely at Random with a 40% missing rate. In the middle is Survey-specific missing rate with a 20% missing rate for Survey 1 and 80% missing rate for Survey 2. On the right is blockwise survey-specific missing rate with a 20% missing rate for Survey 1 and 80% missing rate for Survey 2.

concept called fully conditional specification, in which each incomplete variable is imputed by a different model. It generates multiple imputed datasets that are averaged to retrieve the final imputed data. Since MICEs employ a regression-based approach, hyperparameter tuning was not performed.

2.3.2. KNN

KNNImputer is a method in Python's Scikit-learn package[25] (version 0.22) and was used to study the KNN algorithm. KNNImputer predicts each sample's missing values using the average value from the closest data points in the training set. Hyperparameter tuning was used to select the optimal value for the number of nearest neighbors used during imputation.

2.3.3. MissForest

MissForest[16] (version 1.5) is an R package which uses a random forest approach to impute missing values, building multiple decision trees to make predictions using the other remaining features. By averaging several classification or regression trees, MissForest employs out-of-bag error estimates and can capture complex, non-linear relationships. Hyperparameter tuning was used to select the optimal values for the number of trees and the maximum number of iterations.

2.3.4. MIDAS

MIDASpy[26] (version 1.3.1) is a Python package that was used to study the MIDAS algorithm. It introduces additional missing values into a given dataset and restores these values using an unsupervised neural network called a denoising autoencoder. Then, the resulting model is used to predict the values of the original missing data. Similar to MICE, MIDASpy generates multiple imputed datasets that are averaged to retrieve the final imputed data. Hyperparameter tuning was used to select the optimal values for the input drop, layer structure, and number of epochs.

2.4. Evaluation of imputation performance

For each missing data simulation scenario, missingness was introduced into the complete dataset 10 different times as 10 separate trials. The values in Table 1 correspond to the percentage of subject IDs in the full dataset (with missing values among participants with autism) who are not present in each specific survey. These missing rates were used when generating the missing datasets for the SMR and BSMR simulation scenarios.

The four models were used to impute the missing data, and these imputed values were compared with the true values in the preprocessed complete dataset. In each imputation trial, the RMSE values were calculated for each column using the postResample method from the caret package (Version 6.0 – 94) in R. To retrieve the RMSE value for an imputed column, the following formula was used:

$$\text{RMSE} = \sqrt{\frac{\sum(\hat{y}_i - y_i)^2}{n}}$$

Where $\hat{y}_i$ are predicted values and y_i are observed values. As indicated by the equation, the square of the difference between the predicted and observed value was summed across each item in the column that was imputed. This value was then divided by the total number of imputed items and the square root of this value was stored as the column's RMSE.

These column-specific RMSEs were averaged across all columns in the dataset. Then, these RMSEs were again averaged across the 10 trials for each simulation setting. This resulted in a mean overall RMSE for each simulation scenario. These error values were then compared for every simulation scenario between each imputation method.

SCQ summary score, RBS-R summary score, and DCDQ summary score evaluate the social communication function, severity of repetitive behaviors, and motor functions, respectively, in study participants with autism. They were calculated based on corresponding questionnaires. The RMSE values of these specific mental and behavior summary scores were also compared between the four imputation methods across each simulation scenario.

Finally, the total computation time was assessed for the four imputation methods during the BSMR simulation scenario, which was chosen since it is closest in nature to missingness in real survey data.

3. Results

3.1. Overview of full dataset and missingness patterns

The full dataset used in this study consists of 117,099 study participants with autism. Slightly more than half of the participants (51.3%) did not complete SCQ survey, which screens for social functioning; 63.8% did not complete RBS-R survey on repetitive behaviors; and 72.9% did not complete DCDQ survey on motor functions (Table 1). A total of 34,067 participants have medium missing rates between 20% and 80% among 363 total questions (Table 2), 37,710 participants exhibit low missing rates (<20%), and 45,322 participants exhibit high missing rates (>80%, Table 2).

When compared to female participants, there are slightly more male participants with high and low missing

rates. Around 39% of male participants have high missing rates, which is slightly larger than the 37% of female participants, while 33.5% of male participants have low missing rates, and only around 28% of female participants have low missing rates.

For individuals between ages 2 and 18, around 22% of these participants have medium missing rates. The missing rates of these individuals are more concentrated toward extreme values since around 39% have either low or high missing rates or 22% exhibit medium missing rates. For individuals below 2 years of age, around 40% have medium missing rates. Around 62% of individuals above 18 years of age have medium missing rates, whereas nearly 0% exhibit low missing rates.

Close to half of the self-reported white participants, Native Hawaiian participants, and individuals who identified as "Multiple Races" have low missing rates. The rates of missingness for self-reported African American, Asian, and Native American individuals are concentrated toward the extreme values, with more than 30% exhibiting high missing rates, while <25% of the participants who were self-identified as White or "Multiple Races" reported high missing rates. Those who self-reported themselves as an "Other" race exhibit large amounts of missingness since around 66% have missing rates larger than 80%.

Table 2. Demographic characteristics of sample organized by low (<20%), medium (20 – 80%), and high (>80%) missing rate in SPARK

	Missing rate			*P*-value
	Low missing rate (<20%)	Medium missing rate (20 – 80%)	High missing rate (>80%)	
Number of Subjects	37,710 (32.2)	34,067 (29.1)	45,322 (38.7)	
Sex (%)				<0.001
Male	29460 (33.5)	24,030 (27.3)	34,412 (39.1)	
Female	8,250 (28.3)	10,037 (34.4)	10,910 (37.4)	
Age (%)				<0.001
<2 years	456 (28.5)	636 (39.7)	509 (31.8)	
2 – 5 years	9,773 (38.0)	6,189 (24.1)	9,726 (37.9)	
6 – 11 years	16,511 (39.1)	9,230 (21.9)	16,463 (39.0)	
12 – 18 years	10,966 (38.4)	6,217 (21.7)	11,401 (39.9)	
>18 years	4 (~0.0)	11,795 (62.0)	7,223 (38.0)	
Race (%)				<0.001
White	28,727 (47.3)	17,968 (30.0)	14,093 (23.2)	
African American	2,063 (37.8)	1373 (25.2)	2,021 (37.0)	
Asian	876 (35.0)	645 (25.7)	988 (39.4)	
Native American	180 (37.4)	141 (29.3)	160 (33.3)	
Native hawaiian	55 (43.0)	29 (22.7)	44 (34.4)	
Multiple races	4,155 (48.3)	2,203 (25.6)	2,249 (26.1)	
Other	1654 (4.2)	11,708 (30.0)	25,767 (65.9)	

Note: Proportion of missing variables for each subject was calculated in the full dataset of this study containing 117,099 total participants with autism.

3.2. Sample characteristics of complete dataset and simulation of three missingness patterns

To assess the imputation performance of the four popular missing data imputation methods (MICE, KNN, MissForest, and MIDAS), a preprocessed complete dataset with 15,196 participants with autism (Table 3, details in

Table 3. Sample characteristics in the preprocessed complete dataset containing 15,196 participants

	Number of observations (percentage) or mean (standard deviation)
Number of subjects	15,196
Sex (%)	
Male	11,901 (78.3)
Female	3,295 (21.7)
Age (%)	
<2 years	61 (0.4)
2 – 5 years	3,029 (19.9)
6 – 11 years	8,442 (55.6)
12 – 18 years	3,664 (24.1)
>18 years	0 (0.0)
Race (%)	
White	11,938 (78.6)
African American	656 (4.3)
Asian	331 (2.2)
Native American	71 (0.5)
Native Hawaiian	22 (0.1)
Multiple races	1,649 (10.9)
Other	529 (3.5)
Summary scores (mean [SD])	
SCQ score	21.72 (7.09)
RBS-R score	35.16 (20.50)
DCDQ score	37.87 (12.73)

Notes: This table includes the number of observations and percentage breakdowns of sex, age, and race as well as means and standard deviations for the summary scores of the; SCQ: Social Communication Questionnaire; RBS-R: Repetitive behavior scale-revised; and DCDQ: Developmental coordination disorder questionnaire.

Methods) was first obtained. Around 78% of participants with complete data are male and 22% are female. The male-to-female ratio is 3.5:1, which aligns with the sex ratio among subjects with autism in the general population. About half of the individuals with complete data are between 6 and 11 years of age. Only 0.4% of subjects are under 2 years of age while none are above 18. About 79% of participants were self-identified as white. The category with the second largest number of participants is "Multiple Races" (10.9%), followed by African American (4.3%), "Other" (3.5%), and Asian (2.2%). The number of participants who are Native American or Native Hawaiian are below 1%. In the preprocessed complete dataset, the SCQ, RBS-R, and DCDQ scores have average values of 21.72, 35.16, and 37.87, respectively.

All variables were standardized with a mean of zero and standard deviation of 1 so that the imputation error, calculated as RMSE, can be interpreted as the average deviation of the predicted scores from the true scores in units of standard deviation. To assess the performance of the missing data imputation methods, missing values were introduced to the preprocessed complete dataset with 15,196 participants with autism. First, to simulate the scenario on MCAR, a random subset of values across the entire dataset was converted to missing values. Ten incomplete datasets were generated for each missingness percentage (10 – 90%). Second, to examine the performance of the imputation methods on MNAR patterns, 10 incomplete datasets were randomly generated for the SMR and BSMR simulation scenarios separately. When doing so, the missing rates in the original SPARK dataset were used (Table 1) to reflect the missingness distribution present in the real data.

3.3. Performance of imputation on overall dataset

The four imputation methods were applied to the incomplete datasets in each of the three simulation scenarios (Figure 3). The imputed values were compared with the actual values in the complete dataset, and the RMSE values were calculated. RMSE can be interpreted as the average deviation of the predicted scores from the true scores in units of standard deviation since all variables were standardized. Lower RMSE values correspond to higher accuracy in missing value imputation.

In the MCAR scenario, the imputation error for all models generally rose as the missing rate increased. MissForest has the lowest overall RMSE (ranging between 0.73 and 1.0), outperforming the other methods especially when the missing rate was low (Figure 3, left panel). However, as the percentage of missing values increased, the performance of KNN and MIDAS became comparable to that of MissForest. MICE outperformed KNN and MIDAS between 20% and 60% of random missingness but performed considerably worse than all other models for the remaining missing rates.

In the MNAR scenarios, all models exhibited an increase in imputation error in the BSMR scenario when compared to SMR. MissForest produced the lowest error rate in the SMR scenario, with an RMSE of 0.83, but did not perform as well during the BSMR scenario that simulated blockwise missingness. MissForest also exhibited larger variations in

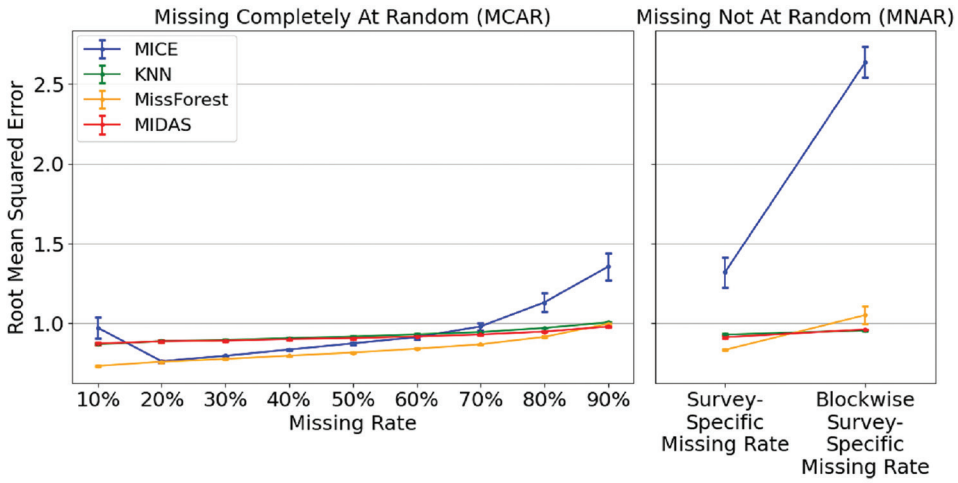


Figure 3. Evaluation of imputation performance based on overall root mean squared error (RMSE). Values across the 10 trials using the missing completely at random simulation scenario (left). Overall RMSE values across the 10 missing not-at-random trials in the survey-specific missing rate and blockwise missingness with Survey-specific missing rate simulation scenarios (right).
Abbreviations: KNN: K-Nearest Neighbors; MICE: Multiple Imputation by Chained Equations; MIDAS: Multiple Imputation with Denoising Autoencoders; MissForest: Non-parametric missing value imputation using Random Forest.

RMSE (standard deviation = 0.056) in the BSMR scenario than in the SMR scenario (standard deviation = 0.0043). For the BSMR scenario, KNN and MIDAS performed the best with an average RMSE of 0.96), outperforming the other methods especially when the missing rate was low (Figure 3, left panel). The variability of the RMSE was also relatively low for both methods, with a standard deviation of 0.0066 for KNN and 1e-6 for MIDAS. MICE performed worse than the other imputation methods in both SMR and BSMR scenarios. Especially in the BSMR scenario, the RMSE value was significantly higher at 2.64 with a relatively large standard deviation of 0.098.

For every simulation scenario, the difference in imputation performance on overall RMSE between KNN and MIDAS was marginal. Both models produced very similar results throughout the experiment and for each simulation scenario besides BSMR, they typically performed slightly worse than MissForest.

3.4. Performance of imputation on mental and behavioral summary scores

For every simulation scenario, the mean and standard deviations of RMSE values for the SCQ, RBS-R, and DCDQ scores were computed across the 10 trials as displayed in Figure 4. The relative performance of the four models was generally consistent across the three summary scores.

In the MCAR scenario, MissForest consistently outperformed KNN and MIDAS when imputing all three summary scores. The MICE model exhibited a steep incline in error as the missing rate was incremented. It performed the best until the missing rate was increased to 50%, after which it was surpassed by the remaining models. MICE is ideal for lower rates of random missingness but begins to perform exponentially worse as the rate gets larger. In fact, the MICE model produced the largest RMSE among the four methods at a 90% missing rate. For missing rates that are 50% and above, MissForest is the ideal model since it had the lowest errors among the four methods.

The MissForest model performed the best in the SMR scenario. However, each method, especially MICE and MissForest, exhibited error rates that rose sharply when the missing values became blocked by survey type in the BSMR scenario. In the BSMR scenario, KNN and MIDAS exhibited the lowest error rates with MissForest performing slightly worse. MICE performed considerably worse than the remaining models in the BSMR scenario.

3.5. Computational time

When comparing the computational times of the four models, the BSMR simulation scenario was used since this environment most closely resembles the missingness patterns in the real data when participants skip an entire survey in SPARK.

As shown in Figure 5, MIDAS and KNN not only had similar overall error rates but also exhibited comparable imputation times of around 10 – 13 min. MissForest had a median imputation time of slightly <30 min. On the other hand, MICE had a median imputation time of around 285 min, which was significantly larger than those of the remaining models. The difference in computational time between implementations in R and Python is negligible.[26]

4. Discussion

The establishment of biobank databases has enabled the collection of self-reported mental and behavioral surveys at scale.[1-3] SPARK has gathered social and behavioral survey data from about 100,000 individuals[1] and there is ongoing collection of more survey data on existing participants. UK Biobank has measurements on lifetime depressive disorder, cognitive function, attention, and impulsivity from about 150,000 participants.[2,27,28] All of Us also has strategic plans to collect mental and behavioral surveys at scale.[3] However, the data quality and statistical power are compromised by missing data. Recent advances in machine learning methods have inspired novel missing data imputation approaches with increased accuracy and computational efficiency.[13-16] Previous studies either have not reviewed these newly developed imputation methods or have not focused on assessing imputation accuracy in mental and behavioral surveys that exhibit blockwise missing structures.[18-22]

Our study provided insights on the missingness pattern in SPARK, a large-scale cohort with autism, and assessed the imputation accuracy and computational time of four popular missing data imputation methods—MICE, KNN, MissForest, and MIDAS. This was done by simulating three missingness scenarios in mental and behavioral surveys, including SCQ, RBS-R, and DCDQ. We observed that 50 – 70% of participants with autism did not complete SCQ, RBS-R, and DCDQ surveys and the dataset exhibited blockwise missing structures. The missing rates also varied by sex, age, and race. Overall, KNN and MIDAS showed relatively stable performance with increasing missing rate in the MCAR scenario and slightly higher imputation error when blockwise missingness is introduced in the MNAR scenarios. The error rate increased more significantly in MICE and MissForest in both MCAR and MNAR scenarios, with a particularly notable surge in error rate for MICE when blockwise missing structures were introduced. When imputing SCQ, RBS-R, and DCDQ summary scores in the MCAR scenario, MICE had the lowest error rate when the missing rate was low, while MissForest had the lowest

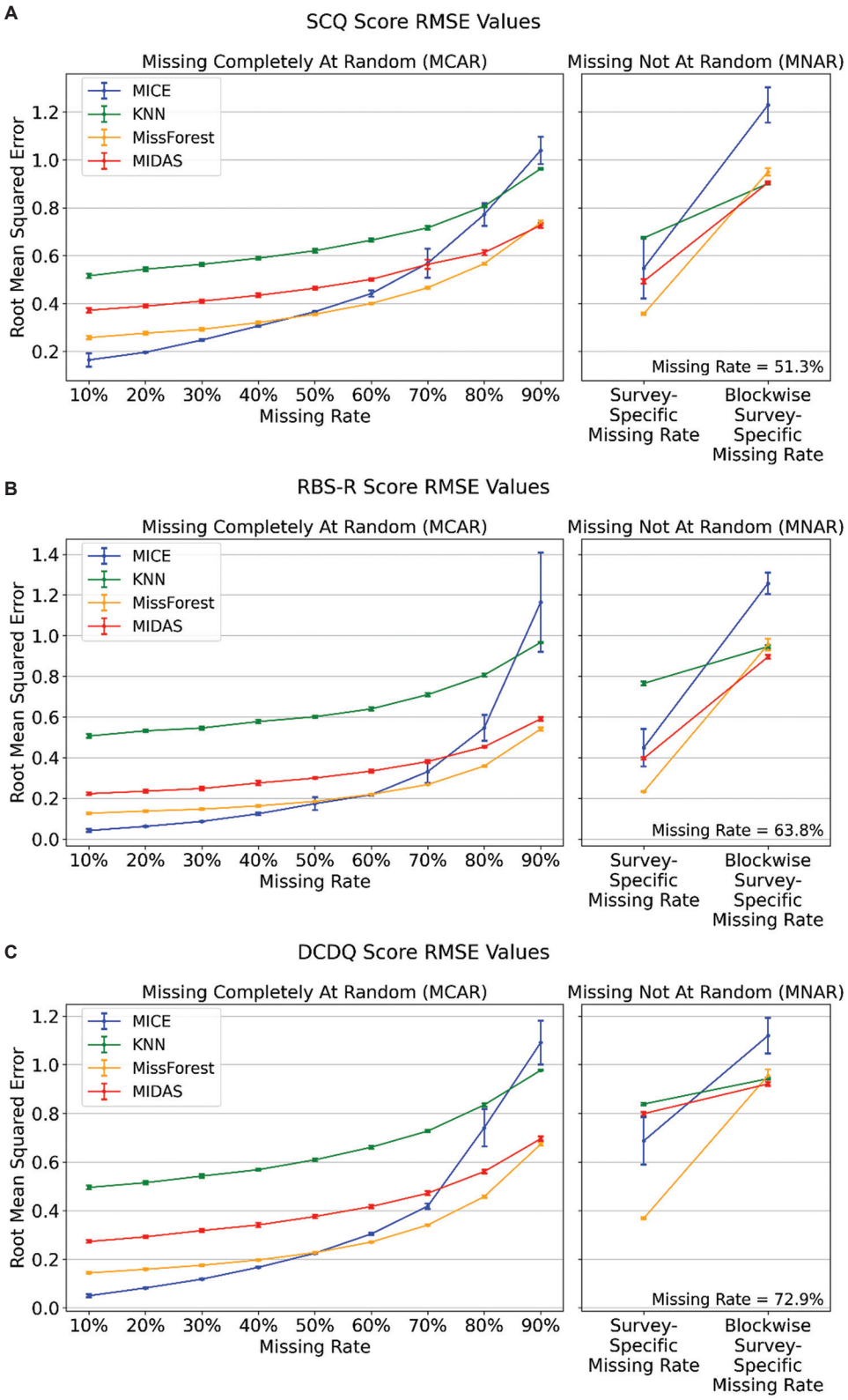


Figure 4. Imputation performance on summary scores from mental health surveys. Root mean squared error (RMSE) values for imputing social communication questionnaire scores (A), Repetitive Behavior Scale-Revised scores (B), Developmental coordination disorder questionnaire scores (C) across the Missing Completely at Random (MCAR) and missing not at random (MNAR) trials. RMSE values for the across the MCAR and MNAR trials.

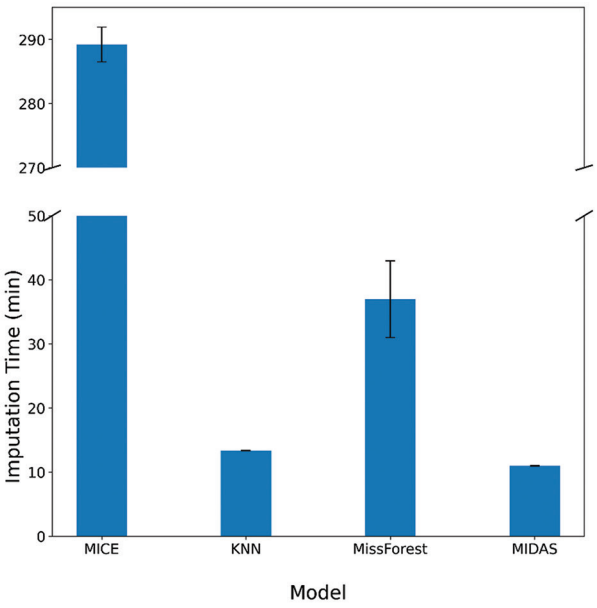


Figure 5. Total imputation times (in minutes) and standard deviations of each model for the 10 trials in the Blockwise Missingness with BSMR scenario. The total sample size is 15,196.
Abbreviations: KNN: K-Nearest Neighbors; MICE: Multiple Imputation by Chained Equations; MIDAS: Multiple Imputation with Denoising Autoencoders; MissForest: Non-parametric missing value imputation using Random Forest; BSMR: Survey-specific missing rate.

error rate when the missing rate was high. However, in the presence of blockwise missingness in the MNAR scenario, MIDAS was consistently the best-performing model across all three summary scores, with KNN and MissForest having similar or slightly higher error rates. The results of this study suggested that some models like MICE are sensitive to high missing rates and blockwise missing structures, while MIDAS and KNN may perform better in the overall dataset and specific summary scores in the presence of blockwise missingness. The average computational times were each 10 min for MIDAS and KNN to impute 15,196 subjects with blockwise missingness, about 35 min for MissForest, and about 290 min for MICE. These results highlight the computational efficiency in machine learning imputation algorithms even in highly complex neural network models in MIDAS. Newly developed imputation models have better optimization in their algorithms and take advantage of parallel computing to reduce the computational time.

Our results show the potential to impute missing data in large-scale databases with mental and behavioral surveys, especially imputing summary scores based on medical history and neurodevelopmental measures. When the data exhibits blockwise missingness, the imputation error increases, but models such as MIDAS and KNN can still provide imputed results that are relatively stable and accurate. This shows that when a block of correlated variables in one survey is completely missing, other related surveys or medical history can also provide relevant information for imputation. The choice of imputation methods may depend on the overall missing rate and missingness patterns in a dataset.

The strength of our study is that a large-scale collection of mental and behavioral surveys in SPARK was utilized to simulate the missingness patterns, particularly with blockwise missing structures that are commonly observed in mental health databases. This study also systematically assessed the latest missing data imputation approaches like MIDAS. The limitation is that the complete data with missing data simulation primarily comes from adolescents. Despite the inclusion of various racial groups in the simulation, most participants are white. Assessment in other types of large-scale mental and behavioral surveys with adults and minority groups is warranted for future studies.

Missing data imputation is widely used in national surveys with mental and behavioral surveys. For example, the National Survey on Drug Use and Health (NSDUH) has been providing imputation-revised variables by the predictive mean neighborhood methods since 1999.[29] There is also the recent phenotype imputation model developed in the UK Biobank, which has shown increased power for genetic studies.[30] As biobanks and national surveys collect more large-scale data on mental and behavioral surveys, missing data imputation will produce more accurate imputed values and become an integral part of analysis to maximize the use of the data.

5. Conclusion

Our study underscores the efficacy of advanced imputation techniques, such as MIDAS and KNN, in addressing missing data within large-scale mental and behavioral surveys. Our findings showcase that for similar databases with mental and behavioral surveys on autism, dementia, and other disorders, machine learning-based imputation methods can be leveraged to effectively recover missing information. This study demonstrates that machine learning methods offer increased performance and faster computation times over traditional algorithms. The performance of these advanced imputation techniques demonstrates their potential to optimize analyses and advance research in mental and behavioral disorders.

Acknowledgments

The authors are extremely grateful to the thousands of individuals and families who are participating in the SPARK. The authors also thank the sites, staff, and

volunteers of the SPARK Clinical Site Network and SFARI for their invaluable contributions.

Funding

This work is supported by Southern California Environmental Health Sciences Center pilot grant from NIH/NIEHS, grant number P30ES007048 (Rob McConnell), and The Tobacco-Related Disease Research Program, grant number T32IR5216 (Xuejuan Jiang) and NIH/NIA, grant number 1RF1AG076124-01A1 (Hussein Yassine).

Conflict of interest

The authors declare that they have no conflicts of interest.

Author contributions

Conceptualization: Chang Shu
Formal analysis: Preethi Prakash
Investigation: Preethi Prakash, Chang Shu
Methodology: Preethi Prakash, Kelly Street, Yufeng Shen, Chang Shu
Writing–original draft: Preethi Prakash, Chang Shu
Writing–review & editing: Kelly Street, Shrikanth Narayanan, Bridget A. Fernandez, Yufeng Shen, Chang Shu

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

SPARK Phenotype Dataset is accessible through an application at SFARI Base (https://base.sfari.org). All software used in this study is publicly available. The code for simulations and analysis can be found at https://github.com/AprilShuLab/MissingDataImputation.

Further disclosure

The paper has been uploaded to medRxiv (doi: 10.1101/2024.05.13.24307231).

References

1. Feliciano P, Daniels AM, Snyder LG, *et al.* SPARK: A US cohort of 50,000 families to accelerate autism research. *Neuron*. 2018;97:488-493.

   doi: 10.1016/j.neuron.2018.01.015

2. Davis KAS, Coleman JRI, Adams M, *et al.* Mental health in UK Biobank - development, implementation and results from an online questionnaire completed by 157 366 participants: A reanalysis. *BJPsych Open*. 2020;6:e18.

   doi: 10.1192/bjo.2019.100

3. Ramirez AH, Sulieman L, Schlueter DJ, *et al.* The all of Us research program: Data quality, utility, and diversity. *Patterns* (*N Y*). 2022;3:100570.

   doi: 10.1016/j.patter.2022.100570

4. Chesnut SR, Wei T, Barnard-Brak L, Richman DM. A meta-analysis of the social communication questionnaire: Screening for autism spectrum disorder. *Autism*. 2017;21:920-928.

   doi: 10.1177/1362361316660065

5. Hooker JL, Dow D, Morgan L, Schatschneider C, Wetherby AM. Psychometric analysis of the repetitive behavior scale-revised using confirmatory factor analysis in children with autism. *Autism Res.* 2019;12:1399-1410.

   doi: 10.1002/aur.2159

6. Van Damme T, Vancampfort D, Thoen A, Sanchez CPR, van Biesen D. Evaluation of the Developmental Coordination Questionnaire (DCDQ) as a screening instrument for co-occurring motor problems in children with autism spectrum disorder. *J Autism Dev Disord.* 2022;52:4079-4088.

   doi: 10.1007/s10803-021-05285-1

7. Jebb AT, Ng V, Tay L. A review of key likert scale development advances: 1995-2019. *Front Psychol*. 2021;12:637547.

   doi: 10.3389/fpsyg.2021.637547

8. Mirzaei A, Carter SR, Patanwala AE, Schneider CR. Missing data in surveys: Key concepts, approaches, and applications. *Res Soc Adm Pharm*. 2022;18:2308-2316.

   doi: 10.1016/j.sapharm.2021.03.009

9. Mack C, Su Z, Westreich D. *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide, Third Edition*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2018.

10. Khan SI, Hoque ASM. SICE: An improved missing data imputation technique. *J Big Data.* 2020;7:37.

    doi: 10.1186/s40537-020-00313-w

11. Phiwhorm K, Saikaew C, Leung CK, Polpinit P, Saikaew KR. Adaptive multiple imputations of missing values using the class center. *J Big Data.* 2022;9:52.

    doi: 10.1186/s40537-022-00608-0

12. De Goeij MCM, van Diepen M, Jager KJ, Tripepi G, Zoccali C, Dekker FW. Multiple imputation: Dealing with missing data. *Nephrol Dial Transplant.* 2013;28:2415-2420.

    doi: 10.1093/ndt/gft221

13. Van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. *J Stat Softw*. 2011;45:1-67.

doi: 10.18637/jss.v045.i03

14. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20:40-49.

    doi: 10.1002/mpr.329

15. Taunk K, De S, Verma S, Swetapadma A. A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. In: *2019 International Conference on Intelligent Computing and Control Systems* (*ICCS*). IEE; 2019.

16. Stekhoven DJ, Bühlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics.* 2011;28:112-118.

    doi: 10.1093/bioinformatics/btr597

17. Lall R, Robinson T. The MIDAS touch: Accurate and scalable missing-data imputation with deep learning. *Polit Anal.* 2022;30:179-196.

    doi: 10.1017/pan.2020.49

18. Shrive FM, Stuart H, Quan H, Ghali WA. Dealing with missing data in a multi-question depression scale: A comparison of imputation methods. *BMC Med Res Methodol*. 2006;6:57.

    doi: 10.1186/1471-2288-6-57

19. Peyre H, Leplège A, Coste J. Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey. *Qual Life Res.* 2011;20:287-300.

    doi: 10.1007/s11136-010-9740-3

20. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. *J Big Data*. 2021;8:140.

    doi: 10.1186/s40537-021-00516-9

21. Xu X, Xia L, Zhang Q, Wu S, Wu M, Liu H. The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Med Res Methodol.* 2020;20:42.

    doi: 10.1186/s12874-020-00932-0

22. Croy CD, Novins DK. Methods for addressing missing data in psychiatric and developmental research. *J Am Acad Child Adolesc Psychiatry.* 2005;44:1230-1240.

    doi: 10.1097/01.chi.0000181044.06337.6f

23. Lee JH, Huber JC Jr. Evaluation of multiple imputation with large proportions of missing data: How much is too much? *Iran J Public Health*. 2021;50:1372-1380.

    doi: 10.18502/ijph.v50i7.6626

24. Petrazzini BO, Naya H, Lopez-Bello F, Vazquez G, Spangenberg L. Evaluation of different approaches for missing data imputation on features associated to genomic data. *BioData Mining*. 2021;14:44.

    doi: 10.1186/s13040-021-00274-7

25. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825-2830.

26. Lall R, Robinson T. Efficient multiple imputation for diverse data in python and R: MIDASpy and rMIDAS. *J Stat Softw*. 2023;107:1-38.

    doi: 10.18637/jss.v107.i09

27. Fawns-Ritchie C, Deary IJ. Reliability and validity of the UK Biobank cognitive tests. *PLoS One*. 2020;15:e0231627.

    doi: 10.1371/journal.pone.0231627

28. Schweren LJS, van Rooij D, Shi H, *et al*. Diet, physical activity, and disinhibition in middle-aged and older adults: A UK biobank study. *Nutrients*. 2021;13:1607.

    doi: 10.3390/nu13051607

29. Grau E, Frechtel P, Odom D, Painter D. A Simple Evaluation of the Imputation Procedures Used in NSDUH. In: *Proceedings of the 2004 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods, Toronto, Ontario, Canada [CD-ROM*]. Alexandria, VA: American Statistical; 2004.

30. An U, Pazokitoroudi A, Alvarez M, *et al.* Deep learning-based phenotype imputation on population-scale biobank data increases genetic discoveries. *Nat Genet*. 2023;55:2269-2276.

    doi: 10.1038/s41588-023-01558-w

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Machine learning-driven prediction of EBNA1 inhibitors against Epstein–Barr virus in nasopharyngeal carcinoma

Lavinia Clarisa Wicklem[1], Siaw San Hwang[1], Bee Theng Lau[1], Mrinal Bhave[2], and Xavier Wezen Chee[1]*

[1]Science Programme, School of Engineering and Science, Swinburne University of Technology (Sarawak Campus), Kuching, Sarawak, Malaysia

[2]Department of Chemistry and Biotechnology, School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Melbourne, Victoria, Australia

Abstract

Nasopharyngeal carcinoma (NPC), particularly prevalent in regions such as Malaysia, is a significant health concern often linked to Epstein-Barr virus (EBV) infection. The EBV nuclear antigen 1 (EBNA1), crucial for EBV survival and NPC tumorigenicity, has emerged as a potential therapeutic target for EBV-positive NPC. In this study, we utilized quantitative structure-activity relationship (QSAR) models to predict potential inhibitors of EBNA1. These models were developed based on the molecular fingerprints of known EBNA1 inhibitors, using both classification and regression approaches. Our QSAR classification models demonstrated consistently high precision, recall, F1 score, and accuracy scores across the training set. The top-performing models, constructed using logistic regression algorithms, achieved perfect precision scores of 1.000 in the test set evaluation. These models' recall, F1 score, and accuracy scores were 0.571, 0.727, and 0.667, respectively. On the other hand, the best-performing model among the regression models was built using the sequential minimal optimization regression algorithm, achieving a correlation coefficient of 0.703. The mean absolute error and root mean square error of this QSAR regression model were 0.173 and 0.217, respectively, whereas the relative absolute error was 0.689. We screened the enamine advanced compound library using this regression model to predict compounds with potential EBNA1 inhibitory effects. This led to the identification of the top 10 compounds with the most promising predicted EBNA1 inhibitory properties.

Keywords: Epstein-Barr virus nuclear antigen 1; Nasopharyngeal carcinoma; Quantitative structure-activity relationship; Inhibitor; Machine learning; Compound screening

***Corresponding author:**
Xavier Wezen Chee
(xchee@swinburne.edu.my)

Citation: Wicklem LC, Hwang SS, Lau BT, Bhave M, Chee XW. Machine learning-driven prediction of EBNA1 inhibitors against Epstein–Barr virus in nasopharyngeal carcinoma. *Artif Intell Health*. 2025;2(1):93-104. doi: 10.36922/aih.4375

Received: July 30, 2024

Revised: September 10, 2024

Accepted: September 23, 2024

Published Online: November 8, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

The drug discovery process involves several stages, starting with the identification of disease targets and the search for small molecules that can modulate these targets. This often involves testing thousands to millions of compounds in various assays, with only a few progressing to animal testing and pre-clinical studies.[1] Conclusively, developing new and effective drugs is tedious, requiring millions of dollars and spanning over a decade.[2]

While high-throughput screening can identify active compounds at low concentrations, it often produces a low hit rate or high false positives.[3] This leads to a significant discrepancy between the number of hits identified and the number of viable lead compounds, which leads to wastage.

One way to negate this problem is using quantitative structure-activity relationship (QSAR) in drug discovery. QSAR is a ligand-based drug design method that uses mathematical models to correlate the chemical features of inhibitors to their bioactivity.[4] QSAR models streamline drug discovery by predicting compound activity based on their structure and properties, allowing researchers to prioritize promising candidates. This targeted approach reduces the need for extensive testing, saving time, resources, and materials while accelerating the research process. It optimizes resource allocation and promotes sustainable research practices by focusing efforts on compounds with a higher likelihood of success.[5] Another advantage is that QSAR aids in designing active molecules in a "greener" way by reducing the need for extensive experimental synthesis and testing on animals. A study examined the toxicity of various ionic liquids (ILs), which have the potential to harm aquatic life. Their study utilized advanced QSAR techniques to develop reliable models for predicting IL toxicity without animal testing. Their QSAR models demonstrated high predictive accuracy, with classification models achieving over 86% accuracy and regression models showing a correlation (R^2) >0.90 in the test data. These high-performance models provided strong predictions and pinpointed the structural elements of ILs contributing to their cytotoxicity. These QSAR models offer valuable tools for designing safer and environmentally friendly ILs.[6] Notably, QSAR-based virtual screening has emerged as a pivotal approach in contemporary scientific investigations, facilitating the identification of potential drug candidates. QSAR has been used to design chalcone derivatives that outperform standard tuberculosis drugs, identify potent neuraminidase inhibitors for influenza A, identify potent inhibitors for 5-HT_{1A} receptors for mood and anxiety disorders, and identify potential antimalarial activity in compounds that have low toxicity towards the mammalian cell.[7-10] QSAR was also used to identify critical structural features enhancing the inhibitory effects of compounds against liver carcinoma cells in tumor-targeting drug studies.[11] In antipsychotic/antidepressant studies, QSAR models have aided in predicting the activities of natural compounds against specific receptors, offering potential alternatives to synthetic drugs.[12] QSAR methodologies were also used to clarify physicochemical factors influencing the activity and cytotoxicity of compounds against human immunodeficiency viruses and influenza viruses in antiviral drug studies.[13,14]

Nasopharyngeal carcinoma (NPC) is strongly associated with Epstein-Barr virus (EBV). NPC typically affects individuals in their mid-40s and is more prevalent in men. It consistently exhibits EBV positivity, regardless of geographic location. Annually, NPC accounts for approximately 90,000 cases and 50,000 deaths recorded globally.[15] Its distribution is unique, with Asian countries representing around 80% of documented cases and mortality rates. In Malaysia, NPC ranks as the fourth most common cancer among males.[16] Among the Bidayuh community, part of Malaysia's indigenous population, the risk of NPC is notably elevated, with men and women experiencing a 2.3-fold and 1.9-fold increase, respectively, compared to other populations during the same period.[17] NPC poses a significant health concern, among which EBV latent infection stands out as a prominent contributor.

EBV is a virus capable of infecting epithelial and B cells, facilitating its persistence within the host and transmission among humans. A critical protein in maintaining viral stability and promoting viral gene expression is called the EBV nuclear antigen 1 (EBNA1). EBNA1 interacts with the oriP region of the EBV genome, forming dimers and complexes crucial for DNA looping and maintaining genome stability.[18,19] In addition, EBNA1 recruits cellular proteins to facilitate DNA replication.[20] EBNA1 binds to the Family of Repeats (FR) element during cell division, tethering EBV episomes to cellular chromosomes for proper segregation.[21-23] EBNA1 also activates EBV gene transcription by interacting with the FR element, with specific regions within EBNA1 being crucial for this function.[24] Moreover, EBNA1 affects several cellular signaling pathways in cell transformation and growth. It amplifies STAT1 signaling, enhances interferon responsiveness, and inhibits the transforming growth factor beta and nuclear factor kappa B pathways, ultimately promoting viral persistence and oncogenesis.[25,26] EBNA1 also disrupts promyelocytic nuclear bodies, impairing DNA repair, p53 activation, and apoptosis in response to DNA damage.[27] This disruption is mediated by interactions with cellular proteins ubiquitin-specific-processing protease 7 (USP7) and casein kinase 2, leading to promyelocytic leukemia protein degradation.[27-29] EBNA1 interacts with USP7, stabilizing its binding and preventing p53 stabilization protease.[30] Consequently, cells expressing EBNA1 exhibit reduced p53 accumulation upon DNA damage, facilitating cell survival and potentially contributing to tumorigenesis.[27] EBNA1 expression also correlates with increased oxidative stress, characterized by elevated reactive oxygen species (ROS) levels and DNA damage. This phenomenon, mediated by the upregulation of ROS-generating enzyme NADPH oxidase 2 (NOX2), may promote genomic instability and tumor development.[31,32]

Due to the involvement of EBNA1 in EBV's persistence and oncogenesis, we decided to deploy QSAR modeling to identify inhibitors targeting EBNA1. At present, QSAR applications in search of EBNA1 inhibitors remain unexplored in the current scientific literature. To bridge this gap, our research aims to identify potential compounds with inhibitory activities against EBNA1 using our QSAR models.

2. Data and methods

2.1. Dataset preparation

We developed the QSAR models using the AID2381 dataset obtained from a study by Gianti *et al.*[33] into molecular descriptors and fingerprints. All the compounds in the dataset demonstrated inhibitory activity toward EBNA1 through *in vitro* studies. The compounds in the database were experimentally evaluated using fluorescence polarization assay and were shown to inhibit EBNA1 selectively. First, we split the dataset into a training set and an external test set with a ratio of approximately 4:1. This yields a training set with 34 compounds and a test set with nine compounds. The compounds from these two datasets were then featured with chemical fingerprints using the PaDEL-Descriptor package. In total, 1024 chemical fingerprints were generated for each chemical compound in both datasets. After conversion into chemical fingerprints, we cleaned the dataset by removing empty rows and columns. In addition, we extracted the bioactivity of the ligands in pIC_{50} format.

2.2. Attribute selection

We constructed the QSAR models using the Waikato Environment of Knowledge Analysis (WEKA) package.[34] WEKA is a software consisting of an extensive collection of machine learning algorithms for data mining and exploration.[35] Before model construction, we performed attribute selection to identify the most relevant features for the model construction.[36] There are two parts to selecting the attributes: Attribute evaluation and search method. The attribute evaluation assesses each attribute related to the output variable within the dataset. We applied two methods of attribute evaluation: CfsSubsetEval (CFS) and ClassifierSubsetEval (CSE).

2.2.1. CFS

This method evaluates the worth of a subset of attributes by considering each feature's predictive ability and the degree of redundancy between them. Subsets of features highly correlated with the class while having low intercorrelation are preferred.[37] To select attributes, the attribute evaluator will employ a search method. The search method systematically explores various combinations of attributes within the dataset, aiming to identify a selection of preferred features. We used two search methods for CFS: Best first (BF) and greedy stepwise (GS). The BF method searches the attribute space by greedy hill climbing augmented with a backtracking facility, while the GS method performs a greedy forward or backward search through the space of attribute subsets.[38,39]

2.2.2. CSE

This method uses an algorithm to estimate the "merit" of attributes.[37] We used several algorithms for CSE to select the top attributes. For classification modeling, we employed algorithms Naïve Bayes (NB), instance-based learner (IBK), J48 Decision Tree (J48), random forest (RF), and logistic regression (LR). For regression modeling, we used linear regression (LRE), simple linear regression (SLR), sequential minimal optimization (SMO) regression, IBK, and RF algorithms. We also employed search methods BF and GS for CSE. For better visualization, we show the attribute selection process in this study (Figure 1).

2.3. Classification QSAR model

After the attribute selection process, we built the classification models using the NB, IBK, J48, RF, and LR algorithms.

2.3.1. Evaluation metrics for classification

The performance of the classification model was evaluated using standard metrics, including precision, recall, F1 score, and accuracy. Precision is a metric that evaluates the accuracy of correct predictions. It is calculated by dividing the number of accurate positive predictions by the total number of positive predictions.[40]

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

where TP is true positive, and FP is false positive.

The recall metric measures the number of actual observations predicted correctly. It is determined by dividing the number of correct positive predictions by the total number of actual positive instances.[40]

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

where TP is true positive, and FN is false negative.

F1 score is a metric that calculates the harmonic mean between precision and recall. The formula of F1 score, which provides a balanced measure of a model's performance, is given as follows:

$$\text{F1 score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} \times \text{recall}} \quad \text{(III)}$$

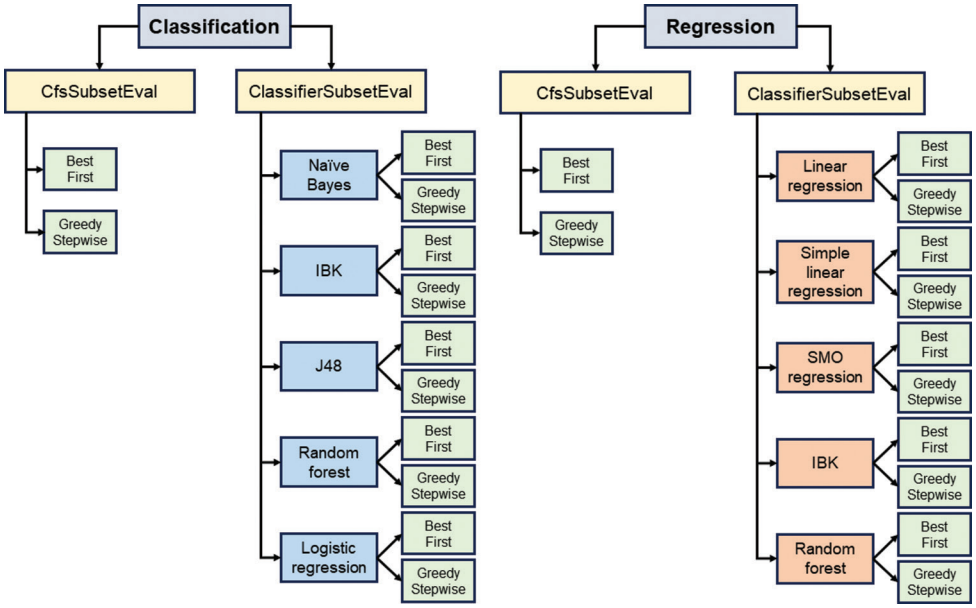


Figure 1. Quantitative structure-activity relationship model process
Abbreviations: IBK: Instance-based learner; SMO: Sequential minimal optimization.

Accuracy is the sum of two accurate predictions divided by the total number of data sets. It measures how often the classifier makes the correct prediction. It is the ratio between the number of correct predictions and the total number of forecasts.[40] We can calculate accuracy using the formula below.

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

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

These four evaluation classification metrics can provide a comprehensive understanding of the performance of the classification QSAR models.

2.4. Regression QSAR model

After the attribute selection process, we built the QSAR regression models using LRE, SLR, SMO regression, IBK, and RF algorithms.

2.4.1. Evaluation metrics for regression

We assessed the regression QSAR models' performance using correlation coefficient (r), mean absolute error (MAE), root mean squared error (RMSE), and relative absolute error (RAE) scores. The R score is a statistical measure of the strength of a linear relationship between two variables. The value of r ranges from −1 to 1. A negative score indicates an inverse correlation between the variables, whereas a positive score means the variables have a positive correlation.[41] Meanwhile, an r value close to 0 indicates a very weak or no linear correlation between the variables.[41] r is calculated as below.]

$$r = \frac{\sum\left[\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right]}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 * \sum\left(y_i - \bar{y}\right)^2}} \quad \text{(V)}$$

where: x_i: each individual x value; $\bar{x}$: mean of all x values; y_i: each individual y value; $\bar{y}$: mean of all y values

MAE score is calculated as the average of the absolute error values between the observed and predicted values. The score ranges from 1 being perfect to 0 being wrong.[42] MAE is calculated as below.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \check{y}_i\right| \quad \text{(VI)}$$

wheren represents the number of predictions; y_i represents the observed values; and $\hat{y}_i$ represents the predicted values.

RMSE is the squared root of the mean of all the errors, which describes the prediction magnitude error.[43] The scores range from 1 to 0, with lower scores preferred. RAE is determined by dividing the sum of absolute errors by the absolute difference between the mean and the actual value. The equation for RMSE is given in the following.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \quad \text{(VII)}$$

wheren represents the number of predictions; y_i represents the observed values; and $\hat{y}_i$ represents the predicted values.

RAE serves as a measure to assess the performance of a predictive model and is represented as a ratio. Lower RAE scores indicate a more effective mode.[44] The equation for calculating RAE is as follows.

$$RAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\sum_{i=1}^{n}|y_i - \bar{y}|} \quad \text{(VIII)}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{(IX)}$$

where n represents the number of observations; y_i represents the observed value; and $\bar{y}$ represents the average of observed values.

These four assessment regression metrics offer a thorough perspective on the performance of regression QSAR models.

2.5. Model deployment

After constructing the QSAR models, we validated all our models using the external test set. The chosen model was then deployed on the enamine advanced compound library. The enamine advanced compound library was similarly featurized with chemical fingerprints using the PaDEL-Descriptor package.

3. Results

3.1. Classification QSAR models

Our study yielded the following results for classification-based machine learning models (Table 1). CFS-LR-BF and CFS-LR-GS models exhibited precision scores of 1.000, recall scores of 0.952, F1 scores of 0.976, and accuracy scores of 0.976. In addition, the CFS-NB-BF and CFS-NB-GS models had precision, recall, and F1 scores all at 0.952 and accuracy scores at 0.953. The CSE-J48-LR-BF model achieved a precision score of 0.955, a recall score of 1.000, an F1 score of 0.977, and an accuracy score of 0.976. Meanwhile, the CSE-J48-IBK-BF model demonstrated a precision score of 0.952, a recall score of 0.952, an F1 score of 0.952, and an accuracy score of 0.953. We visualized the performance of these models using a confusion matrix (Figure 2).

We evaluated our models using an external test set comprising eight compounds (Table 2). The CFS-LR-BF and CFS-LR-GS QSAR classification models demonstrated precision scores of 1.000 and recall scores of 0.571. Both models achieved F1 scores of 0.727 and accuracy scores of 0.667. The CFS-NB-BF and CFS-NB-GS models exhibited precision scores of 1.000 and recall scores of 0.429. Both models achieved F1 scores of 0.600 and accuracy scores of 0.556. Finally, the CSE-J48-LR-BF and CSE-J48-IBK-BF models demonstrated precision scores of 1.000, with recall scores of 0.429. Both models achieved F1 scores of 0.600 and accuracy scores of 0.556. We also presented the results of the test set evaluation using a series of confusion matrices (Figure 3). These visual representations show the models' performance in classifying active and inactive compounds.

3.2. Regression QSAR models

For regression-based models, we obtained the following results. For the training set of CSE-LRE-BF-SMO and CSE-LRE-GS-SMO, both models achieved R scores of 0.992. Both models had MAE values of 0.029 and RMSE values of 0.037. The RAE values for both models were 0.118. For the training set of the CSE-SMO-BF-LRE and CSE-SMO-GS-LRE QSAR regression models, both models achieved R scores of 0.999. Both models had MAE values of 0.004 and RMSE values of 0.005. The RAE values for both models were 0.014. Regarding the training set results for the CSE-SMO-BF-SMO and CSE-SMO-GS-SMO QSAR regression models, we observed that both models achieved R scores of 0.999. Both models achieved MAE values of 0.008 and RMSE values of 0.010. The RAE values for both models were 0.032. We plotted the graphs of experimental pIC_{50} versus predicted pIC_{50} of the compounds in the training set (Figure 4). Consecutively, we evaluated the models on a test set to determine the predictive power of each model.

Table 1. Score for evaluation metric for the training set

	CFS-LR-BF	CFS-LR-GS	CFS-NB-BF	CFS-NB-GS	CSE-J48-LR-BF	CSE-J48-IBK-BF
Precision	1.000	1.000	0.952	0.952	0.955	0.952
Recall	0.952	0.952	0.952	0.952	1.000	0.952
F1 score	0.976	0.976	0.952	0.952	0.977	0.952
Accuracy	0.976	0.976	0.953	0.953	0.976	0.953

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

Figure 2. Confusion matrix for the training set results of (A) CFS-LR-BF, (B) CFS-LR-GF, (C) CSE-J48-BF-LR, (D) CFS-NB-BF, (E) CFS-NB-GS, and (F) CSE-J48-BF-IBK

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

Table 2. Score for evaluation metric for the test set

	CFS-LR-BF	CFS-LR-GS	CFS-NB-BF	CFS-NB-GS	CSE-J48-LR-BF	CSE-J48-IBK-BF
Precision	1.000	1.000	1.000	1.000	1.000	1.000
Recall	0.571	0.571	0.429	0.429	0.429	0.429
F1 score	0.727	0.727	0.600	0.600	0.600	0.600
Accuracy	0.667	0.667	0.556	0.556	0.556	0.556

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

For our external test set results, we observed that the CSE-LRE-BF-SMO and CSE-LRE-GS-SMO achieved R scores of 0.703 and 0.705, respectively. The MAE and RMSE values for both models were 0.173 and 0.217, respectively. Meanwhile, the RAE values for both models were 0.688 and 0.686, respectively. Both the CSE-SMO-BF-LRE and CSE-SMO-GS-LRE QSAR regression models achieved an R score of 0.703 in the test set. The MAE and RMSE values were 0.173 and 0.217, respectively. The RAE values for both models were 0.689. Moving on to the CSE-SMO-BF-SMO and CSE-SMO-GS-SMO QSAR regression models, both models achieved an R score of 0.703 in the test set. The MAE values for both models were 0.173 whereas the RMSE values for both models were 0.217. The RAE values for both models were 0.689. The outcomes of the test set evaluation are depicted through a table summarizing the different evaluation metrics (Table 3) and plots of actual pIC_{50} versus predicted pIC_{50} (Figure 5).

3.3. Deployment of model

Given that our target variable is the pIC_{50} of compounds, we decided to employ a modeling approach that provides numerical outcomes, namely the regression algorithm. Therefore, we chose to deploy the CSE-SMO-BF-LRE model on the enamine advanced library to predict their inhibitory activities against EBNA1. After the enamine advanced library compounds were featured with chemical fingerprints, we predicted their pIC_{50} against EBNA1 using the chosen regression model. The structures of the top 10 compounds are shown in Figure 6. Future work would involve purchasing these ten compounds for experimental validation.

4. Discussion

4.1. Classification QSAR models

We assessed our classification QSAR models' performance using four key metrics: Precision, recall, F1 score, and

Figure 3. Confusion matrix for the test set results of (A) CFS-LR-BF, (B) CFS-LR-GF, (C) CSE-J48-BF-LR, (D) CFS-NB-BF, (E) CFS-NB-GS, and (F) CSE-J48-BF-IBK

Abbreviations: BF: Best first; CFS: CfsSubsetEval; CSE: ClassifierSubsetEval; GS: Greedy stepwise; IBK: Instance-based learner; J48: J48 Decision Tree; LR: Logistic regression; NB: Naïve Bayes.

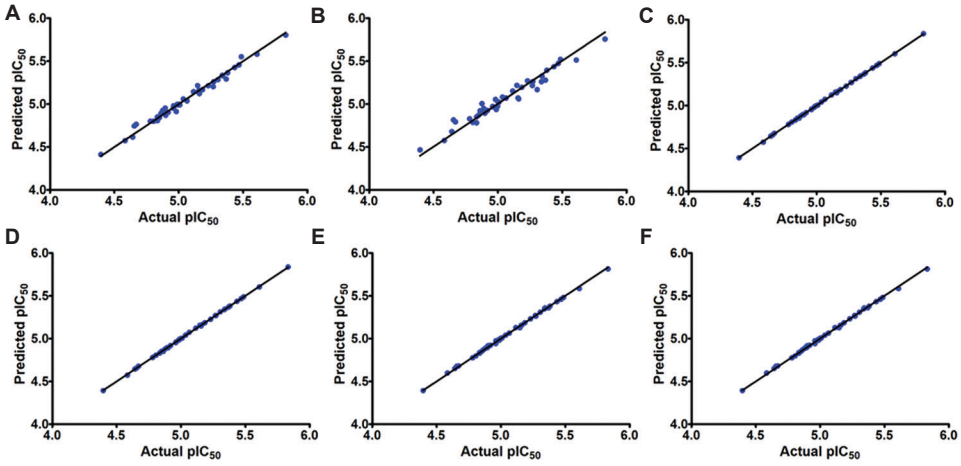


Figure 4. Graphical representation for the training set result for (A) CSE-LRE-BF-SMO, (B) CSE-LRE-GS-SMO, (C) CSE-SMO-BF-LRE, (D) CSE-SMO-GS-LRE, (E) CSE-SMO-BF-SMO, and (F) CSE-SMO-GS-SMO

Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; SMO: Sequential minimal optimization.

Table 3. Score for evaluation metric for the test set

	CSE-LRE-BF-SMO	CSE-LRE-GS-SMO	CSE-SMO-BF-LRE	CSE-SMO-GS-LRE	CSE-SMO-BF-SMO	CSE-SMO-GS-SMO
R	0.703	0.705	0.703	0.703	0.703	0.703
MAE	0.173	0.172	0.173	0.173	0.173	0.173
RMSE	0.217	0.217	0.217	0.217	0.217	0.217
RAE	0.688	0.686	0.689	0.689	0.689	0.689

Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; MAE: Mean absolute error; R: Correlation coefficient; RAE: Relative absolute error; RMSE: Root mean squared error; SMO: Sequential minimal optimization.

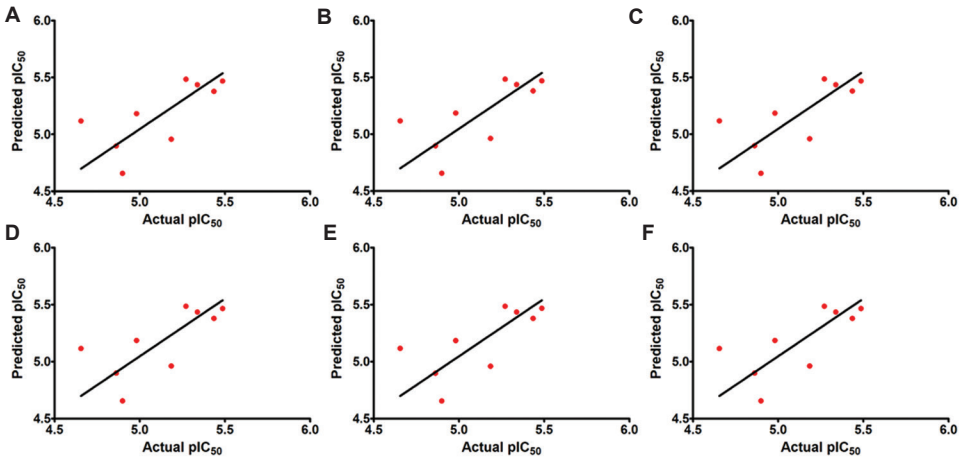


Figure 5. Graphical representation for the test set result for (A) CSE-LRE-BF-SMO, (B) CSE-LRE-GS-SMO, (C) CSE-SMO-BF-LRE, (D) CSE-SMO-GS-LRE, (E) CSE-SMO-BF-SMO, and (F) CSE-SMO-GS-SMO
Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; SMO: Sequential minimal optimization.

Table 4. Score for evaluation metric for the training set

	CSE-LRE-BF-SMO	CSE-LRE-GS-SMO	CSE-SMO-BF-LRE	CSE-SMO-GS-LRE	CSE-SMO-BF-SMO	CSE-SMO-GS-SMO
R	0.992	0.992	0.999	0.999	0.999	0.999
MAE	0.029	0.029	0.004	0.004	0.008	0.008
RMSE	0.037	0.037	0.005	0.005	0.010	0.010
RAE	0.118	0.118	0.014	0.014	0.032	0.032

Abbreviations: BF: Best first; CSE: ClassifierSubsetEval; GS: Greedy stepwise; LRE: Linear regression; MAE: Mean absolute error; R: Correlation coefficient; RAE: Relative absolute error; RMSE: Root mean squared error; SMO: Sequential minimal optimization.

accuracy. Our results highlighted two top-performing classification models, CFS-LR-BF and CFS-LR-GS. Both models exhibited high precision, recall, F1, and accuracy scores. In addition, the rest of the classification models also demonstrated strong performance (Figure 2). Our results showed that all six models accurately and successfully classified active and inactive compounds in the training set. During the external test set evaluation (Table 2), the CFS-LR-BF and CFS-LR-GS QSAR classification models demonstrated perfect precision scores of 1.000, indicating their precision in classifying a compound as active. However, their recall scores were moderate at 0.571, suggesting some active compounds might have been missed. Both models achieved F1 scores of 0.727 and accuracy scores of 0.667, indicating a balanced performance. The CFS-NB-BF and CFS-NB-GS models also exhibited perfect precision scores of 1.000, but their recall scores were lower at 0.429. Both models achieved consistent F1 scores of 0.600 and accuracy scores of 0.556. Finally, the CSE-J48-LR-BF and CSE-J48-IBK-BF models demonstrated perfect precision scores of 1.000, with moderately low recall scores of 0.429. Both models achieved consistent F1 scores of 0.600 and accuracy scores of 0.556. The consistently high precision scores across all models indicate their ability to identify active compounds correctly. However, the variability in recall scores suggests differences in their abilities to capture all true positive instances. While the models excel in minimizing false positive predictions, they may have limitations in identifying all active compounds in the dataset. Considering the scores of all models, we suggest that CFS-LR-BF and CFS-LR-GS are the top QSAR models for classification tasks.

4.2. Regression QSAR models

The performance of our regression-based QSAR models was evaluated using several key metrics: The correlation coefficient (R), MAE, RMSE, and RAE (Table 4). Based on the training set scores for the QSAR regression models, all models achieved high R scores with low MAE and RMSE values. Consequently, all the regression QSAR models demonstrated excellent predictive performance, with high correlation, low error rates, and minimal relative error in the training set. However, a good model cannot be determined solely by good scores on the training set. Therefore, we also evaluated the models on a test set to determine the predictive power of each model. Based on our external test set results, we observed that the CSE-LRE-BF-SMO and CSE-LRE-GS-SMO regression QSAR models

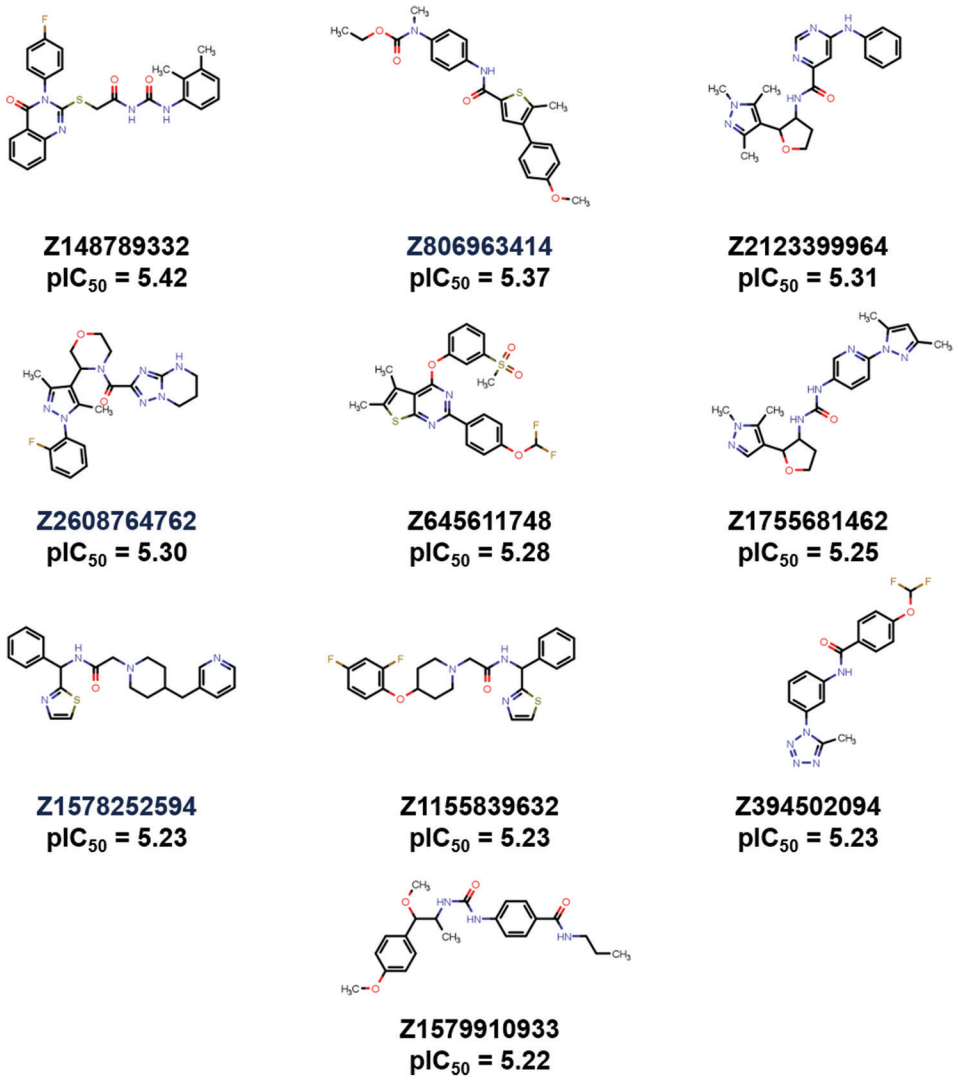


Figure 6. Two-dimensional chemical structures of chosen compounds generated using MarvinSketch 23.12

performed the best. Both models achieved R scores of 0.703 and 0.705, respectively. The MAE and RMSE values for both models were low, with MAE values of 0.173 and RMSE values of 0.217. These error values suggest that the models' predictions deviate from the actual values by a small amount.[45] Meanwhile, the RAE values for both models were moderate, with values of 0.688 and 0.686, respectively. The RAE scores suggest that the models' predictions deviate from the actual values by a moderate percentage relative to the scale of the target variable. For the CSE-SMO-BF-LRE and CSE-SMO-GS-LRE regression QSAR models, both models achieved an R score of 0.703 in the test set. The MAE and RMSE values for both models were low, with MAE values of 0.173 and RMSE values of 0.217. The RAE values for both models were also moderate, at 0.689. Moving on to the CSE-SMO-BF-SMO and CSE-SMO-GS-SMO regression QSAR models, both models achieved an R score of 0.703 in the test set. The MAE and RMSE values for both models were low, with MAE values of 0.173 and RMSE values of 0.217. The RAE values for both models were also moderate, at 0.689. The outcomes of the test set evaluation are depicted through a table summarizing the different evaluation metrics (Table 3) and plots of actual pIC_{50} versus predicted pIC_{50} (Figure 5).

5. Conclusion

This study highlights the potential of QSAR modeling in identifying candidate compounds for inhibiting EBNA1, a key target in addressing EBV-associated diseases such as NPC. Our findings demonstrated that QSAR classification models, particularly CFS-LR-BF and CFS-LR-GS, exhibit strong precision, albeit with moderate recall. This suggests

their effectiveness in identifying active compounds while minimizing false positives. Despite the moderate recall, their balanced F1 scores and moderate accuracy indicate good performance. Similarly, the CSE-SMO-BF-LRE QSAR model captured the relationship between compound bioactivity and chemical fingerprints. Using QSAR for our drug screening process optimized resource allocation and reduced the need for extensive experimental synthesis, aligning with sustainable research practices. Furthermore, our QSAR-based screening of the Enamine Advanced compound library predicted the top 10 compounds with potential inhibitory effects against EBNA1. Further experimental validation of these predicted inhibitors is needed to confirm their efficacy and safety, paving the way for potential therapeutic interventions against EBV-positive NPC.

Acknowledgments

L.C.W. was supported by the Swinburne Sarawak Postgraduate Scholarship.

Funding

This work was supported by the MAKNA Cancer Research Award 2021 given to Xavier Wezen Chee.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Xavier Wezen Chee
Formal analysis: Lavinia Clarisa Wicklem, Bee Theng Lau, Xavier Wezen Chee
Investigation: Lavinia Clarisa Wicklem, Xavier Wezen Chee
Methodology: Siaw San Hwang, Bee Theng Lau, Mrinal Bhave, Xavier Wezen Chee
Writing – original draft: Lavinia Clarisa Wicklem
Writing – review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data will be made available from the corresponding authors upon reasonable request.

Further disclosure

The authors would like to disclose that part of the findings included in the manuscript have been presented on (date: 3 August 2023) at the Swinburne Sarawak Postgraduate Research Conference 2023, (location: Swinburne University of Technology Sarawak).

References

1. Singh N, Vayer P, Tanwar S, Poyet JL, Tsaioun K, Villoutreix BO. Drug discovery and development: Introduction to the general public and patient groups. *Front Drug Discov.* 2023;3:1201419.

   doi: 10.3389/fddsv.2023.1201419

2. Sun J, Warden AR, Ding X. Recent advances in microfluidics for drug screening. *Biomicrofluidics*. 2019;13(6):061503.

   doi: 10.1063/1.5121200

3. Thorne N, Auld DS, Inglese J. Apparent activity in high-throughput screening: Origins of compound-dependent assay interference. *Curr Opin Chem Biol.* 2010;14(3):315-324.

   doi: 10.1016/j.cbpa.2010.03.020

4. Hansch C, Fujita T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc.* 1964;86(8):1616-1626.

   doi: 10.1021/ja01062a035

5. Chatterjee A. 27 - Computational methods and tools for sustainable and green approaches in drug discovery. In: Banik BK, editor. *Green Approaches in Medicinal Chemistry for Sustainable Drug Design.* Amsterdam: Elsevier; 2020. p. 965-988.

   doi: 10.1016/B978-0-12-817592-7.00027-7

6. Gupta S, Basant N, Singh KP. Nonlinear QSAR modeling for predicting cytotoxicity of ionic liquids in leukemia rat cell line: An aid to green chemicals designing. *Environ Sci Pollut Res.* 2015;22:12699-12710.

   doi: 10.1007/s11356-015-4526-3

7. Gomes MN, Braga RC, Grzelak EM, *et al*. QSAR-driven design, synthesis and discovery of potent chalcone derivatives with antitubercular activity. *Eur J Med Chem.* 2017;137:126-138.

   doi: 10.1016/j.ejmech.2017.05.026

8. Lian W, Fang J, Li C, Pang X, Liu AL, Du GH. Discovery of influenza a virus neuraminidase inhibitors using support vector machine and Naïve Bayesian models. *Mol Divers.* 2016;20(2):439-451.

   doi: 10.1007/s11030-015-9641-z

9. Luo M, Wang XS, Roth BL, Golbraikh A, Tropsha A. Application of quantitative structure-activity relationship models of 5-HT1A receptor binding to virtual screening identifies novel and potent 5-HT1A ligands. *J Chem Inf Model.* 2014;54(2):634-647.

   doi: 10.1021/ci400460q

10. Zhang L, Fourches D, Sedykh A, *et al*. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. *J Chem Inf Model.* 2013;53(2):475-492.

doi: 10.1021/ci300421n

11. Kamano Y, Yamashita A, Nogawa T, *et al*. QSAR evaluation of the Ch'an Su and related bufadienolides against the colchicine-resistant primary liver carcinoma cell line PLC/PRF/5. *J Med Chem*. 2002;45(25):5440-5447.

doi: 10.1021/jm0202066

12. Avram S, Stan MS, Udrea AM, Buiu C, Boboc AA, Mernea M. 3D-ALMOND-QSAR models to predict the antidepressant effect of some natural compounds. *Pharmaceutics*. 2021;13(9):1449.

doi: 10.3390/pharmaceutics13091449

13. Ravichandran V, Jain A, Mourya V, Agrawal RK. Prediction of anti-HIV activity and cytotoxicity of pyrimidinyl and triazinyl amines: A QSAR study. *Chem Pap.* 2008;62:596-602.

doi: 10.2478/s11696-008-0072-5

14. Yuan H, Parrill AL. QSAR studies of HIV-1 integrase inhibition. *Bioorg Med Chem.* 2002;10(12):4169-4183.

doi: 10.1016/s0968-0896(02)00332-2

15. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394-424.

doi: 10.3322/caac.21492

16. Azizah AM, Hashimah B, Nirmal K, *et al*. *Malaysia National Cancer Registry Report* (*MNCR*). Putrajaya, Malaysia: National Cancer Institute, Ministry of Health; 2019.

17. Devi BCR, Pisani P, Tang TS, Parkin DM. High incidence of nasopharyngeal carcinoma in native people of Sarawak, Borneo Island. *Cancer Epidemiol Biomarkers Prev.* 2004;13(3):482-486.

18. Yates J, Warren N, Reisman D, Sugden B. A cis-acting element from the Epstein-Barr viral genome that permits stable replication of recombinant plasmids in latently infected cells. *Proc Natl Acad Sci U S A*. 1984;81(12):3806-3810.

doi: 10.1073/pnas.81.12.3806

19. Frappier L, O'Donnell M. Epstein-barr nuclear antigen 1 mediates a DNA loop within the latent replication origin of Epstein-Barr virus. *Proc Natl Acad Sci U S A*. 1991;88(23):10875-10879.

doi: 10.1073/pnas.88.23.10875

20. Chaudhuri B, Xu H, Todorov I, Dutta A, Yates JL. Human DNA replication initiation factors, ORC and MCM, associate with oriP of Epstein-Barr virus. *Proc Natl Acad Sci U S A*. 2001;98(18):10085-10089.

doi: 10.1073/pnas.181347998

21. Harris A, Young BD, Griffin BE. Random association of Epstein-Barr virus genomes with host cell metaphase chromosomes in Burkitt's lymphoma-derived cell lines. *J Virol.* 1985;56(1):328-332.

doi: 10.1128/JVI.56.1.328-332.1985

22. Petti L, Sample C, Kieff E. Subnuclear localization and phosphorylation of Epstein-Barr virus latent infection nuclear proteins. *Virology*. 1990;176(2):563-574.

doi: 10.1016/0042-6822(90)90027-o

23. Lee MA, Diamond ME, Yates JL. Genetic evidence that EBNA-1 is needed for efficient, stable latent infection by Epstein-Barr virus. *J Virol.* 1999;73(4):2974-2982.

doi: 10.1128/jvi.73.4.2974-2982.1999

24. Lupton S, Levine AJ. Mapping genetic elements of Epstein-Barr virus that facilitate extrachromosomal persistence of Epstein-Barr virus-derived plasmids in human cells. *Mol Cell Biol*. 1985;5:2533-2542.

doi: 10.1128/mcb.5.10.2533-2542.1985

25. Wood VHJ, O'Neil JD, Wei W, Stewart SE, Dawson CW, Young LS. Epstein-Barr virus-encoded EBNA1 regulates cellular gene transcription and modulates the STAT1 and TGFbeta signaling pathways. *Oncogene.* 2007;26(28):4135-4147.

doi: 10.1038/sj.onc.1210496

26. Valentine R, Dawson CW, Hu C, *et al*. Epstein-Barr virus-encoded EBNA1 inhibits the canonical NF-κB pathway in carcinoma cells by inhibiting IKK phosphorylation. *Mol Cancer.* 2010;9:1.

doi: 10.1186/1476-4598-9-1

27. Sivachandran N, Sarkari F, Frappier L. Epstein-Barr nuclear antigen 1 contributes to nasopharyngeal carcinoma through disruption of PML nuclear bodies. *PLoS Pathog.* 2008;4(10):e1000170.

doi: 10.1371/journal.ppat.1000170

28. Scaglioni PP, Yung TM, Cai LF, *et al*. A CK2-dependent mechanism for degradation of the PML tumor suppressor. *Cell.* 2006;126(2):269-283.

doi: 10.1016/j.cell.2006.05.041

29. Sivachandran N, Cao JY, Frappier L. Epstein-Barr virus nuclear antigen 1 Hijacks the host kinase CK2 to disrupt PML nuclear bodies. *J Virol.* 2010;84(21):11113-11123.

doi: 10.1128/JVI.01183-10

30. Holowaty MN, Zeghouf M, Wu H, *et al*. Protein profiling with Epstein-Barr nuclear antigen-1 reveals an interaction with the herpesvirus-associated ubiquitin-specific protease HAUSP/USP7. *J Biol Chem*. 2003;278(32):29987-29994.

doi: 10.1074/jbc.M303977200

31. Gruhne B, Sompallae R, Marescotti D, Kamranvar SA,

Gastaldello S, Masucci MG. The Epstein-Barr virus nuclear antigen-1 promotes genomic instability via induction of reactive oxygen species. *Proc Natl Acad Sci U S A*. 2009;106(7):2313-2318.

doi: 10.1073/pnas.0810619106

32. Cao JY, Mansouri S, Frappier L. Changes in the nasopharyngeal carcinoma nuclear proteome induced by the EBNA1 protein of Epstein-Barr virus reveal potential roles for EBNA1 in metastasis and oxidative stress responses. *J Virol*. 2012;86(1):382-394.

doi: 10.1128/JVI.05648-11

33. Gianti E, Messick TE, Lieberman PM, Zauhar RJ. Computational analysis of EBNA1 "druggability" suggests novel insights for Epstein-Barr virus inhibitor design. *J Comput Aided Mol Des*. 2016;30(4):285-303.

doi: 10.1007/s10822-016-9899-y

34. Bouckaert RR, Frank E, Hall M. *WEKA Manual for Version 3-9-1*. Hamilton, New Zealand: University of Waikato; 2016. p. 1-341.

35. Holmes G, Donkin A, Witten IH. Weka: A Machine Learning Workbench. In: *Proceedings of ANZIIS'94-Australian New Zealand Intelligent Information Systems Conference*. IEEE; 1994. p. 357-361.

36. Kononenko I, Hong SJ. Attribute selection for modelling. *Future Gener Comput Syst*. 1997;13(2):181-195.

doi: 10.1016/S0167-739X(97)81974-7

37. Hall MA. *Correlation-based Feature Subset selection for Machine Learning*. Thesis Submitted in Partial Fulfilment of the Requirements of the Degree of Doctor of Philosophy at the University of Waikato; 1988.

38. Hall M, Guetlein M. *BestFirst*; 2019. Available from: https://weka.sourceforge.io/doc.dev/weka/attributeselection/bestfirst.html [Last accessed on 2024 Nov 07].

39. Hall M. *GreedyStepwise*; 2019. Available from: https://weka.sourceforge.io/doc.dev/weka/attributeselection/greedystepwise.html [Last accessed on 2024 Nov 07].

40. Vujović Ž. Classification model evaluation metrics. *Int J Adv Comput Sci Appl*. 2021;12(6):599-606.

doi: 10.14569/IJACSA.2021.0120670

41. Ratner B. The correlation coefficient: Its values range between +1/−1, or do they? *Journal of Target Meas Anal Mark*. 2009;17(2):139-142.

doi: 10.1057/jt.2009.5

42. Tatachar AV. Comparative assessment of regression models based on model evaluation metrics. *Int J Innov Technol Explor Eng*. 2021;8(9):853-860.

43. Gill J, Moullet M, Martinsson A, *et al*. Evaluating the performance of machine-learning regression models for pharmacokinetic drug-drug interactions. *CPT Pharmacometrics Syst Pharmacol*. 2023;12(1):122-134.

doi: 10.1002/psp4.12884

44. Damodharan S, Reddy SV, Sarojamma B. WEKA models for rainfall data. *Int J Emerg Technol Innovat Res*. 2022;9:C111-C119.

45. Tropsha A, Gramatica P, Gombar VK. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. *QSAR Comb Sci*. 2003;22(1):69-77.

doi: 10.1002/qsar.200390007

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

A machine learning approach to unravel client and program-specific effects in opioid treatment retention

Yinfei Kong[1]*, Erick Guerrero[2], Jemima Frimpong[3], Tenie Khachikian[4], Suojin Wang[5], Thomas D'Aunno[6], and Daniel Howard[4]

[1]Department of Information Systems and Decision Sciences, College of Business and Economics, California State University, Fullerton, CA, United States of America

[2]Research to End Health Disparities Corp, I-Lead Institute, Los Angeles, CA, United States of America

[3]New York University Stern School of Business, New York University Abu Dhabi, Saadiyat Island, Abu Dhabi, United Arab Emirates

[4]Department of Psychological and Brain Sciences, College of Arts and Sciences, Texas A&M University, College Station, TX, United States of America

[5]Department of Statistics, College of Arts and Sciences, Texas A&M University, College Station, TX, United States of America

[6]Health Policy and Management, Robert F. Wagner Graduate School of Public Service, New York University, New York, NY, United States of America

***Corresponding author:**
Yinfei Kong
(yikong@fullerton.edu)

Citation: Kong Y, Guerrero E, Frimpong J, *et al.* A machine learning approach to unravel client and program-specific effects in opioid treatment retention. *Artif Intell Health*. 2025;2(1):105-113. doi: 10.36922/aih.3750

Received: May 24, 2024

Revised: September 10, 2024

Accepted: October 25, 2024

Published Online: November 14, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract

This study examines the impact of workforce diversity, particularly the presence of Black/African American staff, on client retention in opioid use disorder (OUD) treatment, recognizing the historically low retention rates among Black and Hispanic populations in such programs. Using a novel machine learning technique called "causal forest," we explored the heterogeneous treatment effects of staff diversity on client retention, aiming to identify strategies that enhance client retention and improve treatment outcomes. Analyzing data from four waves of the National Drug Abuse Treatment System Survey spanning the years 2000, 2005, 2014, and 2017 ($n = 627$), we focus on the relationship between workforce diversity and retention. The findings revealed diversity-related variations in retention across 61 out of 627 OUD treatment programs (<10%), with potential beneficial effects attenuated by other program characteristics. These characteristics include programs that are more likely to be private-for-profit, have lower percentages of Black and Latino clients, lower staff-to-client ratios, higher proportions of staff with graduate degrees, and lower percentages of unemployed clients. Our results suggest that workforce diversity alone is insufficient for improving retention. Programs with characteristics linked to greater retention are better positioned to leverage a diverse workforce to enhance retention, offering important implications for policy and program design to better support Black clients with OUDs.

Keywords: Workforce diversity; Opioid use disorder; Treatment retention; Causal forest; Heterogeneous treatment effect

1. Introduction

The opioid epidemic continues to adversely impact the public health system of the United States. The Centers for Disease Control and Prevention estimates that there were over 81,000 opioid-related overdose deaths in 2023.[1] Increased opioid use disorder (OUD) treatment retention can improve treatment outcomes, including reduced rates of mortality and of relapse.[2-5] Concurrently, retention rates in OUD treatment are highly variable between programs and demographic groups, with 6-month retention rates commonly dropping below 50% for some groups.[4,6] Several studies measuring retention in OUD programs have found lower retention rates among minoritized individuals who identify as Black/African American and as Latino/Hispanic (Black and Latinos, hereafter).[7-9] Other studies have identified subgroup differences between Black, Latino, and White clients, including variations in predictors of retention and the treatment outcomes associated with retention.[10,11] It is therefore important to consider unique differences, particularly of minoritized patients like Black clients, when exploring strategies to boost retention rates in OUD programs.

Past research on the effect of culturally responsive practices on the retention of Black OUD clients has identified promising culturally responsive organizational factors, including offering bilingual language services; developing specific policies and procedures designed to serve minority clients; and having managers who believe in the importance of cultural sensitivity.[12-16] Workforce diversity, defined as having a higher percentages of Black staff members, is thought to improve Black OUD clients' treatment outcomes by fostering a culturally responsive treatment environment.[13,17-20] However, previous studies on the impacts of workforce diversity on OUD client retention have looked for simple associations and have included only a few basic modifying variables, leading to variable retention outcomes.[16,21]

The heterogeneous nature of these results indicates that workforce diversity may have differential impacts on retention rates in OUD programs with different organizational characteristics. We build on prior studies that have suggested that workforce diversity in the absence of other factors, such as high levels of training and education among staff members, may be insufficient to improve treatment outcomes.[17,18] Unpacking the heterogeneity in associations between workforce diversity and treatment retention can help healthcare policymakers, leaders of OUD treatment programs, and researchers to understand which programs would benefit most from the expansion of workforce diversity, and importantly, the additional conditions necessary to optimize the benefits of workforce diversity.

In this study, we applied heterogeneous treatment effect (HTE) estimation methods to understand which workforce diversity characteristics facilitate positive retention effects. HTE estimation is a machine learning method which was originally designed to study variations in the effects of clinical interventions and has been generalized to other applications such as public policy and marketing.[22-25] Heterogeneous treatment effect (HTE) estimation methods, including causal forests, have been effectively applied in fields such as personalized medicine, public policy, and marketing.[26,27] In personalized medicine, HTE helps tailor treatments to individual patients, improving outcomes by accounting for diverse responses. In public policy, it identifies how different populations are impacted by interventions, guiding more equitable policymaking. In marketing, HTE enables businesses to optimize strategies by understanding how various customer segments respond to different campaigns. The strength of HTE methods lies in their ability to handle complex interactions and high-dimensional data, offering deeper insights than traditional regression models.

In this work, we adopted a state-of-the-art HTE estimation method called "causal forest," to examine the heterogeneous impact of workforce diversity on OUD treatment retention.[28,29] Causal forest is a machine learning method that extends the random forest framework to estimate the varying effects of a treatment across different subgroups within a population. This method involves constructing an ensemble of decision trees, where each tree is specifically designed to identify splits in the data that reveal differences in treatment effects between subpopulations. To ensure accurate and unbiased estimates, causal forest uses a technique known as "honesty," where the data used to determine the optimal splits in the trees is separate from the data used to estimate the treatment effects. This approach allows for a detailed exploration of how the impact of a treatment may differ across various segments of the data.

There are several advantages of this method over traditional regression models. First, due to potential high collinearity and a high false discovery rate, only a limited number of interactions can be included in traditional regression models. Second, causal forest provides variance for individually-estimated treatment effects, that is, one can calculate the asymptotic p-values for the statistical significance of treatment effects for each observation.

By examining HTE, we can untangle the various factors that may influence how workforce diversity impacts OUD client retention. The benefit of this study to the field of healthcare, and disparities within this field in particular, includes informing healthcare policies, and practices, on

which program characteristics can be adjusted to maximize the benefits of workforce diversity for OUD client retention. This study is also of relevance to the field of computational science, using machine learning to showcase an application of a novel approach to understanding heterogeneity.

2. Methods

We relied on nationally representative data from the National Drug Abuse Treatment System Survey (NDATSS), a dataset containing eight waves of survey data from outpatient substance use treatment programs (OTPs) from 1988 to 2017.[30,31] Each wave incorporated a large percentage of programs from the previous wave, except programs excluded due to closure. More details on the NDATSS dataset can be found elsewhere.[21] In this paper, we looked at the last four waves of the NDATSS (110 OTPs in 2000, 142 in 2005, 184 in 2014, and 190 in 2017).

2.1. Dependent variable

We used an established measure of retention, the percentage of clients in treatment for more than 3 months in a treatment program, as the dependent variable. This measure has been used in other studies.[4,21,32]

2.2. Independent variables

The key independent variable is workforce diversity, which we define as the percentage of staff self-identified as Black or African American. This measure has been used in other studies.[17,18,21,32] To apply the existing estimation method for HTE, we dichotomized the treatment variable. Thus, we consider programs with more than 20% Black staff as having high workforce diversity. This threshold was chosen because more than 50% of the programs in our sample had less than 20% Black staff. The other relevant independent variables that define the heterogeneity of the treatment effect on client retention rates include program and client characteristics such as percentage of Black clients, percentage of Latino clients, accreditation by The Joint Commission (TJC), ownership status, program type (private-for-profit, private-not-for-profit, public), staff-to-client ratio, proportion of staff who have graduate degrees, percentage of unemployed clients, and whether the program is located in a state that expanded Medicaid coverage.

2.3. Statistical analysis

We conducted a comprehensive comparative analysis of all variables across the four-year period to assess any significant differences or associations. Categorical variables were examined using the Chi-square tests to determine if there were statistically significant associations between variables over time. For continuous variables, we utilized analysis of variance (ANOVA) to compare mean differences across the four years. This approach allowed us to identify patterns, trends, and variations in the data, providing a detailed understanding of how each variable evolved over the study period. To examine the heterogeneity of the association between workforce diversity and retention in OUD treatment, we used the causal forest method in which weights were incorporated to make the data nationally representative.[28,29]

The dataset used in our study was organized at the program level, meaning that each record corresponds to a single program. Therefore, when we refer to percentages of specific client demographics, we are indicating the proportion of those clients relative to the total number of clients within each program.

Causal forest is particularly well-suited for this analysis as it estimates the client and program-specific treatment effects of workforce diversity on retention. By doing so, it highlights how the presence of a diverse workforce might influence retention rates in different programs. In addition, the causal forest method generates variance estimates, which allow us to assess whether the observed treatment effects are statistically significant and different from zero. This approach not only quantifies the impact of workforce diversity but also provides a measure of the confidence we can have in these effects, revealing the conditions under which workforce diversity plays a crucial role in enhancing OUD treatment retention.

3. Results

We found significant differences among variables across the four different years that we examined. Table 1 presents the comparative analysis by year. The percentages of clients in treatment for more than 3 months were significantly different across years ($P < 0.001$). The percentages of Black clients were also significantly different across years ($P < 0.001$), with the percentages of Black clients being lower in the last two waves (2014 and 2017). More programs were from states that expanded Medicaid coverage in 2017 compared with 2014 ($P < 0.001$). There was an increasing trend of program age across years ($P < 0.001$). The results also showed that fewer programs were owned by another organization in the last two waves ($P < 0.001$). The staff-to-client ratio was significantly different across years ($p = 0.024$). The results also showed that the percentages of unemployed clients were higher in the last two waves ($P < 0.001$).

Results from the causal forest method (Table 2) showed that 61 OTPs had statistically significant positive treatment effects for workforce diversity. This means that these 61 OTPs would significantly benefit from having

Table 1. Comparative analysis of Opioid Treatment Programs in NDATSS data

	2000 *n*=1 10	2005 *n*=142	2014 *n*=184	2017 *n*=190
Percentage of clients in treatment more than 3 months***	84.3 (16.5)	87.9 (13.2)	75.2 (30.5)	79.0 (27.7)
More than 20% of Black staff	61 (55.5%)	70 (49.3%)	78 (42.4%)	93 (48.9%)
Client characteristics				
Percentage of Black clients***	29.4 (28.1)	25.5 (26.4)	18.7 (24.1)	20.8 (23.0)
Program characteristics				
Medicaid expansion***	-	-	116 (63%)	140 (73.7%)
TJC accreditation	27 (24.5%)	52 (36.6%)	60 (32.6%)	55 (28.9%)
Program age***	17.8 (11.6)	20.8 (12.0)	23.7 (14.1)	27.0 (15.1)
Owned by another organization***	78 (70.9%)	103 (72.5%)	43 (23.4%)	59 (31.1%)
Type of programs				
Private for-profit	40 (36.4%)	53 (37.3%)	63 (34.2%)	64 (33.7%)
Private not-for-profit	45 (40.9%)	64 (45.1%)	101 (54.9%)	99 (52.1%)
Public	25 (22.7%)	25 (17.6%)	20 (10.9%)	27 (14.2%)
Staff-to-client ratio in percentage*	4.3 (4.1)	3.9 (2.7)	4.2 (4.4)	6.1 (10.6)
Proportion of graduate staff	0.3 (0.2)	0.4 (0.2)	0.3 (0.2)	0.3 (0.2)
Percentage of unemployed clients***	44.6 (23.0)	43.6 (24.7)	54.7 (26.4)	52.1 (25.8)

Notes: The values inside the parentheses in the table that do not have a "%" sign are the corresponding standard deviations. *P<0.05, **P<0.01, ***P<0.001.
Abbreviations: NDATSS: National Drug Abuse Treatment System Survey; TJC: The Joint Commission.

Table 2. Comparative analysis of programs with no or significant benefit from workforce diversity

	No benefit from diversity (*n*=565)	Benefit from diversity (*n*=61)	p-value
Medicaid expansion	233 (41.2%)	23 (37.7%)	0.69188
Year			0.34624
2000	104 (18.4%)	6 (9.8%)	
2005	129 (22.8%)	13 (21.3%)	
2014	163 (28.8%)	21 (34.4%)	
2017	169 (29.9%)	21 (34.4%)	
Percentage of Black clients	24.9 (25.7)	2.6 (2.9)	2.12E-66
TJC accreditation	182 (32.2%)	12 (19.7%)	0.06199
Owned by another organization	261 (46.2%)	22 (36.1%)	0.16922
Type of programs			2.22E-10
Private for-profit	175 (31%)	45 (73.8%)	
Private not-for-profit	298 (52.7%)	11 (18%)	
Public	92 (16.3%)	5 (8.2%)	
Staff-to-client ratio in percentage	5.0 (7.0)	2.0 (1.1)	2.38E-19
Proportion of graduate staff	0.3 (0.2)	0.5 (0.2)	2.43E-05
Percentage of unemployed clients	52.0 (25.2)	27.1 (17.8)	4.31E-16
Program age	23.9 (13.9)	14.5 (10.8)	1.51E-08

Note: The values inside the parentheses in the table that do not have a "%" sign are the corresponding standard deviations.
Abbreviation: TJC: The Joint Commission.

a high percentage of Black staff in terms of increasing the percentage of clients who stay in treatment longer than 3 months (retention). Among the remaining 566 OTPs, 562 did not have statistically significant treatment effects, while

four had statistically significantly negative treatment effects. It is important to note that these four programs with negative treatment effects may potentially represent false discoveries.

The comparison of characteristics of these 61 OTPs with the other 566 OTPs is presented in Table 2. The 61 OTPs that would benefit the most from workforce diversity had significantly lower percentages of Black clients ($P < 0.001$,), were more likely to be private-for-profit ($P < 0.001$), had lower staff-to-client ratio ($P < 0.001$), much higher proportion of staff who had graduate degrees ($P < 0.001$), much lower percentage of unemployed clients ($P < 0.001$), and were more likely to be newer programs ($P < 0.001$). The box plots of the percentage of clients in treatment for more than 3 months, categorized based on the presence of high and low percentage of Black staff, in these 61 OTPs and the other 566 OTPs are presented in Figure 1. Higher percentages of Black staff were associated with the increased percentage of clients in treatment to more than 3 months in these 61 OTPs.

4. Discussion

To study the role of the variation of workforce diversity in improving OUD treatment retention, we explored the heterogeneous treatment effect with a novel machine-learning method called causal forest. Our analytical method aimed to advance understanding of the variation in the association between workforce diversity, that is, percentage of Black staff in an OUD treatment program, and OUD treatment retention (percentage of clients in treatment for more than 3 months).

We found that only a small proportion of the sample, that is, 61 out of 627 OTPs (<10%), would statistically significantly benefit from workforce diversity when it comes to retaining clients. This means that, notwithstanding the level of workforce diversity, characteristics of these programs, other than workforce diversity, would cause them to benefit more from diversity. The characteristics that amplified the impact of workforce diversity on retention included: lower percentages of Black clients, lower staff-to-client ratio, higher proportion of staff who had graduate degrees, and lower percentage of unemployed clients. In addition, those OTPs were more likely to be private for-profit and newer.

The characteristics of these 61 OTPs suggest that the impact of workforce diversity is context-dependent, and its effectiveness is shaped by a combination of program-specific factors. For instance, programs with lower percentages of Black clients might benefit more from diversity because the presence of Black staff could provide critical cultural insights and connections that are otherwise lacking. Similarly, a lower staff-to-client ratio and higher staff education levels could facilitate more personalized and culturally competent care, which is particularly beneficial in diverse workforce environments. These findings imply that workforce diversity alone is insufficient for improving retention across all settings; rather, it is the synergy between diversity and other favorable program characteristics that drives positive outcomes.

Moreover, the observation that private for-profit programs were more likely to benefit from diversity could

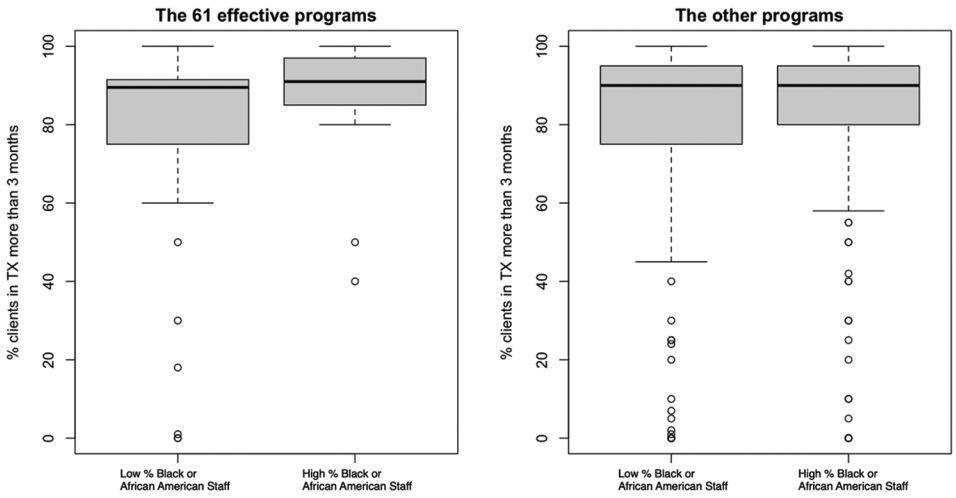


Figure 1. Box plots illustrating the heterogeneous effects of workforce diversity on treatment retention. (Left) Box plots of the percentage of clients in treatment for more than 3 months, categorized based on the presence of high and low percentage of Black staff, in the 61 effective programs. (Right) Box plots of the percentage of clients in treatment for more than three months, categorized based on the presence of high and low percentages of Black staff, in the other 565 ineffective programs.
Abbreviations TX: Texas.

reflect the greater operational flexibility and resource availability in these settings, allowing for more effective implementation of diversity initiatives. This raises important considerations for public and not-for-profit programs, which may need additional support to create environments where workforce diversity can flourish. These programs might require targeted investments in staff training, educational opportunities, and resource allocation to replicate the conditions under which diversity positively impacts retention in private for-profit settings.

The characteristics of these 61 OTPs indicate that workforce diversity is most likely to improve client retention when implemented in less constrained programs, that is, those with attributes often reported in the existing literature to be associated with positive outcomes.[4,32] This may explain why we did not see a significant association between the percentage of Black staff and the percentage of clients in treatment for more than 3 months when considering the full sample of 627 OTPs.

Few studies have examined the general association between Black workforce diversity and treatment retention among Black clients.[16,21] These studies identified significant associations that may have been driven by a small subgroup or population. In ddition, organizational characteristics may alter the impact of workforce diversity in a different direction. Findings in this paper inform rigorous analytical approaches to understand relationships of individual and program features with client outcomes. The benefit of this approach is to help public health policymakers identify OTPs that might benefit from workforce diversity, or alternatively, OTPs with high workforce diversity that could benefit from greater resources.

Our study is also aligned with the national call to diversify the workforce in addiction health services and thus informs how and when diversity could most benefit client-centered outcomes. However, it is also important to recognize that the relationship between workforce diversity and client retention is complex and influenced by multiple factors beyond the demographic composition of the staff. The findings from this study highlight the need for a more nuanced understanding of how diversity interacts with program-specific characteristics to influence outcomes. Future research should continue to explore these interactions, with a particular focus on identifying the conditions under which workforce diversity is most likely to enhance retention, and how these conditions can be fostered across different types of treatment programs.

Policymakers should recognize that while workforce diversity is important, it is not a standalone solution for improving client retention in OUD treatment programs. Policies solely focused on increasing diversity may not yield desired outcomes unless other factors are addressed. This study highlights program characteristics associated with a positive impact of workforce diversity on retention. As such, policymakers may want to allocate resources related to program characteristics that enhance the benefits of a diverse workforce. It is also crucial to strike a balance between resource allocation and diversity goals, as less constrained programs, often linked to positive outcomes, maximize the benefits of diversity. Policies should support adequate resource allocation, including staffing and educational opportunities, while fostering diversity. In addition, targeted strategies should prioritize retention rates among Black clients, addressing their unique challenges through tailored interventions, culturally competent care, and efforts to reduce disparities in access and quality of treatment. Moreover, as the landscape of opioid treatment continues to evolve, it will be essential for policymakers to remain flexible and responsive to new evidence on the factors that influence retention. This includes being open to revising policies and practices as more is learned about the role of workforce diversity in different contexts and the needs of client populations change over time. Overall, policies should consider program characteristics, resource allocation, and diversity goals to improve retention rates, particularly among Black clients, in OUD treatment programs.

Several limitations of this study should be acknowledged. Most existing methods can only estimate the heterogeneous treatment effects for binary variables. Thus, we had to dichotomize the percentage of Black staff to obtain a binary treatment variable. We chose the cutoff of 20% because 48.2% (i.e., about one half) of programs had more than 20% Black staff. Ideally, we would explore the heterogeneity with the original continuous variable, that is, percentage of Black staff. The identified 61 OTPs would have a greater impact on retention given their diverse workforce. However, there may be other heterogeneity among these 61 OTPs in the association between workforce diversity and retention. We did not examine such heterogeneity in this paper because the higher treatment effects on these 61 OTPs were composite effects of several variables. In fact, we cannot observe significant treatment effects by altering the value of just one variable, while keeping the others constant. Moreover, our finding that a lower percentage of Black clients being associated with lower constraints, and therefore greater retention, should be further examined. Future studies should scrutinize this finding to better understand the mechanisms that drive this association and suggest concrete approaches to improve outcomes equally and equitably.

5. Conclusion

Our findings contribute to a deeper understanding of how workforce diversity, particularly in the form of higher

percentages of Black staff, can positively impact retention rates among Black clients in opioid treatment programs (OTPs). This underscores the importance of employing advanced statistical methods to identify and quantify when and how diversity enhances treatment outcomes, especially for minority clients. By doing so, we can better address the unique needs of these clients and optimize program resources to serve minority communities effectively.

As federal and state authorities prepare to allocate substantial financial resources from various sources – including pharmaceutical settlements and new tax revenues – to improve access to opioid treatment,[33,34] it is imperative to understand how these investments can best support OTPs in enhancing overall patient outcomes. This is particularly vital for improving outcomes among minority populations, who often face additional barriers to accessing quality care. A diverse workforce not only reflects the communities these programs serve but also has the potential to foster a more inclusive and supportive treatment environment that can significantly impact retention and recovery. Therefore, fully leveraging the benefits of workforce diversity will be crucial in shaping policies and interventions that maximize the effectiveness of opioid treatment programs, especially in underserved and minority communities.

Acknowledgments

The authors would like to thank Joanna Mendoza, from Texas A&M University, for editing and proofreading this manuscript.

Funding

Support for this research and manuscript preparation was provided by a National Institute on Minority Health and Health Disparities research grant (R01MD014639, CoPIs: Daniel Howard and Erick Guerrero) and a National Institute on Drug Abuse research grant (DA048176, CoPIs: Jeanne Marsh and Erick Guerrero).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Yinfei Kong, Erick Guerrero
Formal analysis: Yinfei Kong, Erick Guerrero, Suojin Wang
Investigation: Yinfei Kong, Erick Guerrero, Suojin Wang
Methodology: Yinfei Kong, Suojin Wang
Writing – original draft: Yinfei Kong, Erick Guerrero
Writing – review & editing: Jemima A. Frimpong, Tenie Khachikian, Suojin Wang, Thomas D'Aunno, Daniel Howard

Ethics approval and consent to participate

This study was reviewed and approved by the Institutional Review Board of Texas A&M University (TAMU IRB#2019-0268DCR).

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. CDC/National Center for Health Statistics. U.S. Overdose Deaths Decrease in 2023, First Time Since 2018; 2024. Available from: https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2024/20240515.htm [Last accessed on 2024 May 24].
2. Bart G. Maintenance medication for opiate addiction: The foundation of recovery. *J Addict Dis*. 2012;31(3):207-225.
   doi: 10.1080/10550887.2012.694598
3. Chan B, Gean E, Arkhipova-Jenkins I, *et al*. Retention strategies for medications for opioid use disorder in adults. *J Addict Med*. 2020;15(1):74-84.
   doi: 10.1097/adm.0000000000000739
4. Timko C, Schultz NR, Cucciare MA, Vittorio L, Garrison-Diehn C. Retention in medication-assisted treatment for opiate dependence: A systematic review. *J Addict Dis*. 2015;35(1):22-35.
   doi: 10.1080/10550887.2016.1100960
5. Williams AR, Samples H, Crystal S, Olfson M. Acute care, prescription opioid use, and overdose following discontinuation of long-term buprenorphine treatment for opioid use disorder. *Am J Psychiatry*. 2020;177(2):117-124.
   doi: 10.1176/appi.ajp.2019.19060612
6. Carroll KM, Weiss RD. The role of behavioral interventions in buprenorphine maintenance treatment: A review. *Am J Psychiatry*. 2017;174(8):738-747.
   doi: 10.1176/appi.ajp.2016.16070792
7. Manhapra A, Petrakis I, Rosenheck R. Three-year retention in buprenorphine treatment for opioid use disorder nationally in the Veterans health administration. *Am J Addict*. 2017;26(6):572-580.
   doi: 10.1111/ajad.12553
8. Proctor SL, Copeland AL, Kopak AM, Hoffmann NG, Herschman PL, Polukhina N. Predictors of patient retention in methadone maintenance treatment. *Psychol Addict Behav*. 2015;29(4):906-917.
   doi: 10.1037/adb0000090

9. Weinstein ZM, Kim HW, Cheng DM, *et al*. Long-term retention in office based opioid treatment with buprenorphine. *J Subst Abuse Treat*. 2017;74:65-70.

   doi: 10.1016/j.jsat.2016.12.010

10. Acevedo A, Garnick D, Ritter G, Horgan C, Lundgren L. Race/ethnicity and quality indicators for outpatient treatment for substance use disorders. *Am J Addict*. 2015;24(6):523-531.

    doi: 10.1111/ajad.12256

11. Mennis J, Stahler GJ, El Magd SA, Baron DA. How long does it take to complete outpatient substance use disorder treatment? Disparities among blacks, hispanics, and whites in the US. *Addict Behav*. 2019;93:158-165.

    doi: 10.1016/j.addbeh.2019.01.041

12. Guerrero EG. Managerial capacity and adoption of culturally competent practices in outpatient substance abuse treatment organizations. *J Subst Abuse Treat*. 2010;39(4):329-339.

    doi: 10.1016/j.jsat.2010.07.004

13. Guerrero EG. Enhancing access and retention in substance abuse treatment: The role of Medicaid payment acceptance and cultural competence. *Drug Alcohol Depend*. 2013;132(3):555-561.

    doi: 10.1016/j.drugalcdep.2013.04.005

14. Guerrero EG, Campos M, Urada D, Yang JC. Do cultural and linguistic competence matter in Latinos' completion of mandated substance abuse treatment? *Subst Abuse Treat Prev Policy*. 2012;7:827-836.

    doi: 10.1186/1747-597x-7-34

15. Guerrero EG, Khachikian T, Kim T, Kong Y, Vega WA. Spanish language proficiency among providers and Latino clients' engagement in substance abuse treatment. *Addict Behav*. 2013;38(12):2893-2897.

    doi: 10.1016/j.addbeh.2013.08.022

16. Guerrero E, Andrews CM. Cultural competence in outpatient substance abuse treatment: Measurement and relationship to wait time and retention. *Drug Alcohol Depend*. 2011;119(1-2):e13-e22.

    doi: 10.1016/j.drugalcdep.2011.05.020

17. Howard DL. Are the treatment goals of culturally competent outpatient substance abuse treatment units congruent with their client profile? *J Subst Abuse Treat*. 2003;24(2):103-113.

    doi: 10.1016/s0740-5472(02)00349-5

18. Howard DL. Culturally competent treatment of African American clients among a national sample of outpatient substance abuse treatment units. *J Subst Abuse Treat*. 2003;24(2):89-102.

    doi: 10.1016/s0740-5472(02)00348-3

19. Jordan A, Jegede O. Building outreach and diversity in the field of addictions. *Am J Addict*. 2020;29(5):413-417.

    doi: 10.1111/ajad.13097

20. Weller BE, Harrison J, Adkison-Johnson C. Training a diverse workforce to address the opioid crisis. *Soc Work Mental Health*. 2021;19(6):568-582.

    doi: 10.1080/15332985.2021.1975014

21. Guerrero EG, Kong Y, Khachikian T, Wang S, D'Aunno T, Howard D. *Workforce Diversity and Disparities in Opioid Treatment Wait Time and Retention*. Research Square. Preprint; 2022.

    doi: 10.21203/rs.3.rs-1651284/v1

22. Alaa A, Schaar M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. *Proc Mach Learn Res*. 2018;80:129-38.

23. Angus DC, Chang CC. Heterogeneity of treatment effect: Estimating how the effects of interventions vary across individuals. *JAMA*. 2021;326(22):2312-2313.

    doi: 10.1001/jfama.2021.20552

24. Kong Y, Zhou J, Zheng Z, Amaro H, Guerrero EG. Using machine learning to advance disparities research: Subgroup analyses of access to opioid treatment. *Health Serv Res*. 2021;57(2):411-421.

    doi: 10.1111/1475-6773.13896

25. Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc Natl Acad Sci U S A*. 2019;116(10):4156-4165.

    doi: 10.1073/pnas.1804597116

26. Hill JL. Bayesian nonparametric modeling for causal inference. *J Comput Graph Stat*. 2011;20(1):217-240.

27. Grimmer J, Messing S, Westwood SJ. Estimating heterogeneous treatment effects and the effects of heterogeneous treatments with ensemble methods. *Polit Anal*. 25(4):413-434.

    doi: 10.1017/pan.2017.15

28. Athey S, Tibshirani J, Wager S. Generalized random forests. *Ann Stat*. 2019;47(2):1148-1178.

    doi: 10.1214/18-AOS1709

29. Wager S, Athey S. Estimation and inference of heterogeneous treatment effects using random forests. *J Am Stat Assoc*. 2018;113(523):1228-1242.

    doi: 10.1080/01621459.2017.1319839

30. D'Aunno T, Park SE, Pollack HA. Evidence-based treatment for opioid use disorders: A national study of methadone dose levels, 2011-2017. *J Subst Abuse Treat*. 2019;96:18-22.

    doi: 10.1016/j.jsat.2018.10.006

31. D'Aunno T, Pollack HA, Frimpong JA, Wuchiett D. Evidence-based treatment for opioid disorders: A 23-year national study of methadone dose levels. *J Subst Abuse Treat*.

2014;47(4):245-250.

doi: 10.1016/j.jsat.2014.06.001

32. Liu J, Storfer-Isser A, Mark TL, *et al*. Access to and engagement in substance use disorder treatment over time. *Psychiatr Serv*. 2020;71(7):722-725.

doi: 10.1176/appi.ps.201800461

33. Campbell B. *A Proposed Legislative Fund Could Help to Close Racial, Health Gap*; 2021. Available from: https://phadvocates.org/wp-content/uploads/2021/06/bkgrnd-for-fund_060321.pdf [Last accessed on 2024 May 24].

34. Haffajee RL. The public health value of opioid litigation. *J Law Med Ethics*. 2020;48(2):279-292.

doi: 10.1177/1073110520935340

ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

BRIEF REPORT

Does improving diagnostic accuracy increase artificial intelligence adoption? A public acceptance survey using randomized scenarios of diagnostic methods

Yulin Hswen[1,2*], Ismaël Rafaï[2], Antoine Lacombe[2], Bérengère Davin-Casalena[3], Dimitri Dubois[4], Thierry Blayac[4], and Bruno Ventelou[2]

[1]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California, United States of America

[2]Aix-Marseille Univiversity, CNRS, AMSE, Marseille, France

[3]Observatoire Régional de la Santé, Provence-Alpes-Côte d'Azur, France

[4]CEE-M, Univ. Montpellier, CNRS, INRAe, Institut Agro, Montpellier, France

(This article belongs to the *Special Issue: Artificial intelligence for diagnosing brain diseases*)

Abstract

This study examines the acceptance of artificial intelligence (AI)-based diagnostic alternatives compared to traditional biological testing through a randomized scenario experiment in the domain of neurodegenerative diseases (NDs). A total of 3225 pairwise choices of ND risk-prediction tools were offered to participants, with 1482 choices comparing AI with the biological saliva test and 1743 comparing AI+ with the saliva test (with AI+ using digital consumer data, in addition to electronic medical data). Overall, only 36.68% of responses showed preferences for AI/AI+ alternatives. Stratified by AI sensitivity levels, acceptance rates for AI/AI+ were 35.04% at 60% sensitivity and 31.63% at 70% sensitivity, and increased markedly to 48.68% at 95% sensitivity (p <0.01). Similarly, acceptance rates by specificity were 29.68%, 28.18%, and 44.24% at 60%, 70%, and 95% specificity, respectively (P < 0.01). Notably, AI consistently garnered higher acceptance rates (45.82%) than AI+ (28.92%) at comparable sensitivity and specificity levels, except at 60% sensitivity, where no significant difference was observed. These results highlight the nuanced preferences for AI diagnostics, with higher sensitivity and specificity significantly driving acceptance of AI diagnostics.

Keywords: Artificial intelligence; AI diagnostics; Neurodegenerative diseases; Machine learning

***Corresponding author:**
Yulin Hswen
(yulin.hswen@ucsf.edu)

Citation: Hswen Y, Rafaï I, Lacombe A, *et al.* Does improving diagnostic accuracy increase artificial intelligence adoption? A public acceptance survey using randomized scenarios of diagnostic methods. *Artif Intell Health.* 2025;2(1):114-120.
doi: 10.36922/aih.3561

Received: May 2, 2024

1st revised: August 2, 2024

2nd revised: September 17, 2024

3rd revised: September 27, 2024

Accepted: September 27, 2024

Published Online: October 18, 2024

Copyright: © 2024 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Background

The integration of artificial intelligence (AI) into health care brings the promise of revolutionizing diagnostic and prognostic capabilities, offering more precise, data-driven insights that can enhance patient outcomes.[1,2] By harnessing AI's ability to analyze large datasets, including electronic health records (EHRs) and digital data concerning consumer behaviors, health-care systems can potentially improve the early detection

and prediction of disease risks with greater accuracy and timeliness. This combination of clinical and behavioral data could enable more personalized diagnostic models.

However, despite the substantial potential AI offers in transforming health care, there remains significant hesitation toward its widespread adoption, particularly in AI-assisted diagnostics. Much of this reluctance stems from concerns about patient privacy and the risks associated with data surveillance.[3] Health-care professionals and patients alike worry that the use of AI in clinical settings could lead to breaches of sensitive information, unauthorized data access, and misuse of personal health data, which are the main factors undermining users' trust in AI-driven systems. This aversion to AI, fueled by privacy concerns, continues to be a major obstacle to its full acceptance in the health-care field.

By examining the resistance to AI, previous research found that generative chatbot AI faces a hesitant adoption.[4] A review of 7912 articles aimed at identifying predictors of AI adoption revealed that perceived usefulness, performance expectancy, trust, and effort were key factors influencing the willingness to use AI in health care.[4] The review also emphasized that no amount of AI could fully replace the value of human interaction or ensure cultural sensitivity. In another study related to AI use in health care,[5] this reluctance was shown to be more pronounced among individuals with limited proficiency in Internet or computer technologies. A noted source of concern stems from the uncertainty surrounding the data sources that power these AI models, leading to skepticism about the reliability and accuracy of the health information they generate. In addition, users express unease over the lack of transparency in how these models operate and the inherent complexity of AI systems. These factors contribute to fears of miscommunication, misinterpretation of health symptoms, and the potential for inaccurate diagnoses. In another related survey, trust in AI adoption was found to be closely linked to regulatory oversight, with performance and communication also playing critical roles in users' willingness to embrace AI applications in health care.[6]

A survey conducted in Sweden showed that only 20% of health-care professionals used AI-based systems in their work, with "trust" emerging as the most critical factor in their willingness to adopt these technologies.[7] A review of 42 studies examining health-care professionals' acceptance of AI revealed widespread concerns, particularly regarding AI's potential for errors, sensitivity, and timely access. In addition, the perceived loss of professional autonomy and challenges in integrating AI into existing clinical workflows were consistently identified as a significant barrier to adoption.[8] These findings highlight that healthcare workers, such as patients, exhibit a degree of AI adoption hesitancy, particularly in its utilization in diagnostics.

In contrast to previous studies, which primarily rely on surveys, this study aims to broaden the existing literature on AI adoption hesitancy by testing AI adoption through randomized scenario-based experiments. This approach allows for a more nuanced understanding of how individuals respond to AI in varied controlled contexts.

2. Methods

This study evaluated the public acceptability of AI-based diagnostic tools and the accuracy trade-offs required to integrate EHRs and digital data in the domain of neurodegenerative diseases (NDs). A survey was conducted on a representative sample of the French adult population (n = 1017) using a quota non-probability sampling method (quotas were on age, gender, socio-professional status, and living area). This collection of data was part of the larger Discrete Choice Experiment[9-11] aiming at unveiling the trade-offs surrounding the decision-making by individuals about neurodegenerative testing. Before agreeing with study participation, all subjects were given comprehensive information regarding the study's purpose, procedures, potential risks, and benefits. The study protocol was reviewed and approved by the Ethics Committee of Aix-Marseille University (approval number: 2022-10-20-009). Written consent was obtained from each of the subjects to participate in this study.

The 1017 participants were exposed to a set of alternative scenarios of testing methods to predict the hypothetical 10-year risk of developing an ND that affects an average of 7% of the population after the age of 65.[12] Through the pool institut ViaVoice, participants were confidentially randomized to scenarios depicting various levels of AI-based diagnostic integration and non-AI traditional laboratory saliva test. The researchers were blinded to participants' identities. The three scenarios of tests included: (1) non-AI diagnostics using a laboratory test with a salivary sample, (2) AI diagnostics incorporating EHRs, defined as "AI," and (3) AI diagnostics incorporating EHR and digital consumer data from mobile devices, thereafter, defined as "AI+." To assess the impact of diagnostic accuracy on participants' preferences, the attributes of sensitivity (true positive rate) at 60%, 70%, or 95%, and 1-specificity (false positive rate) at 5%, 30%, or 40% were also varied. An example of the randomized scenario is shown in Figure 1.

3. Statistical analysis

Of the 5085 scenarios randomly proposed, we selected the pairs (3225) that display a comparison between AI (or AI+)

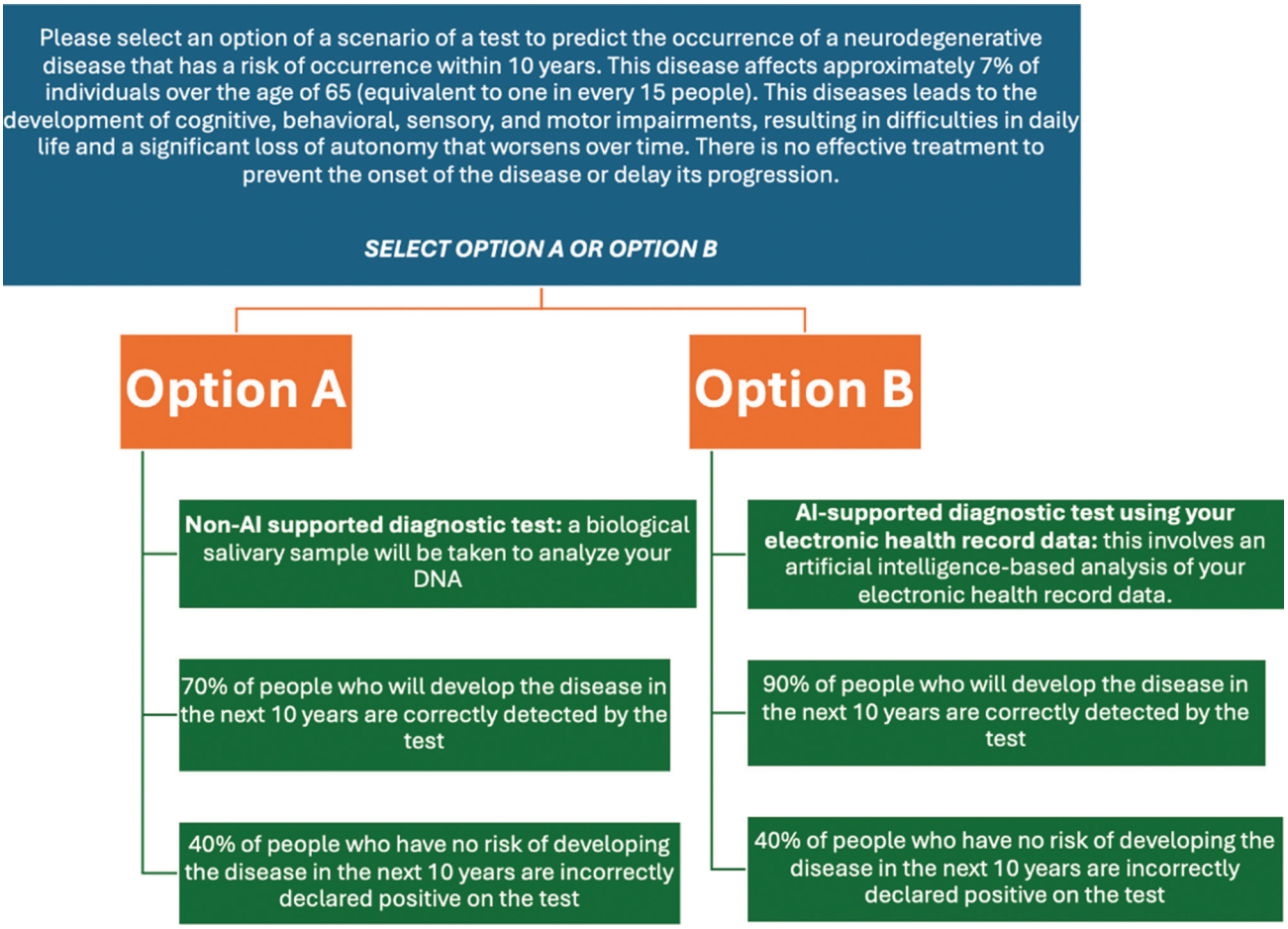


Figure 1. Example of a randomized scenario

and saliva test as the benchmark for further analysis. Then, descriptive analyses were conducted to compare the proportions of acceptance of the AI and/or AI+ option with those of the saliva test option. The differences in acceptance rates between the AI-versus-saliva-test option and the AI+-versus-saliva-test option were analyzed using pairwise z-tests, whereas the differences in proportions between three levels of sensitivity and specificity per type of pairs offered (*i.e.*, AI vs. saliva and AI+ vs. saliva) were compared using Chi-squared tests.[13,14]

4. Results

Contingency tables describing proportions of agreement between the AI and/or AI+ alternative with the saliva test alternative are presented in Table 1. From the 3225 pairs of AI/AI+ against saliva tests that were offered, 1482 are linked to a choice of AI test versus saliva test, whereas 1743 are associated with a choice of AI+ test versus saliva test. Only 36.68% of the answers were pro-AI/AI+ (45.82% for the AI-vs.-saliva-test pairs and 28.92% for the AI+-vs.-saliva-test pairs; proportions were significantly different at a 1% threshold). Figure 2 shows the proportion of AI adoption (vs. saliva test) across different levels of accuracy. Upon stratifying the answers by sensitivity level, we found that 35.04% of the answers were in favor of AI or AI+ when AI's sensitivity was 60% and 31.63% when AI's sensitivity was 70%, and this proportion increased to 48.68% when AI's sensitivity was 95% (Chi-squared test showed significant difference at a 1% threshold). With respect to specificity, 29.68%, 28.18%, and 44.24% of the answers favored AI/AI+ test over saliva test when specificity levels were 60%, 70%, and 95%, respectively (Chi-squared test showed significant differences at a 1% threshold). Finally, when we compared AI-versus-saliva-test option and AI+-versus-saliva-test option per sensitivity or specificity levels, we found significantly higher acceptance rates in the AI group than in the AI+ group from the same sensitivity or specificity level (except when sensitivity is 60%, where we found no significant differences between rates of acceptance when AI-vs.-saliva-test or AI+-vs.-saliva-test options were offered).

Table 1. The proportion of individuals choosing the AI and/or AI+ alternative over the saliva test per sensitivity and specificity levels

	AI versus saliva test	AI+ versus saliva test	AI or AI+ versus saliva test
n	1482	1743	3225
Proportion of yes to AI/AI+	$45.82\%_{+++}$	$28.92\%_{+++}$	36.68%
Sensitivity			
Proportion of yes to AI/AI+ when sensitivity=60%	36.95%***	32.77%**	35.04%***
Proportion of yes to AI/AI+ when sensitivity=70%	$44.33\%^{***}_{+++}$	$26.43\%^{**}_{+++}$	31.63%***
Proportion of yes to AI/AI+ when sensitivity=95%	$59.39\%^{(***)}_{(+++)}$	$30.04\%^{(**)}_{(+++)}$	48.68%***
Specificity			
Proportion of yes to AI/AI+ when specificity=60%	$41.24\%_{+++}$	$18.2\%^{(***)}_{(+++)}$	29.68%***
Proportion of yes to AI/AI+ when specificity=70%	$47.67\%_{+++}$	$12.73\%^{(***)}_{(+++)}$	28.18%***
Proportion of yes to AI/AI+ when specificity=95%	$47.04\%_{+++}$	$41.94\%^{(***)}_{(++)}$	44.24%***

Notes: (1) Chi-square statistical test of difference of acceptance rate per type of scenario offered (AI vs. saliva test or AI+ vs. saliva test): *P<0.1, **P<0.05, and ***P<0.01; (2) Pairwise z-test of difference in proportions per sensitivity or specificity level (60%, 70%, or 95%) between AI-versus-saliva-test and AI+-versus-saliva-test: ^{+}P<0.1, ^{++}P<0.05, ^{+++}P<0.01.
Abbreviations: AI: Artificial intelligence using electronic health records (EHRs) data; AI+: Artificial intelligence using electronic health records (EHRs) data and digital consumer data.

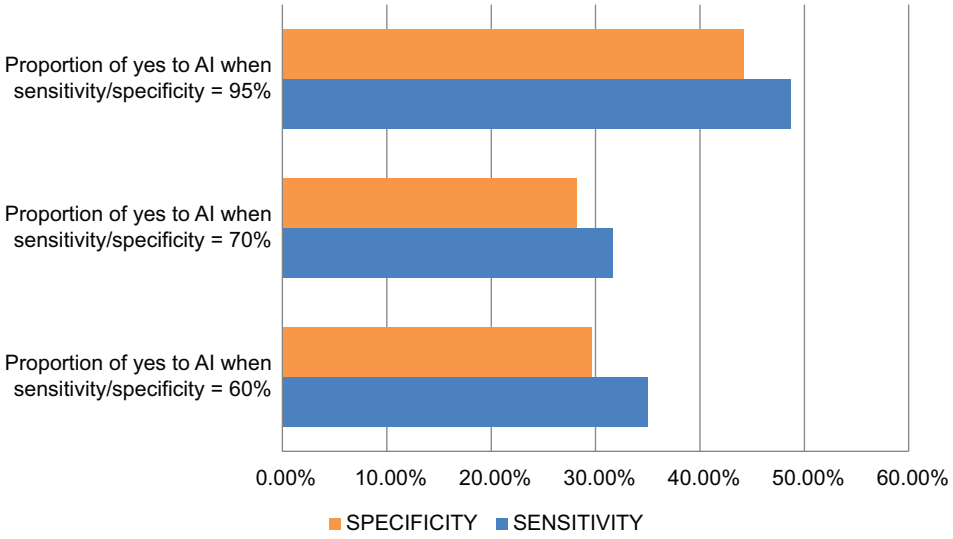


Figure 2. The proportion of artificial intelligence adoption (vs. saliva test) across different levels of accuracy

5. Discussion

The findings of this study provide valuable insights into public acceptance of AI-based diagnostic alternatives compared to conventional saliva tests. By analyzing 3225 pairwise choices, we observed that only 36.68% of participants preferred AI/AI+ alternatives over traditional saliva tests, specifically with a significant preference for AI (45.82%) compared to AI+ (28.92%).

The results strongly showed the influence of AI sensitivity (how well the AI test can identify true positives) on acceptance rates. As AI sensitivity increased from 60% to 95%, support for AI/AI+ diagnostics rose considerably, reaching 48.68% at 95% sensitivity. A similar pattern emerged with specificity (how well an AI test can identify true negatives), where acceptance rates increased from 29.68% at 60% specificity to 44.24% at 95% specificity.

These findings underscore the importance of enhancing diagnostic accuracy in fostering public trust in AI-based tools. Public acceptance of AI diagnostics is closely tied to accuracy levels, and these results suggest that AI tools must meet or exceed a 95% performance threshold to achieve meaningful levels of AI acceptance.

In addition, AI consistently outperformed AI+ across all levels of sensitivity, with the exception of 60% sensitivity, where no significant difference in preference between AI and AI+ was found. This outcome may indicate a hesitancy towards the integration of digital consumer data -this is what AI+ means compared to AI- versus EHR data alone. However, when accuracy approaches a sensitivity level of 95%, the public appears more willing to consider the use of these digital consumer data resources, reflecting a trust deficit that can be mitigated by increased diagnostic performance.

6. Study limitations

It is important to emphasize that, to minimize the biases of physical invasiveness[15] while striving to level the playing field in comparison to AI testing methods, we deliberately chose a salivary test for this study. As a result, our estimates of the public's preference for biological tests may be in fact lower if AI testing was compared to more physically invasive procedures such as brain imaging, cerebrospinal fluid analysis, or blood tests.[16,17] This decision likely shaped the participants' responses, as the less invasive nature of the salivary test may have led them to favor it over more physically invasive testing methods. As a result, the reluctance toward AI diagnostics observed in this study may be less significant when compared to scenarios involving more invasive testing procedures.

Public perceptions of AI adoption are also likely to differ significantly across geographic regions, influenced by varying cultural, economic, and social factors that shape attitudes toward technology. Although previous studies have shown similar AI hesitancies, this study was conducted in France and national differences could result in diverse levels of trust, familiarity, and comfort with AI, thereby affecting how AI technologies are embraced across different nations. Consequently, this variability poses a potential limitation to the generalizability of this study's findings. Factors such as regional regulatory environments, access to technology, socioeconomic disparities, and historical experiences with digital tools could further amplify these discrepancies in AI acceptance. Therefore, our findings must be considered within the diverse global contexts where AI technologies may be implemented. This underscores the importance of future research to examine AI adoption across a broader range of geographic and cultural settings, ensuring greater applicability and relevance.

7. Contributions of this study

Unlike earlier research that has largely focused on survey-based methods, this study expands the body of knowledge on AI adoption by conducting investigations on AI acceptance through randomized, scenario-driven experiments. Using this approach, we can capture a more detailed perspective on how people react to AI in diverse and controlled situations, addressing the broader challenge of AI hesitancy and the complexity of its acceptance in real-world settings. Our findings significantly enhance the current body of research by providing empirical evidence on the threshold of diagnostic accuracy required for AI-driven technologies to achieve widespread public acceptance. By quantifying these levels of accuracy, we offer a framework for understanding the public's expectations of AI in health-care settings. This research not only underscores the importance of reliability and accuracy in AI diagnostics but also highlights the nuanced factors influencing public trust and adoption. In addition, it sheds light on how varying degrees of accuracy can shape public perceptions, offering insights for developers, policymakers, and health-care professionals aiming to bridge the gap between technological advancements and public readiness for AI integration. These insights are particularly valuable in addressing AI hesitancy and ensuring the ethical implementation of AI in health care.

8. Conclusion

Our findings carry important implications for the development and implementation of AI diagnostics in health care. Public hesitation persists as a significant barrier, especially when AI tools are perceived as lacking sufficient accuracy or integrating excessive amounts of personal data. Our results emphasize the critical need for AI developers and health-care providers to prioritize transparency, accuracy, and usability in AI diagnostic technologies. Moreover, educating the public about the potential benefits of AI diagnostics, particularly diagnostic accuracy, could further alleviate concerns and promote broader acceptance.

This study highlights the nuanced preferences of the public for AI diagnostics, with higher sensitivity and specificity acting as key drivers of acceptance. While AI holds considerable potential to transform health-care diagnostics, addressing the public's concerns about accuracy and complexity will be essential to its successful adoption.

Acknowledgments

None.

Funding

The project leading to this publication has received funding from the French government under the "France 2030" investment plan managed by the French National Research Agency (reference: ANR-17-EURE-0020) and from Excellence Initiative of Aix-Marseille University – A*MIDEX. This research also received support from the French National Research Agency (GRANT ANR-20-COVR-00 and ANR-21-JPW2-002), as well as funding from the National Institute of Health T32 grant (5T32MD015070-05).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: All authors
Formal analysis: Bruno Ventelou, Yulin Hswen, Ismaël Rafaï, Antoine Lacombe
Investigation: Bruno Ventelou, Yulin Hswen, Ismaël Rafaï, Antoine Lacombe
Methodology: Thierry Blayac, Dimitri Dubois, Yulin Hswen, Ismael Rafai, Bruno Ventelou, Antoine Lacombe
Writing–original draft: Bruno Ventelou, Yulin Hswen
Writing–review & editing: Ismaël Rafaï, Antoine Lacombe

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Ethics Committee of Aix-Marseille University (approval number: 2022-10-20-009). Written consent was obtained from each of the subjects to participate in this study.

Consent for publication

Written consent was obtained from each of the subjects to publish their data and/or images.

Availability of data

Data used in the study can be obtained from the corresponding author upon reasonable request.

References

1. Reyna MA, Nsoesie EO, Clifford GD. Rethinking algorithm performance metrics for artificial intelligence in diagnostic medicine. *JAMA*. 2022;328(4):329-330.

   doi: 10.1001/jama.2022.10561

2. Aggarwal R, Sounderajah V, Martin G, *et al*. Diagnostic accuracy of deep learning in medical imaging: A systematic review and meta-analysis. *NPJ Digit Med*. 2021;4(1):65.

   doi: 10.1038/s41746-021-00438-z

3. Laux J, Wachter S, Mittelstadt B. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regul Gov*. 2024;18(1):3-32.

   doi: 10.1111/rego.12512

4. Choung H, David P, Ross A. Trust in AI and its role in the acceptance of AI technologies. *Int J Hum Comput Interact*. 2023;39(9):1727-1739.

   doi: 10.1080/10447318.2022.2050543

5. Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digit Health*. 2019;5:2055207619871808.

   doi: 10.1177/205520761987180

6. Esmaeilzadeh P. Use of AI-based tools for healthcare purposes: A survey study from consumers' perspectives. *BMC Med Inform Decis Mak*. 2020;20:1-19.

   doi: 10.1186/s12911-020-01191-1

7. Floruss J, Vahlpahl N. *Artificial Intelligence in Healthcare: Acceptance of AI-based Support Systems by Healthcare Professionals*. Jönköping University. Master Thesis; 2020. Available from: https://www.diva-portal.org/smash/get/diva2:1433298/fulltext01.pdf [Last accessed on 2024 Oct 11].

8. Lambert SI, Madi M, Sopka S, *et al*. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. *NPJ Digit Med*. 2023;6(1):111.

   doi: 10.1038/s41746-023-00874-z

9. De Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: A review of the literature. *Health Econ*. 2012;21(2):145-172.

   doi: 10.1002/hec.1697

10. Szinay D, Cameron R, Naughton F, Whitty JA, Brown J, Jones A. Understanding uptake of digital health products: Methodology tutorial for a discrete choice experiment using the bayesian efficient design. *J Med Internet Res*. 2021;23(10):e32365.

    doi: 10.2196/32365

11. Clark MD, Determann D, Petrou S, Moro D, De Bekker-Grob EW. Discrete choice experiments in health economics: A review of the literature. *Pharmacoeconomics*. 2014;32:883-902.

    doi: 10.1007/s40273-014-0170-x

12. Alawode DO, Heslegrave AJ, Ashton NJ, *et al*. Transitioning from cerebrospinal fluid to blood tests to facilitate diagnosis and disease monitoring in Alzheimer's disease. *J Int Med*. 2021;290(3):583-601.

    doi: 10.1111/joim.13332

13. Greenwood PE, Nikulin MS. *A Guide to Chi-squared Testing*. Vol. 280. United States: John Wiley and Sons; 1996.

14. Campbell I. Chi-Squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. *Stat Med*. 2007;26(19):3661-3675.

doi: 10.1002/sim.2832

15. Freeman D, Lambe S, Yu LM, *et al*. Injection fears and COVID-19 vaccine hesitancy. *Psychol Med*. 2023;53(4):1185-1195.

doi: 10.1017/S0033291721002609

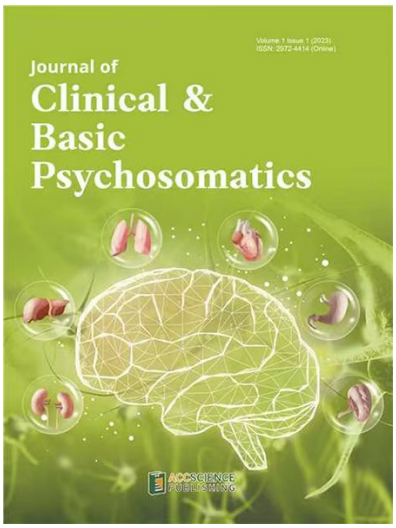
16. McLenon J, Rogers MA. The fear of needles: A systematic review and meta-analysis. *J Adv Nurs*. 2019;75(1):30-42.

doi: 10.1111/jan.13818

17. Von Wedel P, Hagist C. Physicians' preferences and willingness to pay for artificial intelligence-based assistance tools: A discrete choice experiment among german radiologists. *BMC Health Serv Res*. 2022;22(1):398.

doi: 10.1186/s12913-022-07769-x

Our Journals



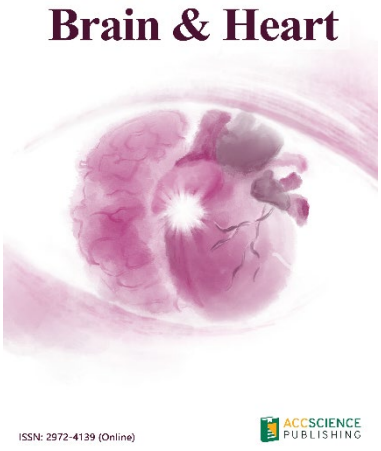
Journal of Clinical and Basic Psychosomatics (JCBP) is a quarterly journal focusing on clinical and basic research on symptoms, assessment, treatment, management, and the mechanism of psychosomatic disorders. *Journal of Clinical and Basic Psychosomatics* covers subject areas, including but not limited to the following:

- Conceptualization and classification of psychosomatic medicine
- Mechanism, biological markers, brain images, and treatment studies
- Psychosomatic reactions, syndromes, disorders, and diseases
- Psychosomatic disorders treated in general hospitals, including endocrinology, neurology, gastroenterology, dermatology, pain management, oncology, rheumatology, and other departments
- Psychological evaluation, management, rehabilitation, resilience training, and psychotherapy for general and specific populations during the pandemic
- Physiological disorders related to psychological factors (eating disorders, sleeping disorders, and sexual dysfunction)
- Somatic symptoms and related disorders and mental disorders due to somatic disease

Brain & Heart focuses on neurocardiology, a neurology and cardiology-based interdisciplinary subject that studies the circulatory mechanism of the human body, as well as the mechanisms of the interplay between the cardiovascular system and the nervous system. The journal's scope includes:

Clinical and basic research on diseases related to the circulatory and nervous systems, such as: orthostatic dizziness, orthostatic hypotension, autonomic dysfunction, and the relationship between the autonomic nervous system and the circulatory function in cerebral degeneration;

Heart-brain research on patients with syncope, autonomic dysfunction, cryptogenic stroke, and stroke with atrial fibrillation; research on the relationship between structural heart diseases and nervous system diseases, the correlation between cardiac electrophysiology and abnormal organizational structures and the pathogenesis of stroke, as well as new ways of diagnosis, treatment and prevention of unexplained stroke.



Start a new journal

Write to us via email if you are interested to start a new journal with AccScience Publishing. Please attach your CV, professional profile page and a brief pitch proposal in your email. We shall inform you of our decision whether we are interested to collaborate in starting a new journal.

Contact: info@accscience.com

https://www.accscience.com/journal/AIH

ACCSCIENCE PUBLISHING

Contact
www.accscience.com
8 Burn Road, #15-03 Trivex, Singapore 369977
Email: editorial@accscience.com
Phone: +65 8182 1586