

Artificial Intelligence in Health



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
9 Raffles Place, Republic Plaza 1 #06-00 Singapore 048619.

Volume 2 • Issue 4 • October 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*



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: Freda Wang

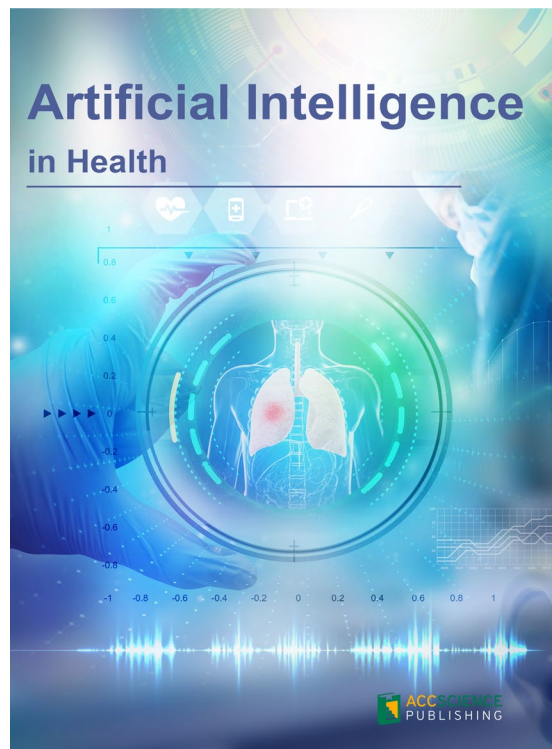
Production Editor: Sharmila Velapasamy

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/2/4>.



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 Editors

Weiping Ding, *China*

Xudong Liu, *China*

Ruiheng Zhang, *China*

Editorial Board

Members*

Adel Al-Jumaily, *Australia*

Zeeshan Ali, *China*

Ahmed Bouridane, *UAE*

Joaquim Carreras, *Japan*

Oscar Castillo, *Mexico*

Faouzi Alaya Cheikh, *Norway*

Xiaojun Chen, *China*

Xiaochun Cheng, *UK*

Krzysztof Jozef Cios, *USA*

Alfredo Cuzzocrea, *Italy*

Di Dong, *China*

Anastasios Dounis, *Greece*

Włodzisław Duch, *Poland*

Ayman El-Baz, *USA*

Adel Elmaghraby, *USA*

Manuel F.G. Penedo, *Spain*

Runwei Guan, *China*

Rémy Guillevin, *France*

Andrew A. Gumbs, *France*

Pankaj Gupta, *India*

A. Ben Hamza, *Canada*

Alexander Hramov, *Russia*

Bin Hu, *China*

Yisen Huang, *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*

Haipeng Liu, *UK*

Nicola Luigi Bragazzi, *Canada*

Xiaoke Ma, *China*

Xuele 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*

Witold Pedrycz, *Canada*

Alexander N. Pisarchik, *Spain*

Dawid Polap, *Poland*

Mihail Popescu, *USA*

Mukesh Prasad, *Australia*

Minghui Qian, *China*

Marek Reformat, *Poland*

Hongliang Ren, *China*

Hassan Rivaz, *Canada*

José Santamaría López, *Spain*

Paulo Adriano Schwingel, *Brazil*

Wei Shao, *China*

Chao Shen, *China*

Patricia A. Shewokis, *USA*

Qiongfeng Shi, *China*

Ali Hassan Sodhro, *Sweden*

L. Stergioulas, *Netherlands*

Jasjit S. Suri, *USA*

Kenji Suzuki, *Japan*

Abdelmalik TALEB-AHMED, *France*

Sukun Tian, *China*

Erfan Babae Tirkolaee, *Turkey*

Miguel Garcia Torres, *Spain*

Igor Tsigelny, *USA*

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*

Hui Yu, *UK*

Yudong Zhang, *UK*

Yu Zhang, *USA*

Wensheng Zhang, *China*

Zhuhuang Zhou, *China*

Shang-Ming Zhou, *UK*

Youth Editorial Board

Members*

Yankai Chen, *USA*

Qiong Chen, *China*

Sibo Cheng, *France*

Bu Chenyang, *China*

Afify Heba, *Egypt*

Jiashuang Huang, *China*

Hongxin Pan, *China*

Yuchen Pan, *China*

Shuo Wang, *China*

*Editorial Board Members as of October 9, 2025

CONTENTS

REVIEW ARTICLE

- 1 Applications of artificial intelligence in acute stroke imaging**
Arjun Kalyanpur, Neetika Mathur

PERSPECTIVE ARTICLE

- 13 Expertise in AI and clinical publishing exposes peer review gaps: A perspective**
Ezra N. S. Lockhart

ORIGINAL RESEARCH ARTICLES

- 22 Accurate early detection of Parkinson's disease from single photon emission computed tomography imaging through convolutional neural networks**
R. Prashanth
- 33 Deep vision transformers in neurodegenerative disease diagnosis using ¹⁸F-fluorodeoxyglucose positron emission tomography scans and anatomical brain atlas**
Pooriya Khorramyar, Amira Soliman, Farzaneh Etmiani, Stefan Bytner
- 47 Comparison of synthetic data generation techniques for obesity level prediction based on dietary habits and physical status**
Hakan Alp Eren, Halil Ibrahim Emek, Sinem Bozkurt Keser
- 75 A hierarchical federated learning-based health stack for future pandemic preparedness**
Rojalini Tripathy, Asmit Balabantaray, Nisarg Shah, Prashant Kumar Jha, Ajay Kumar Gogineni, Atri Mukhopadhyay, Kisor Kumar Sahu, Padmalochan Bera
- 92 Artificial intelligence versus humans: A comparative analysis of time, cost, and performance on a clinical code conversion task**
Carly Hudson, Marcus Randall, Candice Bowman, Anu Joy, Adrian Goldsworthy
- 103 Stratifying autonomic nervous system regulation patterns in healthy men: A machine learning approach**
Wollner Materko
- 114 RefSAM3D: Adapting the Segment Anything Model with cross-modal references for three-dimensional medical image segmentation**
Xiang Gao, Kai Lu
- 129 Leveraging the smarts in your phone: An artificial intelligence-driven iOS application for neurosurgical navigation of external ventricular drains**
Andrew Abumoussa, Benjamin Succop, Carolyn Quinsey, Yueh Lee, Sivakumar Jaikumar

BRIEF REPORT

- 139 Feasibility of artificial intelligence-driven personalized learning for internal medicine residents: Integrating adaptive artificial intelligence in flipped classrooms**
Marcos A. Sanchez-Gonzalez, Noelani-Mei Ascio, Omar Shah, Ashley Matejka, Mark Terrell, Salman Muddassir

REVIEW ARTICLE

Applications of artificial intelligence in acute stroke imaging

Arjun Kalyanpur¹  and Neetika Mathur^{2*} ¹Department of Clinical Radiology, dAlgnostiX/Teleradiology Solutions, Bengaluru, Karnataka, India²Department of Clinical Training and Research, dAlgnostiX/Teleradiology Solutions, Bengaluru, Karnataka, India**Abstract**

Stroke remains a major global public health challenge, representing the second leading cause of death worldwide and a primary contributor to long-term disability. The paradigm “time is brain” underscores the importance of treating stroke patients within the critical window period, ideally within 60 min from symptom onset, to minimize damage and improve outcomes. The integration of artificial intelligence (AI) into stroke imaging has transformed diagnosis and management by increasing speed, accuracy, and efficiency. AI algorithms have been trained to detect acute stroke, assess hemorrhage, detect and quantify midline shifts, calculate automated Alberta Stroke Program Early Computed Tomography Scores, and identify dense middle cerebral artery on non-contrast computed tomography (CT) as well as large vessel occlusions on CT angiograms, with high sensitivity and specificity. AI also aids in treatment guidance and outcome monitoring. This review provides insights into AI applications in acute stroke imaging, including its role in early detection, screening, triage and prioritization, automated image analysis, workflow optimization, and system integration. Despite its benefits, AI adoption faces challenges such as clinical validation, ethical considerations, and integration into existing workflows. Future developments depend on large, diverse, and well-annotated datasets to train more robust AI systems capable of guiding treatment strategies and improving patient outcomes. The seamless integration of cloud-based AI solutions with telerreporting platforms has the potential to revolutionize stroke care by enabling rapid, high-quality radiologic interpretation, even in remote locations.

***Corresponding author:**Neetika Mathur
(neetika.mathur@imagecorelab.com)**Citation:** Kalyanpur A, Mathur N. LLMs-Healthcare: Applications of artificial intelligence in acute stroke imaging. *Artif Intell Health*. 2025;2(4):1-12.
doi: 10.36922/AIH025140025**Received:** March 31, 2025**Revised:** June 17, 2025**Accepted:** July 4, 2025**Published online:** July 24, 2025**Copyright:** © 2025 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.**Keywords:** Artificial intelligence; Stroke imaging; Computed tomography angiograms; Magnetic resonance angiography; Hemorrhage; Teleradiology; Workflow integration**1. Introduction**

Stroke is the second leading cause of mortality and a major global health concern, responsible for about 5.5 million deaths annually, and was the fourth-highest Level 3 cause of disability-adjusted life years (DALYs) in 2021.¹ It is a medical emergency defined by the sudden reduction of blood flow to the brain, leading to cell death and functional impairments. According to the World Health Organization, stroke accounts for nearly 11% of global deaths.² Furthermore, global DALYs due to stroke have surged from 119.89 million in 1990 to 159.86 million in 2021, driven by population growth and

increased life expectancy. This figure is projected to rise to 186.88 million by 2030 and 224.86 million by 2050, as reported by Cheng *et al.*³ in the Global Burden of Disease 2021 study. In India, stroke is the fourth-leading cause of death and the fifth-leading cause of disability. The stroke-related death rate in India has increased from 44 to 55 people per 100,000 population between 1990 and 2021.² According to a study conducted by Pandian *et al.*,⁴ India has the highest DALYs due to stroke among countries in the Southeast Asia Region.

A striking feature in India is that a large proportion of the stroke patients are from the younger population, unlike in developed countries. Nearly 20% of patients hospitalized with a first-time stroke are under 40 years of age.⁵ Younger individuals are increasingly at risk due to sedentary lifestyles, substance use (including tobacco, nicotine, alcohol, and illicit drugs), and stress. Other risk factors involve elevated blood pressure, blood sugar, cholesterol, and body weight.^{6,7}

The paradigm “time is brain” is pivotal in stroke care, as millions of neurons die with each minute that a stroke goes untreated. Therefore, treating stroke patients within the critical window period or the golden hour (within 60 min of symptom onset) is essential. During this time, physicians should administer medication and initiate treatment as quickly as possible.⁸ Beyond this golden hour, irreversible brain damage occurs, making treatment less effective. Treatment strategies include intravenous tissue plasminogen activator for thrombolysis and endovascular treatment (EVT).⁹ Timely intervention is critical to minimize damage and improve outcomes. However, disparities in stroke care persist due to delays in diagnosis, limited access to treatment, and a shortage of radiologists and stroke care experts. Teleradiology has transformed stroke care by enabling rapid, high-quality, and accurate radiologic interpretation, even from remote locations.^{8,10-12} Still, the sharp rise in the volume of radiologic imaging, without a corresponding increase in the number of trained radiologists, necessitates more scalable and efficient solutions.¹³

Neuroimaging is essential for identifying acute strokes and distinguishing between ischemic and hemorrhagic types.¹⁴ Tools such as computed tomography (CT) and magnetic resonance imaging (MRI) are pivotal in detecting, characterizing, and diagnosing strokes. Artificial intelligence (AI) is a rapidly advancing field that offers powerful tools for fast and efficient imaging analysis. Its emergence has enabled the analysis of large datasets, pattern recognition, and prediction with unprecedented speed and accuracy. In healthcare, AI is growing at a rate of 40% per annum and is projected to help decrease healthcare costs by United States Dollar 150 billion by

2026.¹⁵ AI-powered tools further enhance radiologic image analysis, enabling fast and precise identification of ischemic and hemorrhagic strokes, as well as vascular abnormalities, thereby supporting swift and effective stroke management.

Machine learning algorithms assist in analyzing CT angiograms and identifying large vessel occlusions (LVOs) in real time. Studies have shown that these AI tools reduce door-to-treatment times by promptly alerting clinicians.¹⁶ AI can assess imaging data to determine whether a patient is eligible for procedures like mechanical thrombectomy or tissue plasminogen activator administration. AI models also analyze electronic health records, imaging data, and outputs from wearable devices to assess stroke risk. For example, predictive algorithms can detect atrial fibrillation, a major stroke risk factor, from smartwatch electrocardiogram data with high sensitivity.¹⁷

Several studies have evaluated the role of AI in stroke management and patient care.¹⁸⁻²³ A review article by Liu *et al.*¹⁸ highlights the role of AI in areas such as automated segmentation of infarct areas, identification of LVOs, stroke outcome prediction, analysis of hemorrhagic transformation risk, prediction of recurrent ischemic stroke, and automated grading of collateral circulation. Al-Janabi *et al.*¹⁹ provided an overview of the AI tools used to identify strokes and guide acute ischemic stroke care.

This review paper explores the transformative potential of AI in stroke care. It provides an overview of AI applications in acute stroke care imaging, focusing on the advancements in detection and screening, triaging and prioritization, quantification and prognosis, automated image interpretation, and workflow optimization, supported by published review articles on the subject.

2. Methodology

A comprehensive literature search was conducted using the PubMed and Google Scholar databases, focusing on papers evaluating the use of AI in stroke imaging, published between 2014 and 2024. Keywords used for the search included “artificial intelligence in stroke,” “AI in acute stroke,” “AI in hemorrhage,” “AI in ASPECTS score,” “AI in large vessel occlusion,” and “AI in midline shift.”

Studies were included if they focused on the application of AI in stroke imaging, specifically involving acute stroke, hemorrhage detection, Alberta Stroke Program Early Computed Tomography Scoring (ASPECTS), LVO detection, or midline shift assessment. Only original research articles, reviews, and systematic reviews written in English and with full-text availability were considered.

Studies were excluded if they were unrelated to medical imaging or the application of AI in stroke, or if they

were duplicates, conference proceedings, commentaries, editorials, or abstracts without full-text access.

A total of 316 studies were initially retrieved through database searches. After removing 33 duplicates, 283 studies remained for title and abstract screening. Of these, 127 full-text articles were assessed for eligibility based on the inclusion and exclusion criteria. A final total of 78 studies were included in the review.

The study selection process is depicted in [Figure 1](#).

3. Results

AI has introduced a paradigm shift in medical imaging.²⁴ The application of AI in stroke imaging spans multiple domains, including screening, detection, triage, and automated diagnosis of carotid artery disease,²⁵⁻²⁷ brain hemorrhage and infarct segmentation, quantification, and prognosis; distinguishing ischemic from non-ischemic tissue and normal versus infarcted brain;²⁸⁻³⁶ midline shift detection and quantification;³⁷⁻⁴¹ automated ASPECT score calculation;⁴²⁻⁴⁶ and detection of dense middle cerebral artery (MCA) and LVO on CT angiograms.⁴⁷⁻⁵³ Several commercially available AI-integrated workflows have been developed to interpret ischemic stroke imaging automatically ([Table 1](#)). These AI-driven tools enhance workflow integration, optimize radiological interpretation, and improve stroke management.

3.1. AI in stroke screening

Carotid artery stenosis is commonly associated with plaque progression and accounts for 10 – 20% of ischemic

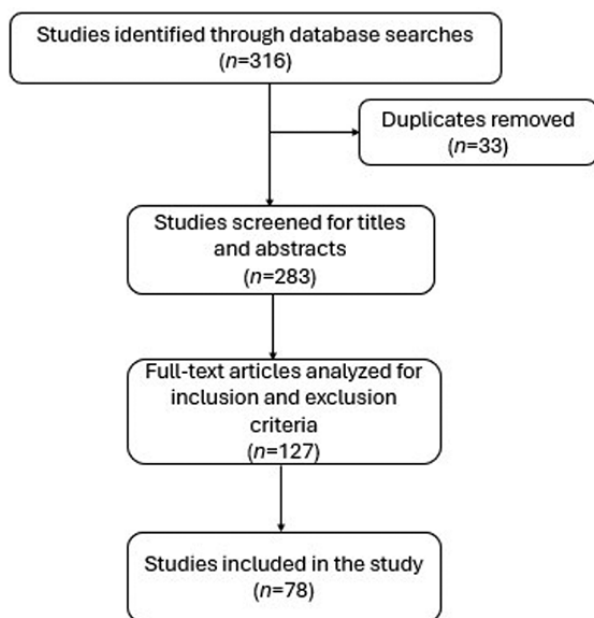


Figure 1. The article search and screening process

strokes due to atherothrombotic disease. Non-invasive imaging techniques used to assess the likelihood of atherosclerotic plaque formation and evaluate lumen diameter reduction include MRI, CT angiograms (CTA), and ultrasound imaging. AI enhances imaging interpretation by identifying even minute plaques that may go unnoticed by radiologists, thereby facilitating timely diagnosis and treatment of carotid artery disease. It also helps standardize the identification and quantification of carotid plaque across various medical imaging centers and among different physicians. A study by Kordzadeh *et al.*²⁵ demonstrated the applicability and precision of AI in detecting carotid artery disease using grayscale static duplex ultrasound images. In their findings, the AI system achieved 91% sensitivity, 86% specificity, and 92% accuracy in identifying normal carotid arteries, and 87% sensitivity, 82% specificity, and 90% accuracy in detecting any degree of carotid artery stenosis. Skandha *et al.*²⁶ conducted a study using echocolor Doppler imaging on the internal carotid arteries of 345 patients and developed a diffusion convolutional neural network to distinguish symptomatic and asymptomatic plaques, achieving an accuracy of 95.66% (area under the curve [AUC] 0.956, $p < 0.0001$). AI algorithms also improve the detection and characterization of carotid plaques through CTA and MRA. In CTA, AI enables early plaque detection, standardizes quantification, and assesses plaque vulnerability. In MRA, AI estimates varying degrees of carotid artery stenosis and automates risk assessment using MRI-based models, such as the high-risk plaque MRI model, which automatically estimates risk scores related to plaque vulnerability.²⁷ These algorithms play a pivotal role as segmentation systems, differentiating between different layers (such as the lumen, outer wall, and lipid core), and various components of atherosclerotic plaque on T1- and proton density-weighted images, enabling precise identification of plaque contours and vulnerable lesions.

3.2. AI in acute stroke imaging

3.2.1. Assessment of hemorrhage

Hemorrhagic strokes, classified based on the location of bleeding, include subarachnoid hemorrhage, intraparenchymal hemorrhage, and intraventricular hemorrhage.²⁸ AI algorithms have shown high sensitivity and specificity in detecting hemorrhages, even in challenging cases involving small bleeds or complex brain anatomy. These tools are capable of segmenting and quantifying hemorrhages, thereby improving classification and localization. For instance, a study by Rava *et al.*²⁸ demonstrated that the AI could automate the detection and triage of patients undergoing non-contrast CT (NCCT)

Table 1. List of artificial intelligence (AI) algorithms in acute stroke imaging, along with their analytical performance metrics

No.	Findings	AI model	Vendor	Sensitivity (%)	Specificity (%)	Accuracy (%)	Cohort size	Study design type	Reference
1	Hemorrhage	AUTOStroke Solution	Canon	93	93	NA	200	Retrospective	Rava <i>et al.</i> ²⁸
		Qure.ai	Qure.ai	NA	NA	NA	21,095	Retrospective	Chilamkurthy <i>et al.</i> ²⁹
		Neural Assist	TeleradTech	92	84	84	21,420	Prospective	-
2	Midline shift	qER-Quant software	Qure.ai	95	95	NA	313,318 head CT	Retrospective	Chilamkurthy <i>et al.</i> ²⁹
		Neural Assist	TeleradTech	84	89	89	22,729	Prospective	-
3	ASPECT score	AI DLAD	D.LABS	65	82	80	258	Retrospective	Chiang <i>et al.</i> ⁴³
		Deep-ASPECTS	Qure.ai	77	99	NA	5,000	Retrospective	Upadhyay <i>et al.</i> ⁴⁴
		RAPID ASPECTS	iSchemaView	NA	NA	NA	100	Retrospective	Maegerlein <i>et al.</i> ⁴⁵
		e-ASPECTS	Brainomix	44	93	87	2,640	Retrospective	Nagel <i>et al.</i> ⁴⁶
4	Dense MCA	Xception Model	viso.ai	82.90	89.70	86.50	18,396	Retrospective	Shinohara <i>et al.</i> ⁴⁸
		Neural Assist	TeleradTech	56.25	94	89.7	22,708	Prospective	-
5	LVO	Viz-LVO	Viz.ai	80.3	82.9	82.70	610	Retrospective	Rodrigues <i>et al.</i> ⁵¹
		AUTOStroke Solution	Canon	73	98	81	303	Retrospective	Rava <i>et al.</i> ⁵²
		RAPID-CTA	Rapid AI	94	76	NA	477	Retrospective	Amukotuwa <i>et al.</i> ⁵³
6	CT Perfusion analysis	Viz CTP	Viz.ai	80	86.20	NA	94 labeled training images and 62 unlabeled testing images	Retrospective	Soun <i>et al.</i> ¹
		e-CTP	Brainomix®	NA	NA	NA	111	Retrospective	Shahrouki <i>et al.</i> ⁵⁷

Note: This table enlists selected examples of AI algorithms currently available on the market and does not represent a complete list. Abbreviations: ASPECTS: Alberta Stroke Program Early Computed Tomography Score; CT: Computed tomography; CTA: Computed tomography angiograms; CTP: Computed tomography perfusion; DLAD: Deep learning-based automatic detection; LVO: Large vessel occlusion; MCA: Middle cerebral artery; NA: Not available.

by classifying them as intracranial hemorrhage positive or negative, with a specificity of 0.93 ± 0.01 , sensitivity of 0.93 ± 0.03 , positive predictive value of 0.85 ± 0.02 , and negative predictive value of 0.98 ± 0.01 . Similarly, the AI algorithm *Neural Assist* by TeleradTech classifies, localizes, and quantifies hemorrhages with 92% sensitivity and 83% specificity, as prospectively studied in a cohort of 21,420 scans. It accurately detects various hemorrhage types with an overall accuracy of 85% (Figures 2A-D and 3). *Neural Assist* processes non-contrast adult head CT Digital Imaging and Communications in Medicine (DICOM) files and analyzes them to detect intracranial hemorrhage, midline shift, cranial fractures, and dense MCA signs. It prioritizes critical scans by generating priority flags and automatically produces a preliminary report to support specialist review (Figure 4). The output is a structured report available in DICOM, PDF, or DOC formats. Additionally, a study by Chilamkurthy *et al.*²⁹ reported that their AI algorithms achieved AUC values

of 0.8977, 0.9559, 0.9194, 0.9161, 0.9044, and 0.9288 for detecting intraparenchymal hemorrhage, intraventricular hemorrhage, intracranial hemorrhage, subdural hematoma (SDH), subarachnoid hemorrhage, and epidural hematoma, respectively.

3.2.2. Detection of midline shift

Midline shift is a crucial indicator of the lateral displacement of midline structures of the brain due to trauma or mass effects resulting from hematomas, tumors, abscesses, or intracranial lesions. It serves as a key prognostic feature in stroke.³⁷ AI tools used to measure midline shift are generally categorized into two types: Symmetry-based approaches, which calculate the curve of the deformed midline, and landmark-based approaches, which detect anatomical landmarks such as the septum pellucidum within specified ventricular regions and measure midline shift accordingly.³⁸ Chilamkurthy *et al.*²⁹

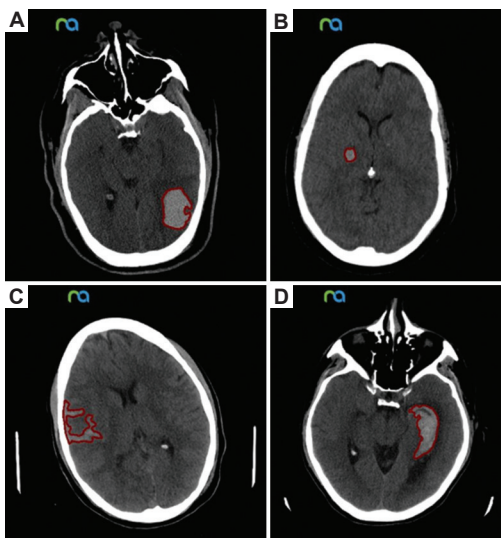


Figure 2. TeleradTech artificial intelligence *Neural Assist* algorithm in detection of hemorrhages: (A) and (B) intraparenchymal hemorrhage, (C) subarachnoid hemorrhage, (D) intraventricular hemorrhage

Hemorrhage findings

Type of hemorrhage	Hemorrhage location	Quantification
Intraparenchymal hemorrhage (IPH)	Right ganglio-thalamus (Slice -13)	Length: 7 mm, Width: 1.2cm



Figure 3. Classification, localization, and quantification of hemorrhage by TeleradTech's artificial intelligence *Neural Assist* algorithm

proposed an AI model that detects midline shift (MLS) with an AUC of 0.9276. Nguyen *et al.*³⁹ developed a deep learning algorithm that attained a case-level midline shift identification AUC of 95.3%, utilizing a testing dataset of 2,545 NCCT head scans, and measured midline shift with an average absolute error of 1.20 mm across 228 midline shift-positive cases. Chen *et al.*⁴⁰ described an automated process using CT imaging to quantify MLS and triage for elevated intracranial pressure. The AI *Neural Assist* algorithm developed by TeleradTech detects midline shift with 84% sensitivity and 89% specificity, based on a cohort of 22,729 patients (Figure 5).

3.2.3. ASPECTS analysis

The ASPECTS is a scoring system generally used to guide treatment strategies for patients presenting with MCA

DRAFT REPORT BY NEURAL ASSIST

Study UID: ...

Case Type: **Positive**

Hemorrhage	Midline Shift	Calvarial Fracture	Dense MCA
Positive	Negative	Negative	Not Evaluated

Hemorrhage Findings

Type of Hemorrhage	Hemorrhage Location	Quantification
Intraventricular hemorrhage (IVH)	Fourth Ventricle (Slice - 7), Third Ventricle (Slice - 14), Left Lateral Ventricle (Slice - 15)	Extensive

Slice - 7 Slice - 14

Figure 4. Draft radiology report generated by TeleradTech's artificial intelligence *Neural Assist* algorithm

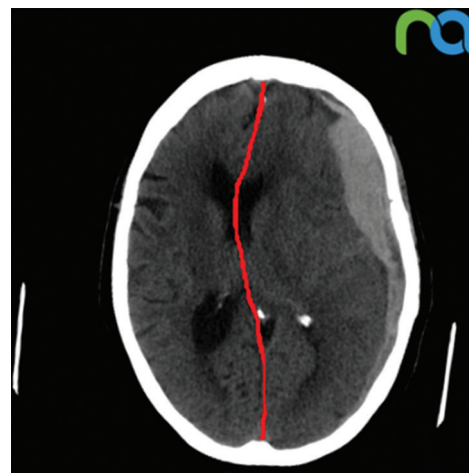


Figure 5. Detection of midline shift by TeleradTech artificial intelligence *Neural Assist* algorithm

ischemic strokes using NCCT brain scans.⁴² AI algorithms assist in automating ASPECTS calculation, enabling rapid and accurate evaluation of acute ischemic stroke severity on NCCT scans and ultimately improving stroke care. For example, Chiang *et al.*⁴³ studied the potential of a deep learning-based automatic detection (DLAD) algorithm for ASPECT scoring on NCCT images in patients with symptoms of acute ischemic stroke. The DLAD achieved 65% sensitivity, 82% specificity, and 80% accuracy in ASPECTS prediction, thus enhancing and expediting

physician decision-making. Upadhyay *et al.*⁴⁴ evaluated an AI algorithm for automated ASPECT scoring, which decreased diagnosis time for NCCT scans and demonstrated a 76.19% agreement with radiologists. AI-assisted ASPECT scoring systems have shown outcomes comparable to, or in some cases better than, manual assessments by clinicians. They demonstrate good to excellent reliability, with intraclass correlation coefficients indicating strong agreement with expert consensus and reference standards. In a study by Maegerlein *et al.*,⁴⁵ AI-generated ASPECTS in acute MCA stroke showed better agreement with predefined consensus scores than human readers alone. AI tools not only reduce inter-observer variability but also enhance clinical decision-making by providing quick and reliable ASPECT scores, which are critical for assessing the severity of acute ischemic stroke and determining patient eligibility for treatments like thrombectomy and thrombolysis.^{46,47} Additionally, features such as heat maps indicate the probability of low attenuation and sulcal effacement.

3.2.4. MCA

In a study conducted by Shinohara *et al.*⁴⁸ on a cohort of patients with acute ischemic stroke, the diagnostic performance of a deep convolutional neural network model (Xception) was evaluated for the identification and prioritization of the hyperdense MCA sign on NCCT. The model demonstrated a sensitivity of 82.9%, specificity of 89.7%, and accuracy of 86.5% using leave-one-case-out cross-validation. Furthermore, the AI *Neural Assist* algorithm developed by TeleradTech detected dense MCA with an accuracy of 89.7% (Figure 6).

3.2.5. LVO

AI algorithms enable rapid and accurate detection of LVO, facilitating timely alerts and swift decision-making for reperfusion treatments or transfer to specialized stroke centers when needed. Various studies have

demonstrated that AI tools can precisely identify LVO on CTA in real time.⁴⁹⁻⁵³ Le *et al.*⁵⁰ demonstrated that a machine learning algorithm used for automated LVO detection on CTA, coupled with secure communication at non-EVT-performing primary stroke centers, significantly reduced door-in-door-out time by promptly alerting clinicians. This intervention increased the number of patients undergoing EVT after transfer, ultimately improving patient outcomes. In a retrospective study by Rodrigues *et al.*⁵¹ found that the AI Viz-LVO Algorithm[®] version 1.4 detected internal carotid artery and MCA-M1 LVOs with a sensitivity of 87.6%, specificity of 88.5%, and accuracy of 87.9% (AUC 0.88). Similarly, Rava *et al.*,⁵² in a study on acute ischemic stroke patients, reported that the ^{AUTO}Stroke Solution LVO achieved 73% sensitivity, 98% specificity, and 81% accuracy in correctly identifying and localizing LVOs. The accuracy, sensitivity, and Matthews correlation coefficients of the algorithm for detecting different occlusion types were as follows: 0.95, 0.90, and 0.89, respectively, for the internal carotid artery; 0.89, 0.77, and 0.78, respectively, for the M1 segment of the MCA; and 0.80, 0.51, and 0.59, respectively, for the M2 segment of the MCA. Additionally, the RAPID CTA AI solution showed strong potential in detecting intracranial LVO, with a sensitivity of 94% and specificity of 76%, as revealed in a study conducted by Amukotuwa *et al.*⁵³

3.3. Perfusion analysis

CT perfusion (CTP) imaging has emerged as a key imaging technique for assessing acute ischemic stroke and determining eligibility for endovascular clot retrieval in cases of LVO.^{54,55} Cerebral blood flow and volume, mean transit time, and other pseudocolor perfusion variables are leveraged to evaluate the condition of ischemic brain tissue. Research by Hu *et al.*⁵⁶ emphasized that the quality of AI-based CTP pseudocolor images was superior compared to the control group ($p < 0.05$), enabling easier, faster, and more precise identification of ischemic strokes, hemorrhagic strokes, and vascular abnormalities. This aids physicians in detecting the infarct location and assessing cerebral blood flow. A retrospective study by Shahrrouki *et al.*⁵⁷ demonstrated the ability of the AI tool e-Stroke Suite (Brainomix[®]) in accurately estimating ischemic core volumes using both NCCT and CTP, with mean volumes of about 21 mL and 20 mL, respectively, in a cohort of 111 patients.^{19,57} Mallon *et al.*⁵⁸ prospectively evaluated the Brainomix[®] e-Stroke AI in 551 patients and found it demonstrated 58.6% sensitivity, 83.5% specificity, and 77% accuracy for acute ischemic stroke. The tool also showed strong concordance in perfusion data for both core and penumbra zones, facilitating rapid and definitive diagnosis.

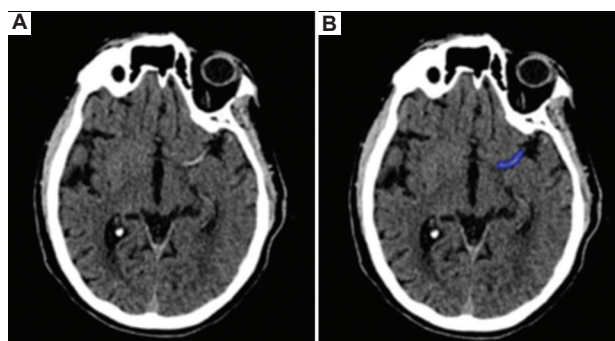


Figure 6. TeleradTech artificial intelligence (AI) *Neural Assist* algorithm for detection of dense middle cerebral artery (MCA): (A) Original image; (B) AI-interpreted image showing dense MCA (blue margin)

3.4. Automation: Workflow and triage

Commercially available algorithms are being integrated into the clinical workflows of numerous large institutions, both in practice and trials, to provide automated triage and segmentation of acute stroke cases. These tools help decrease the workload on radiologists and enhance diagnostic accuracy. Automated ASPECT scoring supports treatment teams in selecting patients for endovascular therapy. Overall, such tools offer rapid and efficient analyses to improve stroke care at both spoke and hub hospitals and reduce the turnaround times in medical workflows. In a retrospective study by Colasurdo *et al.*,⁵⁹ a convolutional neural network was incorporated into the institutional workflow to detect SDH from NCCT head scans. The subdural convolutional neural network showed 91.4% sensitivity, 96.4% specificity, and 95.1% accuracy; for the subgroup with SDH thickness greater than 10 mm, the sensitivity reached 100%.

Retrospective research by Soun *et al.*¹ also demonstrated that the integration of an AI algorithm into the hospital system supported triage of CTA in ischemic stroke cases, enabling automatic identification of LVO cases with 96% sensitivity and 85% specificity, and a turnaround time of 22 min. The AI was accessible in the Picture Archival and Communication System via a web or mobile application.

3.5. Integration with teleradiology workflow

Teleradiology has been addressing the global challenge of radiologist shortages.⁶⁰⁻⁶⁴ The seamless integration of cloud-based AI solutions with telereporting platforms enhances workflow by prioritizing critical cases, sharing automated alerts to stroke teams for prompt action, and extending the benefits of AI across multiple domains, including remote or underserved areas.⁶⁵⁻⁷¹ However, data security poses a significant challenge for cloud-based AI algorithms. Implementing robust cybersecurity systems is pivotal to ensure secure integration of AI into the teleradiology workflow. Another challenge is ensuring that AI outputs are accessible within the teleradiologist's viewer. The use of aggregator platforms and workflows that consolidate outputs from multiple AI tools would support seamless telereporting. Furthermore, leveraging clinical data from teleradiology, incorporating feedback from teleradiologists, and collaborating with AI developers for training, upgrading, and validating AI systems will further streamline integration in real time. A potential obstacle is the lack of adequate infrastructural support for AI integration. Employing Graphic Processing Units would allow efficient and fast processing of large volumes of data for AI development.^{13,72}

3.6. Ethical, legal, and social implications of AI in stroke imaging

The application of AI has revolutionized stroke imaging; however, ethical, legal, and societal implications present barriers that need to be addressed. Potential biases in training data and the decision-making process of AI (often referred to as the “black box” nature) raise ethical and societal concerns. These can be mitigated by implementing a robust framework that emphasizes data security, patient privacy, and fair and equitable access to AI applications in healthcare.⁷³

Multidisciplinary discussions on the advantages and limitations of using AI in healthcare, among all stakeholders, including clinicians, AI developers, administrative personnel, and policymakers, are essential. Standardized protocols and regulations should be established to promote impartiality, clarity, trustworthiness, accountability, confidentiality, and compassion in the development of AI within an ethical framework.⁷⁴

4. Challenges and future directions

The development and deployment of AI platforms in clinical settings have been instrumental in transforming stroke care by lowering mortality and improving quality of life. However, several challenges constrain their widespread adoption. One major challenge is the limited generalizability of datasets, i.e., AI models are often trained on single-center or homogeneous datasets, which may result in underperformance when applied to external populations or systems with different scanners, imaging protocols, electronic health record systems, laboratory equipment, and varying clinical and administrative procedures. To improve generalizability and performance, continuous learning from large, diverse, multicenter, and high-quality annotated datasets is essential.^{29,75}

Another challenge is the “black box” nature of numerous AI models, which limits interpretability, reliability, and transparency in their decision-making processes, hindering widespread acceptance (1). The development and implementation of heat maps, prediction-based modules, user-friendly interfaces, interactive dashboards, and visualization tools can help make AI insights more understandable, thereby addressing the “black box” problem.⁷⁶ The white paper of the Italian Society of Medical and Interventional Radiology emphasizes the urgent need for explainable AI (xAI), which can reveal the rationale behind AI decision-making, offering insights into its strengths, limitations, and potential future performance.⁷⁷ Furthermore, the ongoing training of technologists, radiologists, and physicians through workshops and

continuous medical education is vital for keeping pace with advancements in AI tools and techniques.

The lack of historical records also limits diagnostic accuracy. Integrating multimodal data, including clinical history and laboratory results, with stroke imaging is crucial for prognostic analysis, allowing timely diagnosis, early intervention, treatment guidance, and outcome monitoring.⁷⁸

Finally, regulatory compliance and the integration of AI into clinical workflows are paramount. AI tools must be rigorously validated and approved by regulatory bodies such as the Food and Drug Administration prior to their deployment in clinical settings. Despite these challenges, AI algorithms hold immense promise as transformative tools in stroke care.¹³

5. Conclusion

Acute stroke is a time-sensitive clinical situation where swift assessment and treatment are critical. The refinement of guidelines and protocols, along with the implementation of technologies that reduce time to treatment, will remain central areas of focus in stroke care. The development and integration of AI algorithms into clinical workflows can detect subtle signs of stroke, quantify infarct size, assess collateral status, predict patient outcomes, and guide prognosis and post-stroke recovery planning. AI has revolutionized stroke imaging by improving detection, enabling synchronous communication, and enhancing triage, diagnosis, and prognosis assessment.

Emerging AI technologies should be leveraged with transparency, supported by appropriate legislation and regulation, to enhance both clinical impact and the credibility of these algorithms.

In conclusion, the integration of AI tools into the teleradiology workflow can significantly address global workforce shortages in stroke care and tackle several challenges, including ethical, legal, and societal implications.

Glossary

Term	Definition
Alberta Stroke Program Early Computed Tomography Score (ASPECTS)	A 10-point quantitative scoring system used to assess the extent of early ischemic changes in the brain on computed tomography scans following an acute ischemic stroke
Deep learning	A subset of machine learning that uses multilayered neural networks, known as deep neural networks, to simulate complex decision-making processes similar to those of the human brain
Artificial intelligence vendors	Companies that provide access to their proprietary artificial intelligence models, typically via Application Programming Interfaces (APIs)

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Arjun Kalyanpur

Formal analysis: Neetika Mathur

Investigation: All authors

Methodology: Neetika Mathur

Supervision: Arjun Kalyanpur

Visualization: Arjun Kalyanpur

Writing – original draft: Neetika Mathur

Writing – review & editing: Arjun Kalyanpur

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

- Soun JE, Chow DS, Nagamine M, *et al.* Artificial intelligence and acute stroke imaging. *AJNR Am J Neuroradiol.* 2021;42(1):2-11.
doi: 10.3174/ajnr.A6883
- Behera DK, Rahut DB, Mishra S. Analyzing stroke burden and risk factors in India using data from the Global Burden of Disease Study. *Sci Rep.* 2024;14(1):22640.
doi: 10.1038/s41598-024-72551-4
- Cheng Y, Lin Y, Shi H, *et al.* Projections of the stroke burden at the global, regional, and national levels up to 2050 based on the global burden of disease study 2021. *JAHA.* 2024;13:e036142.
doi: 10.1161/JAHA.124.036142
- Pandian JD, Padma Srivastava MV, Aaron S, *et al.* The burden, risk factors and unique etiologies of stroke in South-East Asia Region (SEAR). *he Lancet Reg Health Southeast Asia.* 2023;17:100290.
doi: 10.1016/j.lanse.2023.100290
- Jones SP, Baqai K, Clegg A, *et al.* Stroke in India: A systematic

- review of the incidence, prevalence, and case fatality. *Int J Stroke*. 2022;17(2):132-140.
doi: 10.1177/17474930211027834
6. Rangamani S, Huliappa D, Kulothungan V, *et al*. Stroke incidence, mortality, subtypes in rural and urban populations in five geographic areas of India (2018–2019): results from the National Stroke Registry Programme. *Lancet Reg Health Southeast Asia*. 2024;23:100308.
doi: 10.1016/j.lansea.2023.100308
7. Tripathi M, Vibha D. Stroke in young in India. *Stroke Res Treat*. 2010;2011:368629.
doi: 10.4061/2011/368629
8. Kalyanpur A, Mathur N. A Teleradiology system for early ischemic and hemorrhagic stroke evaluation and management. *J Clin Interv Radiol ISVIR*. 2023;7(3):183-189.
doi: 10.1055/s-0043-1771379
9. Maier O, Menze BH, Von Der Gablentz J, *et al*. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. *Med Image Anal*. 2017;35:250-269.
doi: 10.1016/j.media.2016.07.009
10. Johnston KC, Worrall BB. Teleradiology assessment of computerized tomographs online reliability study (TRACTORS) for acute stroke evaluation. *Telemed J e-Health*. 2003;9(3):227-233.
doi: 10.1089/153056203322502605
11. Krupinski E. Teleradiology: Current perspectives. *RMI*. 2014;7:5-14.
doi: 10.2147/RMI.S48140
12. Bogner P, Chadaide Z, Lenzsér G, *et al*. Teleradiology-based stroke network in Western and Southern Transdanubia in Hungary. *Orv Hetil*. 2021;162(17):668-675.
doi: 10.1556/650.2021.32097
13. Kalyanpur A, Mathur N. Integration of teleradiology and artificial intelligence: Opportunities and challenges. *MRAJ*. 2024;12(10):1-8.
doi: 10.18103/mra.v12i10.5904
14. Nukovic JJ, Opancina V, Ciceri E, *et al*. Neuroimaging modalities used for ischemic stroke diagnosis and monitoring. *Medicina (Kaunas)*. 2023;59(11):1908.
doi: 10.3390/medicina59111908
15. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: *Artificial Intelligence in Healthcare*. Netherlands: Elsevier; 2020. p. 25-60.
doi: 10.1016/B978-0-12-818438-7.00002-2
16. Hassan AE, Ringheanu VM, Preston L, Tekle WG. Artificial intelligence-parallel stroke workflow tool improves reperfusion rates and door-in to puncture interval. *SVIN*. 2022;2(5):e000224.
doi: 10.1161/SVIN.121.000224
17. Tison GH, Sanchez JM, Ballinger B, *et al*. Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiol*. 2018;3(5):409.
doi: 10.1001/jamacardio.2018.0136
18. Liu Y, Wen Z, Wang Y, *et al*. Artificial intelligence in ischemic stroke images: Current applications and future directions. *Front Neurol*. 2024;15:1418060.
doi: 10.3389/fneur.2024.1418060
19. Al-Janabi OM, El Refaei A, Elgazzar T, *et al*. Current stroke solutions using artificial intelligence: A review of the literature. *Brain Sci*. 2024;14(12):1182.
doi: 10.3390/brainsci14121182
20. El Naamani K, Musmar B, Gupta N, *et al*. The Artificial intelligence revolution in stroke care: A decade of scientific evidence in review. *World Neurosurg*. 2024;184:15-22.
doi: 10.1016/j.wneu.2024.01.012
21. Shafaat O, Bernstock JD, Shafaat A, *et al*. Leveraging artificial intelligence in ischemic stroke imaging. *J Neuroradiol*. 2022;49(4):343-351.
doi: 10.1016/j.neurad.2021.05.001
22. Mouridsen K, Thurner P, Zaharchuk G. Artificial intelligence applications in stroke. *Stroke*. 2020;51(8):2573-2579.
doi: 10.1161/STROKEAHA.119.027479
23. Yedavalli VS, Tong E, Martin D, Yeom KW, Forkert ND. Artificial intelligence in stroke imaging: Current and future perspectives. *Clin Imaging*. 2021;69:246-254.
doi: 10.1016/j.clinimag.2020.09.005
24. Pinto-Coelho L. How Artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering (Basel)*. 2023;10(12):1435.
doi: 10.3390/bioengineering10121435
25. Kordzadeh A, Askari A, Abbassi OA, Sanoudos N, Mohaghegh V, Shirvani H. Artificial intelligence and duplex ultrasound for detection of carotid artery disease. *Vascular*. 2023;31(6):1187-1193.
doi: 10.1177/17085381221107465
26. Skandha SS, Gupta SK, Saba L, *et al*. 3-D optimized classification and characterization artificial intelligence paradigm for cardiovascular/stroke risk stratification using carotid ultrasound-based delineated plaque: Atheromatic™ 2.0. *Comput Biol Med*. 2020;125:103958.
doi: 10.1016/j.combiomed.2020.103958
27. Miceli G, Rizzo G, Basso MG, *et al*. Artificial intelligence in symptomatic carotid plaque detection: A narrative review.

- Appl Sci.* 2023;13(7):4321.
doi: 10.3390/app13074321
28. Rava RA, Seymour SE, LaQue ME, *et al.* Assessment of an artificial intelligence algorithm for detection of intracranial hemorrhage. *World Neurosurg.* 2021;150:e209-e217.
doi: 10.1016/j.wneu.2021.02.134
29. Chilamkurthy S, Ghosh R, Tanamala S, *et al.* *Development and Validation of Deep Learning Algorithms for Detection of Critical Findings in Head CT Scans.* [arxiv preprint]; 2018.
doi: 10.48550/ARXIV.1803.05854
30. Agarwal S, Wood D, Grzeda M, *et al.* Systematic review of artificial intelligence for abnormality detection in high-volume neuroimaging and subgroup meta-analysis for intracranial hemorrhage detection. *Clin Neuroradiol.* 2023;33(4):943-956.
doi: 10.1007/s00062-023-01291-1
31. Chiramal JA, Johnson J, Webster J, *et al.* Artificial Intelligence-based automated CT brain interpretation to accelerate treatment for acute stroke in rural India: An interrupted time series study. *PLOS Glob Public Health.* 2024;4(7):e0003351.
doi: 10.1371/journal.pgph.0003351
32. Gilotra K, Swarna S, Mani R, Basem J, Dashti R. Role of artificial intelligence and machine learning in the diagnosis of cerebrovascular disease. *Front Hum Neurosci.* 2023;17:1254417.
doi: 10.3389/fnhum.2023.1254417
33. Hillis JM, Bizzo BC, Newbury-Chaet I, *et al.* Evaluation of an artificial intelligence model for identification of intracranial hemorrhage subtypes on computed tomography of the head. *SVIN.* 2024;4(4):e001223.
doi: 10.1161/SVIN.123.001223
34. Zhao Z, Zhang Y, Su J, *et al.* A comprehensive review for artificial intelligence on neuroimaging in rehabilitation of ischemic stroke. *Front Neurol.* 2024;15:1367854.
doi: 10.3389/fneur.2024.1367854
35. Savage CH, Tanwar M, Elkassem AA, *et al.* Prospective evaluation of artificial intelligence triage of intracranial hemorrhage on noncontrast head CT examinations. *Am J Roentgenol.* 2024;223(5):e2431639.
doi: 10.2214/AJR.24.31639
36. Vacek A, Mair G, White P, *et al.* Evaluating artificial intelligence software for delineating hemorrhage extent on CT brain imaging in stroke. *J Stroke Cerebrovasc Dis.* 2024;33(1):107512.
doi: 10.1016/j.jstrokecerebrovasdis.2023.107512
37. Agrawal D, Joshi S, Poonamallee L. Automated midline shift detection and quantification in traumatic brain injury: A comprehensive review. *Indian J Neurotrauma.* 2024;21(1):6-12.
doi: 10.1055/s-0043-1777676
38. Liao CC, Chen YF, Xiao F. Brain midline shift measurement and its automation: A review of techniques and algorithms. *Int J Biomed Imaging.* 2018;2018:4303161.
doi: 10.1155/2018/4303161
39. Nguyen NP, Yoo Y, Chekkoury A, *et al.* Brain Midline Shift Detection and Quantification by a Cascaded Deep Network Pipeline on Non-contrast Computed Tomography Scans. In: *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).* United States: IEEE; 2021. p. 487-495.
doi: 10.1109/ICCVW54120.2021.00059
40. Chen W, Belle A, Cockrell C, Ward KR, Najarian K. Automated midline shift and intracranial pressure estimation based on brain CT images. *JoVE.* 2013;74:3871.
doi: 10.3791/3871
41. Yan JL, Chen YL, Chen MY, *et al.* A Robust, fully automatic detection method and calculation technique of midline shift in intracranial hemorrhage and its clinical application. *Diagnostics.* 2022;12(3):693.
doi: 10.3390/diagnostics12030693
42. Pexman JH, Barber PA, Hill MD, *et al.* Use of the alberta stroke program early CT score (ASPECTS) for assessing CT scans in patients with acute stroke. *AJNR Am J Neuroradiol.* 2001;22(8):1534-1542.
43. Chiang PL, Lin SY, Chen MH, *et al.* Deep learning-based automatic detection of ASPECTS in acute ischemic stroke: Improving stroke assessment on CT scans. *J Clin Med.* 2022;11(17):5159.
doi: 10.3390/jcm11175159
44. Upadhyay U, Ranjan M, Golla S, *et al.* *Deep-ASPECTS: A Segmentation-Assisted Model for Stroke Severity Measurement.* [arxiv preprint]; 2022.
doi: 10.48550/ARXIV.2203.03622
45. Maegerlein C, Fischer J, Mönch S, *et al.* Automated calculation of the Alberta stroke program early CT score: Feasibility and reliability. *Radiology.* 2019;291(1):141-148.
doi: 10.1148/radiol.2019181228
46. Nagel S, Sinha D, Day D, *et al.* e-ASPECTS software is non-inferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. *Int J Stroke.* 2017;12(6):615-622.
doi: 10.1177/1747493016681020
47. Dhand S, O'Connor P, Hughes C, Lin SP. Acute ischemic stroke: Acute management and selection for endovascular therapy. *Semin Intervent Radiol.* 2020;37(2):109-118.

- doi: 10.1055/s-0040-1709152
48. Shinohara Y, Takahashi N, Lee Y, Ohmura T, Kinoshita T. Development of a deep learning model to identify hyperdense MCA sign in patients with acute ischemic stroke. *Jpn J Radiol*. 2020;38(2):112-117.
doi: 10.1007/s11604-019-00894-4
49. Brugnara G, Baumgartner M, Scholze ED, *et al*. Deep-learning based detection of vessel occlusions on CT-angiography in patients with suspected acute ischemic stroke. *Nat Commun*. 2023;14(1):4938.
doi: 10.1038/s41467-023-40564-8
50. Le NM, Iyyangar AS, Kim Y, *et al*. Machine learning-enabled automated large vessel occlusion detection improves transfer times at primary stroke centers. *SVIN*. 2024;4(3):e001119.
doi: 10.1161/SVIN.123.001119
51. Rodrigues G, Barreira CM, Bouslama M, *et al*. Automated large artery occlusion detection in stroke: A single-center validation study of an artificial intelligence algorithm. *Cerebrovasc Dis*. 2022;51(2):259-264.
doi: 10.1159/000519125
52. Rava RA, Peterson BA, Seymour SE, *et al*. Validation of an artificial intelligence-driven large vessel occlusion detection algorithm for acute ischemic stroke patients. *Neuroradiol J*. 2021;34(5):408-417.
doi: 10.1177/1971400921998952
53. Amukotuwa SA, Straka M, Smith H, *et al*. Automated detection of intracranial large vessel occlusions on computed tomography angiography: A single center experience. *Stroke*. 2019;50(10):2790-2798.
doi: 10.1161/STROKEAHA.119.026259
54. Alwood BT, Meyer DM, Ionita C, *et al*. Multicenter comparison using two AI stroke CT perfusion software packages for determining thrombectomy eligibility. *J Stroke Cerebrovasc Dis*. 2024;33(7):107750.
doi: 10.1016/j.jstrokecerebrovasdis.2024.107750
55. Albers GW, Lansberg MG, Kemp S, *et al*. A multicenter randomized controlled trial of endovascular therapy following imaging evaluation for ischemic stroke (DEFUSE 3). *Int J Stroke*. 2017;12(8):896-905.
doi: 10.1177/1747493017701147
56. Hu M, Chen N, Zhou X, Wu Y, Ma C. Deep learning-based computed tomography perfusion imaging to evaluate the effectiveness and safety of thrombolytic therapy for cerebral infarct with unknown time of onset. *Contrast Media Mol Imaging*. 2022;2022:9684584.
doi: 10.1155/2022/9684584
57. Shahrouki P, Kihira S, Tavakkol E, *et al*. Automated assessment of ischemic core on non-contrast computed tomography: A multicenter comparative analysis with CT perfusion. *J NeuroIntervent Surg*. 2024;16(12):1288-1293.
doi: 10.1136/jnis-2023-020954
58. Mallon DH, Taylor EJR, Vittay OI, Sheeka A, Doig D, Lobotesis K. Comparison of automated ASPECTS, large vessel occlusion detection and CTP analysis provided by Brainomix and RapidAI in patients with suspected ischaemic stroke. *J Stroke Cerebrovasc Dis*. 2022;31(10):106702.
doi: 10.1016/j.jstrokecerebrovasdis.2022.106702
59. Colasurdo M, Leibushor N, Robledo A, *et al*. Automated detection and analysis of subdural hematomas using a machine learning algorithm. *J Neurosurg*. 2023;138(4):1077-1084.
doi: 10.3171/2022.8.JNS22888
60. Chandramohan A, Krothapalli V, Augustin A, *et al*. Teleradiology and technology innovations in radiology: status in India and its role in increasing access to primary health care. *Lancet Reg Health Southeast Asia*. 2023;23:100195.
doi: 10.1016/j.lansea.2023.100195
61. Božić V. *Radiology, Telemedicine and Artificial Intelligence*; 2023.
doi: 10.13140/RG.2.2.34259.96800
62. Kalyanpur A, Mathur N. The Role of teleradiology in interpretation of ultrasounds performed in the emergency setting. *Digit Diagn*. 2024;5:231-242.
doi: 10.17816/DD624586
63. Kalyanpur A, Meka S, Joshi K, Somashekar Nair HT, Mathur N. Teleradiology in Tripura: Effectiveness of a Telehealth model for the rural health sector. *IJHTI*. 2022;1(2):7-12.
doi: 10.60142/ijhti.v1i02.36
64. Kalyanpur A, Mathur N. Impact of teleradiology on oncological interpretation of PET-CT scans. *Indian J Nucl Med*. 2024;39(6):436-440.
doi: 10.4103/ijnm.ijnm_31_24
65. Kapoor N, Lacson R, Khorasani R. Workflow applications of artificial intelligence in radiology and an overview of available tools. *J Am Coll Radiol*. 2020;17(11):1363-1370.
doi: 10.1016/j.jacr.2020.08.016
66. van de Leemput SC, Prokop M, van Ginneken B, Manniesing R. Stacked bidirectional convolutional LSTMs for deriving 3D non-contrast CT from spatiotemporal 4D CT. *IEEE Trans Med Imaging*. 2020;39(4):985-996.
doi: 10.1109/TMI.2019.2939044
67. Pu A, Wang H, Ying J. Optimized backprojection filtration algorithm for postoperative reduction and analysis of respiratory infection-related factors of pelvic fractures by

- CT imaging. *Sci Program*. 2021;2021:1-10.
doi: 10.1155/2021/3554718
68. Shlobin NA, Baig AA, Waqas M, *et al*. Artificial intelligence for large-vessel occlusion stroke: A systematic review. *World Neurosurg*. 2022;159:207-220.e1.
doi: 10.1016/j.wneu.2021.12.004
69. Zebrowitz E, Dadoo S, Brabant P, *et al*. The impact of artificial intelligence on large vessel occlusion stroke detection and management: A systematic review meta-analysis. *Intell Based Med*. 2024;10:100161.
doi: 10.1016/j.ibmed.2024.100161
70. Murray NM, Unberath M, Hager GD, Hui FK. Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: A systematic review. *J NeuroIntervent Surg*. 2020;12(2):156-164.
doi: 10.1136/neurintsurg-2019-015135
71. Dantas J, Ribeiro G, Dagostin C, *et al*. Can artificial intelligence to detect large vessel occlusion improve patient care? A systematic review and meta-analysis (P5-5.026). *Neurology*. 2024;102(17_supplement_1):6072.
doi: 10.1212/WNL.0000000000206218
72. Leveraging teleradiology with artificial intelligence. *Bull World Health Organ*. 2025;103(2):86-87.
doi: 10.2471/BLT.25.020225
73. Weiner EB, Dankwa-Mullan I, Nelson WA, Hassanpour S. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. *PLOS Digit Health*. 2025;4(4):e0000810.
doi: 10.1371/journal.pdig.0000810
74. Siala H, Wang Y. SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Soc Sci Med*. 2022;296:114782.
doi: 10.1016/j.socscimed.2022.114782
75. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195.
doi: 10.1186/s12916-019-1426-2
76. Marey A, Arjmand P, Alerab ADS, *et al*. Explainability, transparency and black box challenges of AI in radiology: Impact on patient care in cardiovascular radiology. *Egypt J Radiol Nucl Med*. 2024;55(1):183.
doi: 10.1186/s43055-024-01356-2
77. Neri E, Aghakhanyan G, Zerunian M, *et al*. Explainable AI in radiology: A white paper of the Italian Society of Medical and Interventional Radiology. *Radiol Med*. 2023;128(6):755-764.
doi: 10.1007/s11547-023-01634-5
78. Shurrah S, Guerra-Manzanares A, Magid A, Piechowski-Jozwiak B, Atashzar SF, Shamout FE. Multimodal machine learning for stroke prognosis and diagnosis: A systematic review. *IEEE J Biomed Health Inform*. 2024;28(11):6958-6973.
doi: 10.1109/JBHI.2024.3448238

PERSPECTIVE ARTICLE

Expertise in AI and clinical publishing exposes peer review gaps: A perspective

Ezra N. S. Lockhart*

Department of Marriage and Family Sciences, National University, San Diego, California, United States of America

Abstract

Artificial intelligence and machine learning are advancing rapidly in medical and mental health research, yet clinical publishing remains structurally unprepared to evaluate these technologies with the rigor they demand. Despite the rise of AI-driven models for suicide risk prediction and diagnostic assessment, editorial and peer review processes often lack the technical expertise required to assess methodological validity. Drawing on dual fluency in AI and clinical publishing, this perspective identifies a critical gap at the intersection of innovation and editorial oversight. This article reveals how editorial decisions in high-impact psychiatry journals have dismissed valid methodological concerns as “overly technical” and undermined independent scientific critique, drawing on two case studies: one involving a model that differentiates suicidal from non-suicidal self-harm, and another analyzing speech-based suicide risk assessment. These case studies serve as the foundation for a broader critique of editorial decision-making in clinical publishing, revealing persistent structural blind spots in evaluating AI-integrated research. To prevent the pre-mature adoption of flawed models in clinical care, this perspective proposes targeted reforms: recruiting technically proficient reviewers, mandating transparent methodological reporting, and protecting space for independent post-publication evaluation. Without such changes, the integrity of the field and the safety of patients remain at risk.

Keywords: Artificial intelligence; Peer review-research; Ethics-research; Editorial policies; Speech analysis

***Corresponding author:**Ezra N. S. Lockhart
(elockhart@nu.edu)

Citation: Lockhart ENS. Expertise in AI and clinical publishing exposes peer review gaps: A perspective. *Artif Intell Health*. 2025;2(4):13-21. doi: 10.36922/AIH025210049

Received: May 22, 2025**Revised:** June 8, 2025**Accepted:** June 16, 2025**Published online:** July 3, 2025

Copyright: © 2025 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 integration of artificial intelligence (AI) and machine learning (ML) into clinical research is no longer speculative.¹⁻¹⁰ From suicide risk detection to diagnostic classification, AI-driven tools are already shaping the future of mental healthcare.^{11,12} Yet, while the promise of these technologies is real, so are the risks of their pre-mature adoption. The methodological complexity of AI systems demands careful scrutiny, but clinical publishing has not kept pace. Many journals lack both the technical infrastructure and editorial expertise required to evaluate these studies with the rigor they warrant.^{13,14}

As a researcher-clinician with dual expertise in both AI development and clinical psychiatry, I have observed firsthand the challenges posed by this gap. Two critiques I submitted to high-impact psychiatry journals – one challenging an AI model differentiating

suicidal from non-suicidal self-harm, the other critiquing a speech-based suicide risk detection system – were rejected not for inaccuracies in my evaluation, but for being “overly technical” or “lacking clinical relevance.”^{11–31} In one case, editorial processes allowed the original authors to pre-clear critiques, undermining the independence of peer review and suppressing substantive methodological discussion.¹²

These cases are not outliers. They reflect a deeper, systemic issue in how interdisciplinary research is handled in clinical publishing. Through these case studies, this perspective contributes to the ongoing discourse on peer review integrity by identifying structural editorial failures, analyzing their ethical and scientific implications, and proposing reforms to align publishing practices with the technical demands of AI-integrated mental health research.

2. The challenge of evaluating AI in clinical publishing

While transformative, AI and ML methods are not immune to significant flaws.^{10,13,14} Unlike conventional clinical research methods (e.g., randomized controlled trials, cohort studies, case-control studies, cross-sectional studies, case reports, and systematic reviews), AI-driven studies and studies using AI methods demand a nuanced understanding of data science principles, algorithmic transparency, model generalizability, and ethical implications.¹⁹ Peer reviewers and editors in clinical journals, who may not be versed in the complexities of computational models, can unintentionally overlook or misinterpret issues that would be immediately evident to AI specialists.¹³

3. Case study 1: Methodological limitations of Haghish (2025)

This challenge was starkly evident when I submitted some correspondence to a high-impact psychiatry journal regarding a 2025 study by Haghish, titled “Differentiating Adolescent Suicidal and Nonsuicidal Self-Harm with Artificial Intelligence.”¹¹ My critique focused on several key methodological concerns, including class imbalance, model interpretability, and generalizability, all essential to validate that AI models are both scientifically sound and clinically applicable.^{15–29}

Class imbalance is a pervasive problem in supervised ML, especially in sensitive domains, such as adolescent self-harm, where suicidal attempts constitute a small minority of the dataset.^{18,22,23,27,28} While Haghish employed synthetic oversampling techniques¹⁸ (specifically the synthetic minority oversampling technique, SMOTE), these methods – although well-intentioned – carry inherent risks. Oversampling can inflate minority class

representation artificially, leading to overfitting on synthetic samples that do not adequately represent real-world variation.^{18,23} This undermines model robustness and compromises generalizability across unseen populations and clinical settings. In my correspondence, I wrote:

Class imbalance remains one of the most significant challenges in supervised machine learning, particularly in domains, such as adolescent self-harm, where suicide attempts represent a small portion of the dataset. The synthetic oversampling techniques employed, while well-intentioned, may risk overfitting and undermine generalizability.

The clinical adoption of AI models hinges on transparent decision-making processes that clinicians can understand and trust. The original study lacked sufficient interpretability measures to explain how the model attributed importance to various features. I proposed integrating SHAP (SHapley Additive exPlanations) values to provide fine-grained, interpretable insights into feature contributions. SHAP values allow clinicians to see which factors most influenced the model’s predictions in individual cases, facilitating informed clinical judgment and improving acceptance in high-stakes settings.^{16,17} Specifically, I noted:

Integrating SHAP values could enhance the transparency of the model’s feature attribution, making the system more interpretable to clinicians and better suited for high-stakes environments.

Adolescents’ behavioral and clinical profiles vary widely across different populations and healthcare contexts. The study’s model was trained on a relatively homogeneous sample, limiting its applicability elsewhere. I suggested employing transfer learning techniques, which allow models to leverage knowledge from related datasets or tasks to improve performance on new, diverse cohorts.^{24,26} Transfer learning offers a path to improve model adaptability and external validity, a key requirement for any AI tool intended for broad clinical use:

Transfer learning could offer a viable path to improve generalizability, particularly across diverse clinical settings or populations not represented in the original training data.

The letter was ultimately rejected, with editorial feedback stating that these methodological concerns were “outside the journal’s thematic scope.”^{13,14} While editorial discretion is understandable, this dismissal raises deeper issues about how clinical journals vet AI-driven research. By sidelining fundamental questions about model rigor and applicability, the editorial board risks perpetuating the publication of AI studies that lack sufficient scientific and

ethical scrutiny.^{13,14,19-21} This episode exemplifies a recurring problem: clinical journals often lack the necessary expertise, infrastructure, or review frameworks to rigorously evaluate the technical complexities and ethical dimensions of AI and ML in mental health research. Without such mechanisms, flawed models with serious real-world consequences may be accepted uncritically.^{13,14,19-21}

4. Case study 2: Methodological oversights of Ding *et al.* (2025)

The systemic editorial failures seen in the Haghish case were not isolated. A study by Ding *et al.*,¹² titled “Speech-Based Suicide Risk Recognition for Crisis Intervention Hotlines Using Explainable Multi-task Learning,” innovatively applies multi-task learning (MTL) and explainable AI (XAI) to speech-based suicide risk detection in crisis hotline calls. Although innovative, several methodological choices warrant further scrutiny, particularly regarding speech pre-processing, feature extraction, model architecture, and multimodal integration.²⁸

A major concern is the removal of silences longer than 1 s from speech segments. Silences carry important emotional weight in high-stress contexts, indicating hesitation or distress, and their exclusion could lead to loss of critical psychological signals.²⁷ As stated in my letter:

*Silence in speech, particularly during high-stress crisis calls, can carry emotional weight; its removal may obscure indicators of hesitation, distress, or emotional regulation.*²⁵⁻²⁷

Excluding these silences risks discarding valuable psychological signals that are integral to accurately assessing caller emotional state.^{25,27} Transformer models with self-attention mechanisms (e.g., Wav2Vec 2.0) are better suited to capture such long-range dependencies without omitting silent intervals.^{24,28}

The authors also utilized a fixed 5-s segmentation window for feature extraction, which may be too rigid to capture the inherently non-linear and rapidly fluctuating emotional content in crisis speech.^{28,29} I argued:

The use of fixed 5-s segmentation windows may prevent the model from capturing the dynamic fluctuations typical in crisis speech patterns.

More flexible temporal modeling approaches, such as variable-length sequences handled by transformers with multi-head attention, or techniques, such as sliding windows and dynamic time warping, could better capture these rapid emotional transitions.²⁴⁻²⁶

Regarding feature extraction, the research team extracted 178 paralinguistic features but did not clearly

explain their prioritization or integration within the model. Feature interactions in speech emotion recognition are complex, non-linear, and context-dependent. I highlighted that:

*These technical issues have direct implications for suicide risk classification and cannot be dismissed as merely academic.*¹³⁻¹⁵

Advanced feature selection techniques, such as SHAP-based approaches or quantum-behaved particle swarm optimization (QPSO), have shown promise in refining discriminative feature sets to improve performance and interpretability, suggesting avenues for methodological improvement.^{17,23,25}

Regarding model architecture, while Bidirectional Long Short-Term Memory (Bi-LSTMs) capture some temporal dependencies, they have inherent limitations in modeling long-range context.^{24,26} Transformer-based architectures outperform Bi-LSTMs by leveraging multi-head attention and enabling more interpretable focus on critical speech segments.^{24,26} I emphasized:

“The limitations of Bi-LSTM architectures in capturing long-range emotional dependencies” diminish the model’s ability to detect subtle emotional variations over extended speech segments.

Recent pre-trained transformer speech models (e.g., HuBERT) further demonstrate robustness and efficiency in real-world noisy environments, making them preferable for this application.²⁸

Finally, the study’s exclusive reliance on speech features overlooks the benefits of multimodal integration. Combining speech with textual transcriptions or physiological data has been shown to improve emotion detection accuracy and model robustness, especially in complex, high-stakes environments, such as crisis hotlines.^{26,29,30} Despite the substantive and clinically relevant nature of these critiques, the editorial board rejected the letter as “overly technical” and “lacking clinical relevance.” This dismissal highlights a systemic editorial issue wherein rigorous methodological critique of AI models is marginalized, risking the publication of flawed models with potential real-world harms.^{19-21,30,31}

5. Editorial gatekeeping and conflicts of interest

The methodological shortcomings outlined in both case studies are not anomalous oversights but symptomatic of deeper structural failures in editorial practices governing AI in mental health research. A recurring theme in both rejections was that the letters were “overly technical” or

“outside scope” – feedback suggesting that editorial boards often defer scientific vetting to original authors, a process colloquially referred to as *pre-clearance*. While this practice may be intended to streamline correspondence handling, it effectively allows original authors to veto external critique, compromising the neutrality and independence of peer review.³²⁻³⁸

This gatekeeping is further exacerbated by a systemic lack of technical and ethical expertise among clinical journal editors to assess AI-related submissions. As ML models become more complex and deeply integrated into healthcare, editorial boards must be equipped to evaluate not only clinical relevance but also algorithmic validity, interpretability, and fairness.^{30,31} Without such expertise, editorial decisions may inadvertently privilege esthetic novelty or positive results over scientific rigor and replicability.

In many journals, the peer review process itself remains opaque and insufficiently diverse, further contributing to biased publication outcomes. Studies show that increasing gender and international diversity among reviewers correlates with fairer evaluations and higher-quality editorial outcomes.³³⁻³⁵ Yet, even in journals that acknowledge these disparities, few have adopted concrete reforms, such as blind review, reviewer training in AI ethics, or structured checklists for evaluating ML studies.^{19,30,39-45}

As generative AI continues to scale across clinical domains, scholars have increasingly called for the integration of *embedded ethics* into the development, evaluation, and dissemination of medical AI research.³⁹⁻⁴⁵ This approach demands that ethical concerns – such as algorithmic bias, safety, transparency, and explainability – be addressed from the outset, not appended *post hoc*. In this model, ethics is not a checkpoint at the end of the pipeline but a structural element of rigorous scientific inquiry.

Despite these calls, the editorial handling of the critiques toward the works of Haghish¹¹ and Ding *et al.*¹² suggests that present publishing norms fall short of AI-driven studies and studies using AI methods. The absence of substantive engagement with these challenges implies that many journals remain ill-equipped – or unwilling – to enforce ethical scrutiny as part of peer review. Without meaningful reform in areas, such as editorial independence, reviewer training, and conflict-of-interest transparency, flawed AI models may continue to bypass critical evaluation and enter the clinical literature unchallenged.

This failure is not merely procedural. It raises foundational questions about epistemic authority in clinical AI:

- Who determines what constitutes valid evidence?
- Who is accountable when predictive models reinforce structural bias or contribute to diagnostic error?

In the absence of systemic safeguards, the pre-mature adoption of under-evaluated AI tools threatens not just the integrity of the scientific record but the safety and equity of patient care.^{30,31,39,40}

6. Comparative analysis across cases

The rejection of substantive methodological critiques in both the Haghish¹¹ and Ding *et al.*¹² case studies reveals consistent patterns of editorial gatekeeping, technical exclusion, and ethical under-evaluation. While the studies addressed different domains (*i.e.*, text-based versus speech-based suicide prediction), the nature of the overlooked issues and the editorial rationale for rejection were strikingly similar. These cases demonstrate that systemic editorial deficiencies can transcend methodological domain, modality, and even discipline.

Table 1 summarizes the critical methodological concerns raised in each case, mapping them to potential clinical consequences and corresponding editorial responses. This side-by-side view makes visible the shared vulnerabilities in AI health research publication and underscores the urgency for reform in peer review protocols.

Figure 1 shows a conceptual model depicting the multi-layered nature of editorial gatekeeping and its consequences. Critique pre-clearance, limited AI literacy, and narrow definitions of clinical relevance combine to create significant obstacles. Together, these factors build barriers that obstruct scientific accountability.

Together, these cases illustrate a systemic breakdown in editorial accountability. When valid methodological critiques are filtered out by opaque editorial practices or vetoed by original authors, the epistemic integrity of the scientific record is compromised. Moreover, the publication of inadequately vetted AI models has serious clinical and ethical implications.

7. Limitations

In discussing the constraints within editorial decision-making, I recognize several key factors that shape how scholarly work and professional discourse are disseminated. Journals operate within specific editorial frameworks that dictate what content is selected for publication. These policies may prioritize particular research methodologies or thematic focuses, inadvertently shaping which perspectives enter the broader academic conversation.^{20,21,32-38}

The peer-review process, while intended to ensure rigor and credibility, is subject to variability in reviewer expertise,

Table 1. Summary of critiques and editorial responses across two case studies

Dimension	Case study 1 (Haghighi, 2025) ¹¹	Case study 2 (Ding <i>et al.</i> , 2025) ¹²
Data imbalance	Overreliance on SMOTE without evaluation of generalizability.	Not applicable.
Model interpretability	Lacked SHAP-based or interpretable mechanisms.	Feature attribution unclear; 178 features used without ranking.
Generalizability	Single-site data; no external validation or transfer learning.	No discussion of generalizability beyond one speech corpus.
Temporal modeling	Not applicable.	Fixed 5-s windows insufficient for dynamic emotional variance.
Silence removal	Not applicable.	Silences removed, obscuring emotional/psychological cues.
Model architecture	Not discussed.	Bi-LSTM used despite limitation; transformers not explored.
Multimodal design	Not applicable.	Speech-only; no integration of text, physiology, or behavioral context.
Proposed improvement	SHAP, transfer learning, balanced evaluation.	Transformer models, QPSO, SHAP, multimodal fusion.
Potential consequence	Misclassification in adolescent self-harm; clinical misapplication.	Missed suicide risk signals; failure in real-world crisis detection.
Editorial justification	“Outside scope.”	“Overly technical.”

Abbreviations: Bi-LSTM: Bidirectional Long Short-Term Memory; QPSO: Quantum-behaved particle swarm optimization; SHAP: SHapley Additive exPlanations; SMOTE: Synthetic minority over-sampling technique.

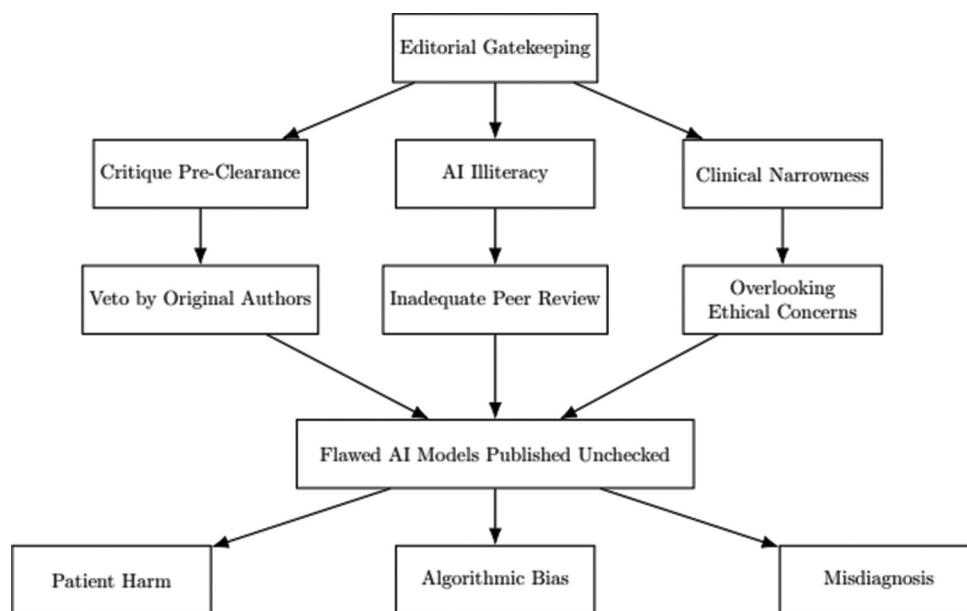


Figure 1. Editorial gatekeeping in artificial intelligence health research

implicit biases, and institutional priorities that may limit the diversity of published viewpoints.³²⁻³⁸ Because of space limitations and journal formatting constraints, the scope of arguments permissible within letters to the editor is frequently restricted; moreover, although such letters serve as a platform for academic discourse, their acceptance remains contingent upon editorial discretion and alignment with the journal’s thematic priorities.^{32,37} The visibility of alternative frameworks within scholarly publishing is influenced by citation networks, funding availability, and institutional affiliations, affecting the accessibility of critical perspectives outside dominant paradigms.^{19,32-38}

Journals, including digital-only platforms, must often balance the volume of valid commentaries they receive against practical considerations, such as editorial resources and thematic coherence, making it unrealistic to publish all submissions regardless of their merit.³² Furthermore, the sensitive nature of mental health data imposes significant privacy constraints that restrict the open sharing of patient-level information. Ethical and legal obligations to protect participant confidentiality limit access to raw datasets, which complicates reproducibility and external validation efforts – challenges well documented in AI healthcare research.³⁹⁻⁴³ These factors underscore the need for adaptive

publication policies and innovative data governance frameworks that balance scientific transparency with the ethical imperatives unique to this field.

8. Calls for reform: Elevating the standards of peer review

To ensure the safe and effective integration of AI into clinical practice, scientific publishing – especially in clinical journals – must reform its approach to reviewing AI and ML research. To that end, the following recommendations are proposed:

- *Expert reviewers for AI methodologies:* Journals should engage data science and AI experts to identify technical flaws and verify the reproducibility, transparency, and robustness of the models.
- *Transparent model evaluation:* Manuscripts must provide explicit details regarding model training, data handling, and algorithm performance while addressing issues, such as class imbalance, bias, and interpretability.
- *Encouraging open data and code:* To facilitate reproducibility, journals should promote the sharing of data and code, enabling independent verification and improvement of AI models.
- *Dedicated spaces for AI methodological critiques:* Creating sections devoted to methodological discussion can encourage healthy academic discourse and improve the quality of published research.
- *Ethical and clinical considerations:* All AI-driven studies should include mandatory sections on ethics—analyzing informed consent, privacy, and potential harm—to ensure safe and responsible applications in clinical settings.

9. Final thoughts: Upholding scientific rigor and ethical standards

As AI continues to permeate healthcare, the imperative for rigorous, methodologically sound research grows ever more urgent. Inaccurate or insufficiently validated AI models risk fatal errors—misclassifying suicide risk, withholding necessary care, or prompting harmful interventions. These are not abstract concerns; they are life-or-death consequences of editorial decisions made today.

Clinical journals serve as critical gatekeepers of scientific integrity, and they must adapt to the challenges posed by the complexity and novelty of AI-driven methodologies. Only through independent, transparent, and technically informed peer review can the scientific community ensure that AI tools are deployed ethically, effectively, and safely in clinical settings. By embracing robust methodological critique rather than dismissing it as “overly technical,”

journals protect not only the rigor of science but also the well-being of vulnerable patients.

Looking ahead, future research should explore interdisciplinary innovations to enhance the robustness, interpretability, and clinical utility of AI models in mental and medical health. Emerging computational frameworks⁴⁶⁻⁴⁹ – such as those based on circular bipolar complex intuitionistic fuzzy linguistic information, Frank power aggregation operators, and MABAC models – have demonstrated success in fields, such as renewable energy analysis and wireless communications. In addition, approaches employing neuro-fuzzy, complex propositional picture fuzzy Sugeno–Weber power aggregation and fractal mathematics, including superior Mandelbrot sets, offer promising avenues for managing uncertainty and improving model transparency.⁵⁰⁻⁵⁷ While these advanced techniques have yet to be widely applied in mental or medical health AI, their adaptation holds potential to address critical methodological challenges, including class imbalance, model interpretability, and generalizability. Integrating such innovations could complement editorial reforms, pushing the field toward more reliable, ethical, and clinically impactful AI and ML applications.

This perspective highlights systemic failures in editorial oversight and offers concrete recommendations to reform peer review processes – reforms essential to maintaining trust in both AI research and its real-world applications. Without such change, the promise of AI risks becoming overshadowed by preventable harm and eroded confidence. Addressing these challenges is not optional; it is a critical responsibility that the scientific community and clinical publishers must urgently embrace to protect both patients and the integrity of mental health research.

Acknowledgments

I extend my heartfelt gratitude to Sean Harty, my first mentor and a lifelong friend. I met Sean in 1997, and his immediate recognition of my aptitude for computer programming and network design changed the trajectory of my life. Over the years, I quite literally followed him across three different companies—a testament to his exceptional technical skill, leadership, and integrity. Sean was not only a guiding force in my early IT and network engineering career but also a constant sounding board, always offering his time, wisdom, and a whiteboard to help map through complex ideas. His mentorship went far beyond professional development; it laid the groundwork for how I think, problem-solve, and lead. Now approaching 30 years of service as Chief Information Officer in local state government, Sean remains one of the most influential figures in my life and a treasured friend of 28 years.

Funding

None.

Conflict of interest

The author has previously submitted critiques to the *Journal of Affective Disorders* regarding AI methodologies in clinical research, which were not accepted for publication. While this perspective discusses editorial practices in AI research – including *Journal of Affective Disorders* – the analysis is conducted independently, without financial or institutional influence. The views expressed reflect methodological and ethical concerns relevant to AI-driven mental health research and do not stem from any personal, professional, or financial stake in the journal or related entities.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The original letters to the editors generated and analyzed in this expert perspective article are available upon request of the corresponding author. The editorial and peer responses are withheld due to editorial policy.

References

1. Umar BU, Ajao LA, Dogo EM, Ajao FJ, Atama M. Artificial intelligence model for prediction of cardiovascular disease: An empirical study. *Artif Intell Health*. 2023;1(1):42-56. doi: 10.36922/aih.1746
2. Nawab K. Artificial intelligence scribe: A new era in medical documentation. *Artif Intell Health*. 2024;1(4):12-15. doi: 10.36922/aih.3103
3. 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*. 2024;2(1):105-113. doi: 10.36922/aih.3750
4. Wicklem LC, San Hwang S, Lau BT, Bhave M, Chee XW. Machine learning-driven prediction of EBNA1 inhibitors against Epstein-Barr virus in nasopharyngeal carcinoma. *Artif Intell Health*. 2024;2(1):93-104. doi: 10.36922/aih.4375
5. 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 a network-centric digital health stack for future pandemics. *Artif Intell Health*. 2024;2(1):29-41. doi: 10.36922/aih.2888
6. Casiraghi JL, Lizio A, Bolognini S, *et al*. Exploring the viability of robotic technology integrated with Vivaldi artificial intelligence for functional assessment in amyotrophic lateral sclerosis. *Artif Intell Health*. 2024;1(4):73-84. doi: 10.36922/aih.3732
7. Schwingel PA, Schwingel D, De Aquino SR, *et al*. An exploratory study on the potential of ChatGPT as an AI-assisted diagnostic tool for visceral leishmaniasis. *Artif Intell Health*. 2024;1(4):97-106. doi: 10.36922/aih.3930
8. Luu MSK, Tuchinov BN, Prokaeva AI, Korobko DS, Malkova NA, Tulupov AA. Discovering predictive features of multiple sclerosis from clinically isolated syndrome with machine learning. *Artif Intell Health*. 2024;1(4):107-122. doi: 10.36922/aih.4255
9. Thomas C, Prasad RR. Health-care app detection using optimized clustering. *Artif Intell Health*. 2024;1(4):16-29. doi: 10.36922/aih.2585
10. Vishwanath AB, Srinivasalu VK, Subramaniam N. Role of large language models in improving provider-patient experience and interaction efficiency: A scoping review. *Artif Intell Health*. 2024;2(2):1-10. doi: 10.36922/aih.4808
11. Haghish EF. Differentiating adolescent suicidal and nonsuicidal self-harm with artificial intelligence: Beyond suicidal intent and capability for suicide. *J Affect Disord*. 2025;378:381-391. doi: 10.1016/j.jad.2025.02.015
12. Ding Z, Zhou Y, Dai AJ, *et al*. Speech based suicide risk recognition for crisis intervention hotlines using explainable multi-task learning. *J Affect Disord*. 2025;370:392-400. doi: 10.1016/j.jad.2024.11.022
13. Gulumbe BH. Obvious artificial intelligence-generated anomalies in published journal articles: A call for enhanced editorial diligence. *Learn Publ*. 2024;37(4):1-5. doi: 10.1002/leap.1626
14. Leveridge M. This editorial about AI in publishing was definitely written by a human. *Can Urol Assoc J*. 2023;17(6):151-152. doi: 10.5489/cuaj.8424
15. Tonmoy STI. Embeddings at BLP-2023 Task 2: Optimizing fine-tuned transformers with cost-sensitive learning for

- multiclass sentiment analysis. In: *Proceedings of the First Workshop on Bangla Language Processing*. 2023. p. 340-346.
16. Qi X, Wang S, Fang C, Jia J, Lin L, Yuan T. Machine learning and SHAP value interpretation for predicting comorbidity of cardiovascular disease and cancer with dietary antioxidants. *Redox Biol.* 2025;79:103470.
doi: 10.1016/j.redox.2024.103470
17. Ahmed U, Jiangbin Z, Almogren A, *et al.* Hybrid bagging and boosting with SHAP based feature selection for enhanced predictive modeling in intrusion detection systems. *Sci Rep.* 2024;14:30532.
doi: 10.1038/s41598-024-81151-1
18. Alkhalwaldeh IM, Albalkhi I, Naswhan AJ. Challenges and limitations of synthetic minority oversampling techniques in machine learning. *World J Methodol.* 2023;13(5):373-378.
doi: 10.5662/wjm.v13.i5.373
19. Ganjavi C, Eppler MB, Pekcan A, *et al.* Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: Bibliometric analysis. *BMJ.* 2024;384:e077192.
doi: 10.1136/bmj-2023-077192
20. Jackson P, Weaver CC, editors. *Rethinking Peer Review: Critical Reflections on a Pedagogical Practice*. Colorado: WAC Clearinghouse; 2023.
21. Smith R. Peer review: A flawed process at the heart of science and journals. *J R Soc Med.* 2006;99(4):178-182.
doi: 10.1177/014107680609900414
22. Joloudari JH, Marefat A, Nematollahi MA, Oyelere SS, Hussain S. Effective class-imbalance learning based on SMOTE and convolutional neural networks. *Appl Sci.* 2023;13(6):4006.
doi: 10.3390/app13064006
23. Araf I, Idri A, Chairi I. Cost-sensitive learning for imbalanced medical data: A review. *Artif Intell Rev.* 2024;57:80.
doi: 10.1007/s10462-023-10652-8
24. Baevski A, Zhou F, Mohamed AR. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. In: *Proceedings of NeurIPS*. 2020.
25. Daneshfar F, Kabudian SJ. Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimed Tools Appl.* 2020;79(1):1261-1289.
doi: 10.1007/s11042-019-08222-8
26. Das R, Singh TD. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput Surv.* 2023;55(13s):1-38.
doi: 10.1145/3586075
27. Pfeifer E, Wittmann M. Waiting, thinking, and feeling: Variations in the perception of time during silence. *Front Psychol.* 2020;11:602.
doi: 10.3389/fpsyg.2020.00602
28. Sharrab YO, Attar H, Eljini MAH, Al-Omary Y, Al-Momani W. Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions. *IEEE Access.* 2025;13:46925.
doi: 10.1109/ACCESS.2025.3550855
29. Vinciarelli A, Pantic M, Bourlard H. Social signal processing: Survey of an emerging domain. *Image Vision Comput.* 2009;27(12):1743-1759.
doi: 10.1016/j.imavis.2008.11.007
30. Choi J, Liu Z, Xu P. Advancing artificial intelligence in mental health: Ethical considerations in machine learning applications. *J Affect Disord.* 2024;379:302-314.
doi: 10.1016/j.jad.2024.03.017
31. Gong Y. Navigating the future of clinical informatics: A focus on patient safety, AI integration, and education. *J Clin Inform.* 2024;1(1):4988.
doi: 10.36922/jci.4988
32. Merriman B. Peer review as an evolving response to organizational constraint: Evidence from sociology journals, 1952-2018. *Am Soc.* 2021;52(3):341-366.
doi: 10.1007/s12108-020-09473-x
33. Murray D, Siler K, Larivière V, *et al.* *Gender and International Diversity Improves Equity in Peer Review*. *bioRxiv* [Preprint]; 2018.
doi: 10.1101/400515
34. Lerback JC, Hanson B, Wooden P. Association between author diversity and acceptance rates and citations in peer-reviewed Earth science manuscripts. *Earth Space Sci.* 2020;7(5):e2019EA000946.
doi: 10.1029/2019EA000946
35. Callahan M. Gender bias and peer review: Annals seeks greater diversity. *Ann Emerg Med.* 2019;74(6):736-739.
doi: 10.1016/j.annemergmed.2019.08.432
36. Drozd JA, Lodomery MR. The peer review process: Past, present, and future. *Br J Biomed Sci.* 2024;81:12054.
doi: 10.3389/bjbs.2024.12054
37. Khalifa AA. A letter to the editor: The peer review process: Past, present, and future. *Br J Biomed Sci.* 2025;81:14125.
doi: 10.3389/bjbs.2024.14125
38. Siau K, El-Omar E. How to write a letter to the editor. *United European Gastroenterol J.* 2020;8(8):981-983.
doi: 10.1177/2050640620956921
39. Rieke N, Hancox J, Li W, *et al.* The future of digital health

- with federated learning. *NPJ Digit Med.* 2020;3(1):119.
doi: 10.1038/s41746-020-00323-1
40. Gallifant J, Fiske A, Levites Strekalova YA, *et al.* Peer review of GPT-4 technical report and systems card. *PLOS Digit Health.* 2024;3(1):e0000417.
doi: 10.1371/journal.pdig.0000417
41. McLennan S, Fiske A, Tigard D, Müller R, Haddadin S, Buyx A. Embedded ethics: A proposal for integrating ethics into the development of medical AI. *BMC Med Ethics.* 2022;23(1):6.
doi: 10.1186/s12910-022-00746-3
42. McLennan S, Fiske A, Celi LA, *et al.* An embedded ethics approach for AI development. *Nat Mach Intell.* 2020;2(9):488-490.
doi: 10.1038/s42256-020-0214-1
43. Fiske A, Prainsack B, Buyx A. Data work: Meaning-making in the era of data-rich medicine. *J Med Internet Res.* 2019;21(7):e11672.
doi: 10.2196/11672
44. Fiske A, Buyx A, Prainsack B. Health information counselors: A new profession for the age of big data. *Acad Med.* 2019;94(1):37-41.
doi: 10.1097/ACM.0000000000002395
45. McLennan S, Lee MM, Fiske A, Celi LA. AI ethics is not a panacea. *Am J Bioeth.* 2020;20(11):20-2.
doi: 10.1080/15265161.2020.1819470
46. Ali Z, Yang MS. Analysis of renewable energies based on circular bipolar complex intuitionistic fuzzy linguistic information with Frank power aggregation operators and MABAC model. *Int J Comput Intell Syst.* 2025;18(1):1-40.
doi: 10.1007/s44196-025-00800-z
47. Tanveer M, Sajid M, Akhtar M, *et al.* Fuzzy deep learning for the diagnosis of Alzheimer's disease: Approaches and challenges. *IEEE Trans Fuzzy Syst.* 2024;32(10):5477-5492.
doi: 10.1109/TFUZZ.2024.3409412
48. Apostolopoulos ID, Papandrianos NI, Papathanasiou ND, Papageorgiou EI. Fuzzy cognitive map applications in medicine over the last two decades: A review study. *Bioengineering (Basel).* 2024;11(2):139.
doi: 10.3390/bioengineering11020139
49. Rafeisangari R, Shiri N. A multi-adaptive neuro-fuzzy inference system with variable thresholds for heartbeat classification. *Artif Intell Health.* 2024;1(4):43-60.
doi: 10.36922/aih.3367
50. Alrasheedi AF, Mishra AR, Alshamrani AM, Rani P, Pamucar D. Decision-analytics-based risk allocation in the micromobility sector: Sugeno-Weber operators and picture fuzzy distance methodology. *Cognit Comput.* 2024;16(6):3122-3148.
doi: 10.1007/s12559-024-10333-w
51. Mandelbrot BB, Evertsz CJ, Gutzwiller MC. *Fractals and Chaos: The Mandelbrot Set and Beyond.* Vol. 3. Berlin: Springer; 2004.
52. Devaney R. *An Introduction to Chaotic Dynamical Systems.* United States: CRC Press; 2018.
53. Devaney RL. *A First Course in Chaotic Dynamical Systems: Theory and Experiment.* United States: CRC Press; 2018.
54. Flake GW. *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems, and Adaptation.* United States: MIT Press; 2000.
55. Barnsley MF. *Fractals Everywhere.* United States: Academic Press; 2014.
56. Schroeder M. *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise.* United States: Courier Corporation; 2009.
57. Falconer K. *Fractal Geometry: Mathematical Foundations and Applications.* United States: John Wiley and Sons; 2013.

ORIGINAL RESEARCH ARTICLE

Accurate early detection of Parkinson's disease from single photon emission computed tomography imaging through convolutional neural networks

R. Prashanth* 

Independent Researcher, Bengaluru, Karnataka, India

Abstract

Early and accurate detection of Parkinson's disease (PD) remains a crucial diagnostic challenge with substantial clinical implications, particularly for ensuring effective treatment and patient management. For instance, a group of subjects with scans without evidence of dopaminergic deficit (SWEDD) who are initially diagnosed as PD but exhibit normal single photon emission computed tomography (SPECT) scans. Over time, follow-up assessments often lead to a revised diagnosis of non-PD. In the meantime, these subjects may receive PD-specific medications that can cause more harm than benefit. In this paper, a case study is presented in which machine learning models are developed and trained on SPECT images to distinguish early PD from healthy controls, as well as to differentiate SWEDD cases from early PD. The case study utilizes a well-known, publicly available dataset and explores several machine learning classifiers, including support vector machines, logistic regression, feed forward neural networks, and convolutional neural networks (CNNs). The CNN model gave the best performance in differentiating PD from healthy subjects. All these models demonstrated strong potential for early differentiation of SWEDD cases from PD. These results suggest that the proposed approach could support clinicians in making more accurate and timely diagnostic decisions.

Keywords: Computer-aided diagnosis; Machine learning; Deep learning; Parkinson's disease; Medical imaging

***Corresponding author:**R. Prashanth
(prashanth.r.iitd@gmail.com)

Citation: Prashanth R. Accurate early detection of Parkinson's disease from single photon emission computed tomography imaging through convolutional neural networks. *Artif Intell Health*. 2025;2(4):22-32.
doi: 10.36922/AIH025040005

Received: January 21, 2025**1st revised:** May 13, 2025**2nd revised:** May 22, 2025**Accepted:** May 30, 2025**Published online:** June 17, 2025

Copyright: © 2025 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

Parkinson's disease (PD) is a progressive neurodegenerative disorder affecting millions of people worldwide and is characterized by the loss of dopaminergic neurons in the substantia nigra.^{1,2} Its prevalence increases with age, impacting approximately 1% of individuals over 60 years.³ The clinical diagnosis of PD is challenging as there are no definitive diagnostic tests and the diagnosis is based on the presence of cardinal symptoms, such as tremor at rest, rigidity, and bradykinesia, along with a subject's response to PD medications.¹ However, these symptoms appear in the later stages of the disease and by the time the patient manifests these symptoms, the patient might have already crossed the early stage of the disease.⁴ Early detection of PD is important because appropriate targeted therapies could be initiated before any major deterioration occurs.⁵

It can also help develop treatments and identify patients eligible for therapeutic clinical trials.⁵

Single photon emission computed tomography (SPECT) imaging using ¹²³I-Ioflupane (DaTSCAN or [123I]FP-CIT) has been shown to increase the diagnostic accuracy of PD, mainly in the earlier stages of the disease, by showing the functional deterioration or dopaminergic deficit in the striatal region of the brain (which is one of the primary regions affected in PD).⁶⁻⁹ The accuracy of diagnosis of PD at an early phase is the poorest based on clinical indices as early symptoms are mild/moderate, unlike in advanced stages of the disease.^{4,5} Furthermore, these symptoms are common in other neurodegenerative disorders like essential tremor and multiple system atrophy, which often leads to misdiagnosis.¹⁰⁻¹² The effects of misdiagnosis are severe as it may lead to unnecessary medical examinations and therapies, and associated side-effects. Recent studies have shown that around 3.6 – 19.6% of clinically diagnosed PD subjects show no dopaminergic deficit, and these subjects are classified as scans without evidence of dopaminergic deficit (SWEDD).¹⁰⁻¹² Subsequent follow-up on these subjects have shown that they neither deteriorate nor respond to levodopa (a primary medication in PD) and that their SPECT scans remain normal in the follow-up imaging. Thus, these subjects were considered highly unlikely of having PD and that the initial diagnosis of PD was incorrect.¹³⁻¹⁵ These studies evidently point out that dopaminergic imaging is highly useful and that an abnormal imaging, at least in cases of diagnostic uncertainty, is strongly supportive of a diagnosis of neurodegenerative Parkinsonism (PS), such as PD.

In clinical practice, SPECT images are usually analyzed by visual inspection and/or by region of interest (ROI) analysis.¹⁶ The visual analysis relies on the judgment of the observer that heavily depends on his expertise, experience, and knowledge.¹⁷ ROI techniques involve outlining or positioning the ROI over the striatum (target region) and the occipital cortex (reference region), and a quantitative measure termed the background subtracted striatal uptake ratio is computed.⁶ Despite the odds, the latter method or the quantitative method is the most acceptable one, since, according to a trial study, it provides an excellent intra- and inter-observer agreement.¹⁸ However, the ROI-based approach relies on manual intervention for placing the ROIs.

There have been many studies that make use of machine learning techniques to develop predictive models from SPECT imaging features for the early detection of PD.^{11,19-31} Segovia *et al.*²⁸ extracted voxels corresponding to the striatum and performed data decomposition

using partial least squares followed by classification into controls and Parkinsonism by means of a support vector machine (SVM) classifier. Illan *et al.*²⁹ also used voxels corresponding to the striatum to train an SVM classifier with a linear kernel to classify controls and PS. Rojas *et al.*³⁰ used voxels corresponding to the striatum and then carried out feature reduction through principal component analysis followed by classification using SVM. Towey *et al.*³¹ performed feature extraction on all voxels through singular value decomposition followed by classification into PS or non-PS. Huertas-Fernández *et al.*³² calculated the bilateral caudate and putamen uptake and asymmetry indices from SPECT images and developed predictive models using logistic regression, SVM, and LDA to classify PD from vascular Parkinsonism. Kim *et al.*²⁰ used image augmentation to increase the size of data and a classifier based on the Inception v3 model that can classify normal from abnormal SPECT scans.

There are also many studies using the SPECT data from the Parkinson's progression marker initiative (PPMI), which is among the most popular, widely used, and largest database for PD,^{11,19,21-27,33,34,45-48} and the same data were used in the present study. Choi *et al.*¹¹ trained a convolutional neural network (CNN), which they called PD net, using SPECT images to classify PD from normal and non-Parkinsonism tremor. They also used the model to classify SWEDD subjects. In their analysis, they used the complete volume data, rather than considering a selected range of slices, due to which the CNN network became complex with many layers. Martínez-Murcia *et al.*²² also used a CNN to differentiate PD from others (healthy normal and SWEDD). They used a threshold-based approach to select sub-volumes from the volume which they later input to the CNN. They observed that due to this sub-volume selection, the complexity of the CNN became small with just two convolutional layers.²² Martínez-Murcia *et al.*²¹ used the features extracted from SPECT images through independent component analysis to train an SVM classifier to distinguish PD from normal. They observed much better performance than their previous work using the voxel-as-features approach. Hirschauer *et al.*¹⁹ used data from different clinical examinations and SPECT imaging, and trained an enhanced probabilistic neural network model to differentiate PD from SWEDD. Oliveira and Castelo-Branco²³ used voxels as features that were extracted based on volumes of interest defined (which required manual intervention), and an SVM classifier was used to classify PD from normal. The standard binding potential features along with other features related to the volume and length of the striatal region from SPECT images was utilized to train an SVM classifier that could classify PD from healthy normal.³³ Ortiz *et al.*²⁴ extracted features from isosurfaces

computed from the ROI and trained a CNN-based model to classify PD from healthy normal. Prashanth *et al.*²⁷ computed shape- and surface-fitting-based features and used machine learning methods to develop classification models to differentiate scans with deficit, as in PD, from scans without deficit, as in normal and SWEDD. Prashanth *et al.*²⁶ also used data from multiple modalities including clinical examinations, laboratory examinations, and dopaminergic imaging, and developed classification models to distinguish early PD from normal. The same researcher group had used the striatal binding ratios to develop classification and prognostic models for PD.²⁵ Zhang *et al.*³⁴ employed multimodal data which included SPECT imaging data to identify different PD subtypes through the long-short term memory (LSTM) networks and dynamic time warping. Shiiba *et al.*⁴⁵ extracted radiomics features including intensity- and texture-based features in the caudate, putamen, and pallidum volumes of interest from the SPECT images and used machine learning methods to classify PD from normal. Tufail *et al.*⁴⁶ developed a 3D CNN model (consisting of 14 layers including 5 convolution, 5 max pooling, and 3 fully connected layers) that is capable of performing multiclass classification of Alzheimer's and PDs using positron emission tomography and SPECT neuroimaging modalities. Majhi *et al.*⁴⁷ used magnetic resonance imaging and SPECT imaging data to train many deep learning models including VGG16, DenseNet, DenseNet-LSTM, and InceptionV3 that are optimized through gray wolf optimization. Khachnaoui *et al.*⁴⁸ trained deep learning models based on EfficientNet-B0, Mobilenet-V2, and a custom CNN with 10 layers (4 convolutional, 4 max pooling) using SPECT images.

However, several limitations that persist in prior work are as follows:

- Use of entire image volumes, increasing model complexity and the risk of overfitting;
- Dependence on explicit feature extraction pipelines followed by machine learning classifiers;
- Manual intervention for ROI placement, reducing reproducibility and scalability;
- Training on small subject cohorts, limiting generalizability;
- Focus solely on binary classification (*e.g.*, PD vs. healthy), with limited attention to diagnostically challenging cases such as SWEDD.

The present study addresses the above limitations by developing a compact CNN-based model optimized through Bayesian hyperparameter optimization to distinguish early PD from healthy controls, as well as to differentiate diagnostically challenging SWEDD cases from those with early PD. Unlike prior approaches that

relied on full volumes or manual feature engineering, this work uses only the most relevant slice(s) from SPECT volumes thereby reducing model complexity and the risk of overfitting. The model was trained and evaluated on data from the PPMI, one of the most extensive and standardized PD imaging databases available. By incorporating SWEDD into the classification task, this work contributes toward differential diagnosis within the Parkinsonian spectrum. The proposed method combines the strengths of automated feature learning with informed slice selection, enabling improved diagnostic accuracy and practical utility for early-stage PD and SWEDD detection.

2. Materials and methods

2.1. Dataset details

The data used in the study were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (<http://www.ppmi-info.org/data>). For up-to-date information, please visit <http://www.ppmi-info.org>. The PPMI is a landmark, large-scale, comprehensive, observational, international, multi-center study that recruits *de novo* (early-untreated) PD patients, and age- and gender-matched healthy subjects to identify PD progression biomarkers.^{4,35}

In this work, SPECT imaging data from the screening visits of 209 healthy normals, 443 early PD, and, 80 SWEDD were used. All the subjects in the three groups are age- and gender-matched to minimize demographic bias. Table 1 shows the age, gender, and Hoehn and Yahr (HY) stage distribution for the three groups. All PD patients were in their early stage (HY stage 1 or 2 with mean \pm standard deviation as 1.50 ± 0.50 ³⁶) and all SWEDD subjects exhibited early-stage PD symptoms (HY stage as 1.46 ± 0.53).

2.2. Image pre-processing by PPMI

All the SPECT scans taken at different PPMI sites undergo a standard pre-processing procedure before they are publicly shared through the database.³⁵ This pre-processing was carried out so that all scans were in the same anatomical alignment (spatially normalized). The process includes reconstruction from raw projection data and attenuation correction, followed by application of a standard Gaussian 3D 6.0 mm filter and then normalization of these images to standard Montreal Neurologic Institute space. These pre-processed scans, which were then shared for public access, were used for this analysis. The analysis pipeline is shown in Figure 1.

2.3. Slice selection

Each SPECT scan consists of 91 transaxial slices (from bottom to top of the head) each of size 109×91 , which

Table 1. Details of the subjects in terms of age, gender, and the HY stage

Gender	Normal		Early PD			SWEDD		
	Count	Age (mean)	Count	Age (mean)	HY stage	Count	Age (mean)	HY stage
Female	73	59.32	157	60.91	1.46±0.50	30	58.16	1.4±0.50
Male	136	61.65	286	62.13	1.53±0.50	50	61.80	1.5±0.54
All	209	60.79	443	61.7	1.51±0.50	80	60.43	1.46±0.53

Note: HY stands for Hoehn and Yahr stage.

Abbreviations: PD: Parkinson's disease; SWEDD: Scans without evidence of dopaminergic deficit.

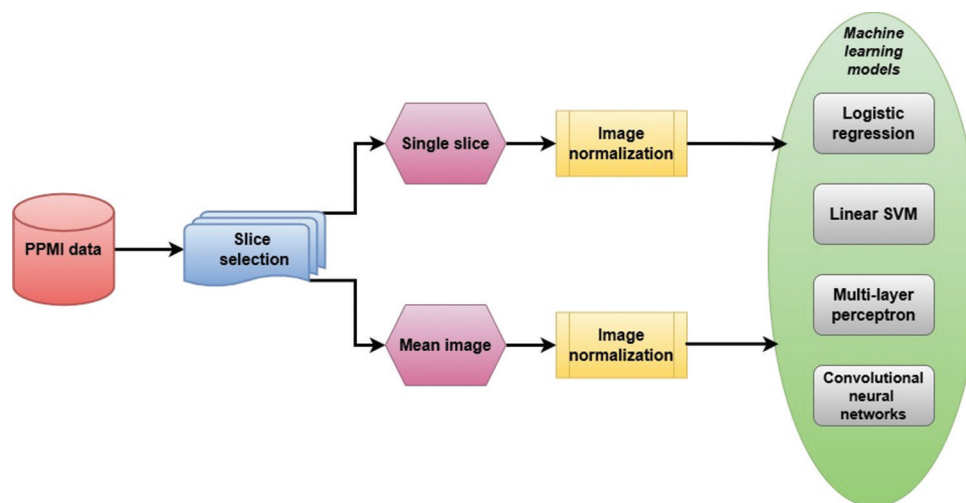


Figure 1. Flowchart of the analysis

Abbreviations: PPMI: Parkinson's Progression Markers Initiative; SVM: Support vector machine.

means each scan is of 3D type with size $91 \times 109 \times 91$. In published literature,²⁷ the areas of striatal activity from SPECT images were analyzed and it was observed that the most relevant striatal activity came from slices 35 to 48, with the highest activity occurring in slice number 41.

In this work, two types of images were used for the analysis.

- Single slice: It is the 41st slice extracted from the SPECT volume, as this is the slice with maximum striatal uptake, making it very relevant for PD detection.
- Mean image: It is the average of slices from 35 to 48 extracted from the SPECT volume as these are the slices that show striatal activity.

Figure 2 shows both the single slice and mean image for the three groups: Normal control, early PD, and SWEDD. Normal scans are characterized by intense, uniform, and symmetric high-intensity regions (corresponding to the caudate and striatum) on both hemispheres that appear as two comma-shaped regions, as evident in Figure 2A and 2C. In PD, dopaminergic neuron deterioration leads to a reduction in the comma-shaped region, which becomes smaller and more circular in shape, as observed in Figure 2B.

2.4. Image normalization

The intensities in the original SPECT image ranged from 0 to $2^{15}-1$. To standardize the data, the selected images (both single slice as well as the mean image) were normalized by dividing the intensity values by $2^{15}-1$, so that the normalized intensity is in the range $[0 - 1]$.

2.5. Data partitioning

Data were divided into two parts, namely, Partition 1 and Partition 2, in the ratio of about 80:20. Partition 1 was used for model training and evaluation using an approach based on cross-validation (10-fold). That is, Partition 1 data were split into 10 folds and then one of the folds became the evaluation set, and the remaining nine folds were used for training the model, with the whole process repeated nine times such that every fold became a test set and the remaining nine folds became the training data. Partition 2 was exclusively used for hyperparameter tuning of the machine learning methods and was not involved in model training or evaluation. All reported performance measures were solely based on the cross-validation output from Partition 1 data. An illustration of the data partitioning is shown in Figure 3.

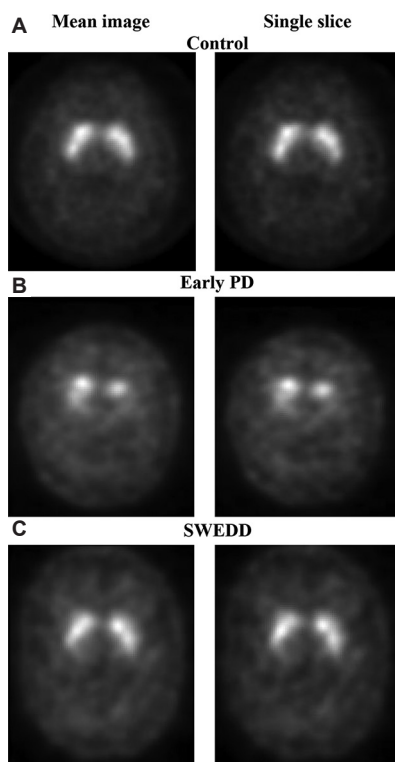


Figure 2. Mean image and single slice image for the three groups: normal control, early PD, and SWEDD. Mean image was created by taking the average of slices from 35 to 48 (from the total 91 slices), and the 41st slice represents the single slice used in the study. Abbreviations: PD: Parkinson's disease; SWEDD: Scans without evidence of dopaminergic deficit.

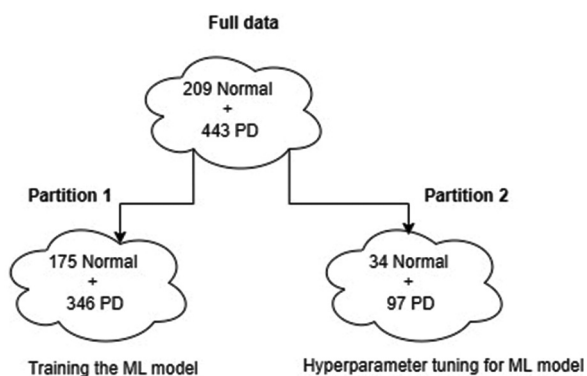


Figure 3. An illustration of data partitioning. Abbreviations: ML: Machine learning; PD: Parkinson's disease.

2.6. Machine learning techniques

The techniques utilized in the study include SVM using the linear kernel,³⁸ logistic regression,³⁹ CNN,³⁷ and multilayer perceptron (MLP).³⁷ MLP is a feed-forward neural network with one or more hidden layers. For both logistic regression and SVM, L1 regularization was employed as

it has the inherent ability of feature selection and thereby enhancing numerical stability.³⁹ For neural network based methods, dropout technique is used that involves randomly dropping out a fraction of neurons during the training process, thereby can help in preventing overfitting.⁴¹

The hyperparameters of these methods are as follows:

- SVM: Regularization of parameter C.
- Logistic regression: Regularization of parameter C.
- CNN: Number of convolutional layers, number of filters in each layer, filter sizes, dropout rate in each layer, number of neurons in the fully connected layer, dropout in the fully connected layer, batch size, and number of epochs.
- MLP: Number of hidden layers, number of neurons in each layer, and dropout rate.

These hyperparameters were estimated using a hold-out set (Partition 2, as explained in the above section), while the models were trained and evaluated using the normalized images from Partition 1 through a 10-fold cross-validation approach.

2.6.1. CNNs for predictive modeling

In this work, a CNN model was created and trained to classify early PD subjects from healthy normal controls. Unlike traditional approaches that rely on handcrafted features such as textures or shapes, CNNs automatically learn hierarchical feature representations directly from raw image data. When trained effectively, CNNs can extract both low-level features (*e.g.*, edges, textures) and high-level abstractions (*e.g.*, disease-relevant patterns), eliminating the need for manual feature engineering.³⁷ A CNN typically consists of a convolutional layer(s), a transformation layer(s), a pooling layer(s), and a fully connected layer(s). In the convolutional layer, filters are applied to regions (tiles) of the input feature map to generate new features. The hyperparameters in this layer are the size of the filter and the number of filters. During training, the CNN learns the optimal filter matrices that enable it to extract meaningful patterns from the data.

After the convolution operation, a transformation function, typically the rectified linear unit, is applied to the convolved feature. This will introduce non-linearity into the model. This is followed by a pooling step, where max pooling is typically carried out, downsampling feature map through the reduction of its spatial dimensions while preserving the most important information. In max pooling, tiles are extracted and the maximum value is taken to generate a new feature map. The max pooling filter size is 2×2 , and the stride was kept as 2 in the study. At last, there is a fully connected layer(s) that performs classification based on the features from the pooling layer.

The hyperparameters in a CNN model need to be fine-tuned for optimal performance and to prevent overfitting. For instance, increasing the number of filters in the convolutional layers can help capture more diverse features, but also leads to higher computational costs and training time. Moreover, beyond a certain point, additional filters may contribute to only minimal improvements in performance.

2.6.2. Fine-tuning: Hyperparameter optimization

Model fine-tuning is important for achieving optimal performance. In this study, all models were subjected to hyperparameter optimization. For SVM and logistic regression, a grid search combined with cross-validation was employed to identify the best configuration. For the neural network-based models – MLP and CNN – a more sophisticated approach was adopted.

Research has shown that Bayesian hyperparameter optimization is more efficient than manual, random, or grid search-based methods, particularly for neural networks, both in terms of performance and the computation time required to identify optimal hyperparameters.⁴⁰ In Bayesian optimization, unlike in random search, it keeps track of past evaluation scores which is used to form a probabilistic model mapping hyperparameters to a probability of a score on the objective function $p(y|x)$. Now this probabilistic model is much easier to optimize than the original objective function, thereby helping in finding the next best set of hyperparameters to evaluate. In this paper, the Tree-structured Parzen Estimator Approach was used to estimate the probabilistic model.⁴⁰ The optimal architecture for CNN and MLP was estimated based on this optimization and is presented in the Results section.

3. Results and discussion

The hyperparameters of classification algorithms were estimated separately for both cases, which are single slice image and mean image, using a hold-out set (Partition 2) which was not used in either training or evaluation of the models. [Table 2](#) shows the estimated hyperparameters.

The hyperparameters estimated vary between the single slice and mean image cases. The regularization parameter C for SVM and logistic regression increased slightly for the mean image case. This indicates that the models benefited from less regularization on the averaged data, and this is because averaging could have led to a reduction in noise and better overall representation of the pattern needed for detection. The CNN configurations estimated in the study are much more compact and efficient as compared to a related work²⁴ where five convolutional layers and three fully connected layers were used.

[Table 3](#) shows the 10-fold cross-validation performance metrics for all methods applied to both single slice and mean image cases. All models demonstrated excellent classification performance, with CNN achieving the highest accuracy. The performance measures obtained for the mean image consistently gave better results as compared to a single slice, except for the MLP model. This is because the mean image likely provides a more comprehensive representation of the striatal region by integrating information across multiple slices, thereby smoothing out noise and capturing more consistent patterns relevant to early PD detection. This richer representation can help these models generalize better and improve classification accuracy.

The metrics obtained here significantly improve the results obtained elsewhere²⁷ and other closely related works.^{11,19-31,33,34} For instance, in Prashanth *et al.*'s work,²⁷ a classification model was developed for the detection of early PD from normal controls using features extracted from SPECT images, achieving an accuracy of 97.29% and an area under the region operating characteristic curve (AUC) of 99.26%. In contrast, our approach achieved an accuracy of 99.08% and an AUC of 99.93% using single slice images – surpassing previous benchmarks. Notably, our method does not rely on any explicit feature extraction. Instead, CNN automatically learns discriminative features directly from the input data through its convolution and pooling operations, highlighting its capacity for effective end-to-end learning.

3.1. Error analysis

Among all the attempted methods, CNN gave the best performance metrics. However, as observed in [Table 3](#), very few records were misclassified. [Figure 4](#) shows examples of images that were misclassified, with [Figure 4A](#) showing a SPECT image from a normal subject but classified as PD, and [Figure 4B](#) showing a SPECT image from a PD subject but classified as normal. It should be noted that a normal scan is characterized by intense, uniform, and symmetric high-intensity regions on both hemispheres that appear as two comma-shaped regions (as observed in [Figure 2A](#)). In the case of [Figure 4A](#), it is observed that the tail or the bottom of the comma-shaped region is less intense as compared to the upper region. This might be an interesting case of misdetection from the CNN model as the model is actually detecting the non-uniformity in the comma-shaped region in the image. Training the network with more images like these can help alleviate these errors. In fact, such errors may even assist clinicians by flagging potentially ambiguous or borderline cases. Similarly, [Figure 4B](#) is a case of misdetection where an early PD case is detected as normal. Here as well, it is an interesting

Table 2. The hyperparameters estimated for machine learning models

Method	Single slice	Mean slice
SVM (linear kernel)	C=0.5	C=1
Logistic regression	C=3	C=5
MLP	Two-layer neural network with the configuration below: <ul style="list-style-type: none"> • Hidden layer 1: Dense layer with 64 neurons, activation=ReLU • Output layer: Dense with 2 neurons, activation=Sigmoid • Batch size: 8 • Number of epochs: 30 • Dropout: None 	Two-layer neural network with the configuration below: <ul style="list-style-type: none"> • Hidden layer 1: Dense layer with 128 neurons, activation=ReLU • Output layer: Dense with 2 neurons, activation=Sigmoid • Batch size: 8 • Number of epochs: 50 • Dropout: 0.2
CNN	<ul style="list-style-type: none"> • Layer 1: Conv2D with 64 filters, kernel size (3×3), activation=ReLU • Layer 2: MaxPooling2D with pool size (2×2) • Dropout: 0.1 • Fully connected layer: Dense with 64 neurons, activation=ReLU • Dropout: 0.3 • Output layer: Dense with 2 neurons, activation=Sigmoid 	<ul style="list-style-type: none"> • Layer 1: Conv2D with 32 filters, kernel size (5×5), activation=ReLU • Layer 2: MaxPooling2D with pool size (2×2) • Layer 3: Conv2D with 32 filters, kernel size (5×5), activation=ReLU • Layer 4: MaxPooling2D with pool size (2×2) • Dropout: 0.1 • Fully connected layer: Dense with 64 neurons, activation=ReLU • Dropout: 0.1 • Output layer: Dense with 2 neurons, activation=Sigmoid

Abbreviations: CNN: Convolutional neural network; MLP: Multilayer perceptron; ReLU: Rectified linear unit; SVM: Support vector machine.

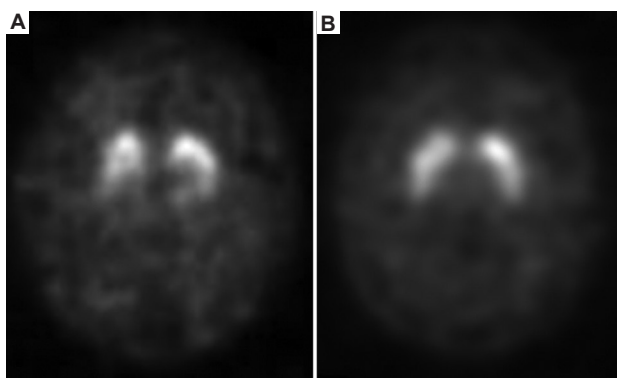


Figure 4. An illustration of misclassifications from the CNN model. (A) Normal detected as early PD. (B) Early PD detected as normal. Abbreviations: CNN: Convolutional neural network; PD: Parkinson's disease.

observation that the characteristics of the image appear similar to the patterns in a normal image which might have caused the misdetection.

3.2. Performance on SWEDD data

The SWEDD data consist of 80 subjects and were input to the machine-learned models. The performance of these methods is given in Table 4. CNN gave the best detection with an accuracy of 95% (76 out of 80). Figure 5 shows the cases of misclassification from the CNN model. It is interesting to observe that all these misclassified images

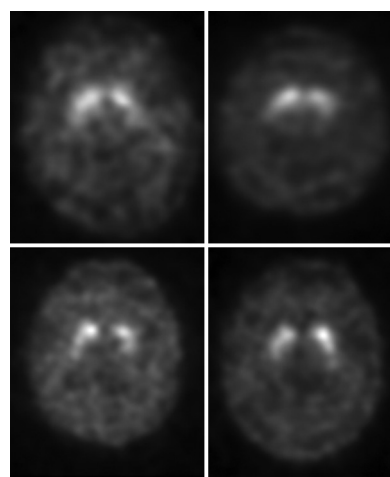


Figure 5. SWEDD images that were misclassified as early PD by the CNN model. Abbreviations: CNN: Convolutional neural network; PD: Parkinson's disease; SWEDD: Scans without evidence of dopaminergic deficit.

show unexpected pattern of dull and uneven comma-shaped regions, which deviates from the bright and even comma-shaped regions seen in normal images. Using the PPMI data for analysis, Choi *et al.*¹¹ observed that a few SWEDD cases showing unusual image pattern were classified as abnormal (or PD), and the diagnosis of the majority of these cases was later changed to clinical PD based on a 2-year follow-up. This finding underscores the potential of machine learning techniques, particularly

Table 3. Performance metrics obtained for different methods for using single slice image and mean image

Method	Confusion matrix	Accuracy	AUC	APR	Precision	Recall	Specificity
A. Single slice image							
SVM	$\begin{bmatrix} 335 & 11 \\ 13 & 162 \end{bmatrix}$	95.39	98.70	95.87	96.26	96.82	92.57
Log Reg	$\begin{bmatrix} 336 & 10 \\ 13 & 162 \end{bmatrix}$	95.58	98.62	96.25	96.28	97.11	92.57
MLP	$\begin{bmatrix} 338 & 8 \\ 4 & 171 \end{bmatrix}$	97.69	99.57	99.11	98.83	97.69	97.71
CNN	$\begin{bmatrix} 340 & 6 \\ 3 & 172 \end{bmatrix}$	98.27	99.78	99.45	99.13	98.27	98.29
B. Mean image							
SVM	$\begin{bmatrix} 338 & 8 \\ 9 & 166 \end{bmatrix}$	96.73	98.78	96.30	97.41	97.69	94.86
Log Reg	$\begin{bmatrix} 338 & 8 \\ 12 & 163 \end{bmatrix}$	96.16	98.74	96.95	96.57	97.69	93.14
MLP	$\begin{bmatrix} 338 & 8 \\ 5 & 170 \end{bmatrix}$	97.50	99.12	96.56	98.54	97.69	97.14
CNN	$\begin{bmatrix} 340 & 6 \\ 2 & 173 \end{bmatrix}$	98.46	99.91	99.80	99.41	98.27	98.86

Notes: The confusion matrix is represented as $\begin{bmatrix} \text{True positive} & \text{False negative} \\ \text{False positive} & \text{True negative} \end{bmatrix}$. Single slice is the 41st slice image and mean image is the image

obtained after taking the mean of all slices numbered from 35 to 48 in the 3D scan. All values are expressed in percentage (%).

Abbreviations: AUC: Area under the region operating characteristic curve; CNN: Convolutional neural network; LogReg: Logistic regression; MLP: Multilayer perceptron; SVM: Support vector machine.

Table 4. Classification of the SWEDD data using different methods

Method	Mean image		Single slice image	
	True negative	False positive	True negative	False positive
CNN	75	5	76	4
LogReg	73	7	73	7
LinearSVM	73	7	73	7
MLP	74	6	73	7

Abbreviations: CNN: Convolutional neural network; LogReg: Logistic regression; MLP: Multilayer perceptron; SVM: Support vector machine; SWEDD: Scans without evidence of dopaminergic deficit.

CNNs, in this domain, as these models are capable of learning subtle and complex patterns from training data and making inferences that may even precede clinical judgment.

4. Limitations and future work

Recent research shows that deep learning techniques such as the CNN could benefit from the latest advances such as data augmentation, which represents a technique used to increase the training data using information from the available training data.⁴² Traditional transformations which include a combination of various affine transformations and using generative adversarial networks⁴³ are effective ways to augment the data. Label smoothing is another advancement that has shown to improve the performance of deep learning models.⁴⁴ In label smoothing, the hard class labels are converted to soft labels. Both data augmentation and label smoothing are methods for regularizing the neural network models, which can help in preventing overfitting and also help networks in converging faster.

Despite the strong performance of the proposed models, it is not without any limitations. First, the model was trained and evaluated solely on the PPMI dataset, and its generalizability to external datasets remains to be validated. Second, the slice selection and averaging strategies used in this study are based on fixed indices that are applicable in the PPMI data, which may not optimally capture relevant features in all subjects in other datasets. Third, while CNNs perform well, they are inherently black-box models, making it difficult to interpret specific feature-driven decisions. Finally, although promising results were achieved in this controlled research setup, further validation is necessary before deployment in clinical environments.

5. Conclusion

Accurate and early detection of PD is a challenging clinical problem. The numerous common symptoms shared by this class of Parkinsonism disorders represent a source of misdiagnosis. Accurate identification of degenerative Parkinsonism from other non-degenerative ones is crucial for effective patient treatment and management. In this work, machine learning models that could classify early PD from healthy normal and also SWEDD from PD using SPECT images were developed. These machine learning models exhibit very high performance, of which the CNN model especially achieves the best performance. These predictive models carry enormous potential to be used in a clinical setting and can act as an aid the clinical diagnostic process.

Acknowledgments

None.

Funding

PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and other funding partners include AbbVie, Allergan, Amathus Therapeutics, Avid Radiopharmaceuticals, Biogen Idec, BioLegend, Bristol-Myers Squibb, Celgene, Denali Therapeutics, GE Healthcare, Genentech, GlaxoSmithKline, Eli Lilly and Company, Lundbeck, Merck & Co., Meso Scale Discovery, Pfizer, Piramal, Prevail Therapeutics, Hoffmann-La Roche, Sanofi Genzyme, Servier, Takeda Pharmaceutical Company, Teva, Verily Life Sciences, Voyager Therapeutics, and UCB (Union ChimiqueBelge).

Conflict of interest

The author declares no competing interests.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

Further disclosure

This work was carried out independently by the author. The author is currently employed at Siemens Healthineers, Bangalore, India, however, the views expressed and the work presented here are solely those of the author and do not reflect the views of the company.

References

1. Jankovic J. Parkinson's disease: Clinical features and diagnosis. *J Neurol Neurosurg Psychiatry*. 2008;79(4):368-376. doi: 10.1136/jnnp.2007.131045
2. Fahn S. Description of Parkinson's disease as a clinical syndrome. *Ann N Y Acad Sci*. 2003;991:1-14. doi: 10.1111/j.1749-6632.2003.tb07458.x
3. Tysnes OB, Storstein A. Epidemiology of Parkinson's disease. *J Neural Transm (Vienna)*. 2017;124(8):901-905. doi: 10.1007/s00702-017-1686-y
4. Marek K, Jennings D, Lasch S, *et al*. The Parkinson progression marker initiative (PPMI). *Prog Neurobiol*. 2011;95(4):629-635. doi: 10.1016/j.pneurobio.2011.09.005
5. Groveman BR, Orrù CD, Hughson AG, *et al*. Rapid and ultra-sensitive quantitation of disease-associated α -synuclein seeds in brain and cerebrospinal fluid by α Syn RT-QuIC. *Acta Neuropathol Commun*. 2018;6:7. doi: 10.1186/s40478-018-0508-2
6. Booij J, Tissingh G, Boer GJ, *et al*. [123I]FP-CIT SPECT shows a pronounced decline of striatal dopamine transporter labelling in early and advanced Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 1997;62(2):133-140. doi: 10.1136/jnnp.62.2.133
7. Booth TC, Nathan M, Waldman AD, Quigley AM, Schapira AH, Buscombe J. The role of functional dopamine-transporter SPECT imaging in Parkinsonian syndromes, part 1. *AJNR Am J Neuroradiol*. 2015;36(2):236-244.

- doi: 10.3174/ajnr.A3971
8. Cummings JL, Henchcliffe C, Schaier S, Simuni T, Waxman A, Kemp P. The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration. *Brain*. 2011;134(11):3146-3166.
doi: 10.1093/brain/awr177
 9. Seibyl J, Jennings D, Grachev I, Coffey C, Marek K. 123-I Ioflupane SPECT measures of Parkinson disease progression in the Parkinson progression marker initiative (PPMI) trial. *J Nucl Med*. 2013;54(1):57-58.
 10. Beach TG, Adler CH. Importance of low diagnostic accuracy for early Parkinson's disease. *Mov Disord*. 2018;33(10):1551-1554.
doi: 10.1002/mds.27485
 11. Choi H, Ha S, Im HJ, Paek SH, Lee DS. Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging. *Neuroimage Clin*. 2017;16:586-594.
doi: 10.1016/j.nicl.2017.09.010
 12. Coarelli G, Garcin B, Roze E, Vidailhet M, Degos B. Invalidation of Parkinson's disease diagnosis after years of follow-up based on clinical, radiological and neurophysiological examination. *J Neurol Sci*. 2019;406:116454.
doi: 10.1016/j.jns.2019.116454
 13. Catafau AM, Tolosa E, DaTscan Clinically Uncertain Parkinsonian Syndromes Study Group. Impact of dopamine transporter SPECT using 123I-Ioflupane on diagnosis and management of patients with clinically uncertain Parkinsonian syndromes. *Mov Disord*. 2004;19(10):1175-1182.
doi: 10.1002/mds.20112
 14. Kupsch AR, Bajaj N, Weiland F, et al. Impact of DaTscan SPECT imaging on clinical management, diagnosis, confidence of diagnosis, quality of life, health resource use and safety in patients with clinically uncertain parkinsonian syndromes: A prospective 1-year follow-up of an open-label controlled study. *J Neurol Neurosurg Psychiatry*. 2012;83(6):620-628.
doi: 10.1136/jnnp-2011-301695
 15. Marek K, Jennings D, Seibyl JP. Long-term follow-up of patients with scans without evidence of dopaminergic deficit (SWEDD) in the ELLDOPA study. *Neurology*. 2005;64:A274.
 16. Benamer TS, Patterson J, Grosset DG, et al. Accurate differentiation of parkinsonism and essential tremor using visual assessment of [123I]-FP-CIT SPECT imaging: The [123I]-FP-CIT study group. *Mov Disord*. 2000;15(3):503-510.
 17. Scherfler C, Nocker M. Dopamine transporter SPECT: How to remove subjectivity? *Mov Disord*. 2009;24(Suppl 2):S721-S724.
doi: 10.1002/mds.22590
 18. Staff RT, Ahearn TS, Wilson K, et al. Shape analysis of 123I-N-omega-fluoropropyl-2-beta-carbomethoxy-3beta-(4-iodophenyl) nortropane single-photon emission computed tomography images in the assessment of patients with parkinsonian syndromes. *Nucl Med Commun*. 2009;30(3):194-201.
doi: 10.1097/MNM.0b013e328314b863
 19. Hirschauer TJ, Adeli H, Buford JA. Computer-aided diagnosis of Parkinson's disease using enhanced probabilistic neural network. *J Med Syst*. 2015;39(11):179.
doi: 10.1007/s10916-015-0353-9
 20. Kim DH, Wit H, Thurston M. Artificial intelligence in the diagnosis of Parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning. *Nucl Med Commun*. 2018;39(10):887-893.
doi: 10.1097/MNM.0000000000000890
 21. Martínez-Murcia FJ, Górriz JM, Ramírez J, Illán IA, Ortiz A, Parkinson's Progression Markers Initiative. Automatic detection of Parkinsonism using significance measures and component analysis in DaTSCAN imaging. *Neurocomputing*. 2014;126:58-70.
doi: 10.1016/j.neucom.2013.01.054
 22. Martínez-Murcia FJ, Ortiz A, Górriz JM, et al. A 3D convolutional neural network approach for the diagnosis of Parkinson's disease. In: *International Work-Conference on the Interplay Between Natural and Artificial Computation*. Cham: Springer; 2017. p. 324-333.
doi: 10.1007/978-3-319-59740-9_32
 23. Oliveira FP, Castelo-Branco M. Computer-aided diagnosis of Parkinson's disease based on [(123)I]FP-CIT SPECT binding potential images, using the voxels-as-features approach and support vector machines. *J Neural Eng*. 2015;12(2):026008.
doi: 10.1088/1741-2560/12/2/026008
 24. Ortiz A, Munilla J, Martínez-Ibañez M, Górriz JM, Ramírez J, Salas-Gonzalez D. Parkinson's disease detection using isosurfaces-based features and convolutional neural networks. *Front Neuroinform*. 2019;13:48.
doi: 10.3389/fninf.2019.00048
 25. Prashanth R, Dutta Roy S, Mandal PK, Ghosh S. Automatic classification and prediction models for early Parkinson's disease diagnosis from SPECT imaging. *Expert Syst Appl*. 2014;41(8):3333-3342.
doi: 10.1016/j.eswa.2013.11.031
 26. Prashanth R, Dutta Roy S, Mandal PK, Ghosh S. High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. *Int J Med Inform*. 2016;90:13-21.

- doi: 10.1016/j.ijmedinf.2016.03.001
27. Prashanth R, Roy SD, Mandal PK, Ghosh S. High-accuracy classification of Parkinson's disease through shape analysis and surface fitting in 123I-Ioflupane SPECT imaging. *IEEE J Biomed Health Inform.* 2017;21(2):794-802.
doi: 10.1109/JBHI.2016.2547901
28. Segovia F, Gorriz JM, Ramirez J, Alvarez I, Jimenez-Hoyuela JM, Ortega SJ. Improved parkinsonism diagnosis using a partial least squares based approach. *Med Phys.* 2012;39(7):4395-4403.
doi: 10.1118/1.4730289
29. Illan IA, Gorriz JM, Ramirez J, Segovia F, Jimenez-Hoyuela JM, Ortega Lozano SJ. Automatic assistance to Parkinson's disease diagnosis in DaTSCAN SPECT imaging. *Med Phys.* 2012;39(10):5971-5980.
doi: 10.1118/1.4742055
30. Rojas A, Górriz JM, Ramírez J, et al. Application of empirical mode decomposition (EMD) on DaTSCAN SPECT images to explore Parkinson disease. *Expert Syst Appl.* 2013;40(7):2756-2766.
doi: 10.1016/j.eswa.2012.11.017
31. Towey DJ, Bain PG, Nijran KS. Automatic classification of 123I-FP-CIT (DaTSCAN) SPECT images. *Nucl Med Commun.* 2011;32(8):699-707.
doi: 10.1097/MNM.0b013e328347cd09
32. Huertas-Fernandez I, Garcia-Gomez F, Garcia-Solis D, et al. Machine learning models for the differential diagnosis of vascular parkinsonism and Parkinson's disease using [(123) I] FP-CIT SPECT. *Eur J Nucl Med Mol Imaging.* 2015;42(1):112-119.
doi: 10.1007/s00259-014-2882-8
33. Oliveira FP, Faria DB, Costa DC, Castelo-Branco M, Tavares JMR. Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson's disease based on [¹²³I] FP-CIT SPECT images. *Eur J Nucl Med Mol Imaging.* 2018;45(6):1052-1062.
doi: 10.1007/s00259-017-3918-7
34. Zhang X, Chou J, Liang J, et al. Data-driven subtyping of Parkinson's disease using longitudinal clinical records: A cohort study. *Sci Rep.* 2019;9(1):797.
doi: 10.1038/s41598-018-37545-z
35. Marek K, Chowdhury S, Siderowf A, et al. The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann Clin Transl Neurol.* 2018; 5(12):1460-1477.
doi: 10.1002/acn3.644
36. Hoehn MM, Yahr MD. Parkinsonism: Onset, progression and mortality. *Neurology.* 1967;17(5):427-442.
doi: 10.1212/wnl.17.5.427
37. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521(7553):436-444.
doi: 10.1038/nature14539
38. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20(3):273-297.
doi: 10.1007/BF00994018
39. Lee SI, Lee H, Abbeel P, Ng AY. Efficient l-1 regularized logistic regression. In: *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI).* 2006. p. 401-408.
40. Bergstra JS, Bardenet R, Bengio Y, Kégl B. Algorithms for hyper-parameter optimization. In: *Advances in Neural Information Processing Systems 24.* 2011. p. 2546-2554.
41. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. *J Mach Learn Res.* 2014;15(1):1929-1958.
42. Perez L, Wang J. *The Effectiveness of Data Augmentation in Image Classification using Deep Learning.* [arXiv Preprint]; 2017.
43. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems 27.* [Preprint]; 2014. p. 2672-2680
44. Muller R, Kornblith S, Hinton GE. When does label smoothing help? In: *Advances in Neural Information Processing Systems 32.* 2019. p. 4696-4705.
45. Shiiba T, Takano K, Takaki A, Suwazono S. Dopamine transporter single-photon emission computed tomography-derived radiomics signature for detecting Parkinson's disease. *EJNMMI Res.* 2022;12(1):39.
doi: 10.1186/s13550-022-00910-1
46. Tufail AB, Ma YK, Zhang QN, et al. 3D convolutional neural networks-based multiclass classification of Alzheimer's and Parkinson's diseases using PET and SPECT neuroimaging modalities. *Brain Inform.* 2021;8(1):23.
doi: 10.1186/s40708-021-00144-2
47. Majhi B, Kashyap A, Mohanty SS, et al. An improved method for diagnosis of Parkinson's disease using deep learning models enhanced with metaheuristic algorithm. *BMC Med Imaging.* 2024;24(1):156.
doi: 10.1186/s12880-024-01335-z
48. Khachnaoui H, Chikhaoui B, Khelifa N, Mabrouk R. Enhanced Parkinson's disease diagnosis through convolutional neural network models applied to spect datscan images. *IEEE Access.* 2023;11:91157-91172.
doi: 10.1109/ACCESS.2023.3308075

ORIGINAL RESEARCH ARTICLE

Deep vision transformers in neurodegenerative disease diagnosis using ^{18}F -fluorodeoxyglucose positron emission tomography scans and anatomical brain atlasPooriya Khorramyar*, Amira Soliman, Farzaneh Etmnani, and Stefan Byttner

Center for Applied Intelligent Systems Research in Health (CAISR Health), The School of Information Technology, Halmstad University, Halmstad, Halland, Sweden

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

This research explores adapting vision transformers (ViTs) to classify neurodegenerative diseases while ensuring their decision-making process is interpretable. We developed a model to classify ^{18}F -fluorodeoxyglucose (^{18}F -FDG) positron emission tomography (PET) brain scans into three categories: cognitively normal (CN), mild cognitive impairment (MCI), and Alzheimer's disease (AD). The dataset utilized in this research contains 580 samples of ^{18}F -FDG PET scans obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The proposed model obtained an F1 score of 81% (macro-average of all classes) on the test dataset, a significant performance improvement compared to the literature. Furthermore, we combined the model's attention maps with the Automated Anatomical Atlas 3 (AAL3), which represents a digital brain map, to identify the most influential areas on the model's predictions and to conduct a regions' importance study as a step toward explainability. We demonstrated that ViTs can achieve competitive performance compared to convolutional neural networks while enabling the development of explainable models without extra computations due to the attention mechanism.

Keywords: Vision transformer; Neurodegenerative disease; ^{18}F -FDG PET; Medical image analysis; Brain scan; Deep neural network

***Corresponding author:**Pooriya Khorramyar
(pookho20@student.hh.se)

Citation: Khorramyar P, Soliman A, Etmnani F, Byttner S. Deep vision transformers in neurodegenerative disease diagnosis using ^{18}F -fluorodeoxyglucose positron emission tomography scans and anatomical brain atlas. *Artif Intell Health*. 2025;2(4):33-46. doi: 10.36922/AIH025140026

Received: March 31, 2025**1st revised:** April 12, 2025**2nd revised:** May 22, 2025**Accepted:** May 26, 2025**Published online:** June 19, 2025**Copyright:** © 2025 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, which provided that 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

Neurodegenerative diseases (NDDs) lead to progressive deterioration and death of neurons, damaging the nervous system and brain. Affecting more than 55 million patients with a yearly increase rate of 10 million new cases worldwide, NDDs are a prominent cause of disability and death.¹ In addition, Alzheimer's disease (AD), as the most widespread form, accounts for 70% of NDD cases and plays a significant role in these statistics.¹ Although NDDs have a heavy impact on healthcare systems and patients' lives, they remain incurable as of today.¹ However, timely diagnosis is pivotal in disease management and improving the patient's quality of life.²

Diagnosing NDDs is exceedingly demanding and requires years of training and experience. Hence, according to some studies, it has been estimated that 75% of NDD cases are undiagnosed worldwide due to various reasons, including the diagnosis complexity.³ Astoundingly, this number rises to 90% in low- and middle-income countries, according to the same analysis.³ Moreover, the growing number of NDD cases could devastate healthcare systems in coming years, according to this study.³ Therefore, innovative and affordable methods are needed to assist doctors and decrease this diagnosis gap.

The rapid progress of artificial intelligence (AI) and its sub-fields has led to outstanding results in different domains, including medical image processing. Thus, researchers attempted to harness the power of deep neural networks (DNNs) in diagnosing NDDs and demonstrated that they could have competitive performance compared to human experts.^{4,5}

The advent of vision transformers (ViTs) resulted in distinguished performance in various computer vision tasks, surpassing traditional approaches like convolutional neural networks (CNNs).⁶ Therefore, their application in NDD diagnosis has been a trending research subject and the focal point of various studies, including this paper. We developed our model based on vanilla ViT, proposed by Dosovitskiy *et al.*⁶, and trained it using ¹⁸F-fluorodeoxyglucose (¹⁸F-FDG) positron emission tomography (PET) brain scans provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI).⁷ The motivation behind our work is as follows:

- Dosovitskiy *et al.*⁶ achieved exceptional results in image classification tasks by applying standard transformers,⁸ utilized in natural language processing (NLP), directly to images with the least possible modifications. In addition to its notable performance, this approach enables vision models to benefit from advancements in the NLP domain, including large language models, because of architectural similarities. Consequently, vanilla ViT⁶ was a rational and sustainable foundation due to its design, performance, and simplicity for investigating what transformer-based vision models accomplish in diagnosing NDDs.
- ¹⁸F-FDG PET scans, which reveal metabolic activities of various brain regions by measuring their glucose consumption, are considered pivotal in diagnosing and discriminating different NDDs, including mild cognitive impairment (MCI) and AD.⁹ Although other brain imaging technologies such as computed tomography (CT) and magnetic resonance imaging (MRI) can expose NDDs too, PET scans have proved to be superior in exposing these brain conditions as soon as possible and earlier than other methods.^{10,11}

Understanding the model's logic is the key to obtaining explainability in the medical domain, as human users must comprehend the reasoning behind each prediction before considering it. Therefore, we combined ViT's attention maps and the Automated Anatomical Atlas 3 (AAL3)¹² brain atlas to develop an explainable model that provides the most critical brain regions in the classification. The proposed model also delivers a heatmap of the input scan, in which the brightness of each pixel corresponds to its significance in the model's decision, overlaid on the original image, allowing the user to investigate pivotal regions further.

Our model achieves an F1 score of 81% (macro-average of all classes) on the test dataset, surpassing other approaches by a considerable gap. Please note that we only analyze our results against comparable studies regarding classes and the type of input brain scans. Furthermore, our proposed ViT has remarkable performance, in contrast to other models, in distinguishing MCI, which has proved to be one of the most challenging brain conditions to diagnose due to its prodromal nature. MCI is a transition stage between cognitively normal (CN) and AD. Consequently, MCI patients may experience some common NDD symptoms, such as memory loss or language problems, but the extent is such that they do not impede daily life.¹³ Therefore, differentiating MCI cases from other categories can be inherently complicated.

Finally, we conducted experiments to reveal the contribution of different brain regions to the model's decisions. Although NDDs can affect various areas, this study showed that some brain regions are significantly more critical in the model's predictions.

To summarize, the contribution and novelty of this research is as follows:

- Introducing a complete data pre-processing and reshaping pipeline for 3D PET scans and brain atlases, allowing for fine-tuning of pre-trained ViTs on this type of data. This step is crucial since most ViTs are pre-trained on natural three-channel RGB images. Therefore, resizing and reshaping 3D data into three channels are essential to match the model's input shape.
- Obtaining competitive performance in ternary NDD classification (CN/MCI/AD) utilizing ¹⁸F-FDG PET brain scans and vanilla ViT.⁶ This approach is beneficial since vanilla ViT mostly shares the same architecture as the standard transformer,⁸ used in NLP. Therefore, these architectural similarities could enable future studies to leverage advancements in NLP.
- Outperforming previous approaches by a noticeable margin (specifically in predicting MCI cases) in

the ternary classification of NDDs solely based on ^{18}F -FDG PET scans.

- Combining the model's attention maps and the AAL3 brain atlas for improved model explainability. Apart from the predicted label, our model provides a heatmap overlaid on the original input scan, highlighting the most influential brain regions to the model's prediction. Furthermore, the model delivers names of the key areas with the assistance of the AAL3 brain atlas.
- Performing a comprehensive brain regions' importance analysis by combining the model's attention maps and AAL3 atlas to find the most influential areas in the model's predictions. This study aims to enhance the model's explainability and suggest key areas in distinguishing various brain conditions.

2. Related works

Deep learning algorithms have shown outstanding results and potential in solving intricate tasks, motivating researchers to employ them for various medical image analysis tasks, including NDD classification.

Before the emergence of transformer-based vision models, such as ViTs,⁶ most researchers had focused on employing CNNs for NDD classification.^{4,5} Etminani *et al.*⁴ proposed a comprehensive data pre-processing pipeline and a 3D CNN model based on VGG16¹⁴ for NDD classification using ^{18}F -FDG PET scans. The authors demonstrated that 3D CNN algorithms could obtain competitive results compared to human readers, outperforming experienced nuclear medicine physicians independently and their consensus.⁴ Furthermore, Etminani *et al.*⁴ focused on explainability and dedicated a part of their research to interpreting the suggested model using an occlusion experiment.¹⁵ Ding *et al.*⁵ developed a CNN established on inception-v3¹⁶ to classify NDDs through brain ^{18}F -FDG PET scans. The authors also compared their model's performance to radiology readers' using a subset of the ADNI and an independent test dataset, which resulted in the model's superior results in both cases. Furthermore, Ding *et al.*⁵ employed the saliency map approach¹⁷ for the model interpretation and analysis. Lozupone *et al.*¹⁸ utilized 2D CNNs and a new explainable AI strategy to develop an interpretable model for classifying NDDs; however, the authors aimed for a two-class classification in their research and used 3D MRI brain scans for designing the model.

The advent of ViTs⁶ and their cutting-edge performance in various computer vision tasks convinced researchers to investigate utilizing them in the medical domain and NDD diagnosis. Khatri and Kwon¹⁹ focused on designing an explainable ViT utilizing self-supervised learning and ^{18}F -FDG PET scans for binary classification of two

MCI sub-categories, namely convertible MCI (MCI-c) and stable MCI (MCI-s), to predict MCI progression to AD. The authors also studied attention regions for model explainability. Shin *et al.*²⁰ proposed applying ViTs on ^{18}F -florbetaben scans for binary and ternary classification of NDDs. Although this type of PET scan, which demonstrates beta-amyloid (β -amyloid) plaques in the brain, has proved beneficial in identifying NDDs, it is often used in research settings.²¹ Therefore, ^{18}F -FDG has remained the most commonly used brain PET imaging technique.¹¹ Xing *et al.*²² developed a multi-modal ViT by combining two types of PET brain scans (^{18}F -FDG and ^{18}F -AV45) for the binary classification of NDDs. Specifically, the proposed model includes two ViTs, each specialized in extracting features of a specific PET type. Then, the extracted features are concatenated and fed into a classifier for the final prediction.²² Similarly, Odusami *et al.*²³ suggested an approach for binary classification of NDDs by fusing MRI and PET brain scans.

Most studies have concentrated on applying ViTs to MRI data. Unlike PET scans, which expose metabolic activities and functions, MRI is supposed to reveal the brain's structure. Therefore, MRI is usually beneficial in diagnosing NDDs at later stages, when the disease causes abnormalities in the physical brain's structure. Lyu *et al.*²⁴ developed a ViT solely based on an MRI dataset for a binary classification task; however, the authors added convolutional layers to their model to obtain better results. Sarraf *et al.*²⁵ proposed OVITAD, an optimized ViT architecture trained on a combination of functional MRI and structural MRI (sMRI) to classify NDDs. Furthermore, the authors used attention maps to achieve better model interpretation. Hoang *et al.*²⁶ focused their study on predicting MCI cases that could potentially progress into AD; therefore, the authors trained their ViT on sMRI data for a binary classification.²⁶ Aghdam *et al.*²⁷ applied a pre-trained pyramid ViT²⁸ to sMRI data to classify CN and AD cases. Kushol *et al.*²⁹ designed Addformer, which utilizes a new fusion transformer block that combines sMRI data in spatial and frequency domains to improve binary classification accuracy. They also visualized the model's attention maps, similar to most ViT-based studies, to gain model explainability. Shah *et al.*³⁰ introduced the multi-modal Bi-vision Transformer (BiViT), a ViT that includes two modules of mutual latent fusion and parallel coupled encoding strategy to enhance feature learning. The authors also utilized MRI data and demonstrated tokens for a better model understanding.

While we aimed for NDD classification in this work, there are some key differences compared to the literature, as follows:

- Achieving competitive performance in ternary NDD

classification while developing the ViT solely on ¹⁸F-FDG PET scans

- Integrating a brain atlas with ViT’s attention maps to gain model explainability and provide more information to the user.

3. Data and methods

3.1. Data acquisition

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Michael W. Weiner, MD, as the principal investigator. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

Figure 1 depicts a 3D raw ¹⁸F-FDG PET scan selected from the ADNI dataset before our pre-processing steps along axial, sagittal, and coronal axes. A thorough description of technical details for each imaging session and phase is available in the ADNI documentation.³¹

The following criteria in choosing ¹⁸F-FDG PET scans from ADNI were considered, similar to Etminani *et al.*:⁴

- CN and AD: We solely selected the most recent scan for each subject if more than one was available
- MCI: We exclusively chose the cases that later developed into AD during the ADNI studies.

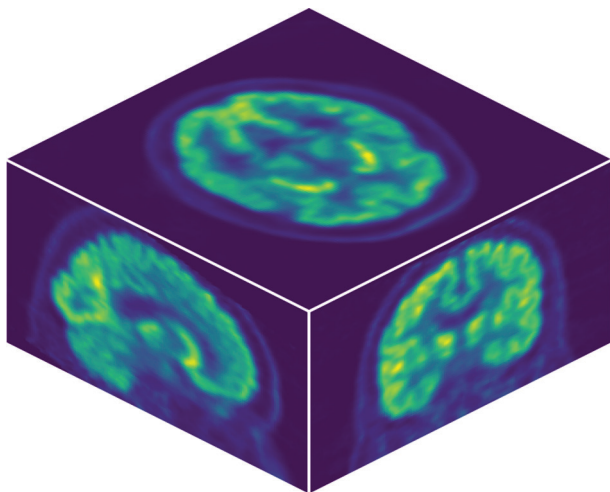


Figure 1. A 3D raw ¹⁸F-FDG PET scan from the ADNI dataset along axial, sagittal, and coronal axes. The ADNI scans differ in voxel intensities, image size, and number of channels since they are obtained using a diverse range of scanners on different sites. Also, each scan contains the subject’s skull, which does not provide beneficial information for our research. Therefore, these scans need pre-processing before utilizing them for model training.

Abbreviations: ADNI: Alzheimer’s Disease Neuroimaging Initiative; ¹⁸F-FDG: ¹⁸F-fluorodeoxyglucose; PET: Positron emission tomography.

Considering this and the first criterion, the dataset includes the last MCI scan before progression to AD.

This sample selection procedure resulted in a dataset of size 580. Table 1 provides the details about the dataset split ratios and the number of samples.

3.2. Brain imaging technologies and techniques

There are several brain imaging technologies with their unique advantages and disadvantages. Thus, in this part, we discuss the rationale behind utilizing ¹⁸F-FDG PET scans in our research.

A PET scan imaging session starts after injecting slight amounts of a radioactive tracer into the subject’s veins, which spreads to the body through the blood flow. The tracer enables the PET scanning device to capture metabolic activities in various tissues and organs, including the subject’s brain.

Although all brain imaging technologies can reveal NDDs when sufficiently developed, PET scans are the best choice for detecting brain conditions at the earliest stages.^{10,11} The reason is that NDDs usually cause abnormal metabolic patterns in some parts of the brain from the very early phases.¹⁰ Therefore, PET imaging often exposes NDDs before other brain imaging technologies, including CT and MRI, due to its focus on the brain’s metabolism.^{10,11}

There are three well-known PET imaging types, namely amyloid, tau, and FDG, each suited for demonstrating special metabolic activities or changes in the brain using different tracers and procedures. Amyloid and tau PET scans, although showing promising results in NDD diagnosis, are commonly used in research settings at the time of writing.²¹ Consequently, ¹⁸F-FDG PET scans that show the brain’s glucose (energy) usage are the most accessible and standard option in NDD diagnosis.

A central objective of our research was to propose a model and set of methods that enable rapid clinical diagnosis of NDDs. Consequently, ¹⁸F-FDG PET scans were the most reasonable choice compared to other imaging technologies since they usually allow for early identification of NDDs.

Table 1. The number of samples per class and data split ratios

Class	Training	Validation	Test
CN	140	20	20
MCI	160	20	20
AD	160	20	20
Sum	460	60	60

Abbreviations: AD: Alzheimer’s disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

3.3. Data pre-processing

The ¹⁸F-FDG PET scans in the ADNI dataset were acquired utilizing different types of imaging devices and during various phases. Therefore, these scans vary significantly in their properties, including the image size, number of channels, and intensities of voxels. Furthermore, they include the subject’s skull, which does not deliver beneficial information regarding NDD diagnosis. Furthermore, raw scans may contain noise or blur due to the patient’s movement or other technical issues. Consequently, we use the pre-processing procedure developed by Etmnani *et al.*,⁴ which employs MATLAB³² and statistical parametric mapping (SPM12)³³ to ensure all scans have the same properties.

The pre-processing steps for each sample are as follows:

- We converted the scan to the NIfTI format
- It was crucial to place the brain approximately in the center of the scan. Therefore, we reoriented and repositioned the brain to set the volume’s origin at the anterior commissure region
- Our dataset included scans of various shapes. Hence, we normalized the scan to ensure all samples had identical spatial size and number of channels
- Using the tissue probability map of SPM12, the brain was segmented
- The last pre-processing stage removed the subject’s skull from the scan. Consequently, we used segmentation maps obtained from the previous step with a filter for skull-stripping.

The pre-processing procedure led to skull-stripped scans of size 79 × 95 × 79, representing channels, height, and width, respectively. Then, the values of voxels were normalized using a min-max scaler across the channels in Python. Finally, we dismissed the initial ten and last nine channels of the 3D scan since they included a tiny fraction of the brain, resulting in 3D scans with the shape of 60 × 95 × 79.

3.4. Data reshaping

According to our experiments and the literature,⁶ pre-training on large amounts of data is crucial to achieving the best

performance using ViTs. However, most large computer vision datasets include natural images with three RGB channels. Therefore, we reshaped the samples to 3 × 570 × 950 in our dataset to utilize transfer learning and available pre-trained models. This procedure constructs a three-channel image, in which every channel depicts the brain along a unique axis (sagittal, coronal, and axial). Figure 2 shows the result of data pre-processing and reshaping steps on a single scan.

3.5. Model architecture and training

Our proposed model has a similar architecture to vanilla ViT, as suggested by Dosovitskiy *et al.*⁶ After training different models with and without transfer learning, we concluded that pre-training is crucial to obtaining excellent results. Therefore, we employed the Hugging Face³⁴ Transformer API and model hub for development. Specifically, the foundation of our model is a base-sized ViT, pre-trained on ImageNet-21k³⁵ and ImageNet 2012,³⁶ respectively.³⁷ Finally, we fine-tuned the model on our ¹⁸F-FDG PET scan dataset to classify NDDs.

Figure 3 illustrates the model’s diagram, inspired by Dosovitskiy *et al.*⁶ First, the scan was resized to 3 × 384 × 384 to match the model’s input shape. Then, the scan was divided into patches of 3 × 32 × 32, flattened, and supplied to a standard transformer along with position embeddings holding the spatial information. Finally, a multilayer perceptron head translated the model’s final hidden state into the probability of classes for the classification task. Table 2 summarizes the model’s specifications.

We employed an AdamW optimizer (learning rate = 5e-5, weight decay = 0.15) for model development. Furthermore, an exponential learning rate decay (γ = 0.9999 per epoch) was used during training. Finally, we selected a weighted cross-entropy as the loss function, in which the weight of each class was the inverse of its frequency during training, as shown below:

$$l(x, y) = L = \{l_1 \dots l_N\}^T \tag{I}$$

$$l_n = -\sum_{c=1}^C w_c \log \frac{\exp(x_{n,c})}{\sum_{i=1}^C \exp(x_{n,i})} y_{n,c} \tag{II}$$

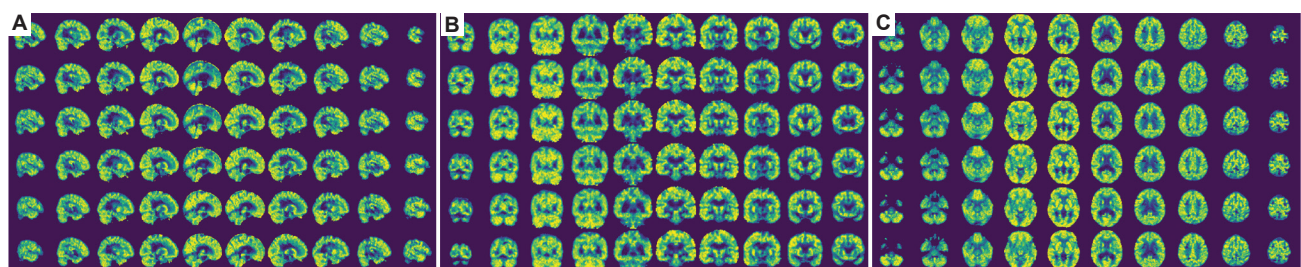


Figure 2. After the initial data pre-processing, we reshaped each scan of 60 × 95 × 79 to a three-channel image of 3 × 570 × 950, in which every channel illustrates the brain along a unique axis. This data reshaping was crucial to utilize transfer learning and pre-train the model on large computer vision datasets that contain natural three-channel RGB images. (A) Sagittal, (B) Coronal, (C) Axial.

$$w_c = \frac{1}{Class\ Frequency} \rightarrow w_{CN} = \frac{1}{140}, w_{MCI} = \frac{1}{160}, w_{AD} = \frac{1}{160} \quad (III)$$

Finally, Algorithm 1 shows the data augmentation process for model training.

Algorithm 1. The data augmentation procedure used in the model training

```

t1←GaussianBlur (kernel_size= (3, 3), sigma = (0.1,2))
t2←GaussianNoise (mean=0, std=0.05)
t3←ColorJitter (brightness=0.1)
t4←ColorJitter (contrast=0.1)
t5←ColorJitter (saturation=0.1)
random_choice←RandomChoice([t1, t2, t3, t4, t5])
transforms←RandomApply([random_choice], p=0.7)
    
```

3.6. Explainability

Explainability is vital in healthcare since experts should understand the reason behind the model’s predictions before considering or counting them. Therefore, we combined the model’s attention maps and the AAL3 brain atlas to discover the most impactful brain regions on the model’s conclusions. During the inference mode, our model follows these steps to provide various details to the user:

- The model extracts the attention map of each input scan and overlays this data on the original image. The outcome of this step is a heatmap of the brain regions,

in which pixel values correspond to their influence on the model’s decision

- The model illuminates pixels with values exceeding 95% of the maximum value using red rectangles. This step enables the user to examine and analyze all key areas in the input scan
- Ultimately, the model overlays the heatmap, extracted in the first step, on the AAL3 atlas and locates pixels with the highest intensity to provide the name of the brain regions that encompass them. Providing these areas’ names is crucial to the user since they are most influential in the model’s prediction.

We reshaped the AAL3 atlas to $3 \times 950 \times 570$ to fit the size of our input scans using the following procedure; the result of which is in Figure 4:

- By overlaying the AAL3 atlas on a pre-processed sample in MRICron,³⁸ we first reshaped AAL3 to $79 \times 95 \times 79$
- It was crucial to verify that both the reshaped atlas and the input scan followed the same coordinate system. Therefore, we loaded the resulting new atlas and the input scan into MRICron again and compared their coordinate system side-by-side. This step ensured that corresponding coordinates referred to the same brain area in both files
- We discarded the first ten and last nine slices from AAL3, similar to the input scans, resulting in a shape

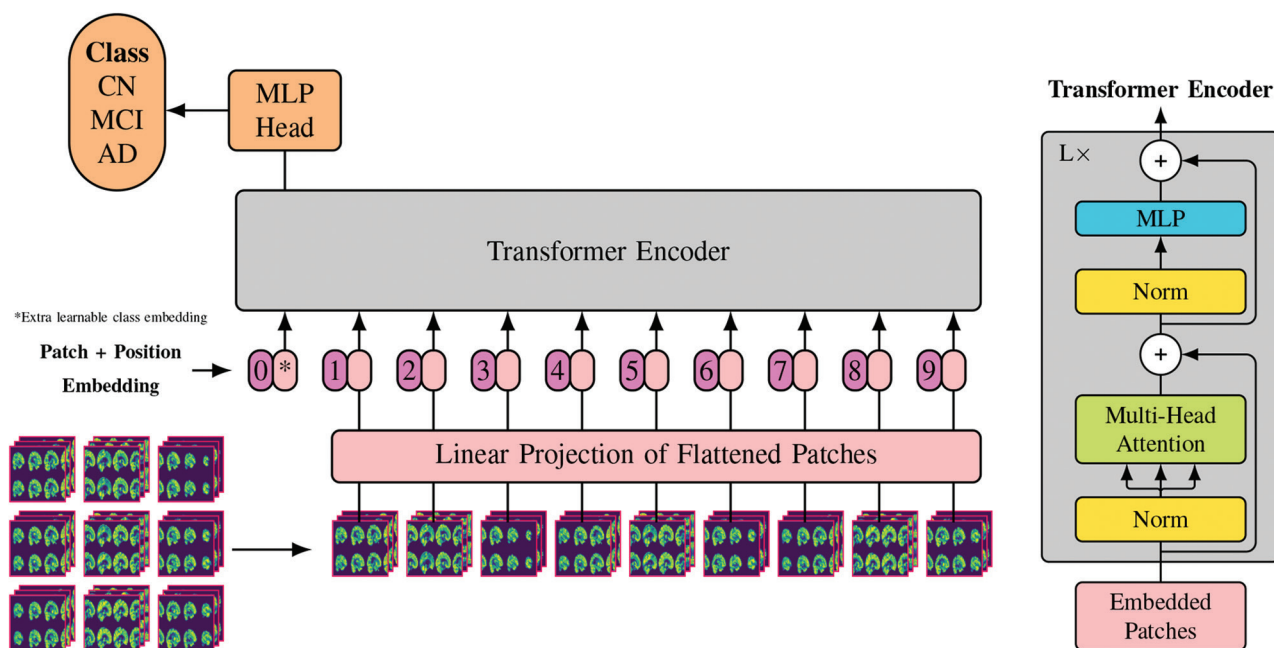


Figure 3. The model architecture is identical to the ViT-Base introduced by Dosovitskiy *et al.*⁶ First, the scan is reshaped into $3 \times 384 \times 384$ to fit the model’s input. Then, it is split into patches of shape $3 \times 32 \times 32$, flattened, and provided to a standard transformer along with position embeddings that contain spatial information. At the last stage, an MLP acts as the classification head to map the final hidden state into the probability of classes. The illustration of the model’s architecture was inspired by Dosovitskiy *et al.*⁶ Abbreviations: AD: Alzheimer’s disease; CN: Cognitively normal; MCI: Mild cognitive impairment; MLP: Multilayer perceptron.

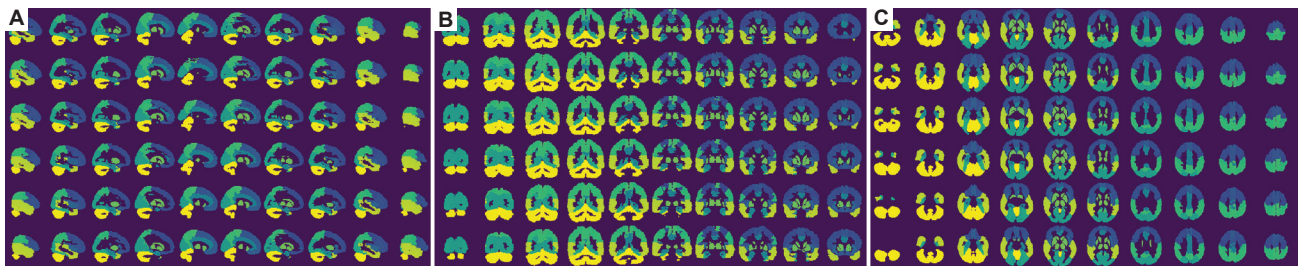


Figure 4. The resized AAL3 brain atlas ($3 \times 950 \times 570$), where channels illustrate regions from three different perspectives. Reshaping the atlas was a critical step since its dimensions should match that of input scans. Each color indicates a different brain area. (A) Sagittal, (B) Coronal, (C) Axial. Abbreviation: AAL3: Automated Anatomical Atlas 3.

of $60 \times 95 \times 79$

- The final stage entailed projecting AAL3 into three channels along different axes, resulting in an image of shape $3 \times 950 \times 570$.

3.7. Regions’ importance study

A vital component of our research was identifying the most critical brain regions to the model’s predictions. Apart from achieving better explainability, this study can help researchers and clinicians pay special attention to these key areas during their examinations.

As mentioned, our model utilizes the AAL3 atlas to provide the attention map and the most critical brain region for each input scan. Therefore, we conducted our study in the following manner:

- We combined the training, validation, and test sets to form a dataset of 580 scans
- After feeding all samples to the model, we saved the attention maps and suggested critical regions for each correctly classified scan in a database
- We considered the occurrence rate of each region in the database as a metric to show its importance in the model’s predictions
- Finally, we calculated the mean of all attention maps to generate a heatmap of brain areas.

4. Results

4.1. Model performance

Figure 5 illustrates the confusion matrix of the model’s predictions. The model performed the best in distinguishing between CN and AD cases with no error. However, classifying MCI cases was challenging for the model, similar to human experts, due to their prodromal state. Specifically, differentiating MCI and AD cases needed the most enhancement with an error of 25%. This error might be due to selecting the last scan of each MCI case that progressed into AD later, which made distinguishing these two classes more challenging. In addition, Table 3 demonstrates the performance of the proposed model in detail using several

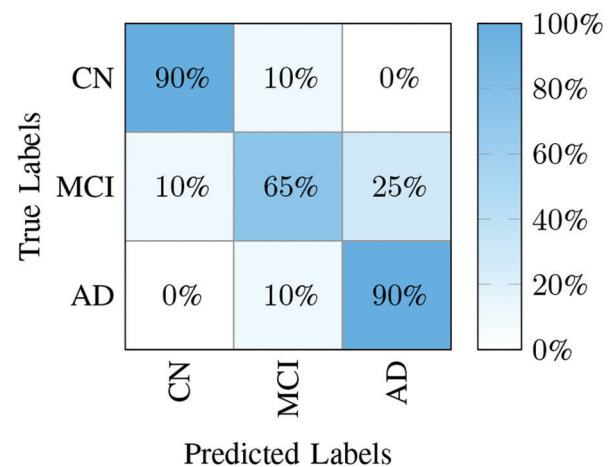


Figure 5. The model’s confusion matrix, illustrating its performance on the test dataset, with values normalized over true labels. The model can perfectly distinguish CN and AD cases with no error. However, classifying MCI is challenging due to its prodromal nature. Abbreviations: AD: Alzheimer’s disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

Table 2. Specifications of the proposed vision transformer

Parameter	Value
Input shape	$3 \times 384 \times 384$
Patch size	$3 \times 32 \times 32$
Layers	12
Hidden size	768
Multilayer perceptron size	3072
Heads	12
Hidden dropout	0.1

metrics. Similar to the confusion matrix, this table reveals the challenge of classifying MCI cases.

Finally, to comprehend the model’s representation of the learned data, we conducted a principal component analysis (PCA) on the last hidden state before SoftMax. Figure 6 illustrates the results of this analysis for the training and test datasets individually.

Table 3. The performance of the proposed model per class using different classification metrics

Class	Sensitivity	Specificity	Precision	F1 score	Accuracy (95% CI)
CN	0.90	0.95	0.90	0.90	82% (72%, 91%)
MCI	0.65	0.90	0.76	0.70	
AD	0.90	0.88	0.78	0.84	
Macro-average	0.82	0.91	0.81	0.81	

Abbreviations: AD: Alzheimer’s disease; CI: Confidence interval; CN: Cognitively normal; MCI: Mild cognitive impairment.

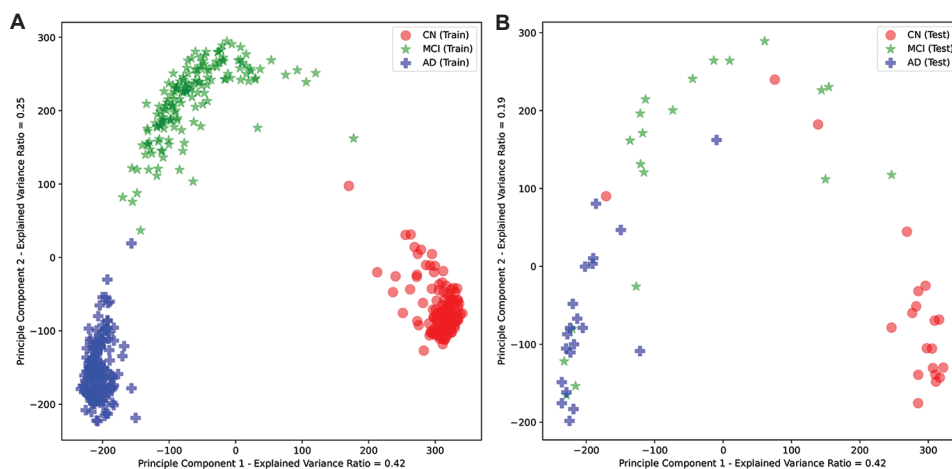


Figure 6. The result of dimensionality reduction on the model’s last hidden state before SoftMax using PCA with two principal components. This analysis is beneficial in gaining insight into the learned representation of data and the model’s performance in distinguishing between classes. The figure illustrates the true labels. (A) Train, (B) Test.

Abbreviations: AD: Alzheimer’s disease; CN: Cognitively normal; MCI: Mild cognitive impairment; PCA: Principal component analysis.

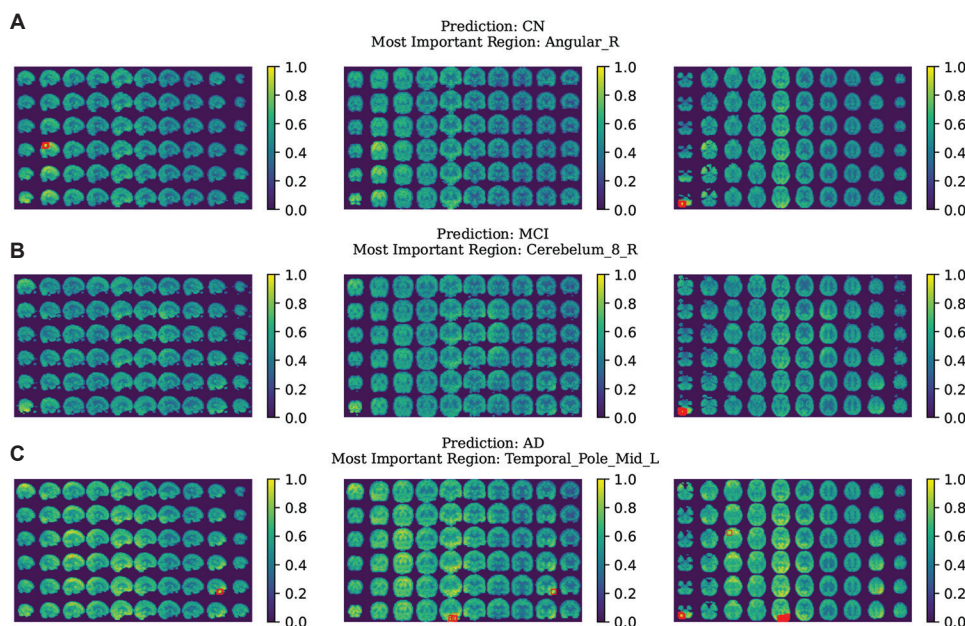


Figure 7. The model’s inference output for three correctly classified samples. Our model provides extra information during inference to obtain explainability and to assist the user in making a diagnosis. (A) CN, (B) MCI, (C) AD.

Abbreviations: AD: Alzheimer’s disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

4.2. Explainability

Figure 7 illustrates the prediction results of three correctly classified scans. As depicted, the model provides the following information to the user in the inference mode:

- The predicted label
- The brain region that has the most influence on the model’s prediction. This information results from locating the pixel with the highest intensity value in an overlay of the attention map and the AAL3 brain atlas
- An overlay of the attention map and the input scan, in which the brightness of each pixel is analogous to its significance in the model’s conclusion. Red rectangles also illustrate regions with attention values greater than 95% of the maximum attention.

In addition to the predicted label, this information enables domain experts to find out the model’s logic and examine the brain’s key areas further.

4.3. Regions’ importance study

Figure 8 illustrates the overall importance of different regions and for predicting each label independently. Please note that Figure 8 only contains the AAL3 regions that our model suggested as crucial at least once, ignoring all other areas without any occurrence during inference. Our model suggests the angular gyrus, known to be heavily affected by MCI and AD,³⁹⁻⁴¹ as the most critical region in

distinguishing the CN class. In addition, the temporal pole is the key area in classifying AD, aligning with previous studies that found all AD patients experience atrophy and other complications in this brain region.⁴² Finally, the proposed model defines the cerebellum as the essential area for MCI classification. Traditionally, this part of the brain did not play a pivotal role in diagnosing NDDs.⁴³ However, recent studies have revealed the significance of the cerebellum in diagnosing MCI and various stages of AD.⁴³ Further investigations also indicate that AD progression causes cerebellar transformations, and this region is central to obtaining significantly better performance in classification tasks.⁴⁴

Figure 9 shows the brain heatmaps, where the brightness of a pixel signifies its impact on the model’s decisions. As indicated in both figures, some brain regions play a substantial role in diagnosing various classes.

5. Discussion

Affecting millions of lives, NDDs are a leading cause of death and disability worldwide.¹ Although remaining mostly incurable, early diagnosis of such conditions is a key to better disease management and enhancing the patient’s quality of life.²

Diagnosing NDDs is challenging, even for proficient nuclear medicine physicians, and requires substantial

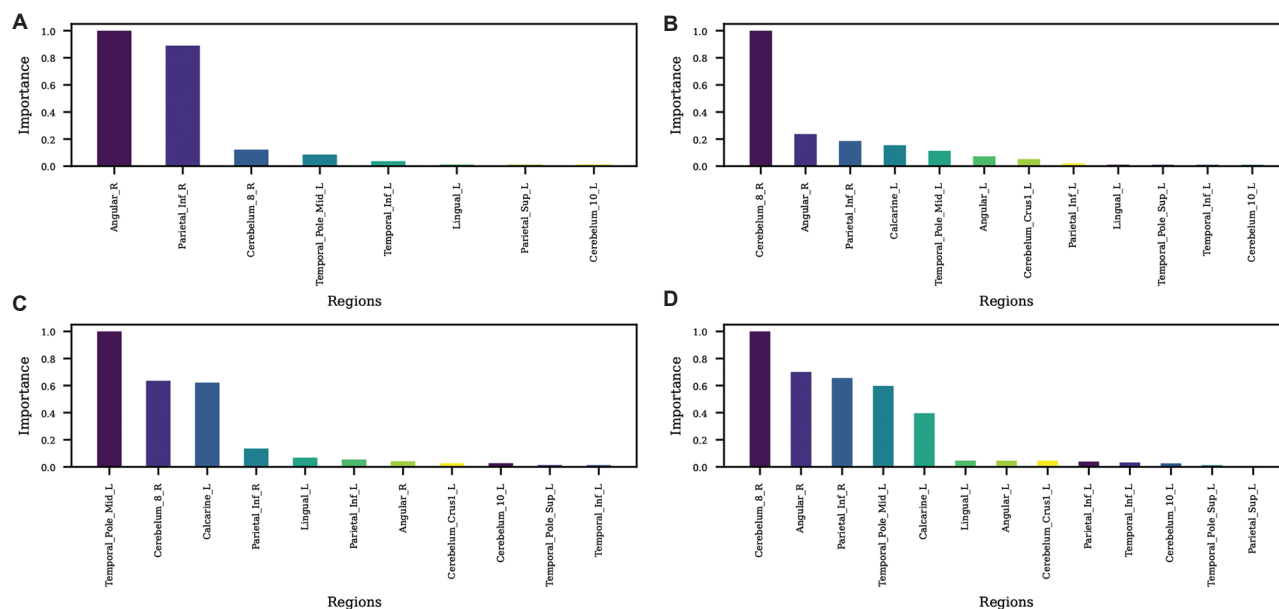


Figure 8. The significance of the AAL3 regions in predicting each class and their overall contributions during our regions’ importance study. After combining the training, validation, and test datasets, we fed the resulting dataset of 580 samples to the model and saved the suggested crucial region for correctly classified scans. Then, we considered the occurrence rate of each region as a metric to show its importance in the model’s diagnoses. Please note we only included the areas suggested by the model as critical at least once, ignoring all other parts without any occurrence during inference. (A) CN, (B) MCI, (C) AD, (D) Overall. Abbreviations: AAL3: Automated Anatomical Atlas 3; AD: Alzheimer’s disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

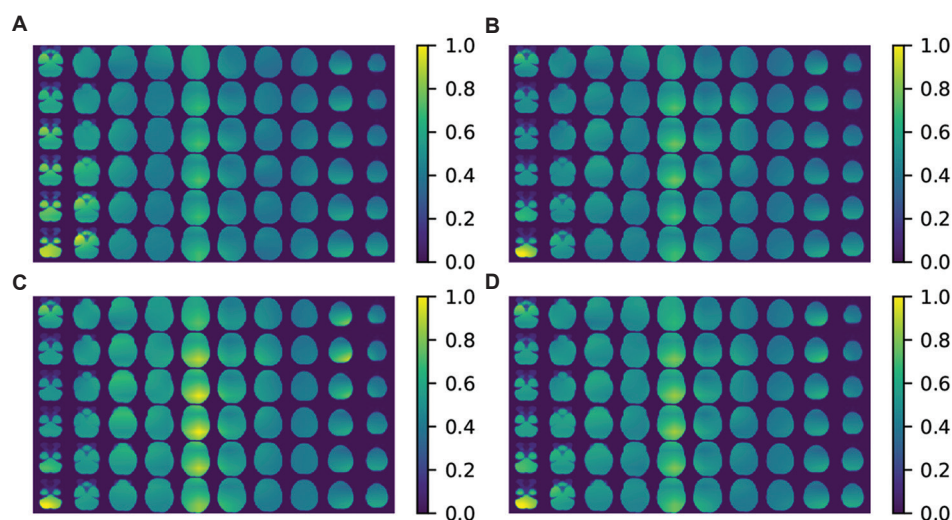


Figure 9. The mean attention maps of brain areas in our regions' importance study, where the brightness of a pixel translates to its significance in the model's predictions. This figure results from merging the training, validation, and test datasets and feeding the final dataset of 580 samples to the model. Then, we saved the model's attention maps for all correctly classified scans. Finally, we computed the mean of saved attention maps to generate a mean attention map for the whole dataset and each class. (A) CN, (B) MCI, (C) AD, (D) Overall.

Abbreviations: AD: Alzheimer's disease; CN: Cognitively normal; MCI: Mild cognitive impairment.

knowledge and training. Therefore, this complexity in disease diagnosis, together with other factors, has left about 75% of NDD patients undiagnosed worldwide, and this number rises to about 90% in low- and middle-income countries, according to some studies.³ In addition to revealing a considerable diagnosis gap, these investigations predict that the fast-growing number of NDD patients could strain healthcare systems in the future.³ Consequently, novel and affordable tools and techniques are required for the final diagnosis of NDD and/or for assisting healthcare providers in this task.

While the rapid progress of AI and its sub-fields revolutionized our lives, researchers have attempted to harness the power of these new technologies in the healthcare domain, including NDD diagnosis. Although most of these research projects utilized long-established approaches and architectures like CNNs, the emergence of ViTs and their groundbreaking performance convinced us to explore the potential of employing this new architecture in NDD classification.

In this work, we developed a model to classify ¹⁸F-FDG PET brain scans into CN, MCI, and AD. Specifically, we designed the model based on the vanilla ViT-Base, introduced by Dosovitskiy *et al.*,⁶ and trained on the ADNI dataset.⁷ Combining the proposed data, pre-processing procedure, training recipe, and transfer learning enabled our model to achieve an F1 score of 81% (macro-average of all classes), significantly outperforming previous approaches. To comprehend the model's performance, we

also compare our results with those presented in other papers, as listed in Table 4. However, a few points worth mentioning for a fair comparison:

- While several papers examined NDD classification using DNNs, we exclusively selected studies that aimed for ternary classification (CN/MCI/AD) using ¹⁸F-FDG PET scans
- Although all chosen studies employed ADNI as their primary dataset, the authors may have used different subsets to train and test the models
- As shown in Table 4, DNNs can surpass physicians when NDD diagnosis is solely based on brain scans. However, domain experts usually consider a comprehensive collection of information for the clinical diagnosis, including the patient's medical history, genetics, blood tests, and cognitive and physical evaluations. Consequently, instead of solely relying on brain scans, doctors and nuclear medicine physicians consider various factors, a practice that poses a substantial advantage over AI models.

According to Table 4, in addition to a significantly higher F1 score, our model excels in distinguishing MCI cases, which proved to be the most challenging condition to classify, compared to human experts and other models.

Apart from improving the model's performance, developing an explainable model was a pivotal goal of this research. Therefore, we integrated the AAL3 brain atlas information into the attention maps. This method resulted in a model that provides the brain region with the highest

Table 4. The performance comparison of our model with others in the literature

Model	Class	Sensitivity	Specificity	Precision	F1 score	F1 score (micro-average)
Ding <i>et al.</i> ⁵ (model CNN)	CN	0.59	0.75	0.60	0.59	0.64
	MCI	0.54	0.68	0.55	0.55	
	AD	0.81	0.94 ^a	0.76	0.78	
Etminani <i>et al.</i> ⁴ (model CNN)*	CN	0.88	0.90	0.81	0.84	0.63
	MCI	0.17	0.94 ^a	0.20	0.18	
	AD	0.91 ^a	0.92	0.83 ^a	0.87 ^a	
Etminani <i>et al.</i> ⁴ (consensus of human readers)* [†]	CN	0.70	0.81	0.64	0.67	0.45
	MCI	0.25	0.75	0.08	0.12	
	AD	0.47	0.90	0.68	0.56	
Our model (ViT)	CN	0.90 ^a	0.95 ^a	0.90 ^a	0.90 ^a	0.81 ^a
	MCI	0.65 ^a	0.90	0.76 ^a	0.70 ^a	
	AD	0.90	0.88	0.78	0.84	

Notes: All results were obtained using ¹⁸F-FDG PET brain scans from the ADNI dataset. Besides acquiring a significantly higher F1 score, our model outperforms others in classifying MCI cases by a considerable margin. *The authors considered an additional DLB class in their paper and tested the model and human readers in a four-class classification task (CN, MCI, AD, DLB). [†]Four professional nuclear medicine physicians with 3, 8, 13, and 16 years of experience. ^aThe highest values per each metric and class.

Abbreviations: AD: Alzheimer’s disease; ADNI: Alzheimer’s Disease Neuroimaging Initiative; CN: Cognitively normal; CNN: Convolutional neural network; DLB: Dementia with Lewy bodies; ¹⁸F-FDG: ¹⁸F-fluorodeoxyglucose; MCI: Mild cognitive impairment; PET: Positron emission tomography; ViT: Vision transformer.

impact on its decision and a heatmap of the input scan, in which the pixels’ intensities illustrate their importance in the model’s prediction. Furthermore, we conducted a study to reveal regions’ significance in the model’s decisions, showing some brain areas are of utmost importance in predicting various conditions.

5.1. Limitations

A key goal of this study was to ensure that the model can distinguish and classify all stages of AD. Therefore, we decided to develop the model solely on MCI cases that later progressed to AD. However, this sample selection might introduce some bias in the dataset regarding the MCI class and result in the model becoming a prognostic MCI-to-AD classifier. In addition, we exclusively relied on the ADNI dataset in this research, which may restrict the model’s out-of-distribution generalization.

Although the critical regions suggested by the proposed model are consistent with the literature to a significant extent, our findings require more examination and validation by medical domain experts.

Finally, we should stress that while DNNs demonstrate promising results in NDD classification, they are limited to their datasets, substantially affecting their generalization performance. In addition, unlike clinical procedures, DNNs do not consider all medical factors and base their predictions on limited data. Therefore, vast improvements and training on diverse datasets are critical to designing robust and clinically

applicable models that can assist experts in NDD diagnosis.

6. Conclusion

We believe this research showcases the extraordinary potential of the ViT architecture in NDDs classification, which surpasses other methods, including CNNs. Apart from their excellent performance, ViTs allow computer vision researchers to benefit from advancements in NLP due to the sharing of the same transformer architecture. Furthermore, ViTs make developing explainable models more feasible by leveraging the attention mechanism.

Acknowledgments

None.

Funding

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Dec 5, 2024 12:30 PM Genentech,

Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research provided funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<https://www.fnih.org/>). The grantee organization is the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California. Farzaneh Etmnani and Amira Soliman are supported by Center for Applied Intelligent Systems Research in Health (CAISR Health) funded by Knowledge Foundation (grant no.: 20200208 01H).

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: All authors

Formal analysis: Pooriya Khorramyar, Amira Soliman

Investigation: All authors

Methodology: Pooriya Khorramyar, Amira Soliman

Writing—original draft: Pooriya Khorramyar

Writing—review & editing: All authors

Ethics approval and consent to participate

Please visit the ADNI documentation at <https://adni.loni.usc.edu/help-faqs/adni-documentation/>.

Consent for publication

Please visit the ADNI documentation at <https://adni.loni.usc.edu/help-faqs/adni-documentation/>.

Availability of data

To download the dataset, please visit the ADNI's web page at <https://adni.loni.usc.edu/>. The code is available on the project's GitHub repository (https://github.com/Pooriya-Kh/NDD_ViT).

References

- World Health Organization (WHO). *Dementia Key Facts*. Available from: <https://www.who.int/news-room/fact-sheets/detail/dementia> [Last accessed on 2025 Mar 30].
- Alzheimer Society of Canada. *The 10 Benefits of Early Diagnosis*. Available from: <https://alzheimer.ca/en/about/dementia/do-i-have/dementia/how/get/tested/dementia/tips/individuals-families-friends/10> [Last accessed on 2025 Mar 30].
- Alzheimer's Disease International. *Over 41 Million Cases of Dementia go Undiagnosed Across the Globe - World Alzheimer Report Reveals*. Available from: <https://www.alzint.org/news/events/news/over/41/million/cases/of/dementia/go/undiagnosed-across-the-globe-world-alzheimer-report-reveals> [Last accessed on 2025 Mar 30].
- Etmnani K, Soliman A, Davidsson A, *et al.* A 3D deep learning model to predict the diagnosis of dementia with lewy bodies, Alzheimer's disease, and mild cognitive impairment using brain ¹⁸F-FDG PET. *Eur J Nucl Med Mol Imaging*. 2022;49(2):563-584.
doi: 10.1007/s00259-021-05483-0
- Ding Y, Sohn JH, Kawczynski MG, *et al.* A deep learning model to predict a diagnosis of alzheimer disease by using ¹⁸F-FDG PET of the brain. *Radiology*. 2019;290(2):456-464.
doi: 10.1148/radiol.2018180958
- Dosovitskiy A, Beyer L, Kolesnikov A, *et al.* *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. [ArXiv Preprint]; 2020.
- The Alzheimer's Disease Neuroimaging Initiative*. *Alzheimer's Disease Neuroimaging Initiative (ADNI)*. Available from: <https://adni.loni.usc.edu> [Last accessed on 2025 Mar 30].
- Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. *Advances in Neural Information Processing Systems*. 2017. p. 30.
- Nobili F, Arbizu J, Bouwman F, *et al.* European association of nuclear medicine and European academy of neurology recommendations for the use of brain ¹⁸F-fluorodeoxyglucose positron emission tomography in neurodegenerative cognitive impairment and dementia: Delphi consensus. *Eur J Neurol*. 2018;25(10):1201-1217.
doi: 10.1111/ene.13728
- Mayo Foundation for Medical Education and Research (MFMER). *Positron Emission Tomography Scan*. Available from: <https://www.mayoclinic.org/tests/procedures/pet/scan/about/pac-20385078> [Last accessed on 2025 Mar 30].
- Cleveland Clinic. *PET Scan*. Available from: <https://my.clevelandclinic.org/health/diagnostics/10123-pet-scan> [Last accessed on 2025 Mar 30].
- Rolls ET, Huang CC, Lin CP, Feng J, Joliot M. Automated anatomical labelling atlas 3. *Neuroimage*. 2020;206:116189.
doi: 10.1016/j.neuroimage.2019.116189
- Mayo Foundation for Medical Education and Research (MFMER). *Mild Cognitive Impairment (MCI)*. Available

- from: <https://www.mayoclinic.org/diseases/conditions/mild/cognitive/impairment/symptoms/causes/syc-20354578> [Last accessed on 2025 Mar 30].
14. Simonyan K, Zisserman A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [ArXiv Preprint]; 2014.
 15. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*. Berlin: Springer; 2014. p. 818–833.
 16. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016. p. 2818–2826.
 17. Simonyan K, Vedaldi A, Zisserman A. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. [ArXiv Preprint]; 2013.
 18. Lozupone G, Bria A, Fontanella F, Meijer FJ, De Stefano C. *AXIAL: Attention-Based Explainability for Interpretable Alzheimer’s Localized Diagnosis using 2D CNNs on 3D MRI Brain Scans*. [ArXiv Preprint]; 2024.
 19. Khatri U, Kwon GR. Explainable vision transformer with self-supervised learning to predict Alzheimer’s disease progression using 18F-FDG PET. *Bioengineering (Basel)*. 2023;10(10):1225.
doi: 10.3390/bioengineering10101225
 20. Shin H, Jeon S, Seol Y, Kim S, Kang D. Vision transformer approach for classification of Alzheimer’s disease using 18F-florbetaben brain images. *Appl Sci*. 2023;13(6):3453.
doi: 10.3390/app13063453
 21. The National Institutes of Health. *How Biomarkers Help Diagnose Dementia*. Available from: <https://www.nia.nih.gov/health/alzheimers/symptoms/and/diagnosis/how/biomarkers-help-diagnose-dementia> [Last accessed on 2025 Mar 30].
 22. Xing X, Liang G, Zhang Y, Khanal S, Lin AL, Jacobs N. Advit: vision transformer on multi-modality pet images for Alzheimer disease diagnosis. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. United States: IEEE; 2022. p. 1–4.
doi: 10.1109/ISBI52829.2022.9761584
 23. Odusami M, Maskeliūnas R, Damaševičius R. Pixel-level fusion approach with vision transformer for early detection of Alzheimer’s disease. *Electronics*. 2023;12(5):1218.
doi: 10.3390/electronics12051218
 24. Lyu Y, Yu X, Zhu D, Zhang L. Classification of Alzheimer’s disease via vision transformer: Classification of Alzheimer’s disease via vision transformer. In: *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments. PETRA ’22*. United States: Association for Computing Machinery; 2022. p. 463–468.
doi: 10.1145/3529190.3534754
 25. Sarraf S, Sarraf A, DeSouza DD, Anderson JAE, Kabia M, The Alzheimer’s Disease Neuroimaging Initiative. OViTAD: Optimized vision transformer to predict various stages of Alzheimer’s disease using resting-State fMRI and structural MRI data. *Brain Sci*. 2023;13(2):260.
doi: 10.3390/brainsci13020260
 26. Hoang GM, Kim UH, Kim JG. Vision transformers for the prediction of mild cognitive impairment to Alzheimer’s disease progression using mid-sagittal sMRI. *Front Aging Neurosci*. 2023;15:1102869.
doi: 10.3389/fnagi.2023.1102869
 27. Aghdam MA, Bozdog S, Saeed F, Alzheimer’s Disease Neuroimaging Initiative. PVTAD: Alzheimer’s disease diagnosis using pyramid vision transformer applied to white matter of T1-weighted structural MRI data. *Proc IEEE Int Symp Biomed Imaging*. 2024;2024:10.
doi: 10.1109/isbi56570.2024.10635541
 28. Wang W, Xie E, Li X, *et al*. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. United States: IEEE; 2021. p. 568–578.
 29. Kushol R, Masoumzadeh A, Huo D, Kalra S, Yang YH. Addformer: Alzheimer’s disease detection from structural MRI using fusion transformer. In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE; 2022. p. 1–5.
doi: 10.1109/ISBI52829.2022.9761421
 30. Shah SMAH, Khan MQ, Rizwan A, Jan SU, Samee NA, Jamjoom MM. Computer-aided diagnosis of Alzheimer’s disease and neurocognitive disorders with multimodal Bi-vision transformer (BiViT). *Pattern Anal Appl*. 2024;27(3):76.
doi: 10.1007/s10044-024-01297-6
 31. *The Alzheimer’s Disease Neuroimaging Initiative. ADNI Documentation*. Available from: <https://adni.loni.usc.edu/help-faqs/adni-documentation> [Last accessed on 2025 Mar 30].
 32. The MathWorks Inc. *Matlab R*; 2016a. Available from: <https://www.mathworks.com> [Last accessed on 2025 Mar 30].
 33. UCL Queen Square Institute of Neurology. *Statistical Parametric Mapping*. Available from: <https://www.fil.ion.ucl.ac.uk/spm> [Last accessed on 2025 Mar 30].
 34. Wolf T, Debut L, Sanh V, *et al*. *HuggingFace’s Transformers: State-of-the-Art Natural Language Processing*; 2020. Available from: <https://arxiv.org/abs/1910.03771> [Last accessed on

- 2025 Mar 30].
35. Ridnik T, Baruch EB, Noy A, Zelnik-Manor L. *ImageNet-21K Pretraining for the Masses*. CoRR. 2021. Available from: <https://arxiv.org/abs/2104.10972> [Last accessed on 2025 Mar 30].
 36. Russakovsky O, Deng J, Su H, *et al*. ImageNet large scale visual recognition challenge. *Int J Comput Vis IJCV*. 2015;115(3):211-252.
doi: 10.1007/s11263-015-0816-y
 37. Google. *Vision Transformer (Base-Sized Model)*. Available from: <https://huggingface.co/google/vit-base-patch32-384> [Last accessed on 2025 Mar 30].
 38. Rorden C, Brett M. Stereotaxic display of brain lesions. *Behav Neurol*. 2000;12(4):191-200.
doi: 10.1155/2000/421719
 39. Li Y, Wang X, Li Y, *et al*. Abnormal resting-state functional connectivity strength in mild cognitive impairment and its conversion to Alzheimer's disease. *Neural Plast*. 2016;2016:4680972.
doi: 10.1155/2016/4680972
 40. Talwar P, Kushwaha S, Chaturvedi M, Mahajan V. Systematic review of different neuroimaging correlates in mild cognitive impairment and Alzheimer's disease. *Clin Neuroradiol*. 2021;31(4):953-967.
doi: 10.1007/s00062-021-01057-7
 41. Salmon E, Collette F, Bastin C. Cerebral glucose metabolism in Alzheimer's disease. *Cortex*. 2024;179:50-61.
doi: 10.1016/j.cortex.2024.07.004
 42. Arnold SE, Hyman BT, Van Hoesen GW. Neuropathologic changes of the temporal pole in Alzheimer's disease and pick's disease. *Arch Neurol*. 1994;51(2):145-150.
doi: 10.1001/archneur.1994.00540140051014
 43. Yang C, Liu G, Chen X, Le W. Cerebellum in Alzheimer's disease and other neurodegenerative diseases: An emerging research frontier. *MedComm (2020)*. 2024;5(7):e638.
doi: 10.1002/mco2.638
 44. Bruchhage MMK, Correia S, Malloy P, Salloway S, Deoni S. Machine learning classification identifies cerebellar contributions to early and moderate cognitive decline in Alzheimer's disease. *Front Aging Neurosci*. 2020;12:524024.
doi: 10.3389/fnagi.2020.524024

ORIGINAL RESEARCH ARTICLE

Comparison of synthetic data generation techniques for obesity level prediction based on dietary habits and physical status

Hakan Alp Eren¹, Halil İbrahim Emek², and Sinem Bozkurt Keser^{2*}¹Department of Software Engineering, Faculty of Engineering and Architecture, Eskişehir Osmangazi University, Eskişehir, Türkiye²Department of Computer Engineering, Faculty of Engineering and Architecture, Eskişehir Osmangazi University, Eskişehir, Türkiye**Abstract**

In the contemporary context of the obesity epidemic and its associated comorbidities, early detection of individuals at risk is critical. Artificial intelligence and machine learning techniques offer substantial potential for automating obesity risk assessment, enabling early diagnosis and intervention. However, the development of robust predictive models is often hampered by limited or imbalanced datasets. Synthetic data generation has emerged as a key solution, allowing the expansion and balancing of data while preserving privacy. Recent surveys highlight that the synthetic minority oversampling technique (SMOTE) is a leading method for data generation in obesity detection. In line with this, our study analyzed the Estimation of Obesity Levels dataset, a dataset from the University of California, Irvine repository, focused on dietary habits and physical condition, which suffers from class imbalance. We compared three synthetic data generation approaches: SMOTE—nominal and continuous, variational autoencoders, and conditional tabular generative adversarial network. We trained multiple classifiers on the generated datasets and evaluated their performance. Classifiers trained on data including height and weight (i.e., body mass index [BMI]-related features) achieved F1-scores of up to 98.16%, as expected due to the direct role of BMI in obesity classification. Crucially, models trained without height and weight still achieved an F1-score of 74.48% when synthetic augmentation was used, demonstrating that useful obesity prediction models can be developed even in the absence of explicit anthropometric measures. These results indicate that synthetic data can enable accurate classification when key features are missing or when data are scarce.

Keywords: Obesity; Synthetic data; Tabular data; Data augmentation; Machine learning; Class imbalance***Corresponding author:**Sinem Bozkurt Keser
(sbozkurt@ogu.edu.tr)**Citation:** Eren HA, Emek Hİ, Keser SB. Comparison of synthetic data generation techniques for obesity level prediction based on dietary habits and physical status. *Artif Intell Health*. 2025;2(4):47-74. doi: 10.36922/AIH025140027**Received:** April 1, 2025**Revised:** June 2, 2025**Accepted:** June 10, 2025**Published online:** June 25, 2025**Copyright:** © 2025 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**

According to the World Health Organization, obesity is defined as the accumulation of fat in the body to an extent that impairs health. The rise in obesity rates has become a growing concern not only in high-income countries but also in middle- and low-income nations.¹ The increasing prevalence of obesity across all age groups is linked to

health issues such as hypertension, diabetes, certain types of cancer, and musculoskeletal disorders.² Globalization has transformed obesity into a global public health challenge, demanding attention and coordinated action in the international policy arena.³ Obesity is a major, yet preventable, global health condition, with a high and rising prevalence among children and adolescents, leading to serious health complications and substantial healthcare costs.⁴ In light of these circumstances, early diagnosis of obesity becomes critically important. By analyzing individual characteristics, it is possible to predict an individual's risk of developing obesity.

Artificial intelligence (AI) methodologies hold great promise for automating obesity risk estimation by enabling early diagnosis and timely intervention.⁵ For example, predictive models based on dietary and lifestyle features can identify individuals at elevated risk even before clinical obesity manifests. However, real-world datasets (especially survey-based ones) are often small, imbalanced, or contain missing values. Synthetic data generation provides a means to overcome these challenges by creating artificial records that replicate the statistical properties of real data.^{6,7} This can improve model training and generalization without compromising patient privacy.⁸

Recent evidence underscores the multifactorial nature of obesity. For instance, early-life nutrition and feeding practices have long-term effects on weight trajectories: exclusive breastfeeding is associated with a lower risk of childhood overweight and obesity.⁹ Studies of adult obesity also emphasize the role of dietary patterns and psychosocial factors. Sobas *et al.*¹⁰ identified distinct dietary patterns (“prudent” healthy diet versus “fast food & alcohol”) among bariatric surgery candidates, with the latter linked to more severe obesity.¹⁰ Colonnello *et al.*¹¹ found that dysfunctional eating behaviors (e.g., night eating, food cravings) correlate with lipid and metabolic abnormalities in obese patients.¹¹ El-Sehrawy *et al.*¹² showed that a high triglyceride-glucose (TyG) index (a marker of insulin resistance) is associated with adverse lipid profiles and disordered eating in obesity.¹² Psychological stress is also implicated – Kuckuck *et al.*¹³ demonstrated that long-term stress (measured by hair cortisone) is associated with hedonic eating tendencies in obese individuals.¹³ Together, these studies highlight that beyond anthropometric measures, a combination of diet quality, metabolic markers, and behavioral patterns influence obesity outcomes. This study does not attempt to discover new causal factors; rather, it focuses on the methodological contribution of using synthetic data to improve obesity prediction models based on available dietary/behavioral features. Specifically, it examines the extent to which predictive accuracy can

be maintained when key features (height and weight) are unavailable.

A review of the literature reveals that synthetic data generation is widely applied, with the synthetic minority oversampling technique (SMOTE) being one of the most commonly used approaches.¹⁴⁻¹⁷ In line with this, the present study analyzes the Estimation of Obesity Levels Based on Eating Habits and Physical Status (EOL) dataset, which suffers from an unbalanced class distribution. To address the issue of limited sample size, various synthetic data generation techniques were employed.^{18,19} Furthermore, an AI system was developed using machine learning (ML) algorithms to estimate obesity levels based on individuals' eating habits and physical status.

The performance of ML models trained on data generated using different techniques – namely variational autoencoders (VAE), generative adversarial network (GAN), and SMOTE – nominal and continuous (SMOTE-NC)—was compared. The next section of this paper presents a summary of related work, including the datasets used, methodologies applied, and results reported. The Materials and Methods section describes the dataset characteristics, synthetic data generation approaches, preprocessing procedures, and interrelationships among attributes. It also details the ML algorithms and evaluation methodology used to assess model effectiveness. The Results and Discussion section presents model outputs through various graphs and tables. Finally, the manuscript concludes with a summary and suggestions for future research.

2. Literature review

Numerous studies in the literature have addressed the problem of obesity detection, with particular emphasis on dataset construction and the development of ML models. [Table 1](#) summarizes the key characteristics of the datasets used in these studies, the ML techniques applied, and the corresponding performance metrics reported.

Palechor *et al.*¹⁴ developed a dataset for obesity level classification using data collected from individuals in Mexico, Peru, and Colombia. The dataset comprises 17 attributes related to eating habits and physical condition. Of the 2,111 instances, 23% were collected directly from users via a web platform, while the remaining 77% were synthetically generated using SMOTE. Classification of individuals was based on their body mass index (BMI) values. Subjects with a BMI below 18.5 were categorized as underweight, those with values between 18.5 and 24.9 as normal weight, and those between 25 and 29.9 as overweight. BMI values of 30 and above indicated obesity, which was further divided into three classes: 30–34.9 as

Table 1. Summary of literature on obesity risk prediction using machine learning (ML) algorithms

Study	Dataset	ML algorithm	Results
Helforoush and Sayyad ¹⁵	UCI Obesity dataset (2,111 samples; 17 features)	ANN + PSO hybrid; compared with baseline regression	The ANN-PSO model achieved an accuracy of 92%, outperforming standard regression methods. SHAP analysis identified weight and height as the most influential features
Ayub <i>et al.</i> ¹⁶	UCI Obesity dataset	Attention Bi-LSTM deep network	The proposed model achieved 96.5% accuracy in obesity classification, surpassing previous approaches. The integration of an attention mechanism enhanced the model's ability to capture feature influence
Shakti <i>et al.</i> ¹⁷	UCI Obesity dataset	Multiple comparisons: k-NN, SVM, RF, GBM, MLP	The MLP achieved the highest accuracy of 97.2%, followed by GB with ~96.2%. These results highlight the advantage of incorporating diverse features to improve classification performance
Yağmur ¹⁸	UCI Obesity dataset	DT + POA (hybrid model)	The hybrid DT-POA model with fuzzy tuning outperformed the baseline DT, demonstrating improved classification performance for obesity levels
Özkurt ¹⁹	UCI Obesity dataset	XGBoost, RF, NB, k-NN, DT (+ SHAP XAI)	XGBoost achieved the highest accuracy of 92%. SHAP analysis identified key predictors, including family history of obesity and vegetable intake
Wang ²⁰	UCI Obesity dataset (height/weight excluded)	Ordinal versus multinomial Logit; LogitBoost; SVM, NB, RF, k-NN	The LogitBoost model achieved the highest performance with ~70% accuracy (Kappa=0.65). Other ML models yielded accuracies ranging from 75% to 79%. The overall lower accuracy was attributed to the exclusion of BMI-related features. Nonetheless, active transportation (e.g., biking), and family history were identified as key predictors
Okpe <i>et al.</i> ²¹	UCI Obesity dataset	Multilayer perceptron ANN	A tuned ANN achieved 97% accuracy in multi-class obesity prediction, demonstrating that high accuracy can be attained with a relatively simple NN architecture
Azad <i>et al.</i> ²²	UCI Obesity dataset	Stacked ensemble (GBM, XGB, etc.) + LIME explanations	The stacking ensemble model achieved ~98% accuracy, outperforming previous models (~97.8%). Model explainability was enhanced through the integration of LIME
Solomon <i>et al.</i> ²³	UCI Obesity dataset	Hybrid voting ensemble (XGBoost + GBM + MLP)	The ensemble model achieved an accuracy of 97.16%, surpassing the single XGBoost model (~96.4%). These results set a high benchmark for future studies in obesity prediction
Kaur <i>et al.</i> ²⁴	UCI Obesity dataset	GB, BME, XGBoost, RF, SVM, k-NN	XGBoost achieved 97.79% accuracy with a 70 – 30 train-test split, followed by GBM with ~97.16%. The results demonstrated the superiority of ensemble methods. In addition, the model provided personalized diet recommendations based on predictive outcomes
Muliawan <i>et al.</i> ²⁵	Kaggle Obesity dataset (2,111 samples; 17 features)	RF	An accuracy of 81.76% was achieved using only eating habit parameters, validating the effectiveness of RF as a screening tool for obesity risk based solely on dietary data
Choudhuri <i>et al.</i> ²⁶	UCI Obesity dataset	Hybrid ML model (combining algorithms)	A hybrid approach was proposed for estimating obesity levels, combining multiple ML techniques. This method improved accuracy compared to individual models and has been cited in subsequent studies for its pioneering contribution
Cervantes and Palacio ²⁷	UCI Obesity dataset (original introduction)	Computational intelligence methods (e.g., ANN, fuzzy)	An early study achieved viable obesity level prediction, laying the groundwork for the application of ML on this dataset and serving as a baseline in later research
Ganie <i>et al.</i> ²⁸	Kaggle Obesity dataset (2,111 samples; 17 features)	Bagged DT, RF, extra tree, XGBoost, GB, CatBoost, voting classifier	The proposed model achieved 98.10% accuracy in obesity classification, outperforming previous approaches. The ensemble of boosting algorithms effectively captured complex patterns in lifestyle data
Nagarajan <i>et al.</i> ²⁹	UCI Obesity dataset	TabNet, XGBoost, RF, MLP, bagging, DT, SVM, k-NN, SGD, AdaBoost, stacking, GB	The proposed model achieved 99.3% accuracy in obesity classification, outperforming previous approaches. The use of SMOTE and deep learning techniques enhanced learning from imbalanced classes
Umoh <i>et al.</i> ³⁰	UCI Obesity dataset	KNN, SVM, bagging, stacking, voting, LR, DT, AdaBoost	The proposed model achieved 93.97% accuracy in obesity classification. Optimization through feature selection techniques improved the model's understanding of dietary and physical habits

(Contd...)

Table 1. (Continued)

Study	Dataset	ML algorithm	Results
Vairachilai <i>et al.</i> ³¹	Kaggle COVID-19 Healthy Diet dataset	Protein Food Item Prediction Regression model	The proposed model achieved high predictive accuracy, with MAPE of 29% for meat and milk and 31% for oil crops and vegetable products. The integration of protein-rich food variables allowed refined modeling of feature influence in obesity prediction
Forte <i>et al.</i> ³²	FITescola® project dataset	CNN	The proposed model achieved 75% accuracy in obesity classification. The inclusion of physical fitness variables improved feature interpretability and overall model performance
Yağın <i>et al.</i> ³³	Physical Activity and Eating Habits dataset from İnönü University; includes alcohol use, device use, and meal frequency	Trained NN with Bayesian optimization	The proposed model achieved 93.06% accuracy in obesity classification, outperforming prior methods. The integration of Bayesian optimization enhanced the model's ability to select critical features
Gözükara Bağ <i>et al.</i> ³⁴	Web-based public dataset on physical activity and nutrition (gender, BMI, diet, etc.)	LR, RF, XGBoost with Bayesian optimization	The proposed model achieved 99.33% accuracy using logistic regression, with improved classification accuracy after feature selection. The inclusion of nutritional and activity data further strengthened the model's predictive capacity

Abbreviations: ANN: Artificial neural network; BME: Bagging meta-estimator; Bi-LSTM: Bidirectional long short-term memory; BMI: Body mass index; CNN: Convolutional neural network; COVID-19: Coronavirus disease 2019; DT: Decision tree; GB: Gradient boosting; GBM: Gradient boosting machine; k-NN: k-nearest neighbors; LIME: Local interpretable model-agnostic explanations; LogitBoost: Logistic regression boosting; LR: Logistic regression; MAPE: Mean absolute percentage error; MLP: Multi-layer perceptron; NB: Naïve Bayes; NN: Neural network; POA: Pelican optimization algorithm; PSO: Particle swarm optimization; RF: Random Forest; SGD: Stochastic Gradient Descent; SHAP: Shapley additive explanations; SVM: Support vector machine; UCI: University of California, Irvine; XAI: Explainable artificial intelligence; XGBoost: Extreme gradient boosting.

obesity type I, 35–39.9 as obesity type II, and 40 or higher as obesity type III.

In the study by Helforoush and Sayyad¹⁵, titled *Hybrid Metaheuristic ANN-PSO*, various ML models were applied for obesity risk prediction. The authors proposed a hybrid artificial neural network optimized using particle swarm optimization (ANN-PSO). When evaluated on the University of California, Irvine (UCI) obesity dataset – which contains 2,111 records and 17 features related to dietary habits and physical conditions – the ANN-PSO model achieved an accuracy of ~92%, outperforming traditional regression models. To enhance interpretability, the study employed Shapley additive explanation analysis, which revealed that weight and height were among the most influential features in predicting obesity levels. These findings highlight the potential of metaheuristic optimization methods to improve the performance of neural networks in personalized obesity risk profiling.

Ayub *et al.*¹⁶ developed an attention-enhanced bidirectional long short-term memory (ABi-LSTM) model to classify individuals into obesity categories using the same dataset. Their deep learning architecture incorporated an attention mechanism to emphasize key features – such as height, weight, and activity level – allowing the model to capture complex patterns within the data. The proposed ABi-LSTM achieved a multiclass classification accuracy of 96.5%, representing a substantial improvement in precision, recall, and F1-score over existing approaches. The authors

highlighted this result as a paradigm shift, demonstrating the effectiveness of attention-based deep sequential models in enabling accurate obesity risk prediction.

Shakti *et al.*¹⁷ evaluated multiple ML frameworks on the UCI obesity dataset, which contains 2,111 instances with 17 attributes related to eating habits and lifestyle factors. The models tested included k-nearest neighbors (k-NN), support vector machine (SVM), random forest (RF), gradient boosting (GB), and a multilayer perceptron (MLP) neural network. Among these, the MLP classifier achieved the highest accuracy at 97.2%, followed closely by GB at ~96.2%. These findings highlight that incorporating diverse features – such as dietary habits and physical activity – alongside robust learning algorithms like neural networks (NNs) can yield high classification performance. The study emphasizes that such levels of accuracy are essential for enabling targeted interventions for individuals at risk of obesity.

Yağmur¹⁸ proposed a hybrid model that combines a decision tree (DT) classifier with the pelican optimization algorithm (POA), a metaheuristic optimization technique, to enhance obesity level classification. Utilizing the 2,111-instance dataset, the model applied fuzzy parameter tuning via POA to optimize the tree's decision thresholds for multiclass categorization. The hybrid DT-POA approach reportedly outperforms the standard DT model in predicting obesity levels. Although the precise accuracy value is not explicitly stated, the author highlights the

model's effectiveness and suggests that it can serve as a robust tool to assist healthcare professionals in obesity risk assessment. This study illustrates how evolutionary optimization algorithms can improve the performance of traditional classifiers in this domain.

Özkurt¹⁹ implemented multiple ML algorithms in conjunction with explainability techniques to predict obesity risk. The study utilized data from 2,111 individuals in the UCI obesity dataset, which contains attributes related to dietary habits and physical conditions. A range of ML classifiers – including DT, RF, Naïve Bayes, k-NN, and extreme gradient boosting (XGBoost) – were evaluated. Among these, the XGBoost model achieved the highest classification accuracy at approximately 92% for obesity level prediction. To enhance interpretability, the author employed Shapley additive explanations to identify key features influencing the model's decisions. The analysis revealed that family history of obesity, vegetable intake, and frequency of between-meal consumption were among the most influential predictors. These findings demonstrate that boosting algorithms, when integrated with explainable AI (XAI) techniques, can deliver both high predictive performance and valuable insights into obesity-related risk factors.

In a related study, Wang²⁰ presented their findings in *E3S Web of Conferences*, focusing on obesity level prediction using lifestyle habit features while deliberately excluding direct anthropometric measures such as height and weight to assess model generalizability. The study evaluated a range of ML algorithms, including logistic regression variants (ordinal and multinomial), ensemble methods (LogitBoost and XGBoost), and standard classifiers (Naïve Bayes, SVM, RF, and k-NN). Among these, the LogitBoost ensemble achieved the highest performance, with an accuracy of ~70% and a Kappa statistic of ~0.65. In contrast, the XGBoost model performed poorly, reaching an accuracy of $\leq 20\%$ due to the exclusion of key features. Other models, such as SVM, k-NN, and RF, achieved accuracies ranging from 75% to 79%. Although these values are lower than those reported in studies that incorporate BMI-related features, the author provided important insights. Specifically, they emphasized that when anthropometric data are unavailable, lifestyle indicators play a critical role in obesity prediction. Feature importance analysis revealed that the mode of transportation (e.g., riding a bike) was the most influential predictor, followed by family history of overweight and frequency of vegetable consumption. This comparative study suggests that even in the absence of direct body measurements, lifestyle-related attributes can still support reasonably accurate obesity risk assessments.

Okpe *et al.*²¹ proposed a multilayer perceptron ANN model for multiclass obesity classification using the UCI

obesity dataset. The study involved a comprehensive set of preprocessing steps, including handling missing values, encoding categorical attributes related to diet and physical activity, and additional data preparation procedures. A feedforward ANN was then implemented in Python and trained on the preprocessed dataset. The model achieved a classification accuracy of 97% across seven obesity categories, indicating its effectiveness in capturing the patterns between eating habits, physical conditions, and obesity outcomes. The authors emphasized that careful data cleaning and hyperparameter optimization were critical to achieving this high level of performance. Their findings highlight that even relatively simple NN architectures can yield accuracy comparable to more complex or ensemble-based models when properly optimized.

Azad *et al.*²² proposed a stacking ensemble model that integrates XAI techniques for obesity risk classification, published in early 2025. In their study, the researchers combined multiple base classifiers within a stacked architecture and employed local interpretable model-agnostic explanations (LIME) to provide local interpretability. The model was evaluated on the standard obesity dataset, achieving an accuracy of ~98%, which slightly outperformed previously reported best-performing models such as GB and XGBoost (~97.8%). Beyond the improved predictive performance, the integration of LIME offered valuable insight into individual predictions, addressing the “black-box” issue. Comparative analysis demonstrated that the proposed approach outperformed all prior studies in terms of classification accuracy. This research highlights the effectiveness of combining diverse classifiers through ensembling and underscores the importance of incorporating XAI techniques to enhance model transparency, particularly in clinical decision-making contexts.

Solomon *et al.*²³ introduced a majority-voting ensemble model composed of GB, XGBoost, and an MLP NN to classify obesity levels. Utilizing the Latin American obesity dataset, their hybrid ensemble achieved an accuracy of 97.16%, surpassing the best-performing individual model (XGBoost), which attained 96.37%. This result, published in *Diagnostics* in 2023, established a high-performance benchmark and has since been frequently cited by 2024 studies as a state-of-the-art reference. By comparing multiple algorithms, the authors demonstrated that an ensemble model can effectively leverage the strengths of its individual components. The majority-voting approach outperformed all single classifiers, highlighting the advantage of combining diverse learning paradigms. The impact of this work is further reflected in subsequent research, such as that of Azad *et al.*,²² which aimed to exceed the benchmark established by this study.

In a seminal study, Kaur *et al.*²⁴ investigated the application of ML algorithms for obesity risk prediction and meal planning. Using the UCI obesity dataset, the researchers applied six ML algorithms – GB, Bagging meta-estimator, XGBoost, RF, SVM, and k-NN – to predict adult obesity risk. The models were evaluated under various train-test split ratios (90/10, 80/20, 70/30, etc.), with ensemble methods consistently demonstrating superior performance. Notably, XGBoost achieved an accuracy of up to 97.79% at the 70:30 split, followed closely by GB at ~97.16%. In contrast, simpler models such as k-NN and SVM showed lower accuracy, ranging from 82% to 87%. The study also featured a diet recommendation component generated based on the model's predictions, demonstrating a practical integration of ML with personalized dietary guidance. This early work established the reliability of ML models – particularly boosting ensembles – in predicting obesity-related outcomes, with the reported accuracy of XGBoost (~97.8%) serving as a benchmark in subsequent literature.

Muliawan *et al.*²⁵ focused on leveraging only eating habit features for obesity risk prediction, employing an RF classifier. The study utilized an open-access version of the 17-feature obesity dataset obtained from Kaggle, placing emphasis on dietary variables (e.g., frequency of high-calorie food consumption and meal frequency) while deliberately minimizing reliance on physical measurements. The RF model achieved an accuracy of 81.76% in distinguishing between high-risk and low-risk individuals. Although this performance is lower than that of models incorporating both dietary and physical attributes, it underscores the critical role of physical features in achieving optimal predictive accuracy. Nonetheless, the findings demonstrate that food intake patterns alone can yield approximately 82% accuracy, emphasizing the potential of ML algorithms in healthcare-related applications. The authors conclude that RF can serve as an effective screening tool in scenarios where detailed anthropometric data are unavailable.

Choudhuri *et al.*²⁶ proposed a hybrid ML model for obesity level estimation, utilizing the UCI obesity dataset. While the paper does not report specific performance metrics, the term “hybrid” suggests a combination of classification and optimization techniques. Subsequent studies have cited this work as an early example of integrating multiple classifiers to enhance prediction accuracy. This study is considered foundational in the adoption of ensemble and hybrid approaches within obesity prediction research. It paved the way for later works – such as that of Helforouh and Sayyad¹⁵ – which further developed and refined these strategies.

In a related vein, the study by Cervantes and Palacio,²⁷ published in *Informatics in Medicine Unlocked* in 2020, is

notable as one of the earliest applications of ML algorithms to the “Obesity Levels” dataset. The researchers employed computational intelligence techniques – potentially including neural networks or fuzzy systems – to estimate obesity levels. This pioneering work catalyzed broader interest in the dataset, contributing to the establishment of baseline results and illustrating the feasibility of obesity classification through ML methods.

Ganie *et al.*²⁸ explored the efficacy of ensemble learning techniques for predicting obesity risk using a publicly available Kaggle dataset focused on lifestyle behaviors. The study applied various ensemble learning methods, including RF, extra trees, XGBoost, and CatBoost, using both bagging and boosting strategies. Among these, XGBoost delivered the highest performance, achieving an accuracy of 98.1% and an F1-score of 96.5%. The findings demonstrate the robustness of ensemble models, particularly boosting techniques, in deriving predictive insights from multi-dimensional lifestyle datasets.

Nagarajan *et al.*²⁹ performed a comparative analysis of several ML and deep learning models for predicting obesity levels using a real-world dataset with 17 features, including demographic and health-related variables. To improve model performance on imbalanced classes, the authors implemented SMOTE. The algorithms tested included TabNet, XGBoost, GB, MLP, and RF. The GB algorithm achieved the highest accuracy of 99.3%, with XGBoost and TabNet following closely at 99% and 98.4%, respectively, validating the effectiveness of ensemble and deep learning models in healthcare data analysis.

Umoh *et al.*³⁰ focused on optimizing various ML classifiers to estimate obesity levels from physical activity and dietary data obtained through structured surveys. The dataset underwent thorough preprocessing, including normalization and feature selection. The study evaluated a range of classifiers, including SVM, GB, DT, and others. Among them, GB emerged as the top-performing model, achieving an accuracy of 97.23%. This research highlighted the significance of integrating robust feature selection with classifier tuning for effective obesity level prediction.

Vairachilai *et al.*³¹ applied the protein intake prediction and response (PIPR) ML model to analyze the impact of dietary behavior on obesity during the COVID-19 pandemic. The dataset included comprehensive lifestyle and nutritional behavior indicators. Multiple ensemble learning algorithms, such as RF and extra trees, were evaluated in the study. The PIPR model stood out with an accuracy of 96.7%, demonstrating its capability to capture nuanced relationships between protein intake and obesity risk and confirming the value of ensemble strategies in obesity prediction tasks.

Forte *et al.*³² developed a deep learning-based NN model aimed at classifying obesity risks among Portuguese adolescents. The model used the FITescola[®] dataset, which includes information on physical fitness levels and BMI percentiles. Leveraging the power of deep learning, specifically convolutional NNs, the study aimed to improve the detection of obesity risk patterns in youth. The proposed model achieved a classification accuracy of 96.3%, showcasing the potential of deep NNs to support early intervention strategies in public health contexts.

Yağın *et al.*³³ proposed a Bayesian-optimized NN for the estimation of obesity levels using a dataset focused on lifestyle factors and eating habits obtained from the UCI ML Repository. The study utilized a feedforward deep NN whose hyperparameters were tuned via Bayesian optimization to maximize predictive accuracy. This optimization improved the network's ability to identify significant patterns in the data by fine-tuning parameters such as learning rate and hidden layers. The final model achieved an accuracy of 96.5%, outperforming earlier approaches and demonstrating the effectiveness of combining NNs with optimization strategies.

Gözükara Bağ *et al.*³⁴ introduced a predictive modeling approach that integrates physical activity and nutritional habit data for classifying obesity levels. They utilized a dataset comprising 2,111 records from the UCI ML Repository, which included variables such as gender, BMI, dietary patterns, and physical activity. The study employed ML algorithms, including RF, k-NN, and XGBoost. Feature scaling and selection techniques were applied to enhance model performance. The highest classification accuracy of 98.87% was achieved using the XGBoost algorithm, underscoring its superiority in handling complex lifestyle-related data for obesity classification.

Several works underscore the impact of diet and lifestyle features on obesity classification. For example, studies using the EOL dataset have identified that eating habits (e.g., frequency of high-calorie food intake, number of meals) and lifestyle choices (e.g., mode of transport, frequency of physical activity) significantly influence obesity level predictions. These findings are consistent with nutrition research showing that “prudent” diet patterns (rich in fruits and vegetables) are linked to lower obesity, whereas fast-food – heavy patterns correlate with higher adiposity.¹⁰ Obesity is closely tied to metabolic syndrome markers. The TyG index study and investigations of oxytocin levels illustrate that blood biomarkers and hormonal factors are often elevated in obesity and associated with eating behaviors.^{12,13}

In addition to SMOTE, various over-sampling techniques have been adapted for multiclass problems in

medicine. Yang *et al.*³⁵ reviewed multiclass oversampling for imbalanced health datasets, noting an emerging trend toward hybrid methods combining SMOTE with other strategies.³⁵ While SMOTE-NC (used in our study) is a straightforward approach that interpolates minority-class samples in mixed-type data, more complex generators like GANs can capture non-linear feature dependencies. Synthetic tabular data in health often requires careful evaluation; we leverage standard classification metrics to assess model performance on generated data.⁷

Recent work on GANs and VAEs shows they can simulate realistic clinical datasets. For instance, standalone reports on conditional tabular GANs (CTGANs) or VAE variants demonstrate their success in reproducing distributions of complex clinical features.^{6,7} However, empirical comparisons of these methods (VAE versus GAN versus traditional oversampling) in specific applications like obesity remain limited, which motivates our empirical study. In summary, while many studies have achieved high accuracy in obesity prediction using ensemble or deep learning models, they typically rely on the original data (often including BMI-related attributes).

3. Materials and methods

3.1. Dataset definition

This study utilized the dataset titled *Estimation of Obesity Levels Based on Eating Habits and Physical Condition*.⁵ The data were collected from individuals in Mexico, Peru, and Colombia, encompassing information on dietary habits, physical conditions, and obesity levels. The dataset contains a total of 2,111 instances and 17 attributes. The first 498 instances were collected directly from users, while the remaining samples were synthetically generated by Palechor *et al.*¹⁴ using SMOTE. All analyses and synthetic data generation in this study were conducted using the 498 user-collected samples. The features included are gender, age, height, weight, family history of obesity, frequent consumption of high-calorie foods, frequency of vegetable consumption, number of main meals, consumption of food between meals, smoking, daily water consumption, calorie tracking, frequency of physical activity, frequency of using technological devices, alcohol consumption, type of transportation used, and obesity level. It is important to note that the dataset contains no missing values. The gender distribution is shown in [Figure 1](#), with 271 males (54.4%) and 227 females (45.6%), indicating a relatively balanced sample.

As illustrated in [Figure 2](#), the data indicate a predominance of affirmative responses, with 300 individuals (60.2%) supporting the proposition and 198 individuals (39.8%) opposing it. The distribution reflects a clear majority in favor of the proposition.

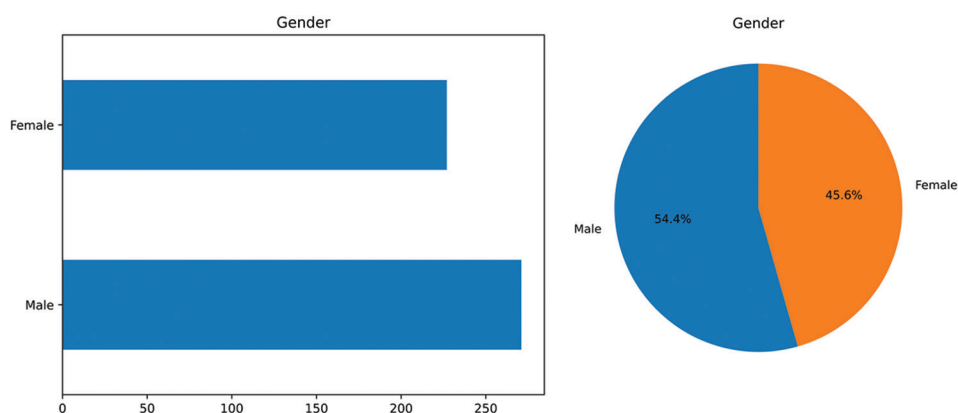


Figure 1. Gender distribution

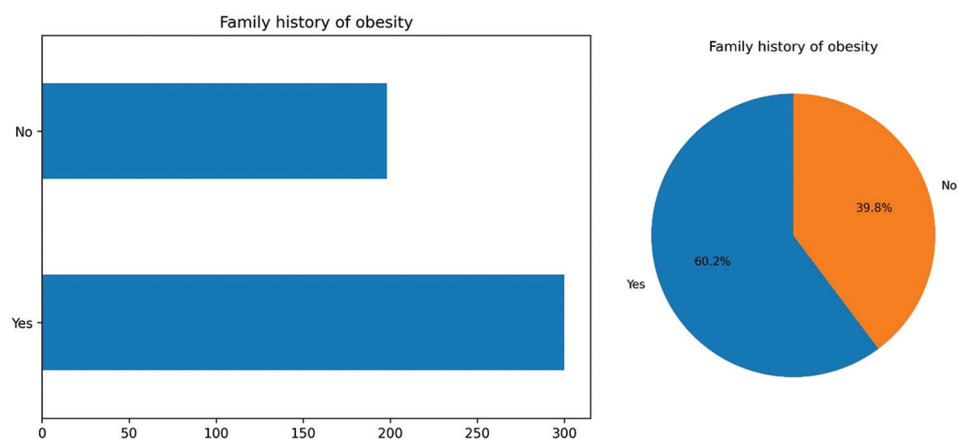


Figure 2. Distribution of family history of obesity

As shown in Figure 3, a similar pattern is observed in responses related to the consumption of high-calorie foods. The “yes” option predominates, with 348 individuals (69.9%) indicating frequent consumption, while the “no” option accounts for the remaining 30.1% of the sample.

The frequency of vegetable consumption was categorized into three response options: sometimes, always, and never. The number of responses for each category was 272, 194, and 32, respectively. As illustrated in Figure 4, these correspond to proportions of 54.6%, 39%, and 6.4%, respectively.

In the distribution of the number of main meals, the option of consuming three meals per day ranked first, with 344 individuals (69.1%). As illustrated in Figure 5, 108 individuals (21.7%) reported consuming only one meal, while 46 individuals (9.2%) reported consuming four meals daily.

As illustrated in Figure 6, the distribution of intermeal food consumption indicates that “sometimes” is the most

frequently selected option, with 289 individuals. The remaining responses include “often” ($n = 136$), “always” ($n = 53$), and “no” ($n = 20$), each representing smaller portions of the sample.

The sample population consisted of 32 smokers and 466 nonsmokers, corresponding to 6.4% and 93.6% of the total, respectively. These proportions are illustrated in Figure 7.

An analysis of the daily water consumption reveals that 135 individuals consume <1 L, 266 individuals consume between 1 and 2 L, and 97 individuals consume more than 2 L/day. The corresponding scatter plot is presented in Figure 8.

A total of 55 individuals reported monitoring their caloric intake, whereas 443 individuals did not. As shown in Figure 9, these correspond to 11% and 89% of the sample, respectively.

The data indicate that 158 individuals (31.7%) engage in physical activity one or 2 days/week, while 162 individuals (32.5%) do not engage in any physical activity. In addition,

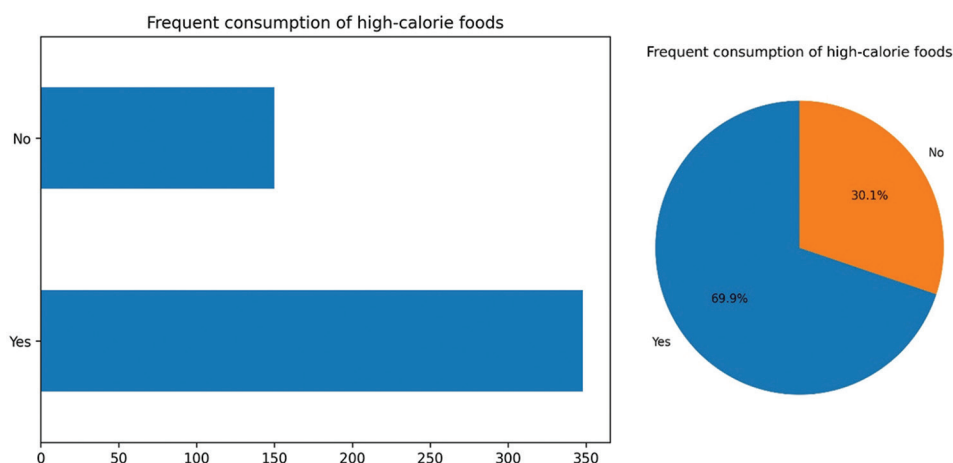


Figure 3. Distribution of frequent consumption of high-calorie foods

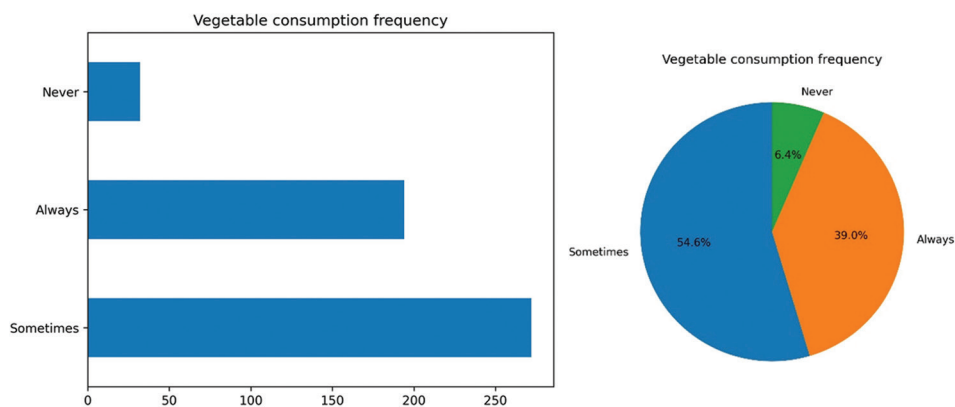


Figure 4. Distribution of frequency of vegetable consumption

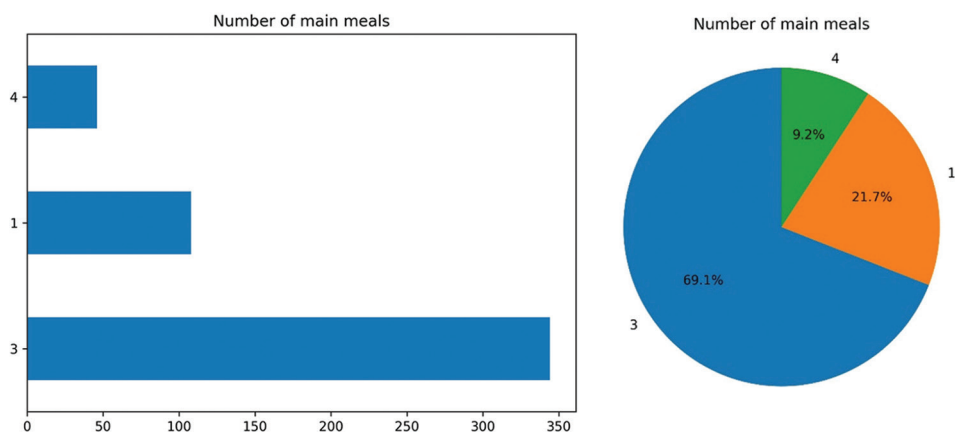


Figure 5. Distribution of the number of main meals

113 individuals (22.7%) are active for 2–4 days, and the remaining 65 individuals (13.1%) participate in physical activity for 4–5 days/week. The distribution of these values is illustrated in Figure 10.

A total of 243 individuals (48.8%) reported using technological devices for 0–2 h/day, whereas 181 individuals (36.3%) indicated daily usage of 3–5 h. As illustrated in Figure 11, the remaining 74 individuals

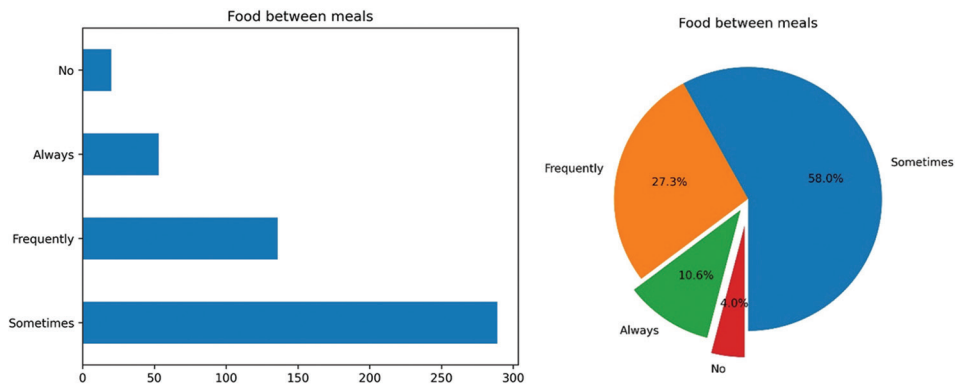


Figure 6. Distribution of food consumption between meals

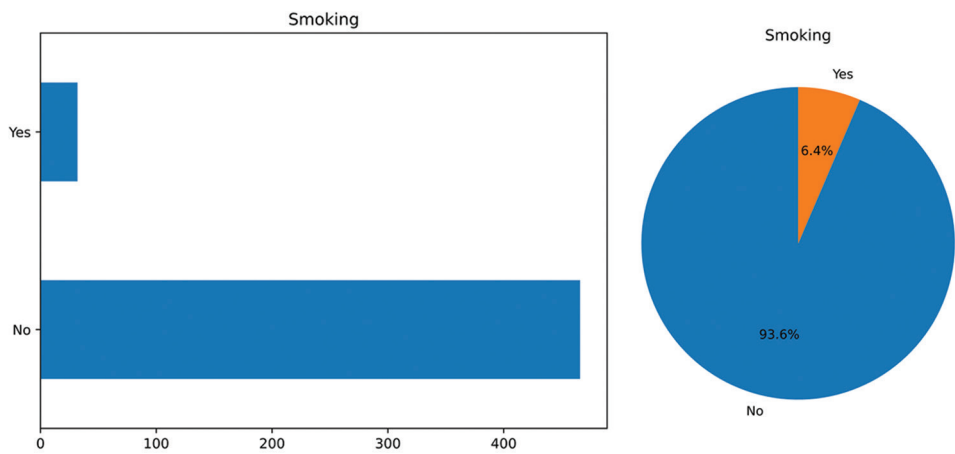


Figure 7. Distribution of smoking

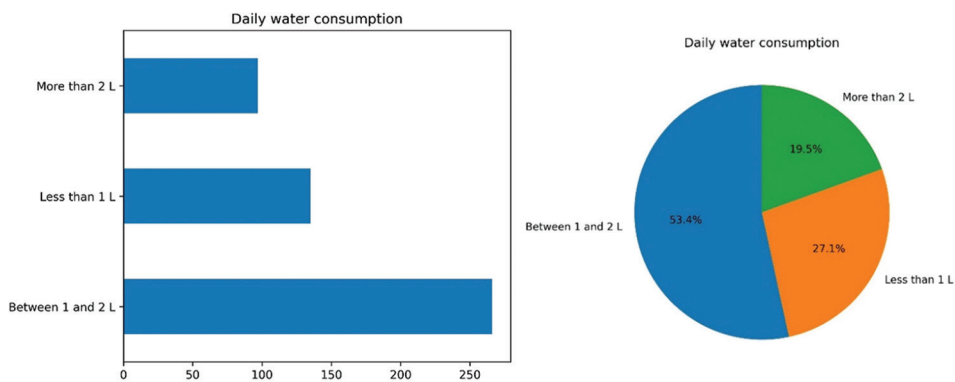


Figure 8. Distribution of daily water consumption

(14.9%) reported using technological devices for more than 5 h/day.

As illustrated in Figure 12, 273 individuals (54.8%) reported occasional alcohol consumption, while 179 individuals (35.9%) stated that they never consumed alcohol. In addition, 45 individuals (9%) reported frequent

alcohol use, and one participant (0.2%) indicated consistent daily alcohol consumption.

Regarding transportation preferences, the data indicate a predominant reliance on public transportation, with 326 individuals (65.5%) selecting this option. Automobile use ranks second, preferred by 99 individuals (19.9%).

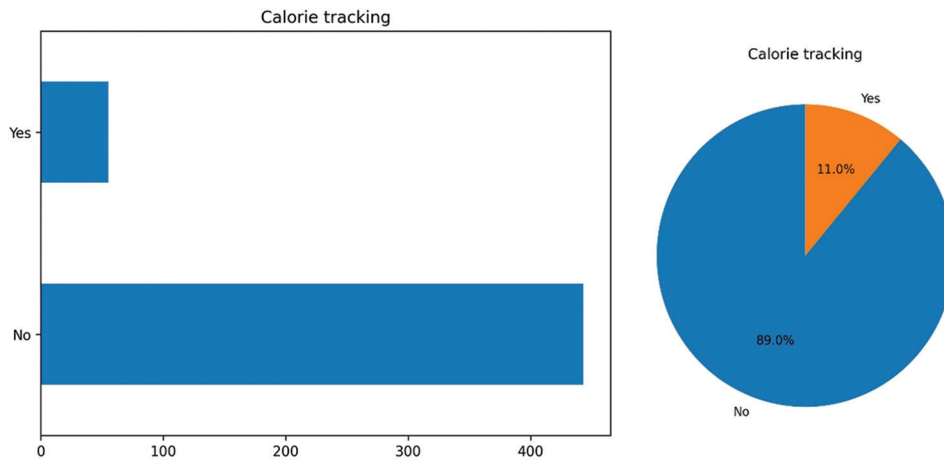


Figure 9. Distribution of calorie tracking

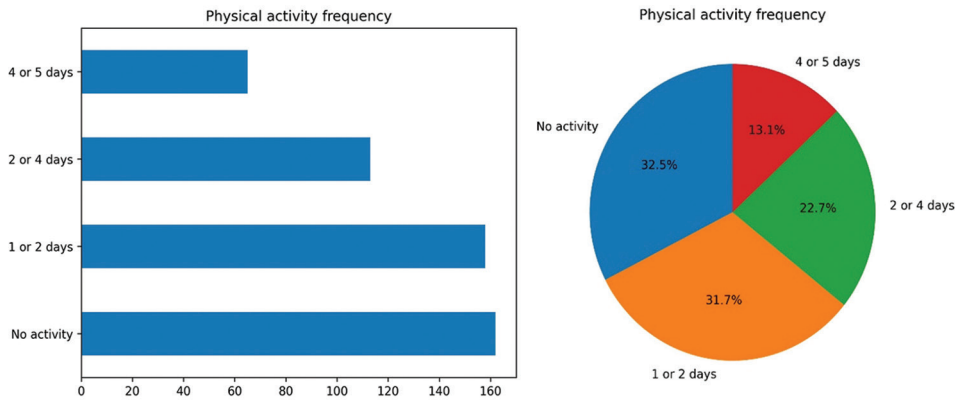


Figure 10. Distribution of physical activity frequency

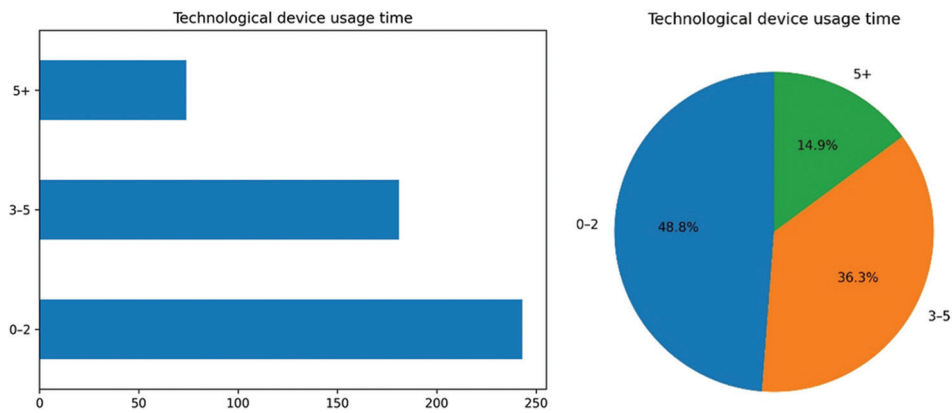


Figure 11. Distribution of duration of technological device use

Walking was chosen by 55 individuals (11%), followed by motorcycles (11 individuals; 2.2%) and bicycles (seven individuals; 1.4%). The distribution of these preferences is illustrated in Figure 13.

The original version of the dataset includes seven obesity level categories as class labels: underweight, normal weight, level I overweight, level II overweight, type I obese, type II obese, and type III obese. Due to class imbalance,

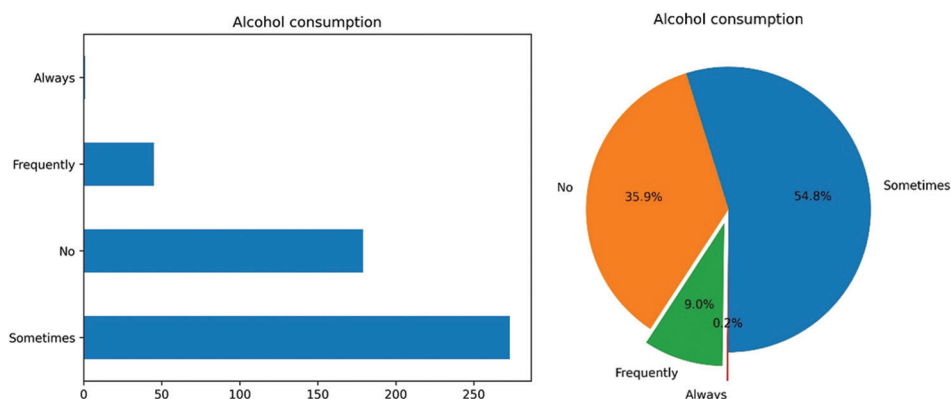


Figure 12. Distribution of alcohol consumption

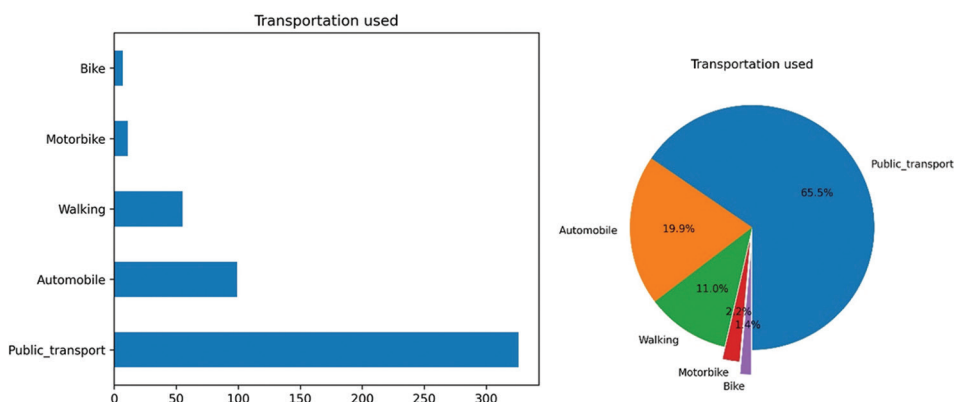


Figure 13. Distribution of transportation modes used

the overweight and obese subcategories were consolidated into single groups, resulting in four final categories. The distribution of the obesity level attribute after this merging is illustrated in Figure 14. According to the revised distribution, 37 individuals (7.4%) were classified as underweight, 284 individuals (57%) as normal weight, 116 individuals (23.3%) as overweight, and 61 individuals (12.2%) as obese.

The age values in the dataset ranged from 14 to 61 years. The distribution of ages is presented in Figure 15. The mean age was 23.15 years, with a standard deviation of 6.72, whereas the median age was 21.

The height values in the dataset ranged from 1.45 to 1.98 m. The mean height was 1.69 ± 0.09 meters, and the median was 1.68 meters. The distribution of height values is illustrated in Figure 16.

The dataset includes weight values ranging from 39 to 173 kg. The mean weight was 69.57 ± 17.01 kg, and the median was 67 kg. The distribution of weight values is depicted in Figure 17.

The mean weight within the underweight category was 53.8 kg for men and 46 kg for women. For individuals in the normal weight category, the mean values were 67.9 kg (men) and 56.5 kg (women). In the overweight group, mean weights were 82.7 kg for men and 71.4 kg for women. Finally, in the obese category, which represents the highest weight class, the averages increased to 106.4 kg for men and 86.3 kg for women. The relationship between obesity levels and weight by gender is illustrated in Figure 18.

As illustrated in Figure 19, individuals who did not consume vegetables and those who avoided eating between meals were predominantly in the younger age group. In addition, data suggest that underweight individuals and those who tracked their caloric intake also tend to be younger.

As illustrated in Figure 20, weight distributions are presented across the different obesity levels. As expected, the categories appeared in a sequential manner, with weight increasing progressively from lower to higher obesity classes.

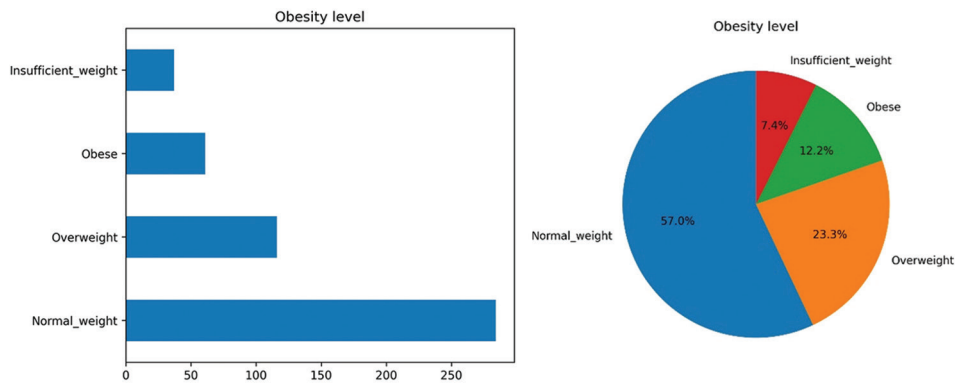


Figure 14. Distribution of obesity level

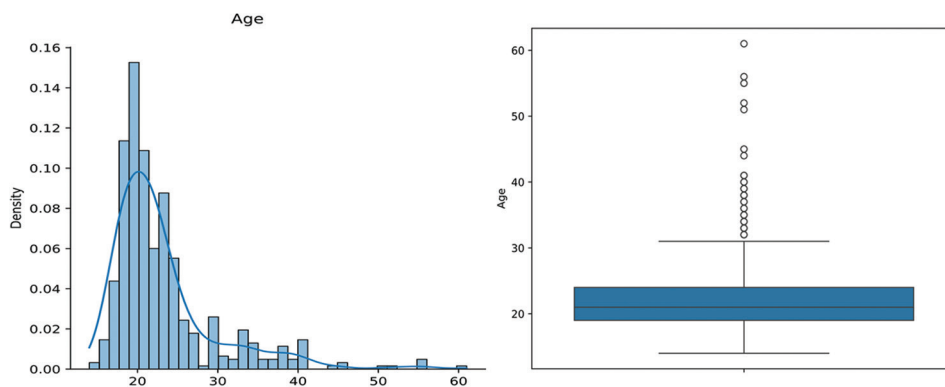


Figure 15. Age distribution

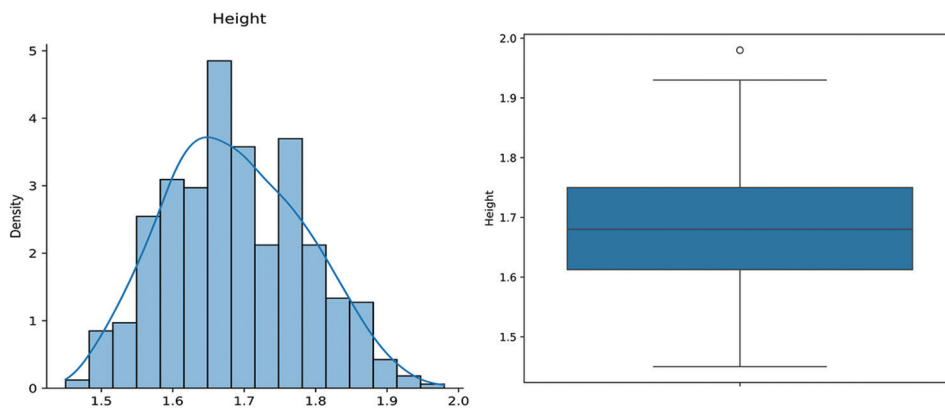


Figure 16. Height distribution

As illustrated in Figure 21, the axes of the graph represent height and weight, with color coding used to indicate different obesity levels. It is evident that, for a given height, the obesity class increased with weight, and male subjects tended to fall into higher obesity categories compared to female subjects. As BMI – the metric used to assign class labels – is directly correlated with height

and weight (as shown in Equation I and Figure 21), the classes appeared to be linearly separable in the graph. However, including these attributes in model training may lead to an overestimation of performance. To illustrate this potential discrepancy, two distinct datasets were used: one that includes height and weight, and another that excludes them.

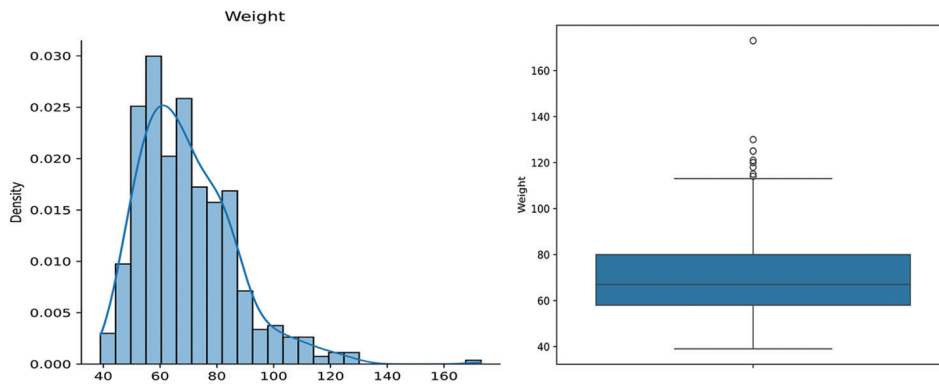


Figure 17. Weight distribution

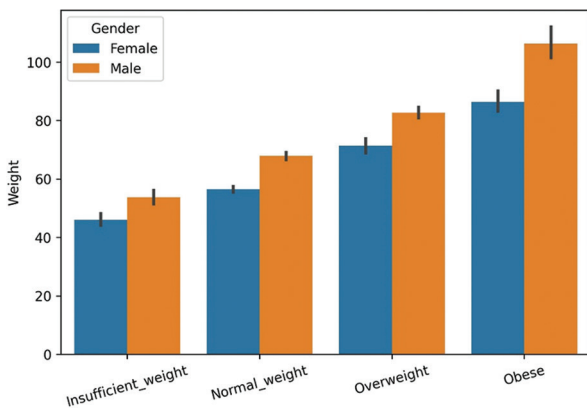


Figure 18. Associations of different levels of obesity with weight by gender

$$BMI = \frac{\text{Weight (in kg)}}{\text{Height}^2 \text{ (in m)}} \tag{I}$$

As illustrated in Figure 22, the rows represent gender, the columns indicate whether individuals tracked their caloric intake, the axes correspond to age and weight, and the colors denote obesity classes. The figure reveals that individuals with higher levels of obesity were predominantly those who did not track calories and exhibited higher weight values. Furthermore, the data suggest that individuals who engaged in calorie tracking tended to be younger.

3.2. Synthetic data generation

The synthetic data generation methods employed in this study included the SMOTE-NC method from the *Imbalanced-learn* library by Lemaître *et al.*³⁶ and the VAE-based tabular VAE (TVAE) and GAN-based CTGAN by Xu *et al.*³⁷ methods in the Synthetic Data Vault (SDV) library by Patki *et al.*³⁸ Given that the majority class in the original dataset consisted of individuals with normal weight (284 samples), synthetic data were generated to

match this sample size in each of the minority classes. After data generation, the final dataset comprised 1,136 instances, with equal representation across the four classes: underweight, normal weight, overweight, and obese.

SMOTE-NC is a variant of the SMOTE designed to address class imbalance by generating synthetic samples through interpolation. Unlike the original SMOTE algorithm, SMOTE-NC is capable of handling both numerical and categorical features, thereby producing synthetic data that more accurately represents the underlying structure of the original dataset. This method improves the diversity and representativeness of the minority class, ultimately contributing to more robust and generalizable model training.³⁹

The TVAЕ is a generative model based on the VAE architecture, specifically designed to handle the heterogeneous nature of tabular data, which often includes a mix of continuous and categorical variables. The model consists of an encoder network that maps the input data into a latent space represented by Gaussian distributions and a decoder network that reconstructs the data from these latent representations. This structure enables TVAЕ to learn complex data distributions and supports conditional data generation by allowing specific attributes to be fixed during the sampling process. Once trained, TVAЕ can generate realistic synthetic tabular data by sampling from the latent space, providing a robust framework for addressing class imbalance and performing data augmentation tasks.⁴⁰

The CTGAN extends the traditional GAN architecture by introducing modifications tailored to the unique characteristics of tabular data. While standard GANs – originally developed for image generation – struggle with the heterogeneity of tabular datasets, particularly due to mixed data types and the presence of discrete variables, CTGAN effectively addresses these limitations.

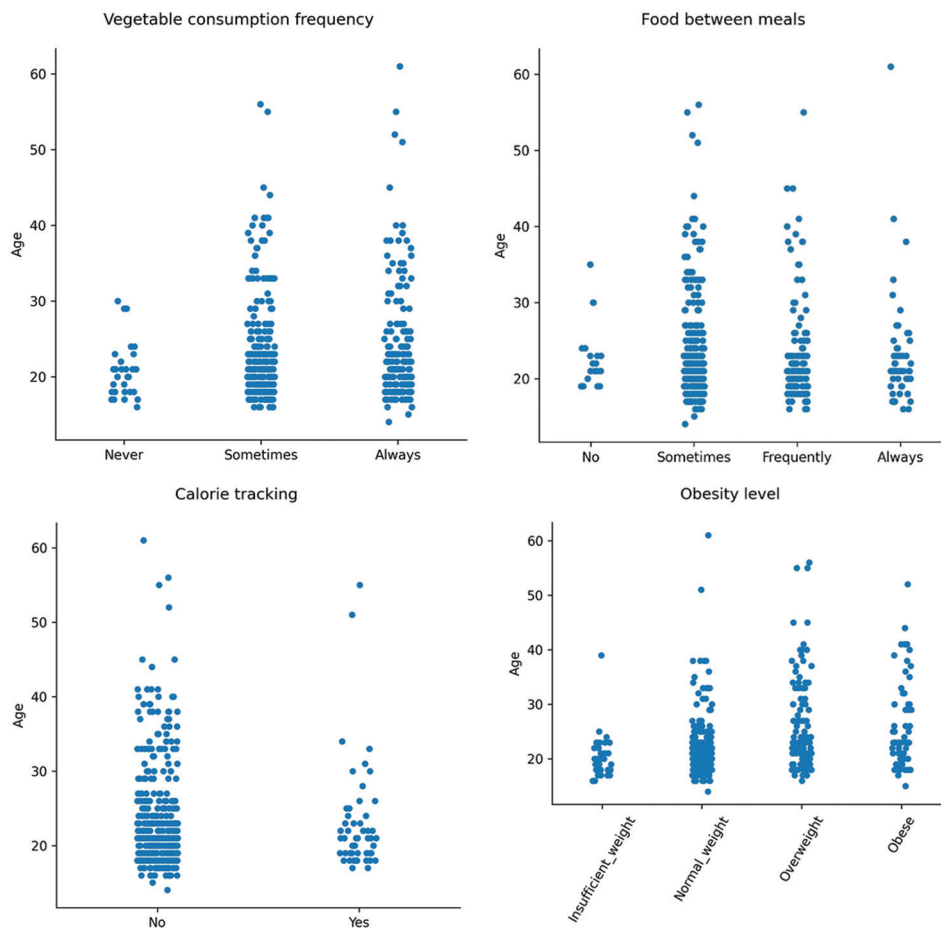


Figure 19. Frequency of vegetable consumption, food consumption between meals, calorie tracking, and obesity level by age

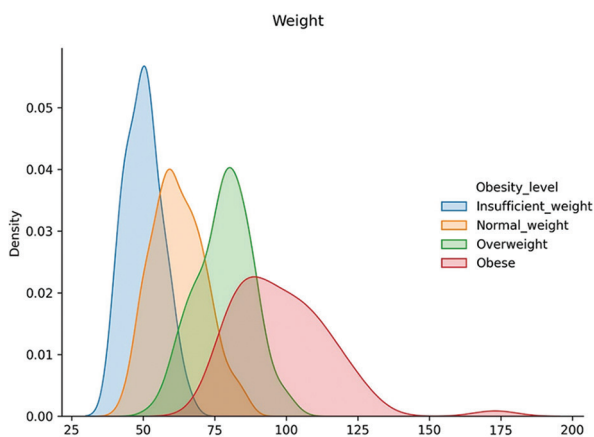


Figure 20. Weight distribution across obesity levels

It incorporates conditional data generation and mode-specific normalization techniques to model complex relationships in tabular data more accurately.⁴¹

The synthetic data generated using SMOTE-NC, which is designed to handle both numerical and categorical variables, includes numerical attributes such as age, height, weight, and number of main meals. These values were initially represented with up to 16 digits after the decimal point. Therefore, appropriate rounding procedures were applied to enhance data consistency. Specifically, the age and number of main meals were rounded to whole numbers, height to two decimal places, and weight to one decimal place. In contrast, no such adjustments were necessary for the synthetic data generated by VAE and GAN-based methods, as this issue did not occur. However, for these NN-based approaches, a separate model was trained for each class. It was observed that training a single model using all class samples resulted in lower performance, highlighting the advantage of class-specific model training in these architectures.

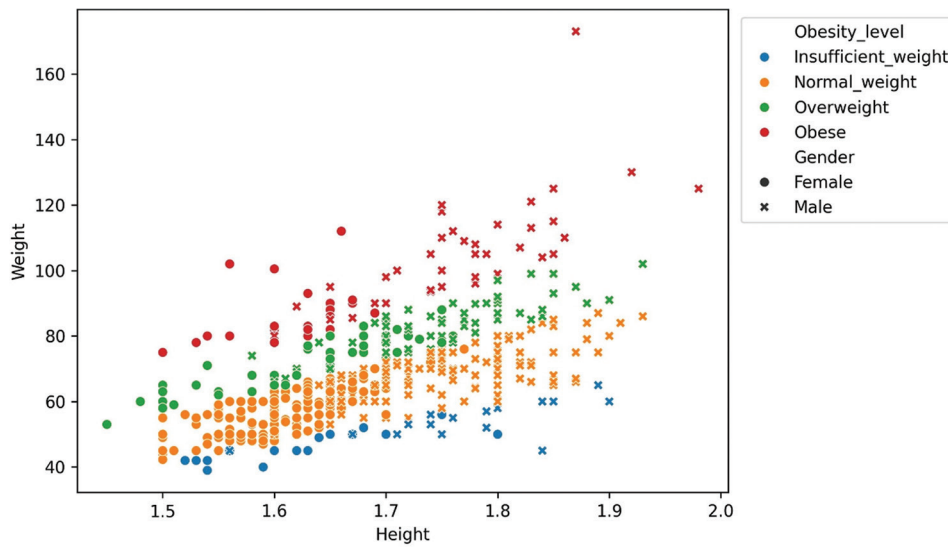


Figure 21. Obesity class distributions by gender on height and weight axes

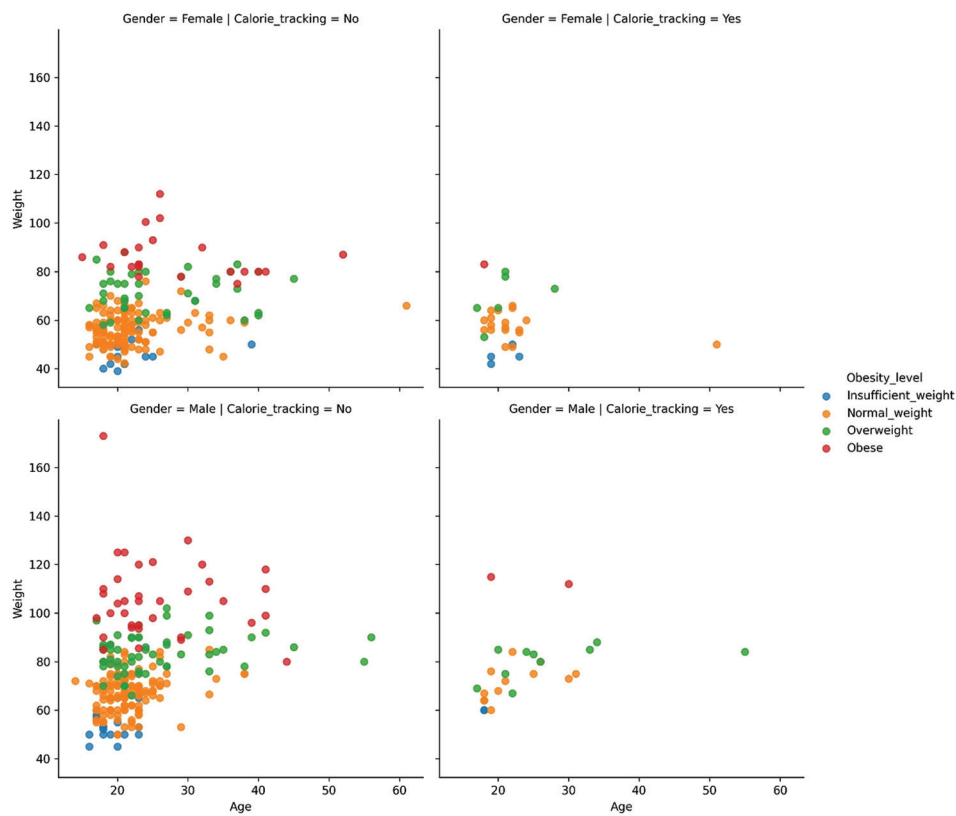


Figure 22. Associations between gender, calorie tracking, and obesity level on the age and weight axes

3.3. Data preprocessing and ML algorithms

Given that the majority of the attributes in the dataset are categorical, both one-hot encoding and label encoding methods were applied to prepare the data for input

into ML algorithms. Among the attributes subjected to one-hot encoding, the gender variable was coded as male = 0 and female = 1. For binary attributes with “yes”/“no” responses, such as family history of obesity, frequent consumption of high-calorie foods, smoking,

and calorie tracking, “no” responses were encoded as 0 and “yes” responses as 1.

Label encoding was applied to attributes that exhibit an ordinal relationship among their values. These attributes include frequency of vegetable consumption, food consumption between meals, daily water consumption, frequency of physical activity, duration of technological device use, alcohol consumption, transportation preference, and obesity level (class label).

For frequency of vegetable consumption, the categories “never,” “sometimes,” and “always” were encoded as 0, 1, and 2, respectively. The responses “no,” “sometimes,” “often,” and “always” for food consumption between meals were encoded as 0, 1, 2, and 3. Daily water consumption categories – “less than 1 liter,” “between 1 and 2 liters,” and “more than 2 liters” – were encoded as 1, 2, and 3. For frequency of physical activity, the categories “no activity,” “1–2 days,” “2–4 days,” and “4–5 days” were assigned the values 0, 1, 2, and 3, respectively. Duration of technological device usage was categorized as “0 – 2 h,” “3 – 5 h,” and “5+ h” and encoded as 0, 1, and 2. Alcohol consumption levels (“no,” “sometimes,” “often,” and “always”) were encoded as 0, 1, 2, and 3.

For transportation preference, two different encoding strategies were tested. In one version, “public

transportation” was assigned the base value 0, whereas in the other version, “walking” was assigned 0. The results of ML models showed no significant difference between these two approaches. As a result, the final version of the dataset adopted the encoding that prioritized “walking,” with values ranging from 0 to 4. The obesity level, designated as the class label, was encoded from 0 to 3, where 0 represents the lowest level (underweight), and 3 represents the highest level (obese).

Following the encoding process, correlation heatmaps were generated for each of the datasets created using SMOTE-NC, TVAE, and CTGAN. These heatmaps were used to visualize the relationships between the attributes, where values close to 1 indicate strong positive correlations and values close to -1 indicate strong negative correlations.

The correlation heatmap generated from the dataset synthesized using the SMOTE-NC method, presented in Figure 23, reveals notable relationships among several anthropometric variables. Significant correlations were observed between height and gender, weight and gender, weight and height, family history of obesity and weight, obesity level and weight, as well as obesity level and family history of obesity. The strongest correlation, with a

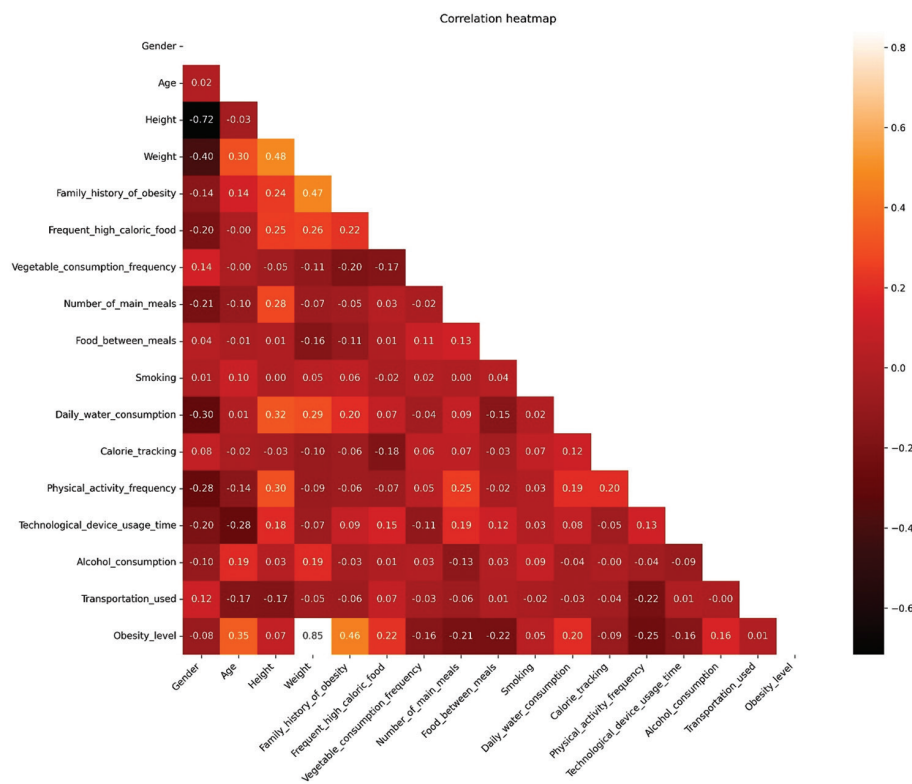


Figure 23. Correlation heatmap for the dataset generated using the synthetic minority oversampling technique—nominal and continuous

coefficient of 0.85, was identified between obesity level and weight, indicating a strong positive relationship.

As illustrated in Figure 24, the correlation heatmap generated from the dataset synthesized using the TVAE method reveals significant associations among various health-related metrics. These included the correlations between height and gender, family history of obesity and weight, obesity level and weight, obesity level and family history of obesity, and obesity level and frequency of physical activity. The strongest correlation was observed between obesity level and weight, with a coefficient of 0.90, indicating a very strong positive relationship.

As shown in Figure 25, the correlation heatmap generated from the dataset synthesized using the CTGAN method reveals notable associations between height and gender, height and weight, and obesity level and weight. A correlation coefficient of 0.84 was observed between obesity level and weight, indicating a strong positive relationship. The consistently high correlation between obesity level and weight across all three datasets (SMOTE-NC, TVAE, CTGAN) can be attributed to the direct role of weight in the calculation of BMI, which serves as the basis for obesity classification, as shown in Equation I and Figure 21.

In addition to encoding, the dataset was standardized using the StandardScaler function from the Scikit-learn library. For each ML algorithm, the training and testing process was repeated 100 times. During each iteration, models were evaluated using multiclass classification metrics: accuracy, precision, recall (sensitivity), and F1-score. These metrics were macro-averaged across classes. The performance metrics used in the study are defined in Table 2.

All metrics were computed using the actual (true) class labels and model predictions on the test set; a “correct prediction” means the predicted class matches the true label. In every run, the random_state parameter was set to values ranging from 0 to 99, based on the current iteration index. Stratified splitting was employed to divide the dataset into training and test sets while preserving class distribution. All classifiers available in the Scikit-learn library were evaluated, and the results of the five models with the highest F1 scores were reported.

4. Results and discussion

This section presents the performance metrics of the models trained on datasets generated using SMOTE-NC, TVAE, and CTGAN – the synthetic data generation techniques

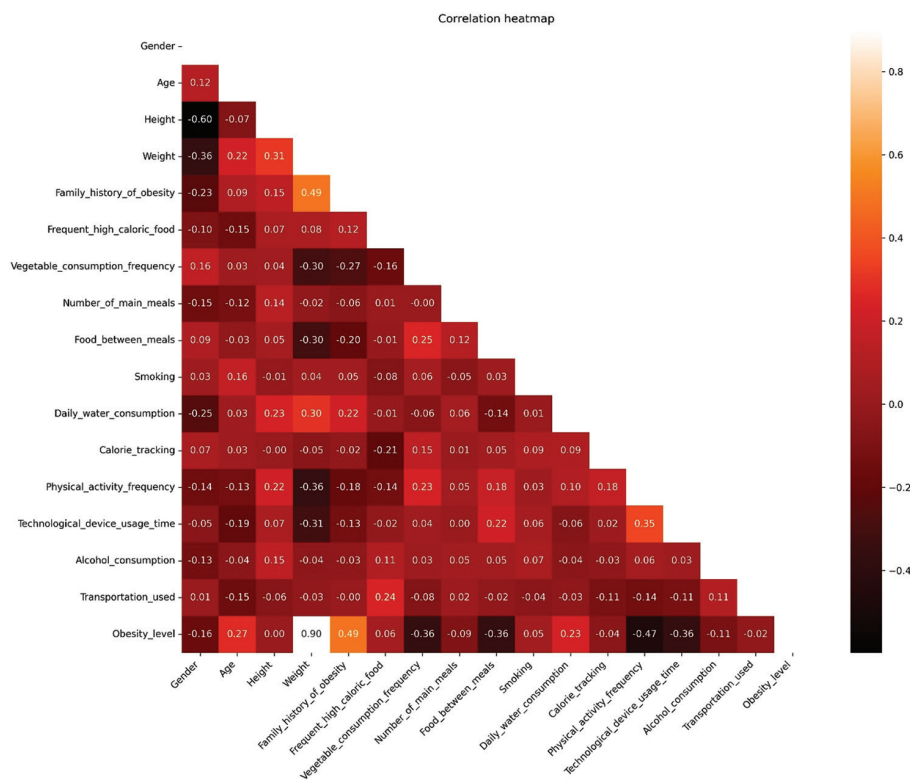


Figure 24. Correlation heatmap for the dataset generated using the tabular variational autoencoder

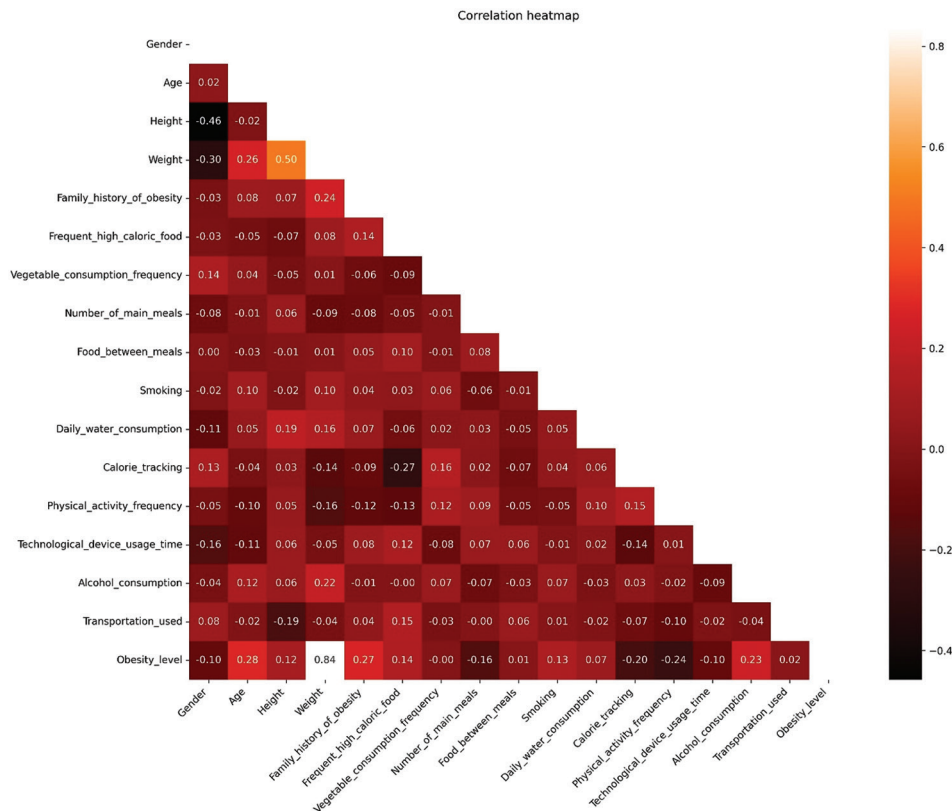


Figure 25. Correlation heatmap for the dataset generated using the conditional tabular generative adversarial network

Table 2. Performance metrics used in model evaluation

Metric	Explanation	Formula
Accuracy (ACC)	Gives the correct prediction rate of the model across all classes	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$
Precision (PRE)	Shows how many positive predictions are actually positive	$PRE = \frac{TP}{TP+FP}$
Recall (REC)	Shows how many true positives are correctly predicted	$REC = \frac{TP}{TP+FN}$
F1-score	Is the harmonic mean of the accuracy and recall metrics	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Notes: TP: True positive; a positive sample correctly predicted as positive. TN: True negative; a negative sample correctly predicted as negative. FP: False positive; a negative sample incorrectly predicted as positive. FN: False negative; a positive sample incorrectly predicted as negative.

employed in this study. Each reported value represents the average performance of 100 independently trained models, utilizing all available classification algorithms in the Scikit-learn library. The results reflect the top five classifiers in

terms of F1-score. Model performance is reported for two scenarios: one excluding the height and weight attributes, and one including them. As shown in Table 3 and Figure 26, the classifiers trained on the SMOTE-NC-generated dataset without height and weight information achieved average performance scores ranging from 70% to 75%.

When height and weight attributes were included, as shown in Table 4 and Figure 27, the average performance increased significantly, with F1 scores reaching up to 98.16%.

As illustrated in Table 5 and Figure 28, the dataset generated using the TVAE method yielded an average performance between 71% and 73% when height and weight attributes were excluded.

MODELS trained on the TVAE-generated dataset that included height and weight features achieved an F1 score of 97.49%. A comprehensive summary of these results is presented in Table 6 and Figure 29.

In the case of the dataset generated using CTGAN – the final synthetic data generation technique – classification models achieved lower performance compared to the other two methods when height and weight attributes were

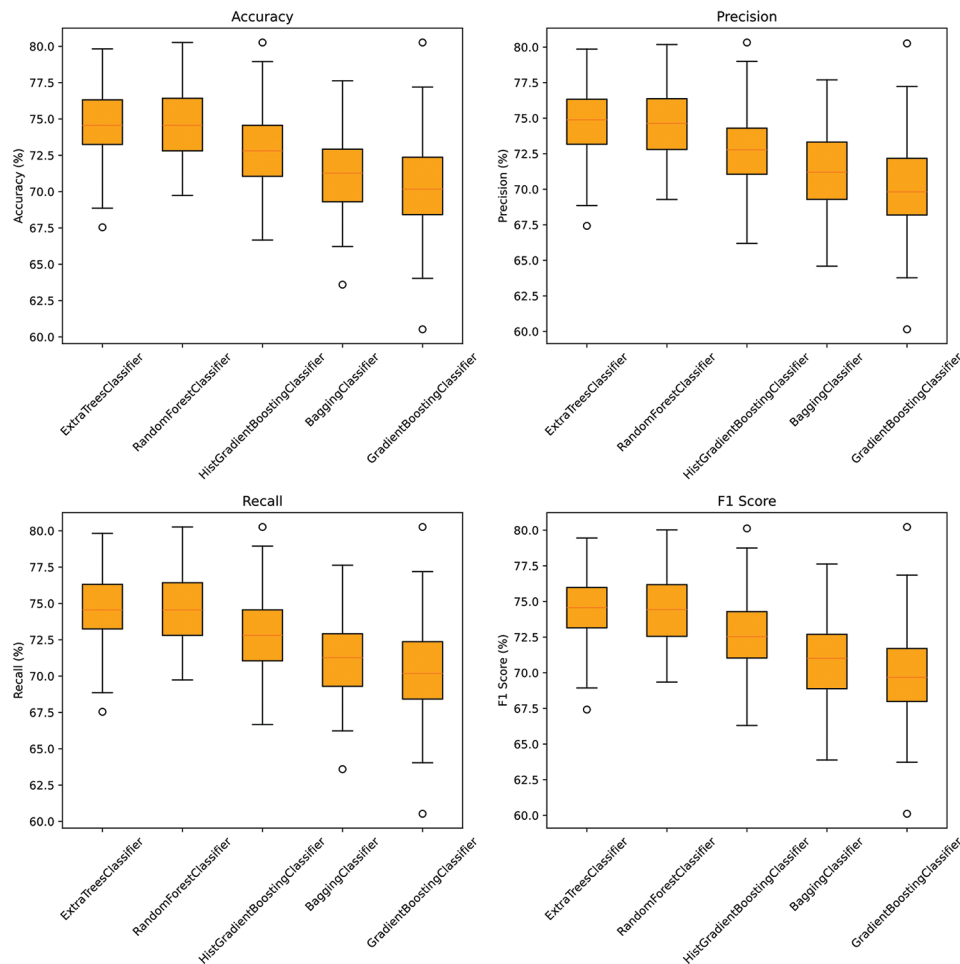


Figure 26. Performance metrics plots of the five most successful classifiers on the SMOTE-NC dataset (excluding height and weight attributes)

Table 3. Average performance metrics of the five most successful classifiers on the synthetic minority oversampling technique – nominal and continuous dataset (excluding height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
ExtraTrees	74.62	74.72	74.62	74.48
RandomForest	74.71	74.62	74.71	74.45
HistGradBoosting	72.87	72.89	72.87	72.70
Bagging	71.24	71.27	71.24	70.98
GradBoosting	70.31	70.19	70.31	69.91

excluded. As shown in Table 7 and Figure 30, the models trained on this dataset reached an average F1 score of approximately 60%.

When height and weight were incorporated as input features, the performance of classifiers trained on CTGAN-generated data became comparable to those trained on

Table 4. Average performance metrics of the five most successful classifiers (using height and weight attributes) on the synthetic minority oversampling technique – nominal and continuous dataset

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LogisticRegCV	98.17	98.21	98.17	98.17
HistGradBoosting	96.61	96.65	96.61	96.61
GradBoosting	95.73	95.79	95.73	95.73
Bagging	94.55	94.64	94.55	94.55
LogisticReg	92.86	93.03	92.86	92.86

data synthesized by other methods. A detailed analysis of Table 8 and Figure 31 showed that F1 scores range between 94% and 97%.

The strong performance of classifiers when using height and weight reaffirms known biology principles: BMI strongly separates obesity classes. A novel insight

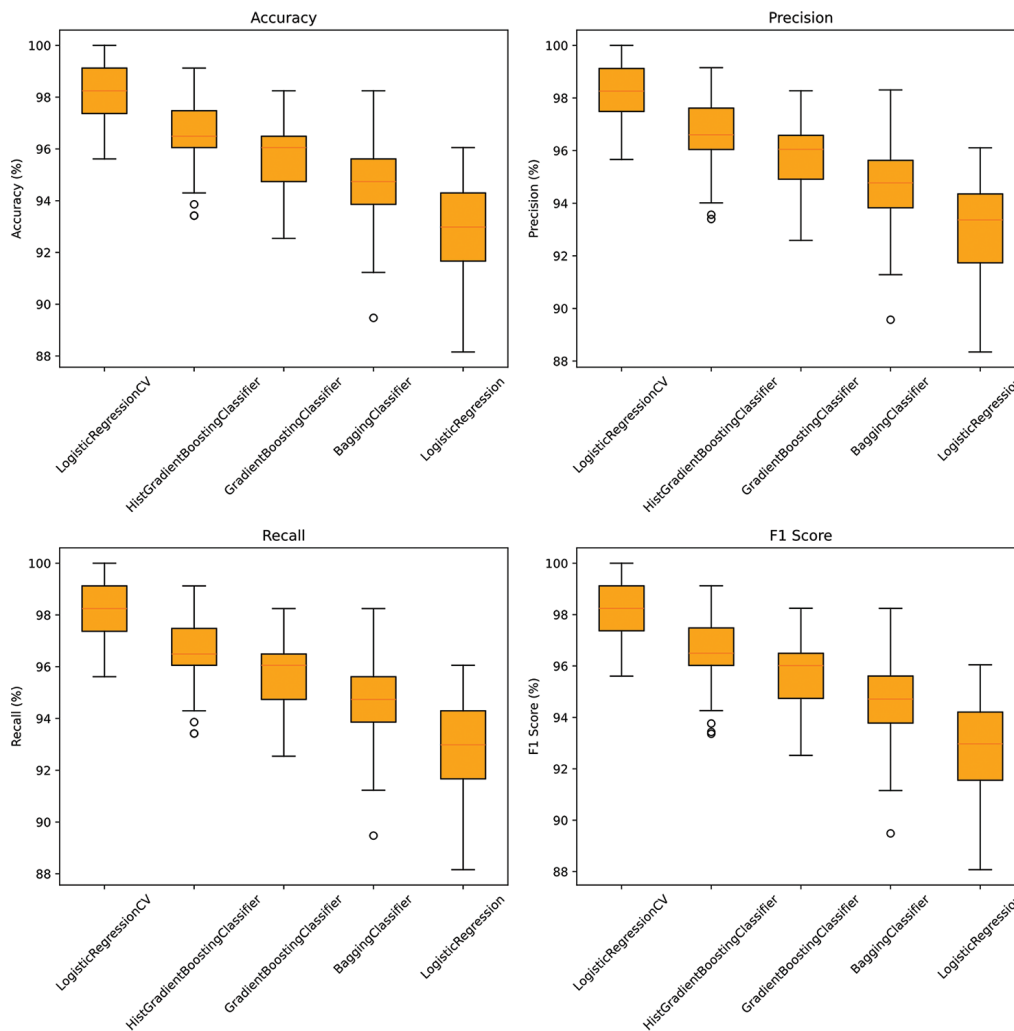


Figure 27. Performance metrics plots of the five most successful classifiers (using height and weight attributes) on the synthetic minority oversampling technique—nominal and continuous dataset

Table 5. Average performance metrics of the five most successful classifiers on the tabular variational autoencoder dataset (excluding height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVC	73.02	74.16	73.02	72.68
NuSVC	72.53	74.32	72.53	72.25
RandomForest	72.12	72.53	72.12	71.77
GradBoosting	71.51	71.31	71.51	71.12
ExtraTrees	71.31	71.66	71.31	71.09

Table 6. Average performance metrics of the five most successful classifiers on the tabular variational autoencoder dataset (using height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LogisticRegCV	97.49	97.54	97.49	97.49
HistGradBoosting	96.08	96.13	96.08	96.07
GradBoosting	94.54	94.59	94.54	94.52
Bagging	94.36	94.44	94.36	94.35
DecisionTree	92.75	92.87	92.75	92.74

from our study is that even without those direct measures, reliable classification (~75% F1 score) is possible by leveraging diet and lifestyle features through synthetic

data augmentation. This finding is clinically relevant, where in many settings (e.g., telehealth surveys, electronic records lacking anthropometric data, or privacy-preserved

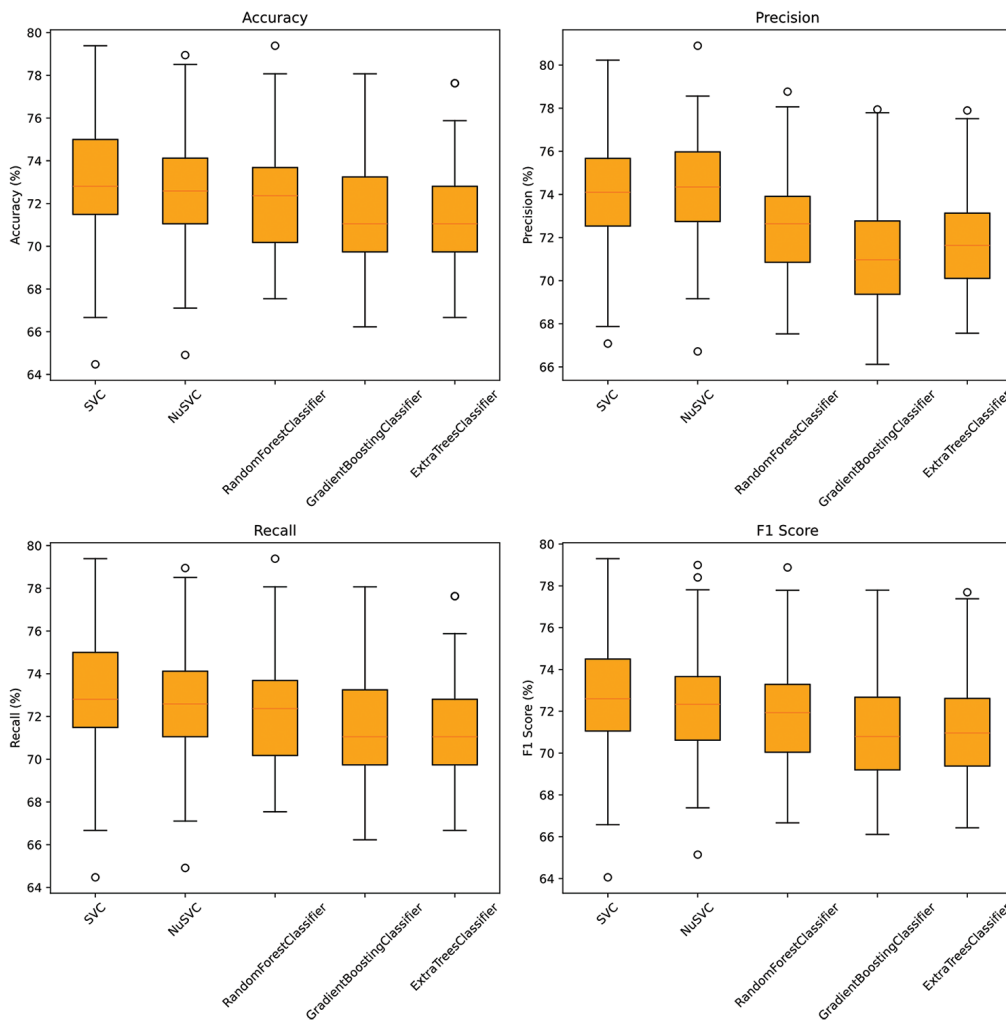


Figure 28. Performance metrics plots of the five most successful classifiers on the tabular variational autoencoder dataset (excluding height and weight attributes)

Table 7. Average performance metrics of the five most successful classifiers on the conditional tabular generative adversarial network dataset (excluding height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
GradBoosting	60.66	60.88	60.66	60.59
HistGradBoosting	59.53	59.80	59.53	59.49
RandomForest	59.25	59.03	59.25	58.92
ExtraTrees	57.40	57.34	57.40	57.19
Bagging	55.70	55.77	55.70	55.43

Table 8. Average performance metrics of the five most successful classifiers on the conditional tabular generative adversarial network dataset (using height and weight attributes)

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LogisticRegCV	97.45	97.50	97.45	97.45
HistGradBoosting	96.09	96.16	69.09	96.09
Bagging	95.53	95.64	95.53	95.53
GradBoosting	95.24	95.34	95.24	95.25
DecisionTree	94.26	94.35	94.26	94.25

research databases), height and weight may be unavailable or missing. Our results suggest that in such cases, synthetic data methods can help build models that still identify

obesity risk with reasonable accuracy. This extends the known correlation of diet/behavior with obesity. For example, higher consumption of fast foods and irregular

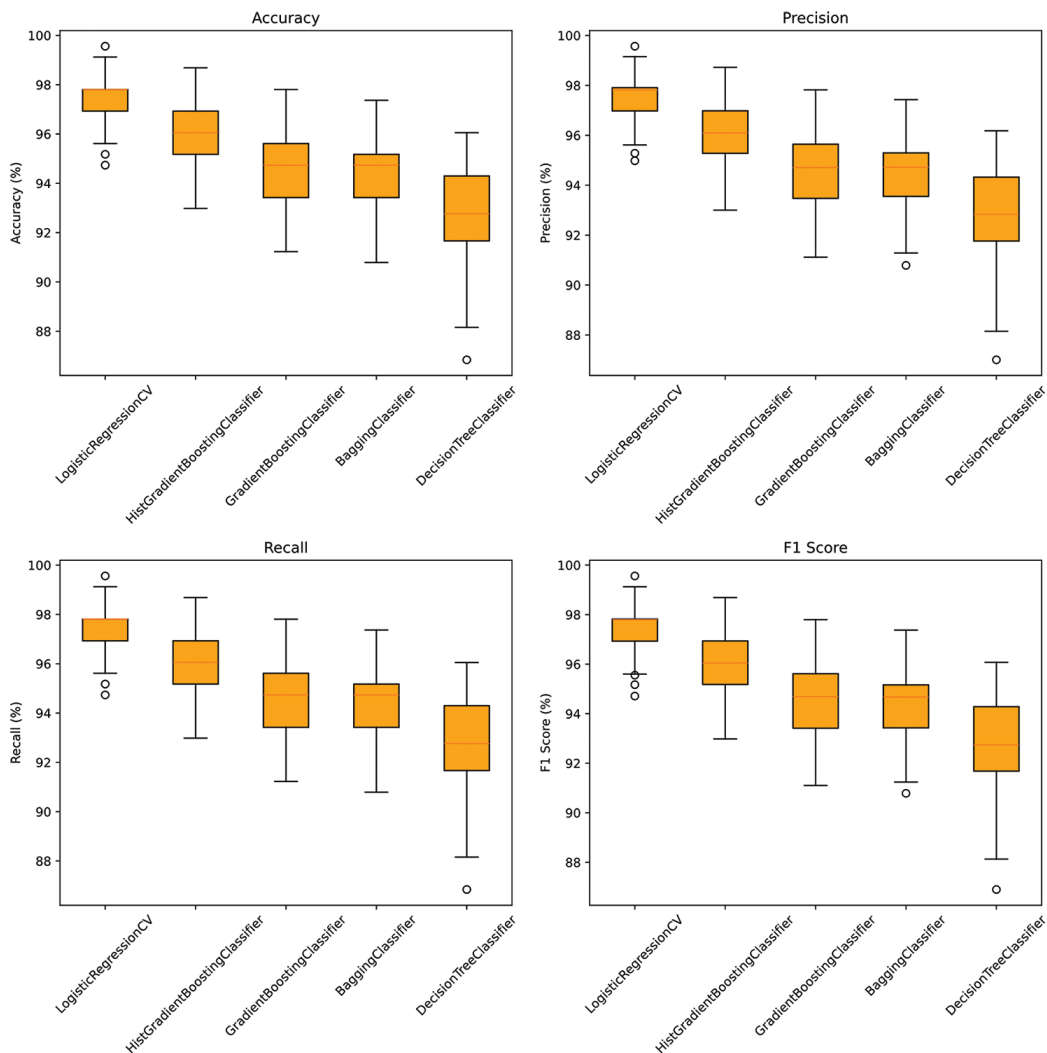


Figure 29. Performance metrics plots of the five most successful classifiers on the tabular variational autoencoder dataset (using height and weight attributes)

meal patterns are associated with higher obesity risk, and our synthetic augmentation appeared to capture these signals effectively for the ML models.

Moreover, these findings align with recent nutrition research. Colonnello *et al.*¹¹ found that dysfunctional eating behaviors (e.g., night eating) are correlated with lipid and metabolic abnormalities; we note that such behaviors are indirectly represented in our features (e.g., meal frequency, alcohol use).¹¹ El-Sehrawy *et al.*¹² reported that elevated TyG index values and disordered eating often co-occur in individuals with obesity, suggesting metabolic–diet linkages.¹² In our models, features related to eating patterns (e.g., intake of high-calorie foods, frequency of snacks) contribute

to predictions, which is consistent with these clinical findings.

Our comparison highlights practical considerations for applying generative data methods in health. Consistent with Hernandez *et al.*,⁷ we found that SMOTE-type oversampling and VAE-based generation can effectively balance and expand tabular health data.⁷ The poorer performance of CTGAN (in the no-BMI case) suggests that GAN-based approaches may require more data or tuning to capture complex categorical relationships in this dataset. Importantly, synthetic data offer benefits beyond model accuracy. Arora and Arora⁸ emphasize that fully anonymized synthetic patient data can “replace the use of real patient data in certain contexts.” In our work, all

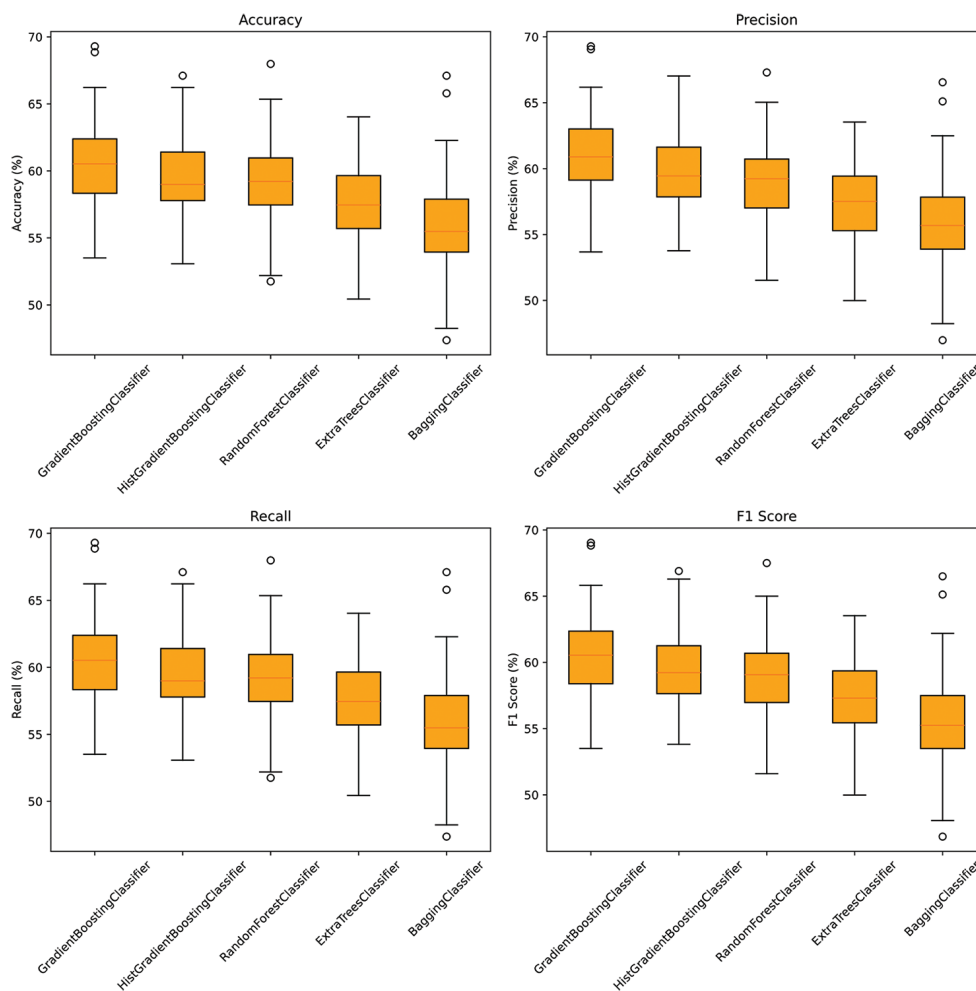


Figure 30. Performance metrics plots of the five most successful classifiers on the conditional tabular generative adversarial network dataset (excluding height and weight attributes)

synthetic examples were derived from the real EOL data, but in principle, such models could be used to generate new plausible patient profiles. This could allow researchers to share or analyze tabular health data while preserving privacy or to simulate large cohorts for training more complex models.

The EOL dataset is cross-sectional and self-reported, which limits causal inference. The synthetic data quality was not evaluated beyond model performance; future work could apply standardized metrics to quantitatively assess the resemblance and privacy of generated samples. We also noted that CTGAN’s underperformance may be due to the limited data size; experimenting with larger or multi-source datasets could test whether GANs become more reliable under such conditions. Clinically, while

our accuracy without BMI (~75% F1) is promising, it may not be sufficient for a standalone diagnosis. Rather, it suggests that such models could serve as preliminary screening tools to flag at-risk individuals for further evaluation.

In summary, our study demonstrates that ML classifiers for obesity can be trained effectively on augmented synthetic data, even when key anthropometric features are absent. This has practical relevance for nutritional and clinical practice, as it implies that an AI tool could estimate obesity risk from just diet and lifestyle information (e.g., survey responses) with reasonable accuracy. It also highlights that synthetic data generation is a viable strategy to mitigate data limitations in health research.

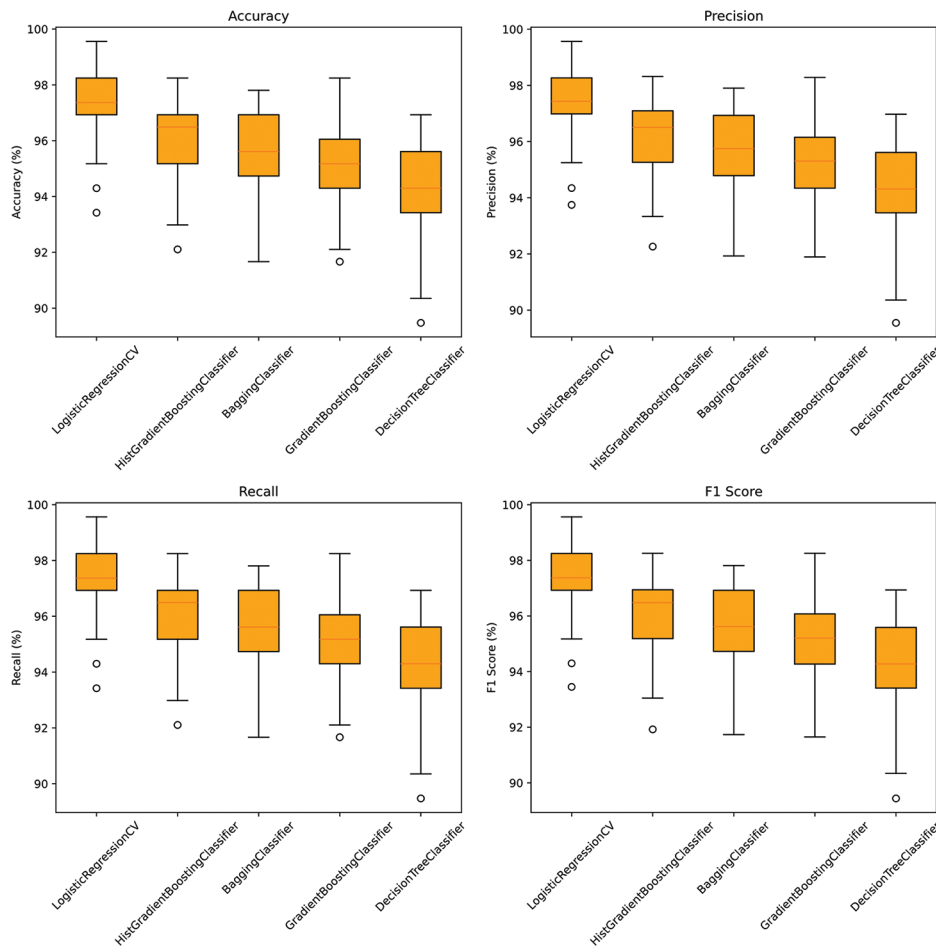


Figure 31. Plots of the performance metrics of the five most successful classifiers on the conditional tabular generative adversarial network dataset (using height and weight attributes)

5. Conclusion and future work

This study demonstrates the effectiveness of training classification models using synthetic data generated through techniques such as SMOTE-NC and TVAE, even when the original dataset is limited in size. A detailed analysis revealed that favorable classification performance can be achieved without the inclusion of height and weight attributes when using synthetic datasets generated by SMOTE-NC and TVAE. However, for the dataset generated using CTGAN, excluding height and weight features results in suboptimal model performance. In contrast, incorporating these features yields significantly improved results across all three datasets, with F1-scores approaching 100%. These findings are particularly important for obesity level prediction, as they indicate that even in the absence of direct anthropometric measures such as height and weight, synthetic data generated using appropriate techniques can support the development of reasonably accurate models

– especially with SMOTE-NC and TVAE. While SMOTE remains a widely adopted technique in the literature for synthetic data generation, this study also highlights the viability of NN-based approaches such as TVAE. In particular, classifiers trained on SMOTE-NC and TVAE datasets (excluding height and weight) achieved an F1 score of approximately 75% on the test set – an outcome not replicated with CTGAN-generated data. Future research directions include: (i) Exploring CTGAN and other generative models on larger or more diverse obesity datasets to improve synthetic fidelity; (ii) integrating additional predictive features (e.g., genetic, microbiome, or detailed metabolic biomarkers) to enhance model relevance; and (iii) conducting prospective validation of synthetic-data-augmented models in clinical or community cohorts to assess their real-world utility in preventive health. We believe that the continued development of synthetic tabular data methods will strengthen AI-driven obesity prevention and nutrition research.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: All authors

Formal analysis: All authors

Investigation: All authors

Methodology: Hakan Alp Eren, Sinem Bozkurt Keser

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

The dataset used in this study is publicly available from the University of California, Irvine ML Repository under the title *Estimation of Obesity Levels Based on Eating Habits and Physical Condition*: <https://archive.ics.uci.edu/dataset/544>.

References









1. Ural D, Kılıçkap M, Gökşülük H, *et al.* Data on prevalence of obesity and waist circumference in Turkey: Systematic review, meta-analysis and meta-regression of epidemiological studies on cardiovascular risk factors. *Turk J Cardiol Arch.* 2018;46(7):577-590.
doi: 10.5543/tkda.2018.62200
2. Yavuz R, Tontuş H. The clinical approach to the obesity in adult, adolescent and pediatric age groups. *J Exp Clin Med.* 2013;30(1s):69-74.
3. Rosengren A. Obesity and cardiovascular health: The size of the problem. *Eur Heart J.* 2021;42(34):3404-3406.
doi: 10.1093/eurheartj/ehab518
4. Dönder E, Önalın E. Definition, epidemiology, and clinical evaluation of obesity. *Firat Med J.* 2018;23(3):1-4.
5. UCI Machine Learning Repository. *Estimation of Obesity Levels Based On Eating Habits and Physical Condition*. Irvine: UCI Machine Learning Repository; 2019.
doi: 10.24432/C5H31Z
6. Shi R, Wang Y, Du M, Shen X, Wang X. *A Comprehensive Survey of Synthetic Tabular Data Generation*. [arXiv Preprint]; 2025.
doi: 10.48550/arXiv.2504.16506
7. Hernadez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods Inf Med.* 2023;62(S01):e19-e38.
doi: 10.1055/s-0042-1760247
8. Arora A, Arora A. Generative adversarial networks and synthetic patient data: Current challenges and future perspectives. *Fut Healthc J.* 2022;9(2):190-193.
doi: 10.7861/fhj.2022-0013
9. Sámano R, Lopezmalo-Casares S, Martínez-Rojano H, *et al.* Early life determinants of overweight and obesity in a sample of Mexico city preschoolers. *Nutrients.* 2025;17(4):697.
doi: 10.3390/nu17040697
10. Sobas K, Suliga E, Bryk P, Gluszek S. Dietary patterns and nutritional status in bariatric surgery candidates—a cross-sectional study. *Nutrients.* 2025;17(4):716.
doi: 10.3390/nu17040716
11. Colonnello E, Libotte F, Masi D, *et al.* Eating behavior patterns, metabolic parameters and circulating oxytocin levels in patients with obesity: An exploratory study. *Eating Weight Disord.* 2025;30(1):6.
doi: 10.1007/s40519-024-01698-w
12. El-Sehrawy AAMA, Khachatryan LG, Kubaev A, *et al.* Triglyceride-glucose index: A potent predictor of metabolic risk factors and eating behavior patterns among obese individuals. *BMC Endocr Disord.* 2025;25(1):71.
doi: 10.1186/s12902-025-01887-3
13. Kuckuck S, Van der Valk ES, Lengton R, *et al.* Long-term hair cortisol and perceived stress are associated with long-term hedonic eating tendencies in patients with obesity. *Psychoneuroendocrinology.* 2025;171:107224.
doi: 10.1016/j.psyneuen.2024.107224
14. Palechor FM, De la Hoz Manotas A. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data Brief.* 2025;25:104344.
doi: 10.1016/j.dib.2019.104344
15. Helforouh Z, Sayyad H. Prediction and classification of obesity risk based on a hybrid metaheuristic machine learning approach. *Front Big Data.* 2024;7:1469981.
doi: 10.3389/fdata.2024.1469981
16. Ayub H, Khan MA, Shehryar Ali Naqvi S, *et al.* Unraveling the potential of attentive Bi-LSTM for accurate obesity prognosis: Advancing public health towards sustainable

- cities. *Bioengineering (Basel)*. 2024;11(6):533.
doi: 10.3390/bioengineering11060533
17. Shakti MAS, Vijayalakshmi M, Kumar N, Vaidhehi M. Analysis on Various Machine Learning Framework for Obesity Level Prediction. In: *Proceedings of the 1st International Conference on Contemporary Global Challenges and Urban Innovations (ICCGUI) IEEE*. Vol. 1; 2024. p. 406-411.
doi: 10.1109/IC-CGU58078.2024.10530812
18. Yağmur N. A hybrid approach to obesity level determination with decision tree and pelican optimization algorithm. *J Sci Rep A*. 2024;57:97-109.
doi: 10.59313/jsr-a.1447814
19. Özkurt C. Examination and evaluation of obesity risk factors with explainable artificial intelligence. *Comput Electron Med*. 2024;1(1):12-17.
doi: 10.69882/adba.cem.2024072
20. Wang X. Predicting obesity risk through lifestyle habits: A comparative analysis of machine learning models. *E3S Web Conf*. 2024;385:05037.
doi: 10.1051/e3sconf/202455305037
21. Okpe OA, Odey JA, Abiodun OJ. A novel multi-class classification of obesity level using artificial neural network. *Int J Adv Multidiscip Res Studies*. 2024;4(3):1374-1379.
22. Azad M, Khan MFK, El-Ghany SA. XAI-enhanced machine learning for obesity risk classification: A stacking approach with LIME explanations. *IEEE Access*. 2025;13:13847-13865.
doi: 10.1109/ACCESS.2025.3530840
23. Solomon DD, Khan S, Garg S, *et al*. Hybrid majority voting: Prediction and classification model for obesity. *Diagnostics (Basel)*. 2023;13(15):2610.
doi: 10.3390/diagnostics13152610
24. Kaur R, Kumar R, Gupta M. Predicting risk of obesity and meal planning to reduce obesity in adulthood using artificial intelligence. *Endocrine*. 2022;78(3):458-469.
doi: 10.1007/s12020-022-03215-4
25. Muliawan A, Fauziah DA, Afrianto E. Obesity risk prediction using random forest based on eating habit parameters. *INSIDE J*. 2024;2(1):13-18.
26. Choudhuri A. A Hybrid Machine Learning Model for Estimation of Obesity Levels. In: *Proceedings of the International Conference on Data Management, Analytics and Innovation*. Vol. 137; 2023. p. 414-423.
doi: 10.1007/978-981-19-2600-6_22
27. Cervantes RC, Palacio ALH. Estimation of obesity levels based on computational intelligence. *Inf Med Unlocked*. 2020;21:100472.
doi: 10.1016/j.imu.2020.100472
28. Ganie SM, Reddy BB, Rege M. An investigation of ensemble learning techniques for obesity risk prediction using lifestyle data. *Decis Analyt J*. 2025;14:100539.
doi: 10.1016/j.dajour.2024.100539
29. Nagarajan SG, Balasubramanian V, Gonugunta P, Gudla SK. Obesity level prediction using deep learning approach-a comparative analysis. *Eng Appl Sci Res*. 2024;51(4):540-554.
30. Umoh PN, Nneji GU, Monday HN, *et al*. Optimizing machine learning classifiers and feature selection techniques for obesity levels estimation using physical habits and dietary data. *World Sci News*. 2024;198:326-353.
doi: 10.1142/WSN198(2024)325-353
31. Vairachilai S, Periyayagi S, Raja SPR. PIPR machine learning model: Obesity impact analysis. *Open Biomed Eng J*. 2024;18(1):1-20.
doi: 10.2174/0118741207289421240430115207
32. Forte P, Encarnação S, Monteiro AM, *et al*. A deep learning neural network to classify obesity risk in portuguese adolescents based on physical fitness levels and body mass index percentiles: Insights for national health policies. *Behav Sci*. 2023;13(7):522.
doi: 10.3390/bs13070522
33. Yağın FH, Gülü M, Görmez Y, *et al*. Estimation of obesity levels with a trained neural network approach optimized by the Bayesian technique. *Appl Sci*. 2023;13(6):3875.
doi: 10.3390/app13063875
34. Gözükarar Bağ HG, Yağın FH, Görmez Y, *et al*. Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits. *Diagnostics*. 2023;13(18):2949.
doi: 10.3390/diagnostics13182949
35. Yang Y, Khorshidi HA, Aickelin U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: Insights for medical problems. *Front Digit Health*. 2024;6:1430245.
doi: 10.3389/fdgth.2024.1430245
36. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J Mach Learn Res*. 2017;18(17):1-5.
37. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular Data using Conditional GAN. In: *Advances in Neural Information Processing Systems*; 2019. p. 32. Available from: <https://proceedings.neurips.cc/paper/2019/hash/254ed7d2de3b23ab10936522dd547b78-abstract.html> [Last accessed on 2024 Dec 12].
38. Patki N, Wedge R, Veeramachaneni K. The Synthetic Data Vault. In: *International Conference on Data Science and Advanced Analytics (DSAA)*; 2016. p. 399-410.
doi: 10.1109/DSAA.2016.49

39. Luo Y, Tao J, Zhu Y, Xu Y. HSS: Enhancing IoT malicious traffic classification leveraging hybrid sampling strategy. *Cybersecurity*. 2024;7(1):11.
doi: 10.1186/s42400-023-00201-9
40. Yadav P, Gaur M, Madhukar RK, Verma G, Kumar P. Rigorous experimental analysis of tabular data generated using TVAE and CTGAN. *Int J Adv Comput Sci Appl*. 2024;15(4):1250-1262.
doi: 10.14569/ijacsa.2024.01504125
41. Huang GL, Wu PY. CTGAN: Cloud transformer generative adversarial network. In: *2022 IEEE International Conference on Image Processing (ICIP)*; 2022. p. 511-515.
doi: 10.1109/ICIP46576.2022.9897229

ORIGINAL RESEARCH ARTICLE

A hierarchical federated learning-based health stack for future pandemic preparedness

Rojalini Tripathy¹, Asmit Balabantaray², Nisarg Shah², Prashant Kumar Jha³, Ajay Kumar Gogineni¹, Atri Mukhopadhyay¹, Kisor Kumar Sahu^{3,4*}, and Padmalochan Bera^{1*}

¹School of Electrical and Computer Sciences, Indian Institute of Technology Bhubaneswar, Odisha, India

²School of Mechanical Sciences, Indian Institute of Technology Bhubaneswar, Odisha, India

³School of Minerals, Metallurgical and Materials Engineering, Indian Institute of Technology Bhubaneswar, Odisha, India

⁴Virtual and Augmented Reality Center of Excellence, Indian Institute of Technology Bhubaneswar, Odisha, India

Abstract

The COVID-19 pandemic, one of the most disruptive global health crises in recent history, exposed critical vulnerabilities in existing healthcare infrastructure. Given the likelihood of future pandemics, it is essential to build a resilient, collaborative, synergistic, data-driven, and intelligent digital healthcare software. It should be meticulously designed and selectively curated to enhance early detection, rapid response, and efficient containment of outbreaks. In this article, we propose a federated learning (FL)-based health stack that prioritizes privacy while fostering collaborative intelligence among hospitals or client nodes. Our framework incorporates hierarchical FL, Byzantine-resilient information-theoretic FL (ByITFL), homomorphic encryption, and blockchain-based smart contracts to ensure secure collaboration among healthcare institutions without sharing raw data. Hierarchical FL leverages multilevel model aggregation to enhance model convergence, scalability, and resilience. ByITFL strengthens security by incorporating trust mechanisms and information-theoretic privacy scoring, while blockchain-based smart contracts ensure transparent, verifiable coordination among participating nodes. Furthermore, deep vulnerability detection using optimized averaged stochastic gradient descent weight-dropped long short-term memory models may further enhance the framework's security, enabling threat identification during decentralized data exchanges. Experimental results show that the proposed hierarchical FL model achieves 94.23% accuracy on the modified National Institute of Standards and Technology dataset, outperforming federated averaging (92.66%) under the same environments. In addition, communication analysis proved that the overall transmission is minimized by collecting updates at local servers before sending them to central servers. Therefore, it is nearly a future-ready technology that can be implemented without many geopolitical issues, even in the case of hypersensitive global situations.

Keywords: Global pandemics; Health stack; Federated learning; Medical data privacy; Machine learning

***Corresponding authors:**

Kisor Kumar Sahu
(kisorsahu@iitbbs.ac.in);
Padmalochan Bera
(plb@iitbbs.ac.in)

Citation: Tripathy R, Balabantaray A, Shah N, *et al.* A hierarchical federated learning-based health stack for future pandemic preparedness. *Artif Intell Health*. 2025;2(4):75-91. doi: 10.36922/AIH025080013

Received: February 21, 2025

1st revised: May 22, 2025

2nd revised: June 8, 2025

Accepted: June 9, 2025

Published online: June 30, 2025

Copyright: © 2025 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, which provided that 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

1.1. General overview: The big picture

The emergence of the COVID-19 pandemic in 2019 shook the foundations of human civilization. While it was not the deadliest pandemic in history, it ranks fifth in terms of death toll, with an official World Health Organization (WHO) estimate of 7 million deaths.¹ When ranked by death toll, the following pandemics are the deadliest, in descending order: (i) the Spanish flu (17 – 100 million deaths, 1918 – 1920),² (ii) the Plague of Justinian (15 – 100 million deaths, 541 – 549),³ (iii) human immunodeficiency virus/acquired immunodeficiency syndrome (approximately 43 million deaths, 1981–present),⁴ and (iv) the Black Death (7 – 35 million deaths, 1346 – 1353).⁴

What sets COVID-19 apart from previous pandemics is that it occurred in a highly globalized, post-internet world, where information could spread instantaneously.⁵ Combined with significant advancements in modern sciences, particularly in medical and computational sciences, it was shocking to see how the pandemic exposed the limitations of human technological capabilities. As argued in previous studies,⁵ although we had the components of a modern technological infrastructure to create a formidable defense, we failed to synergize them effectively to prevent the pandemic's escalation.

Therefore, the central question addressed in this article is: What would the next-generation computational infrastructure look like to effectively combat future pandemics? Such infrastructure should be deployable within current medical record-keeping systems, which involve heterogeneous data types and regulations (mostly privacy-related). To stimulate this discussion, we propose a federated learning (FL)^{6,7}-based, global machine learning (ML) architecture with the potential to address this key issue. While we do not claim this is the ultimate solution (though we believe it may be), the main purpose of this article is to motivate the journey towards that goal. For brevity, we will highlight one or a few methods for each step, although multiple viable solutions may exist. A comprehensive review of FL applications in smart healthcare is presented elsewhere;⁸⁻¹⁰ hence, we will only summarize key technical preliminaries of FL here.

In the absence of a global enforcing agency for administering critical pandemic/medical issues (as the WHO lacks such authority), a decentralized approach offers the most promising pathway for global acceptance and implementation. FL is a decentralized ML approach in which clients (e.g., organizations, healthcare units, mobile devices, and sensors) collaboratively train a shared

model under the coordination of a central server without exchanging raw data.

The FL process typically begins with the central server initializing a global model and distributing its parameters to selected clients. Each client independently trains the model using its local dataset, ensuring data privacy by keeping raw data on the local device. After local training, clients send model updates, such as model gradients, to the central server. The server then aggregates these updates, commonly using methods such as federated averaging (FedAvg),¹¹ to update the global model. The updated model is then redistributed to the clients for the next round of training. This cycle continues iteratively until the model converges or reaches a predefined accuracy threshold. Finally, the trained global model is deployed for use.

As discussed, FL eliminates the need to transfer raw data to a central server; instead, only model parameters are exchanged. This approach significantly reduces the risk of data breaches, as sensitive information remains on the client's side. FL thus upholds data privacy and security, making it one of the highly accepted protocols across countries, communities, and agencies. For this reason, it has been chosen in this article as the foundational model for addressing future pandemics.

We envision that combating future pandemics will require the holistic deployment of the entire computing infrastructure, including, but not limited to, the Internet-of-Things (IoT) devices, sensors, edge computing, and cloud infrastructure. While FL offers privacy-preserving model training, its integration with emerging technologies, such as edge intelligence and IoT, in healthcare introduces new challenges that must be addressed. Recent studies have highlighted the growing need for edge intelligence and distributed learning in health care.^{12,13} For example, research presents an overview of mobile health systems and IoT technologies, focusing on system architecture and data integration strategies.¹⁴ However, centralized or cloud-based models often require raw or partially processed data to be transmitted, raising concerns about data security in sensitive applications, such as health care. In another study,¹³ the authors discussed the role of artificial intelligence in IoT-enabled smart healthcare applications, with a primary focus on lightweight neural network architectures deployed directly on edge devices. Yet, the study does not address the security risks associated with transmitting trained model parameters from edge devices. This omission overlooks potential risks such as adversaries intercepting and reverse-engineering the model parameters to extract sensitive information, or manipulating them to induce bias or abnormal model behavior. In our proposed framework, we use a

privacy-preserving, hierarchically structured FL approach to address these challenges.

We present a secure and scalable FL framework using a hierarchical structure and homomorphic encryption (HE) for privacy-preserving, collaborative model training. Then, we discuss the main features of Byzantine-resilient information-theoretic FL (ByITFL),¹⁴ which offers privacy-preserving aggregation and robust protection against communication-based attacks in FL through standard security techniques. Subsequently, we illustrate an implementation of a deep-learning protocol based on modern convolutional neural network (CNN) architectures that efficiently detect past pandemics like COVID-19 and diseases like lung cancer, with the aim of improving accuracy and computational efficiency by leveraging transfer learning, clustering algorithms, and region-of-interest (ROI) analysis. Blockchain/smart contracts play a key role in establishing digital trust and maintaining transparency, both of which are critical for success in a global endeavor, such as the fight against pandemics. Unfortunately, they are not free from external vulnerabilities. We discuss detecting vulnerabilities in smart contracts using an improved averaged stochastic gradient descent weight-dropped long short-term memory (AWD-LSTM) model, which achieves high accuracy and F1 score by incorporating opcode review analysis and addressing class imbalance in smart contract datasets. In Section 3, we experimentally evaluate our proposed framework and present a detailed cost analysis of the hierarchical FL¹⁵ framework, showing its scalability through efficient communication and computation at client, local server, and central server levels. Communication costs are optimized using hierarchical data aggregation, while computation costs are minimized through localized operations, ensuring scalability and security. In Section 4, we address the issue of non-independent and identically distributed (non-IID) datasets and heterogeneity considerations, both of which are critical for the efficient implementation of our idea.

The FL model that we have described (including hierarchical FL, ByITFL, and HE) is more realistic, as it takes into consideration real-world applications. It is also suggested to use institutional incremental learning¹⁰ and cyclic institutional incremental learning.¹⁰ FL achieves a better rate of model improvement than data-private collaborative learning methods. Moreover, to compare the rates of model improvement, a global validation “dice-over-epoch” (where the model trains over epochs and the metric is computed on both training and validation data) for all collaborative methods showed that FL training converges relatively quickly to the same performance as collaborative data sharing training.

1.2. Technical preliminaries

In this section, we discuss some technical preliminaries used in the architectural design of the proposed hierarchical FL, such as static and dynamic datasets, hierarchical learning,¹⁶ and HE.¹⁷ A static dataset does not change over time, whereas a dynamic dataset is continuously or periodically updated with new data. HE is a privacy-preserving technique that enables computations, such as addition and multiplication, to be performed directly on encrypted data. When data or model parameters are transmitted to a server, there is a risk of interception or tampering by attackers, potentially compromising model integrity and predictions. In the proposed framework, we employ HE to encrypt model parameters before transmission. These encrypted parameters are then aggregated on the server using homomorphic addition and multiplication operations, ensuring both data privacy and secure computation. FL frameworks may suffer from scalability limitations due to communication bottlenecks that arise when a large number of clients frequently transmit model updates to a centralized server. This issue becomes more pronounced in distributed environments, such as health care.

To address this, we employ hierarchical learning.¹⁸ Hierarchical learning is a structured learning process organized across multiple levels, where higher levels aggregate and refine knowledge from lower levels to improve performance and scalability. It has a multi-tiered structure comprising a central server along with multiple local servers at each level. This structure efficiently handles gradient aggregation and transmission in parallel, which reduces overall communication overhead and ensures privacy. [Figure 1](#) represents a hierarchical learning architecture suitable for deployment in the global healthcare system.

We experimentally evaluated our proposed framework using a proxy image classification dataset, implementing a hierarchical FL architecture and transmitting model parameters through HE. We compared its performance against standard FL lacking hierarchical structure and encrypted communication. Our experimental results demonstrate that the hierarchical framework achieved higher model accuracy and reduced training time compared to generic FL. The hierarchical structure effectively distributes the computational and communication load, enabling faster convergence. Additionally, the use of HE ensures that sensitive information remains confidential throughout the training process, thereby satisfying privacy requirements without significantly compromising efficiency. The proposed model is capable of incorporating both static and dynamic data. However, future pandemics may involve highly distributed or heterogeneous data types.

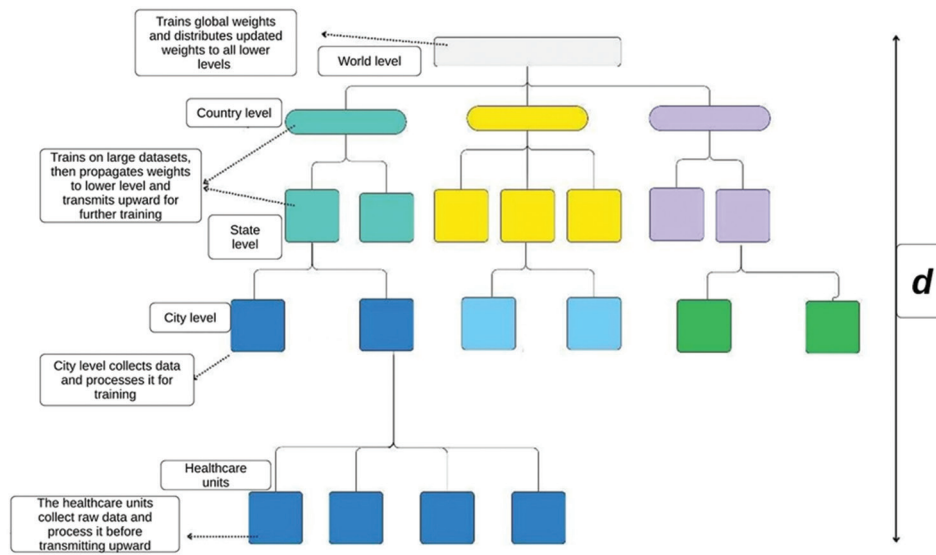


Figure 1. Schematic representation of the hierarchical learning framework implemented in the proposed FL scheme. Here, d represents the depth of the hierarchy, and root may represent the central organization that controls all subordinate levels (e.g., World Health Organization).

Our framework can be extended to train multiple specialized models at the central unit, depending on data characteristics and application requirements. In addition, given the assumption that future pandemics will generate massive volumes of data, we anticipate that the hierarchical structure of our framework can be scaled accordingly by incorporating additional aggregation layers or nodes to efficiently manage the increased load and maintain performance.

2. Proposed FL framework

In this section, we described our proposed secure and scalable FL framework, capable of incorporating dynamic data during model training. In collaborative medical research, to enhance predictive diagnostic accuracy, multiple healthcare data owners aim to train a unified model (or a set of unified models, depending on the nature of a future pandemic). The trained model should be available locally to all collaborators. In scenarios where a single data owner operates multiple healthcare units, privacy concerns are minimized, as such organizations can consolidate their data internally without external exposure. For example, in the United States (US), Kaiser Permanente operates a network of hospitals and clinics as a unified entity. Similarly, in India and internationally, Apollo Hospitals manages an extensive network of hospitals, clinics, and pharmacies. The proposed framework allows these data owners to centralize patient data for analysis and predictive modeling while ensuring strict internal control and compliance with privacy regulations. Figure 2 illustrates the architectural structure of our proposed framework. Here, data owners refer to entities (such as Kaiser Permanente or Apollo

Hospitals), while data units include their affiliated hospitals, clinics, and pharmacies. These data owners interact with designated local servers rather than directly with the central server. Each data owner conducts private training on its localized data and communicates model parameters to the respective local server. The local servers aggregate these parameters and relay them to the central server.

Our proposed hierarchical FL framework effectively addresses security risks during communication phases and at the server level. The hierarchical structure enables scalability by accommodating a wide range of healthcare data, irrespective of its volume.

2.1. Secure model training procedure

The proposed framework consists of three primary entities: the central server, local servers, and local healthcare data holders acting as clients. The central server is responsible for global aggregation, i.e., the final updated parameters will be computed at the central server. The local or regional servers are responsible for aggregation at the cluster or regional level, while the client organizations train their local models on the datasets they manage locally.

Let S and L_i represent the central server and a regional server, respectively, and n denote the number of regional servers, where $1 \leq i \leq n$. Each regional cluster may contain a different number of clients. Let p be the number of clients under L_p , with each client denoted as C_{ij} , where i corresponds to the local server and j represents the client, where $1 \leq j \leq p$. Every client C_{ij} holds a private dataset D_{ij} .

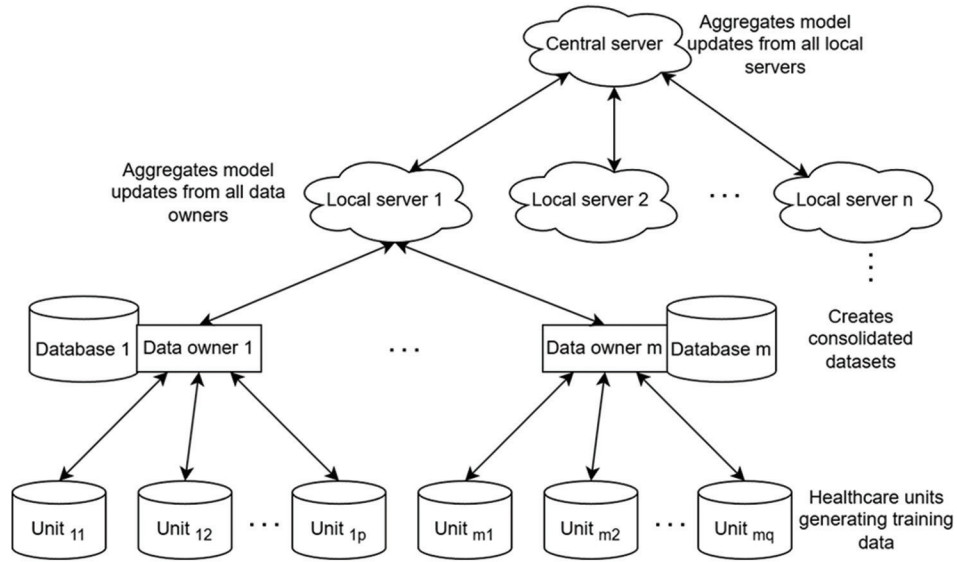


Figure 2. Architectural structure of the proposed hierarchical federated learning

The training process begins with the central server (S) broadcasting the initialized model parameters W_0 to all local servers, ensuring uniformity across local models for their cluster-level distributions. After receiving the initialized model parameters, each C_{ij} trains its local model on a batch of data b_r^{ij} , sampled from its dataset D_{ij} , and generates local model gradients G_r^{ij} for round r . This is represented as:

$$G_r^{ij} \leftarrow \text{train}(W_0, b_r^{ij}) \tag{I}$$

where r represents the current round index, and R is the total number of training rounds. The symbol “ \leftarrow ” denotes the assignment operator, indicating that the result of the operation on the right-hand side is assigned to the variable on the left-hand side. The proposed framework uses HE¹⁹ operations to ensure the security of the model gradients during client-server communications and employs a hierarchical structure to enhance scalability. The depth of the hierarchy can be increased to ensure the required scaling. During the transmission of model gradients, a key-generating authority generates a public-private/secret key pair (pk_r^{ij}, sk_r^{ij}) for each client C_{ij} at the beginning of each round r . Once the keys are received, each client encrypts its local gradients G_r^{ij} using HE. This is represented as:

$$[[G_r^{ij}]] \leftarrow \text{encrypt}(pk_r^{ij}, G_r^{ij}) \tag{II}$$

After encryption, every client C_{ij} transmits the encrypted gradients $[[G_r^{ij}]]$ (here $[[\cdot]]$ represents encrypted values) to its corresponding local server L_i . Upon receiving the encrypted model updates from all clients within the cluster, the local server L_i performs homomorphic

aggregation. As HE supports only addition and multiplication, we use homomorphic addition to find the aggregate sum and then multiply the aggregate sum by the reciprocal of the number of clients in each cluster. This can be represented as:

$$[[G_r^i]] \leftarrow \frac{1}{|p|} \sum_{j \in p} [[G_r^{ij}]] \tag{III}$$

where, $|p|$ represents the cluster size. Following this, each local server L_i transmits aggregated encrypted model parameters $[[G_r^i]]$ from their respective clusters to the central server S. The central server further performs global aggregation on received model parameters. This can be represented as:

$$[[G_r]] \leftarrow \frac{1}{|n|} \sum_{i \in n} [[G_r^i]] \tag{IV}$$

where $|n|$ represents the number of participating local clients. The central server S then sends the aggregated model parameters back to the local servers, which are responsible for disseminating these parameters to the clients in their respective clusters. Before proceeding to the next round of training, each client C_{ij} decrypts the received aggregated encrypted model parameters using the private key sk_r^{ij} for that round. This decryption process can be represented as:

$$G_r \leftarrow \text{decrypt}(sk_r^{ij}, [[G_r]]) \tag{V}$$

Each client C_{ij} then proceeds to train its local model on another batch of data b_{r+1}^{ij} , which may either be static or dynamically collected. This training is performed on the

updated model parameters G_r , generating new model parameters for round $r + 1$, represented as:

$$G_{r+1}^{ij} \leftarrow \text{train}(G_r, b_{r+1}^{ij}) \tag{VI}$$

The hierarchical communication process is repeated using HE for each subsequent round. This process continues until either the final round R is reached or the model converges. The workflow of our proposed framework is diagrammatically represented in Figure 3, and the algorithm is as follows:

Algorithm: 1 Hierarchical federated learning with homomorphic encryption

Require: Initial global model W_0 , rounds R , round count $r = 0$, local servers n , clients C_{ij}

Ensure: Final global model W^*

- 1: Central server S broadcasts W_0 to local servers L_i , where $1 \leq i \leq n$;
- 2: **for** $r = 1$ to R **do**
- 3: Each L_i distributes W_0 to clients C_{ij} ;
- 4: Key authority shares encryption keys (pk_r^{ij}, sk_r^{ij}) ;
- 5: **for** each client C_{ij} **do**
- 6: Each client decrypts $[[G_r]]$ as G_r , where $r \leq 2 \leq R$;
- 7: Train local model on G_r and $b_r^{ij} \subseteq D_{ij}$ to get G_r^{ij} ;
- 8: Encrypt G_r^{ij} and send $[[G_r^{ij}]]$ to L_i ;
- 9: Each L_i aggregates collected $[[G_r^{ij}]]$ as and $[[G_r^i]]$ sends to S ;
- 10: S performs global aggregation to obtain $[[G_r]]$ and updates W_{r+1} ;
- 11: S sends W_{r+1} to local servers for client distribution;
- 12: S broadcasts final model W^* through local servers.

2.2. Secure and reliable network architecture

We discussed the possible security challenges in the FL network by addressing both malicious and unintended disruptions caused by clients or communication vulnerabilities, for example, Byzantine attacks, gradient inversion, and packet sniffing. Solutions such as ByITFL,¹⁴ trust bootstrapping via root datasets (FLTrust),²⁰ and client similarity analysis (e.g., FoolsGold)²¹ ensure secure updates through trust and information-theoretic privacy scores. In addition, a cryptographic framework that blends ByITFL^{14,22} and FedML-HE²³ to address the dual challenges of adversarial client behavior and privacy leakage in medical FL deployments is provided.

2.3. Threat model and security requirements

In FL, systems are susceptible to a range of attacks against security and privacy that compromise model performance and data privacy. The most critical of these issues occurs when malicious clients under an attacker’s control submit

poisoned updates, such as label-flipping or trimmed mean methods, to induce global model divergence. Experiments have shown that only 10% of such malicious clients may reduce model accuracy by up to 33.1%,²⁴ which highlights the need for robustness against such attacks.

Another urgent concern is privacy leakage. Even if raw data is retained locally, leakage of gradient distributions might compromise sensitive information. Attacks on medical imaging data, for example, have resulted in up to 74.24% reconstruction success rates for all data samples in the medical segmentation decathlon liver dataset from model updates²⁵ (for privacy parameter, $\epsilon = 20$). Therefore, robust privacy-preserving mechanisms are essential.

A deployable FL system in healthcare should be robust, i.e., capable of operating in unreliable network environments, especially in wireless configurations with as much as packet loss rates (approximately 10%) for maximum users (approximately 90%).^{26,14} These interruptions cause stale or missing updates, leading to unwanted divergence during model training.

To address the above challenges of privacy and security, a recent article by Xia *et al.*¹⁴ introduces ByITFL, an FL-based privacy-preserving secure aggregation scheme. Here, Byzantine basically refers to clients that are malicious or behave arbitrarily, typically with the intention of poisoning training. ByITFL is designed to maintain the integrity and confidentiality of model updates during aggregation, without relying on computational assumptions. The scheme is information-theoretically private, i.e., even a semi-honest or curious server learns nothing beyond the aggregated gradients. Privacy is quantified using Shannon entropy (H):²² for a finite field F , the local update entropy of a client X , given on the server’s observations, must satisfy

$$H(X|View_{server}) \geq \log|F| \tag{VII}$$

ensuring that individual model inputs are infeasible to reconstruct. Moreover, to be immune to Byzantine attacks, ByITFL incorporates robustness requirements from distributed consensus theory. Specifically, the system can tolerate up to t malicious clients only if the total number of clients N satisfies the inequality:^{14,27}

$$N \geq 3t + 1 \tag{VIII}$$

This barrier ensures that there are sufficient correct clients to regulate the aggregation process so that the system can reject erroneous updates and achieve stable convergence.

2.4. ByITFL cryptographic framework

The ByITFL framework integrates three cryptography primitives to achieve Byzantine resilience and

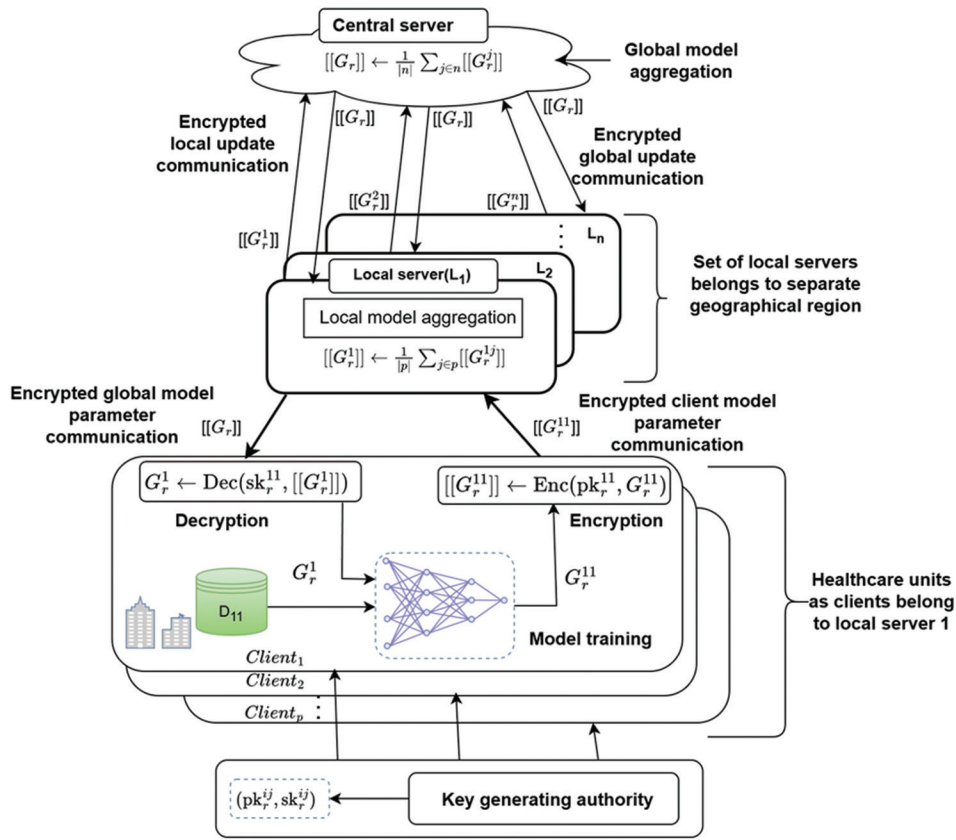


Figure 3. Workflow of the proposed hierarchical federated learning (See Section 2.1 for details)

information-theoretic privacy. Lagrange-coded computing (LCC)²⁷ offers robust aggregation in the presence of hostile clients by spreading gradient computation over participants. For clients tolerating up to malicious actors, LCC encodes gradients as:

$$\bar{g}_i = \sum_{k=1}^K g_k \prod_{1 \leq m \leq K, m \neq k} \frac{\alpha_i - \beta_m}{\beta_k - \beta_m}, m \neq k \quad (\text{IX})$$

where α_i is a client-specific evaluation point, β_m is the Lagrange interpolation point for client m , β_k is the interpolation node ensuring redundancy, g_k is the raw gradient of the k^{th} client, and K is the number of honest clients. This ensures proper aggregation if $N \geq 3t + 1$. Verifiable secret sharing employs Pedersen commitments:

$$C(\bar{g}_i) = g^{\bar{g}_i} h^{ra_i} \text{ mod } P \quad (g, h \in \mathbb{G}, ra_i \leftarrow \mathbb{Z}) \quad (\text{X})$$

Where $C(\bar{g}_i)$ stands for the adherence to the encoded gradient \bar{g}_i , P stands for a large prime number, g and h are the generators of a cyclic group \mathbb{G} , and ra_i is a random number from the set of integers \mathbb{Z} .^{14,22} This enables zero-knowledge proof of gradient consistency without exposing raw updates. To prevent gradient inversion attacks, a

second-degree polynomial ($\overline{\text{ReLU}}$) approximation substitutes the non-differentiable trust scoring with:

$$\overline{\text{ReLU}}(x) = \frac{1}{4}(x^2 + x) \quad (\text{Degree 2 approximation}) \quad (\text{XI})$$

where x is the cosine similarity input. This ensures ϵ -differential privacy ($\epsilon = 0.25$) during trust score computation.^{14,22}

2.5. ByITFL aggregation protocol

The protocol is implemented in three phases. In trust initialization, the server has a root dataset D (e.g., 50 – 100 reference samples) to obtain a reference gradient:

$$g_0 = \frac{1}{|D_{\text{root}}|} \sum_{v_j \in D_{\text{root}}} \nabla_{\theta} \text{Loss}(f_{\theta}(v_j), y_j) \quad (\text{XII})$$

Where v_j and y_j are input-label pairs, θ denotes model parameters, and Loss is the loss function. ∇_{θ} is a gradient operator. It computes the partial derivatives of the loss function with respect to all the parameters in θ , indicating how the model parameters influence the loss. Secure trust

scoring evaluates client updates with a polynomial \overline{ReLU} function:

$$TS_i = \overline{ReLU} \left(\frac{\langle g_0, \bar{g}_i \rangle}{\|g_0\| \cdot \|\bar{g}_i\|} \right) \tag{XIII}$$

Where TS_i is the trust score for client i , g_0 is the reference gradient, and \bar{g}_i is the encoded gradient.²² Here, the dot product $\langle g_0, \bar{g}_i \rangle$ gives the scalar projection, measuring how directionally aligned the two gradient vectors are, while the Euclidian norm in the denominator, which measures the magnitude of the vectors g_0 and \bar{g}_i , normalizes this measure so that it lies in $(-1, 1)$. The updates with $TS_i < 0.1$ can be discarded. For LCC-based aggregation, global gradients are computed as:

$$g_{global} = \frac{\sum_{i=1}^N TS_i \bar{g}_i}{\sum_{i=1}^N TS_i} + \eta \left(\eta \sim \mathbb{N} \left(0, 0.1 \|g_{global}\|^2 \right) \right) \tag{XIV}$$

where η is the re-randomization noise drawn from a zero-mean Gaussian distribution \mathbb{N} with a standard deviation equal to 10% of the global gradient's ℓ_2 norm. Such noise interferes with deterministic patterns across training iterations, making gradient memorization or reverse engineering less likely without compromising update fidelity.^{14,27}

2.6. Integration of FedML-HE with ByITFL for enhanced privacy

The combination of FedML-HE's selective HE²³ and ByITFL's Byzantine-resilient architecture²⁸ offers an optimally balanced solution to FL's dual challenges of privacy and security, providing a secure and reliable network architecture. By using FedML-HE's parameter-level encryption on sensitive gradients that are detected through ByITFL's trust scoring,²⁸ the hybrid solution is able to effectively protect important model updates against inversion attacks²³ while maintaining Byzantine resilience. Empirical evaluations have demonstrated that selective encryption can cut down on communication overhead by up to 10 times for large models such as ResNet-50²³ without a loss in malicious client detection accuracy. Earlier work on Byzantine-resilient secure aggregation architectures points to the possibility of combining cryptographic privacy with adversarial robustness. In addition, FedML-HE's performance-optimized encryption pipeline is optimally compatible with ByITFL's computationally limited architecture. The hybrid solution helps prevent privacy leakage risks in trust-based aggregation while maintaining the system's ability to filter poisoned updates, a circumstance verified by cross-institutional medical FL trials.²³

2.7. Deep-learning protocol

In the previous section, the FL architecture was outlined. However, specific deep-learning models that might be integrated to address future pandemics were not discussed. As future pandemics are uncertain, to illustrate the point, COVID-19 and lung cancer were taken as benchmarks for potential adaptations of deep-learning protocols.

In this context, the study by Gogineni *et al.*²⁹ was chosen as a reference. The study investigated the potential of deep-learning models for automated COVID-19 detection using chest X-ray images, presenting a promising alternative to the current gold standard – reverse-transcription polymerase chain reaction (RT-PCR) test. This choice has two distinct merits. First, COVID-19 represents the most recent pandemic for benchmarking. Second, choosing images as input data type. This choice is crucial as image data is representative of various real-world medical datasets and therefore can be used as a reliable proxy (since the exact nature of future pandemics is unknown). Moreover, images are one of the most complex and prevalent data types in the medical arena; therefore, demonstrating with proxy images offers one of the most effective benchmarking strategies to mimic real-world complexities. Although videos represent more complex data types, they are far less frequent in the medical context and, in a crude sense, can be considered as a sequential stacking of images with time-series data (audios, if any) added in an additional channel.

In the study by Gogineni *et al.*,²⁹ several CNN architectures were implemented, including ResNet34,³⁰ SeResNext50,^{31,32} DenseNet121,³³ and EfficientNet.³⁴ These models were chosen for their distinct advantages in image classification tasks. ResNet34 utilizes skip connections, allowing for efficient training of deep networks, while SeResNext50 incorporates squeeze-and-excitation blocks, which recalibrate channel-wise feature responses for improved representational capacity. Meanwhile, DenseNet121, with its dense connections between layers, facilitates feature reuse and enhances information flow. Finally, EfficientNet models are designed using neural architecture search, optimizing the balance between accuracy and computational efficiency.³⁴ Transfer learning, using the ImageNet dataset,³⁵ was employed to improve model performance on the relatively limited medical image dataset. A learning rate scheduler³⁶ and a one-cycle training policy were also implemented for better convergence and generalization.

The models' performance demonstrated encouraging results.²⁹ ResNet34 and DenseNet121 achieved the highest overall accuracy of 94.09% in classifying images as COVID-19, normal, or pneumonia. This accuracy is considerably higher than the typical 70 – 80% sensitivity

reported for RT-PCR tests. A more detailed analysis revealed that EfficientNet exhibited the highest specificity (95.4%) for COVID-19, while ResNet34 showed the highest sensitivity (94%). Interestingly, EfficientNet showed the highest performance in classifying normal cases, SeResNext50 excelled in classifying pneumonia cases, and ResNet34 was most effective for COVID-19 cases. These results are comparable to, and in some cases outperform, those reported in earlier studies using CNNs for COVID-19 detection from chest X-rays and computed tomography (CT) scans.³⁷⁻³⁹ The present deep-learning approach offers advantages in terms of speed, resource efficiency, and potentially greater accuracy compared to RT-PCR. The approach is particularly appealing for resource-limited settings where rapid screening is critical. Moreover, the decoupled workflow, separating image acquisition and diagnostic evaluation, offers greater operational flexibility and scalability, aligning with the principles outlined in the previous study²⁹ for building a robust pandemic response.

A previous study⁴⁰ explored a novel approach for lung cancer detection, combining an unsupervised clustering algorithm for ROI proposals with CNNs.^{41,42} The modularity optimization-based graph clustering method^{41,42} applied to preprocessed CT scans from the 2016 lung nodule analysis dataset^{43,44} reduces CNN complexity by identifying potential nodule locations. This preprocessing step includes lung segmentation using marker-controlled “watershed segmentation” on Sobel-filtered images, focusing the analysis on relevant areas and reducing the computational burden. The segmented lung regions are then converted into a network, with pixels representing nodes and edges connecting neighboring pixels. This network is then clustered using a modularity function optimized for spatially embedded networks,^{41,42} which has been successfully used previously in analyzing granular assemblies.⁴³ This method effectively identifies nodules based on grayscale-intensity similarity, generating ROI proposals for subsequent CNN analysis.

The proposed method utilizes these ROI proposals to streamline the CNN classification process. Rather than relying on computationally expensive selective search⁴³ or fully labeled datasets required for methods like MaskRCNN,⁴⁵ this approach employs the clustering algorithm to generate a manageable number of fixed-size ROIs. These ROIs are then fed into a CNN, trained using transfer learning initialized with ImageNet weights,³⁵ and optimized using learning rate scheduler techniques.³⁶ This training strategy, employing discriminative learning rates for different layers of the network, facilitates efficient learning of both general and dataset-specific features, which are of great value for the FL learning protocols discussed in this article. Several pre-trained CNN

architectures, including ResNet50,³⁰ SeResNext50,^{31,32} and DenseNet161,³³ were evaluated, with DenseNet161 demonstrating the optimal performance. The CNN classifies each ROI as cancerous or non-cancerous, providing not only diagnostic information but also, combined with the clustering results, an approximate segmentation mask. This approach bypasses the need for computationally intensive semantic segmentation techniques, such as GrabCut⁴⁶ or context-aware masks,⁴⁷ offering a more efficient method for identifying and localizing cancerous nodules within the lung.

2.8. Public data authentication and blockchain

The previous section discussed the need to ensure a smooth communication protocol between the central server, local servers, and data-owning entities, which necessitates entering into a legal contract between the parties. It is important to note that, ideally, the central server should be owned either directly by the United Nations or any of its subsidiaries, such as the WHO. Due to the lack of specific sovereign oversight over the entire protocol, creating digital trust is imperative. Fortunately, a robust technical solution is available to build trust in the digital space between unknown agencies through a decentralized protocol called blockchain. Blockchain-based smart contracts offer an attractive method that creates a transparent technological framework governing the relations between the parties participating in the FL architecture. However, such smart contracts are not free from threats, and in the following, a few novel methods to ensure data safety in smart contracts are discussed.

Detecting vulnerabilities in smart contracts is crucial due to their immutable nature and the potential for significant financial losses, as evidenced by past incidents, including the decentralized autonomous organization⁴⁸ hack and the Parity wallet freeze.⁴⁹ Gogineni *et al.*⁵⁰ constructed a multiclass classifier to detect vulnerabilities in smart contracts by fine-tuning an AWD-LSTM model.⁵¹ The model was fine-tuned using a dataset of 40,877 unique opcode combinations from smart contracts. The smart contracts were classified into four categories: suicidal, prodigal, greedy, and normal. To address the class imbalance, only distinct opcode combinations were retained for normal smart contracts, as they comprised the majority of the dataset and often contained repeated sequences.

The AWD-LSTM model architecture combined a pretrained encoder with a custom classification head, inspired by the ULMFiT method⁵² used in natural language processing. This method achieved an overall accuracy of 91.3% and a weighted average F1 score of

90.0% in classifying smart contract vulnerabilities. The model’s performance was evaluated using various metrics, including precision, recall, F1 score, confusion matrix, and receiver operating characteristic curves, demonstrating its effectiveness in detecting vulnerabilities compared to random guessing.

3. Results

In this section, the cost-effectiveness and performance of the proposed hierarchical FL framework were theoretically and experimentally analyzed. The analyses considered both communication costs and computation costs for all entities involved in the training process: Clients, local servers, and the central server. To demonstrate the utility of the proposed framework, the observed accuracy was compared with the benchmark FL algorithm, FedAvg,⁵³ with fine-tuning.

3.1. Experimental evaluation

The experiments were conducted using the structure of Flower Framework,⁵⁴ extended to support both a standard FedAvg setup and the proposed hierarchical FL framework. Python 3.10.12 (Phyton software foundation, US) was used in a Jupyter Notebook (v7.4.3, Jupyter, US) environment, along with TensorFlow, Pandas, NumPy, and scikit-learn libraries. The TenSEAL library⁵⁵ was utilized to integrate HE computation, supporting encrypted tensor operations using the Cheon-Kim-Kim-Song⁵⁶ scheme from Microsoft’s simple encrypted arithmetic library (SEAL). This simulation setup emulates the interactions between clients and the server in a real-world scenario while maintaining a controlled environment for reproducibility.

The modified National Institute of Standards and Technology (MNIST) image dataset⁵⁷ was employed for experiments, as image data is representative of numerous

real-world medical datasets and can therefore be used as a reliable data proxy. MNIST consists of 70,000 grayscale images, with 60,000 images used for training and 10,000 for testing. Each image is 28×28 pixels and encoded with intensity values ranging from 0 to 255.

For model training, a CNN architecture was designed with an input layer, two hidden layers activated by *ReLU* functions, and an output layer generating a probability distribution over 10 classes using a softmax function. The model was trained for 30 communication rounds. In the baseline FedAvg setup, the simulation involved 12 and 21 clients and a single central server. For the hierarchical FL implementation, two configurations were tested: first, with four local servers, each connected to three client nodes (totaling 12 nodes, matching the first non-hierarchical case), and one central server coordinating the aggregation; second, with three local servers and seven clients per server (totaling 21 nodes, corresponding to the second non-hierarchical case), and a central server. The learning rate was set to 0.01, and each client performed three local training epochs per round. Throughout the training process, both training time and accuracy at each round were monitored to compare model performance. Figure 4a and 4b illustrate the accuracy and training time per round comparison, respectively.

Model accuracy was observed to increase gradually with each communication round for both algorithms. However, the hierarchical FL framework consistently demonstrated higher accuracy due to its intermediate cluster-level aggregation, which is less biased to outlier client data. The proposed framework achieved an accuracy of 94.23%, whereas the FedAvg approach reached 92.66% under the same experimental settings. Training time per round showed minor fluctuations for both algorithms, depending on the data distribution within each training

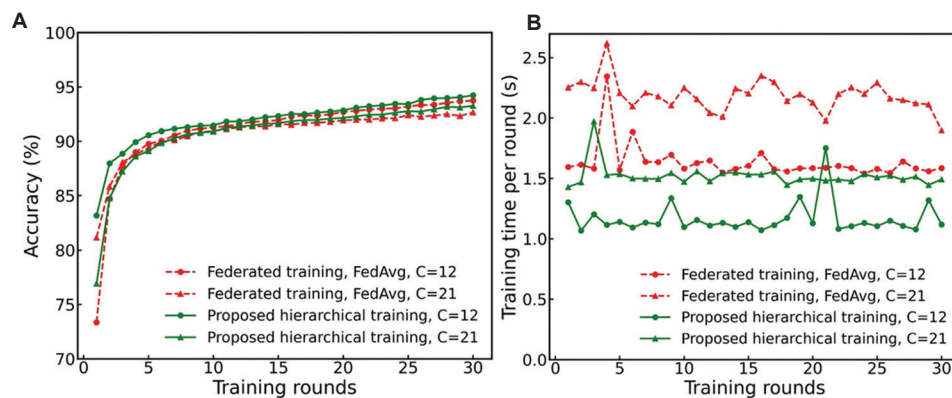


Figure 4. Comparisons between FedAvg and the proposed hierarchical FL framework. (A) Accuracy comparison. Note that the hierarchical FL demonstrates enhanced security compared to FedAvg. (B) Training time per round comparison. Note that the hierarchical FL consumes lesser compute time, demonstrating a higher efficiency than FedAvg.

batch. Nevertheless, the proposed framework exhibited lower training time across all scenarios. This improvement is attributed to the hierarchical structure: whereas FedAvg requires all clients to transmit model updates to the central server concurrently, the hierarchical approach organizes clients into clusters. Each client communicates only with their respective local servers, and the aggregated updates are subsequently forwarded to the central server, thereby reducing communication overhead. Given its time efficiency, the hierarchical model is better suited for global deployment and offers greater scalability compared to FedAvg, particularly in addressing future pandemics. As the hierarchical model outperforms standard FL protocols with complex data types, such as images, it is anticipated to perform effectively with other data types as well, including categorical, numerical, and time-series data, such as cancer classifications, biophysical parameters, and electrocardiography, respectively.

3.2. Communication cost analysis

In the proposed framework, communication occurs at two levels: from client to local server and from local server to central server. All communication is bidirectional, meaning both entities exchange messages during each round. The communication cost of the aforementioned level was analyzed.

3.2.1. Client-to-local server communication

Each client C_{ij} communicated with its assigned local server L_i by transmitting encrypted model gradients $[G_r^{ij}]$ and its associated public key pk_r^{ij} during each round r . The communication cost per client is represented using big-O notation:

$$Cost_{client-to-local} = O(|G_r^{ij}| + |pk_r^{ij}|) \tag{XV}$$

where $|G_r^{ij}|$ is the size of the encrypted gradient and $|pk_r^{ij}|$ is the size of the encryption key for p clients under a single local server. The parameter p was included in the calculation, as it was assumed to represent the maximum number of clients under a local server. The total communication cost is represented as:

$$Cost_{total-client-to-local} = O(p \cdot (|G_r^{ij}| + |pk_r^{ij}|)) \tag{XVI}$$

3.2.2. Local server-to-central server communication

Each local server L_i aggregated encrypted gradients from its clients and transmitted the aggregated model updates $[G_r^i]$ to the central server. The communication cost per local server is represented as:

$$Cost_{local-to-central} = O(|G_r^i| + |pk_r^i|) \tag{XVII}$$

where $|G_r^i|$ is the size of the aggregated gradient and $|pk_r^i|$ is the size of the encryption key for the local server.

For n local servers, the total communication cost is represented as:

$$Cost_{total-local-to-central} = O(n \cdot (|G_r^i| + |pk_r^i|)) \tag{XVIII}$$

3.2.3. Total communication cost

Combining the costs for all client-to-local server and local-to-central server communications, the overall communication cost for the framework is represented as:

$$Cost_{total} = (p \cdot |G_r^{ij}| + n \cdot |G_r^i| + p \cdot |pk_r^{ij}| + n \cdot |pk_r^i|) \tag{XIX}$$

The parameter size at local server L_i , denoted as $|G_r^i|$, includes the presence of p clients, i.e., $|G_r^i| = p \cdot |G_r^{ij}|$, where $|G_r^{ij}|$ represents the parameter size from an individual client j . Therefore, to reflect the total cost across n local servers, $n \cdot |G_r^i|$ is used instead of $n \cdot p \cdot |G_r^{ij}|$.

3.3. Computation cost analysis

The computation cost of the proposed framework was analyzed by evaluating the operations performed at each level: clients, local servers, and the central server. This section provides a detailed computation cost analysis.

3.3.1. Client-side computation

Each client C_{ij} performed local model training on its secured data. The training cost was proportional to the local dataset D_{ij} and the complexity of the model M . Then, it encrypted the model gradients G_r^{ij} using the encryption key pk_r^{ij} . Similarly, it also decrypted the aggregated gradients received from the central server at the end of each round. The computation cost per client is represented as:

$$Cost_{client} = O(|D_{ij}| \cdot M + Enc(|G_r^{ij}|) + Dec(|G_r^{ij}|)) \tag{XX}$$

where $Enc(|G_r^{ij}|)$ is the encryption cost and $Dec(|G_r^{ij}|)$ is the decryption cost for the gradients. Suppose p clients were present under a local server; the total client-side computation cost is represented as:

$$Cost_{total-client} = O(p \cdot (|D_{ij}| \cdot M + Enc(|G_r^{ij}|) + Dec(|G_r^{ij}|))) \tag{XXI}$$

3.3.2. Local server-side computation

Each local server L_i performed the aggregation of encrypted gradients received from all clients in the cluster using homomorphic addition and multiplication, ensuring that the aggregation was performed without decrypting the gradients. If the number of clients assigned to L_i was p , then p additions and a single multiplication were required for aggregation. For simplicity, the time taken for multiplication was assumed as unity. The computation cost per local server is represented as:

$$Cost_{local-server} = O(p \cdot T_{add}) \tag{XXII}$$

where T_{add} represents the time taken for homomorphic addition. For n local servers in the framework, the total local server-side computation cost is represented as:

$$Cost_{total-local-server} = O(n \cdot p \cdot T_{add}) \tag{XXIII}$$

3.3.3. Central server-side computation

Similar to local servers, the central server performed aggregation of gradients received from n local servers. The computation cost at the central server is represented as:

$$Cost_{total-local-server} = O(n \cdot T_{add})$$

3.3.4. Total computation cost

Combining the computation costs at the client, local server, and central server levels, the total computation cost for the proposed framework is represented as:

$$Cost_{total} = O(p \cdot (|D_{ij}| \cdot M + Enc(|G_r^{ij}|) + Dec(|G_r^{ij}|)) + n \cdot p \cdot T_{add} + n \cdot T_{add}) \tag{XXIV}$$

It can be further simplified to:

$$Cost_{total} = O(p \cdot (|D_{ij}| \cdot M + Enc(|G_r^{ij}|) + Dec(|G_r^{ij}|) + n \cdot T_{add})) \tag{XXV}$$

The hierarchical structure optimized both communication and computation by leveraging local servers to consolidate updates before transmitting them to the central server. This ensured scalability, even in scenarios with large datasets and numerous participants.

4. Discussion

In this section, the issue of data heterogeneity in FL is addressed, specifically in the context of datasets distributed across different medical centers or countries. FL encounters significant challenges in real-world medical settings owing to the intrinsic heterogeneity among contributing institutions. Data heterogeneity arises when data distribution varies substantially across clients, leading to non-IID data.⁵⁸ Heterogeneity may manifest as statistical variations, differences in system capabilities, disparities in model architecture, and additional challenges.^{59,60} While the proposed model can directly handle IID datasets, its robustness is demonstrated by showing how it can manage non-IID datasets. Several techniques are proposed to mitigate data imbalance effects and enhance model performance in hierarchical systems designed for managing medical data management.

According to recent studies, and within the context of hierarchical medical data management, generative adversarial networks (GANs), particularly the newly developed robust diffusion models,⁶¹ can effectively achieve uniformity in data availability across medical facilities. Zadeh *et al.*⁶² utilized GANs for cross-modality brain image synthesis, including transformations such as CT to positron emission tomography (PET), CT to magnetic resonance imaging (MRI), MRI to PET, and vice versa.

To address statistical heterogeneity or non-uniform data distribution, which significantly impacts model accuracy, increased communication rounds are often necessary. However, this can introduce bias in the global model, particularly disadvantaging clients with underrepresented data from various institutions. Therefore, aligning the distributions of data across medical centers is critical to mitigating model bias caused by variations in the population of patients or data collection techniques.^{58,63} Class balancing⁶⁴ should be supported with equal representation of all disease classes or conditions across federated nodes to prevent biased learning outcomes. Additionally, standardization of quality⁶⁵ is necessary to normalize data collected via varying equipment and protocols to enhance uniformity and reliability. Moreover, volume balancing⁶⁶ helps prevent dominant contributions from larger hospitals, ensuring equitable learning from all centers.

To fulfill these requirements, GANs, especially robust diffusion models, offer a promising method for establishing data uniformity across hospitals with varied dataset sizes. By generating synthetic images to supplement existing datasets, GANs enable more balanced training with minimal bias. For example, if three medical centers have 500, 400, and 250 data points, respectively, GANs can generate synthetic images to equalize each dataset to approximately 500 data points. Compared to traditional weighted averaging of model parameters, this approach provides a more balanced solution for hierarchical medical system performance. The working principle is based on iterative noise addition and removal, where the generator network analyzes the denoising function to reconstruct the original image.

Despite the benefits of data augmentation via GANs, FL in medical imaging still encounters challenges due to the inherent diversity of imaging data. Scans from different sites vary in scanner type, protocol, and patient demographic, making synthetic data approaches more complex.⁶⁷ Recent FL frameworks, such as distributed synthetic learning, aim to train GANs to produce a single homogeneous dataset of synthetic images for use by all clients,⁶⁷ yet practical concerns remain. Specifically, the application of differential privacy can hamper performance. For example, Kossen *et al.*⁶⁸ reported that enforcing a privacy parameter $\epsilon \approx 7.4$ on GAN-produced angiograms lowered a U-net vessel segmentation's dice score from 0.84 to 0.75.

In addition, GAN-augmented FL models are susceptible to membership inference attacks (MIAs). MIAs allow attackers to deduce whether a particular data point belongs to the training set. For example, Zhang *et al.*⁶⁹ demonstrated class-level and user-level MIAs with GANs, achieving over

90% attack accuracy in FL environments. To mitigate these attacks, a membership inference defense mechanism named DefMIA was introduced, which adds adversarial perturbations to global model parameters, reducing attack accuracy to approximately 50% without affecting model performance. Overall, GAN-based FL remains susceptible to accuracy degradation and privacy concerns when applied to non-IID medical data, necessitating critical evaluation and ongoing refinement.

Model heterogeneity refers to differences across model architectures, capacities, and training approaches across medical centers. Before FL implementation, architectural compatibility must be ensured to facilitate effective knowledge transfer across diverse model structures without compromising performance. Scale adaptability should also be enabled in the platform to accommodate the varying computational capabilities of institutions by enabling flexible model sizes. In addition, knowledge alignment mechanisms are needed to coordinate learning across heterogeneous models. Personalization support should be enabled to address institution-specific requirements or patient groups without compromising the global model's integrity or performance. Therefore, to address model variation, a hierarchical self-distillation approach, such as HierarchyFL,⁷⁰ is recommended. This method facilitates architecture compatibility by enabling resource-limited centers to efficiently learn from advanced models deployed at major sites. However, conventional aggregation techniques, such as FedAvg, cannot be directly applied in this setting. In addition, model architecture inconsistencies reduce knowledge transfer, as some clients may be unable to contribute weights, leading to performance degradation and inconsistency.

System heterogeneity refers to differences in computational resources, storage capacity, and communication bandwidth. Strategies to manage these differences are essential. Resource-conscious participation should be prioritized, with computational overhead adapted to each institution's hardware and network capacity. Efficiency in communication is also required for optimal data transfer, particularly within low or unstable bandwidth settings. Furthermore, storage efficiency should accommodate institutions with limited infrastructure to ensure broad participation. Power efficiency is also important, especially for edge devices, to reduce power consumption with consistent performance within clinical settings. A framework incorporating client selection through reinforcement learning and resource-aware metrics, such as that proposed in personalized FL,⁷¹ is recommended. This framework performs federation tasks dynamically according to each center's available resources,

supporting equitable participation. Nonetheless, system heterogeneity continues to extend training time due to hardware capabilities and network disparities, increase client dropout rates, and introduce latency, all of which hinder overall FL performance. These challenges warrant further research.

5. Conclusion

In this article, a hierarchical FL framework is proposed, designed to enhance scalability, security, and computational efficiency through multilevel aggregation and HE, enabling effective responses to future pandemics. FL is a decentralized ML paradigm in which each client trains a model on local data and shares only model updates with a central server, thus preserving data privacy. The hierarchical structure, consisting of client nodes, local servers, and a central server, supports distributed learning and efficient communication. Local servers perform intermediate aggregation, reducing communication costs and enhancing scalability. The proposed framework supports dynamic data and can be extended to train specialized models at the central unit based on data characteristics. It ensures secure model updates through parameter encryption and seamless integration with ByITFL, which incorporates Byzantine fault tolerance and information-theoretic privacy protection, strengthening the integrity and confidentiality of updates during aggregation.

To address the demands of medical image-based diagnostics, the framework is adaptable to advanced CNN architectures, such as ResNet34, DenseNet121, and EfficientNet, which have demonstrated high accuracy in tasks such as COVID-19 detection. Similarly, in lung nodule detection, the incorporation of modular clustering and transfer learning enhances imaging operations by efficiently identifying ROI, reducing resource usage while improving diagnostic accuracy. On the other hand, blockchain-based smart contracts facilitate trusted coordination among participants, while GANs contribute to standardized data processing. Experiments conducted using the Flower framework and TenSEAL on the MNIST dataset demonstrated improved model accuracy, reduced training time per round, and lower communication overhead due to the scope of optimized intermediate aggregation.

Overall, the proposed framework offers a scalable, privacy-preserving, and computation-efficient solution for digital healthcare. These advancements collectively position the proposed framework as a robust approach for addressing future healthcare challenges and responding effectively to potential pandemics.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Kisor Kumar Sahu

Formal analysis: All authors

Investigation: Rojalini Tripathy

Methodology: All authors

Visualization: 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

Data are available at the following resource: Deng, L. The MNIST database of handwritten digit images for machine learning research. *IEEE signal processing magazine*;2012. 29(6), 141-142. doi: 10.1109/MSP.2012.2211477 and code can be shared upon reasonable request over email: kisorsahu@iitbbs.ac.in.

References

- World Health Organization. *COVID-19 Deaths. WHO COVID-19 Dashboard*. World Health Organization Data; 2023. Available from: <https://data.who.int/dashboards/covid19/deathss> [Last accessed on 2024 Dec 14].
- Roser M. *The Spanish Flu: The Global Impact of the Largest Influenza Pandemic in History*. Our World in Data; 2020. Available from: <https://ourworldindata.org/spanish/flu-largest-influenza-pandemic-in-history> [Last accessed on 2024 Dec 14].
- Mordechai L, Eisenberg M, Newfield TP, Izdebski A, Kay JE, Poinar H. The justinianic plague: An inconsequential pandemic. *Proc Natl Acad Sci U S A*. 2019;116:25546-25554. doi: 10.1073/pnas.1903797116
- Wade L. *From Black Death to Fatal Flu, Past Pandemics Show why People on the Margins Suffer Most*. Science Magazine; 2020. Available from: <https://www.science.org/content/article/black-death-fatal-flu-past-pandemics-show-why-people-margins-suffer-most> [Last accessed on 2024 Dec 14]. doi: 10.1126/science.abc7832
- 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. In: *Computational Modeling and Data Analysis in COVID-19 Research*. United States: CRC Press; 2021. p. 67-94. doi: 10.1201/9781003137481-5
- McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: *Artificial Intelligence and Statistics*. Hamburg: Statista; 2017. p. 1273-1282. doi: 10.48550/arXiv.1602.05629
- Yurdem B, Kuzlu M, Gullu MK, Catak FO, Tabassum M. Federated learning: Overview, strategies, applications, tools and future directions. *Heliyon*. 2024;10:e38137. doi: 10.1016/j.heliyon.2024.e38137
- Nguyen DC, Pham QV, Pathirana PN, *et al*. Federated learning for smart healthcare: A survey. *ACM Comput Surv*. 2022;55(3):1-37. doi: 10.1145/3501296
- Rieke N, Hancox J, Li W, *et al*. The future of digital health with federated learning. *NPJ Digit Med*. 2020;3(1):119. doi: 10.1038/s41746-020-00323-1
- Sheller MJ, Edwards B, Reina GA, *et al*. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci Rep*. 2020;10(1):12598. doi: 10.1038/s41598-020-69250-1
- Sun T, Li D, Wang B. Decentralized federated averaging. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(4):4289-4301. doi: 10.48550/arXiv.2104.11375
- Korzun DG. Internet of things meets mobile health systems in smart spaces: An overview. In: *Internet of Things and Big Data Technologies for Next Generation Healthcare*. Berlin: Springer Nature; 2017. p. 111-129. doi: 10.1007/978-3-319-49736-5_6
- Velichko A, Korzun D, Meigal A. Artificial neural networks for iot-enabled smart applications: Recent trends. *Sensors*. 2023;23(10):4853. doi: 10.3390/s23104853
- Xia Y, Hofmeister C, Egger M, Bitar R. *Byzantine-Resilient Secure Aggregation for Federated Learning without Privacy Compromises*. United States: IEEE; 2024. doi: 10.48550/arXiv.2405.08698
- Stephanie V, Khalil I, Atiquzzaman M, Yi X. Trustworthy privacy-preserving hierarchical ensemble and federated

- learning in healthcare 4.0 with blockchain. *IEEE Trans Ind Inform.* 2022;19(7):7936-7945.
doi: 10.1109/TII.2022.3214998
16. Li G, Hu Y, Zhang M, *et al.* FedHiSyn: A hierarchical synchronous federated learning framework for resource and data heterogeneity. In: *Proceedings of the 51st International Conference on Parallel Processing*. United States: Association for Computing Machinery; 2022. p. 1-11.
doi: 10.1145/3545008.3545065
 17. Wang B, Li H, Guo Y, Wang J. PPFLHE: A privacy-preserving federated learning scheme with homomorphic encryption for healthcare data. *Appl Soft Comput.* 2023;146:110677.
doi: 10.1016/j.asoc.2023.110677
 18. Ooi MPL, Sohail S, Huang VG, *et al.* Measurement and applications: Exploring the challenges and opportunities of hierarchical federated learning in sensor applications. *IEEE Instrum Meas Mag.* 2023;26(9):21-31.
doi: 10.1109/MIM.2023.10328671
 19. Tripathy R, Meshram J, Bera P. HalfFedLearn: A secure federated learning with local data partitioning and homomorphic encryption. *Fut Gener Comput Syst.* 2025;171:107858.
doi: 10.1016/j.future.2025.107858
 20. Cao X, Fang M, Liu J, Gong NZ. *Fltrust: Byzantine-Robust Federated Learning Via Trust Bootstrapping*. United States: Cornell University; 2020.
doi: 10.48550/arXiv.2012.13995
 21. Fung C, Yoon CJ, Beschastnikh I. The limitations of federated learning in sybil settings. In: *23rd International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*. 2020. p. 301-316.
 22. Adilova L, Rosenzweig J, Kamp M. *Information-Theoretic Perspective of Federated Learning*. [Preprint]; 2019.
doi: 10.48550/arXiv.1911.07652
 23. Jin W, Yao Y, Han S, *et al.* *FedML-HE: An Efficient Homomorphic-Encryption-Based Privacy-Preserving Federated Learning System*. United States: Cornell University; 2023.
doi: 10.48550/arXiv.2303.10837
 24. Mozaffari H, Choudhary S, Houmansadr A. Fake or compromised? making sense of malicious clients in federated learning. In: *European Symposium on Research in Computer Security*. 2024. p. 187-207.
doi: 10.48550/arXiv.2403.06319
 25. Ziller A, Mueller TT, Stieger S, *et al.* Reconciling privacy and accuracy in AI for medical imaging. *Nat Mach Intell.* 2024;6(7):764-774.
doi: 10.1038/s42256-024-00858-y
 26. Zhou P, Fang P, Hui P. *Loss Tolerant Federated Learning*. [Preprint]; 2021.
doi: 10.48550/arXiv.2105.03591
 27. Yu Q, Li S, Raviv N, Kalan SMM, Soltanolkotabi M, Avestimehr SA. Lagrange coded computing: Optimal design for resiliency, security, and privacy. In: *22nd International Conference on Artificial Intelligence and Statistics*. [Preprint]; 2019. p. 1215-1225.
doi: 10.48550/arXiv.1806.00939
 28. So J, Güler B, Avestimehr AS. Byzantine-resilient secure federated learning. *IEEE J Select Areas Commun.* 2020;39(7):2168-2181.
doi: 10.48550/arXiv.2007.11115
 29. 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 a network-centric digital health stack for future pandemics. *Artif Intell Health.* 2024;2:29-41.
doi: 10.36922/aih.2888
 30. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. United States: IEEE; 2016. p. 770-778.
doi: 10.1109/cvpr.2016.90
 31. Hu J, Shen L, Albanie S, Sun G, Wu E. *Squeeze-and-Excitation Networks*. United States: Cornell University; 2017.
doi: 10.48550/arxiv.1709.01507
 32. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA: IEEE; 2017. p. 5987-5995.
doi: 10.1109/CVPR.2017.634
 33. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*. United States: IEEE; 2017. p. 4700-4708.
doi: 10.1109/CVPR.2017.243
 34. Tan M, Le QV. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*; 2019. Available from: <https://arxiv.org/pdf/1905.11946> [Last accessed on 2024 Dec 18].
doi: 10.48550/arXiv.1905.11946
 35. Simon M, Rodner E, Denzler J. *ImageNet Pre-trained Models with Batch Normalization*; 2016. Available from: <https://arxiv.org/pdf/1612.01452> [Last accessed on 2024 Dec 18].
doi: 10.48550/arXiv.1612.01452
 36. Smith LN. *Cyclical Learning Rates for Training Neural Networks*. Piscataway: IEEE Xplore; 2017. p. 464-472.
doi: 10.1109/WACV.2017.58

37. 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
38. 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
39. 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 on 2024 Dec 18].
40. Gogineni AK, Kishore R, Raj P, Naik S, Sahu KK. Unsupervised Clustering algorithm as region of interest proposals for cancer detection using CNN. In: *Computational Vision and Bio-Inspired Computing.* Coimbatore: ICCVBIC; 2019. p. 1386-1396.
doi: 10.1007/978-3-030-37218-7_146
41. Kishore R, Gogineni AK, Nussinov Z, Sahu KK. A nature inspired modularity function for unsupervised learning involving spatially embedded networks. *Sci Rep.* 2019;9(1):2631.
doi: 10.1038/s41598-019-39180-8
42. Kishore R, Krishnan R, Satpathy M, Nussinov Z, Sahu KK. Abstraction of meso-scale network architecture in granular ensembles using 'big data analytics' tools. *J Phys Commun.* 2018;2(3):031004.
doi: 10.1088/2399-6528/aab386
43. Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW. Selective search for object recognition. *Int J Comput Vision.* 2013;104:154-171.
doi: 10.1007/s11263-013-0620-5
44. Van Ginneken B, Jacobs C. *LUNA16.* Zenodo; 2019. Available from: <https://luna16.grand-challenge.org/> [Last accessed on 2024 May 20].
doi: 10.5281/zenodo.2595813
45. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *IEEE International Conference on Computer Vision (ICCV).* United States: IEEE; 2017. p. 2980-2988.
doi: 10.48550/arXiv.1703.06870
46. Rother C, Kolmogorov V, Blake A. GrabCut interactive foreground extraction using iterated graph cuts. *ACM Trans Graph (TOG).* 2004;23(3):309-314.
doi: 10.1145/3596711.3596774
47. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: *IEEE Conference on Computer Vision and Pattern Recognition.* United States: IEEE; 2016. p. 2921-2929.
doi: 10.1109/CVPR.2016.319
48. Scharfman J. Decentralized autonomous organization (DAO) fraud, hacks, and controversies. In: *The Cryptocurrency and Digital Asset Fraud Casebook DeFi, NFTs, DAOs, Meme Coins, and Other Digital Asset Hacks.* Vol. 2. Berlin: Springer; 2024. p. 65-106.
doi: 10.1007/978-3-031-60836-0_3
49. Destefanis G, Marchesi M, Ortu M, Tonelli R, Bracciali A, Hierons R. Smart contracts vulnerabilities: A call for blockchain software engineering? In: *International Workshop on Blockchain Oriented Software Engineering (IWBOSE).* United States: IEEE; 2018. p. 19-25.
doi: 10.1109/IWBOSE.2018.8327567
50. Gogineni AK, Swayamjyoti S, Sahoo D, Sahu KK, Kishore R. Multi-Class classification of vulnerabilities in Smart Contracts using AWD-LSTM, with pre-trained encoder inspired from natural language processing. *IOP SciNotes.* 2020;1(3):035002.
doi: 10.1088/2633-1357/abcd29
51. Merity S, Keskar NS, Socher R. *Regularizing and Optimizing LSTM Language Models.* United States: Cornell University; 2017.
doi: 10.48550/arXiv.1708.02182
52. Howard J, Ruder S. *Universal Language Model Fine-Tuning for Text Classification.* United States: Harvard University; 2018.
doi: 10.48550/arXiv.1801.06146
53. Collins L, Hassani H, Mokhtari A, Shakkottai S. Fedavg with fine tuning: Local updates lead to representation learning. In: *Advances in Neural Information Processing Systems.* Vol. 35. United States: MIT Press; 2022. p. 10572-10586.
doi: 10.48550/arXiv.2205.13692
54. Beutel DJ, Topal T, Mathur A, *et al.* *Flower: A Friendly Federated Learning Research Framework.* United States: Cornell University; 2020.
doi: 10.48550/arXiv.2007.14390
55. Benaissa A, Retiat B, Cebere B, Belfedhal AE. *Tenseal: A Library for Encrypted Tensor Operations Using Homomorphic Encryption.* United States: Cornell University; 2021.
doi: 10.48550/arXiv.2104.03152
56. Sathishkumar P, Pugalarasan K, Ponnparamaguru C, Vasanthkumar M. Improving healthcare data security using cheon-kim-kim-song (ckks) homomorphic encryption. In: *2024 International Conference on Knowledge Engineering and Communication Systems.* Vol. 1. United States: IEEE; 2024. p. 1-6.

- doi: 10.1109/ICKECS61492.2024.10616691
57. Deng L. The mnist database of handwritten digit images for machine learning research. *IEEE Sign Process Mag.* 2012;29(6):141-142.
doi: 10.1109/MSP.2012.2211477
58. Jimenez GM, Solans D, Heikkila M, et al. *Non-IID Data in Federated Learning: A Survey with Taxonomy, Metrics, Methods, Frameworks and Future Directions*. United States: Cornell University; 2024.
doi: 10.48550/arXiv.2411.12377
59. Ye M, Fang X, Du B, Yuen PC, Tao D. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Comput Surv.* 2023;56(3):1-44.
doi: 10.48550/arXiv.2307.10616
60. Gao D, Yao X, Yang Q. *A Survey on Heterogeneous Federated Learning*. United States: Cornell University; 2022.
doi: 10.48550/arXiv.2210.04505
61. Müller-Franzes G, Niehues JM, Khader F, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Sci Rep.* 2023;13:12098.
doi: 10.1038/s41598-023-39278-0
62. Zadeh FS, Molani S, Orouskhani M, Rezaei M, Shafiei M, Abbasi H. *Generative Adversarial Networks for Brain Images Synthesis: A Review*. United States: Cornell University; 2023.
doi: 10.48550/arXiv.2305.15421
63. Legler T, Hegiste V, Anwar A, Ruskowski M. Addressing heterogeneity in federated learning: challenges and solutions for a shared production environment. *Procedia Comput Sci.* 2025;253:2831-2840.
doi: 10.48550/arXiv.2408.09556
64. Dai Y, Chen Z, Li J, Heinecke S, Sun L, Xu R. Tackling data heterogeneity in federated learning with class prototypes. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 6. Washington, DC: AAAI Press; 2023. p. 7314-7322.
doi: 10.48550/arXiv.2212.02758
65. Babar M, Qureshi B, Koubaa A. Investigating the impact of data heterogeneity on the performance of federated learning algorithm using medical imaging. *PLoS One.* 2024;19(5):e0302539.
doi: 10.1371/journal.pone.0302539
66. Zhao Y, Li M, Lai L, Suda N, Civin D, Chandra V. *Federated Learning with Non-Iid Data*. United States: IEEE; 2018.
doi: 10.48550/arXiv.1806.00582
67. Chang Q, Yan Z, Zhou M, et al. Mining multi-center heterogeneous medical data with distributed synthetic learning. *Nat Commun.* 2023;14(1):5510.
doi: 10.1038/s41467-023-40687-y
68. Kossen T, Hirzel MA, Madai VI, et al. Toward sharing brain images: Differentially private TOF-MRA images with segmentation labels using generative adversarial networks. *Front Artif Intell.* 2022;5:813842.
doi: 10.3389/frai.2022.813842
69. Jiale ZH, Chengcheng ZH, Xiaobing SU, et al. Membership inference attack and defense method in federated learning based on GAN[J]. *J Commun.* 2023;44(5):193-205.
doi: 10.11959/j.issn.1000-436x.2023094
70. Xia J, Zhang Y, Yue Z, Hu M, Wei X, Chen M. *HierarchyFL: Heterogeneous Federated Learning via Hierarchical Self-Distillation*. United States: Cornell University; 2022.
doi: 10.48550/arXiv.2212.02006
71. Yang H, Li J, Hao M, Zhang W, He H, Sangaiah AK. An efficient personalized federated learning approach in heterogeneous environments: A reinforcement learning perspective. *Sci Rep.* 2024;14:28877.
doi: 10.1038/s41598-024-80048-3

ORIGINAL RESEARCH ARTICLE

Artificial intelligence versus humans: A comparative analysis of time, cost, and performance on a clinical code conversion task

Carly Hudson^{1,2,3*}, Marcus Randall², Candice Bowman^{1,4}, Anu Joy^{4,5}, and Adrian Goldsworthy^{1,6,7}¹Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Queensland, Australia²Bond Business School, Bond University, Gold Coast, Queensland, Australia³Faculty of Medicine and Health, University of New England, Armidale, New South Wales, Australia⁴Mental Health and Specialist Services, Gold Coast Hospital and Health Service, Gold Coast, Queensland, Australia⁵School of Applied Psychology, Griffith University, Brisbane, Queensland, Australia⁶Wesley Research Institute, Brisbane, Queensland, Australia⁷Critical Care Research Group, The Prince Charles Hospital, Brisbane, Queensland, Australia**Abstract**

Healthcare services generate and store large quantities of data, requiring significant resources to manually analyze and gain meaningful insights. Recent advancements in automation tools—such as generative artificial intelligence (GenAI)—provide new opportunities to reduce human labor. This study explores the potential utilization of GenAI for a healthcare data analysis task—specifically, the conversion of clinical data from one diagnostic classification system to another (i.e., the Australian extension of the Systematized Nomenclature of Medicine Clinical Terms to the International Classification of Diseases, 10th Revision, Clinical Modification)—and examines the time and cost benefits of performing this using GenAI compared to a human rater. Conversions were completed using three methods: manual conversion using the National Library of Medicine’s I-MAGIC tool, ChatGPT-4o, and Claude 3.5 Sonnet. The accuracy of the GenAI tools was mapped against the manually extracted codes and examined in terms of a perfect, partial, or incorrect match. Task completion time was recorded and extrapolated to calculate and compare the cost associated with each method. When compared to the manually extracted codes, Claude 3.5 Sonnet yielded the highest level of agreement over ChatGPT-4o, whilst being the most time- and cost-effective. GenAI tools have greater utility than they have currently been given credit for. The automation of big data healthcare analytics, whilst still the domain of humans, is increasingly capable of being undertaken using automation tools with low barriers to entry. The further development of GenAI’s capabilities, alongside the capability of the healthcare system to use it appropriately, has the potential to result in significant resource savings.

Keywords: Data analytics; Diagnostic coding; Generative artificial intelligence; International Classification of Diseases 10th revision; Systematized Nomenclature of Medicine Clinical Terms; SNOMED

***Corresponding author:**
Carly Hudson
(chudson@bond.edu.au)

Citation: Hudson C, Randall M, Bowman C, Joy A, Goldsworthy A. Artificial intelligence versus humans: A comparative analysis of time, cost, and performance on a clinical code conversion task. *Artif Intell Health*. 2025;2(4):92-102. doi: 10.36922/AIH025200045

Received: May 12, 2025

Revised: June 9, 2025

Accepted: June 18, 2025

Published online: July 11, 2025

Copyright: © 2025 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 volume of data generated annually by hospitals and health services far exceeds the analytical capacity of humans.¹ Murphy¹ estimated hospitals produce approximately 50 petabytes (equivalent to 50,000,000 gigabytes) of data each year—97% of which remains unanalyzed or unused. Electronic health records contain a wide range of information, including patient demographics, images, clinical notes, and pathology results. These records offer significant potential for retrospective analysis to support data-driven decision-making and more accurate predictions of service utilization.^{1,2} However, increasingly financially constrained and resource-limited healthcare systems lack the capacity to manually process such large datasets, limiting opportunities to improve healthcare system efficiency.^{1,3}

Generative artificial intelligence (GenAI) refers to a type of artificial intelligence algorithm that enables the creation of new content—such as text, images, video, or audio files—based on a set of training data.^{4,5} GenAI has a wide range of applications, including creating illustrations, writing code, and processing datasets.^{4,7} Additionally, GenAI has the potential to support the analysis of large-scale datasets within healthcare settings.^{5,8} Healthcare has traditionally required significant human labor and expertise, and as such, it has often resisted large-scale efforts for effective automation, particularly in the form of clinical and administrative decision-making.⁹⁻¹² A recent literature review by Li *et al.*¹³ has identified some of the key areas in which GenAI is starting to make an impact within healthcare, including generating discharge summaries,¹⁴ determining appropriate screening procedures for a patient,¹⁵ answering clinical questions, and providing medical education.¹⁶⁻¹⁹

The increasing complexity of global healthcare challenges necessitates new data analysis approaches that can expeditiously and efficiently leverage the vast datasets available to healthcare systems. Recent advancements in automation tools, such as GenAI, provide new opportunities to efficiently complete large-scale healthcare data analytics.²⁰ The widespread implementation of GenAI represents one of the most rapid technological advancements in recent years. OpenAI's²¹ ChatGPT is currently one of the most widely used GenAI tools, with over 100 million online users per week.²² ChatGPT allows users to input prompts, commands, or questions and generates corresponding responses. Its interface is driven by a large language model, a form of natural language processing capable of learning and refining its conversational abilities through both self- and semi-structured training.²³ Data processing is carried out using

large-scale neural networks, incorporating feed-forward and convolutional architectures.²³

Following the widespread success of ChatGPT, competitors have since launched other GenAI tools available to the general public, including Google Gemini,²⁴ Microsoft Copilot,²⁵ and Claude.²⁶ The accuracy and completeness of outputs are limited by the data available to the GenAI model (i.e., what it has been trained on, its access to real-time search capabilities), which may be biased or inaccurate. GenAI tools also have limited knowledge of more specialized topics, resulting in a tendency to “hallucinate”—a phenomenon where a GenAI tool generates information to fill knowledge gaps, thereby decreasing the accuracy of outputs.²⁷ Healthcare professionals require an up-to-date understanding of the current and evolving limitations of GenAI in order to optimally select tasks at which it is likely to excel and to prompt it appropriately.

A key challenge in analyzing large-scale healthcare data is ensuring the consistency of data recording across different health services. Standardized diagnostic coding systems help maintain clinical data uniformity by providing a universal language through which diagnoses can be coded and interpreted consistently across healthcare settings. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)²⁸ is a diagnostic coding system utilized by 48 countries (as of August 2024)²⁹ to capture detailed clinical information on procedures, diseases, and clinical findings. SNOMED CT presents diagnoses using both a numeric code (e.g., “230690007”) and a corresponding descriptor (e.g., “Stroke”). It employs a polyhierarchical structure, in which any given code may belong to one or more “parent” categories (e.g., “asthma” may be categorized under both “respiratory diseases” and “allergic conditions”). While SNOMED CT provides a comprehensive framework for patient-level diagnostic coding—encompassing symptoms, procedures, and clinical observations—the system's complexity can pose challenges for users with limited training.

The International Statistical Classification of Diseases and Related Health Problems (ICD)³⁰ is currently the global standard for coding diagnostic information. ICD focuses on the classification of diseases, disorders, and causes of death using alphanumeric codes. These codes are determined using a hierarchical system, in which codes are categorized by chapters (e.g., F: mental and behavioral disorders) and then further subdivided as more detail is provided (e.g., “F30: mood [affective] disorders,” “F30.9: manic episode, unspecified”). Although the ICD provides less detail than SNOMED CT, its broader categories facilitate population health analytics and provide a standard for international health system comparison.

The ability to convert between diagnostic coding systems has practical applications, particularly within research contexts. For instance, extracting a subset of SNOMED CT codes related to a specific diagnostic grouping (e.g., mental health) is challenging, as there are no broader categories for each condition, unlike ICD codes. This presents challenges when working with large SNOMED CT datasets while attempting to analyze only a subset. Converting diagnostic codes can be a time-consuming task, particularly when this process relies heavily on manual data input and extraction. To the authors' knowledge, it remains unexplored whether GenAI can assist in the conversion of clinical data from one diagnostic coding language to another, such as from SNOMED CT to ICD. Such conversions require specialized knowledge of clinical coding and are labor-intensive to complete manually. Performing diagnostic code conversion tasks using AI models may enable less qualified staff to complete the work in less time, thereby reducing the cost of data processing.

Therefore, this study aims to examine whether publicly accessible GenAI tools – namely ChatGPT-4o and Claude 3.5 Sonnet – can accurately convert clinical diagnostic codes from SNOMED CT to the 10th revision of the ICD (ICD-10). This study also seeks to address the following sub-objectives:

- (i) Compare the level of agreement between ChatGPT-4o and a human rater
- (ii) Compare the level of agreement between Claude 3.5 Sonnet and a human rater
- (iii) Compare the level of agreement between ChatGPT-4o and Claude 3.5 Sonnet
- (iv) Examine the economic benefit, in terms of time and labor cost, of using GenAI to complete this task compared to a human rater.

2. Materials and methods

The SNOMED CT codes used in this study originate from a broader emergency department (ED) dataset, obtained as part of a study investigating mental health presentations to hospital EDs (ethics approval: HREC/2023/QGC/95219). This dataset consists of 19,764 unique SNOMED-CTAU (Australian Extension) numeric codes (e.g., 48694002) and SNOMED-CT-AU names (e.g., “Anxiety reaction”) representing the diagnoses made to the ED over a 3-year period (August 2020 to August 2023). The current evaluation utilizes a randomly selected 10% subset of this data ($n = 1,976$) (Table S1).

To convert the SNOMED CT-AU³¹ codes to ICD-10 Clinical Modification (ICD-10-CM),³² a three-phase approach was employed. First, codes were manually converted by human raters. Second, the codes were

converted using ChatGPT-4o (<https://chatgpt.com/>). Third, the same set of codes was converted using Claude 3.5 Sonnet (<https://claude.ai/>). Both GenAI tools required paid subscriptions at the time of analysis.

The methodology and results of this study were reported in accordance with the METRICS reporting checklist, which outlines standardized reporting metrics – such as model, evaluation, timing, transparency, range of tested topics, randomization, individual factors, query count, and prompt specificity – for GenAI-based studies in healthcare.³³ The completed reporting checklist is listed in Table S2.

2.1. Phase 1: Manual conversion of SNOMED-CT-AU codes

The SNOMED CT-AU codes were manually converted by a team of three raters (AG = 800 codes; AJ = 644 codes; CH = 532 codes). Conversions were performed using the Interactive Map-Assisted Generation of ICD Codes (I-MAGIC) algorithm (<https://imagic.nlm.nih.gov/imagic/code/map>), an online tool that provides a mapping between the two diagnostic coding systems.³⁴ Codes were entered into the tool in the format “SNOMED CT-AU name (SNOMED CT-AU code)” (e.g., “Anxiety reaction [48694002]”), and the corresponding ICD-10-CM code was extracted.

In this study, the I-MAGIC tool was employed as the reference standard against which all other conversion methods were compared. However, some SNOMED CT codes could not be located within the I-MAGIC database. As the dataset utilized the Australian extension of SNOMED CT, while the mapping tool used the standard SNOMED CT list, it is likely that the missing codes were region-specific.³⁵ In such cases, the absence of an equivalent was noted.

2.2. Phase 2: Conversion of SNOMED-CT-AU codes using ChatGPT-4o

ChatGPT-4o²¹ was used to automatically convert the SNOMED CT-AU codes and names into ICD-10-CM codes (completed in August 2024). A Microsoft Excel file containing the SNOMED CT-AU codes and names was uploaded to ChatGPT-4o. The prompt used for the conversion was refined through an iterative process to improve efficiency and reduce the risk of “hallucinations” (i.e., providing false information) and data processing errors.

It was necessary to state that ChatGPT4o could take as much time as required to complete this task, otherwise the message would time out and cease to produce output. Additionally, a limit was observed regarding the number

of codes that could be processed simultaneously without increasing the likelihood of errors, such as incorrect conversions, fabricated inputs, or skipped entries. To manage this, codes were processed in batches of five, with five batches (25 codes in total) converted at a time.

The following final prompt was used:

Please manually and sequentially convert the SNOMED-CT-AU codes into ICD-10-CM codes and names which have not been converted yet. If you need to, complete this task in batches of 5. Take as long as you need. Do not hallucinate, and only convert codes which have been provided to you. Do not create new codes to convert. Provide an update after 5 batches have been completed. Provide a .xlsx file at the end.

Figure 1 presents the final prompt and the corresponding output from ChatGPT-4o.

Following each conversion of 25 codes, ChatGPT-4o needed to be prompted to continue (Figure 2):

Please continue.

Following conversion, ChatGPT-4o generated the requested Microsoft Excel file containing the original input data and the corresponding final output data.

Please manually and sequentially convert the SNOMED CT AU codes into ICD-10-CM codes and names which have not been converted yet. If you need to, complete this task in batches of 5. Take as long as you need. Do not hallucinate, and only convert codes which have been provided to you. Do not create new codes to convert. Provide an update after 5 batches have been completed. Provide a .xlsx file at the end.

Memory updated

I have identified the first batch of 5 SNOMED CT AU codes that require manual conversion to ICD-10-CM codes and names. I will now proceed with the conversion process.

Please hold on while I perform the necessary research and mapping. [...]

Batch 1 of 5 Completed:

1. SNOMED CT AU Code: 15692007
 - SNOMED CT AU Name: Right lateral brow position
 - ICD-10-CM Code: N/A (This term likely represents a specific observation or position rather than a medical diagnosis, and may not have a direct ICD-10-CM equivalent.)
2. SNOMED CT AU Code: 301003007
 - SNOMED CT AU Name: Right middle lobe pneumonia
 - ICD-10-CM Code: J18.1
 - ICD-10-CM Name: Lobar pneumonia, unspecified
3. SNOMED CT AU Code: 111589005
 - SNOMED CT AU Name: Dysfibrinogenaemia
 - ICD-10-CM Code: D68.2
 - ICD-10-CM Name: Hereditary deficiency of other clotting factors (Dysfibrinogenemia can be coded here as it is a clotting disorder.)

Figure 1. ChatGPT-4o prompt and output
 Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

2.3. Phase 3: Conversion of codes from the Australian extension of the Systematized Nomenclature of Medicine Clinical Terms using Claude 3.5 Sonnet

At the time of the study (September 2024), Claude 3.5 Sonnet²⁶ was unable to import or export Microsoft Excel files. Therefore, data were copied and pasted directly into the chat as a means of input. However, there were limitations on the maximum allowable message length. As a result, the data were entered in three batches of 500, followed by a final batch of 476.

The prompt was developed in a similar manner to Phase 2, using an iterative approach. Claude 3.5 Sonnet required fewer instructions to accurately perform the task, allowing the prompt to be developed relatively quickly. However, the limits of output message length restricted Claude 3.5 Sonnet to processing only 50 codes at a time.

During prompt testing, it was noted that Claude 3.5 Sonnet needed to be explicitly instructed not to skip lines, as reflected in the final prompt (Figure 3):

Please convert these SNOMED CT-AU codes and names into ICD-10-CM codes and names. Convert them in sequential order starting from the top and ensuring you do not skip any. I have provided 500 codes in total and need 500 responses at the end. Start with the first 50. If you are unable to convert a code, please state this.

The next five batches of SNOMED CT AU codes have been successfully converted to ICD-10-CM codes and names. The updated dataset has been saved.

You can download the final file using the link below:

[Download the final converted file](#)

Would you like to continue with more conversions, or is there anything else I can assist you with? [...]

Please continue

Memory updated

I will proceed with the next batch of five SNOMED CT AU codes and continue the conversion process. Let's start with the next batch.

Here are the next five SNOMED CT AU codes and their corresponding names:

1. 439127006 - Thrombosis
2. 239778009 - Wrist pyogenic arthritis
3. 197105006 - Perforated diverticulum of duodenum
4. 1244471000168100 - Comminuted fracture of clavicle
5. 247165009 - Retinal detachment - subretinal fluid

I will now proceed with manually converting these to ICD-10-CM codes and names. This will take a moment. [...]

Figure 2. ChatGPT-4o is prompted to continue with the next batch of conversions
 Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

As Claude 3.5 Sonnet was unable to directly export a Microsoft Excel file, it was instead instructed to produce R

code that could be run to generate an output file (Figure 4). The following request was made: “Can you write the code for R to create an Excel file of this data. Write it out in full so it exports all entries 1 – 100.” Due to limitations in output message length, this process was completed in batches of 100. The generated code was then run using R Studio (Posit, USA) to produce the final output file.

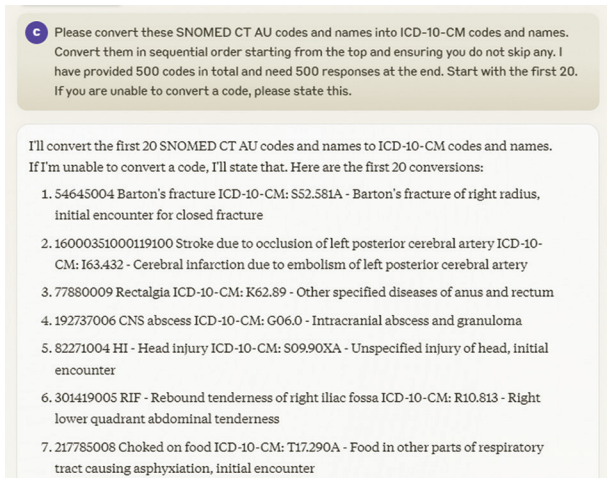


Figure 3. Claude 3.5 Sonnet prompt and output
Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

2.4. Statistical analysis

For the purpose of analysis, the ICD codes were split into three components: (i) letter code (“Level 1”); (ii) major numeric code (before the decimal point: “Level 2”); and (iii) minor numeric code (after the decimal point: “Level 3”) (Table 1).

A pattern-matching program was developed using the C programming language to identify partial and perfect matches among the 1,970 cases between: (i) manual coding and ChatGPT-4o; (ii) manual coding and Claude 3.5 Sonnet; and (iii) ChatGPT-4o and Claude 3.5 Sonnet. The program converted the codes from each method into their component parts. Some manual editing was necessary

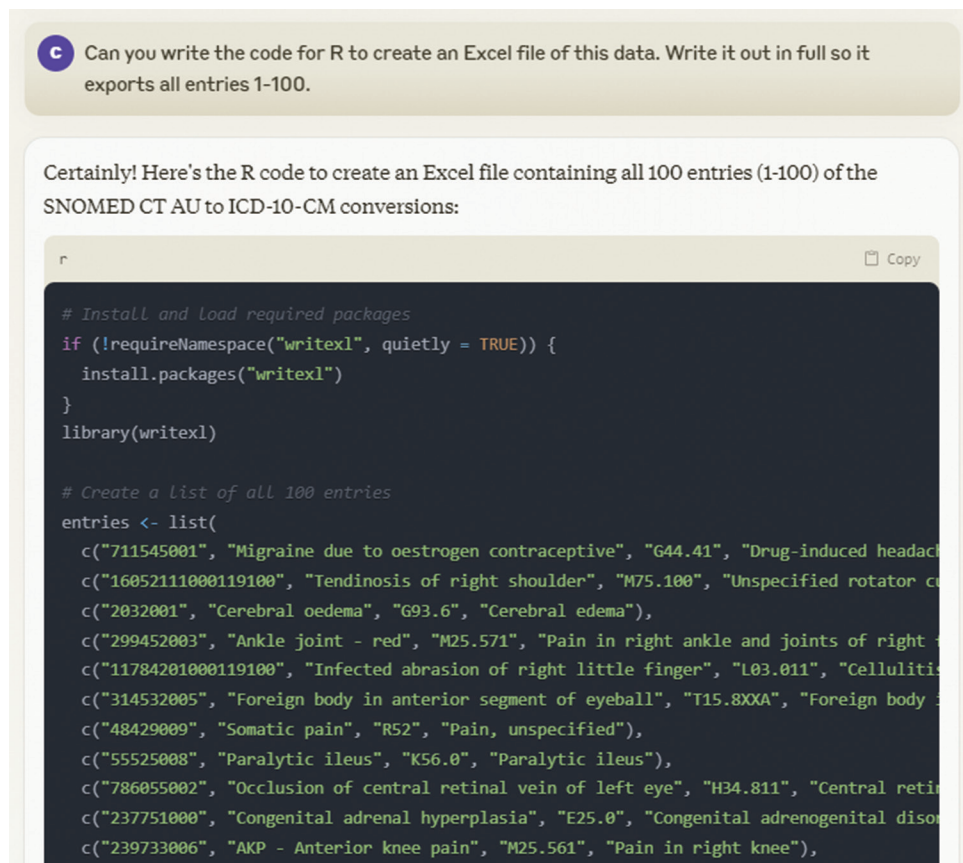


Figure 4. Claude 3.5 Sonnet generating R code to create a Microsoft Excel file
Abbreviations: ICD-10-CM: International Classification of Diseases, 10th Revision, Clinical Modification; SNOMED CT-AU: Australian extension of the Systematized Nomenclature of Medicine Clinical Terms.

to address missing values, which were substituted with null codes.

Using this program, the number of matches found for each comparison was categorized as follows: perfect match, Level 1 partial match, Level 2 partial match, and incorrect match (Table 1). A Chi-squared test of independence was conducted to determine whether there was a statistically significant difference in the number of good matches (perfect and Level 2 partial matches) and poor matches (Level 1 partial and incorrect matches) across the three methods.

2.5. Time and cost analysis

The time required to perform conversions in each phase was recorded to allow for a comparison of the time and

Table 1. Components of the International Classification of Diseases codes used to identify matches

	Level 1	Level 2	Level 3
Example ICD Code	F	30	.9
Perfect match	Yes	Yes	Yes
Level 2 partial match	Yes	Yes	No
Level 1 partial match	Yes	No	No
Incorrect match	No	No	No

Abbreviation: ICD: International Classification of Diseases.

Table 2. Number of correct matches across comparisons

Match category	Manual coding vs. ChatGPT-4o (%)	Manual coding vs. Claude 3.5 Sonnet (%)	ChatGPT-4o vs. Claude 3.5 Sonnet (%)
Perfect match	578 (29.34)	599 (30.41)	757 (38.43)
Level 2 partial match	751 (38.12)	758 (38.48)	763 (38.73)
Level 1 partial match	235 (11.93)	212 (10.76)	230 (11.68)
Incorrect match	406 (20.61)	401 (20.36)	220 (11.17)

Table 3. Time and cost for each method

Method and scenario	Time	FTEs in weeks ^a	Labor cost (AUD)	Cost of GenAI tool (AUD)	Total cost (AUD)
10% subset (n=1976)					
Manual coding	24 h and 31 min	0.64	\$1,279.77	N/A	\$1,279.7731
ChatGPT-4o	5 h and 45 min	0.15	\$300.15	\$30.00	\$330.15
Claude 3.5 Sonnet	3 h and 10 min	0.08	\$165.30	\$30.00	\$195.30
Extrapolation for full dataset (n=19,764)					
Manual coding	245 h and 12 min	6.45	\$12,799.44	N/A	\$12,799.44
ChatGPT-4o	57 h and 30 min	1.51	\$3,001.50	\$30.00	\$3,031.50
Claude 3.5 Sonnet	31 h and 40 min	0.83	\$1,653.00	\$30.00	\$1,683.00

Note: ^aAssumes a 38-h work week.

Abbreviations: FTE: Full time equivalent; N/A: Not available.

labor costs associated with each method. The cost of completing the task was calculated by multiplying the time taken for each method by the hourly wage of a research assistant, which was set at AUD\$52.20/hour (based on the pay rate for a university-employed research assistant, excluding on-costs). Setup costs – namely, the cost of subscribing to ChatGPT-4o or Claude 3.5 Sonnet – were also included in the total cost calculation.

3. Results

Table 2 displays the number of each type of match found for each of the comparisons. A Chi-squared test of independence was conducted to examine differences in the number of good and poor matches among manual coding, ChatGPT-4.0, and Claude 3.5 Sonnet. The analysis reveals a statistically significant difference in agreement across the three comparisons (χ^2 [df = 2] = 56.722, $p < 0.001$).

Agreement on good matches varies considerably between method pairs. The ChatGPT-4.0 and Claude 3.5 Sonnet pair show the highest agreement, producing good matches for 1,520 cases (77.2%) compared to 1,329 cases (67.5%) for manual coding versus ChatGPT-4o and 1,357 cases (68.9%) for manual coding versus Claude 3.5 Sonnet.

Table 3 displays the time and associated cost for a research assistant to perform data conversions using each tool for the 10% subset (n = 1,976) included in this study. It also includes an extrapolated estimate of costs if the entire dataset (n = 19,764) were to be converted from SNOMED to ICD.

Of the three methods used, manual coding was the most time-consuming and costly, taking 24 h and 31 min (AUD\$1,279.7) to convert the subset utilized in this study. When extrapolated to the full dataset, this method is estimated to require 245 h and 12 min, with a labor cost of AUD\$12,799.44.

Of the two GenAI methods, Claude 3.5 Sonnet was the most time- and cost-efficient, requiring 3 h and 10 min (AUD\$195.30, including subscription cost). ChatGPT-4o nearly doubled the time and cost of Claude 3.5 Sonnet, taking 5 h and 45 min (AUD\$195.30, including subscription cost). Regardless, ChatGPT-4o still demonstrated significant time and cost savings compared to manual conversion.

4. Discussion

This evaluation provides a case study to investigate the ability of GenAI tools to process and analyze large-scale healthcare datasets. To the authors' knowledge, this study is the first to challenge GenAI tools to complete a clinical diagnostic coding conversion task and to compare the results against those of a manual rater. Conversion of clinical diagnostic codes to other coding systems, such as the task presented in this study, is a complex and time-consuming task commonly undertaken within healthcare data processing. Therefore, this study highlights an example of a potential use for GenAI within health data analytics.

The analysis in this study examined matches found between the two GenAI tools and the manual rater. The results indicated that the two GenAI tools showed a higher level of agreement than either of them did compared to the manual coding, suggesting that the GenAI methods may employ similar coding strategies or have overlapping strengths in code conversion that differ from manual coding approaches.

However, when interpreting these findings, there are several caveats to consider. For instance, the clinical validity of ICD codes—particularly in cases where these were identified as “partial” or “incorrect matches”—was not assessed. This may have resulted in several potentially valid codes being incorrectly coded. For example, the SNOMED code “314041007 Abdominal pain in early pregnancy” was manually converted to “R10.9 Unspecified abdominal pain”. As this formed the benchmark for comparison between the GenAI tools, conversions made by ChatGPT-4o (“O26.83 Pregnancy related abdominal pain”) and Claude 3.5 Sonnet (“O26.892 Other specified pregnancy related conditions, first trimester”) were considered as incorrect matches.

During the analysis, the GenAI tools identified additional—or arguably better—matches between SNOMED CT and ICD-10-CM. Additionally, there were several cases where the I-MAGIC tool was unable to generate a match for a SNOMED CT code (e.g., “102508009 Well female child”), whereas ChatGPT-4o and Claude 3.5 Sonnet were both able to produce the same alternative

ICD-10-CM code (i.e., “Z00.129 Encounter for routine child health examination without abnormal findings”). This suggests that further formal analysis may demonstrate that GenAI tools outperform human raters. Therefore, it is likely that the results of this study significantly underestimate the accuracy and clinical validity of the matches produced by the GenAI tools.

Despite GenAI tools demonstrating significant time and cost savings, several challenges were noted throughout the conversion process. With regards to ChatGPT-4o, the process of performing the SNOMED CT-AU to ICD-10-CM conversion was not fully automated, nor was it straightforward for someone inexperienced with writing GenAI prompts to perform. When piloting the prompt, ChatGPT-4o tended to skip lines, chunks of data, or “hallucinate” (i.e., produce new input data that was not provided in the dataset). It was therefore necessary to explicitly instruct ChatGPT-4o to “manually and sequentially” convert the provided codes and to “...not hallucinate, and only convert codes which have been provided...” and “...not create new codes to convert.” When completing the final batch of conversions, the output had to be monitored for accuracy. Despite not hallucinating during the task, ChatGPT-4o still produced new input data when it ran out of the codes it had been provided.

When providing additional prompts after the algorithm had performed well, it was beneficial to provide positive reinforcement to inform ChatGPT-4o that it had performed the task correctly. This avoided ChatGPT-4o from changing its original output. There were also instances where ChatGPT-4o would attempt to terminate the task (i.e., “Unfortunately I have run out of time to process additional conversions”) but could be prompted to continue without further issue. These nuances required some level of skill and familiarity with ChatGPT-4o and GenAI prompts.

In terms of the time and labor required, ChatGPT-4o was not simply a “set and forget” solution to a large data task. Due to limitations on the volume of codes it was able to process before sometimes hallucinating, a manual “nudge” (i.e., “Please continue with the next batch”) was required after every 25 codes had been converted. This required continual monitoring of ChatGPT-4o while it was processing to ensure that lines of data were not skipped. Importantly, this renders the task impractical to complete in the background while undertaking other work.

ChatGPT-4o also imposes limits on the number of messages that are permitted within a certain timeframe (40 messages every three h). Given the number of nudges

required to process this data – in addition to further messages to adapt and rectify the prompt if it was not processing correctly – the message limit was quickly reached and required waiting until the window had lapsed before proceeding with the rest of the task. This drastically inflated the timeframe in which the task could be completed.

Claude 3.5 Sonnet provided a more streamlined tool that did not require as much skill or time to produce a prompt. One key limitation of Claude 3.5 Sonnet was the process of importing and exporting data. Unlike ChatGPT-4o, at the time of the study, Claude 3.5 Sonnet did not have the functionality to directly import or export Microsoft Excel files; however, this functionality has since been added with the release of Claude 4.0 Sonnet. Therefore, it was necessary to copy and paste lines of data from the Microsoft Excel file into Claude 3.5 Sonnet. This led to a further limitation, which was the restrictions on both message length and the number of messages permitted. As the amount of data exceeded the input limit, it was necessary to break up the prompt into smaller, more manageable batches of codes (i.e., 500 lines at a time).

Although Claude 3.5 Sonnet did not appear to “hallucinate” with a greater number of conversions, only 50 codes could be converted at a time due to limits on the maximum output message length. This however meant that the message limit (approximately 45 messages every 5 h, dependent on message length) was quickly consumed. Given that Claude 3.5 Sonnet processed codes significantly faster than ChatGPT-4o, this led to a longer waiting period between exceeding the message limit and its renewal. As Claude 3.5 Sonnet was unable to directly export a Microsoft Excel file at the end of the task, this significantly increased the time burden, as it was necessary to produce R Studio code to be run in order to produce the final output dataset. In addition to requiring the worker to have some knowledge of how to run the code in R Studio, this step accounted for the majority of the time taken to complete the task. For instance, it took 1 h and 15 min to complete the code conversion, with the remainder of the time (1 h and 55 min) spent writing and executing the R Studio code. The ability to produce downloadable Microsoft Excel files within Claude 3.5 Sonnet would rectify this limitation, significantly reducing the time and cost required to complete data analysis.

Although other methods are available for large-scale data extraction tasks, such as the creation of Application Programming Interfaces, these may require technical skill and knowledge to set up. These may also be cumbersome and impractical for ad hoc tasks performed by individuals lacking programming skills, particularly those in a busy

clinical or hospital environment.³⁶ GenAI tools remain an accessible and easy-to-use alternative that requires minimal training to achieve a cost- and time-efficient outcome. Additionally, these tools are rapidly improving over time, potentially simplifying the task even further.

4.1. Study limitations

Although this case study provides valuable insights into the use of GenAI to complete a large-scale health data analysis task, several limitations still remain. Firstly, given that this is an Australian dataset, the SNOMED-CT codes came from the Australian edition (SNOMED-CT-AU) whilst the I-MAGIC tool only caters to the standard version. Therefore, this may account for why some codes were unable to be manually converted using the I-MAGIC tool. Additionally, multiple raters were required to complete the manual coding task, thereby introducing potential issues around inter-rater reliability, particularly when coders were less familiar with the task. Furthermore, the I-MAGIC tool currently uses ICD-10-CM and has not yet been updated for the new edition of the ICD (i.e., 11th edition). There is currently no mapping tool available that enables SNOMED CT to be converted to the newer version of the ICD.

In addition, this study only considered ICD-10-CM codes to be “correct” if they either perfectly or partially matched the manual code. Given that the aim of this study was to examine whether this task could be completed using GenAI, it was outside of the scope of the study to manually examine each “incorrect” match to determine whether it was clinically valid. However, this is likely to have significantly impacted the results and led to an underestimation of the level of agreement between the GenAI tools and manual ratings.

A further limitation of this study is the rapid pace at which GenAI tools are being developed and improved. It is likely that in the time since this study was conducted, newer tools have been released that may yield different results in terms of accuracy and processing speed. However, these advancements will likely only improve the overall efficiency and accuracy of GenAI tools.

4.2. Recommendations for future research

There is significant scope for future research within this field. Firstly, further analysis of the produced data from this study is planned to examine the clinical validity of partial or incorrect matches, which will further strengthen the results of this study by producing more accurate ratings between the GenAI and manual coding output. This study used the paid versions of both ChatGPT-4o and Claude 3.5 Sonnet, which offer additional functionalities

and greater processing speed than their free versions. This study could be replicated using the free versions of these tools to compare whether the paid versions yield any difference in terms of level of agreement and processing time. It is also yet to be determined whether the time and cost savings observed in this task would translate to other data conversion tasks. Further studies using GenAI tools are needed to determine whether time and cost differences are consistent across different types of tasks. Additionally, as new GenAI tools—such as DeepSeek³⁷—are released with improvements in speed and functionality, it is recommended that this study be repeated to examine how these improvements impact the speed and accuracy with which this task can be completed. Although ChatGPT-4o and Claude 3.5 Sonnet are not specifically designed for healthcare applications, these tools were selected due to their relatively low cost and wide accessibility. The task presented in this study should also be repeated using GenAI tools specifically designed for clinical or healthcare contexts. Furthermore, the completion of similar tasks using GenAI tools should be considered to further explore their capabilities in healthcare data processing.

5. Conclusion

This study presents a case study demonstrating the use of GenAI tools to complete manual data processing tasks that are typically tedious, time-consuming, costly, and both mentally and physically demanding. The findings highlight that manual processing is often prohibitive in terms of time and cost and that alternative methods – such as the use of GenAI – warrant further exploration. GenAI provides a promising avenue for leveraging the vast quantities of currently unanalyzed healthcare data, with the potential to enhance outcomes for healthcare professionals, researchers, systems, and, most importantly, patients.

Acknowledgments

None.

Funding

This study was supported by an Australian Government Research Training Program Scholarship.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Carly Hudson

Data curation: Carly Hudson, Anu Joy, Adrian Goldsworthy

Formal analysis: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Investigation: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Methodology: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Project administration: Carly Hudson

Resources: Candice Bowman

Supervision: Marcus Randall, Candice Bowman

Writing – original draft: Carly Hudson, Marcus Randall, Adrian Goldsworthy

Writing – review & editing: All authors

Ethics approval and consent to participate

This research was approved by the Human Research Ethics Committee of the Gold Coast Hospital and Health Service (HREC/2023/QGC/95219).

Consent for publication

Not applicable.

Availability of data

Data is available from the corresponding author upon reasonable request.

Further disclosure

This research has been preprinted on Research Square (<https://doi.org/10.21203/rs.3.rs-5143761/v1>).

References

1. Murphy K. How data will improve healthcare without adding staff or beds. Cornell university, institut européen d'administration des affaires. In: Organisation WIP, editors. *Global Innovation Index 2019*. Ch. 8. Geneva: World Intellectual Property Organization; 2019.
2. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: Management, analysis and future prospects. *J Big Data*. 2019;6(1):54.
doi: 10.1186/s40537-019-0217-0
3. Australian Medical Association. *2024 Public Hospital Report Card*; 2024.
4. Feuerriegel S, Hartmann J, Janiesch C, Zschech P. Generative AI. *Bus Inform Syst Eng*. 2024;66(1):111-126.
doi: 10.1007/s12599-023-00834-7
5. Oluwagbenro MB. *Generative AI: Definition, Concepts, Applications, and Future Prospects. Authorea Preprints*; 2024.
6. Banh L, Strobel G. Generative artificial intelligence. *Electron Mark*. 2023;33(1):63.
doi: 10.1007/s12525-023-00680-1
7. Fui-Hoon Nah F, Zheng R, Cai J, Siau K, Chen L. Generative AI and chatGPT: Applications, challenges, and

- AI-human collaboration. *J Inform Technol Case Appl Res*. 2023;25:277-304.
doi: 10.1080/15228053.2023.2233814
8. Chatterjee S, Fruhling A, Kotiadis K, Gartner D. Towards new frontiers of healthcare systems research using artificial intelligence and generative AI. *Health Syst (Basingstoke)*. 2024;13:263-273.
doi: 10.1080/20476965.2024.2402128
9. Yang Y, Ngai EW, Wang L. Resistance to artificial intelligence in health care: Literature review, conceptual framework, and research agenda. *Inform Manage*. 2024;61:103961.
doi: 10.1016/j.im.2024.103961
10. Amin SM, El-Gazar HE, Zoromba MA, El-Sayed MM, Atta MHR. Sentiment of nurses towards artificial intelligence and resistance to change in healthcare organisations: A mixed-method study. *J Adv Nurs*. 2025;81(4):2087-2098.
doi: 10.1111/jan.16435
11. Chaibi A, Zaiem I. Doctor resistance of artificial intelligence in healthcare. *Int J Healthcare Inform Syst Inform (IJHISI)*. 2022;17(1):1-13.
doi: 10.4018/IJHISI.315618
12. Sobaih AEE, Chaibi A, Brini R, Abdelghani Ibrahim TM. Unlocking patient resistance to AI in healthcare: A psychological exploration. *Eur J Investig Health Psychol Educ*. 2025;15(1):6.
doi: 10.3390/ejihpe15010006
13. Li J, Dada A, Puladi B, Kleesiek J, Egger J. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed*. 2024;245:108013.
doi: 10.1016/j.cmpb.2024.108013
14. Patel SB, Lam K. ChatGPT: The future of discharge summaries? *Lancet Digit Health*. 2023;5(3):e107-e108.
doi: 10.1016/S2589-7500(23)00021-3
15. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an adjunct for radiologic decision-making. *MedRxiv*. 2023.
doi: 10.1101/2023.02.02.23285399
16. Ahn C. Exploring ChatGPT for information of cardiopulmonary resuscitation. *Resuscitation*. 2023;185:109729.
doi: 10.1016/j.resuscitation.2023.109729
17. Boßelmann CM, Leu C, Lal D. Are AI language models such as ChatGPT ready to improve the care of individuals with epilepsy? *Epilepsia*. 2023;64(5):1195-1199.
doi: 10.1111/epi.17570
18. Mogali SR. Initial impressions of ChatGPT for anatomy education. *Anat Sci Educ*. 2024;17(2):444-447.
doi: 10.1002/ase.2261
19. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care*. 2023;46(5):e103-e105.
doi: 10.2337/dc23-0197
20. Health-Europe TLR. Embracing generative AI in health care. *Lancet Reg Health Eur*. 2023;30:100677.
doi: 10.1016/j.lanepe.2023.100677
21. Open AI. *ChatGPT*. Available from: <https://chat.openai.com> [Last accessed on 2024 Sep 06].
22. Thorne E. *ChatGPT Hits 100M Weekly Users*. LinkedIn News. Available from: <https://www.linkedin.com/news/story/chatgpt-hits-100m-weekly-users-5808204> [Last accessed on 2024 Sep 09].
23. Thakur K, Barker HG, Khan Pathan AS. *Artificial Intelligence and Large Language Models: An Introduction to the Technological Future*. United States: CRC Press, LLC; 2024.
24. Google AI. *Gemini*. Available from: <https://gemini.google.com> [Last accessed on 2024 Sep 09].
25. Microsoft. *Copilot*. Available from: <https://copilot.microsoft.com> [Last accessed on 2024 Sep 09].
26. Anthropic. *Claude AI*. Available from: <https://claude.ai> [Last accessed on 2024 Sep 06].
27. Waldo J, Boussard S. GPTs and hallucination: Why do large language models hallucinate? *Queue*. 2024;22(4):19-33.
doi: 10.1145/3688007
28. Snomed International. *What is Snomed CT?* Available from: <https://www.snomed.org/what-is-snomed-ct> [Last accessed on 2024 Sep 09].
29. Snomed International. *Members*. Available from: <https://www.snomed.org/members> [Last accessed on 2024 Sep 16].
30. World Health Organisation. *International Statistical Classification of Diseases and Related Health Problems (ICD)*. Available from: <https://www.who.int/standards/classifications/classification-of-diseases> [Last accessed on 2024 Sep 09].
31. National Library of Medicine. *Snomed-CT Browser*. Available from: <https://www.nlm.nih.gov/research/umls/snomed/snomed-browsers.html> [Last accessed on 2025 Jun 09].
32. World Health Organisation. *International Statistical Classification of Diseases and Related Health Problems 10th Revision*. Available from: <https://icd.who.int/browse10/2019/en> [Last accessed on 2025 Jun 09].
33. Sallam M, Barakat M, Sallam M. A preliminary checklist (METRICS) to standardize the design and reporting of studies on generative artificial intelligence-based models in health care education and practice: Development study involving a literature review. *Interact J Med Res*. 2024;13:e54704.

doi: 10.2196/54704

34. National Library of Medicine. *Snomed CT to ICD-10-cm Map*. Available from: <https://www.nlm.nih.gov/research/umls/mapping-projects/snomedct-to-icd10cm.html> [Last accessed on 2024 Sep 06].
35. National Clinical Terminology Service. *What is Snomed CT and the AMT?* Available from: <https://www.healthterminologies.gov.au/understanding-clinical-terminology-landing/what-is-snomed-ct-and-amt> [Last accessed on 2024 Sep 23].
36. Preibisch S. *API Development*. Germany: Springer; 2018.
37. Hangzhou DeepSeek Artificial Intelligence Co. *DeepSeek*. Available from: <https://www.deepseek.com> [Last accessed on 2025 Jun 09].

ORIGINAL RESEARCH ARTICLE

Stratifying autonomic nervous system regulation patterns in healthy men: A machine learning approach

Wollner Materko^{1,2*} ¹Department of Health, Faculty of Health Sciences, Federal University of Amapá, Macapá, Amapá, Brazil²Department of Education, Faculty of Physical Education, Federal University of Amapá, Macapá, Amapá, Brazil**Abstract**

Heart rate variability (HRV) is a critical non-invasive marker of autonomic nervous system regulation and plays an essential role in cardiovascular health. Individual differences in autonomic function necessitate the development of personalized health strategies. This study aimed to develop and validate a method that integrates principal component analysis (PCA) and K-means clustering to identify distinct patterns of autonomic regulation in healthy men using HRV data. A total of 80 young, healthy men (22.0 ± 2.8 years old, 65.2 ± 6.9 kg, and 171.0 ± 6.5 cm) were recruited, and their HRV data were analyzed using time-domain and frequency-domain parameters. PCA was applied to reduce the dimensionality of the HRV data, while K-means clustering was employed to identify distinct autonomic profiles. Silhouette index values were 0.397 for one cluster, 0.481 for two clusters, and 0.556 for three clusters, indicating that the three-cluster solution provided the best fit. Three statistically distinct and physiologically meaningful clusters were identified. Cluster 3 ($n = 19$) demonstrated significantly higher HRV parameters than cluster 1 ($n = 33$) and cluster 2 ($n = 28$) ($p = 0.001$). *Post hoc* analysis further confirms that cluster 1 differed significantly from both cluster 2 and cluster 3 ($p = 0.001$). Based on HRV characteristics, the clusters were characterized as “high vagal tone,” “intermediate vagal tone,” and “low vagal tone.” The “high vagal tone” cluster exhibited the strongest parasympathetic activity, while the “low vagal tone” cluster showed evidence of sympathetic predominance. This study demonstrates a robust approach for stratifying autonomic profiles, highlighting the potential of machine learning in advancing personalized cardiovascular health assessment.

***Corresponding author:**Wollner Materko
(wollner.materko@gmail.com)

Citation: Materko W. Stratifying autonomic nervous system regulation patterns in healthy men: A machine learning approach. *Artif Intell Health*. 2025;2(4):103-113. doi: 10.36922/AIH025050006

Received: January 29, 2025**Revised:** June 24, 2025**Accepted:** June 30, 2025**Published online:** July 28, 2025**Copyright:** © 2025 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.

Keywords: Heart rate variability; Autonomic nervous system; Machine learning; Principal component analysis; K-means clustering

1. Introduction

Cardiovascular diseases (CVDs) represent a significant global health burden, accounting for an estimated 17.9 million deaths annually, according to the World Health Organization.¹ Heart rate monitoring serves as an effective method for detecting cardiac irregularities, such as arrhythmias. Continuous heart rate analysis facilitates early

diagnosis and prevention of CVDs, ultimately reducing morbidity and mortality.²

Heart rate variability (HRV) measures the beat-to-beat fluctuations in R-R intervals (RRI) as recorded by an electrocardiogram (ECG). As a non-invasive metric, HRV provides quantitative insight into the activity of the autonomic nervous system (ANS) by reflecting its modulation of cardiac function.³ Elevated HRV is generally associated with enhanced cardiovascular adaptability and greater resilience to stress, whereas reduced HRV may reflect underlying autonomic dysfunction.⁴ Due to its sensitivity, HRV has emerged as a valuable biomarker for assessing autonomic regulation and cardiac function, playing a vital role in the early detection and prevention of CVDs.⁵

However, the practical application of HRV monitoring, particularly in field settings such as athletic training, is often limited by the cost, complexity, and bulk of conventional ECG equipment.⁶ Wearable technologies, such as the Polar® Heart Rate Monitor Interface Board, provide an affordable, high-quality alternative for acquiring beat-to-beat RRI data in real-world environments.⁷ These devices enable athletes and coaches to access objective physiological feedback, thereby enhancing performance monitoring and supporting personalized, data-driven training adjustments.⁸

The stratification of cardiovascular autonomic function, as assessed by HRV in healthy men, provides valuable insights into ANS regulation and classification of cardiovascular health.⁹ As a non-invasive indicator of autonomic control, HRV quantifies the constant interplay between the sympathetic and parasympathetic nervous system's inputs to the heart. From a clinical perspective, a high level of HRV suggests that the cardiovascular system is resilient and able to cope with stress.¹⁰ Furthermore, understanding the stratification of ANS regulatory patterns may provide valuable insights into the individual cardiovascular health profiles and inform targeted prevention strategies.

The integration of artificial intelligence (AI) with a focus on machine learning offers significant potential to advance the analysis of HRV and deepen our understanding of the regulatory mechanisms of the ANS. This data-driven approach is crucial for shifting the focus from population-level analyses to more precise and personalized evaluations of autonomic health.^{4,5,11-14} which underpins the rationale for this research. The present study aims to explore ANS regulation a critical aspect of cardiovascular health through the application of AI-based methods. By stratifying individuals into distinct groups based on their HRV patterns, this study proposes a novel framework

for categorizing and analyzing variations in autonomic function. Importantly, it investigates heterogeneity within a healthy population, rather than attempting to differentiate between healthy individuals and those with disease. It is hypothesized that, even among asymptomatic young men, there are distinct patterns of autonomic regulation that are physiologically meaningful and may represent different trajectories of cardiovascular health long before clinical disease manifests.

This stratification can be achieved through an integrated analytical framework combining principal component analysis (PCA) and K-means clustering. PCA, an unsupervised dimensionality reduction technique, is used during data pre-processing to maximize variance preservation within a reduced two-dimensional space, thereby facilitating improved visualization and subsequent clustering.^{15,16} The K-means algorithm is then applied to these reduced dimensions to group individuals according to their position in the new analytical space. As an iterative partitioning method, K-means require a pre-specified number of clusters (k) and operate by minimizing intra-cluster distance while maximizing inter-cluster distance to form distinct, coherent groups.^{17,18} This study aims to develop and validate a method that integrates PCA and K-means clustering to identify distinct patterns of autonomic regulation in healthy men using HRV data.

2. State of the art

2.1. HRV

HRV refers to the variation in the time intervals between successive heartbeats.¹⁹ As a well-established indicator of ANS activity, HRV offers a non-invasive means of assessing cardiac autonomic regulation.²⁰ Historically, accurately evaluating autonomic function has presented a significant challenge for cardiologists. A breakthrough occurred in 1981 when Akselrod *et al.*²¹ demonstrated that specific components of HRV correspond directly to parasympathetic and sympathetic nervous system activity. Building on this, a 1987 study by Kleiger *et al.*²² established the clinical prognostic value of HRV, showing that it serves as a predictor of mortality risk following myocardial infarction. In recent years, HRV analysis has been widely recognized as a reliable, non-invasive method for evaluating ANS modulation at the heart's sinus node.²³

Time-domain analysis of HRV quantifies the variability in RRI using direct statistical measurements. Key parameters include the mean normal-to-normal (NN) mean R-R interval (MRR), standard deviation of NN intervals (SDNN), root mean square of successive RRI differences (RMSSD), and the proportion of adjacent NN intervals differing by more than 50 ms (pNN50).²⁴ The

reliability of these metrics is highly dependent on the integrity of the RRI data, as artifacts such as ectopic beats can significantly distort the results. Consequently, the application of filtering techniques is a crucial preprocessing step. A commonly used approach involves excluding any RRI that deviates by more than 20% from the preceding normal interval, thereby ensuring data accuracy.²⁵

Spectral analysis of HRV quantifies the power distribution of different frequency components within the sinus rhythm. The two primary methods employed are the non-parametric Fourier transform which decomposes the signal into constituent sinusoids²⁶ – and the parametric autoregressive model, which estimates the spectrum using a predictive model of the RRI.²⁷ Despite their methodological differences, both methods provide broadly comparable assessments of HRV spectra, and neither demonstrates a clear advantage over the other.

Spectral HRV analysis typically distinguishes two main frequency bands: the high-frequency (HF) band (0.15 – 0.40 Hz), which reflects parasympathetic modulation associated with respiratory sinus arrhythmia, and the low-frequency (LF) band (0.04 – 0.15 Hz), which represents a combination of sympathetic and parasympathetic influences on baroreflex regulation.²⁸ The LF/HF ratio is often calculated to estimate sympathovagal balance, with the HF band serving as an index of parasympathetic tone and the LF band representing integrated autonomic output.²⁹

Similar to the time-domain analysis, the accuracy of HRV spectral analysis is highly dependent on data quality and requires careful handling of arrhythmias. A common approach is to exclude RRIs immediately before and after ectopic beats and replace them with interpolated values based on adjacent, true RRIs. However, the exclusion of more than two RRIs surrounding an ectopic beat is generally avoided due to the risk of compromising signal continuity.³⁰

Following the Task Force report,²⁴ various non-linear metrics have emerged to analyze the complex, multi-causal, and potentially chaotic nature of HRV. These metrics apply techniques such as Lyapunov and Hurst exponents, coarse-grained spectral analysis, detrended fluctuation analysis, and entropy measures to capture the interplay of humoral, hemodynamic, and electrophysiological factors influencing HRV. While their precise physiological interpretations remain under investigation, these methods have shown promising potential in differentiating the effects of conditions such as stress and diabetes on HRV.^{31,32}

Novel metrics based on instantaneous heart rate acceleration and deceleration have also been introduced. In 2006, Bauer *et al.*³³ developed phase-rectified signal

averaging (PRSA), a statistical technique designed to analyze quasi-periodic signals in non-stationary or noisy data. By applying PRSA to NN intervals, they proposed deceleration capacity (DC) and acceleration capacity metrics derived from the coherent averaging of RRIs that exhibit increases or decreases, respectively. These indices aim to assess sympathetic modulation of sinoatrial node acceleration and deceleration, independent of other physiological factors.

The DC index, in particular, has gained significant attention due to its promising clinical implications. Notably, studies have demonstrated its superior predictive power for mortality following acute myocardial infarction compared to the widely used left ventricular ejection fraction.³⁴ Furthermore, a strong relationship has been observed between DC and the risk of sudden cardiac death in individuals with Chagas disease.³⁵ Studies have also reported a significant correlation between DC and high levels of physical conditioning,^{36,37} suggesting its potential use as a valuable marker of cardiovascular health and fitness.

2.2. PCA

PCA is a dimensionality reduction technique that transforms a set of correlated variables into a smaller number of uncorrelated linear combinations, known as principal components (PCs). These components are ordered to capture as much of the total variance in the original dataset as possible.³⁸

The first PC captures the largest proportion of the total variance. The second PC explains the maximum remaining variance, with the constraint that it is uncorrelated with the first. This process continues sequentially, with each subsequent component capturing a decreasing proportion of the remaining variance and remaining uncorrelated with all previously derived components.³⁹

Although PCA can theoretically continue until all variance is accounted for, it is typically stopped after extracting a smaller number of PCs that collectively explain a significant proportion of the total variance.³⁸ The eigenvalue associated with each PC represents the amount of variance it explains higher eigenvalues indicate greater explanatory power.⁴⁰

2.3. Cluster analysis

Cluster analysis encompasses a range of statistical techniques used to group an initially unclassified set of cases, subjects, or objects into relatively homogeneous groups, or clusters, based on observed characteristics.⁴¹ The primary goal is to identify underlying group structures without prior knowledge of group membership. Also

referred to as classification analysis or numerical taxonomy, cluster analysis differs fundamentally from discriminant analysis, which requires predefined groups. In contrast, cluster analysis aims to identify previously unknown groupings inherent in the data.⁴²

The cluster analysis process typically involves a series of key steps: defining the research problem; selecting an appropriate distance or similarity measure; choosing a clustering algorithm; determining the optimal number of clusters; interpreting the characteristics of each identified cluster; and evaluating the validity of the resulting cluster solution.⁴³ Careful selection of variables is essential and should be guided by research hypotheses, prior studies, and the researcher's informed judgment. Similarly, the choice of distance or similarity measure is critical; for instance, Euclidean distance is frequently used.⁴⁴

Clustering methods are broadly classified as hierarchical, non-hierarchical, or two-stage.⁴³ Hierarchical approaches build a nested structure of clusters either agglomeratively (bottom-up) or divisively (top-down) and do not require a predefined number of clusters. The results of these methods are visualized using dendrograms, in which branch lengths indicate inter-cluster distances. In contrast, non-hierarchical methods such as K-means require the number of clusters to be defined in advance. The choice of method depends on the distance measure used, and the resulting clusters must be interpretable and relevant to the research objectives.

3. Materials and methods

3.1. Study population

This cross-sectional study was conducted in Macapá, Brazil, and involved 80 healthy, young male participants (22.0 ± 2.8 years). Participants were recruited based on a low-risk profile for CVD.⁴⁵ Exclusion criteria included smoking, a history of cardiopulmonary disease, or the current use of any medication. All participants provided verbal informed consent before enrollment. The study protocol was approved by the Human Research Ethics Committee of the Federal University of Amapá (CAAE: 50150121.1.0000.0003) and conducted in accordance with the principles of the Declaration of Helsinki⁴⁶ and Resolution 510/2016 of the National Health Council.

3.2. Anthropometric assessment

Before enrollment, all participants received a detailed explanation of the study protocol, including measurement procedures and estimated duration. Participants were instructed to wear appropriate attire (e.g., light clothing, no shoes) and to avoid carrying objects. Height (cm) and weight (kg) were measured using a calibrated mechanical

scale (Filizola, Brazil). All anthropometric measurements were performed by a trained assessor to minimize inter-rater variability and ensure data accuracy.

3.3. Experimental procedure and data acquisition

The tests were conducted in a quiet room maintained at a temperature of 22°C. Participants were instructed to refrain from strenuous physical activity for 24 h and to avoid consuming alcohol, caffeine, or large meals for at least 3 h before their session. Upon arrival at the laboratory, participants rested quietly in a supine position for 10 min while breathing spontaneously. RRIs were recorded throughout this period using a Polar V800 heart rate monitor (Polar, Finland) with a sampling rate of 1,000 Hz. The monitor was positioned over the xiphoid process of the sternum. The first 5 min of data were discarded to allow for signal stabilization, and the subsequent 5 min were used for analysis. The tachograms of RRI were transferred via an infrared interface to Polar Precision Performance SW software version 3.0 (Polar, Finland), which automatically corrected the RRI using a moving average filter. The data were then saved as ".txt" files.

3.4. HRV analysis

For the time-domain analysis, the following parameters were calculated: MRR, SDNN, RMSSD, and the pNN50. For the frequency-domain analysis, spectral analysis was performed using the Welch periodogram method (256-point segments, 128-point overlap, and a Hanning window). This yielded normalized power for the LF (0.04 – 0.15 Hz) and HF (0.15 – 0.40 Hz) bands, both expressed as percentages. All parameters were computed in accordance with the guidelines established by the Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology,²⁴ and were implemented in MATLAB 2020.b (MathWorks, United States).

3.5. Statistical analysis

Descriptive statistics are presented as mean \pm standard deviation. The Shapiro–Wilk test was employed to assess the normality of the data distribution.

An 80×6 matrix of normalized HRV data derived from the RRI tachograms was used for dimensionality reduction to two dimensions using PCA. PCA, a dimensionality reduction technique, transforms correlated variables into uncorrelated PCs via eigenvalue decomposition of the covariance matrix.¹⁵ K-means clustering an unsupervised learning algorithm – was then applied to the reduced-dimensionality data,¹⁶ making it well-suited to classify individuals into distinct groups based on their ANS regulation. Cluster assignment was based on the Euclidean distance metric (Equation 1). Cluster centroids were

iteratively recalculated until convergence (Equation II) by minimizing the sum of squared errors (Equation III).

The following equations were applied in the K-means clustering process:

$$d(p, q) = \sqrt{\sum_{i=1}^n ((q_i - p_i))^2} \tag{I}$$

$$m = \frac{1}{n_k} \sum_{i \in C_k} x_i \tag{II}$$

Where:

- x_i is the HRV values
- n_k is the number of points in C_k
- m is the centroid of a cluster
- d is the Euclidean distance
- p is a data point
- q is a cluster centroid
- q_i is the i -th attribute of data point q
- p_i is the i -th attribute of data point p .

$$J_k = \sum_{i=1}^k \sum_{i \in C_k} (x_i - m_k)^2 \tag{III}$$

Where:

- J_k is the within-cluster sum of squares (WCSS), which is the objective function for K-means
- k is the number of clusters
- m_k is the centroid of cluster k .

To identify distinct subgroups within the dataset, the elbow method was used to determine the optimal number of clusters (k) by evaluating the WCSS, while silhouette analysis measures how well each data point fits within its assigned cluster. Once the optimal k was established, the K-means algorithm partitioned the data by iteratively refining cluster centroids until membership stabilized. The quality of this final partition was visually validated using a silhouette plot, which graphically displays the cohesion and separation of the resulting clusters.⁴⁷ To complement this analysis, agglomerative clustering was conducted using Ward’s linkage method with a Euclidean distance metric, and the output was visualized as a dendrogram.

After identifying three clusters, a one-way analysis of variance was conducted to assess whether there are significant differences in HRV parameters among the groups. Additionally, *post hoc* Tukey’s tests were applied for pairwise comparisons. The magnitude of these differences was evaluated using Cohen’s d effect size, and statistical significance was assessed using 95% confidence intervals (CI) for the mean difference. All statistical analyses were performed in MATLAB 2020b (MathWorks, United States) with a significance level set at $\alpha = 0.05$.

4. Results

Table 1 presents the anthropometric, physical, and HRV data of the participants. Normality tests indicate that age, body mass, height, and MRR follow a Gaussian distribution ($p \geq 0.05$), and the low standard deviations support the homogeneity of the sample. However, SDNN, RMSSD, pNN50, LF, and HF do not follow a normal distribution ($p < 0.05$). These findings highlight the importance of HRV data normalization in the context of machine learning, particularly when applying techniques such as PCA and K-means.

After dimensionality reduction using PCA, a non-hierarchical K-means clustering algorithm was applied to the first two PCs. The algorithm was initialized randomly

Table 1. Anthropometric characteristics and heart rate variability parameters of the participants

Variables	Mean±standard deviation	p -value
Age (years)	22.0±2.8	0.200
Body mass (kg)	65.2±6.9	0.935
Height (cm)	171.0±6.5	0.745
MRR (ms)	935.0±132.2	0.571
SDNN (ms)	62.8±30.9	0.008
RMSSD (ms)	72.7±44.6	0.001
pNN50 (%)	36.6±24.5	0.007
LF (%)	49.0±21.8	0.015
HF (%)	51.8±22.2	0.031

Abbreviations: HF: High-frequency; LF: Low-frequency; MRR: Mean R-R interval; pNN50: The proportion of adjacent normal-to-normal intervals differing by more than 50 ms; RMSSD: The root mean square of successive differences between adjacent intervals; SDNN: The standard deviation of all normal-to-normal intervals.

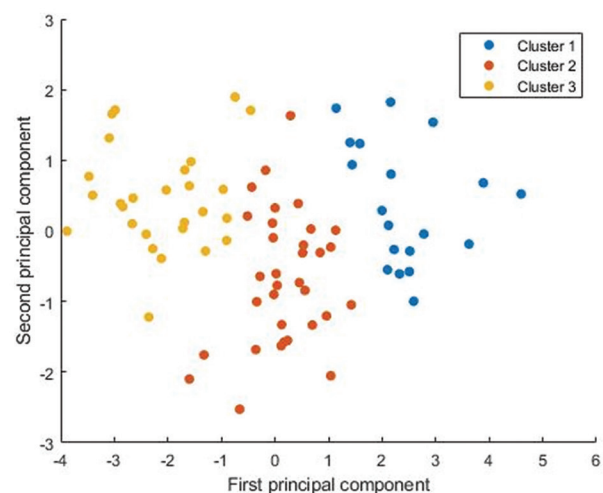


Figure 1. K-means clustering based on principal component coefficients derived from normalized heart rate variability data

and run for 10 iterations, resulting in the partitioning of the data into three distinct clusters, as shown in Figure 1.

Figure 2 shows the WCSS as a function of k , which is used to determine the optimal number of clusters in the dataset. The graph reveals a distinct “elbow” at $k = 3$, where the WCSS reduction plateaus. This suggests that three clusters represent the most appropriate choice, as adding additional clusters beyond this point results in only marginal improvements in cluster compactness. The elbow method, therefore, supports the selection of $k = 3$ as the optimal number of clusters for subsequent analysis.

As shown in Table 2, cluster 3 ($n = 19$) demonstrates significantly higher HRV parameters compared to cluster 1 ($n = 33$) and cluster 2 ($n = 28$) ($p = 0.001$). *Post hoc* analysis further confirms that cluster 1 exhibits significantly different HRV parameters compared to both cluster 2 and cluster 3 ($p=0.001$).

Table 2. Comparison of heart rate variability parameters among the identified clusters

Variables	Cluster 1	Cluster 2	Cluster 3
MRR (ms)	1034.7±129.6	919.2±121.2	886.1±113.0
SDNN (ms)	101.1±24.8	59.9±20.4	40.1±18.1
RMSSD (ms)	132.8±33.0	68.1±25.7	37.5±21.7
pNN50 (%)	67.2±10.7	37.9±16.7	14.4±13.4
LF (%)	34.4±18.0	40.6±15.5	68.9±15.8
HF (%)	69.2±15.6	59.6±15.4	30.8±16.0

Note: Data are presented as mean±standard deviation. Abbreviations: HF: High-frequency; LF: Low-frequency; MRR: Mean R-R interval; pNN50: The proportion of adjacent normal-to-normal intervals differing by more than 50 ms; RMSSD: The root mean square of successive differences between adjacent intervals; SDNN: The standard deviation of all normal-to-normal intervals.

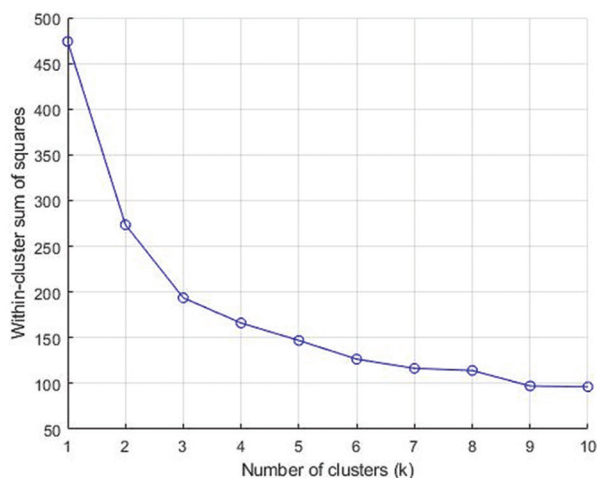


Figure 2. Within-cluster sum of squares plotted against different values of k to determine the optimal number of clusters

Cluster 1 demonstrates a significantly greater parasympathetic profile than cluster 2. This is supported by large and statistically significant differences in mean RRI (mean difference = 122.18; 95% CI = 52.67 – 191.70; $d = 0.99$), SDNN (mean difference = 44.35; 95% CI = 32.40 – 56.31; $d = 2.09$), RMSSD (mean difference = 69.81; 95% CI = 55.80 – 83.81; $d = 2.80$), and pNN50 (mean difference = 32.68; 95% CI = 25.30 – 40.06; $d = 2.49$). HF power – an established marker of vagal tone – is also significantly higher in cluster 1, with a medium-to-large effect size ($d = 0.67$). In contrast, no significant difference in LF power is observed between these two clusters (95% CI = -14.06 – 2.42).

The differences between cluster 1 and cluster 3 are even more substantial. Cluster 1 exhibits significantly higher values across all indices of overall HRV and parasympathetic activity, with very large effect sizes for mean RRI ($d = 1.32$), SDNN ($d = 3.10$), RMSSD ($d = 3.77$), pNN50 ($d = 4.48$), and HF ($d = 3.28$). Furthermore, cluster 1 exhibits significantly lower LF power than cluster 3 (mean difference = -40.18, $d = -2.75$), indicating reduced sympathetic modulation compared to cluster 3.

Although no statistically significant difference in MRR is observed between clusters 2 and 3 (95% CI = -29.45 – 92.78), their autonomic modulation profiles differ significantly. Cluster 2 exhibits significantly greater parasympathetic activity than cluster 3, as evidenced by substantial differences in SDNN, RMSSD, pNN50, and HF power. In addition, cluster 2 exhibits significantly lower LF power ($d = -2.82$), indicating increased vagal tone and reduced sympathetic modulation compared to cluster 3.

The silhouette index, a commonly used measure of cluster quality, was calculated to evaluate the PCA-based grouping of normalized HRV data. The silhouette index values indicate effective cluster separation, with low intra-cluster variability and high inter-cluster dissimilarity. The silhouette scores are 0.397 for one cluster, 0.481 for two clusters, and 0.556 for three clusters. The highest silhouette coefficient is observed at $k = 3$, supporting the selection of three clusters as the optimal solution (Figure 3).

The agglomerative hierarchical clustering dendrogram provides an alternative view of the dataset’s structure (Figure 4). While the longest vertical linkage representing the primary bifurcation suggests a two-cluster solution, further analysis reveals a distinct substructure within one of the main branches. This hierarchical arrangement indicates that a three-cluster model may more accurately reflect the underlying granularity of the data. Thus, the finer details of the dendrogram further support the three-cluster solution identified by the elbow and silhouette

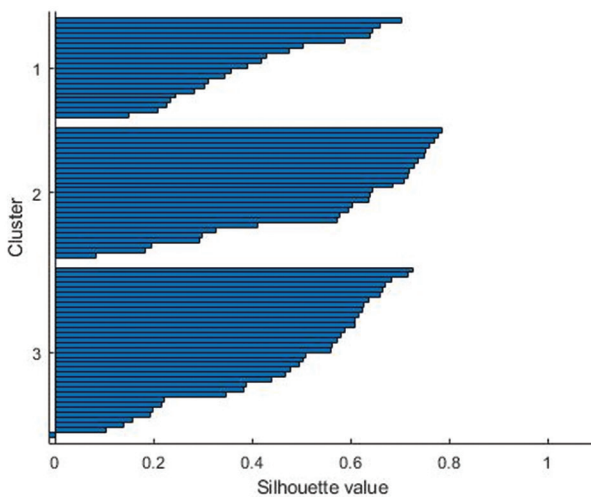


Figure 3. Silhouette plot for identifying the optimal number of clusters

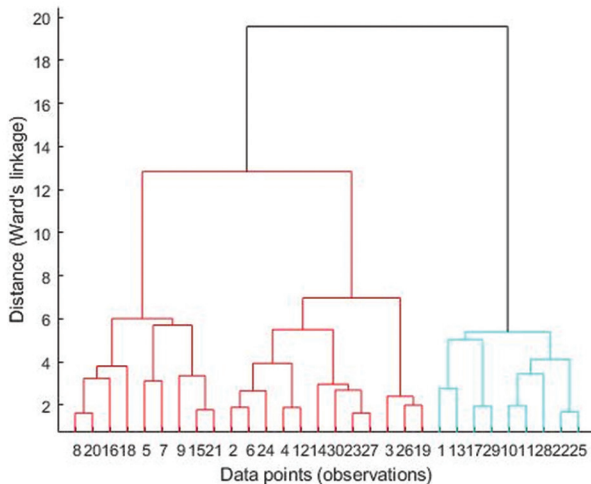


Figure 4. Hierarchical cluster analysis of the heart rate variability dataset

methods, highlighting a nested and interpretable structure within the dataset.

5. Discussion

The primary aim of this study is to develop and validate a method for identifying distinct patterns of autonomic regulation in healthy men, using a combined PCA and K-means clustering approach applied to HRV data. The findings demonstrate that this approach is both feasible and effective in stratifying cardiovascular autonomic function, revealing significant physiological heterogeneity within a homogeneous and healthy sample.

HRV is an important biomarker for assessing cardiovascular health, serving as a non-invasive indicator of the ANS and its modulation of cardiac function.⁴⁸ It reflects the balance between sympathetic and parasympathetic

activity, with implications for both physical and mental well-being.⁴⁹ Reduced HRV is generally associated with a higher risk of CVD and physiological stress,^{3,50,51} whereas elevated HRV is typically indicative of good health.^{38,52-54} The advent of wearable technology has made HRV monitoring more practical, enabling early detection of cardiovascular dysfunction and supporting proactive health management strategies.⁵⁵ However, to date, no studies have examined the use of combined PCA and K-means clustering techniques to identify individual risk profiles based on HRV data.

Pasquini *et al.*⁵⁴ applied PCA to identify five PCs in ANS time series data that collectively explained 75% of the variance during emotional reactivity tasks. K-means clustering revealed five distinct ANS states corresponding to specific emotions, such as awe and sadness.

However, the present study validates a methodological framework integrating PCA and K-means to identify distinct patterns of autonomic regulation in healthy men using HRV data. This approach demonstrates variability in autonomic balance, even within a healthy population. The findings highlight the potential of these techniques to stratify individuals based on cardiovascular autonomic function and suggest that lower HRV profiles may indicate individuals who could benefit from targeted interventions to improve autonomic balance.

A critical component of the analysis involves selecting the optimal number of clusters. Notably, both the elbow method and hierarchical dendrogram support $k = 3$ as a plausible solution, whereas the silhouette index favors a more conservative $k = 2$ solution. Rather than interpreting this discrepancy as a contradiction, it provides valuable insights into the hierarchical structure of the data. The higher silhouette score for $k = 2$ indicates the presence of two primary, well-defined groups. However, selecting $k = 3$, as recommended by the elbow method, provides greater physiological granularity by revealing an “intermediate” group that would otherwise remain undetected in a binary classification. This methodological approach allows for a more detailed and clinically interpretable classification of autonomic regulation.

In addition, these findings offer promising potential for clinical application. The ability to stratify healthy individuals into potential autonomic “risk” profiles provides insights into the development of personalized preventive strategies. A *post hoc* analysis of the three-cluster solution identified distinct, hierarchically ordered autonomic profiles. Cluster 3 (high-vagal-tone profile; $n = 33$) exhibits the highest values across all HRV parameters, indicating enhanced parasympathetic modulation. This profile is typically associated with good cardiovascular health, greater stress resilience, and efficient physical conditioning.

Cluster 2 (intermediate autonomic profile; $n = 28$) represents a moderate autonomic state, with HRV values significantly lower than those of cluster 1 but higher than those of cluster 3. Cluster 1 (low-vagal-tone or sympathetic predominant profile; $n = 19$) exhibits the lowest HRV values of parasympathetic activity and the highest relative LF power. While these individuals are clinically healthy, this profile may indicate a subclinical state of autonomic imbalance or reduced adaptive capacity, potentially indicating elevated long-term cardiovascular risk.

The identification of these three distinct profiles in a young, healthy population represents a central finding of this study. While the literature establishes that reduced HRV is associated with disease, the present findings suggest that a spectrum of autonomic function exists even in the absence of clinical pathology. This observation aligns with previous research, including that by Pasquini *et al.*,⁵⁴ who applied similar techniques to identify autonomic states during emotional reactivity. The present study further highlights interindividual variability at rest, suggesting that these autonomic profiles may represent individual physiological traits.

The use of machine learning, particularly K-means clustering, underscores the potential of AI to enhance HRV analysis and deepen understanding of ANS regulation. This data-driven approach enables the identification of distinct autonomic patterns that may be obscured by traditional statistical methods. The combined PCA and K-means clustering method serves as an effective tool for exploring complex physiological data and identifying meaningful subgroups within a population. Furthermore, the application of machine learning techniques to predict individual risk based on HRV profiles holds significant promise for enabling more personalized and effective preventive strategies.^{4,5,11-13}

It is crucial to interpret these findings within the context of the study's design. Unlike studies that aim to distinguish between healthy individuals and those with cardiac disease, a supervised classification problem this study addresses an unsupervised discovery question: what distinct autonomic profiles exist within a healthy population? This approach is based on the hypothesis that physiologically meaningful patterns may reflect distinct long-term cardiovascular health trajectories, even among asymptomatic individuals, before the onset of clinical disease. Therefore, the identification of three distinct clusters should not be viewed as a limitation due to the absence of a patient control group, but rather as a validation of the study's primary objective, which explores an under-investigated area of autonomic profiling in healthy populations.

However, several limitations should be acknowledged in this study. A key limitation is the homogeneous nature of the sample, which consisted of 80 healthy young men from Macapá, Brazil. Consequently, the findings should be interpreted as proof of concept within this specific demographic. The study emphasizes that the identified cluster patterns may not be directly generalizable to women, other age groups, ethnic backgrounds, or populations with different health and lifestyle characteristics. Therefore, validating the method is a priority, and future studies are recommended to include larger, more diverse, and multi-center cohorts to determine the broader applicability of these autonomic profiles. Furthermore, the cross-sectional design of the study limits the ability to draw conclusions about the temporal stability of these profiles or their predictive value for future health outcomes.

6. Conclusion

This study highlights that the application of combined PCA and K-means clustering to HRV data is a valid and effective approach for identifying distinct patterns of autonomic regulation in healthy men. Three distinct physiological profiles, high vagal tone, intermediate vagal tone, and low vagal tone, were identified, even within a demographically homogeneous cohort. These findings reinforce the importance of HRV as a sensitive biomarker of cardiovascular health and underscore the potential of machine learning techniques to advance personalized preventive strategies. However, future longitudinal studies are needed to examine the stability of these autonomic profiles and their association with long-term health outcomes across more diverse populations.

Acknowledgments

The author would like to thank the Physical Education Department at the Federal University of Amapá for allowing them to use their laboratory facilities and for helping them to recruit participants.

Funding

This research was funded by the Amapá Research Support Foundation through its public call 003/2018, specifically within the "Research Program for the Unified Health System (SUS): management in Health-PPSUS." The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of interest

The author declares no conflicts of interest.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

The study protocol was ethically approved by the Human Research Ethics Committee of the Federal University of Amapá (CAAE: 50150121.1.0000.0003; approval number: 5.121.013) and conducted in accordance with the Declaration of Helsinki. All participants provided verbal informed consent before their inclusion in the study.

Consent for publication

Verbal informed consent was obtained from all participants before their inclusion in the study. The consent form explicitly stated that anonymized data would be analyzed and subsequently used for publication in scientific journals. All data were fully anonymized before analysis, and the results are presented in an aggregated format to ensure that no individual participant can be identified.

Availability of data

The dataset generated and analyzed during the current study is not publicly available, since it contains sensitive clinical information that is subject to data protection regulations.

References

1. Timmis A, Vardas P, Townsend N, *et al.* European society of cardiology: Cardiovascular disease statistics 2021. *Eur Heart J.* 2022;43(8):716-799.
doi: 10.1093/eurheartj/ehab892
2. Chorin E, Hochstadt A, Schwartz AL, Matz G, Viskin S, Rosso R. Continuous heart rate monitoring for automatic detection of life-threatening arrhythmias with novel bio-sensing technology. *Front Cardiovasc Med.* 2021;8:707621.
doi: 10.3389/fcvm.2021.707621
3. Zeid S, Buch G, Velmeden D, *et al.* Heart rate variability: Reference values and role for clinical profile and mortality in individuals with heart failure. *Clin Res Cardiol.* 2024;113(9):1317-1330.
doi: 10.1007/s00392-023-02248-7
4. Materko W, Fernandes DF, Façanha CCR, *et al.* A machine learning approach to developing an accurate stratification of type 2 diabetes mellitus based on heart rate variability parameters using the K-means clustering technique in elderly women. *Gazz Med Ital Arch Sci Med.* 2024;183:44-50.
doi: 10.23736/S0393-3660.23.05096-9
5. Materko W, Fernandes DF, Sadala MN, Pureza D, Alberto AAD, Pena FPS. Evaluation on heart rate variability parameters in elderly with type 2 diabetes mellitus using principal component analysis. *Gazz Med Ital Arch Sci Med.* 2022;181:879-884.
doi: 10.23736/S0393-3660.22.04782-9
6. Gillinov S, Etiwy M, Wang R, *et al.* Variable accuracy of wearable heart rate monitors during aerobic exercise. *Med Sci Sports Exerc.* 2017;49(8):1697-1703.
doi: 10.1249/MSS.0000000000001284
7. Materko W, Dos Reis Façanha CC, Guedes GC, *et al.* Temporal cross-correlation between Polar® heart rate monitor interface board and ECG to measure RR interval at rest. *Isokinet Exerc Sci.* 2024;32(1):59-64.
doi: 10.3233/IES-230061
8. Latino F, Tafuri F. Wearable sensors and the evaluation of physiological performance in elite field hockey players. *Sports (Basel).* 2024;12(5):124.
doi: 10.3390/sports12050124
9. Trevizani GA, Nasario-Junior O, Benchimol-Barbosa PR, Silva LP, Nadal J. Cardiac autonomic changes in middle-aged women: Identification based on principal component analysis. *Clin Physiol Funct Imaging.* 2016;36(4):269-273.
doi: 10.1111/cpf.12222
10. Perrone MA, Volterrani M, Manzi V, Barchiesi F, Iellamo F. Heart rate variability modifications in response to different types of exercise training in athletes. *J Sports Med Phys Fitness.* 2021;61(10):1411-1415.
doi: 10.23736/S0022-4707.21.12480-6
11. Rafeisangari R, Shiri N. A multi-adaptive neuro-fuzzy inference system with variable thresholds for heartbeat classification. *Artif Intell Health.* 2024;1(4):43-60.
doi: 10.36922/aih.3367
12. Nshimiyimana M, Niyogisubizo J, Ninteretse JD. Heartbeat classification using various machine learning models: A comparative study. *Artif Intell Health.* 2024;1(4):61-72.
doi: 10.36922/aih.3543
13. Materko W, Miranda SAM, Bezerra THL, De Oliveira Figueira CAM. Heart rate variability in soccer players and the application of unsupervised machine learning. *Explor Cardiol.* 2025;3:101241.
doi: 10.37349/ec.2025.101241
14. Materko W. Stratification fitness aerobic based on heart rate variability during rest by principal component analysis and K-means clustering. *J Exerc Physiol.* 2018;21(1):91-101.
15. Kaya IE, Pehlivanlı AÇ, Sekizkardeş EG, Ibriki T. PCA based clustering for brain tumor segmentation of T1w MRI images. *Comput Methods Programs Biomed.* 2017;140:19-28.
doi: 10.1016/j.cmpb.2016.11.011
16. Jolliffe IT, Cadima J. Principal component analysis: A review

- and recent developments. *Philos Trans A Math Phys Eng Sci.* 2016;374(2065):20150202.
doi: 10.1098/rsta.2015.0202
17. Ikotun AM, Absalom E, Abualigah L, Abuhaija B, Jia H. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inform Sci.* 2023;622:178-210.
doi: 10.1016/j.ins.2022.11.139
18. Sinaga KP, Yang MS. Unsupervised K-Means clustering algorithm. *IEEE Access.* 2020;8:80716-80727.
doi: 10.1109/access.2020.2992018
19. Ishaque S, Khan N, Krishnan S. Trends in heart-rate variability signal analysis. *Front Digit Health.* 2021;3:639444.
doi: 10.3389/fdgth.2021.639444
20. Kleiger RE, Stein PK, Bigger JT Jr. Heart rate variability: Measurement and clinical utility. *Ann Noninvasive Electrocardiol.* 2005;10(1):88-101.
doi: 10.1111/j.1542-474X.2005.10101.x
21. Akselrod S, Gordon D, Ubel FA, Shannon DC, Barger AC, Cohen RJ. Power spectrum analysis of heart rate fluctuations: A quantitative probe of beat-to-beat cardiovascular control. *Science.* 1981;213(4504):220-222.
doi: 10.1126/science.6166045
22. Kleiger RE, Miller JP, Bigger JT Jr., Moss AJ, The Multicenter Post-Infarction Research Group. Decreased heart rate variability and its association with increased mortality after acute myocardial infarction. *Am J Cardiol.* 1987;59(4):256-262.
doi: 10.1016/0002-9149(87)90795-8
23. Sammito S, Thielmann B, Böckelmann I. Update: Factors influencing heart rate variability-a narrative review. *Front Physiol.* 2024;15:1430458.
doi: 10.3389/fphys.2024.1430458
24. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: Standards of measurement, physiological interpretation and clinical use. *Circulation.* 1996;93(5):1043-1065.
doi: 10.1161/01.cir.93.5.1043
25. Clifford GD, Mcsharry PE, Tarassenko L. Characterizing artefact in the normal human 24-hour RR time series to aid identification and artificial replication of circadian variations in human beat to beat heart rate using a simple threshold. *Comput Cardiol.* 2002;29:129-132.
doi: 10.1109/CIC.2002.1166724
26. Chemla D, Young J, Badilini F, et al. Comparison of fast Fourier transform and autoregressive spectral analysis for the study of heart rate variability in diabetic patients. *Int J Cardiol.* 2005;104(3):307-313.
doi: 10.1016/j.ijcard.2004.12.018
27. Staffini A, Svensson T, Chung UI, Svensson AK. Heart rate modeling and prediction using autoregressive models and deep learning. *Sensors (Basel).* 2021;22(1):34.
doi: 10.3390/s22010034
28. Hayano J, Yuda E. Assessment of autonomic function by long-term heart rate variability: Beyond the classical framework of LF and HF measurements. *J Physiol Anthropol.* 2021;40(1):21.
doi: 10.1186/s40101-021-00272-y
29. Jarczok MN, Weimer K, Braun C, et al. Heart rate variability in the prediction of mortality: A systematic review and meta-analysis of healthy and patient populations. *Neurosci Biobehav Rev.* 2022;143:104907.
doi: 10.1016/j.neubiorev.2022.104907
30. Choi A, Shin H. Quantitative analysis of the effect of an ectopic beat on the heart rate variability in the resting condition. *Front Physiol.* 2018;9:922.
doi: 10.3389/fphys.2018.00922
31. Immanuel S, Teferra MN, Baumert M, Bidargaddi N. Heart rate variability for evaluating psychological stress changes in healthy adults: A scoping review. *Neuropsychobiology.* 2023;82(4):187-202.
doi: 10.1159/000530376
32. Ardashev A, Loskutov A, Passman R, Zhelyakov E, Rytkin E, Efimov I. Theoretical and practical aspects of the nonlinear dynamics' methods of heart rate variability analyses in tachyarrhythmia patients underwent radiofrequency catheter ablation. *Cardiovasc Eng Technol.* 2025;16(2):190-201.
doi: 10.1007/s13239-024-00766-7
33. Bauer A, Kantelhardt JW, Bunde A, et al. Phase-rectified signal averaging detects quasi-periodicities in non-stationary data. *Physica A.* 2006;364:423-434.
doi: 10.1016/j.physa.2005.08.080
34. Bauer A, Kantelhardt JW, Barthel P, et al. Deceleration capacity of heart rate as a predictor of mortality after myocardial infarction: Cohort study. *Lancet.* 2006;367(9523):1674-1681.
doi: 10.1016/S0140-6736(06)68735-7
35. Alberto AC, Pedrosa RC, Zarzoso V, Nadal J. Association between circadian Holter ECG changes and sudden cardiac death in patients with Chagas heart disease. *Physiol Meas.* 2020;41(2):025006.
doi: 10.1088/1361-6579/ab6ebc
36. Materko W, Bartels R, Pecanha T, Lima JRP, Carvalho ARS, Nadal J. Maximum oxygen uptake prediction model based on heart rate variability parameters for young healthy adult

- males at rest. *Open Access Biostat Bioinform.* 2018;2(1):1-7.
doi: 10.31031/oabb.2018.02.000536
37. Nasario-Junior O, Benchimol-Barbosa PR, Pedrosa RC, Nadal J. Assessment of autonomic function by phase rectification of RR-interval histogram analysis in Chagas disease. *Arq Bras Cardiol.* 2015;104(6):450-455.
doi: 10.5935/abc.20150032
38. Greenacre M, Groenen PJF, Hastie T, et al. Principal component analysis. *Nat Rev Methods Primers.* 2022;2(1):100.
doi: 10.1038/s43586-022-00184-w
39. Peres-Neto PR, Jackson DA, Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal.* 2005;49(4):974-997.
doi: 10.1016/j.csda.2004.11.005
40. Cadima J, Jolliffe IT. Loadings and correlations in the interpretation of principal components. *J Appl Stat.* 1995;22(2):203-214.
doi: 10.1080/02664769524627
41. Dalmaijer ES, Nord CL, Astle DE. Statistical power for cluster analysis. *BMC Bioinformatics.* 2022;23(1):205.
doi: 10.1186/s12859-022-04675-1
42. Lawton M, Ben-Shlomo Y, May MT, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. *J Neurol Neurosurg Psychiatry.* 2018;89(12):1279-1287.
doi: 10.1136/jnnp-2018-318337
43. Scitovski R, Sabo K, Martínez-Álvarez F, Ungar Š. *Cluster Analysis and Applications.* Berlin: Springer; 2021.
44. Sneath PHA. A method for testing the distinctness of clusters: A test of the disjunction of two clusters in Euclidean space as measured by their overlap. *Math Geol.* 1975;7(2):123-143.
doi: 10.1007/BF02068305
45. Holtrop J, Bhatt DL, Ray KK, et al. Impact of the 2021 European society for cardiology prevention guideline's stepwise approach for cardiovascular risk factor treatment in patients with established atherosclerotic cardiovascular disease. *Eur J Prev Cardiol.* 2024;31(6):754-762.
doi: 10.1093/eurjpc/zwae038
46. World Medical Association. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA.* 2013;310(20):2191-2194.
doi: 10.1001/jama.2013.281053
47. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math.* 1987;20:53-65.
doi: 10.1016/0167-8396(87)90125-7
48. Raju S, Woo H, Fawzy A, et al. Decreased cardiac autonomic function is associated with higher exacerbation risk and symptom burden in chronic obstructive pulmonary disease. *J COPD Found.* 2023;10(3):328-334.
doi: 10.15326/jcopdf.2023.0410
49. Sundas A, Contreras I, Navarro-Otano J, Soler J, Beneyto A, Vehi J. Heart rate variability over the decades: A scoping review. *PeerJ.* 2025;13:e19347.
doi: 10.7717/peerj.19347
50. Osborne MT, Shin LM, Mehta NN, Pitman RK, Fayad ZA, Tawakol A. Disentangling the links between psychosocial stress and cardiovascular disease. *Circ Cardiovasc Imaging.* 2020;13(8):e010931.
doi: 10.1161/circimaging.120.010931
51. Vaccarino V, Bremner JD. Stress and cardiovascular disease: An update. *Nat Rev Cardiol.* 2024;21(9):603-616.
doi: 10.1038/s41569-024-01024-y
52. Faust O, Hong W, Loh HW, et al. Heart rate variability for medical decision support systems: A review. *Comput Biol Med.* 2022;145:105407.
doi: 10.1016/j.combiomed.2022.105407
53. Lu L, Zhu T, Morelli D, et al. Uncertainties in the analysis of heart rate variability: A systematic review. *IEEE Rev Biomed Eng.* 2024;17:180-196.
doi: 10.1109/rbme.2023.3271595
54. Pasquini L, Noohi F, Veziris CR, et al. Dynamic autonomic nervous system states arise during emotions and manifest in basal physiology. *Psychophysiology.* 2023;60(4):e14218.
doi: 10.1111/psyp.14218
55. Turcu AM, Ilie AC, Ștefăniu R, et al. The impact of heart rate variability monitoring on preventing severe cardiovascular events. *Diagnostics (Basel).* 2023;13(14):2382.
doi: 10.3390/diagnostics13142382

ORIGINAL RESEARCH ARTICLE

RefSAM3D: Adapting the Segment Anything Model with cross-modal references for three-dimensional medical image segmentation

Xiang Gao  and Kai Lu* 

Department of Anesthesiology, Nanjing Drum Tower Hospital, Nanjing University, Nanjing, Jiangsu, China

Abstract

The Segment Anything Model (SAM), originally built on a two-dimensional vision transformer, excels at capturing global patterns in two-dimensional natural images but faces challenges when applied to three-dimensional (3D) medical imaging modalities such as computed tomography and magnetic resonance imaging. These modalities require capturing spatial information in volumetric space for tasks such as organ segmentation and tumor quantification. To address this challenge, we introduce RefSAM3D, an adaptation of SAM for 3D medical imaging by incorporating a 3D image adapter and cross-modal reference prompt generation. Our approach modifies the visual encoder to handle 3D inputs and enhances the mask decoder for direct 3D mask generation. We also integrate textual prompts to improve segmentation accuracy and consistency in complex anatomical scenarios. By employing a hierarchical attention mechanism, our model effectively captures and integrates information across different scales. Extensive evaluations on multiple medical imaging datasets demonstrate that RefSAM3D outperforms state-of-the-art methods. Our work thus advances the application of SAM in accurately segmenting complex anatomical structures in medical imaging.

Keywords: Three-dimensional medical imaging; Cross-modal reference prompt; Volumetric segmentation; Vision transformer

***Corresponding author:**Kai Lu
(961340955@qq.com)

Citation: Gao X, Lu K. RefSAM3D: Adapting the Segment Anything Model with cross-modal references for three-dimensional medical image segmentation. *Artif Intell Health*. 2025;2(4):114-128. doi: 10.36922/AIH025080010

Received: February 17, 2025

Revised: May 1, 2025

Accepted: June 23, 2025

Published online: August 14, 2025

Copyright: © 2025 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 image segmentation is a fundamental task in medical imaging, primarily aimed at identifying and extracting specific anatomical structures, such as organs, lesions, and tissues, from medical images. This process is crucial for numerous clinical applications, including computer-aided diagnosis, treatment planning, and disease progression monitoring. Accurate image segmentation provides precise volumetric and shape information about target structures, which is essential for further clinical applications such as disease diagnosis, quantitative analysis, and surgical planning.¹⁻³

Currently, recent breakthroughs in foundational models for image segmentation^{4,5} have yielded transformative results, leveraging extensive datasets to capture general representations that exhibit exceptional generalizability and performance. However, despite these strides, significant challenges arise when applying these models, particularly

the Segment Anything Model (SAM), to medical image segmentation. For example, Huang *et al.*⁶ demonstrated that SAM performs suboptimally on medical data, especially with objects that have irregular shapes or low contrast. Three main factors limit SAM's effectiveness in this domain. First, medical images, which often differ significantly from natural images, tend to be smaller, irregular in shape, and low in contrast, complicating direct application of the model. Second, medical structures typically have blurred or indistinct boundaries, whereas SAM's pre-training data includes predominantly well-defined edges, reducing segmentation accuracy and stability. Finally, medical imaging data often exists in three-dimensional (3D) forms with rich volumetric details. Yet, SAM's hint engineering was developed for two-dimensional (2D) data, limiting its ability to leverage 3D spatial features essential in medical contexts.

To enhance SAM's performance in medical imaging tasks, it is crucial to adapt and fine-tune the model to address domain-specific challenges. Recent studies have shown that parameter-efficient transfer learning (PETL) techniques, such as Low-Rank Adaptation⁷ and Adapters,⁸ are effective in this context. For instance, Med-Tuning⁹ reduces the domain gap between natural images and medical volumes by incorporating Med-Adapter modules into pretrained visual foundation models. SAMed¹⁰ employs the Low-Rank Adaptation fine-tuning strategy to adjust the image encoder, prompt encoder, and mask decoder of the SAM, achieving a balance between performance and deployment cost. However, these approaches predominantly focus on pure 2D adaptation, not fully exploiting the 3D information inherent in volumetric medical data. Nowadays, research is gradually shifting focus to better utilize the extensive data available in the 3D domain. The related methodologies can be categorized into two main approaches: one relies on prompt design based on SAM,¹¹⁻¹³ and the other achieves fully automatic segmentation when the segmented objects exhibit relatively regular shapes and positions.^{14,15} The automatic prompt generation fails to leverage specialized medical knowledge and struggles to capture critical features due to blurred boundaries and small targets in medical images. These limitations result in suboptimal performance of automated methods, indicating further optimization.

In this paper, we propose Ref-SAM3D, an innovative approach that integrates textual prompts to enhance segmentation accuracy and consistency in complex anatomical scenarios. By incorporating text-based cues, our method enables SAM to perform referring expression segmentation within a 3D context, allowing the model to process both visual inputs and semantic descriptions for more intelligent segmentation strategies. We introduce

a hierarchical attention mechanism that significantly improves the model's ability to capture and integrate information across different scales. This mechanism focuses on critical feature layers while filtering out irrelevant data, thereby enhancing segmentation precision and robustness, particularly in complex 3D structures. By integrating information across multiple scales, the model achieves a nuanced understanding of volumetric data, leading to more precise medical image segmentation. In addition, we adapt the visual encoder to handle 3D inputs and enhance the mask decoder for direct 3D mask generation, bridging the gap between SAM's 2D architecture and the demands of 3D medical imaging. This adaptation is crucial for ensuring the model's applicability and effectiveness in this domain. We evaluate our approach on multiple medical imaging datasets, demonstrating its superior performance compared to state-of-the-art methods. Our experiments highlight the effectiveness of our model in accurately segmenting complex anatomical structures, thereby advancing the application of SAM in medical imaging. The contributions of our work are as follows:

- (i) We introduce a cross-modal reference prompt generation mechanism that integrates text and image embeddings into a unified feature space, facilitating effective cross-modal interaction.
- (ii) We develop a hierarchical attention mechanism that significantly improves the model's ability to capture and integrate information across different scales, leading to improved segmentation precision and robustness, particularly in complex 3D structures.
- (iii) We achieve state-of-the-art results across multiple benchmarks, demonstrating superior performance in 3D medical image segmentation tasks.

2. Related work

2.1. Vision foundation models (VFMs)

With the rapid development of foundation models in computer vision, recent research has focused on leveraging large-scale pre-training to create adaptable models with zero-shot and few-shot generalization capabilities.¹⁶⁻¹⁹ These VFMs draw inspiration from language foundation models like generative pre-trained transformers (GPT) series, showing remarkable adaptability across domains and tasks using pre-training and fine-tuning paradigms.²⁰ Notable examples include the Contrastive Language-Image Pre-training (CLIP) model²¹ and the A Large-scale Image and Noisy-text embedding (ALIGN) model,²² which employ image-text pairs to achieve zero-shot generalization across tasks such as classification and video understanding. Building on these foundations, segmentation-specific models such as the segment-everything-everywhere model²³ and SegGPT²⁴ have emerged to address more

complex tasks. The segment-everything-everywhere model enhances VFM capabilities by introducing a universal prompting scheme that enables semantic-aware open-set segmentation, expanding their use in real-world scenarios. SegGPT, in turn, standardizes segmentation data and employs in-context learning for both images and videos, allowing it to handle diverse segmentation tasks without requiring additional task-specific training. Complementing these advances, DINOv2²⁵ scales up Vision Transformer (ViT) pre-training by increasing data and model size, producing more general and transferable visual features that simplify fine-tuning across a wide range of tasks, further broadening VFM applicability. The SAM⁴ is one of the most notable VFMs for general-purpose image segmentation. Pre-trained on 11 million images and 1 billion masks, SAM enables interactive, prompt-driven zero-shot segmentation across a wide variety of visual tasks. Its impressive versatility has made it a key model for applications such as image segmentation, inpainting, and tracking. However, it still faces limitations in specific domains such as medical imaging, camouflage detection, and shadow segmentation.²⁶

2.2. Adaptation of the SAM in medical imaging

The adaptation of SAM for medical imaging has evolved rapidly, driven by its impressive zero-shot performance in natural image segmentation. Initial evaluation studies²⁷⁻³⁰ examined SAM's applicability to medical image segmentation, but its performance often fell short due to the domain gap between natural and medical images. For instance, He *et al.*²⁸ noted a performance gap of up to 70% in Dice scores compared to domain-specific models. This highlighted the need for task-specific fine-tuning. Following this, research attention shifted from evaluation to the adaptation of SAM for medical images.^{12,13,15,17} Several studies have experimented with fine-tuning SAM by modifying its prompt design to handle the specific characteristics of medical data. SAM-Med2D,³¹ for example, leveraged more comprehensive prompts, including points, bounding boxes, and masks, to optimize SAM for 2D medical image segmentation, whereas the medical SAM adapter¹² incorporated point prompts and adapters to inject medical domain knowledge into SAM's architecture. Although these approaches enhanced SAM's performance, the creation of prompts for each 2D slice of 3D medical data proved to be labor-intensive. Efforts to adapt SAM for 3D medical image segmentation have focused on overcoming this limitation. MedLSAM³² and SAM3D³³ applied SAM to 3D datasets, with approaches such as SAMed¹⁰ and Med-Tuning⁹ employing techniques such as Low-Rank Adaptation to fine-tune SAM for 3D tasks. However, most of these methods have not fully

addressed the critical need to account for 3D volumetric or temporal information, which is vital for medical image segmentation. Innovations such as 3DSAM-Adapter¹³ and modality-agnostic SAM (MA-SAM)³⁴ have incorporated 3D convolutional adapters to transform SAM's 2D architecture into one capable of recognizing 3D structures. Similarly, SAMMed3D¹¹ introduced a framework to generate 3D prompts from 2D points, helping SAM process volumetric data more effectively. The success of these 3D adaptations highlights the importance of leveraging spatial information for more accurate segmentation. Recent trends indicate a shift toward prompt-free or semiautomatic systems, like AutoSAM Adapter,¹⁵ which aim to maintain SAM's zero-shot capabilities while minimizing manual prompt generation.

2.3. PETL

With the widespread adoption of foundational models, PETL has garnered significant attention. PETL methods can be categorized into three main groups. One approach is addition-based methods, which involve integrating lightweight adapters or prompts into the original model. These adapters or prompts allow the fine-tuning of only a small number of additional parameters, enabling the model to adapt to specific tasks while preserving the majority of its pre-trained weights. This approach minimizes the computational overhead associated with training large models, as only the newly introduced components require optimization.^{9,35} Another strategy focuses on specification-based methods, which prioritize the identification and tuning of a small proportion of influential parameters from the original model. This method often employs techniques such as sensitivity analysis to determine which parameters have the most significant impact on the model's performance for a given task. By selectively updating these parameters, specification-based methods aim to achieve efficient adaptation while reducing training burden and maintaining high performance levels.^{10,13} In addition, reparameterization-based methods leverage low-rank representations to minimize the number of trainable parameters during the fine-tuning process. Techniques such as Low-Rank Adaptation and factorized tuning allow models to maintain their expressive power while significantly reducing the number of parameters that need to be adjusted. This approach not only enhances efficiency but also enables strong performance across various PETL tasks, as it effectively captures the essential features required for adaptation.⁷ Recently, PETL techniques have been successfully utilized to adapt VFMs for a wide range of downstream tasks, including image classification, object detection, and, notably, medical image segmentation. Researchers have explored ways to fine-tune vision models

efficiently while addressing the unique challenges posed by these complex tasks.³⁵⁻³⁷

2.4. Image segmentation by referring expressions

Referring image segmentation is a task that involves segmenting a specific object in an image based on a natural language description. This task requires the model to understand both the visual content of the image and the semantic meaning of the text, making it a challenging problem at the intersection of computer vision and natural language processing. With the advent of large-scale vision-language models, the performance of referring image segmentation has significantly improved. Models such as CLIP³⁸ and ALIGN³⁹ leverage large datasets of image-text pairs to learn joint embeddings that can be used for various vision-language tasks, including referring image segmentation. These models have demonstrated strong zero-shot and few-shot capabilities, enabling them to generalize well to unseen tasks and datasets. Recent advances have seen the adoption of transformer architectures for referring expression-based image segmentation. Transformer-based models, such as the ViT,⁴⁰ have been adapted to this task by integrating textual information into the visual processing pipeline. Ding *et al.*⁴¹ introduced a vision-language transformer approach that leverages transformer and multi-head attention mechanisms to establish deep interactions between vision and language features, significantly enhancing holistic understanding. Similarly, cross-modal attention mechanisms have become a key component in modern referring image segmentation models. These mechanisms enable the model to effectively combine visual and textual features by computing attention scores between the two modalities. Li *et al.*⁴² introduced the hierarchical dense attention module to fuse hierarchical visual semantic information with sparse embeddings to obtain fine-grained dense embeddings, and an implicit tracking module to generate a tracking token and provide historical information for the mask decoder.

3. Method

3.1. Overview of Ref-SAM3D

The original SAM, built on a 2D ViT, is proficient in capturing global patterns within 2D natural images. However, its applicability is limited when it comes to medical imaging modalities such as computed tomography (CT) and magnetic resonance imaging (MRI), which involve 3D volumetric data. In these contexts, 3D information is essential for applications such as organ segmentation and tumor quantification, as the characteristics of these structures must be captured from a 3D perspective. Relying solely on 2D views can result

in reduced accuracy due to potential boundary blurring and non-standard scanning postures. Moreover, medical images differ significantly from natural images in both content and structure, demanding higher anatomical precision and detail. Directly applying segmentation models trained on natural images to medical domains thus yields limited effectiveness. Figure 1 shows the proposed method, RefSAM3D.

3.2. 3D volumetric input processing

To enhance SAM's performance in medical imaging tasks, the model needs to be adapted and fine-tuned to accommodate the domain-specific challenges. We introduced a 3D image adapter to enable SAM's processing of volumetric data.

We first modified the visual encoder to handle 3D volumetric inputs. Given a 3D medical volume $V \in R^{C \times D \times H \times W}$, where C , D , H , and W denote the channel, depth, height, and width, respectively, we extracted the 3D features through the following steps.

3.2.1. Patch embedding

We approximated a $k \times k \times k$ convolution (with $k = 14$) by employing a combination of $1 \times k \times k$ and $k \times 1 \times 1$ 3D convolutions. The $1 \times k \times k$ convolution was initialized with pre-trained 2D convolution weights, which remain frozen during fine-tuning. To manage the complexity of the model, we applied depth-wise convolutions for the newly introduced $k \times 1 \times 1$ convolutions, reducing the number of parameters that require tuning.

3.2.2. Positional encoding

In the pre-trained ViT model, we introduced an additional learnable lookup table with dimensions $(C \times D)$ to encode the positional information for 3D points (d , h , and w). By summing the positional embedding from the frozen (h, w) table with the learnable depth-axis embedding, we provided accurate positional encoding for the 3D data.

3.2.3. Attention block

The attention block was directly adjusted to accommodate 3D features. For 2D inputs, the query size was (B, HW, C) , which is easily modified to (B, DHW, C) for 3D inputs while retaining all pretrained weights. We adopted a sliding window mechanism, similar to that in the Swin Transformer, to mitigate memory overhead resulting from the increased dimensionality, optimizing the model's performance and memory footprint.

3.2.4. Bottleneck

As in other studies, we enhanced the bottleneck layer to better adapt to 3D tasks. Specifically, we replaced

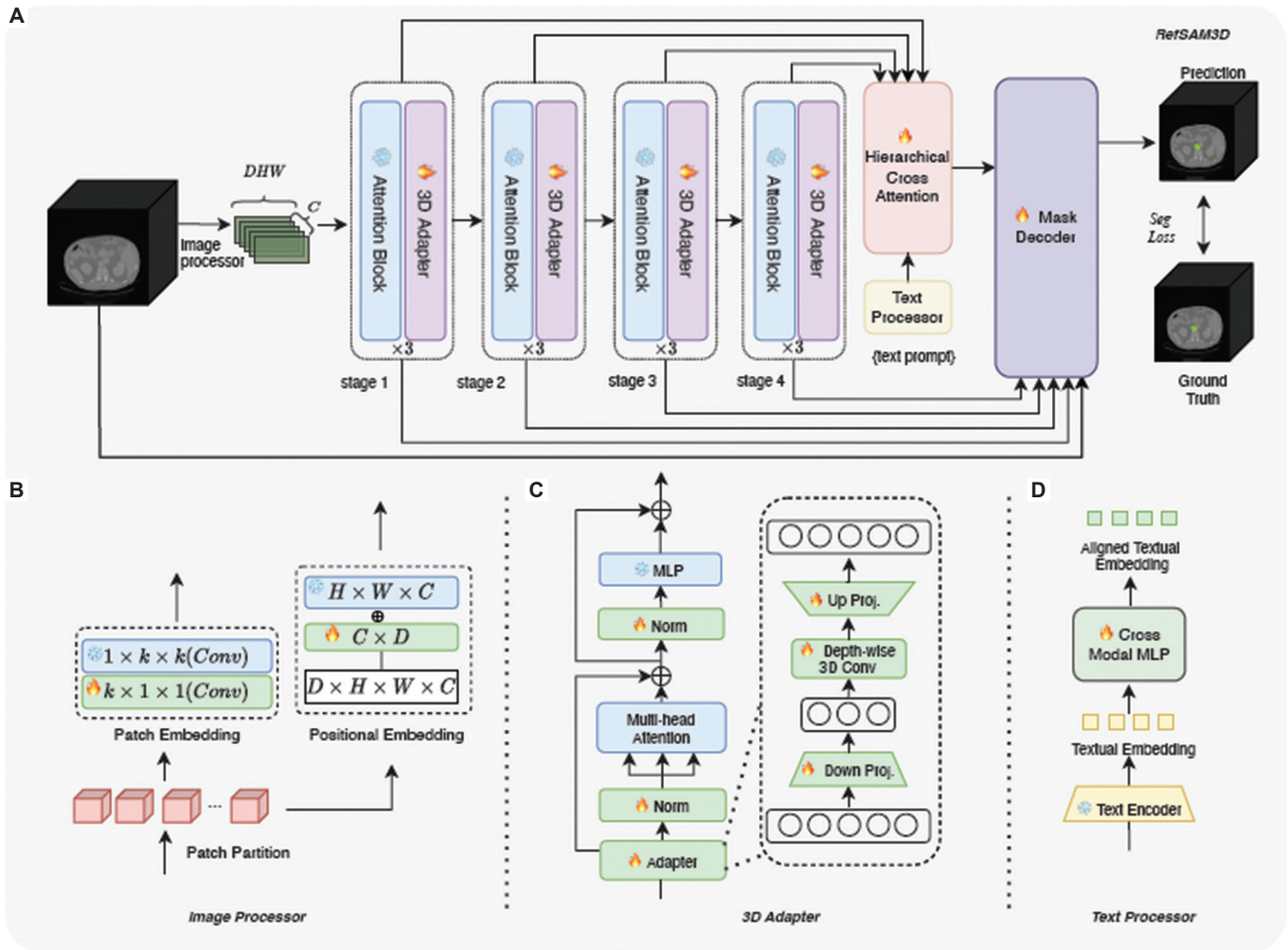


Figure 1. The proposed RefSAM3D method. (A) The overview of our proposed RefSAM3D for three-dimensional (3D) medical image segmentation, which integrates hierarchical cross-attention between image and text modalities to generate accurate segmentation predictions. (B) The design of the image processor, which includes patch partitioning, convolution-based patch embedding, and positional embedding, is used to process volumetric 3D medical data. (C) The framework of the 3D adapter incorporates multi-head attention, depth-wise 3D convolution, and up/down projection for efficient feature extraction and adaptation. (D) The pipeline of the text processor encodes textual prompts and aligns them with visual embeddings using a cross-modal multilayer perceptron for enhanced segmentation guidance.

2D convolutions with 3D ones and trained these layers from scratch to improve performance. To avoid the computational expense of fully fine-tuning a 3D ViT, we employed a lightweight adapter for efficient fine-tuning. The adapter comprised a down-projection and an up-projection linear layer, formulated as shown in Equation I:

$$\text{Adapter}(X) = X + \text{Act}(XW_{\text{Down}})W_{\text{Up}} \quad (\text{I})$$

where $X \in \mathbb{R}^{N \times C}$ represents the input feature, $W_{\text{Down}} \in \mathbb{R}^{C \times N}$ and $W_{\text{Up}} \in \mathbb{R}^{N \times C}$ are the down-projection and up-projection layers, and $\text{Act}(\cdot)$ is the activation function. In addition, we incorporated depth-wise convolutions after the down-projection layer to enhance 3D spatial awareness.

3.3. Cross-modal reference prompt generation

3.3.1. Text encoder

Within the SAM framework, we carefully designed a text encoder to process textual prompts related to image segmentation tasks. Specifically, we employed the text encoder from the CLIP model, which can convert input textual prompts, such as “perform liver segmentation,” into corresponding text embedding vectors.

The textual prompt was first tokenized into a sequence of tokens $T = t_{ii}^L = 1$. These tokens were then input into the CLIP text encoder to obtain the final embedding representation. The output of the text encoder is expressed as the formula shown in Equation II:

$$\mathcal{F}_e = \varepsilon_t(T) \in \mathbb{R}^{L \times C_e} \quad (\text{II})$$

Here, \mathcal{F}_e is the sequence of L word embeddings, each with C_e dimensions, i.e., $\mathcal{F}_e = f_{i=1}^L$, where each word is represented by a C_e -dimensional embedding. By applying a pooling operation over these word embeddings, we obtained a sentence-level embedding $\mathcal{F}_e^s \in \mathbb{R}^{C_e}$.

3.3.2. Cross-modal projector

While text embeddings derived from pre-trained language models capture rich semantic representations, a significant gap exists between these representations and those obtained from visual encoders. This semantic disparity poses challenges in cross-modal fusion, as the two modalities do not naturally reside in the same embedding space. To address this, we adopted a strategy inspired by vision-and-language bidirectional encoder representations from transformers, wherein we employed a multilayer perceptron to align the text and image embeddings. This allows both modalities to be projected into a unified feature space, enabling more effective interaction. Specifically, for each word embedding f_i in \mathcal{F}_e , the sparse embedding can be obtained by adopting the cross-modal multilayer perceptron (Equation III):

$$\mathcal{F}_i^s = MLP(f_i) \in \mathbb{R}_{C_e} \tag{III}$$

3.3.3. Image feature extraction

As previously mentioned, we integrated lightweight adapters into our 3D SAM to efficiently adapt the model for processing volumetric medical images. In this step, we extracted the features produced by each attention block as cross-attention visual hierarchical features.

Let $V_i \in \mathbb{R}^{B \times D_i \times H_i \times W_i \times C}$ denote the output of the i^{th} attention block, where B is the batch size, and H_i , W_i and D_i represent the height, width, and depth of the feature maps, respectively. This extraction allowed us to leverage the unique focus of each attention block on different aspects of

the input data, capturing a rich representation of 3D spatial patterns. The adapted features are computed as Equation IV:

$$V_i' = Adapter_i(V_i), \quad \forall i \in \{1, 2, \dots, N\} \tag{IV}$$

where $N = 4$. We can obtain a collection of image features, as depicted in Equation V:

$$V' = V_1', V_2', \dots, V_N' \tag{V}$$

3.3.4. Hierarchical cross-attention

The hierarchical cross-attention architecture is designed to integrate multi-level visual features with textual inputs, enabling a deeper understanding of cross-modal data in 3D tasks such as medical image analysis. By extracting hierarchical features from each attention block in a 3D SAM, the architecture leverages the fact that each layer focuses on different aspects of the input data, from low-level details to high-level semantics. This structure enhances the model's ability to relate complex 3D spatial patterns with corresponding textual prompts, improving cross-modal understanding. Figure 2 shows the hierarchical cross-attention architecture.

In this architecture, the inputs include both the hierarchical image features, $V' = V_1', V_2', \dots, V_N'$, derived from each attention block, and a textual prompt T , which encodes the semantic information. These inputs are fused through a cross-attention mechanism where each layer of visual features interacts with the textual input, allowing mutual enrichment of modalities. The output is a cross-modal prompt that combines visual and textual information, which can be fed into SAM's prompt encoder to guide tasks such as segmentation or object detection in 3D medical images.

In the hierarchical cross-attention architecture, the cross-attention mechanism is designed to facilitate interaction between the hierarchical image features and

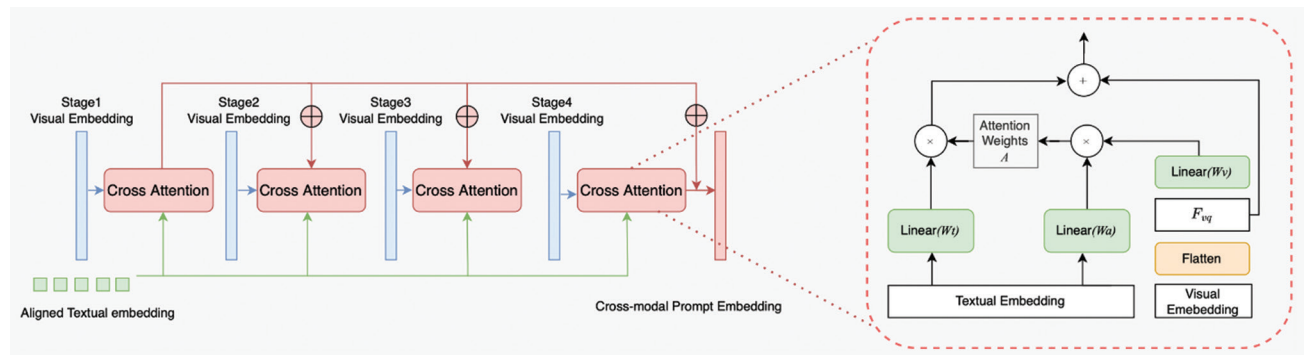


Figure 2. The structure of the cross-modal prompt embedding module. The left part illustrates the overall architecture, where hierarchical visual embeddings from four stages interact with aligned textual embeddings using cross-attention mechanisms to generate cross-modal prompt embeddings. The right part details the cross-attention mechanism, showing how attention weights are computed to align textual and visual embeddings through linear transformations and fusion, enabling effective multi-modal integration for downstream tasks.

the textual prompt. As mentioned above, V_i' represents the adapted feature maps extracted from the i th attention block, and the textual prompt is $T \in \mathbb{R}^{B \times L \times C}$.

The cross-attention process can be formally expressed as follows (Equation VI). For each hierarchical feature F_i' ,

we computed the attention scores A_i with respect to the text T :

$$A_i = \text{softmax} \left(\frac{Q_i K^T}{\sqrt{d_k}} \right) \tag{VI}$$

where $Q_i \in \mathbb{R}^{B \times D_i H_i W_i \times C}$ are the queries derived from F_i' , and $K \in \mathbb{R}^{B \times L \times C}$ are the keys derived from the textual prompt T . The dimensionality d_k represents the size of the keys, which is a scaling factor to ensure stable gradients during training. The attention output O_i for each feature block can then be computed as in Equation VII:

$$O_i = A_i V_i \tag{VII}$$

Where V_i denotes the values corresponding to F_i' and is similarly dimensioned as F_i' . The final output from the cross-attention mechanism can be represented as Equation VIII:

$$O = [O_1, O_2, \dots, O_N] \in \mathbb{R}^{B \times DHW \times C} \tag{VIII}$$

resulting in a combined output that integrates both visual and textual information across multiple layers. This enriched representation was then utilized as a cross-modal prompt in the subsequent stages of SAM's prompt encoder, effectively bridging the gap between visual features and semantic understanding derived from text.

3.4. Lightweight mask decoder

The original SAM mask decoder comprises merely two transformer layers, two transposed convolution layers, and a multilayer perception layer. In the context of 3D medical image processing tasks, we replaced the 2D convolutions with 3D convolutions to enable direct 3D mask generation. Given that many anatomical structures or lesions in medical images are relatively small, it is often necessary to achieve higher resolution images to ensure better distinction of the segmented elements.

In the image encoder of the SAM, the patch embedding process of the transformer backbone embeds each 16×16 patch into a feature vector, resulting in a 16×16 down-sampling of the input. The SAM mask decoder employs two consecutive transposed convolution layers to up-sample the feature map by a factor of four. However, the final prediction generated by SAM still has a resolution that is four times lower than the original input shape. To address

this problem, we employed progressive up-sampling, making moderate adjustments to the SAM decoder by integrating two additional transposed convolution operations. With each layer up-sampling the feature maps by a factor of two, the four transposed convolutional layers progressively restored feature maps to their original input resolution. In addition, we introduced a multilayer aggregation mechanism, designing a network akin to a "U-shaped" architecture. We combined intermediate feature maps from stages 1–4 during the image encoder phase with prompts generated during the cross-modal reference prompt generation phase to enrich the mask features. After up-sampling the mask feature map to the original resolution, we concatenated it with the original image and used another 3D convolution to fuse the information and generate the final mask to better leverage information from the original resolution.

4. Experiments

4.1. Experimental setup

We conducted a comprehensive evaluation of our segmentation method across four medical image segmentation tasks, encompassing three distinct imaging modalities: CT-based tumor segmentation, MRI-based cardiac segmentation, and multi-organ segmentation from multi-modal datasets. Our approach was rigorously compared against state-of-the-art methods on CT imaging tasks. In addition, we assessed our method's performance on MRI cardiac and multi-organ segmentation tasks, providing a thorough analysis of its generalization capabilities and conducting an in-depth ablation study to elucidate the contributions of its constituent components.

4.1.1. Datasets

The kidney tumor segmentation (KiTS21) dataset⁴³ is a comprehensive collection designed for the segmentation of kidneys, tumors, and cysts in CT imaging. It comprises 300 publicly available training cases and 100 withheld testing cases. The dataset is formatted in 3D CT with files stored in the .nii.gz format. The image dimensions exhibit significant variability, with voxel spacing ranging from (0.5, 0.44, 0.44) mm to (5.0, 1.04, 1.04) mm and sizes ranging from (29, 512, 512) to (1,059, 512, 796). The dataset includes annotations for three anatomical structures: kidneys, tumors, and cysts. These structures are consistently present across all training cases, with cysts appearing in 49.33% of the cases. This dataset serves as a critical resource for advancing automated segmentation techniques in medical imaging analysis.

The Medical Segmentation Decathlon (MSD) pancreas tumor dataset¹² consists of 281 contrast-enhanced

abdominal CT scans with annotations for both the pancreas and pancreatic tumors. This dataset is part of the MSD pancreas segmentation challenge. Each CT volume has a resolution of 512×512 pixels, with the number of slices per scan ranging from 37 to 751. The authors filtered the dataset to retain only the axial view images containing more than 5% pancreatic content. Consistent with previous studies, we merged the pancreas and pancreatic tumor masks into a single entity for segmentation.

The liver tumor segmentation benchmark (LiTS)⁴⁴ dataset is a publicly available benchmark dataset focused on liver and liver tumor segmentation. It was created to evaluate and compare the performance of automated liver and liver tumor segmentation algorithms. The LiTS dataset comprises 201 abdominal CT scans, of which 194 contain liver lesions. The dataset is divided into 131 training cases and 70 testing cases. The resolution and quality of the CT images vary, with axial resolutions ranging from 0.56 mm to 1.0 mm and z-direction resolutions ranging from 0.45 mm to 6.0 mm.

The MSD colon dataset⁴⁵ is a publicly available benchmark dataset focused on primary colon cancer segmentation from CT images. The dataset consists of 190 abdominal CT scans in total, which are divided into 126 training cases and 64 testing cases. Each case is annotated with segmentation masks identifying the primary colon cancer regions.

For cardiac segmentation, we utilized the multi-modality whole heart segmentation (MM-WHS) Challenge 2017 dataset,⁴⁶ which contains 20 CT and 20 MRI scans with pixel-level ground-truth annotations. These scans were collected in a real clinical setting and include five anatomical labels: left ventricle blood cavity, right ventricle blood cavity, left atrium blood cavity, right atrium blood cavity, and ascending aorta. In our experiments, only the CT scans were used, which contain between 177 and 363 slices, each with a resolution of 512×512 pixels and voxel spacing ranging from 0.3 to 0.6 mm.

The Beyond the Cranial Vault (BTCV) challenge dataset⁴⁷ comprises 30 CT volumes, each manually labeled with 13 different abdominal organs. The number of slices per scan ranges between 85 and 198, with a slice thickness varying between 2.5 mm and 5.0 mm. All scans have an axial resolution of 512×512 , whereas the in-plane resolution varies from $0.54 \times 0.54 \text{ mm}^2$ to $0.98 \times 0.98 \text{ mm}^2$. We followed the data split proposed by Tang *et al.*,⁴⁸ utilizing 24 cases for training and 6 cases for testing.

For evaluating the model's generalization ability, we also used the multi-modality abdominal multi-organ segmentation challenge (AMOS22) dataset.⁴⁹ This dataset

includes abdominal CT and MRI scans from different patients, with each scan annotated for 15 organs. In line with the approach in MA-SAM, we limited our evaluation to the 12 organs common to both the AMOS22 and BTCV datasets. For generalization testing, we utilize 300 CT scans and 60 MRI scans from the AMOS22 training and validation sets.

4.1.2. Implementation details

We implemented our method and benchmarked it against baseline models using PyTorch (version 2.7.1) and the medical open network for AI framework, specifically utilizing SAM-B for all experiments, which employs ViT-B as the image encoder backbone. The training was conducted on an NVIDIA A40 GPU (United States) with a batch size of 1, using the AdamW optimizer with a linear learning rate scheduler for a total of 200 epochs. The initial learning rate was set to $1e-4$, with a momentum of 0.9 and a weight decay of $1e-5$. Data preprocessing involved adjusting the isotropic spacing to 1 mm. For data augmentation, we applied various transformations, including random rotation, flipping, erasing, shearing, scaling, translation, posterization, contrast adjustments, brightness modifications, and sharpness enhancements. During training, we also sampled foreground and background patches at a 1:1 ratio. For single-organ cancer segmentation, we assessed our method's performance through comparisons with state-of-the-art volumetric segmentation and fine-tuning techniques, using the dice coefficient and normalized surface dice (NSD) as evaluation metrics, similar to Sam-med3d [11]. For multi-organ segmentation, we employed the dice coefficient and Hausdorff distance (HD) as evaluation metrics. For each dataset, we designed specific text prompts to guide the segmentation process, as shown in Table 1. These prompts were carefully crafted to provide clear anatomical context while maintaining consistency across different organs and pathologies.

4.2. Comparison with state-of-the-art methods

Our method was extensively evaluated against a wide range of state-of-the-art 3D medical image segmentation techniques on both CT and MRI datasets. These techniques include the convoluted neural network-based no new (nn)U-Net⁵⁰—an automated configuration framework evolved from the U-Net architecture⁵¹—and the Swin U-Net transformers (Swin-UNETR),⁵² which employs a hierarchical encoder structure for 3D segmentation tasks. Furthermore, we also considered nnFormer,⁵³ a model that integrates both local and global volumetric self-attention mechanisms, and UNETR++,⁵⁴ which enhances segmentation accuracy and efficiency through

the introduction of an efficient pairing attention module. In addition, we compared our approach with 3D UNet-eXpanded Network (UX-Net),⁵⁵ a method designed to create a simple, efficient, and lightweight network that combines the capabilities of hierarchical transformers with the advantages of ConvNet modules. We also evaluated SAM-B, which is the base model of SAM trained on natural images and directly applied to medical images without adaptation. Finally, our method was benchmarked against the latest SAM adaptation techniques, including 3DSAM-adapter¹³—a promptable 3D medical image segmentation model—and MA-SAM,³⁴ a framework that utilizes parameter-efficient fine-tuning strategies and 3D adapters.

The results presented in Table 2 demonstrate that our proposed Ref-SAM3D method consistently outperformed

other approaches across a wide range of tasks, achieving the highest scores in nearly all scenarios, particularly excelling in challenging tumor types. In kidney tumor segmentation, despite challenges such as low contrast with surrounding tissues, blurred boundaries, and high morphological heterogeneity, Ref-SAM3D achieved a dice score of 95.53% and an NSD of 99.45%, surpassing other methods. For pancreatic tumors, which constitute less than 0.5% of CT images and exhibit diverse shapes, Ref-SAM3D achieved a dice score of 82.42%, representing a 2.12% improvement over existing state-of-the-art techniques. In liver tumor segmentation, Ref-SAM3D attained a dice score of 80.10%, effectively handling variations in grayscale and irregular shapes. Despite the extensive distribution and complex anatomical structure of colorectal cancer lesions,

Table 1. Datasets used in our experiments and their corresponding prompt content descriptions

Task	Dataset name	Prompt content
Kidney tumor segmentation	KiTS21 Challenge	CT images, kidneys, tumors, and cysts segmentation, spacing (0.5, 0.44, 0.44) mm to (5.0, 1.04, 1.04) mm, dimensions (29, 512, 512) to (1,059, 512, 796)
Pancreas tumor segmentation	MSD pancreas	CT images, pancreas tumor segmentation, resolution 512×512, slices 37–751
Liver tumor segmentation	LiTS dataset	CT images, liver tumor segmentation, axial resolution 0.56–1.0 mm, z-direction resolution 0.45–6.0 mm
Colon cancer segmentation	MSD colon dataset	CT images, colon cancer segmentation, and abdominal scans
MRI cardiac segmentation	MM-WHS Challenge	MRI images, cardiac structure segmentation (LVC, RVC, LAC, RAC, AA), resolution 512×512, voxel spacing 0.3–0.6 mm
Abdominal multi-organ segmentation	BTCV Challenge	CT images, abdominal organ segmentation (13 organs), slice thickness 2.5–5.0 mm, in-plane resolution 0.54×0.54 mm ² to 0.98×0.98 mm ²
Multi-modality abdominal multi-organ segmentation	AMOS22 dataset	CT and MRI images, abdominal organ segmentation (15 organs), varying modalities and resolutions

Abbreviations: AA: Ascending aorta; AMOS: Abdominal Multi-Organ Segmentation; BTCV: Beyond the Cranial Vault; CT: Computed tomography; MM-WHS: Multi-Modality Whole Heart Segmentation; MRI: Magnetic resonance imaging; MSD: Medical Segmentation Decathlon; LAC: Left atrium blood cavity; LiTS: Liver Tumor Segmentation Benchmark; LVC: Left ventricle blood cavity; RAC: Right atrium blood cavity; RVC: Right ventricle blood cavity.

Table 2. Comparison with classical medical image segmentation methods on four tumor segmentation datasets

Methods	Kidney tumor		Pancreas tumor		Liver tumor		Colon cancer	
	Dice	NSD	Dice	NSD	Dice	NSD	Dice	NSD
nnU-Net	73.07	77.47	41.65	62.54	60.10	75.41	43.91	52.52
Swin-UNETR	65.54	72.04	40.57	60.05	50.26	64.32	35.21	42.94
UNETR++	56.49	60.04	37.25	53.59	37.13	51.99	25.36	30.68
nnFormer	45.14	42.28	36.53	53.97	45.54	60.67	24.28	32.19
3D UX-Net	57.59	58.55	34.83	52.56	45.54	60.67	28.50	32.73
SAM-B (10 pts/slice)	40.07	34.96	30.55	32.91	8.56	5.97	39.14	42.70
3DSAM-adapter (10 points/volume)	74.91	84.35	57.47	79.62	56.61	69.52	49.99	65.67
MA-SAM (1 relaxed 3D bounding box/slice)	93.38	98.91	80.30	97.19	75.23	92.31	65.45	81.40
Ref-SAM3D	95.53	99.45	82.42	98.41	80.10	93.23	70.14	88.90

Note: All data presented as percentages (%).

Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

Ref-SAM3D achieved a dice score of 70.14%, marking a 10.11% increase over current technologies. It is noteworthy that traditional methods like nnU-Net perform well on certain tasks, yet overall, they fall short compared to newer methods such as Ref-SAM3D. Particularly, when dealing with tumors that have blurred boundaries and diverse morphologies, Ref-SAM3D demonstrated significant advantages. These findings underscore the exceptional performance of Ref-SAM3D in addressing a variety of complex medical image segmentation challenges. Figure 3 shows the qualitative visualizations of these tasks.

In the domain of multi-organ segmentation, we conducted experiments on the BTCV dataset. The Ref-SAM3D approach demonstrated exceptional capability, achieving a dice score of 97.1% for spleen segmentation, as shown in Table 3, which surpasses all comparative methods. The left and right kidneys attained dice scores of 96.1% and 94.9%, respectively. The esophagus achieved a dice score of 85.2%, surpassing other methods, whereas the liver and stomach achieved scores of 97.3% and 94.1%, respectively. Furthermore, Ref-SAM3D demonstrated efficiency in handling complex anatomical structures, such as the pancreas and aorta, achieving dice scores of 87.5% and 92.3%, respectively. Ref-SAM3D achieved an average HD value of 2.34, underscoring its superior boundary precision. Figure 4 shows qualitative visualizations of BTCV tasks. From the qualitative visualization results, Ref-SAM3D demonstrated superior performance in multi-organ segmentation tasks. The method accurately

identified and segmented boundaries between different organs, maintaining high segmentation precision even in cases with blurred organ boundaries or complex anatomical structures. Notably, Ref-SAM3D maintained stable segmentation performance for both small organs such as the pancreas, and elongated structures, such as the aorta, further validating the reliability of the quantitative evaluation metrics.

In addition, in the context of cardiac tumor segmentation using MRI, as shown in Figure 5, a qualitative assessment of predicted masks from various segmentation models indicates that our AutoSAM Adapter produced visually superior results, especially in terms of boundary precision, when compared to existing state-of-the-art methods.

4.3. Generalization evaluation

To assess the generalization capabilities of Ref-SAM3D, we conducted comprehensive experiments across heterogeneous datasets and imaging modalities. Our evaluation framework encompassed two distinct scenarios: cross-modality generalization on the AMOS22 dataset (comprising both CT and MRI modalities) and cross-dataset adaptation using the MM-WHS cardiac imaging dataset.

In the zero-shot generalization experiments, we evaluated the model's transferability by applying our Ref-SAM3D, trained exclusively on the BTCV CT dataset, to the AMOS22 dataset without any additional fine-tuning. The quantitative results demonstrated remarkable

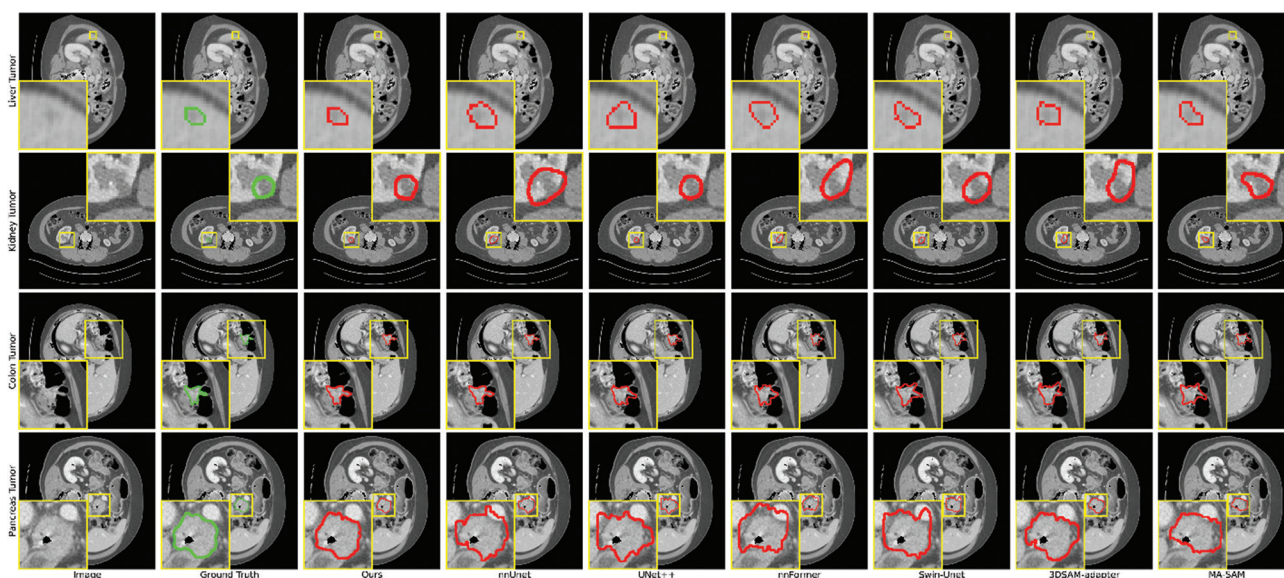


Figure 3. Qualitative visualizations of the proposed method and baseline approaches on liver tumor, kidney tumor, pancreas tumor, and colon cancer segmentation tasks
 Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network

Table 3. Comparison of abdominal multi-organ segmentation results

Metric	Method	Spleen	R.Kd	L.Kd	GB	Eso.	Liver	Stomach	Aorta	IVC	Veins	Pancreas	AG	Average
Dice (%)	nnU-Net	97.0	95.3	95.3	63.5	77.5	97.4	89.1	90.1	88.5	79.0	87.1	75.2	86.3
	Swin-UNETR	95.6	94.2	94.3	63.6	75.5	96.6	79.2	89.9	83.7	75.0	82.2	67.3	83.1
	UNETR++	94.2	92.1	95.4	65.0	75.9	96.9	88.3	85.5	84.9	76.1	81.8	71.3	83.95
	nnFormer	93.5	94.9	95.0	64.1	79.5	96.8	90.1	89.7	85.9	77.8	85.6	73.9	85.6
	3D UX-Net	94.6	94.2	94.3	59.3	72.2	96.4	73.4	87.2	84.9	72.2	80.9	67.1	81.4
	3DSAM-adapter	94.3	96.1	94.1	62.9	79.9	96.1	83.8	88.4	85.3	75.6	83.1	69.4	84.1
	MA-SAM	96.7	95.1	95.4	68.2	82.1	96.9	92.8	91.1	87.5	79.8	86.6	73.9	87.2
	Ref-SAM3D	97.1	94.9	96.1	70.3	85.2	97.3	94.1	92.3	88.8	80.4	87.5	75.1	88.3
HD (%)	nnU-Net	1.07	1.19	1.19	7.49	8.56	1.14	4.84	14.11	2.87	5.67	2.31	2.23	4.39
	Swin-UNETR	1.21	1.41	1.37	2.25	5.82	1.70	13.75	5.92	4.46	7.58	3.53	3.40	4.37
	UNETR++	5.99	1.23	1.33	5.99	10.37	33.12	5.23	8.23	2.14	10.34	3.12	2.13	7.44
	nnFormer	78.03	1.41	1.43	3.00	4.92	1.38	4.24	7.53	4.02	6.53	2.96	2.76	9.95
	3D UX-Net	3.17	1.59	1.26	4.53	13.92	1.75	19.72	12.53	3.47	9.99	3.70	4.11	6.68
	3DSAM-adapter	3.38	1.23	1.21	2.23	5.43	1.15	4.00	6.47	7.88	5.18	4.71	3.94	3.90
	MA-SAM	1.00	1.19	1.07	1.59	3.77	1.36	3.87	5.29	3.12	3.25	3.93	2.57	2.67
	Ref-SAM3D	1.30	1.32	1.00	1.21	3.18	1.23	3.77	4.12	2.30	3.12	3.08	2.44	2.34

Abbreviations: 3D: Three-dimensional; AG: Average; Eso.: Esophagus; GB: Gall bladder; HD: Hausdorff distance; IVC: Inferior vena cava; L.Kd: Left kidney; nn: No new; R.kd: Right kidney; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

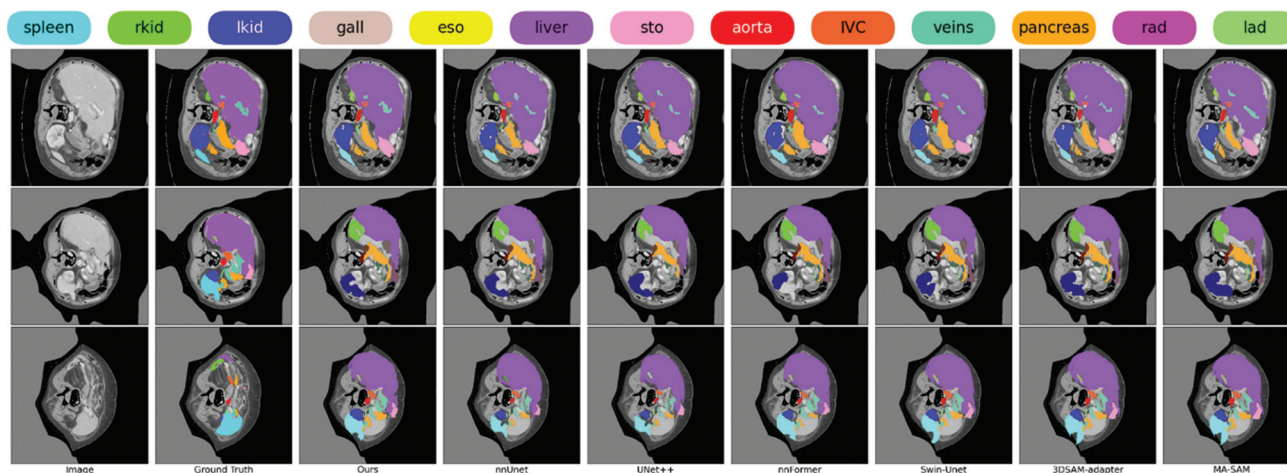


Figure 4. Qualitative visualization of segmentation results generated from our Ref-SAM3D method and other state-of-the-art methods on the Beyond the Cranial Vault dataset. Rkid and Lkid refer to the right and left kidneys, respectively. Sto, rad, and lad stand for stomach, respectively. Abbreviations: 3D: Three-dimensional; IVC: Inferior vena cava; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

performance, achieving a mean dice coefficient of 85.7% on CT images, indicating robust generalization across different CT acquisition protocols and patient cohorts. Notably, in the challenging cross-modality scenario of MRI segmentation, our model maintained substantial performance with a dice score of 63.2% ($\pm 3.1\%$), significantly surpassing baseline methods, including nnU-Net (12.1%) and Swin-UNETR (15.3%).

Furthermore, when employing a five-shot fine-tuning strategy on the AMOS22 MRI data, Ref-SAM3D exhibited even more impressive results, achieving a dice score of 84.1% (Figure 6). This represents a substantial improvement over the fine-tuned versions of nnU-Net (72.4%) and Swin-UNETR (75.3%), demonstrating the model’s superior adaptability and learning efficiency with minimal additional training data. These results underscore

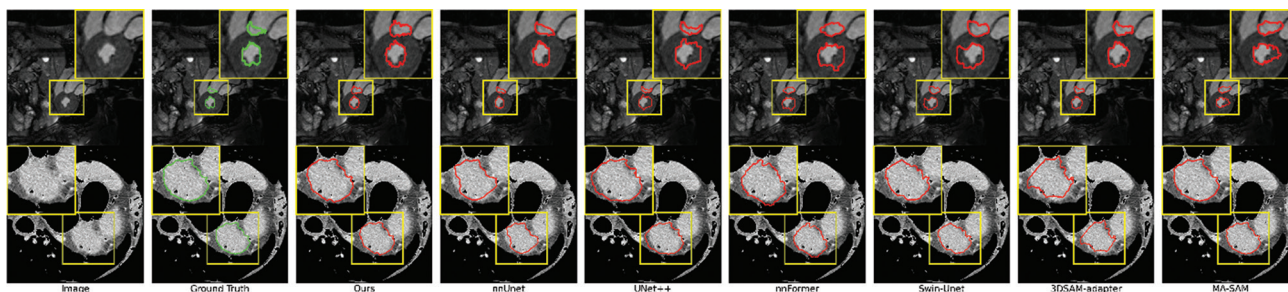


Figure 5. Qualitative visualization of segmentation results generated from different methods for magnetic resonance imaging cardiac tumor segmentation. Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

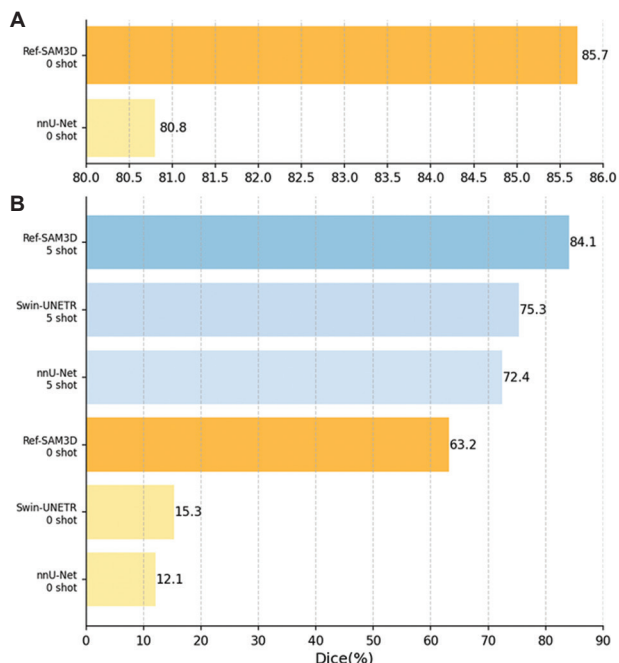


Figure 6. Comparison of zero-shot and five-shot generalization performance of Ref-SAM3D, nnU-Net, and Swin-UNETR on AMOS22 data. (A) Computed tomography (CT) and (B) magnetic resonance imaging (MRI) data.

Abbreviations: 3D: Three-dimensional; nn: No new; NSD: Normalized surface Dice; SAM: Segment Anything Model; UNETR: U-Net Transformers; UX-Net: UNet-eXpanded Network.

Ref-SAM3D’s robust generalization capabilities and its potential as a versatile solution for medical image segmentation across different imaging modalities.

These experimental findings clearly demonstrate Ref-SAM3D’s robust performance across different datasets and imaging modalities. The model’s strong zero-shot generalization capabilities and impressive few-shot learning results suggest its practical value in real-world medical applications, where adapting to diverse imaging conditions with minimal additional training is essential. These characteristics position Ref-SAM3D as a promising

Table 4. Ablation on each key component in our method

Parameters	Dice (%)	Hausdorff distance (%)
Ref-SAM3D	88.3	2.34
Without a text prompt	72.3	7.31
Without a cross-modal projector	80.1	4.22
Without hierarchical fusion	74.1	6.33

solution for clinical deployment, particularly in scenarios requiring flexible and efficient medical image analysis tools.

4.4. Ablation study

4.4.1. Effects of text prompt

The text prompt in our Ref-SAM3D model provided essential semantic guidance by bridging textual descriptions and visual features, enabling better interpretation of anatomical structures. The results, as shown in Table 4, without this component, the model’s performance dropped significantly, with dice score decreasing from 88.3% to 72.3% (−16.0%) and HD increasing from 2.34% to 7.31% (+4.97%). This substantial degradation demonstrates that the text prompt is crucial for leveraging linguistic context to achieve precise medical image segmentation.

4.4.2. Effects of cross-modal projector

The cross-modal projector in Ref-SAM3D plays a vital role in aligning textual and visual inputs, facilitating effective integration of multi-modal information for improved segmentation. By harmonizing these inputs, the projector enhanced the model’s ability to utilize semantic context from text alongside visual data. As shown in Table 4, removing this component resulted in an 8.2% decrease in dice score (from 88.3% to 80.1%) and an HD increase from 2.34% to 4.22%. These results confirm that when the cross-modal projector is removed, the model relies on unaligned embeddings, which can lead to less effective feature integration.

Table 5. The ablation experiments of each stage under the hierarchical cross-attention

Stages	Dice (%)	Hausdorff distance (%)
All stages	88.3	2.34
Stages 1 and 4	78.5	2.76
Stages 2 and 4	82.1	2.62
Stages 3 and 4	85.4	2.48
Stage 4 only	73.78	2.89

4.4.3. Effects of hierarchical cross-attention mechanism

The hierarchical fusion mechanism in Ref-SAM3D is pivotal for integrating information across various encoder layers, enabling the model to capture detailed, multi-level semantic features essential for precise segmentation. Ablation studies, summarized in Table 4, demonstrate the significance of this mechanism. Removing the hierarchical fusion led to a sharp decline in segmentation accuracy, with the dice coefficient dropping from 88.3% to 74.1%, and the HD increasing from 2.34% to 6.33%. This underscores the mechanism's role in effectively combining features across layers for better performance.

Moreover, Table 5 provides a systematic evaluation of each block level's contribution to the model. The results reveal that utilizing all layers (Stage 1–4) achieved the best performance, with a dice score of 88.3% and an HD of 2.34%. In contrast, excluding specific layers led to varied performance declines, with the shallow layers contributing significantly to contextual information and deeper layers enhancing fine-grained details. For example, when only deeper layers (Stages 3 and 4) were used, the dice score dropped to 78.5%, and the HD increased to 2.76%. In contrast, including only the shallow layers (Stages 1 and 2) yielded a dice score of 73.78% and an HD of 2.89%.

These findings underscore the necessity of a comprehensive fusion approach. Each layer's unique contributions—from the broad contextual cues in shallow layers to the detailed semantic information in deeper layers—work synergistically to enhance the model's ability to capture complex anatomical structures, ultimately improving overall segmentation accuracy and robustness.

5. Conclusion

We present Ref-SAM3D, a 3D-adapted SAM framework that synergizes cross-modal prompting and hierarchical attention to address medical segmentation challenges in volumetric imaging. Our model establishes a bidirectional interaction between visual data and semantic text descriptions, enabling intelligent segmentation through joint reasoning over volumetric imaging and clinical context. Three key innovations drive our methodology:

(i) A cross-modal reference prompt generator that fuses text and image embeddings into a unified feature space through adaptive alignment, significantly enhancing spatial-semantic correlation, (ii) a multi-scale hierarchical attention mechanism that dynamically prioritizes critical anatomical features across dimensional scales while suppressing irrelevant noise, significantly improving segmentation robustness in intricate 3D topologies, and (iii) a volumetric architecture adaptation that transforms SAM's native 2D processing into true 3D computation through depth-aware convolutions and recursive mask refinement, effectively bridging the dimensional gap in medical imaging analysis. Extensive validation demonstrates state-of-the-art performance on complex segmentation tasks. While our approach is highly effective, future work is needed to focus on improving computational efficiency to enable real-time clinical applications, exploring semi-supervised learning techniques to address the challenge of limited labeled data. Overall, our method holds significant promise as a generalizable and robust segmentation framework, offering both fully automatic and promptable segmentation capabilities for a wide range of 3D medical imaging applications.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Xiang Gao

Data curation: Xiang Gao

Investigation: Xiang Gao

Methodology: Xiang Gao

Visualization: Xiang Gao

Writing—original draft: Xiang Gao

Writing—review & editing: Kai Lu

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Data will be made available upon request to the corresponding author.

References

1. Obuchowicz R, Strzelecki M, Piorkowski A. Clinical applications of artificial intelligence in medical imaging and image processing-A review. *Cancers (Basel)*. 2024;16(10):1870.
doi: 10.3390/cancers16101870
2. Addimulam S, Mohammed MA, Karanam RK, et al. Deep learning-enhanced image segmentation for medical diagnostics. *Malays J Med Biol Res*. 2020;7(2):145-152.
3. Khalifa M, Albadawy M. AI in diagnostic imaging: Revolutionising accuracy and efficiency. In: *Computer Methods and Programs in Biomedicine Update*. Vol. 5; 2024.
4. Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2023. p. 4015-4026.
5. Zou X, Yang J, Zhang H, et al. Segment Everything Everywhere all at Once. *arXiv Preprint arXiv: 2304.06718*; 2023.
6. Huang Y, Yang X, Liu L, et al. Segment anything model for medical images? *Med Image Anal*. 2024;92:103061.
doi: 10.1016/j.media.2023.103061
7. Hu EJ, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. *arXiv preprint:2106.09685*, 2021.
8. Poth C, Sterz H, Paul I, et al. Adapters: A unified library for parameter-efficient and modular transfer learning. In: Feng Y, Lefever E, editors. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore*; 2023. p. 149-160.
9. Shen J, Wang W, Chen C, et al. Medtuning: A New Parameter-efficient Tuning Framework for Medical Volumetric Segmentation. *arXiv Preprint arXiv: 2304.10880*; 2024.
10. Zhang K, Liu D. Customized Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv: 2304.13785*; 2023.
11. Wang H, Guo S, Ye J, et al. Sam-med3d: Towards General-purpose Segmentation Models for Volumetric Medical Images. *arXiv preprint arXiv: 2310.15161*; 2024.
12. Wu J, Ji W, Liu Y, et al. Medical Sam Adapter: Adapting Segment Anything Model for Medical Image Segmentation. *arXiv preprint arXiv: 2304.12620*; 2023.
13. Gong S, Zhong Y, Ma W, et al. 3dsamadapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Med Image Anal*. 2024;98:103324.
14. Xie B, Tang H, Duan B, Cai D, Yan Y. MaskSam: Towards Auto-prompt Sam with Mask Classification for Medical Image Segmentation. *arXiv preprint arXiv: 2403.14103*; 2024.
15. Li C, Khanduri P, Qiang Y, Sultan RI, Chetty I, Zhu D. Autoprosam: Automated Prompting Sam for 3d Multi-Organ Segmentation. *arXiv preprint arXiv: 2308.14936*; 2024.
16. Zhang Y, Jiao R. Towards Segment Anything Model (sam) for Medical Image Segmentation: A Survey. *arXiv preprint arXiv: 2305.03678*; 2023.
17. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun*. 2024;15(1):654.
doi: 10.1038/s41467-024-44824-z.
18. Shaharabany T, Dahan A, Giryas R, Wolf L. Autosam: Adapting Sam to Medical Images by Overloading the Prompt Encoder. *arXiv preprint arXiv: 2306.06370*; 2023.
19. Na S, Guo Y, Jiang F, Ma H, Huang J. Segment any Cell: A Sam-Based Auto-Prompting Finetuning Framework for Nuclei Segmentation. *arXiv preprint arXiv: 2401.13220*; 2024.
20. Min B, Ross H, Sulem E, et al. Recent advances in natural language processing via large pre-trained language models: a survey. *ACM Comput Surv*. 2024;57(1):1-45.
doi: 10.1145/3605943
21. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*;2021.
22. Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*; 2021.
23. Zou X, Yang J, Zhang H, et al. Segment everything everywhere all at once. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA*; 2023.
24. Wang X, Zhang X, Cao Y, Wang W, Shen C, Huang T. Seggpt: Segmenting Everything in Context. *arXiv Preprint arXiv: 2304.03284*; 2023.
25. Oquab M, Darcet T, Moutakanni T. Dinov2: Learning Robust Visual Features without Supervision. *arXiv Preprint arXiv: 2304.07193*; 2024.
26. Wang Y, Zhou W, Mao Y, Li H. Detect any Shadow: Segment Anything for Video Shadow Detection. *arXiv preprint arXiv: 2305.16698*; 2023.
27. Deng R, Cui C, Liu Q, et al. Segment Anything Model (sam) for Digital Pathology: Assess Zero-Shot Segmentation on Whole Slide Imaging. *arXiv preprint arXiv: 2304.04155*; 2023.
28. He S, Bao R, Li J, et al. Accuracy of Segmentanything Model (sam) in Medical Image Segmentation Tasks. *arXiv preprint arXiv: 2304.09324*; 2023.
29. Hu C, Li X. When Sam Meets Medical Images: An Investigation of Segment Anything Model (Sam) on Multi-Phase Liver Tumor Segmentation. *arXiv preprint arXiv: 2304.08506*; 2023.
30. Zhou T, Zhang Y, Zhou Y, Wu Y, Gong C. Can Sam Segment Polyps? *arXiv preprint arXiv: 2304.07583*; 2023.

31. Cheng J, Ye Y, Deng Z, et al. *Sam-med2d*. *arXiv preprint arXiv: 2308.116184*; 2023.
32. Lei W, Wei X, Zhang X, Li K, Zhang S. *Medlsam: Localize and Segment Anything Model for 3D CT Images*. *arXiv preprint arXiv: 2306.14752*; 2024.
33. Yang Y, Wu X, He T, Zhao H, Liu X. *Sam3d: Segment Anything in 3D Scenes*. In: *International Conference on Computer Vision*; 2023.
34. Chen C, Miao J, Wu D, et al. *Ma-sam: Modality-agnostic sam adaptation for 3D medical image segmentation*. *Med Image Anal.* 2024;98:103310.
35. Pan J, Lin Z, Zhu X, Shao J, Li H. *St-adapter: Parameter-Efficient Image-to-Video Transfer Learning*. *arXiv preprint arXiv: 2206.13559*; 2022.
36. Muksimova S, Umirzakova S, Baltayev J, Cho YI. *RI-cervix-net: A hybrid lightweight model integrating reinforcement learning for cervical cell classification*. *Diagnostics (Basel)*. 2025;15(3):364.
37. Jia M, Tang L, Chen BC, et al. *Visual Prompt Tuning*. *arXiv Preprint arXiv: 2203.12119*; 2022.
38. Radford A, Kim JW, Hallacy C, et al. *Learning transferable visual models from natural language supervision*. In: *International Conference on Machine Learning*. PMLR; 2021. p. 8748-8763.
39. Jia C, Yang Y, Xia Y, et al. *Scaling up visual and vision-language representation learning with noisy text supervision*. In: *International Conference on Machine Learning*. PMLR; 2021, pp. 4904–4916.
40. Dosovitskiy A. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. *arXiv preprint arXiv:2010.11929*; 2020.
41. Ding H, Liu C, Wang S, Jiang X. *Vision-language transformer and query generation for referring segmentation*. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*; 2021. p. 16321-16330.
42. Li Y, Zhang J, Teng X, Lan L, Liu X. *Refsam: Efficiently Adapting Segmenting Anything Model for Referring Video Object Segmentation*. *arXiv Preprint arXiv: 2307.00997*; 2024.
43. Heller N, Isensee F, Trofimova D. *The kits21 Challenge: Automatic Segmentation of Kidneys, Renal Tumors, and Renal Cysts in Corticomedullary-phase ct*. *arXiv Preprint arXiv: 2307.01984*; 2023.
44. Bilic P, Christ P, Li HB, et al. *The liver tumor segmentation benchmark (LiTS)*. *Med Image Anal.* 2023;84:102680. doi: 10.1016/j.media.2022.102680
45. Antonelli M, Reinke A, Bakas S, et al. *The medical segmentation decathlon*. *Nat Commun.* 2022;13(1):4128. doi: 10.1038/s41467-022-30695-9
46. Zhuang X, Li L, Payer C. *Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge*. *Med Image Anal.* 2019;58:101537. doi: 10.1016/j.media.2019.101537
47. Landman B, Xu Z, Iglesias J, Styner M, Langerak T, Klein A. *Miccai multi-atlas labeling beyond the cranial vault-workshop and challenge*. Vol. 5. In: *Proceeding MICCAI Multi-Atlas Labeling Beyond Cranial Vault-Workshop Challenge*; 2015. p. 12.
48. Tang Y, Yang D, Li W, et al. *Self-supervised Pre-training of Swin Transformers for 3d Medical Image Analysis*. *arXiv preprint arXiv:2111.14791*; 2022.
49. Ji Y, Bai H, Yang J, et al. *Amos: A Large-scale Abdominal Multiorgan Benchmark for Versatile Medical Image Segmentation*. *arXiv preprint arXiv:2206.08023*; 2022.
50. Isensee F, Petersen J, Klein A, et al. *nnU-net: Self-Adapting Framework for u-net-Based Medical Image Segmentation*. *arXiv preprint arXiv: 1809.10486*; 2018.
51. Ronneberger O, Fischer P, Brox T. *U-net: Convolutional networks for biomedical image segmentation*. In: Wells WM 3rd, Frangi AF, editors. *Medical Image Computing and Computer-Assisted Intervention - MICCAI, Nassir Navab, Joachim Hornegger*; 2015.
52. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth H, Xu D. *Swin unetr: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images*. *arXiv preprint arXiv: 2201.01266*; 2022.
53. Zhou HY, Guo J, Zhang Y, et al. *nnformer: Volumetric medical image segmentation via a 3D transformer*. *IEEE Trans Image Process.* 2023;32:4036-4045. doi: 10.1109/TIP.2023.3293771
54. Shaker A, Maaz M, Rasheed H, et al. *Unetr++: Delving into Efficient and Accurate 3D Medical Image Segmentation*. *arXiv Preprint arXiv: 2212.04497*; 2024.
55. Lee HH, Bao S, Huo Y, Landman BA. *3D ux-net: A Large Kernel Volumetric Convnet Modernizing Hierarchical Transformer for Medical Image Segmentation*. *arXiv Preprint arXiv: 2209.15076*; 2023.

ORIGINAL RESEARCH ARTICLE

Leveraging the smarts in your phone: An artificial intelligence-driven iOS application for neurosurgical navigation of external ventricular drains

Andrew Abumoussa¹, Benjamin Succop^{2*}, Carolyn Quinsey³, Yueh Lee⁴, and Sivakumar Jaikumar¹¹Department of Neurosurgery, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America²Department of Neurosurgery, School of Medicine, Duke University, Durham, North Carolina, United States of America³Department of Neurosurgery, School of Medicine, University of Missouri, Columbia, Missouri, United States of America⁴Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, United States of America(This article belongs to the *Special Issue: Sensors and circuits for AI in health*)***Corresponding author:**Benjamin Succop
(ben.succop@duke.edu)**Citation:** Abumoussa A, Succop B, Quinsey C, Lee Y, Jaikumar S. Leveraging the smarts in your phone: An artificial intelligence-driven iOS application for neurosurgical navigation of external ventricular drains. *Artif Intell Health*. 2025;2(4):129-138. doi: 10.36922/aih.8195**Received:** December 25, 2024**1st revised:** June 10, 2025**2nd revised:** July 27, 2025**Accepted:** August 13, 2025**Published online:** September 23, 2025**Copyright:** © 2025 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**

External ventricular drain (EVD) placement is a critical neurosurgical procedure traditionally performed freehand, with inherent risks of malposition, infection, and hemorrhage. Recent advances in artificial intelligence (AI), particularly in medical imaging and real-time computer vision, have enabled the development of portable navigation tools that may enhance accuracy, safety, and bedside accessibility. This study evaluated whether iOS devices equipped with a TrueDepth camera could perform real-time object and facial recognition, tracking, and semantic segmentation of computed tomography (CT) scans for non-immobilized heads to guide EVD placement via a custom AI-driven application. A custom iOS application was developed to provide a complete, real-time surgical navigation experience on an iPhone or iPad Pro. Three AI models were trained, tuned, and validated: a semantic segmentation model for brain anatomy, a semantic segmentation model for facial features, and an object detection model for a custom EVD stylet attachment. GPU programming accelerated on-device real-time, continuous registration while optimizing power consumption. A UNet convolutional neural network trained on eight 1 mm head CTs achieved 98.3% testing and 98.2% validation accuracy using a 50/50 test-validation split, segmenting a thin-cut CT in 3 s on an iPhone 12 Pro. Point cloud merging of patient anatomy took 4 seconds with an initial depth scan of 30,000 points, updating in real time with a cumulative error of 1×10^{-8} cm. Transfer learning-powered EVD tracking, trained for 1,000 epochs, achieved an intersection over union of 1.0 and 0.98 for the detection model, with inference times of 800 μ s on Apple's Neural Engine. This feasibility study demonstrates that iOS devices with TrueDepth cameras can enable real-time, continuous surgical navigation for EVD stylets.

Keywords: External ventricular drain; Surgical navigation; Artificial intelligence; Machine learning

1. Introduction

Surgical navigation has become a crucial tool in neurosurgery, enabling accurate localization and targeting of lesions within the brain and spine to improve surgical precision and patient outcomes.¹⁻³ While traditional navigation methods relied on intraoperative imaging, computer-assisted navigation systems have become increasingly common and popular.⁴⁻⁷ These systems, however, typically require costly, proprietary computing to run navigation software and are often bulky and cumbersome.

Innovations in computation, particularly artificial intelligence (AI), have paved the way for the development of higher-accuracy, lower-cost navigation techniques. Convolutional neural networks (CNNs), in particular, have revolutionized image recognition across industries, including neurosurgery.⁸⁻¹⁰ CNNs are neural network-based machine learning, also known as deep learning models, that are now the standard for computer vision, particularly for identifying and recognizing objects or features via pixel analysis.¹¹⁻¹³ In neurosurgery, CNN applications include, but are not limited to, automatic segmentation of vertebral bodies and intervertebral discs in magnetic resonance imaging (MRI)¹⁴⁻¹⁷ and computed tomography (CT),¹⁸ measurement of Cobb angles from X-rays,^{19,20} diagnosis of vertebral fractures,²¹ enhanced diagnosis and classification of brain tumors,^{22,23} and intraoperative co-registration of two-dimensional and three-dimensional (3D) imaging.²⁴ AI-based navigation software powered by U-Net can improve efficiency and performance, potentially eliminating the need for expensive, cumbersome hardware in the operating room or at the bedside.

More recently, augmented reality (AR) has emerged as a valuable tool for facilitating surgical care in the operating room and beyond. AR merges real-world images with virtual objects generated by computer graphics in real time.²⁵ In the operating room, AR has found applications in visualizing tumors and surrounding anatomical structures in a number of oncological settings to facilitate safe resection.²⁶⁻³⁴ In neurosurgery, AR applications include guiding pedicle screw placement in spine surgery and visualizing anatomy during minimally invasive surgery.³⁵⁻³⁸ Outside the operating room, AR has proven effective in surgical and medical education, including virtual surgical training.^{39,40}

Bedside neurosurgical procedures represent a setting where such navigational innovations could be particularly beneficial, as limited space and urgent or emergent scenarios often preclude the use of traditional operating room stereotactic setups. One such urgent or emergent, historically blind-placement procedure is external ventricular drain (EVD) insertion. An EVD is a temporary catheter system inserted into the cerebral ventricles to

divert cerebrospinal fluid (CSF) for both therapeutic and diagnostic purposes. It is most commonly used to manage elevated intracranial pressure in conditions such as subarachnoid hemorrhage, traumatic brain injury, and hydrocephalus.⁴¹ EVDs allow continuous monitoring of intracranial pressure and facilitate sampling of CSF to guide treatment of structural obstructions, hemorrhage-related complications, and other conditions.⁴²

The placement procedures involve incising the scalp and drilling a small burr hole, typically at Kocher's point—approximately 10–11 cm posterior to the nasion and 2–3 cm lateral to the midline—followed by careful advancement of a catheter 6–7 cm until CSF return is noted. This location is presumed to be in the ipsilateral lateral ventricle near the foramen of Monro, which connects the lateral ventricles to the inferior third ventricle.⁴³ The catheter is then tunneled subcutaneously for at least 6 cm to reduce infection risk, secured with stitches or staples, and connected to an external drainage and intracranial pressure monitoring system. Most institutions confirm placement with a post-operative brain CT scan⁴³ (Figure 1). Despite its ubiquity in neurocritical care, EVD placement carries notable risks, including hemorrhage, infection, and malposition. Misplacement of EVDs is associated with higher infection and hemorrhage rates, the need for revision procedures, increased healthcare costs, and prolonged hospital stay (PMID: 36434852).⁴¹

EVD placement has the potential to be optimized by AI-based neuronavigation. AI navigation tools have already been integrated into other specialties for bedside procedures, such as ultrasound-guided vascular access and

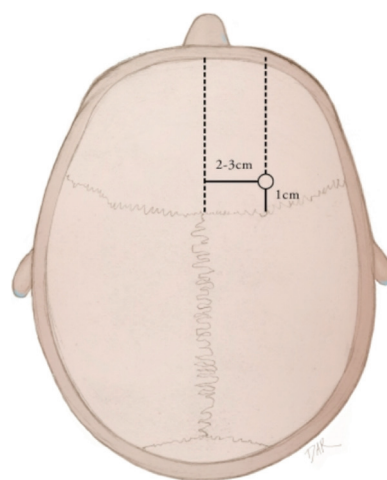


Figure 1. Landmarks for external ventricular drain placement. The classic trajectory is Kocher's point, 11–12 cm behind the nasion on the mid-pupillary line, 1 cm anterior to the palpated coronal suture, with a trajectory orthogonal to the ipsilateral medial canthus and ipsilateral tragus.

portable diagnostic imaging.⁴⁴ The iOS-based platform presented in the present study could similarly serve as a foundational tool in resident education by providing real-time, immersive feedback on trajectory and depth during catheter placement. Coupled with the low hardware footprint of smartphones, it allows for seamless integration of safety and navigation features without imposing a significant burden on workflow or cost.

The cognitive load and decision-making complexity inherent in high-stakes procedures such as EVD placement are often underestimated, particularly for junior providers, trainees, or advanced practitioners who perform the procedure infrequently. AI-based systems can alleviate some of this procedural burden by offering real-time alignment cues, trajectory verification, and visual reinforcement through augmented overlays. This human-machine collaboration reduces reliance on rote memorization or abstract spatial reasoning, thereby lowering error rates, especially during periods of fatigue, night shifts, or acute crisis scenarios.

The present study seeks to investigate whether a custom-designed AI application for mobile devices, specifically an iOS device equipped with a TrueDepth camera, can provide instantaneous navigation by identifying and tracking an EVD stylet in real time, with potential future application as a bedside navigation tool. We developed an iOS application leveraging the optimized computational hardware of Apple devices and performed simulated navigated procedures on specific models (iPhone 12 Pro, 13 Pro, 14 Pro, and M1 and M2 iPad Pro). We evaluated whether these devices could meet the computational requirements for computer-assisted navigation, the resolution and accuracy they could achieve, and the technical feasibility of performing these procedures on battery-powered devices. Accuracy was then compared to that of a traditional navigation system. We hypothesize that our custom application will provide real-time, accurate surgical navigation on an iPhone, encouraging further exploration of its use in EVD placement and other cranial neurosurgical procedures both at the bedside and in the operating room. The ultimate goal of this investigation is to integrate existing technologies in registration and object tracking into a single custom application capable of performing EVD navigation on an iOS device at the bedside, thereby enabling timely neurosurgical navigation without requiring a complex setup that delays urgent or emergent patient care.

2. Data and methods

2.1. Application design and development

The present study involved the development of an iOS application capable of performing iOS-assisted

neurosurgical navigation. To the best of our knowledge, no free and reliable mobile neuronavigation system currently exists that can provide real-time neuronavigation in emergency settings. The innovation of this work lies in producing an iOS application that enables instantaneous patient registration on standard mobile devices (iPhone 12, iPhone 13, or iPad Pro), offering a free neuronavigation platform to assist clinicians with the placement of EVDs without requiring stereotactic immobilization, reference arrays, or fiducials. To evaluate the features available on iOS-powered devices, anonymized patient data were obtained from an open-source repository.⁴⁵

The initial step was to identify the essential components of a computer-assisted navigation procedure. These included: (i) Processing of pre-procedural scans, (ii) real-time detection and tracking of the patient, (iii) object detection and localization of surgical instruments, and (iv) the ability to map both patient anatomy and the surgical device to imaging data (Figure 2). The overarching goal was to achieve real-time, continuous registration with minimal surgeon input. Accordingly, the user interface was designed to reduce manual interaction, creating a seamless experience.⁴⁶ The application integrates multiple programming environments: Python (3.10.10, Python Software Foundation, USA) and TensorFlow (2.12.0, Google, USA) for machine learning models, C++ (17.0.0, Apple, USA) and Metal (Metal 3, Apple, USA) for performance optimization, and Swift (5.9.2, Apple, USA) for the iOS application framework. These were unified using the Xcode Integrated Development Environment (15.4.0, Apple, USA) to build and test the app.

2.2. iOS true depth camera

The iOS TrueDepth camera, typically used for the Face ID feature, uses light detection and ranging (LiDAR, a remote sensing method) to capture accurate topographic data. It projects and analyzes thousands of laser points, measuring their reflection time to create a depth map, which is then coupled with an infrared image. These images are processed by Apple's Neural Engine (compatible chips include A11, A12 Bionic, A12X Bionic, A13 Bionic, A14 Bionic, and A15 Bionic) and compared to the enrolled representation.⁴⁷ Although Apple does not report the depth accuracy of the iOS True Depth camera, independent sources estimate it to be approximately 2% at a distance of 3 m.⁴⁸

2.3. AI model creation and training

Two models were developed for the critical steps of surgical navigation:

- (i). A semantic segmentation model for head CT scans.
- (ii). An object detection model to track EVD catheters.

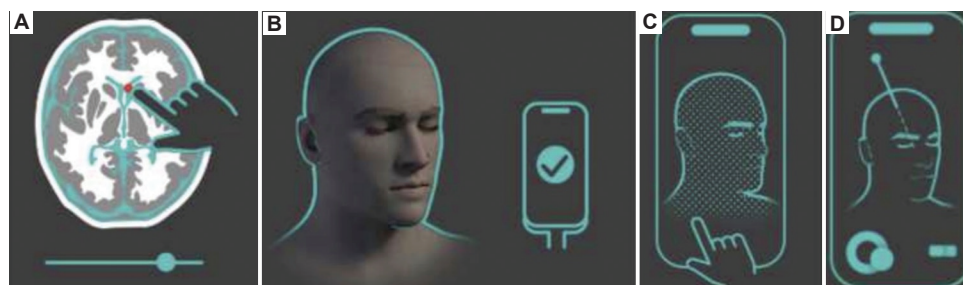


Figure 2. Schematic representation of neuronavigation using an iOS mobile device. (A) Anatomic review and target selection. (B) An iOS device mounted with the patient's head in view. (C) Evaluation of registration parameters. (D) Procedural interface for alignment and depth guidance.

Industrial-strength models, optimized for iOS systems, were rapidly trained using transfer learning. For the head CT segmentation model, the open-source QURE AI CQ500 dataset was used, consisting of 491 head CTs acquired on multiple scanners.⁴⁵ Eight scans were manually annotated to segment key anatomical structures relevant to trauma surgery: background, extracranial soft tissue, skull, brain, and ventricles. These annotations yielded 4,096 CT slices, which were used for model training, validation, and testing, allowing for the identification of each component within an individual CT scan. Our model combines these segmentations with the point cloud generated by iOS devices, fusing point cloud segments with the pre-operative CT scan using GPU acceleration. This fusion process is performed iteratively and then creates the optimal alignment, continuously registering and re-registering during the procedure, as the patient's head is not rigidly secured. The system operates without user input, ensuring instantaneous, continuous neuronavigation and allowing accurate overlay of the EVD catheter location without requiring head fixation or a reference array.

For the EVD object detection model, 937 unique images of EVDs with a visually distinct dodecahedron attachment were acquired in an ICU setting using the front-facing camera of an iOS device to simulate bedside placement. The dodecahedron, attached to the distal end of the EVD, contained a unique QR code on each face to facilitate identification. These images were segmented and used to train a feature-based machine learning algorithm to localize the dodecahedron in space through the iOS device's video feed. The model was externally validated with 200 additional images obtained in non-ICU settings. In total, 700 randomly selected images were used for training and 237 for validation.

Two models were trained: a full network and a transfer-learning model using YOLOv2 (YOLOv2, Ultralytics, USA), both optimized for iOS systems to achieve real-time, accurate navigation of the EVD and its custom dodecahedron attachment.⁴⁹ To confirm robustness and guard against overfitting or class imbalance, we performed

a randomized 10-fold cross-validation with a 90% train/10% testing split.

2.4. Model performance

Segmentation models were assessed for accuracy in training and validation cohorts. In addition, we evaluated the initial time required for point cloud merging with the patient's anatomy and quantified cumulative error following scaling, alignment, and rotation. Finally, performance metrics included intersection over union (I/U), varied intersection over union (I/U), and inference times of the iOS application during real-time EVD tracking.

3. Results

We developed an application capable of performing all steps of surgical navigation on iOS devices. The application can load and display head CTs in DICOM or NIFTI format. The surgeon selects the target by touching the mobile device's screen, and the device stores that information for the 3D transformations necessary to perform the navigated procedure (Figure 3).

While the surgeon views the scan of interest, segmentation is performed in the background by a UNet CNN trained on the eight 1 mm head CTs. This model achieved 98.3% testing accuracy and 98.2% validation accuracy using a 50/50 test-validation split. To confirm that the model was not overfitting and remained robust against class imbalance, we performed randomized 10-fold cross-validation with a 90% train/10% testing split, yielding an average validation accuracy of 98.3% across folds. Segmentation requires 30 ms per slice on a standard iPhone 12, or approximately 3 s per scan, and provides the data for surface merging (Figure 4).

The surgeon then mounts the phone in front of the patient's head and navigates to the next screen. The video feed semantically segments the largest head in view and captures 3D data from the TrueDepth camera. The application allows the surgeon to inspect the TrueDepth image in 3D to ensure scan adequacy before merging.

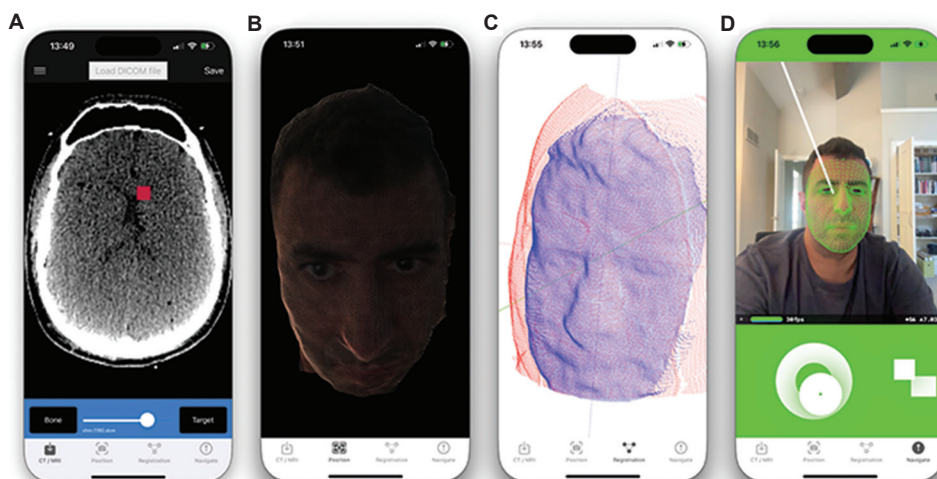


Figure 3. iOS application screenshots during use. (A) DICOM viewer for targeting. (B) Point cloud obtained following device mount. (C) Registration review to inspect the point-cloud merge. (D) Augmented reality-driven navigation interface with alignment and depth guidance.

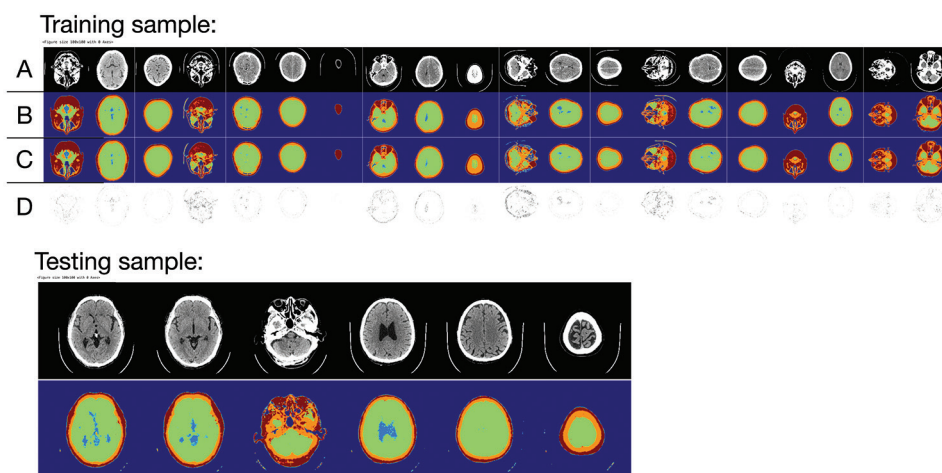


Figure 4. Results of the semantic segmentation model (SSM). Training samples provide four rows of information: (A) original scan, (B) predicted segmentation, (C) ground-truth segmentation, and (D) error map. Testing samples demonstrate performance on previously unseen data. The SSM achieved an accuracy of 98.3% for testing and 98.2% for validation when segmenting background (purple), extracranial soft tissue (red), bone (orange), neural tissue (green), and ventricles (blue).

Once accepted, the phone performs a point-cloud merge by aligning the segmented head CT with the 3D TrueDepth scan. The registration algorithm applies scaling, alignment, and rotation to achieve a coded threshold of 1×10^{-8} cm average difference between the two-point clouds (Figure 2). The initial merge requires an average of 3.8 s, after which updates are performed at 60 merges per second in the background, synchronized with the 60-fps display of the navigated screen.

The final navigated display provides the surgeon with an AR view of the patient, a projection of the target trajectory, and an alignment interface for navigating the specialized EVD stylet (Figure 5). Training of the tracking model for 1,000 epochs resulted in an I/U of 1.0 and a varied I/U of

0.98 for the YOLOv2 model, with inference times of 800 μ s on Apple’s Neural Engine.

4. Discussion

The performance of our AI algorithms, combined with the successful implementation of a functioning application running these models on local hardware, suggests that iOS devices can feasibly provide a complete neurosurgical navigation experience. This innovation has the potential to significantly improve the accessibility, efficiency, and cost-effectiveness of surgical navigation, particularly in resource-limited settings. For example, it could bring navigation directly to the bedside, enhancing accuracy in procedures such as EVD placements, which currently carry error rates of up to 25% with the standard blind, landmark-based

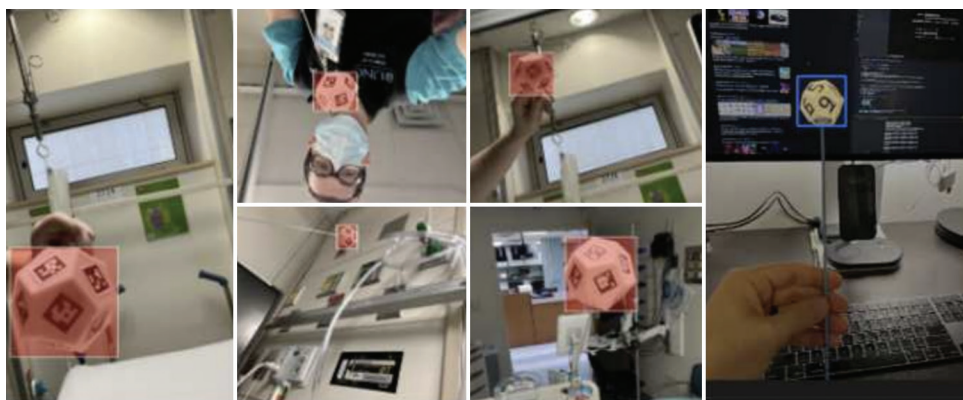


Figure 5. Example of a dodecahedron attached to an external ventricular drain stylet. Sample training data/validation data are shown with red segmentations; testing data are shown with blue bounding boxes. The model successfully tracked and localized the different markings on the dodecahedron.

technique.^{41,50} While experienced providers may not require such assistance in straightforward cases, AI-based navigation could improve safety and efficacy for trainees and in patients with challenging anatomy. By improving first-pass accuracy, AI-assisted systems like ours have the potential to significantly reduce downstream complications and associated costs.

Equally important is the role of AI in medical education. Systems equipped with explainable AI features can serve not only as navigational tools but also as digital mentors—offering real-time procedural feedback, recording attempts for later analysis, and integrating with curricula to track skill acquisition longitudinally. This capacity aligns with emerging research on competency-based training frameworks that leverage AI for both assessment and remediation.

Looking ahead, our findings support the argument for more democratized, hardware-agnostic AI integration into surgical care. Unlike legacy stereotactic systems that cost hundreds of thousands of dollars, require sterilized hardware, and demand specialized personnel, our mobile solution operates on a standard smartphone at no additional cost once deployed. This approach not only reduces barriers to implementation in large hospital systems but may also revolutionize emergency neurosurgical interventions in remote or battlefield environments.

A major design goal of this project was to maximize ease of use, accuracy, and portability. By leveraging commercially available iOS hardware, the system enables real-time, continuous patient registration and a full navigation experience without the need for complex, expensive, or proprietary hardware. The integration of AI-based navigational tools into bedside workflows therefore supports not only improved accuracy but also enhanced procedural safety. For instance, misdirected

EVD placement can cause parenchymal hemorrhage, intracranial hypertension from occlusion, or delayed CSF drainage—all potentially preventable with better visualization.

Moreover, once adopted, this technology could also serve as a real-time procedural documentation tool. By capturing trajectory data, timestamps, and alignment metrics, the application could offer medico-legal protection for providers and support quality improvement initiatives. It could further contribute to a growing body of procedural analytics that may be mined for insights into improving technique, developing personalized risk profiles, and enabling population-level outcome modeling through federated learning frameworks. In the future, real-time procedural metrics could be incorporated into credentialing, maintenance of certification, and residency milestone assessments.

Although current navigation systems enhance surgical safety, they also have significant limitations. These systems are large and bulky, occupying valuable space in operating or procedural rooms. They often rely on rigidly attached reference arrays that are registered only at the beginning of a case, making them prone to errors if anatomical shifts occur or if the patient's position changes relative to the reference array.⁴ In addition, traditional navigation systems require substantial user interaction, which can potentially introduce operator error and inconsistencies during critical steps.⁶ In contrast, the automated and standardized nature of this study's custom iOS application minimizes user-dependent variability and enables more consistent, accurate navigation through continuously updated registration of non-immobilized subjects. The iOS application also reduces the need for additional personnel or large, costly equipment, making navigation feasible in bedside and space-constrained settings where it was previously impractical or impossible.

It is important to acknowledge, however, that not all practicing neurosurgeons will adopt or require this technology in their workflow. Surgeons with extensive clinical experience, including hundreds of EVD placements, may find little practical benefit in an additional technological layer. Indeed, the tactile and anatomical intuition developed over decades cannot be easily replaced or replicated by software. This application is therefore not intended to supplant surgical judgment; rather, it is designed to augment procedural safety and education, particularly for trainees, early-career providers, and those who infrequently perform EVD placements.

Trainees often face steep learning curves in ventricular catheterization, a task further complicated by anatomical variability, patient movement, and the urgency of emergent settings. The traditional apprenticeship model, while time-tested, provides variable exposure and feedback. By delivering real-time visual guidance, trajectory alignment, and error detection, AI-based navigation can shorten the time needed to achieve competence, reduce patient risk, and improve trainee confidence. Recent research supports this potential, showing that access to simulation and navigational feedback correlates with faster skill acquisition and lower complication rates in neurosurgical training programs.⁵¹

Registration remains the key step for mapping a patient's physical anatomy to preprocedural imaging. Conventional systems often rely on as few as 10 points to generate a rigid transformation between real-world and radiographic coordinates.⁶ In contrast, our proprietary system leverages the full 30,000 points provided by Apple's TrueDepth camera, corrects for device-specific intrinsic properties, and performs transformations unique to each video frame. This allows anatomy to move relative to the camera while maintaining continuous registration. These operations are accelerated by the on-device GPU,⁵² minimizing computational burden on the mobile device. Further study is warranted to quantify the impact on iOS battery performance.

Overall, the use of transfer learning allowed us to leverage pre-trained models developed on large datasets to train our models on relatively small datasets, resulting in high accuracy and robustness. For the head CT segmentation model, we leveraged the well-studied open-source QURE. AI CQ500 dataset, consisting of 491 head CTs obtained on multiple scanners, which improved the generalizability of our model compared to training on institutional CTs alone.⁴⁵ The data collection and annotation processes for both models were time-consuming and required expert knowledge. These models further provide opportunities for further iteration and the incorporation of richer datasets, such as MRI, or for segmentation of additional anatomic

structures and pathologies (e.g., tumors, vasculature, cranial nerves). With Apple's CoreML architecture, the models are efficiently accelerated, facilitating rapid inferences that drive the app.

The use of AR in neurosurgical navigation offers a more intuitive and immersive experience for the surgeon while minimizing the fatigue often associated with virtual reality solutions.³ The iOS application utilizes AR to provide real-time visualization of surgical trajectories, anatomic boundaries, and feedback regarding the system's accuracy. These cues may help provide surgeons with an "X-ray vision–" like understanding of patient-specific anatomy, thereby increasing confidence during procedures.

While the present study focused on the feasibility of using an iOS application for navigated EVD placements, there are many potential future applications for this technology. For example, the iOS application could be used for remote surgical guidance, allowing a surgeon in one location to guide another surgeon through a procedure using shared visualization. In addition, it could serve as a training and educational tool, providing a realistic and immersive simulation environment for trainees to practice neurosurgical navigation or to explore anatomy through an alternative medium. The next step will be to compare the application's navigational accuracy against the gold standard of head CT for EVD placement in cadaveric models.

There are several limitations to this study that should be acknowledged. First, this is a proof-of-concept feasibility study. While our model successfully demonstrated that an iOS application can track an EVD in real time, its effect on EVD placement accuracy and clinical outcomes remains unknown. Given the feasibility design, only a limited set of evaluation metrics and parameters were analyzed, which will be expanded in future studies. The next step will involve cadaveric testing to rigorously evaluate the accuracy and safety of the iOS application in a controlled setting. In addition, the small sample size limits the generalizability of findings, underscoring the need for larger studies. Furthermore, the application's overall clinical utility in navigating EVD placement cannot be determined until it is directly compared with the gold standard of head CT, which will be the subject of subsequent research.

Nonetheless, the present study demonstrates the potential of iOS devices to improve neurosurgical navigation, particularly for trainees and inexperienced providers, and establishes a foundation for future research in this area.

5. Conclusion

The goal of this investigation was to integrate existing technologies in registration and object tracking into a

single custom application that can perform EVD navigation on an iOS device at the bedside. This approach facilitates navigation by neurosurgical providers without requiring complex setups that delay urgent or emergent patient care. Our data demonstrate that such an endeavor is feasible, with the custom iOS application achieving high accuracy and near-instantaneous results.

The development of a handheld iOS application for neurosurgical navigation represents a promising advancement in the field. Importantly, its greatest value is likely not for seasoned neurosurgeons who routinely perform EVD placements, but for those with less frequent exposure—such as residents, junior faculty, or providers who take call infrequently. By offering real-time, AR-based guidance and objective trajectory verification, the application holds significant potential as both a safety tool and an educational aid.

Although not every surgeon will find it necessary to adopt this system, its role in surgical education and simulation could support the development of procedural competence among lower-volume practitioners. Ultimately, this may enhance patient safety, reduce complications, and lower costs without adding logistical burdens or delaying emergent care. While further research is needed to evaluate its performance against gold standards such as CT in cadaveric and clinical studies, early data suggest that this technology may improve the accessibility and cost-effectiveness of neurosurgical navigation.

Acknowledgments

We would like to thank Xian Boles for assistance with figure illustration and 3D modeling.

Funding

This work was graciously funded by UNC Health's Innovation Pilot Grant (Grant no.: 29201).

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Andrew Abumoussa, Sivakumar Jaikumar, Carolyn Quinsey

Formal analysis: Andrew Abumoussa, Benjamin Succop, Yueh Lee

Investigation: Andrew Abumoussa, Benjamin Succop

Methodology: Andrew Abumoussa, Benjamin Succop, Yueh Lee

Software: Andrew Abumoussa

Writing—original draft: Benjamin Succop

Writing—review & editing: Andrew Abumoussa, Carolyn Quinsey, Sivakumar Jaikumar

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. Grunert P, Darabi K, Espinosa J, Filippi R. Computer-aided navigation in neurosurgery. *Neurosurg Rev.* 2003;26:73-99; discussion 100-1.
doi: 10.1007/s10143-003-0262-0
2. Mezger U, Jendrewski C, Bartels M. Navigation in surgery. *Langenbecks Arch Surg.* 2013;398:501-514.
doi: 10.1007/s00423-013-1059-4
3. Tagaytayan R, Kelemen A, Sik-Lanyi C. Augmented reality in neurosurgery. *Arch Med Sci.* 2018;14(3):572-578.
doi: 10.5114/aoms.2016.58690
4. Rahmathulla G, Nottmeier EW, Pirris SM, Deen HG, Pichelmann MA. Intraoperative image-guided spinal navigation: Technical pitfalls and their avoidance. *Neurosurg Focus.* 2014;36(3):E3.
doi: 10.3171/2014.1.FOCUS13516
5. Moiraghi A, Pallud J. Intraoperative ultrasound techniques for cerebral gliomas resection: Usefulness and pitfalls. *Ann Transl Med.* 2020;8(8):523.
doi: 10.21037/atm.2020.03.178
6. Khoshnevisan A, Allahabadi NS. Neuronavigation: Principles, clinical applications and potential pitfalls. *Iran J Psychiatry.* 2012;7(2):97-103.
7. Harwick E, Singhal I, Conway B, Mueller W, Treffy R, Krucoff MO. Pinless electromagnetic neuronavigation during awake craniotomies: Technical pearls, pitfalls, and nuances. *World Neurosurg.* 2023;175:e159-e166.
doi: 10.1016/j.wneu.2023.03.045
8. Anwar SM, Majid M, Qayyum A, Awais M, Alnowami M, Khan MK. Medical image analysis using convolutional neural networks: A review. *J Med Syst.* 2018;42:226.
doi: 10.1007/s10916-018-1088-1
9. Ronneberger O, Fischer P, Brox T. *U-net: Convolutional Networks for Biomedical Image Segmentation.* Berlin: Springer; 2015. p. 234-241.

10. Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computat.* 2017;29(9):2352-2449.
11. Gu J, Wang Z, Kuen J, *et al.* Recent advances in convolutional neural networks. *Pattern Recogn.* 2018;77:354-377.
12. Albawi S, Mohammed TA, Al-Zawi S. *Understanding of a Convolutional Neural Network.* New York: IEEE; 2017. p. 1-6.
13. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging.* 2018;9:611-629.
doi: 10.1007/s13244-018-0639-9
14. Huang J, Shen H, Wu J, *et al.* Spine Explorer: A deep learning based fully automated program for efficient and reliable quantifications of the vertebrae and discs on sagittal lumbar spine MR images. *Spine J.* 2020;20(4):590-599.
doi: 10.1016/j.spinee.2019.11.010
15. Lehnen NC, Haase R, Faber J, *et al.* Detection of degenerative changes on MR images of the lumbar spine with a convolutional neural network: A feasibility study. *Diagnostics (Basel).* 2021;11(5):902.
doi: 10.3390/diagnostics11050902
16. Cheng YK, Lin CL, Huang YC, *et al.* Automatic segmentation of specific intervertebral discs through a two-stage multiresunet model. *J Clin Med.* 2021;10(20):4760.
doi: 10.3390/jcm10204760
17. Lessmann N, Van Ginneken B, De Jong PA, Išgum I. Iterative fully convolutional neural networks for automatic vertebra segmentation and identification. *Med Image Anal.* 2019;53:142-155.
doi: 10.1016/j.media.2019.02.005
18. Janssens R, Zeng G, Zheng G. *Fully Automatic Segmentation of Lumbar Vertebrae from CT Images using Cascaded 3D Fully Convolutional Networks.* New York: IEEE; 2018. p. 893-897.
19. Caesarendra W, Rahmaniar W, Mathew J, Thien A. Automated Cobb angle measurement for adolescent idiopathic scoliosis using convolutional neural network. *Diagnostics (Basel).* 2022;12(2):396.
doi: 10.3390/diagnostics12020396
20. Sun Y, Xing Y, Zhao Z, Meng X, Xu G, Hai Y. Comparison of manual versus automated measurement of Cobb angle in idiopathic scoliosis based on a deep learning keypoint detection technology. *Eur Spine J.* 2022;31:1969-1978.
doi: 10.1007/s00586-021-07025-6
21. Small J, Osler P, Paul A, Kunst M. CT cervical spine fracture detection using a convolutional neural network. *AJNR Am J Neuroradiol.* 2021;42(7):1341-1347.
doi: 10.3174/ajnr.A7094
22. Seetha J, Raja SS. Brain tumor classification using convolutional neural networks. *Biomed Pharmacol J.* 2018;11(3):1457.
23. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans Med Imaging.* 2016;35(5):1240-1251.
24. Abumoussa A, Gopalakrishnan V, Succop B, *et al.* Machine learning for automated and real-time two-dimensional to three-dimensional registration of the spine using a single radiograph. *Neurosurg Focus.* 2023;54(6):E16.
doi: 10.3171/2023.3.FOCUS2345
25. Eckert M, Volmerg JS, Friedrich CM. Augmented reality in medicine: Systematic and bibliographic review. *JMIR Mhealth Uhealth.* 2019;7(4):e10967.
doi: 10.2196/10967
26. Brookes MJ, Chan CD, Baljer B, *et al.* Surgical advances in osteosarcoma. *Cancers (Basel).* 2021;13(3):388.
doi: 10.3390/cancers13030388
27. Kraeima J, Glas HH, Merema BBJ, Vissink A, Spijkervet FK, Witjes MJ. Three-dimensional virtual surgical planning in the oncologic treatment of the mandible. *Oral Dis.* 2021;27(1):14-20.
doi: 10.1111/odi.13631
28. Bobak MJ, Weber MW, Doellman MA, *et al.* Modafinil activates phasic dopamine signaling in dorsal and ventral striata. *J Pharmacol Exp Ther.* 2016;359(3):460-470.
doi: 10.1124/jpet.116.236000
29. Lavé A, Meling TR, Schaller K, Corniola MV. Augmented reality in intracranial meningioma surgery: Report of a case and systematic review. *J Neurosurg Sci.* 2020;64(4):369-376.
doi: 10.23736/S0390-5616.20.04945-0
30. Lee C, Wong GKC. Virtual reality and augmented reality in the management of intracranial tumors: A review. *J Clin Neurosci.* 2019;62:14-20.
doi: 10.1016/j.jocn.2018.12.036
31. Gerard IJ, Kersten-Oertel M, Petrecca K, Sirhan D, Hall JA, Collins DL. Brain shift in neuronavigation of brain tumors: A review. *Med Image Anal.* 2017;35:403-420.
doi: 10.1016/j.media.2016.08.007
32. Inoue D, Cho B, Mori M, *et al.* Preliminary study on the clinical application of augmented reality neuronavigation. *J Neurol Surg A Cent Eur Neurosurg.* 2013;74(2):71-76.
doi: 10.1055/s-0032-1333415
33. Tabrizi LB, Mahvash M. Augmented reality-guided neurosurgery: Accuracy and intraoperative application of an image projection technique. *J Neurosurg.* 2015;123(1):206-211.
doi: 10.3171/2014.9.JNS141001

34. Cabrilo I, Sarrafzadeh A, Bijlenga P, Landis B, Schaller K. Augmented reality-assisted skull base surgery. *Neurochirurgie*. 2014;60(6):304-306.
doi: 10.1016/j.neuchi.2014.07.001
35. Molina CA, Phillips FM, Poelstra KA, Colman M, Khoo LT. A cadaveric precision and accuracy analysis of augmented reality mediated percutaneous pedicle implant insertion. *Spine J*. 2020;20(9):S74.
36. Burström G, Persson O, Edström E, Elmi-Terander A. Augmented reality navigation in spine surgery: A systematic review. *Acta Neurochir (Wien)*. 2021;163:843-852.
doi: 10.1007/s00701-021-04708-3
37. Yuk FJ, Maragos GA, Sato K, Steinberger J. Current innovation in virtual and augmented reality in spine surgery. *Ann Transl Med*. 2021;9(1):94.
doi: 10.21037/atm-20-1132
38. Vadalà G, De Salvatore S, Ambrosio L, Russo F, Papalia R, Denaro V. Robotic spine surgery and augmented reality systems: A state of the art. *Neurospine*. 2020;17(1):88-100.
doi: 10.14245/ns.2040060.030
39. Parsons D, MacCallum K. Current perspectives on augmented reality in medical education: Applications, affordances and limitations. *Adv Med Educ Pract*. 2021;12:77-91.
doi: 10.2147/AMEPS.249891
40. Williams MA, McVeigh J, Handa AI, Lee R. Augmented reality in surgical training: A systematic review. *Postgrad Med J*. 2020;96(1139):537-542.
doi: 10.1136/postgradmedj-2020-137600
41. Muralidharan R. External ventricular drains: Management and complications. *Surg Neurol Int*. 2015;6(Suppl 6):S271-S274.
doi: 10.4103/2152-7806.157620
42. Chau CYC, Craven CL, Rubiano AM, et al. The evolution of the role of external ventricular drainage in traumatic brain injury. *J Clin Med*. 2019;8(9):1422.
doi: 10.3390/jcm8091422
43. Huyette DR, Turnbow BJ, Kaufman C, Vaslow DF, Whiting BB, Oh MY. Accuracy of the freehand pass technique for ventriculostomy catheter placement: Retrospective assessment using computed tomography scans. *J Neurosurg*. 2008;108(1):88-91.
doi: 10.3171/jns.2008.108.01.0088
44. Brattain LJ, Pierce TT, Gjestey LA, et al. AI-enabled, ultrasound-guided handheld robotic device for femoral vascular access. *Biosensors (Basel)*. 2021;11(12):522.
doi: 10.3390/bios11120522
45. Chilamkurthy S, Ghosh R, Tanamala S, et al. Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study. *Lancet*. 2018;392(10162):2388-2396.
doi: 10.1016/S0140-6736(18)31645-3
46. Bevan N. *Classifying and Selecting UX and Usability Measures*. Toulouse, France: Institute of Research in Informatics of Toulouse (IRIT); 2008. p. 13-18.
47. Support A. *About Face ID Advanced Technology*. Available from: <https://support.apple.com/en-us/102381#:~:text=Face%20ID%20works%20best%20when,camera%20can%20see%20your%20eyes> [Last accessed on 2025 Aug 31].
48. Depth estimation Technology in iPhones. Available from: <https://www.opencv.ai/blog/depth-estimation#:~:text=Depth%20in%20iPhone,TrueDepth%20camera%2C%20and%20Scene%20Geometry> [Last accessed on 2025 Aug 31].
49. Sang J, Wu Z, Guo P, et al. An improved YOLOv2 for vehicle detection. *Sensors*. 2018;18(12):4272.
doi: 10.3390/s18124272
50. Kakarla UK, Kim LJ, Chang SW, Theodore N, Spetzler RF. Safety and accuracy of bedside external ventricular drain placement. *Neurosurgery*. 2008;63(1 Suppl 1):ONS162-ONS166; discussion ONS166-ONS167.
doi: 10.1227/01.neu.0000335031.23521.d0
51. Patel EA, Aydin A, Cearns M, Dasgupta P, Ahmed K. A systematic review of simulation-based training in neurosurgery, part 1: Cranial neurosurgery. *World Neurosurg*. 2020;133:e850-e873.
doi: 10.1016/j.wneu.2019.08.262
52. Sakai D, Joyce K, Sugimoto M, et al. Augmented, virtual and mixed reality in spinal surgery: A real-world experience. *J Orthop Surg*. 2020;28(3):2309499020952698.
doi: 10.1177/2309499020952698

BRIEF REPORT

Feasibility of artificial intelligence-driven personalized learning for internal medicine residents: Integrating adaptive artificial intelligence in flipped classrooms

Marcos A. Sanchez-Gonzalez^{1*}, **Noelani-Mei Ascio²**, **Omar Shah²**, **Ashley Matejka³**, **Mark Terrell⁴**, and **Salman Muddassir²**¹LECOM School of Health Services Administration, Bradenton, Florida, United States of America²Internal Medicine Program, HCA Florida Oak Hill Hospital, Brooksville, Florida, United States of America³Research and Development, QHSLab, Inc., West Palm Beach, Florida, United States of America⁴Department of Medical Education, Lake Erie College of Osteopathic Medicine, Erie, Pennsylvania, United States of America***Corresponding author:**Marcos A. Sanchez-Gonzalez
(msanchez-gonzalez@lecom.edu)**Citation:** Sanchez-Gonzalez MA, Ascio N, Shah O, Matejka A, Terrell M, Muddassir S. Feasibility of AI-driven personalized learning for internal medicine residents: Integrating adaptive AI in flipped classrooms. *Artif Intell Health*. 2025;2(4):139-145. doi: 10.36922/AIH025130023**Received:** March 25, 2025**1st revised:** May 23, 2025**2nd revised:** May 30, 2025**3rd revised:** June 4, 2025**4th revised:** June 13, 2025**Accepted:** June 16, 2025**Published online:** July 16, 2025**Copyright:** © 2025 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**

Medical residency training faces persistent challenges in delivering individualized learning experiences. While flipped classroom models promote engagement, they often lack real-time, personalized feedback. Artificial intelligence (AI)-driven platforms offer a promising solution by dynamically adapting content to residents' evolving needs. This study evaluated the feasibility and effectiveness of integrating adaptive AI beings into a flipped classroom model for internal medicine residents. The AI-powered platform, edYOU, incorporated a personalized ingestion engine to customize learning content and an intelligent curation engine to ensure content integrity. Residents interacted with AI beings capable of adjusting real-time content delivery based on performance and progress. Learning outcomes were assessed using platform engagement metrics, simulation-based quiz results, and resident feedback. Among eligible residents, 92% actively used the platform, spending an average of 32.3 h (a few minutes to 148 h). A significant positive correlation was observed between time spent on the platform and quiz performance ($r = 0.63$, $p < 0.001$), with 82.6% of educational topics engaged. Learners spent more time on difficult content areas, highlighting the system's ability to adapt to individual challenges. Integrating AI into the flipped classroom proved feasible and was associated with improved engagement, learning efficiency, and academic performance. These results support using AI-enhanced educational tools to foster tailored, learner-centered experiences in graduate medical education. Further research is warranted to optimize implementation strategies and evaluate the long-term impact of AI-driven learning environments on resident development and competency outcomes.

Keywords: Artificial intelligence; Personalized learning; Internal medicine; Flipped classroom; Residency training; Medical education

1. Introduction

Artificial intelligence (AI) has recently transformed medical education by offering personalized, adaptive learning experiences and content delivery. Traditional medical residency training programs (graduate medical education) face challenges in meeting the diverse learning needs of trainees, as standardized approaches often fail to accommodate individual differences in knowledge acquisition and clinical preparedness.¹ Flipped classroom models, which shift passive learning outside the classroom and emphasize active engagement during instructional sessions, have demonstrated improved learner engagement, comprehension, and new language acquisition.^{2,3} The evolution of AI in education has followed both incremental and disruptive paths, supporting personalized learning through increasingly sophisticated cognitive models that could enhance the flipped classroom methodology.⁴ When enhanced by AI, flipped classroom models further personalize instruction, offer real-time feedback, and significantly accelerate the acquisition of new skills and languages by adapting content to individual learning needs and preferences.^{3,5}

Traditional flipped classroom models, while effective in promoting active learning, often fail to account for the heterogeneity of learner progression, intrinsic cognitive load, and the unpredictable demands of clinical training environments.⁶ In fact, traditional flipped classroom models are inherently limited by their reliance on pre-designed, static, repetitive content and a lack of responsive feedback mechanisms.^{1,7} These models often fail to account for the heterogeneity of learner progression, cognitive load, and the dynamic nature of clinical training environments. As a result, residents may struggle to bridge understanding gaps or receive timely clarification on complex topics, particularly when self-directed study occurs outside of scheduled instructional time. In this context, AI-driven educational platforms offer a compelling advancement by introducing real-time adaptability into the flipped classroom structure. A systematic review of AI applications in higher education reveals widespread use in intelligent tutoring, predictive analytics, and learning support tools.⁸ These systems can analyze user input, track performance patterns, and deliver content dynamically tailored to individual learning trajectories.

Through automated feedback loops and continuous assessment, AI platforms can identify areas of difficulty, reinforce core concepts, and adjust content delivery accordingly – features that are otherwise challenging to implement in traditional didactic or blended formats.¹ While early studies have demonstrated the feasibility and learner satisfaction associated with AI-enhanced

instruction in undergraduate medical education, their application in graduate medical education, particularly within the time-constrained and high-stakes context of residency training, remains underexplored.⁹ Given the increasing demands for competency-based education and personalized learning environments in graduate medical education, integrating AI into residency curricula may represent a transformative step in optimizing knowledge acquisition, supporting just-in-time learning, and ultimately improving patient care outcomes across healthcare systems.¹⁰⁻¹² In this vein, through real-time adaptation, the AI beings provide personalized support, ensuring residents receive targeted reinforcement in areas of difficulty. Prior research has demonstrated that AI-assisted learning enhances engagement, reduces study time, and improves performance in formative as well as summative assessments.^{13,14} Given the increasing demand for competency-based medical education, AI-driven platforms have the potential to bridge gaps in traditional learning methods by offering scalable, data-driven solutions tailored to individual learning needs.

Accordingly, the present study sought to evaluate the feasibility of integrating adaptive AI beings into a flipped classroom model for internal medicine residents. Addressing the aforementioned deficiencies of earlier models, the AI-driven platform used in this study leverages natural language processing and machine learning algorithms to assess learner's progress and optimize instructional delivery. This novel integration aims to advance the pedagogical utility of flipped learning by offering a scalable, data-informed solution that delivers real-time personalization within the demanding context of graduate medical education.

2. Data methods

2.1. Study design and setting

This study employed a feasibility design to evaluate integrating an AI-driven learning platform into an internal medicine residency program. The AI-enhanced flipped classroom model was implemented at HCA Florida Oak Hill Hospital's Internal Medicine residency program, with participation of residents from post-graduate years 1 to 3. The 6-month educational intervention examined resident engagement, learning efficiency, and performance changes in preparation for in-service examinations. Residents were introduced to the AI platform through detailed orientation sessions and continuous access to the platform throughout the study period. The platform provided adaptive learning pathways tailored to each resident's performance. Engagement was monitored through various metrics, including total study time, frequency and duration

of interactions, and completion rates of educational modules. This study also assessed subjective feedback to gauge resident satisfaction and perceived benefits of the AI-driven learning approach. Data collection extended to performance outcomes on simulation-based quizzes, aligning engagement metrics with academic progress.

2.2. AI-driven learning platform

The study utilized the edYOU platform, an adaptive, AI-based educational system designed to personalize learning experiences in graduate medical education. The platform's adaptive content delivery is grounded in key principles of adaptive hypermedia, which personalize learning pathways based on user modeling.¹⁵ The platform also integrates two core components: The personalized ingestion engine (PIE) and the intelligent curation engine (ICE). The PIE technology, which follows adaptive learning principles, continuously curates diverse instructional materials from validated academic sources.¹⁶ It uses natural language processing to tailor content delivery based on each learner's demonstrated knowledge, response patterns, and engagement history. This dynamic tailoring allows the system to adjust the difficulty and sequencing of content in real time, supporting individualized progression through the curriculum. The ICE, in turn, ensures the integrity, relevance, and safety of the educational material by implementing automated content validation protocols, including toxicity filtering, bias mitigation, and source verification. The above-mentioned safeguards were purposely designed to maintain academic rigor while minimizing the risk of misinformation or inappropriate content. Medical residents interacted with AI beings (Figure 1) capable of conducting naturalistic dialogue, providing immediate, context-sensitive feedback, and tracking learner progress through analytics-driven personalization. This combination of adaptive delivery and content governance enabled a responsive, structured learning environment aligned with competency-based medical education principles.

2.3. Implementation and data collection

Medical residents were introduced to the AI-driven flipped classroom model through an initial orientation session that provided an overview of the platform's capabilities, including navigation, interaction protocols with the AI beings, and expectations for independent learning. Following orientation, residents were encouraged to utilize the platform beyond scheduled didactic sessions to supplement their self-directed study. The platform's design allowed for asynchronous interaction, enabling learners to access content, receive feedback, and revisit complex concepts at their own pace. Engagement metrics



Figure 1. Adaptive AI beings on the edYOU platform
Abbreviation: AI: Artificial intelligence.

– including total time spent on the platform, percentage of topics completed, frequency of interactions with AI beings, and depth of follow-up queries – were passively and systematically logged by the platform's analytics infrastructure. These objective measures of engagement were paired with performance on simulation-based formative assessments and standardized in-service examination scores to evaluate learning outcomes. In addition, subjective feedback was gathered through post-intervention surveys, which included quantitative ratings and open-ended questions to assess perceived usefulness, ease of use, and the platform's ability to support individualized learning. This multi-modal evaluation strategy allowed data triangulation across usage, performance, and user perception domains.

2.4. Statistical analysis

Descriptive statistics were employed to summarize resident engagement and platform utilization metrics, including total hours spent on the platform, number of completed modules, and frequency of interactions with AI beings. Pearson correlation coefficients were calculated to assess the association between platform usage and academic outcomes, with quiz performance as the primary dependent variable. Statistical significance was determined a priori at a two-tailed alpha level of $p < 0.05$. All analyses were conducted using the IBM Statistical Packages for the Social Sciences Statistics for Windows, version 28.0 (IBM Corp., Armonk, NY, USA).

3. Results

Data are expressed as mean and standard deviation unless otherwise specified. Descriptive statistics are summarized in Table 1.

Table 1. Summary of descriptive statistics

Variable	Mean	SD	Min	Max
Total hours	15.5	40.91	0.18	124.41
Total correct	10.89	11.52	1	34
Questions attempted	28.44	21.32	6	73
Quiz accuracy	0.31	0.21	0.05	0.71
Average of time per question (s)	658.39	564.05	153	1827.77

Abbreviation: SD: Standard deviation.

A total of 92% of eligible internal medicine residents actively engaged with the AI-driven learning platform over the 6-month study period. Residents spent an average of 32.3 h interacting with the AI beings, with engagement ranging from a few minutes to 148 h (Figure 2). The platform was used most frequently in the evenings, with 78% of participants accessing it thrice weekly, demonstrating successful integration into residents’ study routines.

During a 6-month pilot, the internal medicine residents consistently engaged with the AI-powered learning platform, with 78% accessing it at least thrice weekly, primarily in the evenings. The control chart shows a mean session time of 10.31 min and an upper control limit of 77.08 min, reflecting occasional high-engagement periods.

Analysis of performance outcomes revealed a strong positive correlation between platform usage and quiz performance ($r = 0.63, p < 0.001$), indicating that greater engagement with the AI-driven flipped classroom model was associated with improved knowledge retention and test scores. Residents who dedicated more time to AI-based learning achieved quiz accuracy rates of up to 85%, while those with lower engagement had significantly lower scores (Table 2).

In addition, 82.57% of the educational topics were actively engaged, and residents spent more time on challenging subject areas, suggesting that the AI beings effectively guided individualized learning paths. Subjective feedback from residents indicated that the platform was intuitive, adaptable, and beneficial for reinforcing complex concepts, with many participants expressing a preference for AI-driven learning over traditional self-study methods.

Total time spent on the platform averaged 5.42 h, with ± 1.93 h SEM, showing variability in engagement (Figure 3). Some residents had significantly higher usage, reflecting diverse study behaviors.

4. Discussion

The primary objective of this study was to evaluate the feasibility of integrating an AI-driven personalized learning

Table 2. Pearson correlation matrix of resident engagement and quiz performance variables

Variables	Total hours	Questions attempted	Quiz accuracy
Questions attempted	0.784*		
Quiz accuracy	0.308	0.492	
Average time per question (s)	0.357	0.279	0.889**

Notes: Values represent Pearson correlation coefficients (r); * $p < 0.05$; ** $p < 0.01$.

platform into internal medicine residency training. Our findings indicate that the adaptive AI beings significantly enhanced resident engagement and exam performance. In addition, the strong correlation between platform usage and quiz performance ($r = 0.63, p < 0.001$) underscores the effectiveness of AI-driven education in improving learning outcomes while echoing findings from educational data mining, which links behavioral metrics to learning outcomes.¹⁷ These results suggest that AI-enhanced flipped classrooms can provide personalized, data-driven learning experiences that improve study efficiency while maintaining educational rigor.¹⁸ Given the increasing reliance on competency-based education models in graduate medical training, this study highlights the potential role of AI in optimizing individualized learning pathways and supporting residents and educators in the transition to adaptive, technology-enhanced instruction.^{2,19}

The present findings align with previous research demonstrating the benefits of AI-driven learning in medical education. Studies have shown that AI-based adaptive learning models improve knowledge retention, engagement, and self-directed study habits.^{1,5} The association between AI engagement and quiz performance is consistent with prior work, which identified AI-driven feedback mechanisms as a key contributor to improved assessment scores and learner’s confidence.¹³ Similarly, AI-supported flipped classrooms have enhanced active learning and self-efficacy, reinforcing our study’s conclusions regarding AI-driven personalized learning approaches.²⁰ However, our results diverge from studies that report mixed student reception of AI integration, where concerns regarding algorithmic bias and the accuracy of AI-generated content were noted. These discrepancies highlight the need for continued refinement of AI-assisted learning models, focusing on transparency in how content is curated and personalized. As seen in clinical AI applications, a lack of clarity in algorithmic processes can hinder trust and adoption. Similarly, educational AI systems must ensure explainability and accountability to gain acceptance and deliver equitable outcomes. Clear

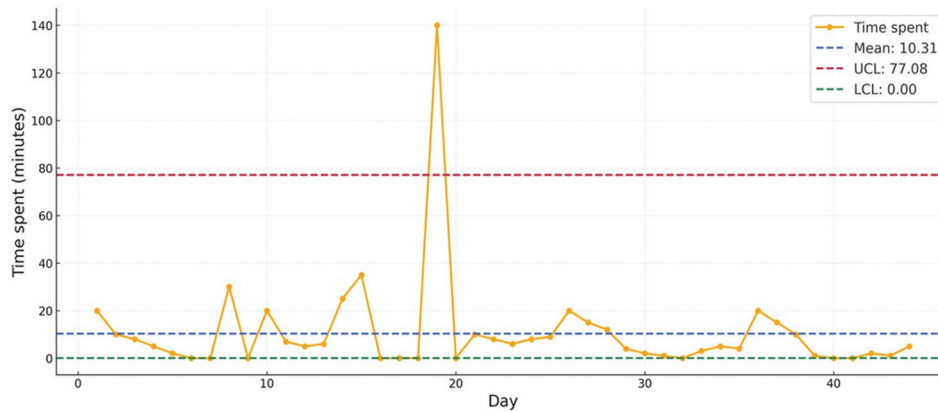


Figure 2. Control chart for time spent on the AI-powered learning platform
Abbreviations: AI: Artificial intelligence; LCL: Lower control limit; UCL: Upper control limit.

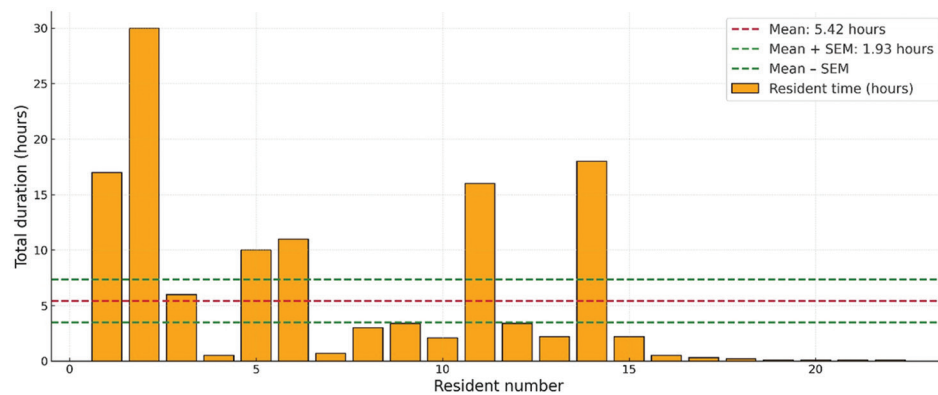


Figure 3. Total duration per resident on the AI-powered learning platform, expressed as mean and SEM, with the outliers removed
Abbreviations: AI: Artificial intelligence; SEM: Standard error of the mean.

governance around content curation is essential for scaling these tools responsibly.²¹

In the present study, we posit that the ICE played a critical role in upholding ethical standards, mitigating misinformation, and ensuring the pedagogical validity of AI-generated educational content. By integrating automated safeguards such as source verification, toxicity filtering, and bias mitigation, ICE directly addresses concerns raised in prior literature regarding the reliability of AI-driven instructional material.^{22,23} This technological advancement represents an ethically sound, meaningful evolution over earlier AI-enhanced learning systems by combining adaptive personalization with structured content governance. In addition, early AI technologies such as Verbot™ demonstrated the potential for digital agents to enhance classroom engagement and delivery.²⁴ As such, the platform supports individualized learning trajectories and reinforces the academic integrity necessary for implementation in medical education.²⁵

While this study provides valuable insights into AI-enhanced medical education, several limitations must be acknowledged. First, the educational research study was conducted within a single internal medicine residency program, without a comparison or control group, potentially limiting the generalizability of findings across different specialties and training environments. Second, although engagement and quiz performance were strongly correlated, long-term educational outcomes, such as board examination performance or clinical decision-making improvements, were not assessed. In addition, resident perceptions of AI-based learning were collected through surveys. However, the study did not include qualitative interviews or focus group discussions, which could have provided more profound insights into learner’s experiences and preferences. Recent studies have shown that generative AI tools can support learner’s motivation and improve knowledge retention in higher education contexts.²⁶ Future research should explore multi-

institutional studies, integrate longitudinal performance tracking, and incorporate resident and faculty perspectives on AI adoption in medical education.

5. Conclusion

Integrating AI-driven adaptive learning platforms presents several opportunities for improving residency training and, potentially, quality of care. Program directors and medical educators should consider incorporating AI-assisted flipped classrooms to supplement traditional didactic instruction, allowing residents to engage with material at their own pace while receiving real-time feedback on areas requiring improvement. In addition, AI platforms should be designed with human oversight mechanisms, ensuring that educational content remains accurate, unbiased, and aligned with competency-based training standards. To further enhance engagement, residency programs should integrate faculty development initiatives that train educators in AI-assisted pedagogical strategies, fostering collaborative learning models that combine AI-driven insights with expert mentoring. As AI technologies evolve, ongoing evaluation of their educational impact, scalability, and ethical use is essential. Future research should also examine how AI-based learning affects long-term clinical performance, interprofessional collaboration, and patient outcomes – areas critical to the future of medical education and healthcare delivery.

Acknowledgment

None.

Funding

The authors declare that edYOU provided the e-learning platform for the study. No external grants or additional financial support were received for this article's research, analysis, or publication.

Conflict of interest

The authors declare that they have no competing interests.

Author contributions

Conceptualization: Marcos A. Sanchez-Gonzalez

Data curation: Noelani-Mei Ascio, Omar Shah, Salman Muddassir

Investigation: Marcos A. Sanchez-Gonzalez, Noelani-Mei Ascio, Omar Shah, Salman Muddassir

Methodology: Ashley Matejka, Mark Terrell

Supervision: Marcos A. Sanchez-Gonzalez

Writing – original draft: Marcos A. Sanchez-Gonzalez

Writing – review & editing: All authors

Ethical approval and consent to participate

This study qualifies for exemption from IRB review under the U.S. Department of Health and Human Services (HHS) Human Subject Regulations Decision Charts. Research involving educational tests, survey procedures, interview procedures, or observation of public behavior, where identifiable information is not recorded or disclosure would not place subjects at risk. As such, this study does not require formal approval by an Institutional Review Board or the acquisition of informed consent, in accordance with federal guidelines.

Consent for publication

This educational research project used only aggregated, fully de-identified assessment data, with no protected health information (PHI), personally identifiable information, or images of participants. As such, this study qualifies for exemption from IRB review under the U.S. Department of Health and Human Services (HHS) Human Subject Regulations Decision Charts.

Availability of data

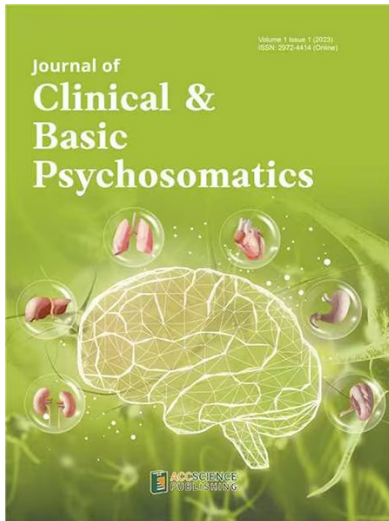
Data are available from the corresponding author upon reasonable request.

References

1. Lo CK, Hew KF. A review of integrating AI-based chatbots into flipped learning: New possibilities and challenges. *Front Educ.* 2023;8:1-7.
doi: 10.3389/educ.2023.1175715
2. Dushyanthen S, Zamri NI, Chapman W, Capurro D, Lyons K. Evaluation of an interdisciplinary educational program to foster learning health systems: Education evaluation. *JMIR Med Educ.* 2025;11:e54152.
doi: 10.2196/54152
3. Dave D, Raval V. *AI-Powered Flipped Classrooms for English Language Learning.* 2024.
4. Roll I, Wylie R. Evolution and revolution in artificial intelligence in education. *Int J Artif Intell Educ.* 2016;26(2):582-599.
doi: 10.1007/s40593-016-0110-3
5. Chan KS, Zary N. Applications and challenges of implementing artificial intelligence in medical education: Integrative review. *JMIR Med Educ.* 2019;5(1):e13930.
doi: 10.2196/13930
6. Barrera Castro GP, Chiappe A, Ramírez-Montoya MS, Alcántar Nieblas C. Key barriers to personalized learning in times of artificial intelligence: A literature review. *Appl Sci.* 2025;15(6):3103.

- doi: 10.3390/app15063103
7. Lin H, Chen Q. Artificial intelligence (AI) -integrated educational applications and college students' creativity and academic emotions: Students and teachers' perceptions and attitudes. *BMC Psychol.* 2024;12(1):487.
doi: 10.1186/s40359-024-01979-0
 8. Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education - where are the educators? *Int J Educ Technol Higher Educ.* 2019;16(1):39.
doi: 10.1186/s41239-019-0171-0
 9. Sanchez-Gonzalez M, Terrell M. Flipped classroom with artificial intelligence: Educational effectiveness of combining voice-over presentations and AI. *Cureus.* 2023;15(11):e48354.
doi: 10.7759/cureus.48354
 10. Vergheze BG, Iyer C, Borse T, Cooper S, White J, Sheehy R. Modern artificial intelligence and large language models in graduate medical education: A scoping review of attitudes, applications & practice. *BMC Med Educ.* 2025;25(1):730.
doi: 10.1186/s12909-025-07321-5
 11. Sriram A, Ramachandran K, Krishnamoorthy S. Artificial intelligence in medical education: Transforming learning and practice. *Cureus.* 2025;17(3):e80852.
doi: 10.7759/cureus.80852
 12. Lee YM, Kim S, Lee YH, *et al.* Defining medical AI competencies for medical school graduates: Outcomes of a delphi survey and medical student/educator questionnaire of South Korean medical schools. *Acad Med.* 2024;99(5): 524-533.
doi: 10.1097/acm.0000000000005618
 13. Birks S, Gray J, Darling-Pomranz C. Using artificial intelligence to provide a 'flipped assessment' approach to medical education learning opportunities. *Med Teach.* 2024;1-8.
doi: 10.1080/0142159x.2024.2434101
 14. Michael J, Rovick A, Glass M, Zhou Y, Evens M. Learning from a computer tutor with natural language capabilities. *Interact Learn Environ.* 2003;11(3):233-262.
doi: 10.1076/ilee.11.3.233.16543
 15. Brusilovsky P, Millán E. User models for adaptive hypermedia and adaptive educational systems. In: Brusilovsky P, Kobsa A, Nejdl W, editors. *The Adaptive Web: Methods and Strategies of Web Personalization.* Berlin, Heidelberg: Springer; 2007. p. 3-53.
 16. Ejjami R. The adaptive personalization theory of learning: Revolutionizing education with AI. *J Next Gener Res.* 2024;1(1):1-18.
doi: 10.70792/jngr5.0.v1i1.8
 17. Baker R, Inventado PS. *Educational Data Mining and Learning Analytics.* Berlin: Springer; 2014. p. 61-75.
 18. Corte-Real A, Nunes T, Caetano C, Almiro PA. Cone beam computed tomography (CBCT) technology and learning outcomes in dental anatomy education: E-learning approach. *Anat Sci Educ.* 2021;14(6):711-720.
doi: 10.1002/ase.2066
 19. Herodotou C, Muirhead DK, Aristeidou M, *et al.* Blended and online learning: A comparative study of virtual microscopy in higher education. *Interact Learn Environ.* 2020;28(6):713-728.
doi: 10.1080/10494820.2018.1552874
 20. Hatwalne PA, Chaudhary SS, Prayagi SV, Adkane RV, Vairagade S. *Comparative investigation of BOPPPS-AI Integrated Flipped Classroom Method and Conventional Teaching Method in Mechanical Engineering Education.* United States: IEEE; 2024. p. 1-5.
 21. Mennella C, Maniscalco U, De Pietro G, Esposito M. Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon.* 2024;10(4):e26297.
doi: 10.1016/j.heliyon.2024.e26297
 22. Yaacoub A, Tarnpradab S, Khumprom P, Assaghir Z, Prevost L, Da-Rugna J. *Enhancing AI-Driven Education: Integrating Cognitive Frameworks, Linguistic Feedback Analysis, and Ethical Considerations for Improved Content Generation.* [arXiv Preprint]; 2025.
 23. Peng J, Shen W, Rao J, Lin J. *Automated Bias Assessment in AI-Generated Educational Content using CEAT Framework.* [arXiv Preprint]; 2025.
 24. Gorby GL. Use of verbot technology to enhance classroom lecture. *Acad Med.* 2001;76(5):552-553.
doi: 10.1097/00001888-200105000-00097
 25. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med.* 2019;25(1):30-36.
doi: 10.1038/s41591-018-0307-0
 26. Monzon N, Hays FA. Leveraging generative artificial intelligence to improve motivation and retrieval in higher education learners. *JMIR Med Educ.* 2025;11:e59210.
doi: 10.2196/59210

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:

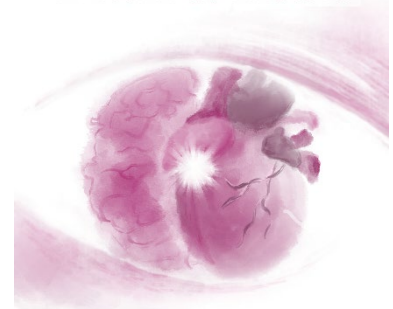
- Conceptualization and classification of psychosomatic medicine
- Mechanism, biological markers, brain images, and treatment studies
- Psychosomatic reactions, syndromes, disorders, and diseases
- Psychosomatic disorders treated in general hospitals, including endocrinology, neurology, gastroenterology, 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.

Brain & Heart



ISSN: 2972-4139 (Online)

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



Contact

www.accscience.com

9 Raffles Place, Republic Plaza 1 #06-00 Singapore 048619

Email: editorial@accscience.com

Phone: +65 8182 1586